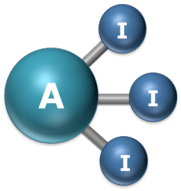


Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Nearer the nearsightedness principle:

Large-scale quantum chemical calculations

Project Report
Project Dates: 21/06/2021 - 27/08/2021
Univerity of Southampton

Project Student: András Vékássy, University of Southampton
Supervised by: Professor Chris-Kriton Skylaris, University of Southampton

Report Date: 28/09/2021

Nearer the nearsightedness principle:

Large-scale quantum chemical calculations

AI3SD-Intern-Series:Report-11_Vékássy
Report Date: 28/09/2021
DOI: 10.5258/SOTON/AI3SD0140
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Project Details

Title	Nearer the nearsightedness principle: Large-scale quantum chemical calculations
Project Reference	AI3SD-FundingCall3_012
Supervisor Institution	University of Southampton
Project Dates	21/06/2021 - 27/08/2021
Website	https://www.southampton.ac.uk/compchem/about/index.page
Keywords	DFT, linear-scaling, density kernel

2 Project Team

2.1 Project Student

Name and Title	Mr András Vékássy
Employer name / University Department Name	University of Southampton, Department of Chemistry
Work Email	av1g19@soton.ac.uk
Website Link (if available)	www.linkedin.com/in/andras-vekassy-baa1a8193/

2.2 Project Supervisor

Name and Title	Professor Chris-Kriton Skylaris
Employer name / University Department Name	University of Southampton, Department of Chemistry
Work Email	C.Skylaris@soton.ac.uk
Website Link (if available)	www.southampton.ac.uk/chemistry/about/staff/cks.page

2.3 Researchers & Collaborators

- **Dr Arihant Bhandari**, University of Southampton, Post-doc at the Skylaris Group.
  Providing help with implementing code and debugging in ONETEP.
- **Mr Davide Sarpa**, University of Southampton, PhD student at the Skylaris Group.
  Giving advice and answering general questions related to computational chemistry and programming, including Python, Bash scripting, version control (Git), data analysis.

3 Lay Summary

Within the realm of quantum theories, Density Functional Theory (DFT) is the most widely used method for chemical and material calculations. Despite DFT being computationally more feasible to implement than other quantum chemical approaches, large-scale simulations are still prohibited by the fact that computational effort scales cubically with the size of the system. To overcome this problem, Nobel prize winner Walter Kohn introduced a new principle he called "the nearsightedness of electronic matter" [1], in which he describes why interactions between electrons are negligible at large atomic distances.

ONETEP, a world-leading DFT software achieves linear-scaling by harnessing the principle of nearsightedness. In practice, ONETEP uses localised atomic orbitals [2], and truncates entries of the density kernel based on a single distance parameter [3]. The latter means that electronic interactions of atoms above the specified kernel cut-off are disregarded. This does speed up calculations considerably, albeit comes with a great price. Consider hydrogen and sulphur. We chemists know that these elements differ in electronic properties and thus applying one cut-off to all atoms is chemically insensitive. To produce faster and more accurate calculations, we studied polypropylene ($C_{72}H_{146}$) and T4 lysozyme [4] and developed a more fine-tuned, element sensitive truncation scheme in ONETEP.

4 Aims and Objectives

The current density kernel truncation scheme in ONETEP applies one distance parameter to all atoms, which - given that atoms differ in electronic properties - is chemically insensitive. Moreover, as the user decreases the cut-off, ONETEP encounters complications in its optimisation procedures and obtains less accurate results.

To gain a better understanding of the effects of kernel truncation, we performed total energy calculations on a polymer ($C_{72}H_{146}$), a nanoparticle ($Na_{128}Cl_{128}$ nanorod) and a protein (T4 Lysozyme [5]). We aimed to identify the limitations of the kernel cut-off and find out when does truncation result in a worsening of the quality of results. We studied sparsity patterns of the density kernel, convergence rates of ONETEP's inner and outer loop and compared changes in the components of the total energy.

Based on the collected data we proposed a new truncation scheme which applies atom-pair kernel cut-offs, meaning it truncates entries of the density kernel depending on the chemical elements involved. By doing so we were hoping to perform chemically more accurate and computationally faster calculations, that is calculations with lower total energy and better convergence. We also considered using the more fine-tuned cut-offs as Machine Learning descriptors in future research, as that would rapidly speed up calculations.

