Pushing the Limit of Quantum Mechanical Simulation to the Raman Spectra of a Biological System with 100 Million Atoms

Honghui Shang[†*], Ying Liu[‡*], Zhikun Wu[‡], Zhenchuan Chen[‡], Jinfeng Liu[§], Meiyue Shao[¶],
Yingzhou Li[||], Bowen Kan[†], Huimin Cui[‡], Xiaobing Feng[‡], Yunquan Zhang[†‡],
Donald G. Truhlar[††], Hong An[†], Xiao He[‡‡*], Jinlong Yang[†*]

[†]Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, China

[‡]State Key Lab of Processors, Chinese Academy of Science; University of Chinese Academy of Sciences, Beijing, China

[§]School of Science, China Pharmaceutical University, Nanjing, China

[¶]School of Data Science and MOE Key Laboratory for Computational Physical Sciences, Fudan University, Shanghai, China

[||]School of Mathematical Sciences and MOE Key Laboratory for Computational Physical Sciences, Fudan University, Shanghai, China

[†‡]Chinese Academy of Science; University of Chinese Academy of Sciences, Beijing, China

[††]Department of Chemistry, University of Minnesota, Minneapolis, MN, USA

[‡‡]Shanghai Frontiers Science Center of Molecule Intelligent Syntheses,
School of Chemistry and Molecular Engineering, East China Normal University, Shanghai, China.

[*]Corresponding authors

***Abstract*—Raman spectroscopy offers invaluable insights into the chemical composition and structural characteristics of various materials, making it a powerful tool for structural analysis. However, accurate quantum mechanical simulations of Raman spectra for large systems, such as biological materials, have been limited due to immense computational costs and technical challenges. In this study, we developed efficient algorithms and optimized implementations on heterogeneous computing architectures to enable fast and highly scalable *ab initio* simulations of Raman spectra for large-scale biological systems with up to 100 million atoms. Our simulations have achieved nearly linear strong and weak scaling on two cutting-edge high-performance computing systems, with peak FP64 performances reaching** 400 **PFLOPS on** 96,000 **nodes of new** Sunway **supercomputer and** 85 **PFLOPS on** 6,000 **node of** ORISE **supercomputer. These advances provide promising prospects for extending quantum mechanical simulations to biological systems.**

***Index Terms*—Raman spectra, all-electron quantum perturbation simulation, heterogeneous architectures, scalability.**

I. Justification for ACM Gordon Bell Prize

We performed unprecedentedly full *ab initio* Raman spectra simulations of real biological system with up to 100 million atoms, which is $> 10000\times$ improvement w.r.t state-of-the-art quantum mechanical methods. Double precision performance of 399.9 PFLOPS is achieved on 96,000 nodes of new Sunway.

II. Performance Attributes

Category of achievement	Scalability peak performance
Type of method used	Non-linear equations, Explicit
Results reported on the basis of	Whole application including I/O
Precision reported	Double Precision
System scale	Measured on Full System
Measurement mechanism	Timer and FLOP count

III. Overview of the Problems

Every year, modern progress in molecular biology and medicinal chemistry relies more and more on simulation. We can take proteins as an illustrative example. Protein function underlies almost all biological function. X-ray structures of proteins give hints about the structure-function relationships that underlie health and medicine, but simulation provides deeper insights. The importance of simulation in protein science was recognized by the award of the 2013 Nobel Prize in chemistry to Karplus, Levitt, and Warshel *"for the development of multiscale models for complex chemical systems. . . . They managed to make Newton's classical physics work side-by-side with the fundamentally different quantum physics. Previously, chemists had to choose to use either or. The strength of classical physics was that calculations were simple and could be used to model really large molecules. Its weakness, it offered no way to simulate chemical reactions. For that purpose, chemists instead had to use quantum physics. But such calculations required enormous computing power and could therefore only be carried out for small molecules."* [1] The specific advance cited in the prize-winning work was the development of combined quantum mechanics and molecular mechanics (QM/MM). In QM/MM one carries out calculations for chemical reactions in proteins, but QM is used only for a smaller subsystem, and the rest of the large system is treated by classical methods. The next grand challenge is to apply QM to the entire system, that is the area of the present work.

Raman spectroscopy is increasingly being applied to biological systems [2] because it can provide chemical and composition information for proteins in essentially all physical states, and it can probe the structural changes in proteins that result from protein–ligand interactions. The distinctive inelastic scattering observed in Raman spectroscopy results from the interaction between light and the vibrations of chemical bonds. The spectral data derived from this interaction allow for the identification of material compositions and structures. Raman spectroscopy has several desirable features, including non-destructive sampling, rapid molecular detection, and sensitivity to subtle structural variations. It has applications in physics, chemistry, materials science, and biomedical research. As shown in Fig. 1, Raman spectroscopy emerges as a potent tool for characterizing biological molecules without the need for labeling. It enables the extraction of biochemical and structural information, facilitating the identification of entire biochemical processes, such as metabolic pathways and dynamics.

Biological systems are highly complex, and their Raman spectra can be influenced by various factors such as conformational changes, interactions with surrounding molecules, and environmental conditions. Molecular dynamics simulations, which are often used to simulate biomolecular systems, conventionally rely on generally-parameterized force fields to describe the interactions between atoms. However, general parametrization may be inaccurate for important site-specific interactions and polarization effects, leading to discrepancies between simulated and experimental spectra. This motivates the development of improved computational methods and models, and here we present a parameter-free *ab initio* approach that involves a direct connection between atomic structure and detailed spectral characteristics and provides a way to enhance our fundamental understanding of atomic and molecular interactions in complex systems. *Ab initio* simulation of Raman spectra necessitates calculating up to the third-order derivative of the electronic energy, involving simultaneous consideration of electric-field and atomic-displacement perturbations. For very large systems, density-functional theory (DFT) [3] is the most practical approach offering high accuracy, and we here apply *ab initio* density-functional perturbation theory (DFPT) [4–6] for direct simulation of Raman spectroscopy. Biological systems are complicated not only by large size but also by solvent effects because biological processes take place in aqueous solution. Capturing these effects in simulations adds another layer of complexity. In this study, we solvate a realistic protein system with an explicit water box for accurate QM simulation of its Raman spectrum including solvent effects. The inclusion of explicit water molecules significantly increases the size of the system.

Accurate and efficient QM calculations on large systems containing up to thousands of atoms are a grand challenge. The number of arithmetic operations in an application of *ab initio* DFT/DFPT to a large system nominally increases as the third power of system size, and this is the main reason why calculations have typically been limited to small systems. However, recent years have seen the emergence of efficient methods known as fragmentation approaches for applying QM to large systems. These methods are grounded on the principle of "chemical locality", which — when applied to a macromolecule — posits that a local region of the molecule is only weakly influenced by distant atoms. Leveraging this principle yields the Quantum Fragmentation (QF) algorithm that involves dividing the macromolecule into smaller subsystems or fragments. The total energy or molecular properties of the entire system can then be obtained by combining contributions from individual fragments in an appropriate fashion.

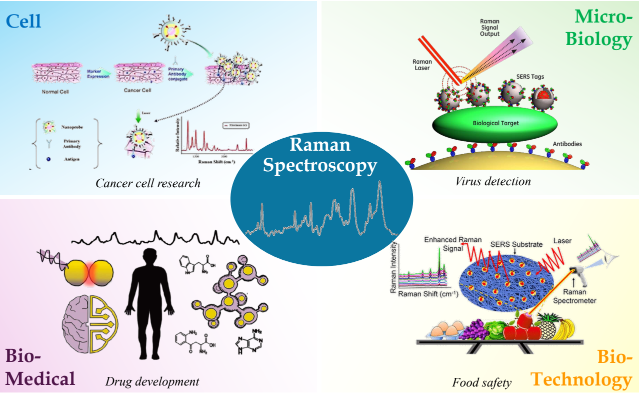


Fig. 1. Examples of the application of Raman spectroscopy for chemical biology research.

To make the calculations practical for large biological systems, the QF algorithm must be implemented efficiently across various supercomputers. In this study, we present a robust new algorithm and a massively parallel implementation of this method in the QF-Raman program, enabling full QM calculations of Raman spectra to scale up to an unprecedented level of 100 million atoms. This implementation can be executed on two typical modern supercomputers with different architectures, i.e., the ORISE supercomputer equipped with GPUs, and the new-generation Sunway supercomputer equipped with many-core accelerators. Major advances are:

- A highly scalable, large-scale QF-Raman implementation has been developed, introducing a three-level parallelization hierarchy that enables very large biological fragments to be served by available processes in an organized yet low-overhead manner.
- Code portability has been achieved with OpenCL, i.e., a cross-platform programming framework. With OpenCL, QF-Raman can execute across various supercomputers equipped with different heterogeneous accelerators.
- Load balance has been achieved among various-sized fragments (sub-systems) with a system-size-sensitive packing strategy that is capable of dynamically packing various fragments into a task and adjusting the task granularity according to the current workload.
- Efficient workload offloading has been achieved with an elastic scheme that gathers and packs together scattered computationally-intensive calculations into a profitable workload to be offloaded to accelerators, since the execution time of each such calculation is too short for offloading alone. To achieve higher hardware utilization on accelerators, calculations are elastically packed according to their computational strength.
- Performance improvements of each fragment simulation

have been achieved by leveraging symmetry in the linear algebra steps, to reduce the number of BLAS operations.

