Predicting Heat of Combustion from SMILES with Machine Learning

Braden Garrett
Utah Valley University
Computer Science Department
Orem, Utah
11027233@uvu.edu

Jingpeng Tang, Ph.D.
Utah Valley University
Computer Science Department
Orem, Utah
jtang@uvu.edu

Xi Chen, Ph.D.
Utah Valley University
Computer Science Department
Orem, Utah
xi.chen@uvu.edu

Keywords—machine learning, Neural Net, Chemistry, Heat of Combustion, SMILES

I. Abstract

Accurately predicting the heat of combustion of chemical compounds is essential for applications in energy research, material science, and chemical engineering. Traditionally, bomb calorimetry is used to determine this value; however, this method is time-consuming, costly, and poses safety risks. In this study the use of machine learning is explored for heat of combustion prediction of molecules directly from their structural representations. A dataset was compiled from multiple sources, including experimental data from the National Institute of Standards and Technology (NIST) and PubChem, with preprocessing to standardize sample format and remove samples with non-standard data. A neural network was developed and trained using ChemBERTa-generated[10] embeddings of molecular SMILES representations. Initial results using a limited dataset produced an R^2 value of 0.522 and a mean squared error (MSE) of 9,970,441, suggesting the need for additional data. A synthetic dataset was created by duplicating the existing set. Using this dataset the model's performance significantly improved, achieving an 80% reduction in MSE and an R^2 of 0.896. These results demonstrate the feasibility of using machine learning to approximate combustion energy, reducing reliance on experimental methods. Future work will focus on expanding the dataset, refining hyperparameters, and improving model scalability with GPU acceleration to enhance predictive accuracy further.

II. Introduction

Heat of combustion measures the energy released when a substance undergoes complete combustion in an oxygen-rich environment under controlled conditions. This energy results from both the breaking of chemical bonds in the reactants and the formation of new bonds in the combustion products, with the net energy release depending on the types of bonds and how the atoms are arranged within the molecule. Bomb calorimetry is a well-established technique to determine the heat of combustion of a substance by measuring the change in thermal energy during combustion. This technique provides insights into the strength of chemical bonds within molecules [1] and can also be applied to calculate the caloric content of food [2]. In a bomb calorimeter, a sample is combusted in a sealed, high-pressure chamber surrounded by a known quantity of water. The reaction occurs in excess oxygen and at a constant volume. The result is obtained by tracking the temperature increase of the surrounding water. While highly accurate this process is time-intensive and involves safety risks due to high-pressure and high-temperature conditions.

The outcome of a bomb calorimetry experiment is intrinsically linked to the molecular composition and the types of bonds present within a substance. Variations in bond energies, molecular symmetry, and atom organization all influence the total heat released during combustion. All these factors are encoded in a Simplified Molecular Input Line Entry System (SMILES). These SMILES are string representations of molecules. For example, acetic acid (CH_3CO_2H), the primary acid in vinegar, is represented by the SMILES string CC(O)=O. Due to the high amount of encoded information in a simple string format, these SMILES are great candidates for leveraging computational methods to predict calorimetric outcomes. Advances in machine learning have demonstrated that models trained on molecular representations can capture the nuanced relationships between molecular structure and energy release, thereby offering a promising alternative to direct experimental measurements [3].

III. Literature Review

The foundation of this project lies in leveraging machine learning to predict molecular properties from SMILES representations. Nakajima et al. demonstrated the potential of machine learning in predicting the bond dissociation enthalpy of hypervalent iodine compounds with high accuracy [5]. Their study employed models such as random forest and support vector machines, showcasing a more accurate and efficient alternative to traditional approaches like the Mendeleev equation or experimental bomb calorimetry for hypervalent iodine compounds. For comparison, the Mendeleev equation (eq. 1) predicts the heat release from combustion as follows:

$$q\,(MJ/kg) = 0.339\,C + 1.035\,H + 0.109\,S - 0.109$$

Equation 1 - Mendeleev Equation

Kiran et al. further illustrated the power of ML by training a neural network on SMILES data to predict the standard enthalpy of formation, which is the energy needed to form a compound, of hydrocarbons [7]. Using k-fold cross-validation, they achieved robust results while targeting specific molecular classes.

Building on these foundational works, this project aims to develop a general-purpose machine learning model to predict the heat of combustion for a broad spectrum of molecules. Unlike the aforementioned studies, which were constrained to particular compound types, our approach seeks to create a more versatile tool by utilizing a diverse training dataset. The accuracy of the proposed model will inherently depend on the

quality and scope of the training data, which must capture the wide variability in molecular structures and bond energies. By employing advanced neural network architectures and comprehensive validation techniques, this work aims to provide a safer, faster, and more cost-effective alternative to experimental calorimetry while maintaining an accuracy within 5% of real-world measurements.

