

Constructing Random Ensembles of Fermionic Systems

Shaket Chaudhary

Thesis Advisor: Professor James Daniel Whitfield

Dartmouth College

Department of Computer Science

May 30, 2019

Abstract

The use of a computer has become a necessity for the modeling of quantum systems. Advances in computational power and algorithms have allowed scientists to approximate the structures and properties of various molecules. Yet as the size of the quantum system grows, the computational time and space complexity grows exponentially. An alternative approach was proposed in a separate study by Dave Wecker which forgoes the expensive computational requirements for molecule simulations. Rather than calculating out the terms of the Hamiltonian in Schrodinger's equation, artificial molecules were modeled by randomly sampling from probability distribution functions. In this paper we explore the validity of this approach by testing our own calculated energies of the hydrogen molecule against the randomly sampled Wecker ensemble. We find that overall the distributions generated randomly range from being largely off from reality to being of a similar enough distribution that the study invites further investigation. This suggests there may be value in random sampling so long as we choose the appropriate probability distribution with the correct parameters.

1 Introduction

The use of computers for computational chemistry has been very important since the computer's inception. It is only for simple cases such as the particle in a box, the harmonic oscillator, and the hydrogen atom for which we have analytic solutions. Apart from these, the quantum many-body problem cannot be solved analytically. Therefore, the aid of a computer simulation is necessary to calculate the structures and properties of molecules and solids. Combined with theoretical physics and chemistry, we are able to use ab initio and semi-empirical methods to approximate molecular behavior. Yet even with the exponential growth of computing speed, the sheer amount of operations necessary make computational chemistry problems hard. In short, due to quantum mechanics, we need the state space to represent all possible superpositions of particles.

For this research paper, we explore if we can characterize electronic structures of molecules using other methods. In a study by [8], they simulate artificial molecules by randomly sampling from various distributions. However, in a follow up paper, this methodology was abandoned as they felt it was not a viable method to simulate the electronic structure of molecules. We therefore reopen the investigation by first

comparing random sampling results to actual integral computations of the hydrogen molecule. From there we hope to uncover a better parametrization of the molecules behavior.

1.1 Overview

The paper is organized as follows. In section 2, we start off by familiarizing the reader with the necessary background we will use. A general knowledge of linear algebra is assumed. When dealing with molecules and atomic systems, it is convenient to use the atomic units. From there we give the reader a background on quantum mechanics so that they may later understand what the electronic structure problem is. Throughout the paper, dirac or bra-ket notation will be commonly used and we give a brief review of this as well. Finally, any special variables or symbols, such as the Hamiltonian will be touched upon.

After the reader familiarizes the language of quantum mechanics, we delve into electronic structure in section 3. The overall problem is introduced which due to its calculation complexity requires approximations such as using the Born-Oppenheimer approximation, the Hartree-Fock approximation, the discrete basis set approximation. Near the end of the chapter we arrive at the calculations of interest for this paper, the one electron and two electron integrals.

Section 4 provides a brief overview of how we further compute the calculations of these integrals. We introduce the notion of basis sets used to generate the wave functions necessary for the integral calculations. Section 5 then reviews the idea of Dave Wecker and details his intuition. That is, perhaps we can model these integral calculations by randomly sampling from certain distribution functions.

It is only after all this background that we detail the experimental setup in section 6. We recreate the random sampling method of Wecker. The methodology of plotting these results into a discretized graph is touched upon. Then the distributions will be compared against computed integrals and we investigate the results in section 7. Section 8 therefore ends with our conclusions and determines possible further directions for study.'

2 Quantum Mechanics Review

2.1 Atomic Units

When dealing with quantum and molecular systems, it is more convenient to use the unit system known as atomic units rather than SI units. This simplifies many procedures due to setting the elementary charge on a proton, the mass of an electron, Planck's reduced constant (angular momentum), and the electrostatic force constant of Coulomb's law, all to unity. A table is provided for an overview of converting between systems.

[7]

Physical quantity	Conversion factor	Value (SI)
Length	a_0	$5.2918 \cdot 10^{-11}$ m
Mass	m_e	$9.1095 \cdot 10^{-31}$ kg
Charge	e	$1.6022 \cdot 10^{-19}$ C
Energy	\mathcal{E}	$4.3598 \cdot 10^{-18}$ J
Angular momentum	\hbar	$1.0546 \cdot 10^{-34}$ J-s

2.2 Quantum Theory

Quantum mechanics is a fundamental theory of physics which deals with describing reality at the smallest scales. There are many aspects of quantum theory, however the following postulates should be sufficient background to understand the electronic structure problem which will be introduced in the next section

First, we can describe the state of a quantum mechanical system completely by its wave function Ψ . Wave functions obey the principle of superposition and exist in a Hilbert space. In other words, we can represent the wave function as a linear combination of other wave functions. When ϕ_n are the eigenstates and $\Psi = \sum c_n \phi_n$, the wave functions form an orthonormal basis that span the Hilbert space. [4]

For every observable in classical mechanics, such as energy, there is a corresponding linear Hermitian operator in quantum mechanics. The Hermitian property, $A = (A^*)^T$, implies the eigenvalues will be real. For energy, the Hermitian operator is the Hamiltonian \hat{H} . When a measurement is performed, the system collapses to one of its eigenvectors and we measure the associated real eigenvalue. The probability of measuring an eigenvalue E_n of a Hermitian operator, is $|c_n|^2$.

The previously mentioned wavefunction Ψ must satisfy the Schrödinger eigenvalue equation: $\hat{H}\Psi = E\Psi$. Solutions to this equation evolve in time through the unitary operator $|\psi\rangle = e^{-i\hat{H}t} |\psi_0\rangle$. Here, we introduce bra ket notation, also known as Dirac notation for conveying the language of quantum mechanics.

It is worth noting that a wave function needs to either be symmetric if the system represents distinguishable particles such as bosons, and antisymmetric if the system represents indistinguishable particles such as fermions. For this paper, we deal with systems of electrons, which are fermions and thus indistinguishable particles. Electrons additionally have a property called spin. In essence, this is an intrinsic form of angular momentum. The spin of the electron is quantized, it can either take on two values, $+1/2$ or $-1/2$, which we call “spin up” or “spin down”. [4]

Finally we touch upon Dirac notation. It is convenient method to represent wave function. Say we have a wave function $|\psi(x)\rangle$, then we can represent this in Dirac as $|\psi\rangle$, where this new notation is the ket vector. If there is an associated coefficient with the ket say $c|\psi\rangle$, then we can formulate the bra vector as $c^* \langle\psi|$. Here c^* is the complex conjugate of the original coefficient. The bra vector is the ket vector transposed. This leads to the notion of the inner product. A transposed vector multiplied by the vector will lead to a scalar. We need to additionally preserve normalization in order to preserve our probabilistic interpretation of quantum mechanics. That is if we take the inner product of a ket vector with itself it will be 1, $\langle\psi|\psi\rangle = 1$. If the ket vector and bra vector happen to be orthogonal then $\langle\psi|\phi\rangle = 0$.

