JOURNAL OF
EMERGING INVESTIGATORS

Article

Modeling Hartree-Fock approximations of the Schrödinger Equation for multielectron atoms from Helium to Xenon using STO-nG basis sets

Krish Gangal[1]*, İpek Gerz[2]*, Tanvi Goyal[1]*, Ajeeth Iyer[3]*, Vaibhav Vaiyakarnam[4]*, Larry McMahan[5]

[1]Irvington High School, Fremont, California
[2]Modern Eğitim Fen Schools, Beşiktaş, Ulus, Türkiye
[3]Rio Americano High School, Sacramento, California
[4]The Quarry Lane School, Dublin, California Aspiring Scholars Directed Research Program
[5]Aspiring Scholars Directed Research Program, Fremont California
*equal contributors

SUMMARY

The energy of an atom is extremely useful in nuclear physics and reaction mechanism pathway determination but is challenging to compute. Thus, we aimed to synthesize regression models for Pople Gaussian expansions of Slater-type Orbitals (STO-nG) atomic energy vs. atomic number scatter plots to allow for easy approximation of atomic energies without using computational chemistry methods. Using the Hartree-Fock method, we calculated atomic energies for elements helium to xenon (He to Xe) using STO-nG electron orbital basis set models. After calculating atomic energies for each basis set, we plotted them as tables of atomic energy (in Hartrees) versus the atomic number. We hypothesized that there would be a non-linear correlation in the scatter plot. We first formed the regressions using data from helium to krypton (He to Kr), and then we added new data from Kr to Xe. We then calculated the old and new coefficients of determination, and the difference between them. Due to their common use in modeling, exponential, sinusoidal, and quadratic regressions were tested to model the data. The data supported the hypothesis that the scatter plots had non-linear correlations. Sinusoidal and quadratic regressions had higher initial coefficients, suggesting higher viability, while sinusoidal regressions also had the highest average final coefficients, and the lowest change in coefficients. The data indicated that of the

INTRODUCTION

The term atomic energy refers to the energy carried by an atom, including its nucleus and electrons. The energy of an atom is vital in nuclear energy studies, emission series predictions, and determining reaction pathways for molecules (1). In the case of reaction pathways, the energies of the electrons can be used to determine the vibrational frequencies of the bonds and determine if desired compounds are stable or not. If they are not stable, alternatives to the compounds, known as analogs, can be modeled such that they have similar chemical properties to the original compound, but are also stable. Such calculations are crucial for pharmaceutical production and material synthesis.

However, calculations require considerable time, as unlike calculating potential energies for particles on a macroscopic scale, electrons and subatomic particles tend to act as both waves and particles at the quantum level. This phenomenon is known as wave-particle duality (2). Rather than treating the electrons as exact particles or pure waves, they are considered “clouds” that display properties of both. Due to this nature, it was determined that measurements for such bodies possessed inherent uncertainties. For example, the more precise a calculation of the position of an electron is, the less precise its momentum calculations are (3). Thus, these clouds are modeled as probability distributions, represented by what is known as a wavefunction. The wavefunctions of electrons within atoms are calculated using the Schrödinger equation, which consequently allows for calculations of other properties such as atomic energy (4).

The Schrödinger equation is highly accurate when calculating single-electron atoms and ions’ electron wavefunctions and atomic energies (5). However, when progressing to multielectron atoms, due to the electrostatic repulsion between the negatively charged electrons, the wavefunctions of the electrons are dependent on each other, preventing the calculation of accurate atomic energies using the Schrödinger equation (6).

Numerous computational techniques exist to approximate the equation. The two most common are Quantum Monte Carlo techniques, which use repeated random sampling to calculate atomic energies, and self-consistent field (SCF) techniques, which start with an initial guess for the atomic energy and then apply a recursive algorithm to get a converged result up to a certain number of decimal places (7). Of these two, SCF methods provide a simplified approach to calculating atomic energies in their treatment of wavefunctions (8).

We utilized Hartree-Fock, the first SCF method devised. This was due to its lower runtime compared to other SCF methods, such as post-Hartree-Fock (PHF). The Hartree-

Table 1. Regression equations.

	Exponential Regression $E_{atom} = a\exp(bZ)+c$			Quadratic Regression $E_{atom} = a(Z+b)^2+c$				Sinusoidal Regression $E_{atom} = a\sin(bZ+c)+d$		
Basis Set	a	b	c	a	b	c	a	b	c	d
STO-2G	-711.494	0.0455887	993.695	-3.10222	-6.44133	-55.4310	$9.72970\cdot10^{10}$	$7.98550\cdot10^{-6}$	1.57074	$9.72970\cdot10^{10}$
STO-3G	-280.179	0.06759	394.637	-2.68347	-4.41653	-25.8120	$1.13920\cdot10^{11}$	$6.86374\cdot10^{-6}$	1.57077	$1.13920\cdot10^{10}$
STO-4G	-715.727	0.0455534	999.055	-3.11220	-6.42516	-55.7641	$1.44880\cdot10^{13}$	$6.55460\cdot10^{-7}$	1.57079	$1.44880\cdot10^{13}$
STO-5G	-280.043	0.0675985	394.471	-2.68283	-4.41766	-25.7985	$3.62780\cdot10^{10}$	$1.21615\cdot10^{-5}$	1.57074	$3.62780\cdot10^{10}$
STO-6G	-280.179	0.067594	394.637	-2.68347	-4.41653	-25.8120	$1.20160\cdot10^{11}$	$6.68316\cdot10^{-6}$	1.57077	$1.20160\cdot10^{11}$

NOTE: Coefficients of formulated calculated exponential, quadratic, and sinusoidal regressions using atomic energy (in Hartrees (Ha)) vs. atomic number from elements He-Kr, using Desmos.

Fock method has applications in calculating not only multielectronic atomic energies and quantum states, but also molecular orbitals, stability, and vibrational modes. It is used in biochemistry, pharmaceuticals, organic and inorganic chemistry, and material science, among other fields (9).

