PS 10: Final Project – Final Paper
Automated Computation Of Molecular Properties From First Principles

Alexander Lin
Supervised by Dr. Michael Mavros

December 19, 2018

1 Introduction

In the world of chemistry, we are greatly interested in the ability to accurately solve the Schrödinger equation. For molecular structures, solutions to the Schrödinger equation allow us to characterize many interesting properties, such as total energies, ionization potentials, and interatomic bond lengths, to name a few.

However, since there do not exist exact, analytical solutions to many-body electron systems, the scientific community has resorted to computational approaches that make numerical approximations. One of the most famous Schrödinger solvers that has been used throughout history is the Hartree-Fock algorithm [13]. By using a series of approximating methods – such as Born-Oppenheimer, the single Slater Determinant, and the variational method – Hartree-Fock allows us to compute an *upperbound* to the true ground state energy E of any given molecule, using nothing but the Cartesian coordinates and nuclear charges of its constituent atoms.

In this project, we automate the implementation of Hartree-Fock for small molecules and derive interesting, experimentally-verified properties from first principles. We begin by testing Hartree-Fock's ability to recover bond lengths of very simple inorganic molecules such as hydrogen fluoride and nitrogen gas. Next, we evaluate how well the algorithm is able to calculate the total energy of various multi-electron atoms.

Finally, we investigate the efficacy of Hartree-Fock in solving the Schrödinger equation for the GDB-13 dataset, an exhaustive enumeration of over 970 million organic and druglike molecules containing up to 13 atoms of cabon, nitrogen, oxygen, sulfur, and chlorine that are saturated with hydrogens [2]. We focus on a subset of this dataset – namely 59 small molecules with up to four non-hydrogen atoms from QM7b[1], which was created to reduce the original dataset to a more manageable number of structures while maintaining the rich diversity of GDB-13 [8]. For each organic molecule, we compare a Hartree-Fock calculation of its total energy and first ionization potential to ground-truth values. We also analyze trends

[1] Freely available at http://quantum-machine.org/datasets/.

within specific organic families such as alkanes and alkynes to show that Hartree-Fock can successfully recover these trends. We conclude the analysis with experimental evidence that the asymptotic running time of Hartree-Fock is $\mathcal{O}(K^4)$, where K is the number of orbitals considered by the algorithm for a given molecule.

The rest of this paper is organized as follows: Section 2 explains the mathematical theory behind the Hartree-Fock algorithm. Section 3 details the process of implementing the algorithm, along with some technical specifications. Section 4 presents the main results, thereby providing some evidence of the algorithm's utility. And finally, Section 5 concludes the paper and touches on some potential future work.

2 Theory

In this section, we heavily utilize the notation of [12] and [13]. We highly recommend interested readers to consult either of these comprehensive sources for additional information about Hartree-Fock theory.

2.1 Initial Approximations

Let us start with the Hamiltonian $\hat{H}$ and corresponding ground-state energy E_{tot} for a multi-electron system. The Hamiltonian can be characterized by five main components,

$$\hat{H} = \hat{T}_N(\boldsymbol{R}) + \hat{T}_e(\boldsymbol{r}) + \hat{V}_{NN}(\boldsymbol{R}, \boldsymbol{Z}) + \hat{V}_{eN}(\boldsymbol{r}, \boldsymbol{R}, \boldsymbol{Z}) + \hat{V}_{ee}(\boldsymbol{r}), \tag{1}$$

where $\hat{T}_N, \hat{T}_e$ respectively describe the kinetic energies of the nuclei and electrons; and $\hat{V}_{NN}, \hat{V}_{eN}, \hat{V}_{ee}$ respectively describe the Coulombic potential energies of nucleus-nucleus repulsion, electron-nucleus attraction, and electron-electron attraction. Here, $\boldsymbol{R} = \{\boldsymbol{R}_1, \ldots, \boldsymbol{R}_M\}$ and $\boldsymbol{r} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N\}$ are matrices that hold three-dimensional coordinates for the M nuclei and N electrons of the molecule in question. The vector $\boldsymbol{Z} = \{Z_1, \ldots, Z_M\}$ denotes the charges for the M nuclei. Note that $\boldsymbol{R}, \boldsymbol{Z}$ are inputs to the algorithm, whereas $\boldsymbol{r}$ is characterized by the wavefunction. In general, we will use $\{i, j, k\}$ to index electrons and $\{A, B, C\}$ to index nuclei.

The first approximation taken by Hartree-Fock is Born-Oppenheimer, which drops $\hat{T}_N$ from the Hamiltonian. The other four terms can be expanded as,

$$\hat{T}_e(\boldsymbol{r}) = -\frac{1}{2}\sum_{i=1}^{N} \nabla_i^2, \tag{2}$$

$$\hat{V}_{NN}(\boldsymbol{R}, \boldsymbol{Z}) = \sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}}, \tag{3}$$

$$\hat{V}_{eN}(\boldsymbol{r}, \boldsymbol{R}, \boldsymbol{Z}) = -\sum_{A=1}^{M}\sum_{i=1}^{N} \frac{Z_A}{r_{Ai}}, \tag{4}$$

$$\hat{V}_{ee}(\boldsymbol{r}) = \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}}, \tag{5}$$