3 Electronic Structure Methods

3.1 Electronic Structure Problem

A primary interest in computational chemistry lies in solving the non-relativistic time-independent Schrödinger equation.

$$\hat{H}|\Psi\rangle = E|\Psi\rangle$$

Here \hat{H} is Hamiltonian operator for a given system composed of the nuclei and electrons described by their position vectors \mathbf{R}_A and \mathbf{r}_e . We can expand the Hamiltonian to describe the interactions between the components of the system. The terms that encompass the Hamiltonian are:

1. The kinetic energy of the electrons.
2. The kinetic energy of the nuclei.
3. The Coulomb attraction between electrons and nuclei.
4. The repulsion between electrons.
5. The repulsion between nuclei.

Then we arrive out our expanded Hamiltonian:

$$\hat{H} = -\sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{A=1}^M \frac{1}{2M_A} \nabla_i^2 - \sum_{i=1}^N \sum_{A=1}^M \frac{Z_A}{r_{iA}} + \sum_{i=1}^N \sum_{j>i}^N \frac{1}{r_{ij}} + \sum_{A=1}^M \sum_{B>A}^M \frac{Z_A Z_B}{R_{AB}}$$

[7] In the equation above, we have N electrons and M nuclei. M_A is the ratio of the mass of the nucleus A to the mass of an electron. Z_A is the atomic number of nucleus A. Additionally we define r_{iA} as the distance between the i th electron and A th nucleus, r_{ij} as the distance between the i th and j th electron, and R_{AB} as the distance between the A th nucleus and B th nucleus. All distances are irrespective of orientation, that is in terms of the absolute value.

3.2 Born-Oppenheimer Approximation

To simplify the electronic structure problem we first invoke the Born-Oppenheimer approximation. In short, we make the assumption that the motions of the nuclei and electrons are separable. The justification lies in the fact that if we take the consideration that $m_e \ll M_A$ often in the magnitude of thousands and consequently the velocity of the electrons being much greater than of nuclei. Then it follows that the kinetic energy of electrons are relatively vastly greater than that of the nuclei. Thus, the first step of the approximation neglects the kinetic energy of the nucleus. From here, the other terms of the Hamiltonian will merely use the nuclear coordinates as parameters. More importantly, we can write the Hamiltonian as a product of the electronic and nuclear components. For the purposes of this paper, we are merely interested in the electronic Hamiltonian. Then we arrive at our simplified electronic structure problem:

$$\hat{H}_{elec}(\mathbf{R}_n) = -\sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{i=1}^N \frac{Z_A}{r_{iA}} + \sum_{i=1}^N \frac{1}{r_{ij}} = T_e + V_{en} + V_{ee}$$

Here we have the nuclear coordinates labelled by \mathbf{R}_n . Here we assign the variables T_e , V_{en} , and V_{ee} as the electronic kinetic energy, electronic-nuclear potential energy, and electronic potential energy between two electrons respectively. From here, solving for electronic Hamiltonian gives us the energy eigenvalues. However, solving for large molecules is still intractable using classical computation as the Hilbert space of the Hamiltonian scales exponentially.

3.3 Hartree-Fock

Using the assumption of the previously mentioned Born-Oppenheimer approximation, we can then use the Hartree-Fock procedure. This procedure is a mean-field solution, that is we reduce the interaction of many particles to one particle and its interaction with the average potential of the other particles. Then we can

reduce the repulsion term for a given electron as a function of its own location within this mean field. Reflecting this, the Hamiltonian is:

$$\hat{H}_{elec} = \sum_{i=1}^N h(i) = - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{i=1}^N \frac{Z_A}{r_{iA}} + V_{HF}$$

This approximation gives us the eigenstates for one electron Hamiltonians, $h(i)$, and produce a series of equations of the form $h(i)\chi_i = e_i\chi_i$. This is simply another form of the eigenvalue equation. The Hartree Fock procedure minimizes the energy to solve for the set of vectors while the entire set remains an orthonormal basis.

In other words, linear combinations of atomic orbitals are used to specify molecular orbitals. More than likely the actual system of interest will contain more than one electron. Then it is important to note that neither atomic nor molecular orbitals can describe this system, but rather N molecular orbitals are needed to obtain N electron wave functions.

These N molecular orbitals additionally have to be antisymmetric. Due to electrons being fermions, they are indistinguishable. Then if two electrons are exchanged, the wave function has to change signs. To approach this anti-symmetry requirement, first note that we can label the coordinates of an electron using three spatial coordinates, and a spin coordinate, $x_i = (r_i, \omega_i)$, where we combine all three spatial coordinates into the first variable r. Then let us consider two non interacting electrons which can occupy one of two orbitals, χ_1 and χ_2 . We represent this as either

$$|\psi_1\rangle = |\chi_1(x_1)\chi_2(x_2)\rangle = |\chi_1\chi_2\rangle$$

$$|\psi_2\rangle = |\chi_2(x_1)\chi_1(x_2)\rangle = |\chi_2\chi_1\rangle$$

And then an equal superposition of these two states gives us the following antisymmetric wave function with respect to exchanging the electrons.

$$|\psi_{anti}\rangle = \frac{1}{\sqrt{2}}(\psi_1 - \psi_2)$$

This can be achieved through Slater determinants, where if a row is interchanged, the determinant changes sign. Then if we use the rows as the electronic coordinate x_k and the columns as orbitals u_j , we can fit the N wave functions into a matrix. [9]

$$\Psi(x_1, \dots, x_n) = \frac{1}{\sqrt{N!}} \det \begin{pmatrix} u_1(x_1) & \dots & u_m(x_1) \\ \vdots & \dots & \vdots \\ u_1(x_n) & \dots & u_m(x_n) \end{pmatrix}$$

Here for simplicity we ignore the dependence on nuclear coordinates. Additionally, a scalar $\frac{1}{\sqrt{N!}}$ is applied in order to normalize the wave functions.

3.4 Matrix Elements of the Electronic Hamiltonian

Going back to the electronic Hamiltonian, we can consider one electron interaction terms and two electron interactions. The one electron interaction terms are $T_e + V_{en}$, that is the kinetic energy of the single electron

and the potential attraction energy between the nucleus and the electron. The two electron interaction term is then just V_{ee} , or the potential repulsive energy between two electrons. Using the matrix elements defined in the previous section we can then formulate the one electron integral:

$$h_{ij} = \langle u_i | (T_e + V_{en}) | u_j \rangle = \int dx_1 u_i^*(x_1) \left(\frac{1}{2} \nabla_i^2 - \sum_{A=1}^M \frac{Z_A}{r_{iA}} \right) u_j(x_1)$$

We can similarly formulate the two electron integral as:

$$h_{ijkl} = \langle u_i | \langle u_j | (T_e + V_{ee}) | u_k \rangle | u_l \rangle = \int dx_1 dx_2 \frac{u_i^*(x_1) u_j^*(x_2) u_k(x_2) u_l(x_1)}{r_{12}}$$

We note that $h_{ijkl} = h_{jilk}$ since relabeling electrons does not change the value of the integral. Additionally, when the spin-orbitals are real, $h_{ijkl} = h_{ljki} = h_{lkji} = h_{ikjl}$. It is these integral calculations that are of interest for the purposes of this paper. Specifically we compare the distributions of the calculations with the results from another paper in which values are randomly sampled from 1 of 5 distributions to simulate the statistics of molecules.