The foundation of the SCF method is the initial guess for the electron wavefunctions. These initial models are known as basis sets and are divided into Slater-type orbitals (STO) and Gaussian-type orbitals (GTO). STOs function as modeled approximations of the polar form functions of an electron orbital, and the time required for STO-based SCF calculations is greater than GTO-based ones. GTOs are easier to calculate but less accurate than STO basis sets (10).

While individual GTO functions are not accurate enough for SCF calculations, they can be summed together through linear combinations to approximate STOs in what is known as a contracted-GTO (CGTO) basis set (11). The size of a CGTO basis set is the number of GTO orbitals summed together for one electron orbital. The smallest CGTO basis sets, also known as minimal basis sets, are the Pople STO-nG basis sets, where n represents the size of the set (12).

Smaller basis sets, especially the STO-3G to STO-6G basis sets, are commonly used for basic atomic and molecular energy calculations due to their short calculation times. Although the accuracy of the calculations increases as the size (or the number of functions approximating each electron orbital) of the basis set increases, the program runtime rises (13). Moreover, the smaller STO-nG basis sets are used as foundations for other specialized basis sets (14). Thus, the smaller STO-nG basis sets are still commonly used.

The applicability of atomic and molecular energy calculations calls for more efficient methods of calculating SCF energies, especially for non-computational chemistry researchers. Thus, we sought to investigate effective and intuitive methods for atomic energy calculations. First, we decided to use a Python-based computational chemistry library for calculations (PySCF) and observe its potential in calculating atomic and molecular energies from scratch for future investigations.

Second, for atomic energies alone, we aimed to model a relationship between atomic number and STO-nG atomic energy. We sought to determine how to calculate the atomic energies for the common STO-nG basis sets without SCF methods. Moreover, STO-nG basis set calculations have yet to be standardized beyond Xe, as there are numerous different STO-nG basis set variations caused by adding extra functions to model the increased number of orbitals beyond Xe. Thus, our models also aimed to identify the feasibility of predicting the STO-nG energies of elements beyond Xe. It must be noted that these models should be interpreted at whole-number values of atomic number only, and do not have any significance at fractional values.

Our experiment aimed to calculate atomic energies for elements from He to Xe using STO-nG basis sets (from n=2 to n=6). We hypothesized that the increase in the atomic energies would be non-linear, as the energy of interelectronic Coulombic repulsion, especially between valence and inner shielding electrons, would increase faster than the energy of the individual electrons and protons themselves, leading to a non-linear atomic energy increase. We found that the data supported our hypothesis, and that sinusoidal regressions seemed to fit the data best out of the tested regressions.

RESULTS

We aimed to calculate the STO-nG energies using PySCF, then use them to plot atomic energy versus atomic number and see if any regressions could be feasibly modeled. Exponential, quadratic, and sinusoidal regressions were selected for testing, as the three models were fundamental and commonly used functions in modeling.

Regression equations were first formulated using atomic energy data from He to Kr (**Table 1**). The feasibility of the regression models was evaluated by calculating the coefficient of determination (R^2) for the data sets, with a maximum degree of precision of four significant figures as offered by the software. The R^2 values of the quadratic and sinusoidal regressions remained at a constant value of 0.9996 for all five basis sets, while the coefficient varied for the exponential regression, with a mean of 0.9983 and a standard deviation of 0.0002 (**Table 2**). STO-2G was found to have the lowest exponential regression coefficient value of 0.9981, while STO-5G and STO-6G had the highest value of 0.9985.

Table 2. Regression original coefficients of determination.

	Coefficients of Determination before New Data (R^2)						
Regression	**STO-2G**	**STO-3G**	**STO-4G**	**STO-5G**	**STO-6G**	**Average**	**Standard Deviation**
Exponential	0.9981	0.9982	0.9982	0.9985	0.9985	0.9983	0.0002
Sinusoidal	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.0000
Quadratic	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.0000

NOTE: Calculation of coefficients of determination for the five basis sets, with regressions calculated with atomic energy (Ha) vs. atomic number data from elements He-Kr, using Desmos.

After the approximative capacity of the regressions was tested through regression formulation using the first 35 elements from He, the predictive power of the regression models was evaluated. Keeping the formulated regressions fixed, the data set was changed to include all the atomic energy values up to element fifty-four, Xe. With eighteen new data points, the R^2 values of the regressions were recalculated. The new R^2 values decreased from the original values, except for STO-4G exponential and STO-4G sinusoidal models, where they remained the same. The exponential regressions had the lowest mean new R^2 value of 0.8997, with a standard deviation of 0.0899. The sinusoidal regressions had the highest mean value of 0.9959, with a standard deviation of 0.0039. The quadratic regressions had a mean new value of 0.9826, with a standard deviation of 0.0270 (**Table 3**).

To understand if there was a relationship between basis set size and an increase in the magnitude of the residuals of the regressions modeling the basis sets, the average difference in the coefficients of determination was also calculated for each basis set. The differences between the old and the new R^2 values were calculated to quantitatively observe the change in the degree of regression correlation with the atomic energy vs. atomic number data set for each basis set. The mean difference for exponential regressions was the largest in magnitude, at -0.0986, with a standard deviation of 0.0896. The mean difference for sinusoidal regressions was the lowest in magnitude, at -0.0037, with a standard deviation of 0.0039. The mean difference for quadratic regressions was -0.0170, with a standard deviation of 0.0270 (**Table 4**). The average difference for all three regressions of a basis set had the lowest magnitude for the STO-2G basis set regressions at -0.0013, and the second lowest for the STO-4G basis set. In contrast, the other three basis sets had similar differences, with STO-5G possessing the highest average difference of -0.0595 (**Figure 1**). On further inspection of the regression functions, it was also observed that the method of least squares parameters for the STO-2G and STO-4G basis sets for the exponential and quadratic models were similar, while the parameters for the other three basis sets were similar, with these two sets of regressions having notable differences in their parameters.