- We have evaluated QF-RAMAN on two typical supercomputers using real biological system with up to 100 million atoms, and we compared the results with the experiment. Full systems are used on both supercomputers: the ORISE supercomputer with up to 24,000 GPUs, and the new-generation Sunway supercomputer with 96,000 many-core accelerators (i.e., 37,440,000 cores).

The optimization techniques outlined in this work are transferable to other DFT/DFPT codes exhibiting similar computational characteristics.

IV. CURRENT STATE OF THE ART

As shown in Fig. 2, various physical properties can be evaluated within the QM framework, using density functional theory (DFT), the total energy of the ground state ($\mathrm{E}^{(0)}$) and the force ($\mathrm{E}^{(1)}$, energy first derivative) on individual atoms can be evaluated. However, to directly relate physical properties to experimental observations, one needs $\mathrm{E}^{(2)}$ (energy second derivative) needed for calculating polarizability and phonon spectroscopy) and $\mathrm{E}^{(3)}$ (energy third derivative needed for Raman Spectroscopy). Theses higher derivatives, describe the system's response to external perturbations, and they can be calculated by density-functional perturbation theory (DFPT). Several computational codes have been developed to perform DFT and DFPT calculations, each employing different basis functions or discretization schemes. Plane wave (PW) basis function-based codes, such as Quantum ESPRESSO [7], VASP [8], ABINIT [9], and Qbox [10] (2006 Gordon Bell prize) are widely used. All-electron Gaussian atomic orbital-based methods are implemented in codes like Gaussian [11], CRYSTAL [12], and the all-electron numerical atomic orbital method in DMol [13] and FHI-aims [14], which can offer low-order scaling calculation, while needing a large number of basis functions to achieve results that are close to the complete basis set limit. Uniform real-space grid-based codes include Octopus [15], RSDFT [16] (2011 Gordon Bell Prize) and DFDFT [17] (2016 Gordon Bell finalist). Finite elements basis functions are adopted in DFT-FE [18,19] (2019 Gordon Bell finalist and 2023 Gordon Bell Prize). Mixed basis codes have also been developed, examples as ONETEP [20], which utilizes periodic sinc functions; BigDFT [21], which combines wavelets with localized support functions; CONQUEST [22], which uses B-spline functions with localized support functions (SF); and CP2K [23], which combines Gaussian and PWs. Both CONQUEST and CP2K have achieved simulations with 1 million atoms [22,23]. In 2023, CP2K [24] pushed the limit of DFT energy and force calculations to 83 million atoms by using a non-orthogonal local submatrix method.

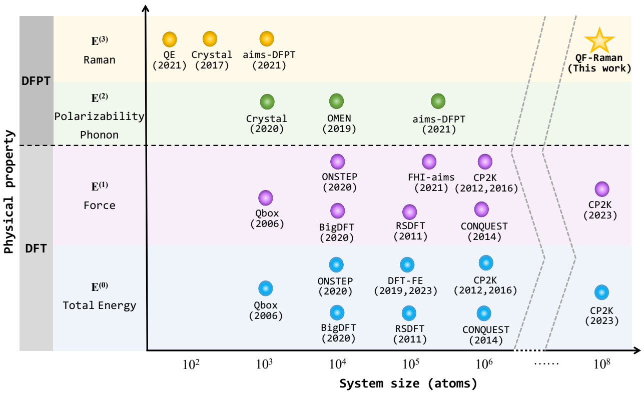


Fig. 2. The accessible system size at various levels of increasingly energy derivatives in *ab initio* electronic structure calculations.

On the other hand, accurate and efficient density functional perturbation theory (DFPT) calculations on large systems remain a significant challenge. In quantum mechanical perturbation theory, the introduced perturbations can disrupt the boundary conditions of periodic systems, and atomic displacements lead to changes in the entire basis set. Consequently, constructing the related matrix elements becomes very complicated, making DFPT optimizations much more intricate than DFT. As a result, DFPT based Raman spectroscopy simulations have been limited to systems with only hundreds of atoms [25], while the aims-DFPT has pushed the Raman spectra simulation limit to 3,000 atoms [26] (2021 Gordon Bell finalist). To further extend the reach of DFPT calculations, Quantum Fragmentation (QF) algorithms [27,28] can be employed for applying QM to large systems. The total energy and molecular properties of the entire system can then be obtained by appropriately combining contributions from individual fragments. For example, the fragment molecular orbital (FMO) method has enabled correlated calculations of 146,592 atoms at the second-order Møller–Plesset perturbation theory (MP2) level using 27,600 GPUs [28]. These fragmentation techniques offer a promising avenue for extending the applicability of DFPT to larger systems and enabling accurate and efficient calculations of their properties. Overall, the current state-of-the-art electronic structure methods are either limited by the order of energy derivatives that can be computed or the scalability and size of the simulations that can be performed. Many existing approaches are limited to calculating only zero and first order energy derivatives, which restricts their applicability to a narrow range of physical properties. Conversely, methods capable of computing higher-order derivatives often suffer from poor scalability and, thus are unsuitable for large-scale simulations of complex systems. To address these limitations, we have developed a novel framework that integrates the QF approach with DFPT, which allows us to circumvent the limitations of conventional electronic structure methods and access a wider range of molecular properties, specifically those relevant to Raman spectroscopy.

A. *Theory of the* QF-RAMAN *simulation*

In the QF-RAMAN approach, the protein is divided into amino acid-based fragments by cutting through each peptide bond except the first and the last; as a result, a protein with N amino acids is cut in $N-3$ places, generating $N-2$ naked residues. Each naked residue a_k is capped on each side with the nearest residue of the formerly connected residues on that side; these two residues are called conjugate caps and are denoted as Cap^*_{k-1} and Cap_{k+1}. Fragment k is then defined as $\mathsf{Cap}^*_{k-1}a_k\mathsf{Cap}_{k+1}$. Since each cap appears twice, one also forms $\mathsf{Cap}^*_k\mathsf{Cap}_{k+1}$ and subtracts its energy to avoid double counting. Hydrogen atoms are added to terminate all dangling bonds. Each water molecule is a fragment. In addition, "generalized concaps" are introduced to include the two-body QM interaction energies between sequentially non-neighboring residues that are spatially in close contact, in particular, if the minimal distance between any two atoms among the two non-neighboring fragments i and j is within a predefined distance threshold λ (in this work λ is set to 4 Å).

Using the DFPT method, the Raman spectra can be calculated through harmonic approximation [6,29]. The second derivative of the total energy with N amino acids and N_{GC} generalized concaps with respect to nuclear coordinates $\mathbf{r}_I$ and $\mathbf{r}_J$ can be calculated [30] by equation Eq. (1),

$$\begin{aligned}
E^{(2)} \approx & \sum_{k=1}^{N-2} E^{(2)}_{\mathrm{F}_k} - \sum_{k=1}^{N-3} E^{(2)}_{\mathrm{CC}_k} + \sum_{k=1}^{M} E^{(2)}_{w_k} \\
& + \sum_{\substack{k=1,\ i,j>i+2,\\ |\mathbf{r}_i-\mathbf{r}_j|\le\lambda}}^{N_{GC}} \left(E^{(2)k}_{ij} - E^{(2)k}_{i} - E^{(2)k}_{j}\right) \\
& + \sum_{\substack{k=1,\\ |\mathbf{r}_{a_i}-\mathbf{r}_{w_j}|\le\lambda}}^{M_{aw}} \left(E^{(2)k}_{a_i w_j} - E^{(2)k}_{a_i} - E^{(2)k}_{w_j}\right) \\
& + \sum_{\substack{k=1,\\ |\mathbf{r}_{w_i}-\mathbf{r}_{w_j}|\le\lambda}}^{M_{ww}} \left(E^{(2)k}_{w_i w_j} - E^{(2)k}_{w_i} - E^{(2)k}_{w_j}\right), \qquad (1)
\end{aligned}$$

where $E^{(2)}$ refers to $\frac{\partial^2 E}{\partial \boldsymbol{r}_I \partial \boldsymbol{r}_J}$, E_{F_k} denotes the self-energy of the fragment k, E_{CC_k} is the energy of $\mathsf{Cap}^*_k\mathsf{Cap}_{k+1}$, E_{w_k} is the one-body of each water molecule k, $E^k_{ij} - E^k_i - E^k_j$ is the two-body QM interaction energy between residues i and j in the kth generalized concap, M is the number of water molecules,M_{aw} is the number of two-body interactions between residues and water molecules that are within the distance threshold λ, $E^k_{a_i w_j} - E^k_{a_i} - E^k_{w_j}$ is their interaction energy, M_{ww} is the number of two-body interactions between pairs of water molecules that are within the distance threshold λ, $E^k_{w_i w_j} - E^k_{w_i} - E^k_{w_j}$ is their two-body interaction energy. Then, vibrational frequencies and their corresponding normal modes can be obtained by diagonalizing the mass-weighted Hessian matrix ($\mathbf{H} = \frac{1}{\sqrt{M_I M_J}}\frac{\partial E^2}{\partial \mathbf{r}_I \partial \mathbf{r}_J}$) of the whole system. The polarizability α can also be derived [31] with Eq. (1) by defining $E^{(2)}$ with α. By applying the chain rule, the derivatives of the polarizability with respect to the normal coordinates ($Q_p = \sum_{I=1}^{N_{\text{atom}}}\sum_{j=x,y,z}(e_{Ij,p}/\sqrt{M_I})r_{Ij}$, where $e_{Ij,p}$ is the eigenvector for the mass-weighted Hessian matrix) can be obtained using the equation below,