4 Simulation of the Hydrogen Molecule

4.1 Basis Sets

We apply electronic structure methods to the hydrogen molecule. To model the electronic wave function of hydrogen, it is convenient to use a basis set to represent the wave function. The hydrogen wave functions have the form $\psi_{nlm}(r, \theta, \phi) = R_{nl}(r) Y_{ml}(\theta, \psi)$. Here $R_{nl}(r)$ represent the radial part of the wave function and is a polynomial function with variable r representing the distance to the origin. Similarly $Y_{ml}(\theta, \psi)$ represents spherical harmonic part of the wave function respectively. To simplify this polynomial wave function, we arrive at Slater Type Orbitals (STO) represented as

$$\psi_{STO}(r_A, \theta, \phi; n, m, l, \zeta) = N r_A^{n-1} e^{-\zeta r_A} Y_{ml}(\theta, \psi)$$

The parameter ζ is the orbital exponent, and large values of the parameter leads to small radial extent while small values lead to large radial extent. The STO basis set is thus a choice to use to represent hydrogen wave functions. The main advantage is at small distances from the origin, Slater Type Orbitals have a cusp that correctly models hydrogen behavior and also decays exponentially. However, hydrogen like atoms lack many-electron interactions and so the STO basis set has its flaws. Then we can introduce the Gaussian Type Orbital (GTO) basis set.

$$\psi_{GTO}(r_A, \theta, \phi; n, m, l, \alpha) = N r_A^{n-1} e^{-\alpha r_A^2} Y_{ml}(\theta, \psi)$$

The difference between STO and GTO is the replacement of $e^{-\zeta r}$ with $e^{-\alpha r^2}$. Then when comparing Slater Type Orbitals with Gaussian Type Orbitals, the squaring of the exponent leads to a smooth curve over the origin rather than the correct cusp which STO displayed. However this can be alleviated by using a linear combination of GTOs to create a sharper curve near the origin. Thus, when choosing a basis set, the number of basis functions is also of importance. Ideally, an infinite amount of basis sets would be used to model the electronic wave functions of molecules. Not surprisingly however, the computational difficulty scales with the number of basis functions.

Basis sets can be classified into different types. First we have the minimal basis set. These are where one function is used per atomic orbital. These basis sets are often computationally cheap, but come at the cost of providing rough estimates. Second, we have the split-valence basis sets. In a molecule, the valence electrons are the main electrons responsible for molecular bonding. Thus, the split-valence basis sets take into account the importance of the valence electrons and often represent the valence electron orbitals with more than one function. Third, we have correlation-consistent basis sets. These are often much larger than the previously mentioned basis set types. There are more basis set types. However, for the purposes of this paper, our experiment only selects basis sets from these three types when modeling the hydrogen molecule.

5 The Wecker Ensemble

In a paper performed by Dave Wecker et al [8], they analyze the computational requirements for performing a quantum full configuration interaction algorithm. Their study involves the full quantum configuration of a molecule. However as almost all molecules are intractable, the computational requirements scale exponentially with the increasing number of orbital interactions. Thus the current landscape involves simulating molecular ensembles and then comparing these results to experimental values. For their molecular ensembles, Wecker et al randomly sampled from various distributions. They assert that this method creates artificial molecules that represent the statistics of interaction terms found in real molecules. In their appendix they detail 5 different distribution functions.

1. $p(h_{pp}) = u_{[-10,0]}(h_{pp})$
2. $p(h_{pq}) = u_{[-1,1]}(h_{pq})$
3. $p(h_{pqqp}) = u_{[-0.5,0.5]}(h_{pqqp})$
4. $p(h_{pqqr}) = \frac{1}{2 \cdot 0.2} e^{-|h_{pqqr}|/0.2}$
5. $p(h_{pqrs}) = \frac{1}{2 \cdot 0.1} e^{-|h_{pqrs}|/0.1}$

The first two equations listed above are for the one-electron integrals. As can be seen for one electron integrals, if one were to take the diagonal matrix elements, Wecker et al believe that a uniform distribution ranging from -10 to 0 would be a good representation. Similarly, if one were to take the off diagonal elements of the one electron integral, it would be akin to sampling from a uniform distribution from -1 to 1. For the two electron integrals, it can be seen that the uniform distribution is still sampled from, this time ranging from -0.5 to 0.5, for 4D matrix elements where the inner two indices and outer two indices match. However for the rest of the 4d matrix elements, sampling is done from an exponential distribution about 0.

6 Experimental Setup

Again, the goal of the experiment was to compare the distribution of the randomly sampled data based on the Wecker paper against the actual calculated integral values for a molecule. We start with the simplest case, the hydrogen molecule. The hydrogen molecule is diatomic, with two protons and two electrons. An additional benefit is that with the chosen hydrogen molecule, we can verify the computer program calculated integrals with validated experimental data. The experiment was performed in Python 3.6 using the library

matplotlib [1] to generate histograms for comparing the data. Additionally, the numpy [5] library was used for matrix and vector operations, and the scipy library [3] for distribution comparison. Additionally, there are numerous publicly available electronic structure libraries in Python and we chose to use PySCF. [6]

6.1 Basis Sets Selection

For the experiment we use 7 basis sets to approximate the one electron and two electron integrals for the hydrogen molecule. First, the minimal basis sets STO-3G and STO-6G are used. Next, we chose two split-valence basis sets 3-21G and 6-31G. Finally, three correlation consistent basis sets, cc-pVDZ - Double-zeta, cc-pVTZ - Triple-zeta, and cc-pVQZ - Quadruple-zeta. The table below summarizes this.

Basis Set Summary	
Basis Set	Number of Orbitals
STO-3G	2
STO-6G	2
3-21G	4
6-31G	4
cc-pVDZ	10
cc-pVTZ	28
cc-pVQZ	60

Here, H2 is meant to convey the values associated with the one electron integral. Similarly, H4 is meant to convey the values associated with the two electron integral. The shapes denote the dimensions of the 2D and 4D matrices. Then in order to compare the distributions, random distributions based on the Wecker paper are created for each basis set matching the dimensions of the shapes of the integrals.

We can immediately see that based on the basis sets chosen, there will be a large gap in the number of terms. With the minimal basis sets, we will have merely 4 and 16 terms for the one electron and two electron integrals respectively. Compare that to the basis set of cc-pVQZ and we see that the one electron integral already has 3600 terms, and the two electron integral 12960000 terms. Then we can suspect that the results of most interest will be from the larger basis sets, however it comes at the cost of being much more computationally expensive as can be seen by the exponential growth of size as a dimension increases.

6.2 Bin Size Selection

After calculation of the results by following the Hartree Fock procedure as well as generating an equally sized matrix of randomly sampled numbers, it is necessary to be able to sort through the data. The values in the matrix more or less can be continuous, then to discretize the values, we perform binning. The matplotlib library was then used to display the binned information in the form of a histogram. However, this lead to the discussion of what was the optimal method for creating bins.

Ideally we want to automate the binning process for all the matrices rather than look at each distribution one by one and only then determine by arbitrarily changing the bins. Naturally this lead to two ideas. First, we could determine with a formula what the bin width should be and then let the number of bins be determined by the bin width and the range of the data. Another idea was we let the Python program choose the bin width, but instead we determine the bin numbers and feed this number to the program.