DISCUSSION

As the graph suggests, the data of STO-nG atomic energy vs. atomic number does not form a linear correlation, thereby supporting our hypothesis. The coefficients of determination of the three models suggest how feasible they are as models for the data set – the higher the coefficient, the greater the correlation between the data points and the regression model. The mean was calculated for each regression to get an average value of how well they correlated with the atomic energy data sets. We observed that the quadratic and the sinusoidal regressions had a higher coefficient of determination, while the exponential regression had lower coefficients of determination. This suggests that the quadratic and sinusoidal regressions correlate better with the data set than the exponential regressions.

The quadratic and sinusoidal regressions were observed to have equal coefficients of determination. It must be noted that the sinusoidal regressions specifically have function parameters orders of magnitude below the range of the data presented. This may suggest that in the context of the data, the data set range may be small enough for the sinusoid function to model it at lower parameters. Further testing must be done for sinusoidal functions to confirm this.

There are multiple ways in which prediction models can

Table 3. Regression new coefficients of determination.

	Coefficients of Determination after New Data (R^2)						
Regression	**STO-2G**	**STO-3G**	**STO-4G**	**STO-5G**	**STO-6G**	**Average**	**Standard Deviation**
Exponential	0.9971	0.8297	0.9982	0.8367	0.8367	0.8997	0.0895
Sinusoidal	0.9976	0.9969	0.9996	0.9894	0.9958	0.9959	0.0039
Quadratic	0.9986	0.994	0.9344	0.993	0.993	0.9826	0.0270

NOTE: Calculation of coefficients of determination for the 5 basis sets for the calculated regressions, after the addition of new atomic energy (Ha) vs. atomic number data from elements Kr-Xe, using Desmos.

Table 4. Regression coefficient differences.

	Difference in Coefficients of Determination (R^2)						
Regression	**STO-2G**	**STO-3G**	**STO-4G**	**STO-5G**	**STO-6G**	**Average**	**Standard Deviation**
Exponential	-0.0010	-0.1685	0	-0.1618	-0.1618	-0.0986	0.0896
Sinusoidal	-0.0020	-0.0027	0	-0.0102	-0.0038	-0.0037	0.0039
Quadratic	-0.0010	-0.0056	-0.0652	-0.0066	-0.0066	-0.0170	0.0270

NOTE: Calculation of difference between final and initial coefficients of determination for the calculated atomic energy (Ha) vs. atomic number regressions for the 5 basis sets (i.e., coefficients before and after new data was added) using Desmos.

be tested. One of the most commonly used is establishing a training set, and then testing using a test set. In this case, our training set was selected as He to Kr, and our test set as He to Xe. To compare how well the correlation of the prediction model fit with the testing set, the difference in the coefficient of determination was calculated between the testing and training sets. These differences represent the addition of new data that affected the correlation of the data set.

If additional data is added and the coefficient decreases, then it implies that the new elements in the training set have introduced greater unexplained error and are not modeled as well, suggesting that further predictions beyond the test set would have even more residual error. For all three models, we found that the addition of new data generally led to a decrease in the coefficient of determination, implying that further prediction would not be possible due to increasing error. The average magnitude of difference in values for the exponential regressions after and before the new data for all basis sets was the highest among the three regressions. The data suggests that sinusoidal regressions had the least average magnitude of difference in values, followed by quadratic regressions.

Of the three regressions, sinusoidal regressions had the least change in correlation after new data was added and remained the most consistent in data correlation after the new data is included. This implies that of the three regressions, sinusoidal regressions were the most viable for predicting the atomic energies of elements beyond Xe, as they generated the least residual error when new data was tested. For all the basis sets, the quadratic and sinusoidal regressions saw a gradual increase in the magnitude of negative residuals concerning to the atomic energy vs. atomic number data set, in contrast, the exponential regressions saw a sharp increase in the magnitude of positive residuals (graphs with residuals in **Appendix 3**).

However, the testing of the model raises several questions and areas for further research. Firstly, the sinusoidal model already demonstrated a great difference in the order of magnitude of the regression parameters and the range of the atomic energy dataset. Thus, further modeling for the sinusoidal regression is required to see if it truly provides meaningful results. Moreover, there were two functions where the coefficient of determination did not change with the addition of the test set. These are the STO-4G exponential and sinusoidal sets. For the sinusoidal sets, the lack of a change in the coefficient of determination may be attributed to its small parameters allowing it to be a good fit, nonetheless. However, two questions are raised: why does the coefficient remain constant for the exponential regression when the rest of the basis sets see great differences in the coefficients for the exponential model, and why do both anomalies are occurring with the STO-4G set? We hypothesized that this correlation was either an error present in the method of least squares used to determine the regressions by Desmos or that there could be a relationship between the basis set and the general correlation of the three tested regressions with that set's atomic energy data.

We found that the average coefficient difference for the STO-2G and STO-4G sets was much lower in magnitude than the average coefficient difference for the other three basis sets, with the difference in magnitude being greatest for the STO-5G data set. This was supported by a similar method of least squares parameters for STO-2G and STO-4G sets' regressions that were not found in the other sets' regressions. The observation of this data suggests that the parameters generated by the method of least squares technique on Desmos have reached different convergent minimums for the different basis sets. Further testing is required on other regressions to support the functions formulated using Desmos.