$$\begin{aligned}
\frac{\partial\alpha}{\partial Q_p} &= \sum_{I=1}^{N_{\text{atom}}}\sum_{j=x,y,z}\frac{\partial\alpha}{\partial\xi_{Ij}}\frac{\partial\xi_{Ij}}{\partial Q_p} \qquad (1 \le p \le N_f) \\
&= \sum_{I=1}^{N_{\text{atom}}}\sum_{j=x,y,z}\frac{\partial\alpha}{\partial\xi_{Ij}}e_{Ij,p}, \qquad (2)
\end{aligned}$$

where N_{atom} is the total number of atoms, N_f is the number of vibrational degrees of freedom, and ξ_{Ij} is the mass-weighted cartesian coordinate defined as

$$\xi_{Ij} = \sqrt{M_I}r_{Ij}, \qquad (j = x, y, z) \qquad (3)$$

here $\partial\alpha/\xi_{Ij}$ can be obtained as a linear combination of corresponding derivatives of fragment properties in a similar fashion. Using these derivatives, the orientation-averaged Raman intensities contributed from pth eigenvalue can be calculated by [32]:

$$R_p \propto \frac{3}{2}\Big(\sum_{i=x,y,z}\frac{\partial\alpha^{ii}}{\partial Q_p}\Big)^2 + \frac{21}{2}\sum_{i,j=x,y,z}\Big(\frac{\partial\alpha^{ij}}{\partial Q_p}\Big)^2. \qquad (4)$$

The polarizabilities (α) are computed with the response density evaluated by the DFPT implemented in FHI-aims [6,29].

B. Algorithmic Challenges

The QF-RAMAN algorithm stated in Section IV-A, introduces new challenges arising from its fragmentation approach.

The load balance challenge arises from subsystem size variation, which leads to significant differences in simulation time among fragments. For example, decomposing the SARS-CoV-2 spike protein (shown in Fig. 7) generates fragments whose sizes range from 9 to 68 atoms, resulting in computational cost differences of a factor of 19×. Therefore, balancing fragments with various sizes among massive parallel processes is a major challenge in the QF-RAMAN simulations.

The heterogeneous-acceleration challenge arises because, although the fragment calculations are nominally easy to be parallelized, the small fragment sizes lead to greatly shortened execution times such that heterogeneous acceleration would not be profitable due to non-neglectful overheads. For example, evaluating the Hamilton matrix of a medium-sized (e.g., 40-atom) fragment spends 85% of its execution time on 2,400 GEMM invocations that are scattered among other calculations, with each GEMM executing for only ~ 0.01 CPU seconds, whose computational strength is far too small to fully utilize heterogeneous computing power, such as GPU. Therefore, packing computationally-intensive calculations for heterogeneous acceleration is another challenge.

The large-scale Raman spectra solver challenge arises, since traditional methods for the calculation of Raman spectra rely on diagonalizing the mass-weighted Hessian matrix, which becomes computationally infeasible for systems with hundreds of millions of degrees of freedom. For example, a 100-million-atom system would require diagonalizing a 300 million by 300 million matrix, far beyond current computational capabilities. This limitation has impeded accurate Raman predictions for large biomolecules and complex materials.

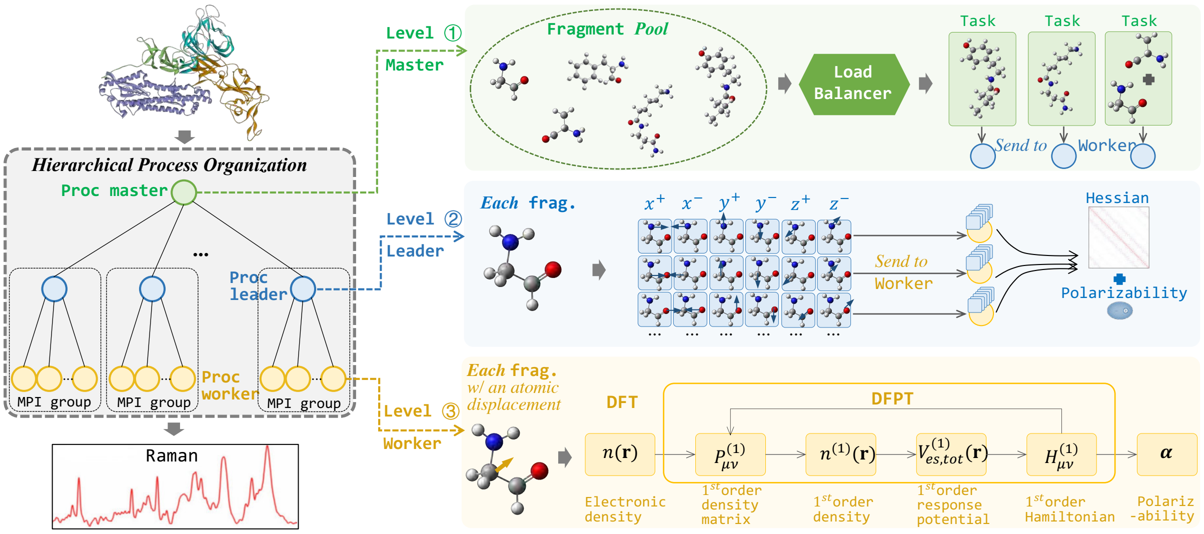


Fig. 3. Hierarchical parallelization framework with a three-level process organization of master–leader–worker.

V. Innovations

QF-Raman involves several innovations for efficient Bio-Raman simulation on modern supercomputers. Section V-A describes a hierarchical parallelization framework, enabling massive fragments to be served by available processes in an organized yet low-overhead manner. Section V-B introduces a system-size-sensitive load balance strategy, capable of dynamically adjusting task granularity according to the system size of un-processed fragments, to overcome the load balancing challenge described in Section IV-B. Section V-C proposes an elastic workload offloading scheme, enabling scattered computationally intensive calculations with short execution times to be gathered and packed into workloads that are profitable for offloading to accelerators, to overcome the heterogeneous-acceleration challenge described in Section IV-B. Section V-D presents a way to use symmetry in the linear algebra steps to improve the per-fragment computational efficiency by reducing the number of BLAS operations. Section V-E proposes a novel approach that avoids full matrix diagonalization, enabling Raman calculations for unprecedentedly large systems. Our method reformulates the Raman spectral intensity expression and uses advanced numerical techniques to reduce computational costs while maintaining accuracy.

Innovations above are independent of the accelerator micro-architecture, thereby can be easily applied across various supercomputers, ensuring the code portability of QF-Raman.[1]

A. Hierarchical Parallelization Framework

QF-Raman uses a three-level process hierarchy of master–leader–worker, as shown in the left part of Fig. 3.

The first level, i.e., the master process, is responsible for decomposing the input protein into fragments and packing them into tasks to be distributed to leader processes, as colored green in Fig. 3. As stated in Section IV-B, the major challenge for the master is the difficulty of balancing parallelized workloads of subsystems (i.e., fragments), whose computational strength may vary greatly due to differences in system sizes. QF-Raman embeds a load balancer in the master, which packs fragments into tasks with a system-size-sensitive load balancing strategy (details in Section V-B).

The second level, i.e., the leader process, is responsible for generating a set of atomic displacements for a given fragment (received from the master) and invoking its workers. As colored blue in Fig. 3, all atomic displacements are equally partitioned and statically assigned to workers, since the computational strength of a given fragment does not change when one of its atoms displaced. The Hessian matrix and polarizability of the given fragment, are obtained after the leader collects all results from its workers.

The third level, i.e., the worker process, is responsible for applying DFT and DFPT simulation on a fragment with a given atomic displacement, as colored orange in Fig. 3. It includes four time-consuming phases: calculation of response density matrix $(P^{(1)}_{\mu\nu})$, real-space integration of the response density $(n^{(1)}(\mathbf{r}))$, Poisson solver for the response potential $(v^{(1)}_{es,tot}(\mathbf{r}))$, and calculation of response Hamiltonian $(H^{(1)}_{\mu,\nu})$, and some of them can be significantly accelerated with heterogeneous accelerators such as GPUs.