There is various literature in statistical analysis dealing with binning. However, we came across two results. One was proposed in 1979 by [2].

$$W = 3.49\sigma N^{-1/3}$$

Here, we define the width of the bins as W . N represents number of values in the data set. Then the methodology is straightforward. The 3.49 was the result of the study testing out different constants. Another similar formula was created by Freedman and Diaconis.

$$W = 2(IQR)N^{-1/3}$$

Here IQR is simply the interquartile range. This formula is noted to be more robust than the previous formula. These formulas worked well in general however with integrals containing a large amount of 0 values, these didn't provide good enough graphs.

Then for some of the graphs in which this binning method wasn't sufficient, I opted to use Sturge's formula. This is for determining the number of bins rather than the bin width. The formula for Sturge's is:

$$k = \lceil \log_2 N \rceil + 1$$

Here k is the resulting estimate for the bin number. This method however didn't provide as smooth histograms as the bin width methods but sometimes was deemed necessary.

6.3 Distribution Comparison

From here the methodology becomes less formulaic. The first step is simply to individually analyze the distribution shape visually of the different integral calculations. Further improvements could be made by overlaying the histogram plots. One metric that is easy to gather would be the center of the distribution, the mean could be calculated based on the data. In general for all distributions the minimum, maximum, and mean values will be calculated.

Since the goal is to determine whether the calculated integrals match the distribution of the associated randomly sampled distribution, it follows we can run statistical tests on the two distributions. Here, the two sample Kolmogorov–Smirnov test may prove useful. This test is used to determine whether two different distributions differ. The distributions are assumed to be sampled from a continuous function. Thus, our data fits this category. The statistic is given as:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

Here we have F representing the different distribution functions. The subscripts 1 and 2 represent the first and second distribution with n and m elements respectively. Then SUP_x is referring to the supremum. The supremum of a subset S of a partially ordered set T is the least element in T that is greater than or equal to all elements of S , if such an element exists. The null hypothesis is then rejected based on the alpha level and the following formula:

$$D_{n,m} > c(\alpha) \sqrt{\frac{n+m}{nm}}$$

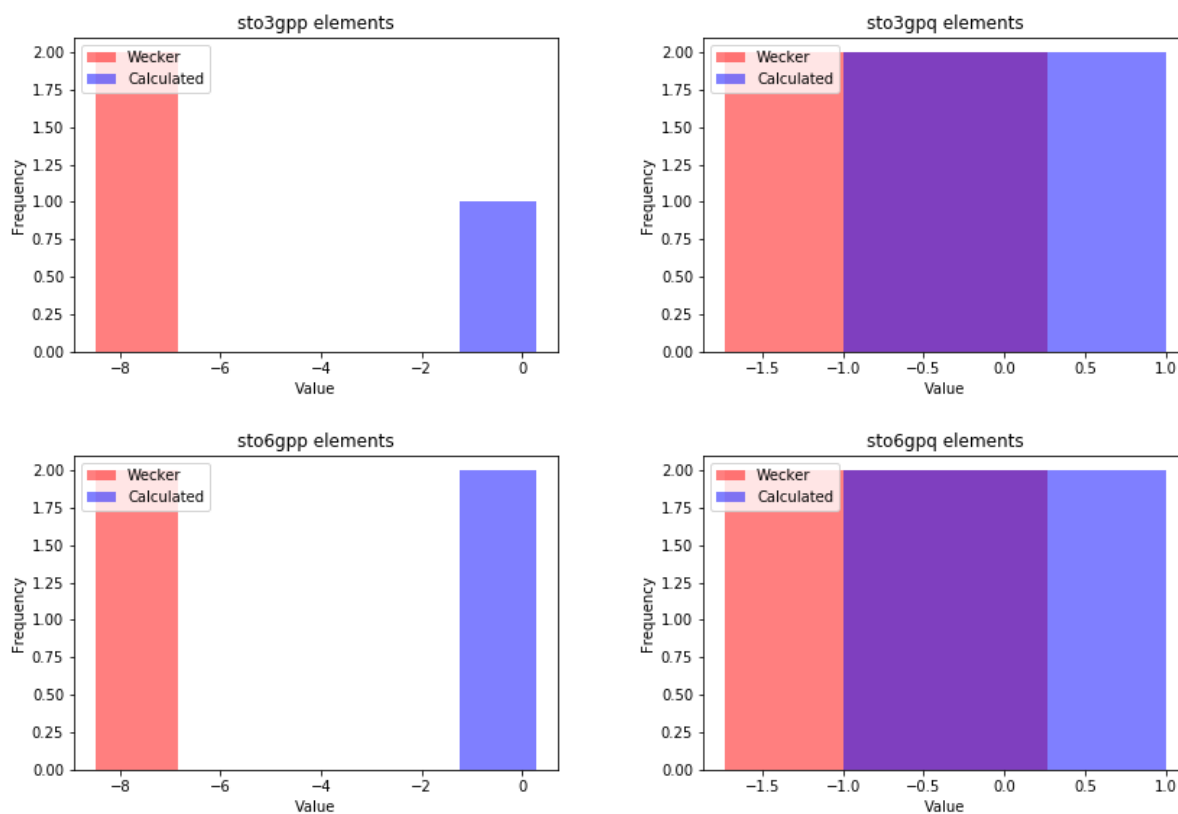
Where $c(\alpha) = \sqrt{-\frac{1}{2} \log_2(\alpha)}$. Then if the distributions don't match, that is, they have a low probability of having been sampled from the same distribution function, the regular Kolmogorov–Smirnov test can be applied. Here, the question is asked whether a data set differs from a certain known distribution function. This will be used to test if we think a given distribution belongs to a known one.

7 Results

In this section, we display the results of comparing the Wecker distributions to the calculated integral distributions. Additionally, we denote the analysis that was unsuccessfully implemented. The majority of the graphs were generated automatically. However when generating the histograms for the correlation consistent basis sets, there were a large amount of zero terms and these had to be taken out of the binning process in order to be able to see the behavior against the Wecker distributions properly. The x axis of the histograms display the values that the integrals or sampling took on and the y axis display the frequency or size of the bin. Additionally, if the data set was large enough to generate a meaningful plot, a table of statistical measures for the integrals is provided.