The data supported our hypothesis that the plots of STO-nG Hartree-Fock atomic energy vs. atomic number would

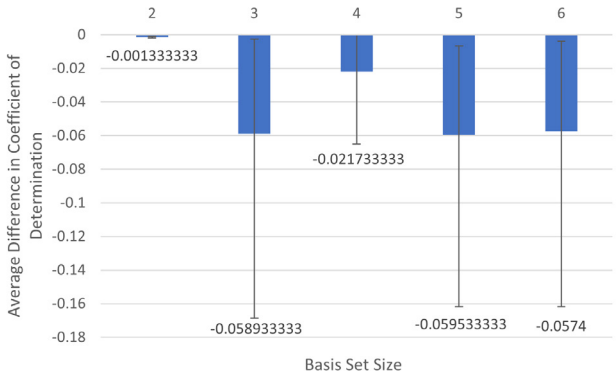

Figure 1. STO-nG basis set average coefficient difference vs. basis set size. A bar graph displaying the variation of the average regression coefficient of determination difference for each basis set, with respect to the size of the basis set. Error bars show mean ± sd.

be non-linear. When testing the regressions, we found that the sinusoidal regression and the quadratic regression had a greater initial correlation with the data than the exponential regression. After introducing the test set, we discovered that exponential regressions had positive residuals, while quadratic and sinusoidal regressions had negative residuals. The exponential functions had greater decreases in the coefficient of determination when new data was introduced compared to the quadratic and sinusoidal functions, suggesting that they were susceptible to more residual error if used to predict atomic energies beyond Xe.

However, these suggestions also came under scrutiny due to further analysis showing that the models for the STO-2G and STO-4G database had different parameters when compared to the other basis sets and were also generally better correlated compared to the other three basis sets. Moreover, the parameters of the sinusoidal set were much lower in order of magnitude than the magnitude of the atomic energies themselves, implying that the sinusoidal models may not be as effective or meaningful in their predictions.

Basic future research developing on this experiment would involve conducting more trials in generating the same regressions using Desmos, and comparing them to regressions generated by other software, to identify whether the anomalies identified were software issues or actual trends. More functions can be tested for regression modeling, such as Gaussian bell-curves, to observe if there are functions with higher coefficients of determination for the data sets. Moreover, future work could also compare PySCF data to previously calculated STO-nG energies, which our research group plans to do in our upcoming research.

MATERIALS AND METHODS

Schrödinger equation and Hartree-Fock method

The Schrödinger equation is a partial differential equation that solves for the energy of an electron in an atom. The Schrödinger equation, where the calculation of the Hamiltonian (which represents the total energy of the atom) is independent of time, can be written as an eigenvalue equation (Equation 1).

$$\hat{H}\psi = E\psi \tag{1}$$

$\hat{H}$ represents the Hamiltonian of the atom, ψ represents the wavefunction of the electron, and E represents the energy of the electron.

The Hartree-Fock method acts as an approximating computational extension to the Schrödinger equation for multielectron atoms. Firstly, the many-electron wavefunction is approximated to be a product of the orbital wavefunction of each electron in the atom (Equation 2) (16).

$$\psi(r_1, r_2, \ldots, r_n) = \prod_{i=1}^{n} \varphi(r_i) \tag{2}$$

The multielectron wavefunction ψ for n electrons is the product of the atomic orbital wavefunctions φ of the electrons.

However, this allows for the existence of two electrons of the same energy level and angular momentum in an atom, which is not possible as per the Pauli exclusion principle (17). Thus, the spin states of the electrons must also be considered, which is done by calculating the Slater determinant of the many-electron wavefunction approximation. This allows for the calculation of the electron energies and the multielectron wavefunction in terms of the spin orbitals of the electrons (18).

The spin orbitals are then calculated through the variational method, where the wavefunction of electron *i* is first calculated independently, then the wavefunction for *j* is calculated using the field of *i* as the average field, and the process is repeated, switching between the wavefunctions of *i* and *j* until the ground state electron energy has been minimized. This process can be conducted for any number of electrons.

After the spin orbitals have been determined, they can be substituted into the Hartree-Fock equation to solve for the energy of one electron spin-orbital (Equation 3).

$$f_i\psi_i = \varepsilon_i\psi_i \tag{3}$$

This is an analog to the Schrödinger equation for a single electron spin-orbital. ψ_i represents the wavefunction of electron *i*, ε_i represents the energy of the electron, and f_i represents the Fock operator for the electron, which is the analog to the Hamiltonian for the atom but includes the energy of interelectronic Coulombic repulsion.

This can then be converted into a matrix equation calculable for the atomic energy, through the Roothaan-Hall equation, an analog of the Schrödinger equation for the Hartree-Fock method (Equation 4) (19).

$$FC = SC\varepsilon \tag{4}$$

F represents the Fock matrix, the sum of H (the core Hamiltonian matrix) and G (the interelectronic Coulombic repulsion matrix). S represents the overlap matrix, calculating the overlap in electron orbitals. C represents the orbital coefficients, which is a linear combination of the calculated spin orbitals of the electrons. ε represents the diagonal energy matrix, which stores the values of the individual ground state energies of the electrons.

Thus, the atomic energy can be calculated using the Roothaan-Hall equation. This calculation can be repeated recursively, using the results of a calculation as the basis for another Hartree-Fock calculation, until the atomic energy converges for a certain number of decimal digits. Thus, Hartree-Fock, and other similar recursive methods are called the Self-Consistent Field (SCF) methods.

PySCF algorithm structure

PySCF is a peer-reviewed Python library with C optimizations facilitating Hartree-Fock and other SCF

calculations using GTO basis sets. The PySCF library functioned as the foundation of our research algorithms (20). While PySCF is preinstalled with basis sets for common elements, it does not provide bases for larger elements. Thus, STO-nG basis set data was obtained from the BSE library, which called on data from the Basis Set Exchange GTO database (21). Using this, atomic energy states were calculated using PySCF for He to Xe using the BSE STO-nG basis sets, and then displayed as LaTeX tables using the Matplotlib library (22).

As opposed to getting the data from online sources, PySCF was used explicitly for future projects that the research group aims to work on. To investigate ways in increasing the accuracy and decreasing program runtime for atomic, and in the future, molecular calculations, PySCF's calculations were tested and compared with previous studies, and its runtimes were also gauged separately to identify if it would be a viable program library for calculations run from scratch. In this light, future testing will also observe PySCF's calculations for smaller and larger organic molecules, wherein data for many compounds cannot be found in online databases yet.