B. System-Size-Sensitive Load Balance

The master balances workloads among leaders by dynamically adjusting task granularity according to the system size of un-processed fragments. Fig. 4(a) illustrates the workflow of the load balancer, which receives two types of signals, namely fragment-completed signal with $frag\_id$ and leader-available signal with $leader\_id$, from each leader, and sends task assignment signals with $task\_id \rightarrow leader\_id$

[1] QF-Raman inherits the accelerator-code portability from a previous work [33], which has re-written aims-DFPT [26] using OpenCL [34], i.e., a cross-platform heterogeneous programming framework.

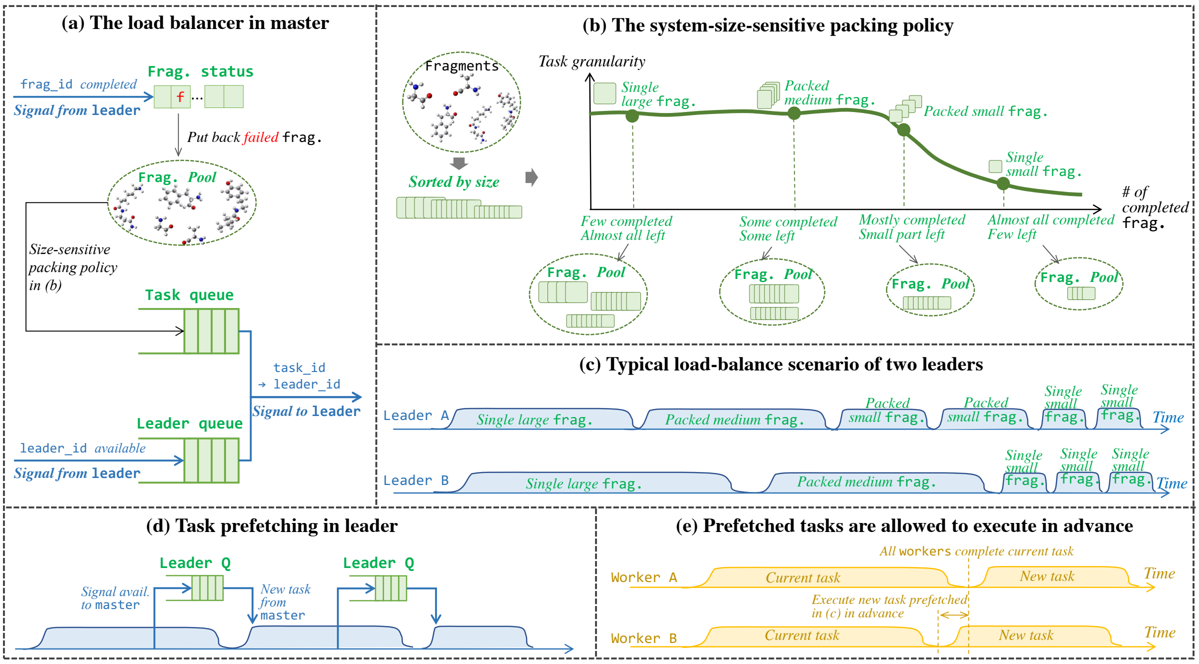


Fig. 4. System-size-sensitive load balancing strategy to concurrently process fragments with a large subsystem size variation.

to leaders. Upon receiving a fragment-completed signal, the master changes the status of $frag\_id$ from "processing" to "completed", and fragments processed for a long time but not yet completed are marked as "un-processed" again, and put back into the pool of un-processed fragments. Upon receiving a leader-available signal, the master pushes its $leader\_id$ in the leader queue, waiting for task assignment. A task queue is used for keeping all ready tasks, with each task generated by picking several fragments from the un-processed pool and packing them together under a system-size-sensitive policy shown in Fig. 4(b), which shows how the system-size-sensitive packing policy packs fragments into tasks. First, the master process sorts the fragments according to their sizes, and it treats each large fragment as a task, since larger fragments require a longer time for calculation. After all large fragments have been sent to execute, the master process packs several medium-sized fragments together as a single task, to avoid frequent communication to leaders. When there are only a few small fragments left, the master gradually reduces the task granularity by decreasing the number of packed fragments until it reaches the minimum granularity; this generates a set of fine-grained tasks to adjust unbalanced leader workloads. In this way, lightly-loaded leaders (e.g., leader A in Fig. 4(c)), which are probably idle at the beginning of the processing of small fragments, are assigned with relatively larger tasks (i.e., packed small fragments). In contrast, heavily-loaded leaders (e.g., leader B in Fig. 4(c)), which are probably busy with their previous tasks until only very few small fragments are left, are assigned with very small tasks (i.e., a single small fragment) so that they may finish at a similar time as those previously lightly-loaded leaders.

Additionally, each leader prefetches its next task to reduce the inter-task idle time, as shown in Fig. 4(d). While a current task is executing, a leader may signal the master with its ability to queue for its next task in advance so that its next task will be assigned upon the completion of the current one. In this case, workers which have completed the current task are allowed to continue with the prefetched new task, while other workers are still busy with the current task, as in Fig. 4(e).

C. Elastic Workload Offloading

We have developed an efficient workload offloading strategy, capable of elastically packing together a few scattered short calculations into a single workload that is profitable for offloading to accelerators, according to the computational strength of those calculations. This has been achieved with a set of loop transformations, and Fig. 5 showcases a typical scenario in which a loop contains both CPU-friendly calculations and GEMM that is accelerator-friendly (denoted as orange circle and purple hexagon respectively) in its body, so that massive GEMM invocations are scattered among CPU-friendly calculations, with each GEMM (i.e., from a single loop iteration) executed for an extremely short time. To generate a workload that is profitable for offloading, QF-RAMAN first strip-mines a loop into a set of strips. In each strip, the intermediate data between the CPU-friendly calculations and multiple GEMM invocations (denoted as gray rectangular) are privatized for each loop iteration in the strip, for the purpose of distributing the CPU-friendly calculations into a CPU-loop and gathering the scattered GEMM invocations into an offloading-loop in this strip. Finally, GEMM invocations in the offloading-loop are elastically batched into several workloads (i.e., *batched-*

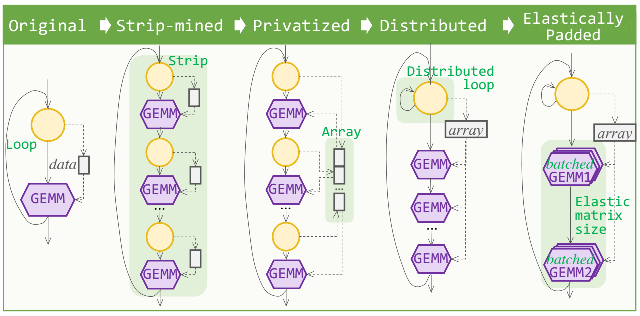
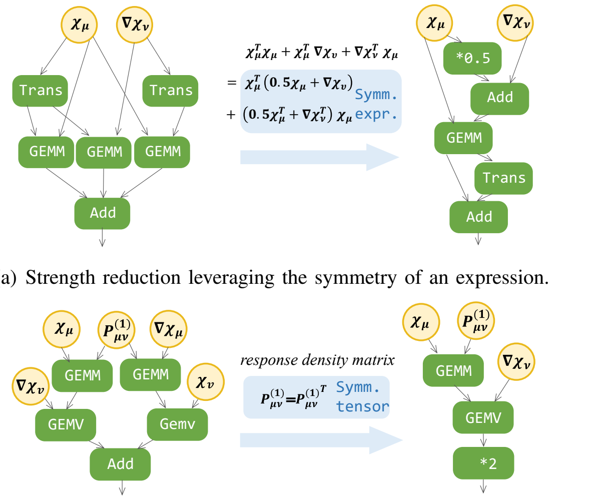


Fig. 5. Elastic workload offloading is capable of packing scattered calculations into a profitable workload, according to their computational strength.



(a) Strength reduction leveraging the symmetry of an expression.

(b) Strength reduction leveraging the symmetry of a tensor.

Fig. 6. Data-flow graphs before/after symmetry-aware strength reduction.

GEMM), with GEMM of similar computational strength (i.e., matrix sizes) batched together.

D. Symmetry-Aware Strength Reduction

We have reduced the computational strength of a few time-consuming integrations, which typically require the sequential invocation of several BLAS operators (e.g., GEMM), by using symmetry in expressions or tensors as shown in Fig. 6. In Fig. 6(a), the expression $\chi_\mu^\mathsf{T}\chi_\mu + \chi_\mu^\mathsf{T}\nabla\chi_\nu + \nabla\chi_\nu^\mathsf{T}\chi_\mu$ needs to be evaluated for computing the first-order Hamiltonian matrix $H_{\mu\nu}^{(1)}$, requiring three GEMM invocations. The expression can equivalently transformed into a symmetric expression that adds two matrices $\chi_\mu^\mathsf{T}(\chi_\mu/2 + \nabla\chi_\nu)$ and $(\chi_\mu^\mathsf{T}/2 + \nabla\chi_\nu^\mathsf{T})\chi_\mu$ which are transposes, reducing the computational strength to only one GEMM (for computing one of the two matrices, $\chi_\mu^\mathsf{T}(\chi_\mu/2 + \nabla\chi_\nu)$ in our case). In particular, the matrix sizes of the GEMM after optimization, remain the same as those of each GEMM before optimization, thereby the computational strength for evaluating this expression, has been reduced by 2/3. In Fig. 6(b), the expression $\chi_\mu P_{\mu\nu}^{(1)}\nabla\chi_\nu + \nabla\chi_\mu P_{\mu\nu}^{(1)}\chi_\nu$ needs to be evaluated to calculate the gradient of response density $\nabla\rho^{(1)}$, involving 2 GEMM's and 2 GEMV's. Given that the response density matrix $P_{\mu\nu}^{(1)}$ is symmetric and χ_μ, χ_ν are transposes, the 2 GEMM's and 2 GEMV's are reduced to 1 GEMM and 1 GEMV.