5 Theory

5.1 Density Functional Theory

What is Density Functional Theory (DFT) and why is it so widely used? To answer this question, let's have a short overview of its derivation. DFT originates from the notorious many-body problem in quantum mechanics. Let's have a system with N electrons and M nuclei with coordinates $\mathbf{r}$ and $\mathbf{R}$, respectively.

$$\hat{H}_{tot}\Psi_{tot}(\mathbf{r},\mathbf{R}) = E_{tot}\Psi_{tot}(\mathbf{r},\mathbf{R}) \tag{1}$$

Firstly, the Born-Oppenheimer approximation [6] is applied to the Schrödinger equation 1, which treats nuclei as stationary particles. This allows separation of the total Hamiltonian into an electronic and a nuclear one. Furthermore, instead of solving a differential equation with 6N coordinates we encounter two equations, each with 3N variables.

$$\hat{H}_{tot} = \hat{H}_e + \hat{T}_n \tag{2}$$

We can separate further the electronic Hamiltonian, into the operators for kinetic energy and nuclei-nuclei, nuclei-electron and electron-electron repulsion.

$$\hat{H}_e = \hat{T}_e + \hat{V}_{nn} + \hat{V}_{ne} + \hat{V}_{ee} \tag{3}$$

We can prove that applying the new, partitioned Hamiltonian to the total wave function creates an electronic and a nuclear wave function.

$$\hat{H}_e\Psi_e(\mathbf{r},\mathbf{R}) = E_e(\mathbf{R})\Psi_e(\mathbf{r},\mathbf{R}) \tag{4}$$

For completeness, the total energy is calculated after obtaining the electronic energy

$$(\hat{T}_n + E_e(\mathbf{R}))\Psi_n(\mathbf{R}) = E_{tot}\Psi_n(\mathbf{R}) \tag{5}$$

Our focus from now on will be the electronic wave function as the nuclear terms are computed after dealing with the electronic structure, placing the nucleii effectively in a cloud of electrons. A mathematical obstacle occurs as there is no straightforward way to solve the electronic if it contains the electron-electron repulsion. Here we introduce a new approximations: the Hartree-Fock method [7].

Hartree-Fock theory (HF) remains in the domain of electronic wave functions and starts by

neglecting the electron-electron repulsion. The instantaneous interactions described by V_{ee} are substituted by the interactions between one electron and a continuous charge distribution generated by the N-1 electrons. HF builds the wave function using Slater determinants [7], which are linear combinations of spin orbitals $\{\phi_N\}$. The determinants are carefully constructed to obey the Pauli exclusion principle [8].

$$\psi_{trial}(\chi_1, \chi_2, ..., \chi_N) = \frac{1}{\sqrt{N!}}|\phi_1\phi_2...\phi_N| \quad (6)$$

HF takes advantage of the variational principle [7, 9] via numerically optimising the parameters of the trial wave function, that is the constants that form the Slater determinant.

$$\langle\psi_{trial}|\,\hat{H}_{HF}\,|\psi_{trial}\rangle = E_{HF} \geq E_{tot} \quad (7)$$

The optimisation is done in an iterative fashion. Each electron is optimised in an average potential created by the overall cloud of electrons [10] to reduce the total energy. V_{ee} is therefore approximated in a self-consistent field V_{ext} [11] and the true solution is never obtained. The difference that arises between the total energy and the Hartree-Fock energy is called correlation energy.

$$E_{tot} - E_{HF} = E_{corr} \quad (8)$$

Turns out, the correlation energy is vital for us chemists. Methods such as post-HF introduce corrections for the correlation energy [[12]], and although they are chemically more accurate, their computational cost poses serious limitations [13].

Can we reduce the number of variables somehow? Yes! Wave functions depend on 4 variables but we can take a different approach and construct density ($\rho(\mathbf{r}, \mathbf{r}')$) from them.

$$\rho(\mathbf{r}, \mathbf{r}') = \langle\psi(\mathbf{r})|\psi(\mathbf{r}')\rangle \quad (9)$$

Density defined this way remains in the territory of quantum mechanics. Hohenberg and Kohn showed that there is a one-to-one correspondance (injectivity [10]) between the density and the external potential, and thus between the wave function and the external potential. In slightly more technical words, $V_{ext}(\mathbf{R})$ is a unique functional[1] of the density. Consequently, the Hamiltonian together with energy is fixed by the external potential - but strictly for the ground state!

$$E_0[\rho_0] = E_{ne}[\rho_0] + T[\rho_0] + E_{ee}[\rho_0] \quad (10)$$

Recall that due to the Born-Oppenheimer approximation the nuclear-nuclear repulsion is a constant, so we are only dealing with kinetic energy, and interactions such as electron-electron repulsion and nuclei-electron attraction. We also replaced the electron-electron interactions with an average potential. This so far is great, but the true density is not known! We return to an idea we have seen before: the variational principle. The second Hohenberg-Kohn theorem proves that "the functional that delivers the ground state energy of the system, delivers the lowest energy if and only if the input density is the true ground state density, ρ_0" [10]. A trial density ρ_{trial} (that satisfies quantum chemical conditions) therefore presents an upper bound to the ground state density and the true ground state energy can be approximated via minimisation procedures. Further discussion on DFT will be omitted, the keen reader can further dive into DFT by exploring Kohn-Sham DFT [10, 16] and the plane-wave pseudopotential method [17], both of which ONETEP and linear-scaling DFT is built upon. The important take away is that energy and all of its components can be accurately calculated from the external potential, ultimately from the electronic structure. Cubic-scaling inherently arises from the orthonormality constraint posed on Kohn-Sham orbitals [7, 18, 19].

5.2 Nearsightedness

To overcome the issue of cubic-scaling, Nobel prize winner Walter Kohn introduced a new principle he called the nearsightedness of electronic matter (NEM) [1].