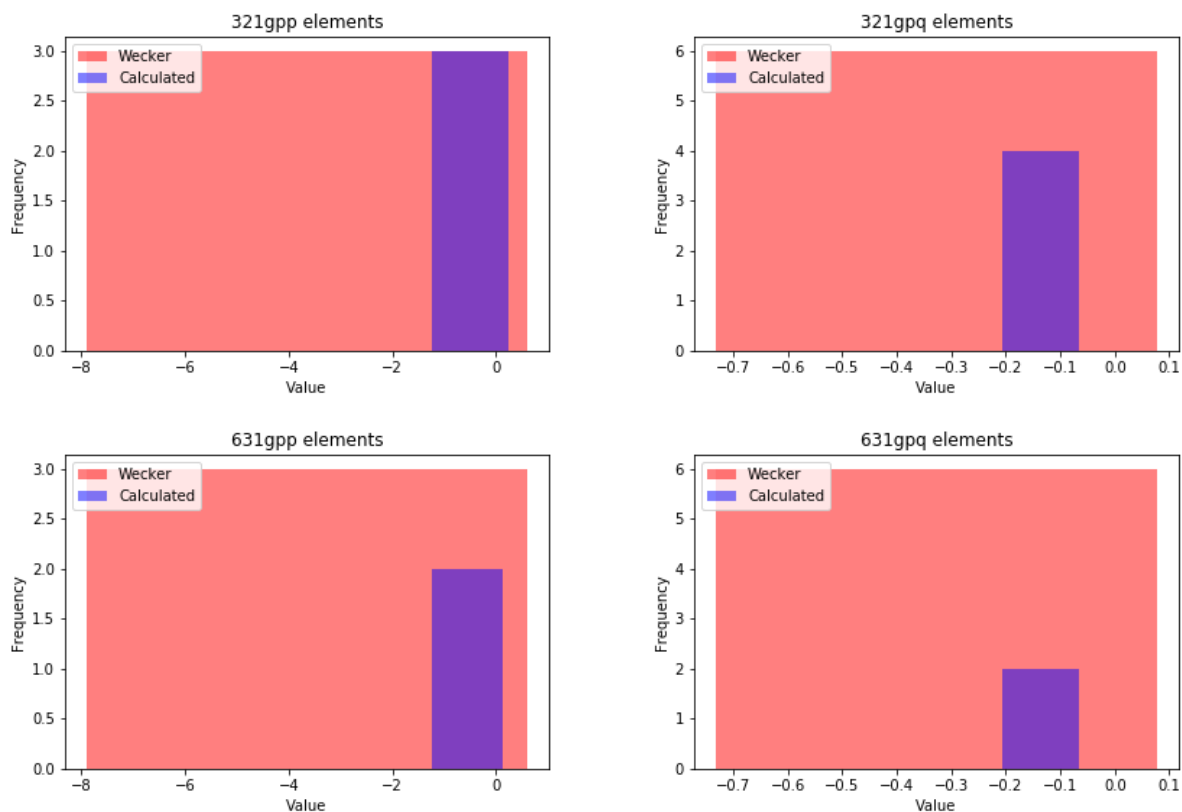
7.1 One Electron Integral Results

7.1.1 Minimal Basis Sets



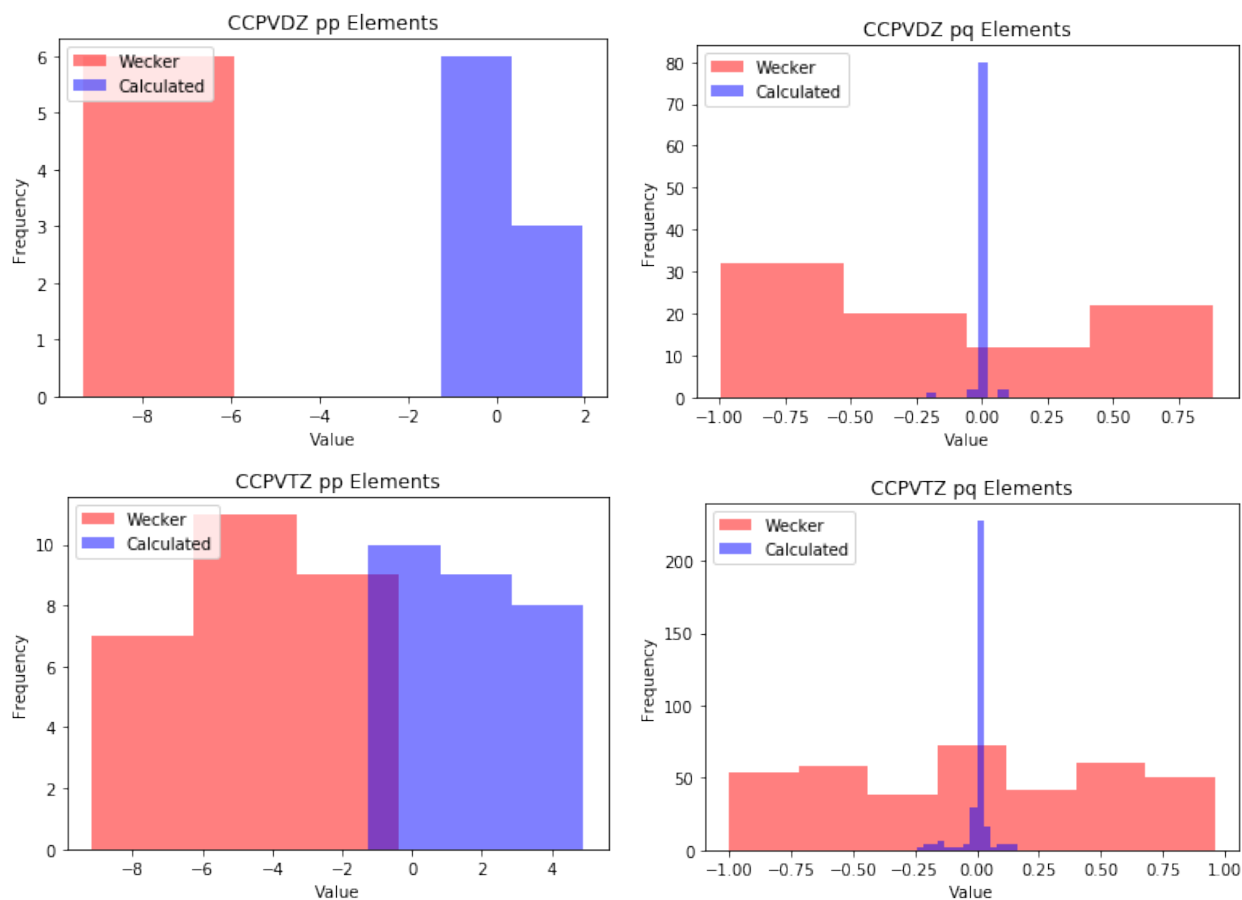
In general, the plots from the minimal basis sets STO-3G and STO-6G didn't provide much room for analysis. As can be seen from the heights of the histograms, there were at most two data points. Given that the matrix representation of the distributions were 2×2 , it follows that we would not obtain meaningful results from only 4 data points.

7.1.2 Split-Valence Basis Sets



The split-valence basis sets also suffered from the same problems as the minimal basis sets. For one electron integrals there were only 16 elements in the 4×4 matrix. We do notice however that the randomly sampled Wecker distribution in all cases had a larger bin width. This implies the data was much less spread out for the integral distributions. This very small range proved to be problematic when trying to determine the bin width using the standard deviation as it often rounded down to 0. The integrals often were in the order of 10^{-16} magnitude in size. This lead to the question of where to cutoff the number and simply make it 0. This problem would also later have to addressed in the correlation consistent basis sets.