where $R_{AB} = \|\boldsymbol{R}_B - \boldsymbol{R}_A\|_2$ is an internuclear distance, $r_{ij} = \|\boldsymbol{r}_j - \boldsymbol{r}_i\|_2$ is an electron-electron distance, and $r_{Ai} = \|\boldsymbol{r}_i - \boldsymbol{r}_A\|_2$ is an nucleus-electron distance. One immediate observation from Equation 3 is that the operator $\hat{V}_{NN}$ has no dependence on electron coordinates $\boldsymbol{r}$; therefore, we can simply calculate this quantity at the beginning of the algorithm and leave it to the side. It follows that the electronic Schrödinger equation may be simplified as

$$\hat{H}_{ele}\Psi(\boldsymbol{r};\boldsymbol{R},\boldsymbol{Z}) = \left[\hat{T}_e(\boldsymbol{r}) + \hat{V}_{eN}(\boldsymbol{r},\boldsymbol{R},\boldsymbol{Z}) + \hat{V}_{ee}(\boldsymbol{r})\right]\Psi(\boldsymbol{r};\boldsymbol{R},\boldsymbol{Z}) = E_{ele}\Psi(\boldsymbol{r};\boldsymbol{R},\boldsymbol{Z}), \tag{6}$$

where the total energy of the multi-electron system $E_{tot} = E_{ele} + V_{NN}$ is the sum of the electronic and nuclear energies.

The antisymmetry principle states that for a system of fermions, the wavefunction must be antisymmetric with respect to changes in position *and* spin of any two fermions [12]. To satisfy this principle, we must introduce a new variable (i.e. the spin coordinate ω) for each electron and define the wavefunction Ψ in terms of $\boldsymbol{x} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, where each $\boldsymbol{x}_i = \{\boldsymbol{r}_i, \omega_i\}$. The following equation describes antisymmetry for any two electrons i, j,

$$\Psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_j, \dots, \boldsymbol{x}_N) = -\Psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_j, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_N). \tag{7}$$

We would like to express Ψ as some aggregated function of single-electron, molecular wavefunctions $\chi_1, \dots, \chi_n$ to make calculations easier. Perhaps the most straightforward way to do this while satisfying Equation 7 – and its immediate corollary, the Pauli exclusion principle – is to let Ψ be a Slater determinant,

$$\Psi(\boldsymbol{x};\boldsymbol{R},\boldsymbol{Z}) = \frac{1}{\sqrt{N!}}\begin{vmatrix} \chi_1(\boldsymbol{x}_1) & \chi_2(\boldsymbol{x}_1) & \cdots & \chi_N(\boldsymbol{x}_1) \\ \chi_1(\boldsymbol{x}_2) & \chi_2(\boldsymbol{x}_2) & \cdots & \chi_N(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\boldsymbol{x}_N) & \chi_2(\boldsymbol{x}_N) & \cdots & \chi_N(\boldsymbol{x}_N) \end{vmatrix}. \tag{8}$$

That is, we assume Ψ to be an anti-symmetric product-sum. It is this assumption that makes Hartree-Fock a *mean field approximation*, which means that each electron feels the average repulsive cloud of other electrons, but not the individual effects. For this reason, the Hartree-Fock model fails to capture certain real-world phenomena, such as London dispersion, that occur between specific sets of electrons. There exists an entire body of literature on post-Hartree-Fock methods for improving this approximation [1].

2.2 Hartree-Fock Energy

Following Sherrill's notation [12], we can compactly re-express Equations 2 and 4 as a one-electron operator $\hat{h}$,

$$\hat{T}_e(\boldsymbol{r}) + \hat{V}_{eN}(\boldsymbol{r},\boldsymbol{R},\boldsymbol{Z}) = \sum_i \left(-\frac{1}{2}\nabla_i^2 - \sum_A \frac{Z_A}{r_{iA}}\right) = \sum_i \hat{h}(i), \tag{9}$$

and Equation 5 as a two-electron operator $\hat{v}$,

$$\hat{V}_{ee}(\boldsymbol{r}) = \sum_{i<j} \hat{v}(i,j) = \sum_{i<j} \frac{1}{r_{ij}}. \tag{10}$$

It follows that we can simply re-write the electronic Hamiltonian of Equation 6 as

$$\hat{H}_{ele} = \sum_i \hat{h}(i) + \sum_{i<j} \hat{v}(i,j). \tag{11}$$

In finding the ground-state energy of a molecule, the goal of Hartree-Fock is to solve the following optimization problem,

$$E_{HF} = \min_{\Psi} E_{ele} = \min_{\Psi} \langle \Psi | \hat{H}_{el} | \Psi \rangle = \min_{\chi_1, \dots, \chi_N} \sum_{i=1}^{n} \langle i | \hat{h} | i \rangle + \sum_{i=1}^{n} \sum_{j=i+1}^{n} [ii|jj] - [ij|ji], \tag{12}$$

where

$$\langle i | \hat{h} | i \rangle = \int_{\mathbb{R}} \chi_i^*(\boldsymbol{x}) \hat{h}(i) \chi_i(\boldsymbol{x}) d\boldsymbol{x}, \tag{13}$$

$$[ii|jj] = \int_{\mathbb{R}^2} \chi_i^*(\boldsymbol{x}_1) \chi_i(\boldsymbol{x}_1) \hat{v}(i,j) \chi_j^*(\boldsymbol{x}_2) \chi_j(\boldsymbol{x}_2) d\boldsymbol{x}_1 \boldsymbol{x}_2, \tag{14}$$

$$[ij|ji] = \int_{\mathbb{R}^2} \chi_i^*(\boldsymbol{x}_1) \chi_j(\boldsymbol{x}_1) \hat{v}(i,j) \chi_j^*(\boldsymbol{x}_2) \chi_i(\boldsymbol{x}_2) d\boldsymbol{x}_1 \boldsymbol{x}_2. \tag{15}$$