Kohn's approximation takes advantage of the fact that "local electronic properties, such as the density n(r), depend significantly on the effective external potential only at nearby points" [20], i.e. interactions between electrons are negligible at large atomic distances. As a consequence of NEM, the density matrix decays exponentially.

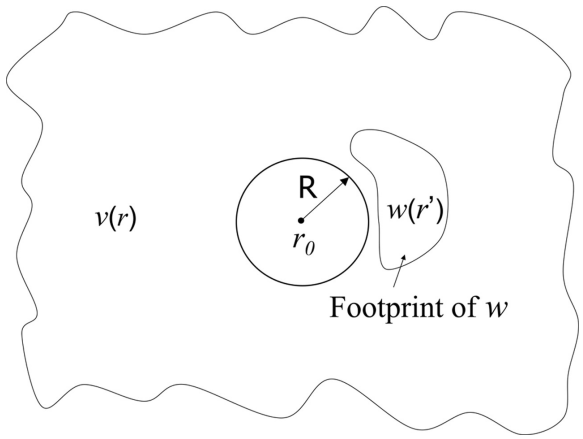


Figure 1: Schematic of nearsightedness by Kohn [20]. Mathematically, $\rho(\mathbf{r}, \mathbf{r}') \sim e^{-\gamma|\mathbf{r}-\mathbf{r}'|} \to 0$ as $|\mathbf{r} - \mathbf{r}'| \to 0$.

[1] A functional is a function of a function. For a mathematically more rigorous definition [14, 15].

5.3 ONETEP

ONETEP (Order-N Electronic Total Energy Package) [3] is a world-leading DFT software that achieves linear-scaling by harnessing the principle of nearsightedness.

In practice, ONETEP implements NEM as follows. Basis sets for elements are built using Non-Orthogonal General Wannier Functions (NGWFs) [2], which are then used in combination with the density kernel $\mathbf{K}^{\alpha\beta}$ to construct the density matrix $\rho(\mathbf{r}, \mathbf{r}')$.

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{\alpha,\beta} \phi_\alpha(\mathbf{r}) \mathbf{K}^{\alpha\beta} \phi^*_\beta(\mathbf{r}')$$

NGWFs are localised orbitals which aid linear-scaling calculations from a computational perspective [21]. Perhaps the most significant methods that make ONETEP distinct from other DFT software is the truncation of the density kernel. Given that elements of p(r,r') decay exponentially, they are negligible at large atomic distances and can be set to zero.

$$\rho(\mathbf{r}, \mathbf{r}') = 0 \text{ when } |\mathbf{r} - \mathbf{r}'| < r_{\text{cut}} \qquad (11)$$

Truncation creates a sparse matrix which can be stored and used for calculations efficiently. Sparsity is a great feature for O(N) scaling but excluding elements of the density kernel makes calculations less accurate by nature of the variational principle. In other words, there is a trade-off between accuracy and computational cost.

Other computational algorithms such as parallelization and sparse matrix algebra [22] also aid O(N) memory and CPU cost.

As seen previously, electronic structure theory [7] often comprises self-consistent field calculations. ONETEP's calculations in particular consist of an initialisation phase followed by two, embedded self-consistent loops [3]. In the initialisation phase, ONETEP obtains the initial NGWFs and creates the initial $\mathbf{K}^{\alpha\beta}$ based on a list of overlaps between elements. After initialisation, the calculation enters the two nested loops. The inner loop, also known as the LNV loop [23, 24], is responsible for optimizing the density kernel. The outer loop is optimizing the NGWFs jointly with the density kernel [25]. The loops are designed to maintain linear-scaling whilst achieving cubic-scaling accuracy - understanding their embedded structure is essential to be able to locate the density kernel in ONETEP's source code.

6 Methodology

We studied three systems: a polymer, a nanoparticle, and a protein. The intern built a polypropylene molecule ($C_{72}H_{146}$) in Avogadro [26] and a NaCl nanorod ($Na_{128}Cl_{128}$, rock-salt, a=5.64Å) using the Atomic Simulation Environment Python library [27]. The polymer was chosen as it is a simple and forgiving system calculation-wise, meanwhile the nanoparticle is easily scalable and can facilitate O(N) testing. A T4 lysozyme complex was specifically chosen from a DFT case-study [5] to include a real-life system in the dataset. Inputfiles and xyz files can be found in the supplementary documentation.

We performed single-point energy calculations on all three systems. Space filling [28] was turned off to prevent ONETEP from rearranging the elements' IDs. The number of LNV iterations were set to 10 for the polymer and the nanorod, and to 30 for the protein. This was done as the main focus of this project was the density kernel and its optimisation. The nanorod proved to be problematic to model and considering the time limitations for the project, we eventually decided to omit the NaCl structure from the dataset. Similarly, to save computational time we stopped the lysozyme calculations after one NGWF CG iteration. Additional information on keywords and setting up calculations can be found at www.onetep.org.