7.1.3 Correlation-consistent Basis Sets

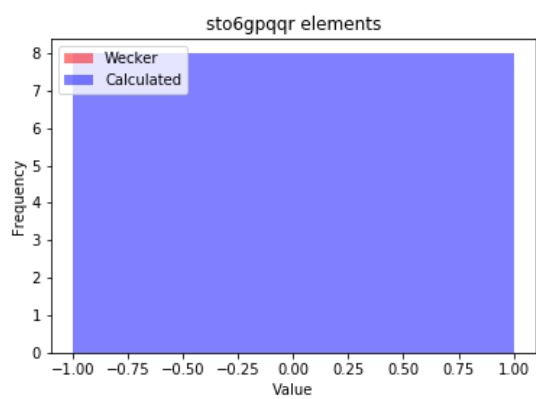
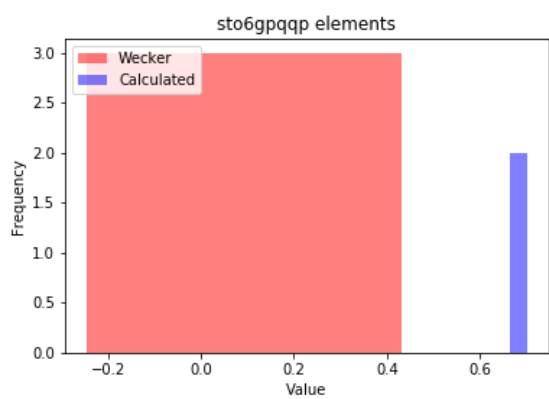
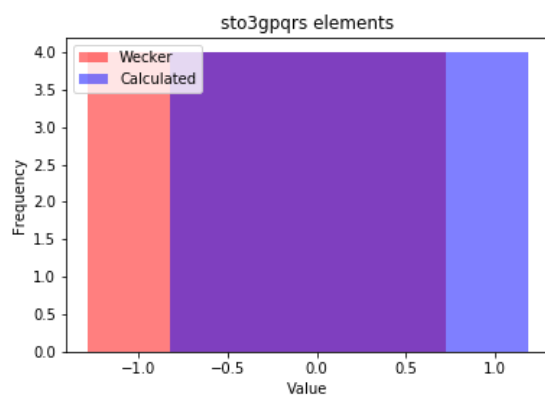
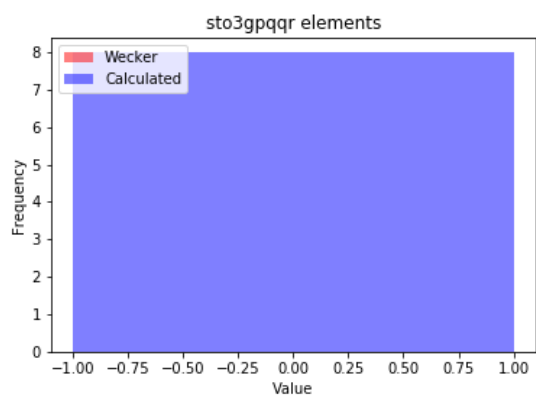
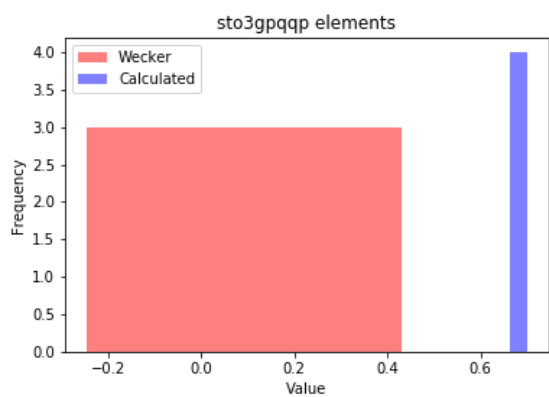


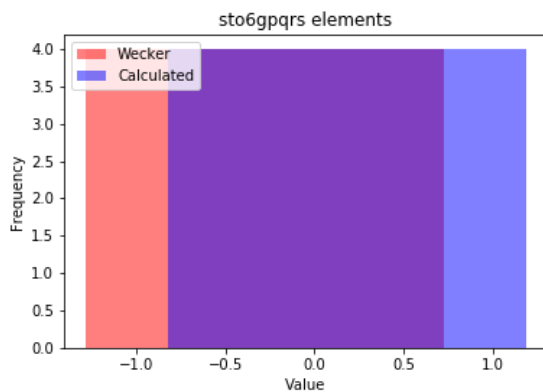
Distribution	Minimum	Maximum	Mean	Standard Deviation
cc-pVDZ h_{pp}	-1.2508	2.3319	0.2944	0.9890
cc-pVDZ h_{pq}	-0.2104	0.1757	0.0040	0.0499
cc-pVTZ h_{pp}	-1.2528	5.5175	1.8122	1.7897
cc-pVTZ h_{pq}	-0.2410	0.1713	-0.0015	0.0337

Due to the growing size of the number of orbitals, we can start to assess whether the random sampling do indeed mirror the statistics of real molecules. For the diagonal elements of the cc-pVDZ and cc-pVTZ, it is interesting to note that qualitatively, the distributions seem to be quite similar albeit with their centers off. However, the off diagonal elements are completely different. The actual integrals have the cusp behavior about 0. The majority of the elements are centered at 0 and exponentially then decay on both sides.

7.2 Two Electron Integral Results

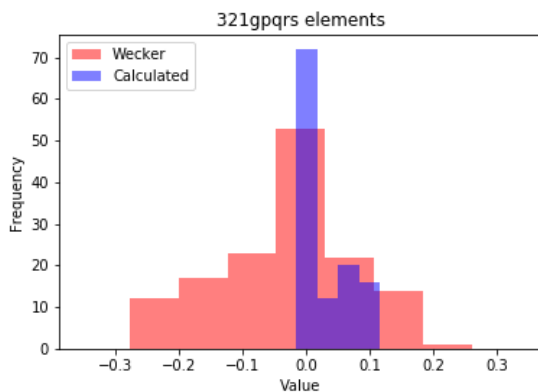
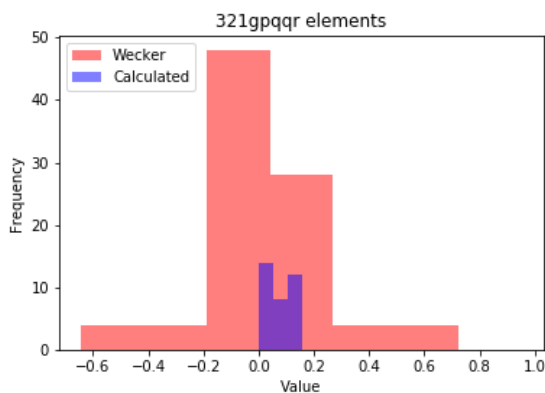
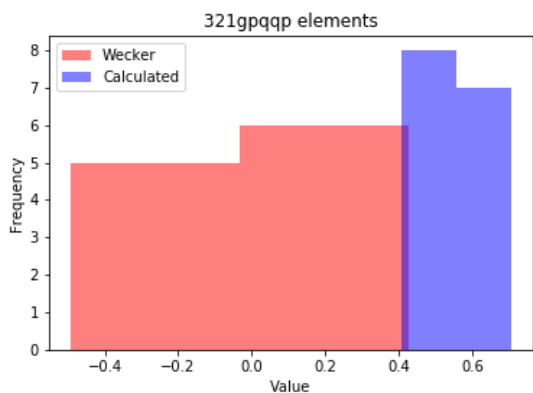
7.2.1 Minimal Basis Sets