Working through Lagrange's method of undetermined multipliers for this optimization, as detailed in [11], we arrive at the eigenvalue problem,

$$f(\boldsymbol{x}_1)\chi_i(\boldsymbol{x}_1) = \epsilon_i \chi_i(\boldsymbol{x}_1), \tag{16}$$

where the Fock operator f is defined by

$$\begin{aligned} f(\boldsymbol{x}_1)\chi_i(\boldsymbol{x}_1) = h(\boldsymbol{x}_1)\chi_i(\boldsymbol{x}_1) &+ \sum_{j \neq i} \left[\int |\chi_j(\boldsymbol{x}_2)|^2 \frac{1}{r_{12}} d\boldsymbol{x}_2 \right] \chi_i(\boldsymbol{x}_1) \\ &- \sum_{j \neq i} \left[\int \chi_j^*(\boldsymbol{x}_2) \chi_i(\boldsymbol{x}_2) \frac{1}{r_{12}} d\boldsymbol{x}_2 \right] \chi_j(\boldsymbol{x}_1), \end{aligned} \tag{17}$$

and ϵ_i is the energy of molecular orbital i. The operator involves the integration of complicated expressions, so to make things analytically tractable, we introduce a basis set of easy-to-integrate atomic orbitals $\tilde{\chi}_1, \dots, \tilde{\chi}_K$. Typically, $\tilde{\chi}_\mu$ is a linear combination of Gaussians whose coefficients have been optimized to fit Slater-type orbitals; we elaborate more on this in Section 3. In doing so, we employ the variational principle – another source of approximation.

We now have that each molecular orbital i is a linear combination of atomic orbitals (i.e. MO-LCAO method) with coefficients $C_{1i}, \dots, C_{Ki}$,

$$\chi_i = \sum_{\mu=1}^{K} C_{\mu i} \tilde{\chi}_\mu. \tag{18}$$

From this, we can rewrite Equation 16 as the Hartree-Fock-Roothaan equations,

$$\sum_\nu F_{\mu\nu} C_{\nu i} = \epsilon_i \sum_\nu S_{\mu\nu} C_{\nu i}, \tag{19}$$

where we have the more tractable integrals,

$$S_{\mu\nu} = \int \tilde{\chi}_\mu^*(\boldsymbol{x}_1)\tilde{\chi}_\nu(\boldsymbol{x}_1)d\boldsymbol{x}_1, \tag{20}$$

$$F_{\mu\nu} = \int \tilde{\chi}_\mu^*(\boldsymbol{x}_1)f(\boldsymbol{x}_1)\tilde{\chi}_\nu(\boldsymbol{x}_1)d\boldsymbol{x}_1. \tag{21}$$

In matrix form, Equation 19 can be written as

$$\boldsymbol{FC} = \boldsymbol{SC\epsilon} \tag{22}$$

Equation 22 is a peculiar eigenvalue equation, because $\boldsymbol{F}$ depends on $\boldsymbol{C}$ and vice-versa. This means that both cannot be optimized simultaneously; instead, the Hartree-Fock algorithm must alternately update these two matrices until convergence.

2.3 Hartree-Fock Algorithm

The Hartree-Fock algorithm [13] can be divided into two main parts – (1) integration and (2) iteration. The integration part tends to dominate in terms of computation time.

During the integration part, there are four main integrals of interest that need to be pre-computed. The first is the overlap integral $S_{\mu\nu}$ for every pair of atomic orbitals μ, ν, as described by Equation 20. The other three – kinetic energy $T_{\mu\nu}$, nuclear-electron attraction $V_{\mu\nu}^{\text{nucl}}$, and electron-electron repulsion $[\mu\nu|\lambda\sigma]$ – are involved in the Fock integral of Equation 21. Their expressions come straight from the operators – as defined by Equations 2, 4, and 5 – applied to the atomic basis functions,

$$T_{\mu\nu} = \int \tilde{\chi}_\mu^*(\boldsymbol{x}_1)\left[-\frac{1}{2}\nabla_1^2\right]\tilde{\chi}_\nu(\boldsymbol{x}_1)d\boldsymbol{r}_1, \tag{23}$$

$$V_{\mu\nu}^{\text{nucl}} = \int \tilde{\chi}_\mu^*(\boldsymbol{x}_1)\left[-\sum_A \frac{Z_A}{r_{A1}}\right]\tilde{\chi}_\nu(\boldsymbol{x}_1)d\boldsymbol{r}_1, \tag{24}$$

$$[\mu\nu|\lambda\sigma] = \int\int \tilde{\chi}_\mu^*(\boldsymbol{x}_1)\tilde{\chi}_\nu(\boldsymbol{x}_1)\left[\frac{1}{r_{12}}\right]\tilde{\chi}_\lambda^*(\boldsymbol{x}_2)\tilde{\chi}_\sigma(\boldsymbol{x}_2)d\boldsymbol{r}_1\boldsymbol{r}_2. \tag{25}$$

After these integrals are computed, the algorithm proceeds by alternately changing $\boldsymbol{F}$ and $\boldsymbol{C}$. A full description is given by [13] and summarized in Algorithm 1. Note that this is the restricted Hartree-Fock procedure, which treats electrons as paired fermions and works for an even number of electrons.