A general program for a set of elements and an STO-nG basis set is structured as follows. After importing the libraries, the STO-nG basis sets for the elements are imported from the BSE. Then, the Hartree-Fock function is defined, taking in inputs of the element symbol, basis set, and number of unpaired electrons, and calculating the Hartree-Fock energy of the atom via PySCF. This energy is saved in a list, along with the element symbols. This list is then printed as a LaTeX table by Matplotlib, with columns of 'Element Symbol' and 'STO-nG Atomic Energy.'

This algorithm was modified for each STO-nG basis set (2G to 6G) and used to calculate and display the energies of the elements. To allow for simultaneous execution of the programs, they were divided into the elemental groups of He-Ne, Na-Ar, K-Ca & Rb-Sr, Sc-Zn, Ga-Kr, Y-Cd, and In-Xe (see **Appendix 1**). Consequently, 42 LaTeX tables were created for each elemental group and basis set (see **Appendix 3**). These were then converted into Microsoft Excel tables using OCR software, and then used to plot regressions of atomic energy vs. atomic number.

Regression formulation and evaluation

After the conversion of the LaTeX files into Microsoft Excel tables, the data was reorganized into one table with the columns of atomic number (Z), and STO-nG atomic energy in Hartrees (E_{atom}/Ha), with one column for each n from 2 to 6 (see **Appendix 2**). Data from He-Kr (elements 2-36) was first

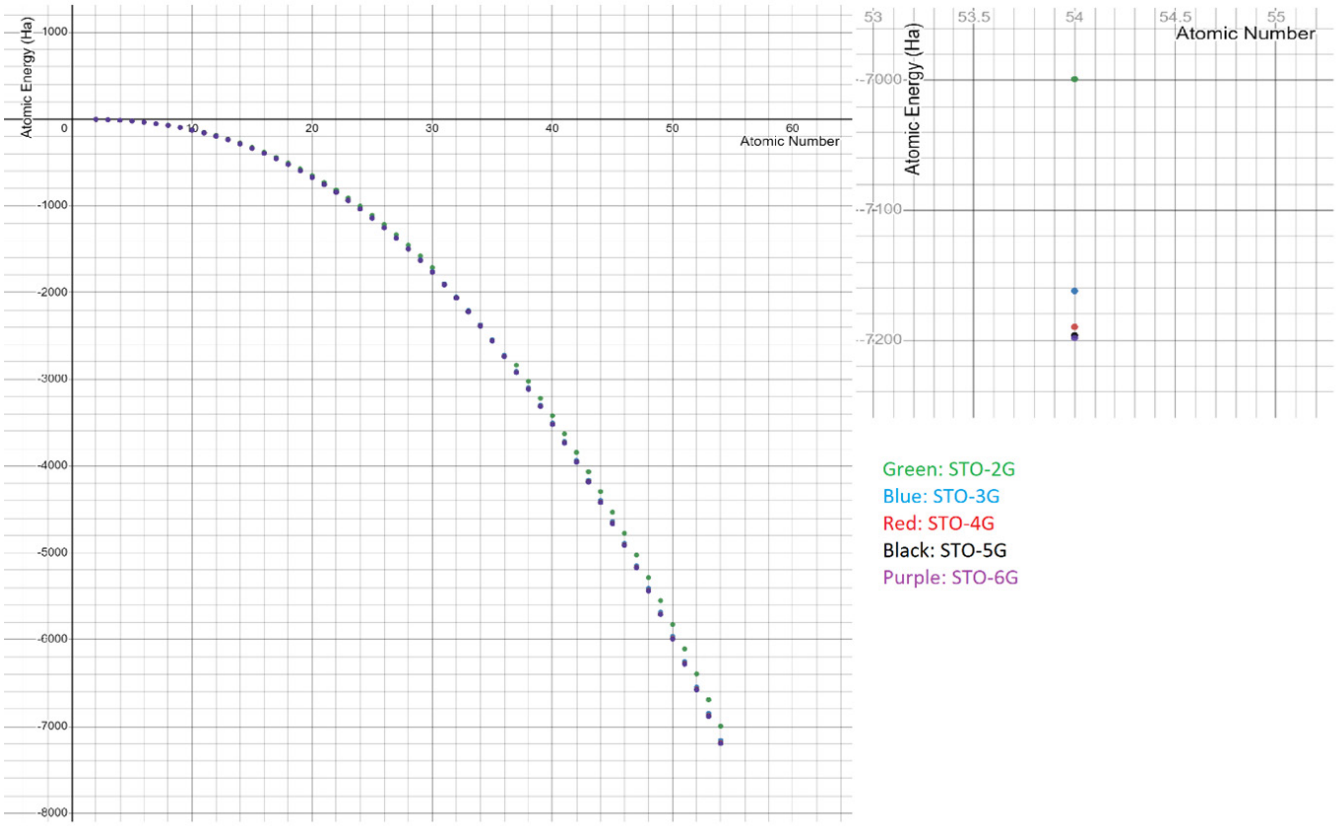

Figure 2. STO-nG Hartree-Fock atomic energy (Hartrees) vs. atomic number for elements, He-Xe. Desmos scatter plot data of STO-nG Hartree-Fock energy vs. atomic number. The image on the left shows all 58 data points. The right side of the figure is a magnification at atomic number 54, corresponding to the energies for Xe, to show the distribution of the 5 basis sets' calculated energies. To view each graph in more detail, see **Appendix 3.**

processed into scatter plots for each STO-nG basis set, and the correlation of the plots was qualitatively categorized.

The complete list of STO-nG atomic energies vs. atomic number that we calculated and tabulated can be found in **Appendix 2**. Graphing the data yielded non-linear scatter plots (**Figure 2**). Basis sets were first imported from the Basis Set Exchange for Hartree-Fock energy calculations. This data was tabulated as LaTeX table images by Matplotlib, and then converted into editable Excel tables using online optical character recognition (OCR) software. From here, data for each basis set (n=2 to 6) from He to Kr was used as a training set to model exponential, quadratic, and sinusoidal regressions. All the data from He to Xe was then used in testing the prediction capabilities of the model. This process has been represented as a flowchart (**Figure 3**).