E. The Efficient Solver for Raman Spectra

Conventional methods for calculating Raman spectra require the computation of all eigenvectors, which is computationally infeasible for large-scale matrices. To overcome this challenge, we rewrite the Raman spectra intensity as

$$I \propto \sum_p \left(\delta(\omega - \omega_p) \left(\frac{\partial \alpha}{\partial Q_p} \right)^2 \right) = \mathbf{d}^\mathsf{T} \delta(\omega - \mathbf{H}) \mathbf{d}, \qquad (5)$$

where ω represents the frequency along the x-axis of the spectrum, and ω_p's refer to the eigenfrequencies of the system. Here $\mathbf{d} = \frac{\partial \alpha}{\partial \xi_{Ij}}$ is the derivative of polarizability vector, and $\mathbf{H}$ is the mass-weighted Hessian matrix ($\mathbf{H} = \frac{1}{\sqrt{M_I M_J}} \frac{\partial E^2}{\partial \mathbf{r}_I \partial \mathbf{r}_J}$), α represents any component of the polarizability tensor α^{ij}, where i and j can each be x, y, or z. We solve this problem using the Lanczos algorithm incorporated with the generalized averaged Gauss quadrature (GAGQ) technique [35]. A k-step Lanczos procedure with the initial vector $\mathbf{q}_1 = \mathbf{d}/|\mathbf{d}|$ produces

$$\mathbf{H}[\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k] = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k]\mathbf{T}_k + [\text{rank } 1], \qquad (6)$$

where $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$ are orthonormal vectors, and $\mathbf{T}$ is a $k \times k$ symmetric tridiagonal matrix. Then for any smooth function $f(\cdot)$, the matrix functional $\mathbf{d}^\mathsf{T} f(\mathbf{H})\mathbf{d}$ can be approximated by

$$\mathbf{d}^\mathsf{T} f(\mathbf{H})\mathbf{d} \approx |\mathbf{d}|^2 \big(f(\mathbf{T})\big)_{1,1}, \qquad (7)$$

i.e., the $(1,1)$-entry of a $k \times k$ matrix $f(\mathbf{T})$ scaled by $|\mathbf{d}|^2$. The $k \times k$ matrix $f(\mathbf{T})$ is sufficiently small so that it is computed by diagonalizing $\mathbf{T}$. The dominating cost of the Lanczos algorithm is k sparse matrix–vector multiplications with $\mathbf{H}$. When the GAGQ technique is incorporated, $f(\mathbf{T})$ is replaced by a $(2k-1) \times (2k-1)$ matrix $f(\widehat{\mathbf{T}})$, which is also computed by diagonalizing the augmented matrix $\widehat{\mathbf{T}}$. We refer to [35,36] for details. The Lanczos algorithm with GAGQ is more accurate than the standard Lanczos algorithm, with negligible additional cost. Therefore, we use this algorithm to compute the Raman spectra. In our setting, we simply set

$$f(\mathbf{H}) = g_\sigma(\omega - \mathbf{H}) \approx \delta(\omega - \mathbf{H}), \qquad (8)$$

where the Gaussian function $g_\sigma(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-t^2/(2\sigma^2)\big)$ is a regularized approximation of $\delta(t)$.

F. Other HPC Optimizations

In addition to the innovations above, other optimizations are also involved in per-fragment performance improvement.

Aggregated data transfer. On ORISE, multiple blocks of data required by multiple GPU kernels in each worker process, are aggregated into one large block and transferred once, for higher PCIe bandwidth utilization. This is not required on Sunway, since the offloaded workloads can access the same memory space with the host thread.

Asynchronous data movement. On Sunway, computations and memory accesses are overlapped, leveraging double buffer and DMA (direct memory access), to hide the latency between on-chip memory and off-chip memory.

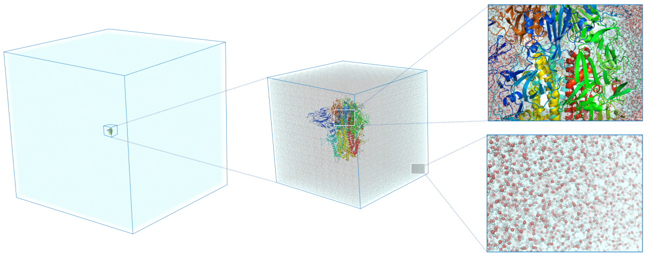


Fig. 7. The SARS-CoV-2 spike protein in water with up to 100 million atoms.

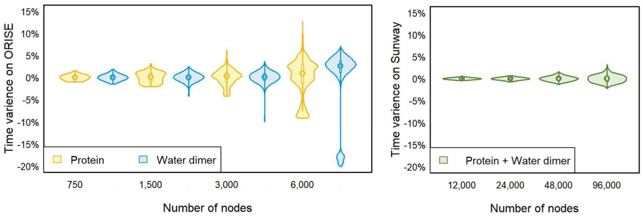


Fig. 8. Execution time variation across massive computing nodes.

VI. How performance was measured

A. Physical and Biological Test Systems

Fragments of the Spike glycoprotein (S protein) in SARS-CoV-2 [37] (PDB number is 7DF3) are used in our evaluation. The protein has 3,180 residues and, including the surrounding water, the simulation has 101,299,008 atoms. The simulations were carried out with distance thresholds of 4 Å for protein two-body QM interactions, 4 Å for protein-water QM interactions, 4 Å for water-water QM interactions. There are 3,171 conjugate caps and 11,394 generalized conjugate caps. There are 3,088 two-body interactions between a residue and a water molecule within the distance threshold, $128,341,476$ water–water interactions within the distance threshold.

B. Systems and Environment

QF-Raman is evaluated on two typical supercomputers with different architectures. The first one is the new ORISE supercomputer on which each computing node is equipped with a 32-core 2.50 GHz x86 CPU and accelerated by 4 HIP-based GPUs interconnected by the PCIe, with each GPU consisting of 4,096 cores in 64 CUs. Computing nodes are connected using an Infiniband network. For compilation, rocm [38] is used. The second one is the new generation Sunway supercomputer, i.e., the latest machine in Sunway family. It contains 96,000 nodes connected via a customized network. Each node has a SW26010-pro heterogeneous CPU, which consists of 390 cores including 6 managing cores and 384 accelerating cores, total 37,440,000 cores in the full system. For compilation, swgcc and swcl [39] are used.

C. Measurement

The DFPT time per cycle (i.e., the wall clock time used for calculating a single DFPT loop) is used for performance measurement, with DFPT applied to each fragment in each leader process (as shown in the right-bottom part of Fig. 3). The "DFPT time per cycle" includes all the time used in a single DFPT loop (IO included). Setup time, such as the setup of the system and MPI initialization and finalization, is not included.

VII. Performance Results

A. Step-By-Step Optimizations

In this section, evaluations have been performed for innovations introduced in Sections V-B, V-C and V-D.

1) System-Size-Sensitive Load Balancing: Fig. 8 gives the execution time variation across massive computing nodes, with respect to their average execution times.

ORISE. The evaluation is conducted on two systems: water dimer and protein respectively, with each water dimmer consisting of same-sized fragments of 6 atoms, and protein having various-sized fragments ranging from 9 to 35 atoms. With the system-size-sensitive load balancing strategy introduced in Section V-B, QF-Raman is able to balance various-sized protein fragments, leading to a time variance of only $-1\% \sim +1.5\%$ on 750 nodes. The time variance increases with the number of nodes, i.e., $-2.1\% \sim +3.2\%$, $-4.3\% \sim +6.2\%$ and $-9.2\% \sim +12.7\%$ on 1,500 nodes, 3,000 nodes, and 6,000 nodes, respectively for the protein, but still remains vary small when considered in light of the huge execution time difference (i.e., 5.4×) between the smallest 9-atom fragments and the biggest 35-atom fragments. In particular, execution times of the even-sized water dimer fragments vary more significantly across nodes than the various-sized protein fragments, since the scheme that allows prefetched tasks to be executed before the completion of current ones (shown in Fig. 4(e)) is disabled, for the purpose of showcasing its effects.