3 Implementation

Our implementation of Hartree-Fock follows Algorithm 1 and is written from scratch. All code has been open-sourced and made available on GitHub[2]. The algorithm is implemented using the Python programming language. Almost everything involves the NumPy computing package, but we do use the SciPy computing package as well to calculate the Boys function [5], which is briefly needed in the integration of Gaussians.

[2] See https://github.com/al5250/ps_10_final_project for project code.

Algorithm 1 Restricted Hartree-Fock

```
Input: nuclear coords $\boldsymbol{R}$, charges $\boldsymbol{Z}$, atomic basis functions $\tilde{\chi}_\mu$, num of electrons $N$
Compute nuclear-nuclear repulsion $V_{NN}$.
Compute integrals $S_{\mu\nu}$, $T_{\mu\nu}$, $V_{\mu\nu}^{\text{nucl}}$, $[\mu\nu|\lambda\sigma]$.
Construct orthonormal basis transformation matrix $\boldsymbol{X}$ from $\boldsymbol{S}$ using canonical method.
Initialize density matrix $\boldsymbol{P} = \boldsymbol{0}$.
Initialize Fock matrix $F_{\mu\nu} = T_{\mu\nu} + V_{\mu\nu}^{\text{nucl}}$.
while $\boldsymbol{P}$ has not converged do
    Calculate transformed Fock matrix $\boldsymbol{F}' = \boldsymbol{X}^T\boldsymbol{F}\boldsymbol{X}$.
    Diagonalize $\boldsymbol{F'}$ to get eigenvectors $\boldsymbol{C'}$ and eigenvalues $\boldsymbol{\epsilon}$.
    Calculate $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{C'}$.
    Compute density matrix $P_{\mu\nu} = \sum_{i=1}^{N/2} C_{\mu i}C_{\nu i}$.
    Compute matrix $G_{\mu\nu} = \sum_{\lambda,\sigma} P_{\lambda\sigma}\,(2 \cdot [\mu\nu|\sigma\lambda] - [\mu\lambda|\sigma\nu])$.
    Compute Fock matrix $F_{\mu\nu} = T_{\mu\nu} + V_{\mu\nu}^{\text{nucl}} + G_{\mu\nu}$.
    Compute electronic energy $E_{ele} = \sum_{\mu,\nu} P_{\mu\nu} \cdot (T_{\mu\nu} + V_{\mu\nu}^{\text{nucl}} + F_{\mu\nu})$.
end while
Compute total energy $E_{tot} = E_{ele} + V_{NN}$.
Output: $E_{tot}, \boldsymbol{\epsilon}, \mathbf{F}, \mathbf{P}, \mathbf{T}, \mathbf{V}^{\text{nucl}}$
```

3.1 Atomic Basis Set – STO-3G

The specific atomic orbital basis set that we choose to use is STO-3G, which was compiled by [4] and [10]. This is a minimal basis set; that is, a basis set that has the smallest possible size for reasonable results. Each atom's orbital $\tilde{\chi}_\mu$ is a probability distribution function over possible locations $\boldsymbol{r}_i$ for a given electron i; it follows the form of a linear combination between three 3-dimensional Gaussians g_1, g_2, g_3 whose parameters $\alpha_{p\mu}$ and coefficients $d_{p\mu}$ have been pre-fit to a Slater-type orbital,

$$\tilde{\chi}_\mu(\boldsymbol{r}_i) \propto \sum_{p=1}^{3} d_{p\mu} \cdot g_p(\boldsymbol{r}_i; \boldsymbol{R}_\mu, \alpha_{p\mu}). \tag{26}$$

Each Gaussian g_p is centered at the nucleus $\boldsymbol{R}_\mu$ of atom μ and has an inverse-variance (sometimes also called "exponential") parameter $\alpha_{p\mu}$ that governs the spread of the distribution. The Gaussian may take different forms depending on the orbital's azimuthal and magnetic quantum numbers,

$$g_p^{1s}(\boldsymbol{r}_i; \boldsymbol{R}_\mu, \alpha_{p\mu}) = \left(\frac{8\alpha_{p\mu}^3}{\pi^3}\right)^{1/4} \exp(-\alpha_{p\mu} r_{i\mu}^2), \tag{27}$$

$$g_p^{2p_x}(\boldsymbol{r}_i; \boldsymbol{R}_\mu, \alpha_{p\mu}) = \left(\frac{128\alpha_{p\mu}^5}{\pi^3}\right)^{1/4} (x_i - X_\mu) \exp(-\alpha_{p\mu} r_{i\mu}^2), \tag{28}$$

$$g_p^{2p_y}(\boldsymbol{r}_i; \boldsymbol{R}_\mu, \alpha_{p\mu}) = \left(\frac{128\alpha_{p\mu}^5}{\pi^3}\right)^{1/4} (y_i - Y_\mu) \exp(-\alpha_{p\mu} r_{i\mu}^2), \tag{29}$$

$$g_p^{2p_z}(\boldsymbol{r}_i; \boldsymbol{R}_\mu, \alpha_{p\mu}) = \left(\frac{128\alpha_{p\mu}^5}{\pi^3}\right)^{1/4} (z_i - Z_\mu) \exp(-\alpha_{p\mu} r_{i\mu}^2). \tag{30}$$

Note that in these equations, we implicitly unpack the vectors $\boldsymbol{r}_i = [x_i, y_i, z_i]^T$ and $\boldsymbol{R}_\mu = [X_\mu, Y_\mu, Z_\mu]^T$.