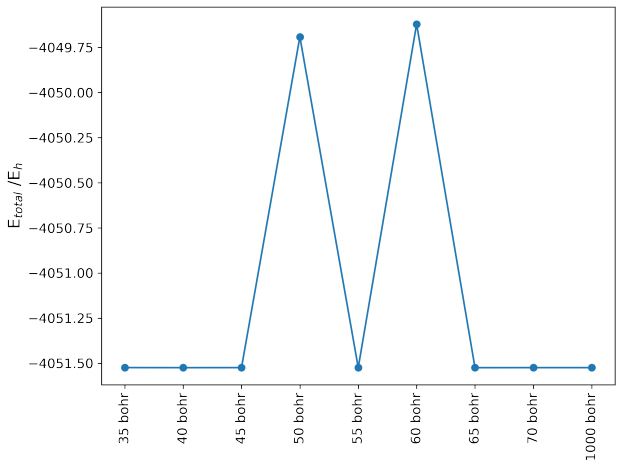


Figure 2: The NaCl calculations did not converge when using 35 and 65 bohr kernel cut-offs. Abnormalities were also witnessed in optimisation parameters.

6.1 Density kernel

Our first objective was to output the density kernels. As highlighted previously the density kernel occurs in two modules in the source code: the lnv_mod (inner loop) and the ngwf_cg_mod (outer loop). As ONETEP was written in Fortran, most matrices are in binary form during calculations. Density kernels are stored in SPAM3 matrices [29] and can only be obtained via calling the sparse_show_matrix() subroutine (Appendix B.1) from the sparse module. SPAM3 type sparse matrices are stored in SPAM3_EMBED array, which stores subsystem matrices. Further-

more, the SPAM3_EMBED matrices are stored in a SPAM3_EMBED_ARRAY (aux in the source code). This additional layer of embedding includes spin channels and k-points [29]. Given that we have no subsystems, our matrix will be aux%kern%m(1:1)%m(1:1). Each output file's name includes the LNV and NGWF CG iteration number and also shows elements with their ID and NGWFs (e.g. C 1 4 is the first carbon's fourth NGWF).

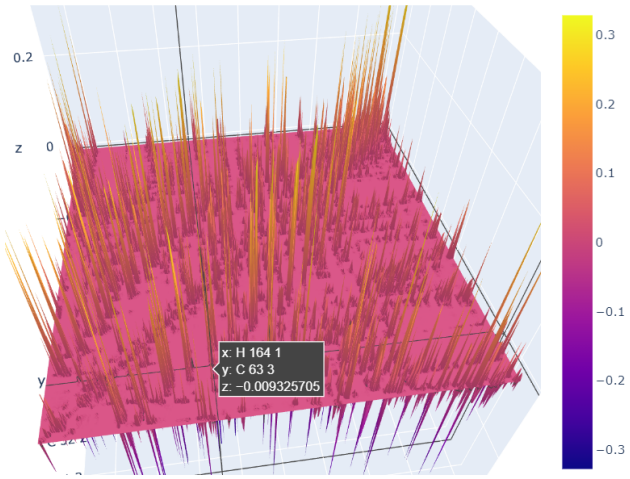
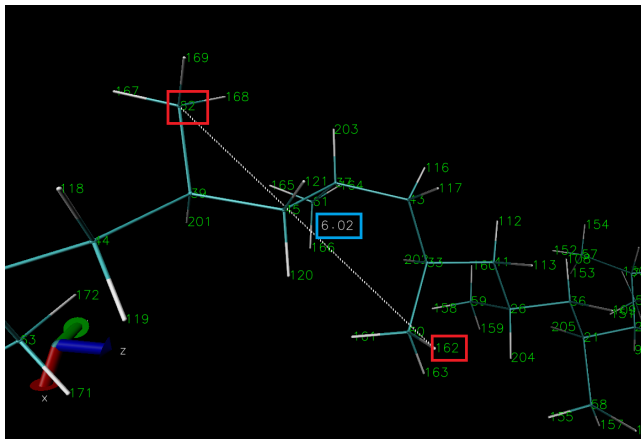


Figure 3: 3D visualisation of the density kernel (above). The Plotly (www.plotly.com) Python library allows loading in labels for the x and y axes and displaying them by hovering the mouse over the plot. We programmed TK Console in VMD [30, 31] to display the IDs of elements (below), and were able to check manually what distance corresponds to the displayed density value. In VMD, element numbering starts from zero. The element IDs and the distance (Å) were highlighted with red and blue, respectively.



The collected SPAM3 matrices were visualised in two and three dimensions (3), which prompted Mr Davide Sarpa the following idea. In molecular dynamics [32], radial distribution functions (RDF, often denoted g(r)) express the probability of finding atoms as a function of distance [33]. In a similar fashion we decided to investigate the distribution of density values and their distance dependence. We have chosen MDAnalysis [34], a Python library to compute distances between atoms. The distance matrices were obtained via the distance_array(atomgroup1,atomgroup2) function [35], see Appendix B.2. We used the distance arrays in combination with the converged SPAM3 matrices to investigate the "range" of interactions between electrons. We set thresholds for the entries of the converged density kernel and kept distances in a separate array that correspond to density values above the threshold. Moreover, our analysis distinguished atom-pairs (e.g. carbon-hydrogen, sulphur-sulphur) rather than treating all chemical elements equivalent as did the old truncation scheme. The obtained atom-pair distances formed the parameters the new truncation scheme:

$$\rho(\mathbf{r}_i, \mathbf{r}_j) = 0 \text{ when } |\mathbf{r}_i - \mathbf{r}_j| < r_{ij}$$

where i and j are chemical elements. The cut-off parameters can be found in the supplementary documentation.