Only three regression models were tested due to time constraints. The quadratic model was also selected due to its foundation in the Rydberg equation, which postulated a direct relation between electron energy and the square of the atomic number (15). While the Rydberg equation is accurate only for monoelectronic ions and atoms and includes other factors such as the shielding of the protons from the valence electrons by the inner electrons, testing was nonetheless done to observe the accuracy of the quadratic model for multielectronic atoms as well.

Testing was conducted using Desmos. To test our hypothesis, the approximative power of the regressions was first tested by constructing the selected regressions for data from He to Kr (elements 2-36) and calculating the R^2 values, and then the predictive power of the regressions was tested by adding in data from Rb to Xe (elements 37-54). From here, the new R^2 values were compared with the older values, and the difference between them was calculated and tabulated for an analysis of the model's predictive capabilities, and to see if any regression fit consistently for all 5 STO-nG basis sets.

Coefficient of determination calculation

For the data set of atomic energies used to build the regression (represented as values $y_1, y_2, \ldots, y_n$ with mean $\bar{y}$), and the predicted set of energies by the modeled function for those atomic numbers (represented as $f_1, f_2, \ldots, f_n$), the residual error for a specific data point is calculated as the difference between actual and predicted values (Equation 5).

$$e_i = y_i - f_i \tag{5}$$

For an x-axis value (in this case, atomic number) i, there is a corresponding atomic energy in the data set of y_i, a predicted atomic energy by the model of f_i, and a residual of e_i. This error is positive if the actual data point exceeds the predicted value, and vice versa for negative error.

Using the residuals, the coefficient of determination can be calculated for a regression model (Equation 6).

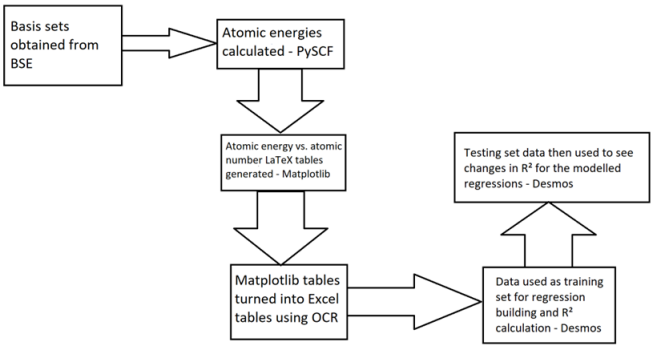

Figure 3. Method flowchart. Simple flowchart depicting the steps in the procedure. BSE – Basis Set Exchange, OCR – Optical Character Recognition.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{6}$$

The coefficient of determination R^2 can be calculated as one minus the fraction of variance unexplained, which represents the error unaccounted for by the model. This fraction is calculated as the sum of the residuals squared over the total sum of squares, which is the sum of the squares of the differences between the dataset energy values and the dataset mean (23).

The coefficient of determination measures how well a regression model fits a dataset. As the magnitudes of the residuals decrease, the value of the coefficient increases. Thus, the maximum coefficient of determination value is 1, and the minimum is 0. Consequently, Desmos uses the method of least squares to form the regressions, which calculates the coefficients for a function to minimize the sum of squares of the residuals (24). This is done through an initial approximation of the coefficients, which then undergoes a recursive process to optimize the coefficients until they converge at a minimum value for the square of the residuals.

General procedural improvements

To increase the efficiency of the experimental procedure, modifications to the investigation are to be instituted for future research and experiment repetition. A major inconsistency was the transfer of data between various applications – first, the data was organized through LaTeX tables as .png images created by Matplotlib. Then, the data was converted into Microsoft Excel .xls tables to allow for editing via OCR, and finally, it was transferred to Desmos for graphing. This three-step approach was adopted due to the strengths of each application: Matplotlib presented the easiest tabular data arrangement approach in Python compared to other tabulation libraries, while Excel allowed for easy data manipulation and division as required on spreadsheets, and Desmos provided intuitive graphical representation and regression validation methods.

However, each transfer step also introduced an area of potential random error in the final regression model built. First, when converting from .png to .xls, the OCR software

rendered certain numbers as letters, such as the digit '5' in the image being converted to 'S' in the .xls file. Moreover, certain images yielded a conversion between the digits '1' and '7' from the image to the file. On Desmos, the regression functions generated only had constants of six significant precision figures, while the determination of coefficients were only calculated to four significant figures. Procedural controls were established to ensure that the error generated at each step was prevented or minimized, such as checking the OCR results with the original image numerous times, by all researchers, to correct any typographical errors and to confirm that the data was accurately converted between file formats.

While Matplotlib and Microsoft Excel also had graphing abilities, Desmos was selected over the two due to the ability to rescale and size the scatter plot graph generated easily and to add new data to the table to check the change in the *R2* value. However, in future experimentation, changes will be made to decrease the time spent on data transfer (such as in the OCR conversion and rechecking), and to increase the precision of measurements. To eliminate the need for the OCR, the data will be outputted by PySCF as a .txt file for easy transfer to a spreadsheet in the immediate future. Further investigation needs to be conducted on potential table-generating Python libraries that allow for copying and pasting of data from the output to a spreadsheet.

Our future research will shift to specialized scientific data processing and regression-building software, such as Vernier Software and Technology's Logger *Pro* 3. While the procedure for regression validation on Logger *Pro* 3 involves more steps than on Desmos, it provides a degree of precision of nine decimal places for all regression calculations. Logger *Pro* also has numerous presets for fundamental regression function building, such as for exponential and sinusoidal regression formulation for dependent variable vs. independent variable scatter plots. Thus, the number of data transfer steps, and the level of random error in the results will be decreased.