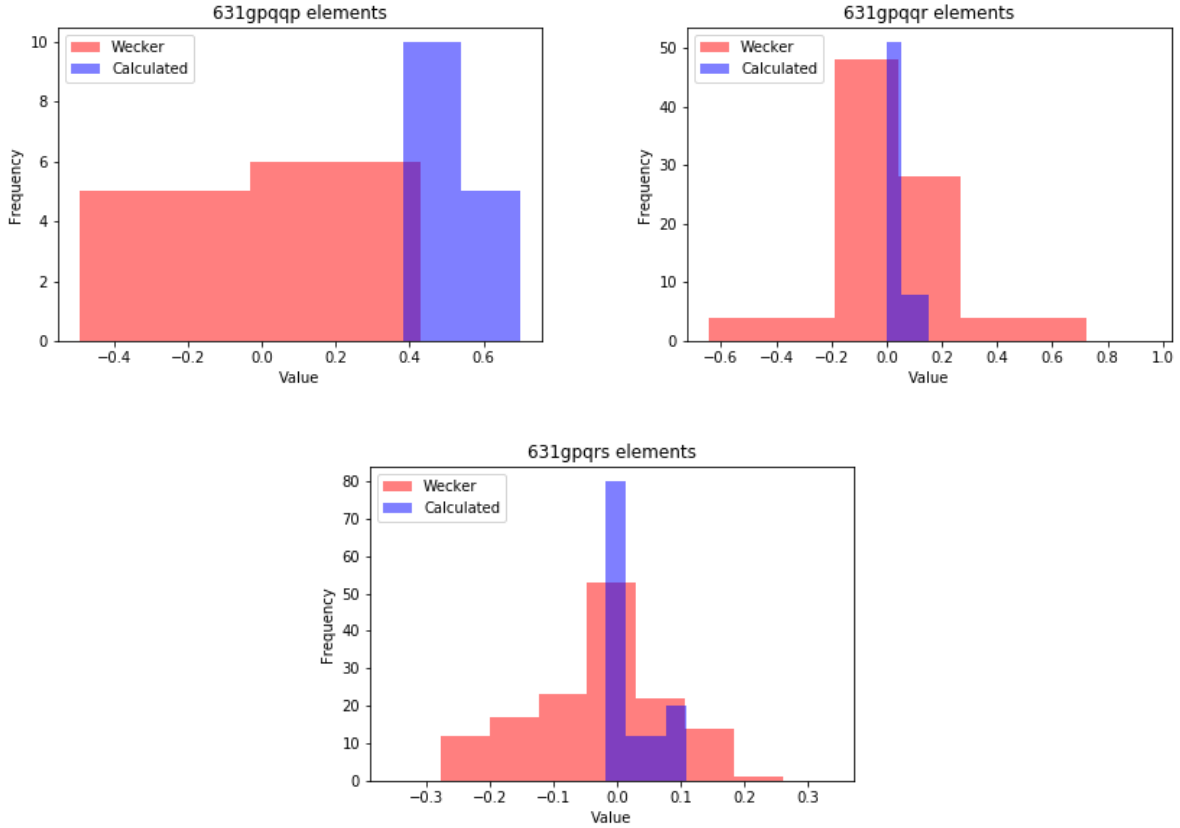




The histograms for all the minimal basis sets for two electron integrals are all grouped together as they are almost identical. The elements denoted by h_{pqqp} are all off the uniform sampling and have a bin number of only 1. The h_{pqqr} elements, the integrals have a distribution ranging from -1 to 1. Finally, the h_{pqrs} elements have a large overlap. At the same time, due to the small size of the data, these results won't provide much insight just as was the case for the one electron integrals.