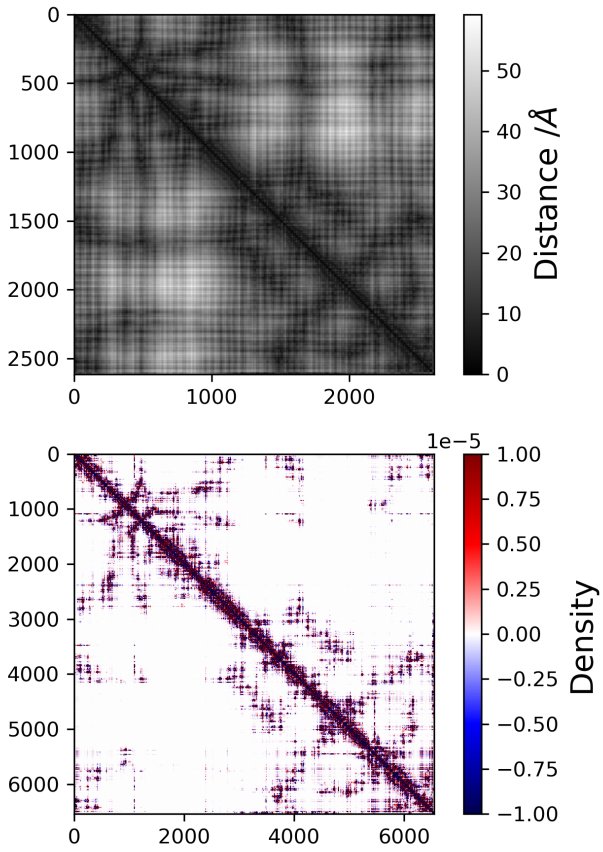


Figure 4: Distance array between all atoms (top) and the sparsity pattern of the converged density kernel (bottom, ∞ cut-off) of T4 Lysozyme L99A/M102Q. Due to nearsightedness, the vast majority of kernel elements are zero and/or negligible.

6.2 Atom-pair kernel cut-off

In ONETEP, the sparse_initialise_mod is responsible for creating the initial sparsity pattern for the density kernel via the parallel_strategy_list_cross_overlaps subroutine. The pattern is assembled from a list of overlaps between atoms (overlaps(jsub,isub)) in accordance with the truncation scheme in use. The atom-pair cut-offs we added essentially reset the radii upon the formation of the overlap list (Appendix B.3). Hence, the old truncation scheme does not interfere with

the new one, the two can be used simultaneously. The idea of creating a block (refer to inputfiles) for the inputfile was dismissed given the limited timespan for the project.

We were able to confirm that the new cut-off method was functioning by setting all atom-pair distances to 25 bohr and compare the results to a calculation using the traditional kernel_cutoff set to 25 bohr on the polymer. Another way to verify that truncation indeed took place was done by looking at the components of the overlaps(jsub,isub) array: max_overlaps (integer), num_overlaps(:) (1D array), overlap_list(:,:) (2D array). These components all correspond to the chemical elements involved, not the density kernel (Appendix C).

6.3 Loops

Our second goal was to study the effects of truncation on the inner and outer loops. With the help Bash scripting, we gathered data from ONETEP's output files (polypropylene.out, complex_2001.out), including components of and the total energy itself, and optimization parameters such as (error in the) RMS gradient and [H,K] commutator [36]. Occupancies [24], which ought to be between 0.00 and 1.00 were also great indicators of ONETEP's converge rates. All gathered data can be found in the supplementary documentation.

Datasets: zip files: SPAM3 matrices, elements (overlap_list) and .out files? (from calc and bash) Software: ONETEP, Python (ASE, MDAnalysis, Plotly), Bash, Fortran, VMD, Avogadro

7 Results

The lysozyme calculation lasted approximately 10 minutes longer owing in part to the plethora of conditionals in the new cutoff (Appendix B.3). We also observed in case of the protein that the density kernel filling was still above 70% for density thresholds such as 10^{-3}, suggesting that the density kernel was not as sparse as expected. The abundance of elements relative to the system size also had a large effect on the cut-off parameters. The lysozyme complex only had 6 sulphurs, and thus the S-S interaction disappeared at density thresholds higher than 10^{-2}.

The new truncation scheme did not perform better for the polymer, we witnessed promising yet contradictory results for the protein. In comparison with the preliminary calculations involving similar density kernel fillings (ref table Y), calculations with the new cut-offs achieved lower total energies (refer to excel sheets in supplementary documentation) with worse convergence rates. Furthermore, the calculation with the largest threshold, that is with the sparsest density kernel yielded to the lowest total energy, contradicting expectations in line with the variational principle. At the same time, the error in the RMS gradient and the [H,K] commutators were higher than expected - occasionally by magnitudes.