ACKNOWLEDGEMENTS

We would like to thank the Aspiring Scholars Directed Research Program, which provided us the opportunity, resources, and funding to pursue our investigation; Shreya Gundani, Shivani Rajagopalan, and Krish Gangal, researchers who laid a foundation for our investigation; and Archith Iyer, Jay Wu, Katherine Xie, Pareekshith Krishna, Rushil Shah, and Sweekrit Bhatnagar, researchers who inspired us to pursue this investigation field.

Received: August 24, 2022
Accepted: September 7, 2023
Published: October 5, 2023

REFERENCES

1. Bartlett, Rodney J., and John F. Stanton. "Applications of Post-Hartree-Fock Methods: A Tutorial." *Reviews in Computational Chemistry*, 2007, pp. 65–169. doi.org/10.1002/9780470125823.ch2.
2. de Broglie, Louis. "Waves and Quanta." *Nature*, vol. 112, no. 2815, Oct. 1923, pp. 540-540. doi.org/10.1038/112540a0.
3. Heisenberg, Werner. *Encounters with Einstein: And Other Essays on People, Places, and Particles*. Princeton University Press, 1989.
4. Schrödinger, Erwin. "An Undulatory Theory of the Mechanics of Atoms and Molecules." *Physical Review*, vol. 28, no. 6, Dec. 1926, pp. 1049–70. doi.org/10.1103/PhysRev.28.1049.
5. Drake, Gordon. "High Precision Calculations for Helium." *Springer Handbook of Atomic, Molecular, and Optical Physics*, Springer New York, 2006, pp. 199–219. doi: doi.org/10.1007/978-0-387-26308-3_11.
6. Hanson, D.M. *et al*. "6.7: The Helium Atom Cannot Be Solved exactly". *Chemistry LibreTexts*, 2020.
7. Hammond, B. L., *et al*. *Monte Carlo Methods in Ab Initio Quantum Chemistry*. World Scientific, 1994.
8. David, Grégoire, *et al*. "Self-Consistent Field Methods for Excited States in Strong Magnetic Fields: A Comparison between Energy- and Variance-Based Approaches." *Journal of Chemical Theory and Computation*, vol. 17, no. 9, Sept. 2021, pp. 5492–508. doi.org/10.1021/acs.jctc.1c00236.
9. Santos, Cleydson B. R., *et al*. "Application of Hartree-Fock Method for Modeling of Bioactive Molecules Using SAR and QSPR." *Computational Molecular Bioscience*, vol. 04, no. 01, 2014, pp. 1–24. doi.org/10.4236/cmb.2014.41001.
10. Magalhães, Alexandre L. "Gaussian-Type Orbitals versus Slater-Type Orbitals: A Comparison." *Journal of Chemical Education*, vol. 91, no. 12, Dec. 2014, pp. 2124–27. doi.org/10.1021/ed500437a.
11. "Electronic Wave Functions - I. A General Method of Calculation for the Stationary States of Any Molecular System." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 200, no. 1063, Feb. 1950, pp. 542–54. doi.org/10.1098/rspa.1950.0036.
12. Hehre, W. J., *et al*. "Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules." *The Journal of Chemical Physics*, vol. 56, no. 5, Mar. 1972, pp. 2257–61. doi.org/10.1063/1.1677527.
13. Pietro, William J., *et al*. "Molecular Orbital Theory of the Properties of Inorganic and Organometallic Compounds. 2. STO-NG Basis Sets for Fourth-Row Main-Group Elements." *Inorganic Chemistry*, vol. 20, no. 11, Nov. 1981, pp. 3650–54. doi.org/10.1021/ic50225a013.
14. Matsuzaki, Rei, *et al*. "Construction of Complex STO-NG Basis Sets by the Method of Least Squares and Their Applications." *Theoretical Chemistry Accounts*, vol. 133,

no. 9, Sept. 2014, p. 1521. doi.org/10.1007/s00214-014-1521-6.
15. Rydberg, J. R. "Investigations of the composition of the emission spectra of chemical elements." *Kongliga Svenska Vetenskaps-Akademiens Handlingar*, vol. 23, no. 11, 1889, pp. 1-177.
16. Hartree, D. R. "The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods." *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, no. 1, Jan. 1928, pp. 89–110. doi.org/10.1017/S0305004100011919.
17. Pauli, Wolfgang. "Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren." *Zeitschrift für Physik,* vol. 31, 1925, pp. 765-783.
18. Fock, Vladimir. "Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems." *Zeitschrift für Physik*, vol. 61, no. 1, 1930, pp. 126-148.
19. "The Molecular Orbital Theory of Chemical Valency VIII. A Method of Calculating Ionization Potentials." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 205, no. 1083, Mar. 1951, pp. 541–52. doi.org/10.1098/rspa.1951.0048.
20. Sun, Qiming, *et al*. "PySCF: The Python-based Simulations of Chemistry Framework." *WIREs Computational Molecular Science*, vol. 8, no. 1, Jan. 2018. doi.org/10.1002/wcms.1340.
21. Pritchard, Benjamin P., *et al*. "New Basis Set Exchange: An Open, Up-to-Date Resource for the Molecular Sciences Community." *Journal of Chemical Information and Modeling*, vol. 59, no. 11, Nov. 2019, pp. 4814–20. doi.org/10.1021/acs.jcim.9b00725.
22. Hunter, John D. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, vol. 9, no. 3, 2007, pp. 90–95. doi.org/10.1109/MCSE.2007.55.
23. Heinisch, O. "Steel, R. G. D., and J. H. Torrie: Principles and Procedures of Statistics. (With special Reference to the Biological Sciences.) McGraw-Hill Book Company, New York, Toronto, London 1960, 481 S., 15 Abb.; 81 s 6 d." *Biometrische Zeitschrift*, vol. 4, no. 3, 1962, pp. 207–08. doi.org/10.1002/bimj.19620040313.
24. "Nonlinear Regressions – Desmos Help Center." *Desmos*, help.desmos.com/hc/en-us/articles/360042428612-Nonlinear-Regressions. us/articles/360042428612-Nonlinear-Regressions.