There exist other basis sets STO-PG for $P = 1, 2, \ldots, 6$, but STO-3G is perhaps the most widely used. Our code is general enough to apply to any STO-PG basis set. Also, there are more complicated basis sets such as 4-31G and 6-31G* [10], but these are outside the scope of this study.

3.2 Integration

Since STO-3G atomic orbitals are fully characterized by Gaussians, we can apply rules for integration of Gaussian products to compute $S_{\mu\nu}$, $T_{\mu\nu}$, $V_{\mu\nu}^{\text{nucl}}$, and $[\mu\nu|\lambda\sigma]$. The key insight is that integration of a product of two Gaussian probability distribution functions yields yet another Gaussian; thus, integration of any two Gaussians can be done in $\mathcal{O}(1)$-time.

For atomic orbitals with higher azimuthal quantum numbers (e.g. p-type orbitals), we must use recurrence relations to perform the integrations. The mathematics become trickier and we defer the reader to [5] for more information[3]. In particular, we follow the McMurchie-Davidson scheme in our code. Since the results of Section 4 only involves small molecules, our implementation currently handles molecules containing elements up to the third row of the periodic table; some modifications will need to be made in order to handle d-type and f-type orbitals.

The integration is also the rate-determining step of the Algorithm 1. Specifically, computing $[\mu\nu|\lambda\sigma]$ will take $\mathcal{O}(K^4)$-time, where K is the number of atomic orbitals across all atoms in the molecule. There are methods such as pre-screening electron-electron repulsion integrals [5] that cut down the computation time; we implement pre-screening in our code, but for small molecules, we have found this modification to make little empirical difference.

3.3 Code Description

The main code is modularized in three Python files:

- `utils.py` – useful helper functions for parsing STO-3G basis set file, computing normalization coefficients and other useful quantities, etc.
- `integrals.py` – functions for calculating the four integrals in the integration step of the algorithm
- `hf.py` – the main Hartree-Fock algorithm as detailed in Algorithm 1; the primary wrapper function that covers all computation is called `hartree_fock`.

There are also iPython notebooks in the repository that contain sample code for the experimental results detailed in Section 4. Basis sets are found in the `basis_sets` directory, while the data (including QM7b [8]) is found in the `data` directory.

[3] The author of the text also has a set of accompanying slides at http://folk.uio.no/helgaker/talks/SostrupIntegrals_10.pdf

4 Results

Using Hartree-Fock, we first reproduce some of the bond length results in [13] for very small molecules. Then, we calculate ground-state energy levels for all multi-electron atoms up to the third row in the periodic table and cross-reference the approximations with the true values in [3]. And finally, we provide a set of results on a dataset of 59 small organic molecules from QM7b [8]. All Hartree-Fock calculations are performed using the STO-3G basis set [10].

4.1 Equilibrium Bond Length

For diatomic molecules, the equilibrium bond length is the distance r between the two atoms such that E_{tot} is minimized. For the molecules FH and N_2, we use our implementation of Algorithm 1 to plot the curve corresponding to E_{tot} as a function of r in Figure 1.

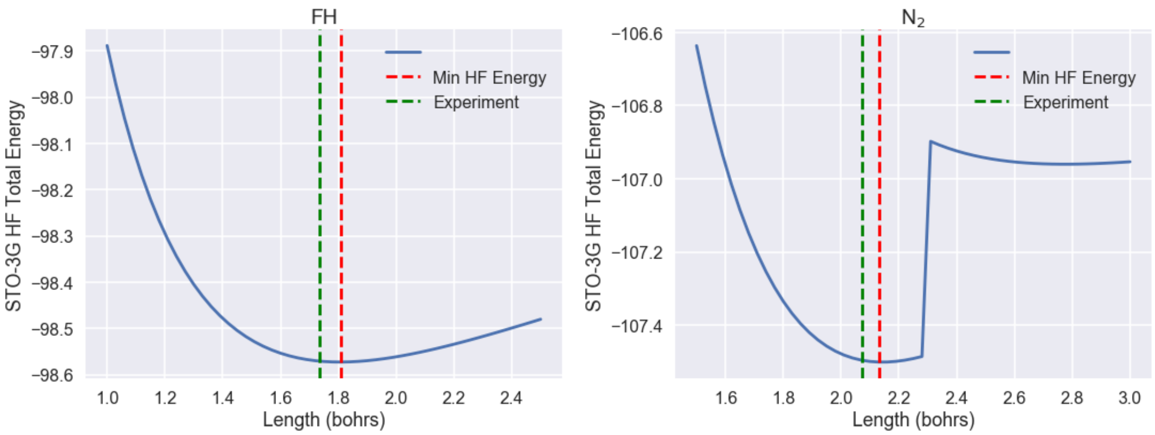

Figure 1: Total energy curves as a function of internuclear distance for diatomic molecules FH (left) and N_2 (right). The dotted green line denotes the minimum of the blue curve, which was computed from first principles via Hartree-Fock. The dotted red line denotes the true equilibrium bond length as found in experimental settings [13]. The errors are less than 10% for both molecules.