7.2.2 Split-Valence Basis Sets

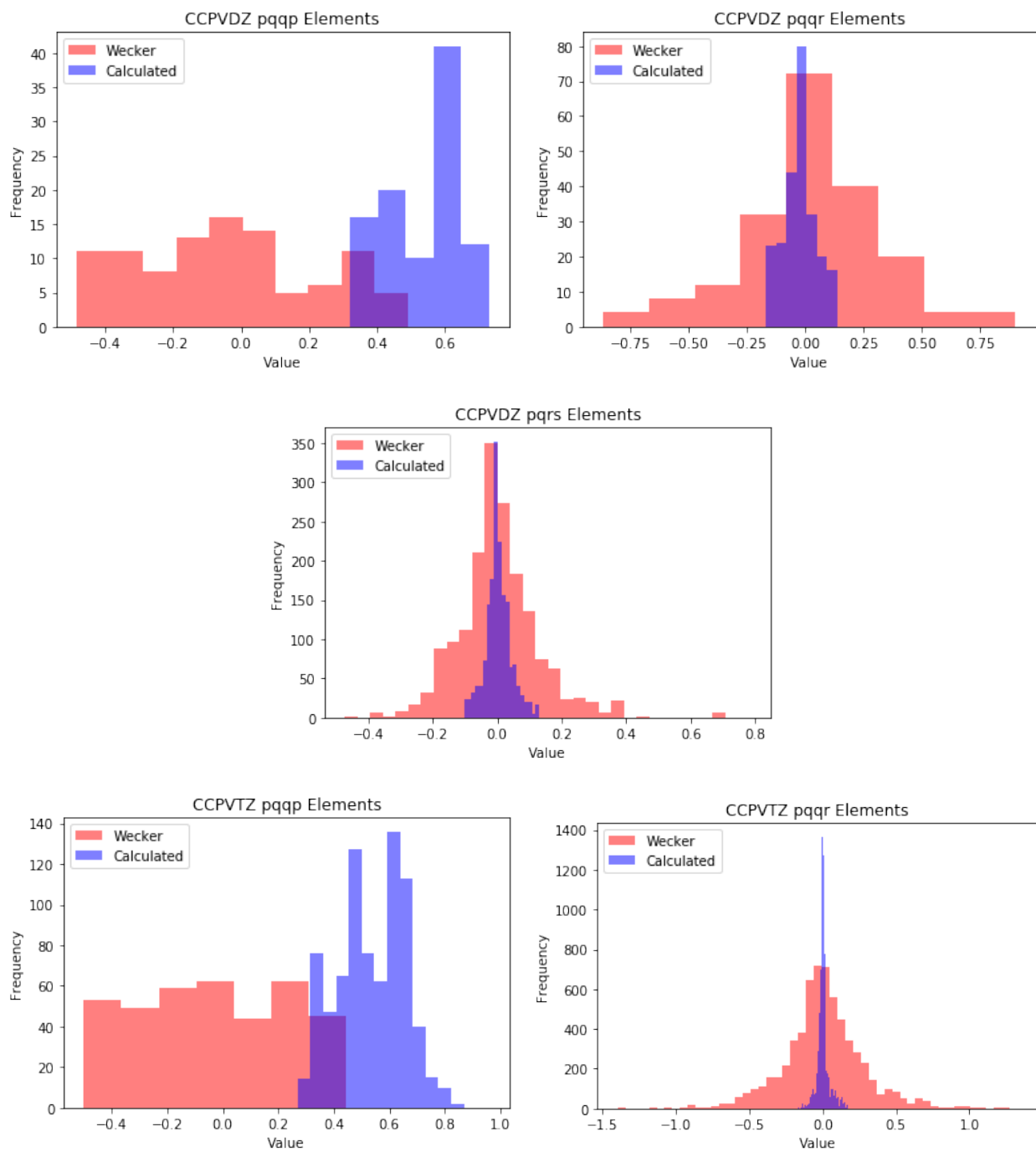


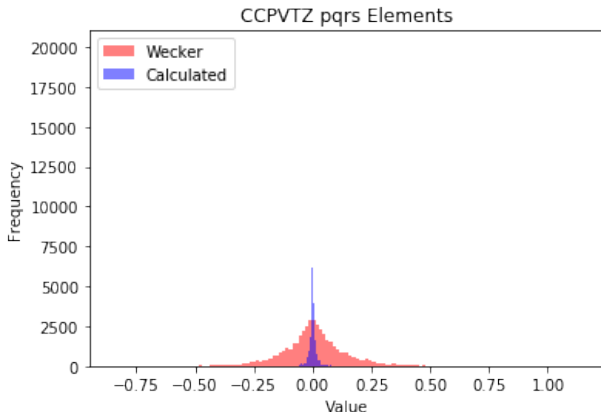


Distribution	Minimum	Maximum	Mean	Standard Deviation
321-G h_{pqqp}	0.4091	0.7480	0.5381	0.1065
321-G h_{pqqr}	0.0000	0.2033	0.0433	0.0679
321-G h_{pqrs}	-0.0167	0.1414	0.0380	0.0497
631-G h_{pqqp}	0.3813	0.7404	0.5155	0.1140
631-G h_{pqqr}	0.000	0.2015	0.0417	0.0664
631-G h_{pqrs}	-0.0193	0.1376	0.0357	0.0480

This time, the split-valence basis sets provided a clearer view of the distributions. We again note the centering of the uniform distributions are off, this time occurring for the h_{pqqp} elements. In the h_{pqqr} elements Wecker overestimates the width of the distribution. Additionally, it isn't clear if the h_{pqqr} elements are actually following an exponential pattern that Wecker proposed they would. It is important to note that originally, the histograms for the h_{pqqr} and h_{pqrs} elements were much more sharp, almost representing a straight vertical line. This was due to the large amount of values near zero around the order of magnitude of 10^{-16} . The values that were of this magnitude were removed from the binning process allowing easier visual comparison to the Wecker distribution.

7.2.3 Correlation-consistent Basis Sets





Distribution	Minimum	Maximum	Mean	Standard Deviation
cc-pVDZ h_{pqqp}	0.3204	0.7657	0.5383	0.1085
cc-pVDZ h_{pqqr}	-0.1701	0.1623	-0.0020	0.0303
cc-pVDZ h_{pqrs}	-0.1029	0.1382	0.0007	0.0183
cc-pVTZ h_{pqqp}	0.2700	0.9796	0.5358	0.1215
cc-pVTZ h_{pqqr}	-0.1894	0.2203	0.0008	0.0186
cc-pVTZ h_{pqrs}	-0.1865	0.2438	0.0001	0.0071

The correlation-consistent basis sets display generally the same shape that was starting to form from the split-valence basis sets. For the h_{pqqp} elements, we can clearly see the centers of the distributions are consistently to the right of the Wecker distribution. Yet the uniform shape that was beginning to appear in the split-valence basis sets begin to disappear and end up with a histogram with numerous peaks and troughs. The h_{pqqr} and h_{pqrs} elements show somewhat similar distributions. Thus, this may indicate that the distribution functions the integrals for h_{pqrs} and h_{pqqr} may be of the same distribution function but with different parameterizations. Especially when compared to the previous histograms and matrix elements, the distributions of the integrals seem to follow much more closely with the Wecker distribution. Again, the same problem arose where a large amount of values near zero were removed. A value was determined to be zero if it was on the order of magnitude 10^{-16} or smaller.

7.3 Inconclusive Results

Unfortunately, the Kolmogorov–Smirnov test always ended up with a two tailed probability value near zero, implying the Wecker ensemble and the integral calculations aren't of the same distribution function. This value occurred regardless of whichever distributions were compared. The next step of testing out known distributions against the integrals however also proved unsuccessful. Tests for uniform distributions and exponential distributions gave probabilities of near zero. Overall, the KS test didn't end up a useful method for determining parameters for fitting the integral distributions. It is unclear where the error in performing the tests arose from.

Additionally, the final basis set cc-pVQZ ended up too computationally expensive. Use of a more powerful computer, or simply more time is necessary.

8 Conclusion

The most conclusive results naturally came from using the largest basis sets. Thus, further investigation can be easily extended and implemented. In this experiment, only a relatively simple molecule was explored. We can achieve more data points with more complicated molecules such as H_2O , CO_2 , and FeS_2 all of which the Wecker paper simulated in their own experiments. Due to time constraints, the ability to simulate larger systems and/or use larger basis sets was eliminated. However, given the time these would provide more meaningful results on how to best parameterize the integral calculations. There is hope for parameterization, as the h_{pqqr} and h_{pqrs} hinted at. Both the distributions seemed to have the same sharp shape. Simply observing the shape of the histograms however should be seen as a starting point and not a robust methodology. As the failure to perform the KS test displayed, a further understanding of statistical tools is necessary in order to derive any meaningful parameterization. Thus, there may some value in further exploring Wecker’s random sampling method for imitating artificial molecules.

References

- [1] J. D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing In Science & Engineering 9.3* (2007), pp. 90–95.
- [2] Alan Julian Izenman. “Recent Developments in Nonparametric Density Estimation”. In: *Journal of the American Statistical Association* 86.413 (1991), pp. 205–224. ISSN: 01621459. URL: <http://www.jstor.org/stable/2289732>.
- [3] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001–. URL: <http://www.scipy.org/>.
- [4] D. McIntyre, C.A. Manogue, and J. Tate. *Quantum Mechanics*. Pearson Education, 2012. ISBN: 9780321850003. URL: <https://books.google.com/books?id=CfMsAAAAQBAJ>.
- [5] Travis Oliphant. *NumPy: A guide to NumPy*. USA: Trelgol Publishing. [Online; accessed {today}]. 2006–. URL: <http://www.numpy.org/>.
- [6] Qiming Sun et al. *PySCF: the Python-based simulations of chemistry framework*. 2017. DOI: 10.1002/wcms.1340. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1340>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1340>.
- [7] A. Szabo and N.S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Books on Chemistry. Dover Publications, 1996. ISBN: 9780486691862. URL: <https://books.google.com/books?id=6mV9gYzEkgIC>.
- [8] Dave Wecker et al. “Gate-count estimates for performing quantum chemistry on small quantum computers”. In: *Phys. Rev. A* 90 (2 Aug. 2014), p. 022305. DOI: 10.1103/PhysRevA.90.022305. URL: <https://link.aps.org/doi/10.1103/PhysRevA.90.022305>.
- [9] James Daniel Whitfield. *At the intersection of quantum computing and quantum chemistry*. 2011.