APPENDIX 1

Link to GitHub repository with the Python code for the algorithms generated using PySCF-BSE-Matplotlib for Hartree-Fock energy calculation and LaTeX tabular organization: https://github.com/ZarseemDyartes/PySCF-Atomic-Energy-Calculations.

APPENDIX 2

	Atomic Energy (E/Ha)				
Atomic Number (Z)	STO-2G	STO-3G	STO-4G	STO-5G	STO-6G
2	-2.702157146	-2.807783957	-2.83572599	-2.843769817	-2.846292095
3	-7.070820937	-7.315526006	-7.376840257	-7.394279132	-7.39993123
4	-13.89023661	-14.3518804	-14.46236256	-14.49316935	-14.50336112
5	-23.39528431	-24.14898866	-24.32912956	-24.37836956	-24.39429456
6	-36.06027444	-37.19839255	-37.47532927	-37.54919293	-37.5723641
7	-52.09700681	-53.71901019	-54.11377517	-54.21730961	-54.24911199
8	-71.57230462	-73.80415026	-74.33687433	-74.4749011	-74.51681631
9	-95.01508371	-97.98650503	-98.68168655	-98.85972134	-98.91325302
10	-122.746034	-126.6045251	-127.4867437	-127.7100967	-127.7767383
11	-155.1599512	-159.6682113	-160.7020065	-160.9572012	-161.0339378
12	-191.4415463	-197.0073545	-198.2596384	-198.568031	-198.6600649
13	-232.0976666	-238.8583582	-240.3343878	-240.7025685	-240.8132643
14	-277.4348738	-285.4662112	-287.1846191	-287.6144056	-287.7446309
15	-327.4946451	-336.8687695	-338.8444164	-339.3387656	-339.4896405
16	-382.3200557	-393.1302194	-395.3853654	-395.9492771	-396.1222861
17	-442.2092562	-454.5421925	-457.1127484	-457.7536431	-457.9504176
18	-507.2492732	-521.2228808	-524.1111039	-524.831786	-525.054179
19	-577.5016492	-593.0773887	-596.2010433	-596.9703008	-597.2072987
20	-652.5536168	-669.9888706	-673.4512749	-674.306045	-674.5707042
21	-732.6267758	-752.0012	-755.8315068	-756.7673674	-757.0580355
22	-818.0237591	-839.5533112	-843.7899486	-844.8188212	-845.1287776
23	-909.1963154	-933.0118294	-937.6697945	-938.7973008	-939.1340302
24	-1005.746566	-1032.074417	-1037.187566	-1038.423019	-1038.803672
25	-1108.799934	-1137.332377	-1142.90814	-1144.256307	-1144.67137
26	-1217.055675	-1248.302373	-1254.395328	-1255.790826	-1256.597396
27	-1331.803591	-1366.04936	-1372.686712	-1374.269047	-1374.688109
28	-1452.626838	-1490.041629	-1496.878246	-1498.85689	-1499.342889
29	-1577.449908	-1620.246154	-1627.958355	-1629.793988	-1629.94936
30	-1712.014001	-1756.276904	-1764.598857	-1766.571917	-1767.1723
31	-1911.823716	-1900.7285	-1909.70188	-1911.823716	-1912.465321
32	-2063.452695	-2051.636256	-2061.193026	-2063.452695	-2064.137844
33	-2221.829645	-2209.263673	-2219.427425	-2221.829645	-2222.559467
34	-2386.888284	-2373.527343	-2384.334495	-2386.888284	-2387.665509
35	-2558.797411	-2544.636781	-2556.092314	-2558.797411	-2559.62205
36	-2737.700782	-2722.706	-2734.837159	-2737.700782	-2738.575159
37	-2837.263872	-2907.603852	-2920.365163	-2923.380771	-2924.304351

38	-3024.488274	-3099.116726	-3112.57479	-3115.75302	-3116.727726
39	-3218.230132	-3297.330817	-3311.471817	-3314.816158	-3315.846302
40	-3419.430948	-3503.034204	-3517.936371	-3521.468846	-3522.554709
41	-3627.688153	-3715.867459	-3731.528793	-3735.240708	-3736.373581
42	-3842.946448	-3935.895633	-3952.37077	-3956.261279	-3957.460721
43	-4065.137271	-4163.205165	-4180.474272	-4184.549008	-4185.807667
44	-4294.52878	-4397.721394	-4415.823885	-4420.089768	-4421.409448
45	-4531.481719	-4639.720512	-4658.695841	-4663.145744	-4664.529713
46	-4775.631762	-4889.314051	-4909.147318	-4913.81582	-4915.265153
47	-5027.284453	-5146.351184	-5167.094754	-5171.974344	-5173.492091
48	-5286.514538	-5411.532716	-5433.171046	-5438.262076	-5439.846968
49	-5551.431926	-5682.777384	-5705.539973	-5710.872906	-5712.540153
50	-5825.817739	-5963.206129	-5986.929084	-5992.490671	-5994.230442
51	-6107.796331	-6251.362462	-6276.06321	-6281.856448	-6283.669667
52	-6397.236056	-6547.122355	-6572.805663	-6578.830564	-6580.717253
53	-6694.306672	-6850.676247	-6877.377171	-6883.643178	-6885.606172
54	-6999.072052	-7162.10421	-7189.851626	-7196.3656	-7198.406703

APPENDIX 3

Link to GitHub repository with all 35 generated STO-nG basis set atomic energy vs. atomic number graphs from He-Xe, and the generated Desmos scatterplots, regression formulation, and regression validation. This data was added in the appendix due to exceeding the figure limit if all graphs and .png Matplotlib tables were included in the paper. https://github.com/ZarseemDyartes/Matplotlib-LaTeX-Tables-and-Desmos-Scatter-Plots.