Discovery of the Predictors of the Standard Heats of Formation of Group 1 and 7 Compounds: A Heuristic Genetic Algorithm with Multiple Regression

Swarup Sai Swaminathan

(i)

As a student in the Medical Sciences Specialized Learning Center of Freehold High School in Freehold, New Jersey, I was given the opportunity to conduct independent research for the complete duration of my junior year. Having been given permission to complete any type of research, I wanted to challenge myself by working in a unique research field: I attempted to combine my knowledge and interest in the life sciences with computer science. A friend suggested the use of regression modeling in my project, as he advised that empirically estimating values was becoming essential in today's scientific world. With helpful guidance from friends, teachers, and advisors, I continued to narrow down my field of research.

Having conducted some exploration on the internet in addition to background research at the library of The College of New Jersey (TCNJ), I desired to use regression modeling to estimate some phenomenon which contained no formula or equation for calculation. The first topic I considered was solubility. However, there seemed to have been a substantial amount of research already conducted concerning this property. After speaking with my science teachers and perusing through science textbooks, I narrowed down the field of interest to one topic: heats of formation. The great range of their values, unpredictability, and lack of a formula made this topic perfect for the research I desired. Additionally, the purpose of my research would be two-fold, as the variables used in the equation would provide speculation as to what atomic properties affected these values.

The background research of my work mostly concerned data gathering. As I was attempting to estimate heat of formation values using a regression equation, I had to compile the data of several variables from various sources, such as the CRC Handbook of Chemistry and Physics as well as the NIST-JANAF Thermochemical Tables. This data collection was completed at the TCNJ library as well as at my high school, with the assistance of my science teachers. However, my background research also involved investigating current methods of estimation in the field of computational chemistry. It was here that I ran into several obstacles, as current methods involved advanced levels of physics and calculus. With great effort and help from teachers, I absorbed the material and attributed the appropriate amount of information into my research.

Nonetheless, the majority of the actual project was conducted at my residence and on my personal computer, as it involved writing a computer program in the Java programming language. I had to input the compiled data into the program, and use it to discover the best regression equation.

Completing this research work demanded great effort. I had to dedicate a large amount of time in order to understand the intricacies of the Java programming language. Although I had learned programming skills during my freshman year of high school, I had to further my skills by learning Java functions. I had to understand how to integrate external classes into my program where necessary. This evolved into a challenging yet rewarding exercise. Although I spent tireless nights attempting to discover flaws in my program and adding aesthetically pleasing characteristics such as the Graphical User Interface (GUI), successful results brought great satisfaction.

Completing this research most certainly opened my eyes to the world of research and the applications of science in today's world. While high school science classes teach the basics principles in these vast fields of study, research allows students to truly benefit from and observe the roles that science and technology play. With this introduction to this fascinating world, I am convinced that I will continue research in various fields during my college and medical school years.

In retrospect, there are several steps I took that resulted in the successful completion of this research work. The most important key to success was perseverance, and the desire to challenge oneself. If a student uses the opportunity of researching a topic of their choice to the fullest, he or she will benefit greatly. Additionally, choosing a field of interest is important, as the student will be dealing with this field for several months and possibly years. Finally, completing unique research is paramount, as this will allow students to bridge the gap between mathematics and the sciences. If a student researches a topic of interest with great persistence, he or she will greatly increase their chance of success.

(ii)

A topic discussed in high school chemistry classes, the heat of formation (ΔH_f°) is the amount of heat released or absorbed when a mole of a compound is formed from its elements. For example, the ΔH_f° value of water, H_2O, is the measured change in heat when two hydrogen atoms combine with an oxygen atom to form water. Heats of formation play an important role in the outcome of chemical and pharmaceutical products such as semiconductors, medicines, and even nuclear therapy. Since the ΔH_f° values for

several compounds are unknown and no simple equation exists for their calculation, much research has been conducted on empirical methods.

This has resulted in the creation of calculus-based methods, involving several advanced yet costly methods. No inexpensive theoretical model exists for the direct estimation of heats of formation. Thus, this research attempted to create an accurate theoretical model that estimated ΔH_f° of certain compounds, using appropriate and available chemical properties as the variables in a regression model. Additionally, little is known concerning which properties actually affect heats of formation at the atomic level. This research also attempted to discover these properties, which will be referred to as the "determinants."

Regression models are used to predict values. For example, Y = 3A + 4B + 5C is a regression model, where A, B, and C are the variables or predictors that predict a value for Y, which is some other property. Similarly, I attempted to estimate the heat of formation values of Group 1 and 7 compounds using chemical properties as the predictors. Group 1 and 7 compounds are formed from Group 1 elements, such as sodium, potassium, and cesium, and their combination with Group 7 elements, such as fluorine, chlorine, and bromine. An example of a Group 1 and 7 compound is sodium chloride, NaCl.