8 Conclusions & Future Work

The new truncation scheme was successfully implemented in ONETEP – we are one step nearer the nearsightedness principle. Calculations on the T4 lysozyme's showed promising results, but the new cut-offs need to be tested in a research setting without pausing runs after one NGWF CG iteration. For instance, reaction enthalpy could be calculated which is comparable to experimental results. Regarding the nested loops, the poorer convergence rates remain to be studied in more detail as at this point it is not clear if the lower total energies are an anomaly or indeed we produced more accurate calculations. It is worth noting that the lowest total energy with the new truncation in use is still 65.43 $kJmol^{-1}$ higher than the calculation with ∞ cut-off. Future research should also investigate how changing the NGWF radii affects the density kernel and its distance dependence.

Provided the new truncation scheme performs well, using the atom-pair cut-offs as Machine Learning descriptors could make calculations significantly faster and more accurate, which would ultimately safe a lot of time and computational effort.

9 Outputs, Data & Software Links

In the supplementary documentation the following can be found: Bash and Python scripts that were used to collect, analyse and visualise data; MS Word documents and MS Excel sheets summarising data relating to the truncation schemes; inputfiles for calculations. "Readme_*.txt" files and comments provide further guidance.

10 Acknowledgements

We acknowledge the AI3SD Network+ for funding and the University of Southampton for use of the IRIDIS High Performance Computing Facility in completion of this work. The intern would like to thank Professor Chris-Kriton Skylaris, Dr Arihant Bhandari and Mr Davide Sarpa for the support throughout the project.

References

[1] W. Kohn, "Density functional and density matrix method scaling linearly with the number of atoms," *Physical Review Letters*, vol. 76, no. 17, pp. 3168–3171, 1996. DOI: 10.1103/PhysRevLett.76.3168.

[2] C. K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Diéguez and M. C. Payne, "Nonorthogonal generalized Wannier function pseudopotential plane-wave method," *Physical Review B - Condensed Matter and Materials Physics*, vol. 66, no. 3, pp. 1–12, 2002. DOI: 10.1103/PhysRevB.66.035119.

[3] J. C. A. Prentice, J. Aarons, J. C. Womack, A. E. A. Allen, L. Andrinopoulos, L. Anton, R. A. Bell, A. Bhandari, G. A. Bramley, R. J. Charlton, R. J. Clements, D. J. Cole, G. Constantinescu, F. Corsetti, S. M. Dubois, K. K. B. Duff, J. M. Escartín, A. Greco, Q. Hill, L. P. Lee, E. Linscott, D. D. O'Regan, M. J. S. Phipps, L. E. Ratcliff, Á. R. Serrano, E. W. Tait, G. Teobaldi, V. Vitale, N. Yeung, T. J. Zuehlsdorff, J. Dziedzic, P. D. Haynes, N. D. M. Hine, A. A. Mostofi, M. C. Payne and C.-K. Skylaris, "The ¡scp¿ONETEP¡/scp¿ linear-scaling density functional theory program," *The Journal of Chemical Physics*, vol. 152, no. 17, p. 174 111, May 2020. DOI: 10.1063/5.0004445.

[4] L. Gundelach, T. Fox, C. S. Tautermann and C. K. Skylaris, "Protein-ligand free energies of binding from full-protein DFT calculations: convergence and choice of exchange-correlation functional," *Physical Chemistry Chemical Physics*, vol. 23, no. 15, pp. 9381–9393, 2021. DOI: 10.1039/d1cp00206f.

[5] J. Dziedzic, S. J. Fox, T. Fox, C. S. Tautermann and C. K. Skylaris, "Large-scale DFT calculations in implicit solvent - A case study on the T4 lysozyme L99A/M102Q protein," *International Journal of Quantum Chemistry*, vol. 113, no. 6, pp. 771–785, 2013. DOI: 10.1002/qua.24075.

[6] D. Terme, S. Beitrag, D. Grund and V. Standpunkte, "Der physik 1.," 1927.

[7] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, ser. Dover Books on Chemistry. Dover Publications, 2012.

[8] W. Pauli, "Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren," *Zeitschrift für Physik*, vol. 31, no. 1, pp. 765–783, 1925. DOI: 10.1007/BF02980631.

[9] P. W. Atkins, J. De Paula and J. Keeler, *Atkins' Physical Chemistry*. Oxford University Press, 2018.

[10] J. F. Stanton, *A Chemist's Guide to Density Functional Theory By Wolfram Koch (German Chemical Society, Frankfurt am Main) and Max C. Holthausen (Humbolt University Berlin). Wiley-VCH: Weinheim. 2000. xiv + 294 pp. $79.95. ISBN 3-527-29918-1*, 11. 2001, vol. 123, pp. 2701–2701. DOI: 10.1021/ja004799q.

[11] H. Ehrenreich and M. H. Cohen, "Self-consistent field approach to the many-electron problem," *Physical Review*, vol. 115, no. 4, pp. 786–790, 1959. DOI: 10.1103/PhysRev.115.786.