Sunway. Fragments from both protein and water dimmer are processed together, since Sunway offers more computing nodes. When using 12,000 nodes, 72,000 processes are created for execution, and their execution times distribute within the range of $-0.4\% \sim +0.4\%$ of their average, and when more nodes are used, i.e., 24,000 nodes, 48,000 nodes, and 96,000 nodes, the time variance remains satisfying, i.e., $-2.3\% \sim +3.2\%$ in the worst case. Workloads are more balanced on Sunway compared with ORISE, due to the co-location of protein and water dimer.

2) Symmetry-Aware Strength Reduction: Fig. 9 illustrates the step-by-step speedups of first applying symmetry-aware strength reduction (introduced in Section V-D) and then applying elastic workload offloading (introduced in Section V-C), across various-sized protein fragments. As stated in Section V-A, each fragment is simulated in a worker process.

As shown by blue bars in Fig. 9, the DFPT cycle can be significantly accelerated by optimizing BLAS calls on both supercomputers. On ORISE, the optimization yields $3.0-4.4\times$ of speedups across various-sized protein fragments, 3.7× on average. On Sunway, similar speedups, i.e., upto 6.0× and averaged 3.7×, can be achieved.

3) Elastic Workload Offloading: Speedups of elastic workload offloading (introduced in Section V-C) are given by

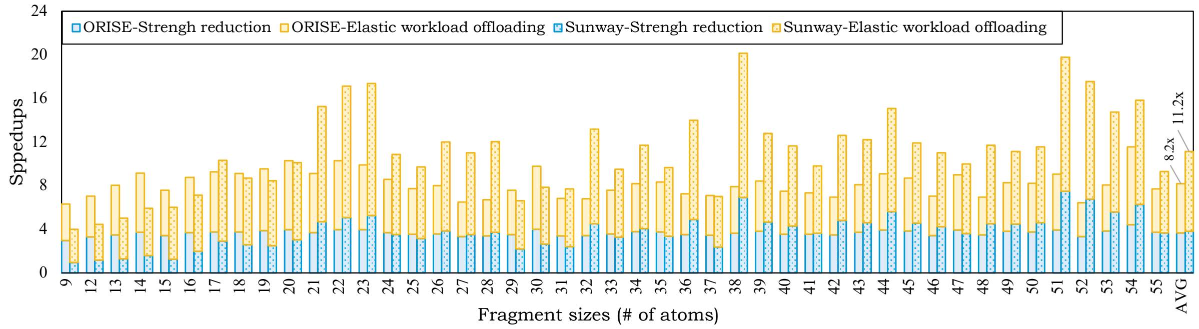


Fig. 9. Step-by-step speedups of applying symmetry-aware strength reduction (Section V-D) and elastic workload offloading (Section V-C).

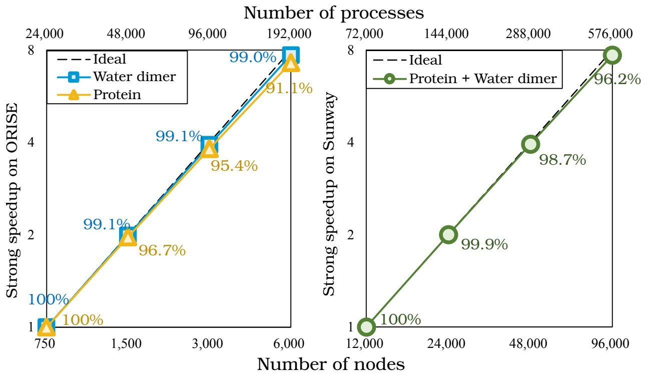


Fig. 10. Strong scaling results on two supercomputers for system in Fig. 7.

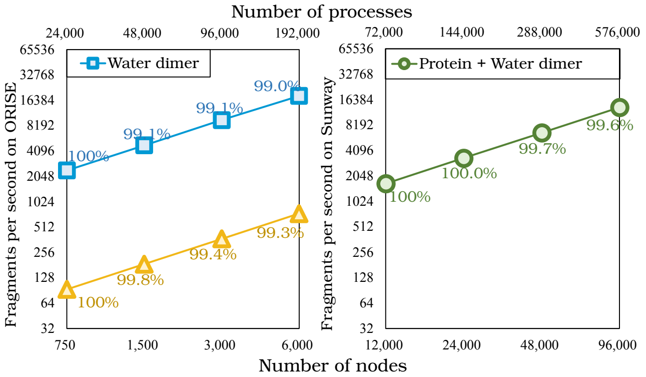


Fig. 11. Weak scaling results on two supercomputers for system in Fig. 7.

orange bars in Fig. 9. For protein fragments with various subsystem sizes, QF-RAMAN is able to gather scattered short calculations and pack at least 64 of them as a workload, for better utilization of parallel computational powers on accelerators. Performances are further boosted on top of symmetry-aware strength reduction, on both supercomputers. On ORISE, speedups are improved to 6.3×–11.6× for an overall DFPT cycle, 8.2× on average. On Sunway, along with symmetry-aware strength reduction, speedups up to 16.2× have been achieved, averaged at 11.2×. In particular, invocations to GEMM are batched with a stride of 32, that is, each matrix with original size $M \times N$ is first padded to $32 \cdot \lceil M/32 \rceil \times 32 \cdot \lceil N/32 \rceil$, and then all GEMM with the same size after padding are batched into a single workload.

B. Scalability Results

1) Strong Scaling: Fig. 10 gives the evaluation results of strong scaling on two supercomputers.

ORISE. Two systems of water dimers and a protein are calculated, using from 750 to 6,000 computing nodes (i.e., 24,000 to 192,000 processes). For the water dimer, a speedup of 1.98× is achieved with 1,500 computing nodes compared with 750 nodes, yielding a parallel efficiency of 99.1%. This efficiency remains satisfying at 3,000 and 6,000 nodes, since workloads are balanced with all fragments having 6 atoms. For protein, using 1,500 computing nodes improves its performance by 1.93× over 750 nodes, leading to a parallel efficiency of 96.7%. It drops slightly compared with water dimer, the reason lies in unbalanced workloads with fragment sizes varying from 9 to 35 atoms. However, with our system-size-sensitive packing strategy introduced in Section V-B, QF-RAMAN is able to obtain significant strong speedups when more nodes are involved, yielding impressive parallel efficiencies of 95.4% and 91.1% with 3,000 and 6,000 nodes respectively, even with those fragments whose system sizes vary by a large amount.

Sunway. Fragments of protein and water dimer are processed together, since Sunway provides more computing nodes. They are calculated using 12,000 to 96,000 nodes (i.e., 72,000 to 576,000 processes), thus the evaluation involves up to 37,440,000 cores. Compared with using 12,000 nodes, a parallel efficiency of 99.9% is obtained with 24,000 nodes. Impressively, using 48,000 and 96,000 nodes can achieve parallel efficiencies of 98.7% and 96.2% respectively, with 96,000 approaching the total number of available computing nodes in the full system of Sunway (96,300 nodes).

2) Weak Scaling: Fig. 11 gives the evaluation results of weak scaling on two supercomputers.

ORISE. Two systems of water dimer and protein are calculated, using from 750 to 6,000 computing nodes (i.e., 24,000 to 192,000 processes). For water dimer, whose fragments have the same size of 6 atoms, 3,343,536 fragments (with atomic displacement) have been processed by 750 nodes, yielding a throughput of 2,406.3 fragments per second. When both computing nodes and fragments are doubled, i.e., to 1,500/3,000/6,000 nodes and

TABLE I
DOUBLE PRECISION PERFORMANCES ON TWO SUPERCOMPUTERS.

Platform	Part	TFLOPS on single accelerator	PFLOPS (% of FP64 peak)
ORISE	$n^{(1)(\boldsymbol{r})}$	1.11 − 3.93	85.27 (53.8%)
	$H^{(1)}_{\mu,\nu}$	0.95 − 3.27	71.56 (45.2%)
Sunway	$n^{(1)(\boldsymbol{r})}$	2.10 − 4.82	311.17 (23.2%)
	$H^{(1)}_{\mu,\nu}$	2.44 − 4.87	399.90 (29.5%)

6,691,536/13,387,536/25,885,440 fragments respectively, the throughput keeps increasing to 4,772.2/9,546.6/18,445.1 fragments per second, yielding impressive weak scaling efficiencies of 99.1%/99.1%/99.0%. For protein, whose fragments have a significant system size variance from 9 to 35 atoms, 88,800 fragments have been processed by 750 nodes, yielding a throughput of 93.2 fragments per second. This is significantly lower than water dimer, due to larger system size and higher complexity in each fragment. Similarly to water dimer, when both computing nodes and fragments are doubled, satisfying weak scaling efficiencies have been obtained, reaching 99.8%/99.4%/99.3% respectively.

Sunway. Similar to strong scaling evaluation above, fragments of water dimers and protein are processed together. To start, 4,151,294 such mixed fragments are calculated using 12,000 nodes, with a throughput of 1,661.3 fragments per second. For 24,000 nodes, the number of fragments has been doubled to 8,302,588, leading to a throughput of 3,324.3 fragments per second, with 100.0% weak scaling efficiency. When further scaling to 48,000/96,000 nodes respectively, QF-RAMAN is capable of processing 6,626.9/13,239.8 fragments per second, yielding scaling efficiencies of 99.7%/99.6%.