After selecting approximately twenty properties based on availability and postulation concerning their affect on the heat of formation value, data was collected from various sources, such as the CRC Handbook of Chemistry and Physics and the NIST-JANAF Thermochemical Tables. These properties include molecular mass, number of electrons, bond length, density, boiling and melting points, combine atomic radii, x-ray

absorption energy, electron binding energy, and molar heat capacity. This data was then placed in a Microsoft Excel spreadsheet. These properties were potential determinants of the heats of formation.

The Java computer program then became the main focus of this research. By applying a genetic algorithm, the regression equation that consisted of four out of the twenty predictors which best estimated the heats of formation would be found. A genetic algorithm is a computer search procedure that mimics the mechanics of evolution and natural biological processes such as mutation and crossover. Essentially, the predictors chosen in a model would act similar to genes during the process of reproduction, and through "mutating" and exchanging predictors, the best model would be discovered.

The adjusted R^2 statistic was used to quantify the accuracy of a model. Adjusted R^2 values range from 0 to 1, where a value close to 1 indicates a nearly perfect model. The computer program was written to generate 50 regression models, each containing four randomly selected predictors. After the adjusted R^2 values for all 50 models were calculated, the models were manipulated using the genetic algorithm and after approximately twenty iterations, the optimal regression model was discovered.

Model 1, the most successful model, had the equation: $\Delta H_f^\circ = -2143.649 + 0.061(\textit{Thermal Conductivity}) + 0.306(\textit{Critical Temperature}) + 0.830(\textit{Bond Length}) + 19.654(\textit{Molar Heat Capacity})$. The regression equation of Model 2, the second best model, was: $\Delta H_f^\circ = -1461.990 - 0.047(\textit{Boiling Point}) + 16.124(\textit{Molar Heat Capacity}) + 0.288(\textit{Critical Temperature}) - 4.074(Z_{\textit{eff}})$.

Model 1, had an adjusted R^2 value of 99.6%, while Model 2 had a value of 99.5%. Once again, adjusted R^2 values that are close to 1 are highly desired, which indicates the strength of these results.

A Graphical User Interface, or GUI, was then developed to increase the flexibility for the user, allowing the user to control several features, such as the location of the file containing all molecular data, the number of models to be generated in one iteration, the number of compounds, the number of variables, the number of models to undergo mutations and crossover, and even where the program's output should be placed. A window would appear when the program ran, prompting the user to complete these fields.

Because the best model for Group 1 and 7 compounds may not have been the best for all compounds, both predictor sets were tested on Group 2 and 7 compounds, which are composed of a Group 2 element and a Group 7 element, resulting in a compound such as barium fluoride, BaF_2. The MINITAB Student Edition statistical software package was then used to find the regression equation and its statistics for the two sets of predictors. That is, two models based on Group 2 and 7 data were created with each using the same respective predictors used in Models 1 and 2.

The regression equation using the predictors of Model 1 was: ΔH_f° = -5360.500 + 0.368(Thermal Conductivity) + 1.000(Critical Temperature) – 5.249(Bond Length) + 56.750(Molar Heat Capacity). This model, Model 1A, had an adjusted R^2 value of 66.6%. The regression equation using the predictors in Model 2 was: ΔH_f° = 458.300 – 0.017(Critical Temperature) + 1.202(Molar Heat Capacity) – 0.351(Boiling Point) – 17.380(Z_{eff}). The adjusted R^2 value of this model, Model 2A, was 93.2%.

It was concluded that the multiple linear regression method was tremendously successful, as the ΔH_f° of any compound can be accurately predicted from one of the two sets of predictors. The first set includes bond length, molar heat capacity, thermal conductivity, and critical temperature, while the second is comprised of molar heat capacity, boiling point, critical temperature, and Z_{eff}. The models can be used to predict ΔH_f° of other compounds, including those that have unknown or undocumented values for its ΔH_f°.

The importance of the predictors chosen by the program must be emphasized, as these properties are potential determinants of ΔH_f°. The selection of bond length as a predictor indicates that ΔH_f° must be somewhat related to the distances between atoms. The relationship between ΔH_f° and intermolecular forces becomes evident with the selection of boiling point. The presence of molar heat capacity, thermal conductivity, and Z_{eff} in the models illustrates the importance of electrons in determining ΔH_f° and how tightly bound they are to the nucleus. In addition, the large coefficient for molar heat capacity in the models indicates its importance in influencing ΔH_f°. Finally, the significance of pressure, atomic velocity, and intermolecular distance in affecting ΔH_f° is apparent by the inclusion of critical temperature. All these properties appear to be crucial in determining the value of ΔH_f°.

Viewing the Group 2 and 7 models as an examination of the predictors chosen, Model 2A appeared to be the stronger model, as it produced a higher adjusted R^2 value. Although Model 1 produced a higher adjusted R^2 than Model 2 for Group 1 and 7 compounds, the Model 2 predictors strongly suggest that their bases are the determinants of ΔH_f°.

Overall, this research was used to discover the optimal regression model that best estimated the heat of formation values of Group 1 and 7 compounds. This work also yielded possible properties that affect these values at the atomic level. With such a model, experimental and expensive methods can be eliminated, leading to more cost efficient and time efficient methods in the field of computational chemistry.