[12] P. W. Langhoff, M. Karplus and R. P. Hurst, "Approximations to Hartree—Fock Perturbation Theory," *The Journal of Chemical Physics*, vol. 44, no. 2, pp. 505–514, Jan. 1966. DOI: 10.1063/1.1726717.

[13] D. L. Strout and G. E. Scuseria, "A quantitative study of the scaling properties of the Hartree-Fock method," *The Journal of Chemical Physics*, vol. 102, no. 21, pp. 8448–8452, 1995. DOI: 10.1063/1.468836.

[14] A. N. Kolmogorov, S. V. Fomin and S. V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, ser. Dover books on mathematics v. 1. Dover, 1999.

[15] D. Daners, "Introduction To Functional Analysis," *Physics Today*, vol. 12, no. 6, pp. 48–50, 1959. DOI: 10.1063/1.3060855.

[16] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects," *Phys. Rev.*, vol. 140, no. 4A, A1133–A1138, Nov. 1965. DOI: 10.1103/PhysRev.140.A1133.

[17] P. Haynes, "Linear-scaling methods in ab initio quantum-mechanical calculations," Ph.D. dissertation, University of Cambridge, 1998.

[18] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, ser. Density-functional Theory of Atoms and Molecules. Oxford University Press, USA, 1994.

[19] S. Mohr, L. E. Ratcliff, L. Genovese, D. Caliste, P. Boulanger, S. Goedecker and T. Deutsch, "Accurate and efficient linear scaling DFT calculations with universal applicability," *Physical Chemistry Chemical Physics*, vol. 17, no. 47, pp. 31 360–31 370, 2015. DOI: 10.1039/c5cp00437c.

[20] E. Prodan and W. Kohn, "Nearsightedness of electronic matter," pp. 0–3, 2005.

[21] Á. Ruiz-Serrano, N. D. Hine and C. K. Skylaris, "Pulay forces from localized orbitals optimized in situ using a psinc basis set," *Journal of Chemical Physics*, vol. 136, no. 23, 2012. DOI: 10.1063/1.4728026.

[22] N. D. Hine, P. D. Haynes, A. A. Mostofi, C. K. Skylaris and M. C. Payne, "Linear-scaling density-functional theory with tens of thousands of atoms: Expanding the scope and scale of calculations with ONETEP," *Computer Physics Communications*, vol. 180, no. 7, pp. 1041–1053, 2009. DOI: 10.1016/j.cpc.2008.12.023.

[23] X.-P. Li, R. W. Nunes and D. Vanderbilt, "Density-matrix electronic-structure method with linear system-size scaling," *Phys. Rev. B*, vol. 47, no. 16, pp. 10 891–10 894, Apr. 1993. DOI: 10.1103/PhysRevB.47.10891.

[24] P. D. Haynes, C. K. Skylaris, A. A. Mostofi and M. C. Payne, "Density kernel optimization in the ONETEP code," *Journal of Physics Condensed Matter*, vol. 20, no. 29, 2008. DOI: 10.1088/0953-8984/20/29/294207.

[25] C. K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne, "Implementation of linear-scaling plane wave density functional theory on parallel computers," *Physica Status Solidi (B) Basic Research*, vol. 243, no. 5, pp. 973–988, 2006. DOI: 10.1002/pssb.200541328.

[26] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek and G. R. Hutchison, "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform," *Journal of Cheminformatics*, vol. 4, no. 1, p. 17, Dec. 2012. DOI: 10.1186/1758-2946-4-17.

[27] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, "The atomic simulation environment{\textemdash}a Python library for working with atoms," *Journal of Physics: Condensed Matter*, vol. 29, no. 27, p. 273 002, Jun. 2017. DOI: 10.1088/1361-648x/aa680e.

[28] M. Challacombe, "A general parallel sparse-blocked matrix multiply for linear scaling SCF theory," *Computer Physics Communications*, vol. 128, no. 1-2, pp. 93–107, Jun. 2000. DOI: 10.1016/S0010-4655(00)00074-6.

[29] R. John and M. Centre, "Embedding high-level quantum mechanical approaches within linear-scaling density functional theory Supervisors : Declaration of Originality Declaration of Copyright," 2020. DOI: 10.25560/80436.

[30] W. Humphrey, A. Dalke and K. Schulten, "Sartorius products," *Journal of molecular graphics*, vol. 14, no. October 1995, pp. 33–38, 1996.

[31] Y. Wang, A. Kiziltas, P. Blanchard and T. R. Walsh, "Calculation of 1D and 2D densities in VMD: A flexible and easy-to-use code," *Computer Physics Communications*, vol. 266, p. 108 032, 2021. DOI: 10.1016/j.cpc.2021.108032.

[32] F. Jensen, *Introduction to Computational Chemistry*. Wiley, 2017.

[33] D. Marx and J. Hutter, *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*. Cambridge University Press, 2009.

[34] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf and O. Beckstein, "MDAnalysis: a toolkit for the analysis of molecular dynamics simulations.," eng, *Journal of computational chemistry*, vol. 32, no. 10, pp. 2319–2327, Jul. 2011. DOI: 10.1002/jcc.21787.