Notice that for N_2, we see a sudden jump in the Hartree-Fock energy at $r \approx 2.3$. This is because the algorithm fails to converge to the true global minimum for the electronic energy E_{ele} for $r \geq 2.3$; the algorithm could very well be identifying the 1st excited state, for example, instead of the ground state.

At this point, a natural question to ask if how limiting is the specific basis set we chose? Could a different basis set (e.g. one more sophisticated than STO-3G) yield better results? How much better would these results be? For the sake of completeness, we show bond length results across different basis sets in Table 1. These results are taken from [13].

	STO-3G	4-31G	6-31G*	Near-HF-limit	Experiment
FH	1.807	1.742	1.722	1.696	1.733
N_2	2.143	2.050	2.039	2.013	2.074

Table 1: Equilibrium bond length results derived using the Hartree-Fock algorithm for different basis sets. All values are recorded in bohrs and taken from [13]. The results are not completely consistent with experimental data because of Hartree-Fock's inability to capture electron correlation effects [1].

4.2 Atomic Energy

Another use of Hartree-Fock is to calculate the total energy of a multi-electron atom. However, given the *restricted* nature of the algorithm that was implemented, we must input an even number of N electrons. Needless to say, not all neutral atoms have an even number of electrons. Thus, we use the following scheme:

- If atom X has an even number of N electrons, then simply use Algorithm 1 to find $E_{tot}(X)$.
- If atom X has an odd number of N electrons, then X^- must have an even number of $N + 1$ electrons. Thus, we use Algorithm 1 to find $E_{tot}(X^-)$. By Koopman's Theorem, the first ionization potential of an atom or molecule can be approximated by the negative of its highest occupied molecular orbital (HOMO) energy [13]. The HOMO energy is an output of Algorithm 1, specifically $\epsilon_{(N+1)/2}(X^-)$. It follows that an approximation to $E_{tot}(X)$ is

$$E_{tot}(X) \approx E_{tot}(X^-) - \epsilon_{(N+1)/2}(X^-). \tag{31}$$

Figure 2 illustrates a comparison between the atomic energy as calculated by Hartree-Fock and a value that is essentially exact [3] for atoms up to atomic number $Z = 18$.

4.3 Analysis of Small Organic Molecules

In this extended analysis, we report results for a small dataset of organic molecules containing up to four non-hydrogen atoms that are either carbon, nitrogen, or oxygen. There are 59 molecules in total; the breakdown is 1 molecule with 1 non-hydrogen, 3 molecules with 2 non-hydrogens, 12 molecules with 3 non-hydrogens, and 43 molecules with 4 non-hydrogens.

We begin by comparing differences between Algorithm 1's calculation of the total molecular energy and ground-truth ”exact” values calculated using density functional theory that are provided by [8], the original source of the dataset. The results are presented in Figure 3. The mean absolute error is 7591 kcal per mol and the median absolute error is 7971 kcal per mol. Notice that one of the organic molecules has an usually high error – this is because the Hartree-Fock algorithm failed to converge within 100 iterations during the calculation of its total energy. In general, the Hartree-Fock calculations are not very accurate, but the trends and patterns can be recovered.

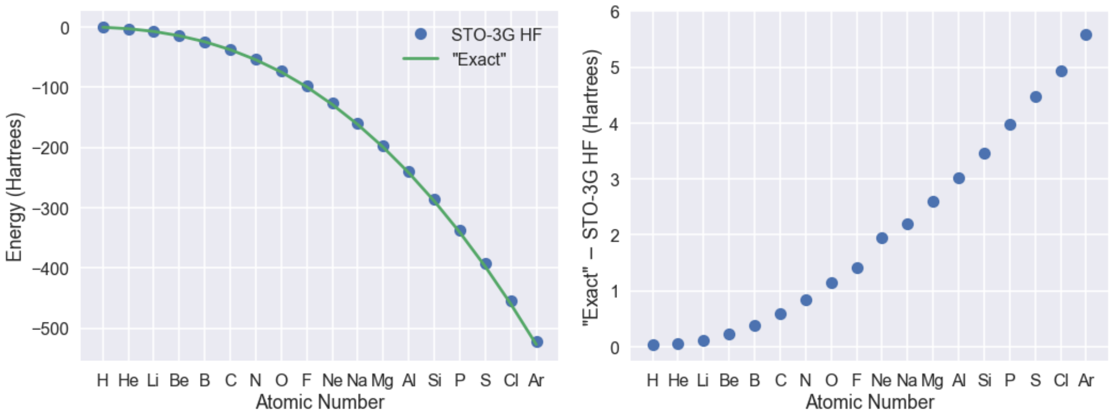

Figure 2: A comparison of values calculated by a scheme based on Algorithm 1 and values taken from an essentially exact source [3] for the total energy of a multi-electron atom (left). The difference between the exact value and the calculated value seems to increase as the number of electrons increases, which is intuitive (right).