C. Peak Performance

The double precision performance for simulating a fragment varies with its system size. Among the four parts in each DFPT cycle (illustrated in the right bottom part of Fig. 3), calculations of response density ($n^{(1)}(\mathbf{r})$) and response Hamiltonian ($H^{(1)}_{\mu,\nu}$) are extremely time-consuming, e.g., contributing to 93.1% of total execution time in a 49-atom segment on ORISE, thereby double precision performances of these two parts are reported in Table I, using the S protein system.

ORISE. When calculating the response density ($n^{(1)}(\mathbf{r})$), 1.11 − 3.93 TFLOPS (on a single GPU) can be achieved across all fragments with various sizes, e.g., 1.99 TFLOPS for a 15-atom fragment, and 2.98 TFLOPS for a 35-atom fragment. Given the fragment size distribution of decomposing the S protein, the performance of its simulation on 24,000 GPUs could thus be estimated to reach 85.27 PFLOPS in double precision, i.e., 53.8% of measurable peak performance on ORISE. When calculating the response Hamiltonian ($H^{(1)}_{\mu,\nu}$), similar performances have been achieved, i.e., 71.56 PFLOPS on the full system with an FP64 efficiency of 45.2%.

Sunway. On a single SW26010-pro processor (including 390 cores), the response density ($n^{(1)}(\mathbf{r})$) can be calculated at 2.10–4.82 TFLOPS across all fragments, and the response Hamiltonian ($H^{(1)}_{\mu,\nu}$) at 2.44–4.87 TFLOPS. Therefore, the double precision performance on the full system of 96,000 nodes could be estimated as 311.17 and 399.90 PFLOPS respectively, and FP64 efficiencies reach 23.2% and 29.5%.

VIII. APPLICATION

This section discusses the application of QF-RAMAN to SARS-CoV-2 spike protein. The spike protein is integral to the virus's ability to infect host cells. It facilitates viral entry into human cells by binding to the angiotensin-converting enzyme 2 (ACE2) receptor present on the surface of human cells, particularly in the respiratory tract. This interaction initiates the fusion of the viral envelope with the host cell membrane, enabling viral entry and subsequent infection. The SARS-CoV-2 spike protein is a pivotal component of the virus that shapes its infectivity, virulence, and susceptibility to intervention. Understanding its role and properties is crucial for developing effective countermeasures against COVID-19 and advancing our broader understanding of viral pathogenesis and immunity. Raman spectra analysis of the spike protein holds promise in furnishing crucial structural insights. Utilizing the PBE density functional and a "light" basis set, the Raman spectrum for the spike protein was computed. The calculated spectrum is compared to experimental data for comparison, are delineated in Fig. 12. The smearing of the theoretical Raman spectra was set to 5 cm^{-1} for gas phase protein and set to 20 cm^{-1} for water and protein with water.

For the protein in the gas phase, as illustrated in Fig. 12(a), there is a good agreement between the calculated and experimental Raman spectra [40], with characteristic patterns being easily visible despite minor differences in intensities. The Raman band around 1030 cm^{-1} is related to the breathing modes of phenylalanine (Phe) residues in proteins, and the band around 1450 cm^{-1} refers to CH2 bending vibration. The simulation successfully reproduces these distinct spectral features. In the amide III spectral region (coupled C–N stretching and N–H bending vibrations, around 1200–1360 cm^{-1}), the two bands exhibit lower intensity in the experimental measurement compared to our calculation. In the amide I region, the relative intensity of the calculated spectra is lower than the measured spectra. Fig. 12(b) shows Raman spectra of the spike protein in an explicit water box (in total 101,298,995 atoms), the Raman signals from the protein are obscured by the spectral contributions of water while the peaks associated with C–H stretching vibrations (around 2900 cm^{-1}) remain discernible. By monitoring these peaks, researchers can obtain information about protein conformational changes and dynamic behavior. We have also computed the Raman spectra of a 101,250,000-atoms model of pure water. In these spectra, we observed the characteristic peaks of O–H bending and O–H stretching vibrations. Furthermore, we observed the emergence of peaks in the low-frequency region. These low-frequency features can be attributed to two-body interactions and the increased number of atoms introduced in the simulation. Simulations with a larger number of atoms can more accurately represent the microscopic structure and interactions within the complex

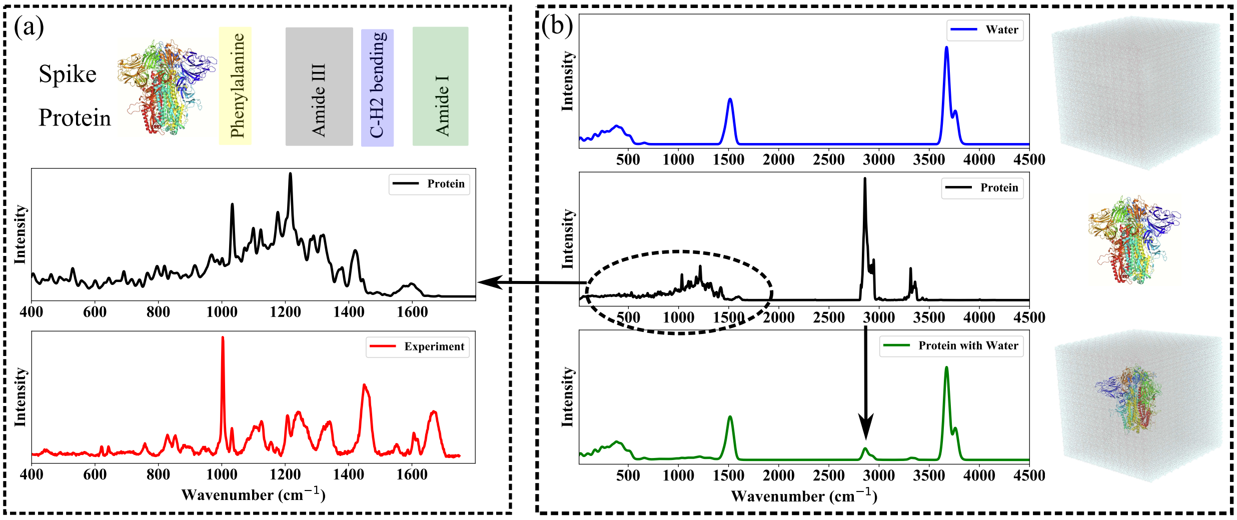


Fig. 12. (a) Simulated Raman spectra of the spike protein in the gas phase (black), the comparison with experimental gas phase data (red) is also given. (b) Simulated Raman spectra of the water (101,250,000 atoms) (blue), the spike protein (49,008 atoms) in the gas phase (black), and the protein with explicit water (101,299,008 atoms) box (green). For visualization, a subset of 1,017,621 atoms is shown in the protein with explicit water.

system. In our future work, we plan to employ fragment-based methods in conjunction with molecular dynamics simulations to obtain more accurate Raman spectra that better capture the features of liquid water.

The impact of studying the SARS-CoV-2 spike protein using QF-RAMAN spectroscopy is multifaceted and significant. The QF-RAMAN spectroscopy allows for a detailed characterization of the spike protein's structure, including its secondary and tertiary structures. This helps in understanding how the protein folds and functions, which is critical for viral attachment and entry into host cells. By analyzing the Raman spectra, researchers can identify how mutations in the spike protein affect its structure and functionality. This is crucial for tracking the evolution of the virus and its various strains. Raman spectra can be used to test the efficacy of these drugs by observing changes in the spike protein's structure upon drug binding. Moreover, insights from Raman spectroscopy can inform the selection of spike protein segments that elicit strong immune responses, improving vaccine design. Raman spectroscopy can be used to develop rapid diagnostic tools that detect the presence of the spike protein or its fragments in clinical samples, facilitating early and accurate diagnosis of COVID-19. The QF-RAMAN technique can help differentiate between different variants of SARS-CoV-2 based on structural differences in the spike protein, aiding in epidemiological tracking and response. The methodologies developed for studying the SARS-CoV-2 spike protein can be applied to other viral proteins, enhancing our overall understanding of viral mechanisms and pathogenesis. The integration of QF-RAMAN spectroscopy in virology can foster cross-disciplinary research, combining insights from physics, chemistry, and biology to tackle viral diseases more effectively. Overall, the impact of applying QF-RAMAN spectroscopy to the SARS-CoV-2 spike protein extends beyond immediate applications in COVID-19 research, and it provides tools and knowledge that can be leveraged in the broader contexts of viral pathology and biomedical innovation.