[35] A.-r. Allouche, "Software News and Updates Gabedit — A Graphical User Interface for Computational Chemistry Softwares," *Journal of computational chemistry*, vol. 32, pp. 174–182, 2012. DOI: 10.1002/jcc.

[36] C. K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne, "Implementation of linear-scaling plane wave density functional theory on parallel computers," *Physica Status Solidi (B) Basic Research*, vol. 243, no. 5, pp. 973–988, 2006. DOI: 10.1002/pssb.200541328.

Appendices

A Glossary

ASE - Atomic Simulation Environment
CG - Conjugate gradient
CPU - Central processing unit
DFT - Density Functional Theory
HF - Hartree-Fock
HK - Hohenberg-Kohn
LNV - Li, Nunes and Vanderbilt
MO - Molecular orbitals
NEM - Nearsightedness of electronic matter
NGWF - Non-Orthogonal General Wannier Functions
ONETEP - Order-N Electronic Total Energy Package
RMS - Root mean square (gradient)
VMD - Visual Molecular Dynamics

B Code

B.1 Outputting SPAM3 matrices

```
fileunit=utils_unit() ! Unique identifier for output file
write(SPAM3_lnv_out,'("SPAM3_ngwf",i0,"_lnv",i0,".out")') ngwf_cg_iter,iteration
open(unit=fileunit,file=SPAM3_lnv_out,form="formatted")
call sparse_show_matrix(aux%kern%m(1,1)%m(1,1),fileunit,show_elems=.true.)
close(fileunit)
```

B.2 Distance arrays - MDAnalysis

```
# Creating a universe for polypropylene from xyz file
inputfile_poly = input('Please specify path to inputfile\n')
u = mda.Universe(inputfile_poly)
# Atomgroups (object)
u_all = u.select_atoms('all')
# Distance array
dist_arr_all = distances.distance_array(u_all.positions,u_all.positions)
```

B.3 Atom-pair kernel cut-off

B.3.1 Initial density kernel

```
! Density kernel
[...]
! Sparsity pattern (in sparse_inital_mod)
call parallel_strategy_list_cross_overlaps(overlaps(jsub,isub), &
          elems_sfc1,elems_sfc2,mdl%cell,'Fixed','Fixed', &
          0.5_DP*range,pair_specific=.true.)
```

The pair_specific logical (boolean) argument is automatically set to false if not specified to avoid bugs in other parts of the code.

B.3.2 Cut-off

ONETEP creates the sparsity pattern in two big loops. The first and second loops are responsible for the num_overlaps array and the overlap_list array, respectively. Both loops were modified as follows:

```
! Square distance between two atoms, 1-2-3 stand for x-y-z coordinates
dist_sq = adiff(1)*adiff(1)+adiff(2)*adiff(2)+adiff(3)*adiff(3)
! AV : Atom-pair cut-off
if (loc_pair_specific == .true.) then
    ! local variables for elements
    local_el1 = elements1(iat)%symbol
    local_el2 = elements2(jat)%symbol
    ! N-N
    if (local_el1 == 'N'.and.local_el2 == 'N') then
        radii(1,iat) = 79.738_DP
        radii(2,jat) = 79.738_DP
    ! N-H
    else if ((local_el1 == 'N'.and.local_el2 == 'H').or.&
        (local_el1 == 'H'.and.local_el2 == 'N')) then
        radii(1,iat) = 69.893_DP
        radii(2,jat) = 69.893_DP
    [...]
    else if (local_el1 == 'O'.and.local_el2 == 'O') then
        radii(1,iat) = 86.673_DP
        radii(2,jat) = 86.673_DP
    else
        write(stdout,*)"Houston, pair-specific cutoff not found."
    end if
end if
cutoff = radii(1,iat) + radii(2,jat)
! AV : cutoff need to be halved
if (loc_pair_specific == .true.) then
    cutoff = cutoff/2
end if
! adding atoms to num_overlaps if distance is below cut-off
if (dist_sq <= cutoff*cutoff) then
```

```
    num_my_overlaps(iat) = num_my_overlaps(iat)+1
end if
```

The second loop ends differently after cut-off:

```
! adding atoms to overlap_list if distance is below cut-off
if (dist_sq <= cutoff*cutoff) then
    if (.not. overlapped(jat)) then
        num_my_overlaps(iat) = num_my_overlaps(iat)+1
        my_overlap_list(num_my_overlaps(iat),iat) = jat
        overlapped(jat) = .true.
    end if
end if
```

C Overlaps(jsub,isub)

Kernel cut-off /bohr	∞ (1000)	25
max_overlaps	218	102
num_overlaps(:)	218 218 218 [...] 218	52 54 99 [...] 52
overlap_list(:,:)	C1: 1 2 3 [...] 127 128 C2: 1 2 3 [...] 127 128 ⋮ H218: 1 2 3 [...] 127 128	C1: 1 2 5 [...] 217 218 C2: 56 60 61 [...] 201 202 ⋮ H218: 1 2 5 [...] 217 218

The polypropylene has 72+146=218 atoms which explains the figures in the ∞ cut-off column. As soon as truncation is applied, the density kernel becomes sparse. C1 and H128 are close to each other, which explains their similar pattern in the overlap_list(:,:) row.