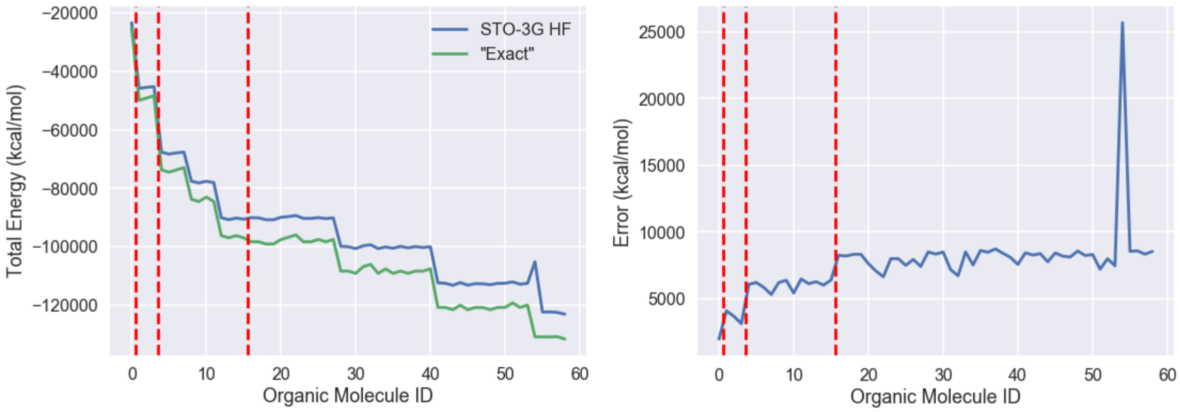

Figure 3: A comparison of total energy values calculated by a scheme based on Algorithm 1 and values taken from an essentially exact source [8] (left). The dotted lines denote the delineations between 1 non-hydrogen, 2 non-hydrogen, 3 non-hydrogen, and 4 non-hydrogen structures. As expected, the error increases as the complexity of the system increases (right). Calculations for organic molecule #54 failed to converge.

The most accurate set of results in this section are Hartree-Fock calculations of first ionization potentials for the organic molecules. Given a molecule with N electrons, we apply Koopman's Theorem to return $-\epsilon_{N/2}$ as the ionization potential [13]. We compare our calculated values to those of [8]. The complete accuracy results are given in Figure 4. The mean absolute error is 3.6 eV with a mean absolute percentage error of 36%. The median absolute error is 2.5 eV with a mean absolute percentage error of 24%. The trends across certain organic families are also consistently maintained by the Hartree-Fock algorithm, which is shown in Figure 5.

Finally, we perform an analysis of the running time of the algorithm as a function of the number of atomic orbitals K. In Section 2, we analytically derived this running time as $\mathcal{O}(K^4)$. In Figure 6, we empirically show that this is the case.

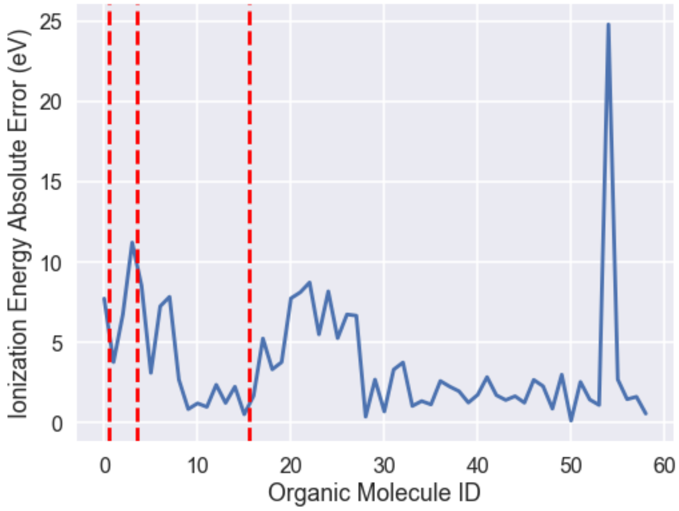

Figure 4: The absolute error in first ionization potentials for different small organic molecules. Ground truth is taken as density functional theory values reported in [8]. Again, the one outlier is a result of a failure of the algorithm to converge.

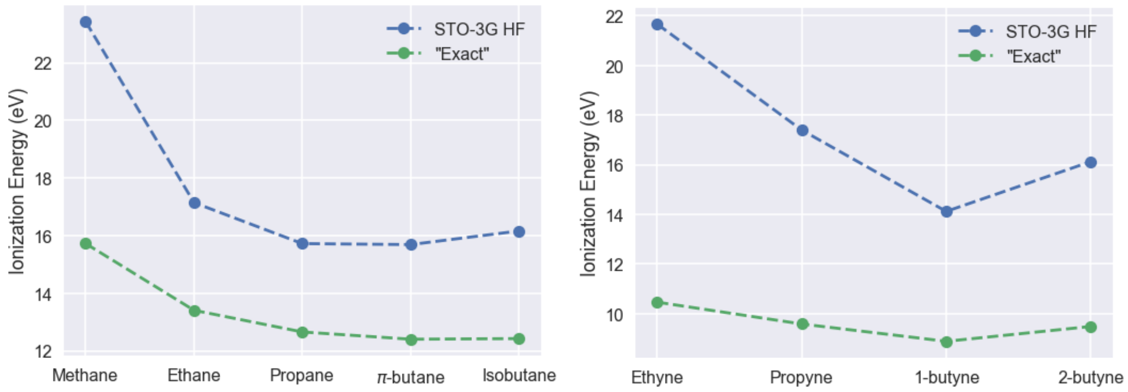

Figure 5: Ionization energy calculations across different organic families – alkanes (left) and alkynes (right). The correct trends are maintained by the Hartree-Fock algorithm.