The past decade has witnessed a tremendous increase in the computing power of high-end supercomputers, and the mainstream design philosophy, due to hardware technology and power limitations, is to pursue higher FLOPs with many-core architectures, e.g., tensor cores. As a result, supercomputers today are not performance-friendly for algorithms with frequent global communications (e.g., traditional QM simulations), requiring a rethinking of algorithm design and parallel implementation for scientific applications. With respect to algorithm design, the idea of "divide-and-conquer" [41] is practical in many areas based on various types of localities. The proposed QF-RAMAN provides a successful example of this idea, with chemical locality exploited. To the best of our knowledge, it marks the pioneering QM calculation for Raman spectra involving as many as more than 100 million atoms, scaled to full systems on two advanced supercomputers. Given its near linear scalability as shown in Fig. 10 and Fig. 11, we see no intrinsic obstacles for even larger systems. With respect to parallel implementation, traditional scaling bottlenecks of communications and memory consumption have shifted to load balance under the "divide-and-conquer" framework, which is overcome by a system-size-sensitive load balance strategy (Section V-B) tailored to the physical problem in QF-RAMAN. Also, calculations in a division (a fragment in this work) could be too small to fully utilize the computing power of accelerators (e.g., GPUs), and QF-RAMAN involves an elastic workload offloading scheme (Section V-C) for this. The techniques above are easy to apply on other supercomputers, since they are independent of any specific architecture. The advances outlined in this study show the viability of *ab initio* simulation of Raman spectra for exascale and post-E machines. QF-RAMAN can be further leveraged for applications to biochemical dynamics, thereby providing a more comprehensive treatment of biological systems.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (T2222026, U23B2020, 62232015, 62090024, 22273023, 22288201) and Shanghai Frontiers Science Center of Molecule Intelligent Syntheses. The numerical calculations in this study were carried out on the ORISE and new Sunway Supercomputer. Computing resources are financially supported by Laoshan Laboratory (No. LSKJ202300305).

References

[1] "https://www.nobelprize.org/prizes/chemistry/2013/press-release."

[2] H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis *et al.*, "Using raman spectroscopy to characterize biological materials," *Nat. Protoc.*, vol. 11, pp. 664–687, 2016.

[3] W. Kohn, "Nobel lecture: electronic structure of matter-wave functions and density functionals," *Rev. Mod. Phys*, vol. 71, pp. 1253–1266, 1999.

[4] X. Gonze and C. Lee, "Dynamical matrices, born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory," *Phys. Rev. B*, vol. 55, pp. 10 355–10 368, 1997.

[5] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, "Phonons and related crystal properties from density-functional perturbation theory," *Rev. Mod. Phys.*, vol. 73, pp. 515–562, 2001.

[6] H. Shang *et al.*, "Lattice dynamics calculations based on density-functional perturbation theory in real space," *Comput. Phys. Commun.*, vol. 215, pp. 26–46, 2017.

[7] P. Giannozzi *et al.*, "Quantum ESPRESSO: a modular and open-source software project for quantum simulations of materials," *J. Phys. Condens.*, vol. 21, p. 395502, 2009.

[8] G. Kresse and J. Hafner, "Ab initio molecular dynamics for liquid metals," *Phys. Rev. B*, vol. 47, pp. 558–561, 1993.

[9] A. H. Romero, D. C. Allan, B. Amadon, G. Antonius, T. Applencourt *et al.*, "ABINIT: overview and focus on selected capabilities," *J. Chem. Phys.*, vol. 152, 2020.

[10] F. Gygi *et al.*, "Large-scale electronic structure calculations of high-Z metals on the BlueGene/L platform," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2006.

[11] M. Frisch *et al.*, "Gaussian 16," 2016, Gaussian Inc. Wallingford CT.

[12] R. Dovesi *et al.*, "The CRYSTAL code, 1976–2020 and beyond, a long story," *J. Chem. Phys.*, vol. 152, p. 204111, 2020.

[13] B. Delley, "An all-electron numerical method for solving the local density functional for polyatomic molecules," *J. Chem. Phys.*, vol. 92, p. 508, 1990.

[14] V. Blum *et al.*, "Ab initio molecular simulations with numeric atom-centered orbitals," *Comput. Phys. Commun.*, vol. 180, pp. 2175–2196, 2009.

[15] X. Andrade *et al.*, "Real-space grids and the octopus code as tools for the development of new simulation approaches for electronic systems," *Phys. Chem. Chem. Phys.*, vol. 17, pp. 31 371–31 396, 2015.

[16] Y. Hasegawa *et al.*, "First-principles calculations of electron states of a silicon nanowire with 100,000 atoms on the K computer," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, 2011.

[17] J. L. Fattebert *et al.*, "Modeling dilute solutions using first-principles molecular dynamics: computing more than a million atoms with over a million cores," in *International Conference for High Performance Computing, Networking, Storage and Analysis*, November 2016, pp. 12–22.

[18] S. Das *et al.*, "Fast, scalable and accurate finite-element based ab initio calculations using mixed precision computing: 46 PFLOPS simulation of a metallic dislocation system," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2019.

[19] S. Das, B. Kanungo *et al.*, "Large-scale materials modeling at quantum accuracy: ab initio simulations of quasicrystals and interacting extended defects in metallic alloys," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2023.

[20] J. C. A. Prentice *et al.*, "The ONETEP linear-scaling density functional theory program," *J. Chem. Phys*, vol. 152, no. 17, p. 174111, 05 2020.

[21] L. E. Ratcliff, *et al.*, "Flexibilities of wavelets as a computational basis set for large-scale electronic structure calculations," *J. Chem. Phys.*, vol. 152, p. 194110, 2020.

[22] A. Nakata *et al.*, "Large scale and linear scaling DFT with the CONQUEST code," *J. Chem. Phys.*, vol. 152, p. 164112, 2020.

[23] T. D. Kühne *et al.*, "CP2K: An electronic structure and molecular dynamics software package—Quickstep: Efficient and accurate electronic structure calculations," *J. Chem. Phys.*, vol. 152, 2020.

[24] R. Schade, T. Kenter, H. Elgabarty, M. Lass, T. D. Kühne, and C. Plessl, "Breaking the exascale barrier for the electronic structure problem in ab-initio molecular dynamics," *Int. J. High Perform. Comput. Appl.*, vol. 37, no. 5, p. 530–538, 2023.

[25] S. Salustro *et al.*, "Comparison between cluster and supercell approaches: the case of defects in diamond," *Theor. Chem. Acc.*, vol. 136, no. 4, pp. 1–13, 2017.

[26] H. Shang *et al.*, "Extreme-scale ab initio quantum Raman spectra simulations on the leadership HPC system in China," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021.

[27] B. Wang, K. R. Yang, X. Xu, M. Isegawa, H. R. Leverentz, and D. G. Truhlar, "Quantum mechanical fragment methods based on partitioning atoms or partitioning coordinates," *Acc. Chem. Res.*, vol. 47, pp. 2731–2738, 2014.

[28] G. M. J. Barca *et al.*, "Scaling correlated fragment molecular orbital calculations on Summit," in *International Conference for High Performance Computing, Networking, Storage and Analysis*, 2022.

[29] H. Shang *et al.*, "All-electron, real-space perturbation theory for homogeneous electric fields: theory, implementation, and application within DFT," *New J. Phys.*, vol. 20, p. 073040, 2018.

[30] J. Liu and X. He, "Recent advances in quantum fragmentation approaches to complex molecular and condensed-phase systems," *WIREs Comput. Mol. Sci*, vol. 13, no. 3, p. e1650, 2023.

[31] J. Liu, J. Z. H. Zhang, and X. He, "Fragment quantum chemical approach to geometry optimization and vibrational spectrum calculation of proteins," *Phys. Chem. Chem. Phys.*, vol. 18, pp. 1864–1875, 2016.

[32] E. B. Wilson, J. C. Decius, and P. C. Cross, *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*. Mineola, NY: Dover Publications Inc., 1980.

[33] Z. Wu *et al.*, "Portable and scalable all-electron quantum perturbation simulations on exascale supercomputers," in *International Conference for High Performance Computing, Networking, Storage and Analysis*, 2023.

[34] "https://www.khronos.org/opencl."

[35] M. Shao *et al.*, "A structure preserving Lanczos algorithm for computing the optical absorption spectrum," *SIAM J. Matrix Anal. Appl.*, vol. 39, no. 2, pp. 683–711, 2018.

[36] L. Reichel, M. M. Spalević, and T. Tang, "Generalized averaged Gauss quadrature rules for the approximation of matrix functionals," *BIT*, vol. 56, pp. 1045–1067, 2016.

[37] C. Xu *et al.*, "Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM," *Sci. Adv.*, vol. 7, no. 1, p. eabe5575, 2021.

[38] AMD ROCm Release Notes, Advanced Micro Devices, Inc., Santa Clara, CA, 2016.

[39] M. Wu *et al.*, "Bandwidth-aware loop tiling for dma-supported scratchpad memory," in *Proceedings of the ACM International Conference on Parallel Architectures and Compilation Techniques*, ser. PACT '20, 2020, p. 97–109.

[40] J. Huang *et al.*, "On-site detection of SARS-CoV-2 antigen by deep learning-based surface-enhanced Raman spectroscopy and its biochemical foundations," *Anal. Chem.*, vol. 93, pp. 9174–9182, 2021, pMID: 34155883.

[41] D. E. Keyes, Y. Saad, and D. G. Truhlar, Eds., *Domain-Based Parallelism and Problem Decomposition Methods in Computational Science and Engineering*. Philadelphia, PA: SIAM, 1995.