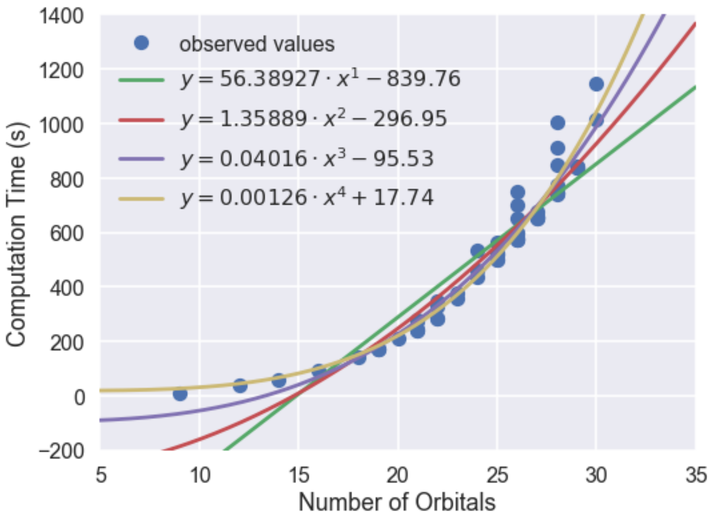

Figure 6: Fitting different polynomial curves to the running time of the algorithm versus number of atomic orbitals. From this is analysis, it is clear that $\mathcal{O}(K^4)$ fits the best.

5 Discussion and Conclusion

In conclusion, we fully implemented the Hartree-Fock algorithm from scratch in Python. Using this algorithm, we provided several results on calculating interesting properties, such as equilibrium bond length, total molecular energy, and first ionization potential. In general, the Hartree-Fock algorithm provides useful calculations that are not exact, yet uncover interesting and accurate trends.

In future work, there are several avenues for improvement. For example, we can implement *unrestricted* Hartree-Fock to directly calculate energies for molecules and atoms containing single-electron orbitals. To improve accuracy, we can also try implementing other basis sets such as 4-31G or 6-31G*, which may give better results than the minimal STO-3G. We can also explore implementation of post-Hartree-Fock methods such as coupled cluster, which takes electron correlation into account to improve algorithm accuracy. However, it is well-known that the running time for coupled cluster may be as high as $\mathcal{O}(K^8)$ [1], which means that we would need to take into account a tradeoff between speed and accuracy. With improved accuracy, we can explore the calculation of other interesting molecular properties, such as atomization energy, that are hard to calculate accurately under the current Hartree-Fock paradigm.

Finally, there is interest in seeing if other, non-physics-based approaches hold any promise in calculating these molecular energies accurately. The $\mathcal{O}(K^4)$ running time of Hartree-Fock is certainly a limitation for larger molecules; perhaps statistics-based approaches such as machine learning could be useful in more efficiently calculating these values. This is certainly an active area of research, especially with the success of deep learning neural networks in making progress towards many different regression problems [7, 8, 9, 6].

References

[1] Rodney J Bartlett and John F Stanton. Applications of post-hartreefock methods: A tutorial. *Reviews in computational chemistry*, pages 65–169, 1994.

[2] Lorenz C Blum and Jean-Louis Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *Journal of the American Chemical Society*, 131(25):8732–8733, 2009.

[3] Carlos F Bunge, Jose A Barrientos, and A Vivier Bunge. Roothaan-hartree-fock ground-state atomic wave functions: Slater-type orbital expansions and expectation values for z= 2-54. *Atomic data and nuclear data tables*, 53(1):113–162, 1993.

[4] David Feller. The role of databases in support of computational chemistry calculations. *Journal of computational chemistry*, 17(13):1571–1586, 1996.

[5] Trygve Helgaker, Poul Jorgensen, and Jeppe Olsen. *Molecular electronic-structure theory*. John Wiley & Sons, 2014.

[6] Kyle Mills, Michael Spanner, and Isaac Tamblyn. Deep learning and the schrödinger equation. *Physical Review A*, 96(4):042113, 2017.

[7] Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole V Lilienfeld, and Klaus-Robert Müller. Learning invariant representations of molecules for atomization energy prediction. In *Advances in Neural Information Processing Systems*, pages 440–448, 2012.

[8] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, 2013.

[9] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[10] Karen L Schuchardt, Brett T Didier, Todd Elsethagen, Lisong Sun, Vidhya Gurumoorthi, Jared Chase, Jun Li, and Theresa L Windus. Basis set exchange: a community database for computational sciences. *Journal of chemical information and modeling*, 47(3):1045–1052, 2007.

[11] C David Sherrill. An introduction to hartree-fock molecular orbital theory. *Georgia inst. of technology*, 2000.

[12] C David Sherrill. A brief review of elementary quantum chemistry. *Georgia Institute of Technology, School of Chemistry and Biochemistry*, 2001.

[13] Attila Szabo and Neil S Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.