IV. METHODOLOGY

Data was collected from two sources and combined into a single dataset to maximize the training data available. The first source was a paper published by the National Institute of Standards and Technology [8], which reports the heat of combustion and other relevant physical properties for approximately 450 molecular compounds. The second source was PubChem, where the support team provided guidance on filtering the database for molecules with a reported heat of combustion values. A JSON document was generated from PubChem and parsed to extract the desired values along with the corresponding CID (compound identifier) for each molecule. The data format is shown in figure 1. All relevant data were extracted from these sources and stored in CSV format (figure 2). Using the CID, the SMILES strings were retrieved via a PubChems API call. Because the data were not originally represented in standard units and conversion to standard units would be a long process that is different for each data point, a significant portion was excluded to ensure consistency across all molecules. The resulting combined dataset comprised 285 data points, each containing a SMILES string and an associated heat of combustion value reported in kJ/mol. From these samples, 20 were removed and set apart as a testing dataset.

```
1.  {
2.          "SourceName": "Hazardous Substances
    Data Bank (HSDB)",
3.          "SourceID": "30",
4.          "Name": "NITROGLYCERIN",
5.          "Description": "The Hazardous
    Substances Data Bank (HSDB) is a toxicology
    database that focuses on the toxicology of
    potentially hazardous chemicals.
6.          "URL":
    "https://pubchem.ncbi.nlm.nih.gov/source/hsdb/3
    0",
7.          "LicenseURL":
    "https://www.nlm.nih.gov/web_policies.html"
8.          "Data": [
9.            {
10.             "TOCHeading": {
11.               "type": "Compound",
12.               "#TOCHeading": "Heat of
    Combustion"
13.             },
14.             "Description": "PEER REVIEWED",
15.             "Reference": [
16.               "O'Neil, M.J. (ed.). The Merck
    Index - An Encyclopedia of  Chemicals, Drugs,
    and Biologicals. 13th Edition,
    Whitehouse  Station, NJ:  Merck and Co., Inc.,
    2001., p. 1185"
17.             ],
18.             "Value": {
19.               "StringWithMarkup": [
20.                 {
21.                   "String": "1580 cal/g"
22.                 }
23.               ]
24.             }
25.           }
26.         ],
27.         "ANID": 2,
28.         "LinkedRecords": {
29.           "CID": [4510]
30.         }
31.       },
```

Figure 1 - PubChem JSON Data

```
1.  Smile,Heat of Combustion
2.  [H][H],-285.826
3.  [O][O],0.0
4.  OO,-98.48
5.  O,0.0
6.  N,-383.181
7.  NN,-622.34
```

Figure 2 - Final Data Format

A synthetic dataset was created by duplicating all existing points 3 times to create a combined total of 795 data points. This set was used to create embeddings and train the model with great improvement to results that will be discussed later on in the paper.

Although SMILES strings inherently encode valuable structural and chemical information about molecules, the raw data were further preprocessed to create representations more conducive to machine learning. We employed ChemBERTa, a large language model pre-trained on chemical data, to generate molecular embeddings. In parallel, we converted the same dataset into ECFP4 fingerprints[11]. We then compared the performance of models trained on these two different representations. Notably, all models consistently exhibited lower mean squared error when using the ChemBERTa embeddings compared to the ECFP4 fingerprints, suggesting that the learned embedding space more effectively captures the nuances of molecular structure relevant for predicting heat of combustion. This improvement in accuracy led to the use of ChemBERTa for embeddings generation on all experiments ran.

To determine which model architecture would yield the most accurate predictions while maintaining a reasonable training time, multiple regression models were developed and evaluated. For initial testing, models from SciKit-Learn[12], a widely used machine learning library, were implemented. Among these, a Support Vector Machine (SVM) demonstrated the best performance, achieving the highest accuracy and the fastest training time compared to the other models tested. Building on these results, testing was expanded to include a Convolutional Neural Network (CNN) implemented in Keras[13]. The CNN exhibited substantial improvements in predictive accuracy, as measured by mean squared error and R^2 values, surpassing the SVM. Given its superior performance, the CNN became the primary focus of further research and model refinement.

Model Type	**R^2**	**MSE**
Random Forest	0.09610	7,353,301
XGBoost	-0.04456	12,084,850
SVM	0.17999	9,525,979
CNN	0.522	9,970,441

Table 1 - Model Statistics

Figure 3 shows the path of data form SMILE to output in the Neural Network that achieved the highest accuracy with respect to R^2. Training time for the neural network was 1 minute and 30 seconds for 6000 epochs using 5 rounds of k-fold validation. Another minute was needed to generate the embeddings using the ChemBERTa model. All processing was done on a AMD Ryzen 9 49000HS CPU.

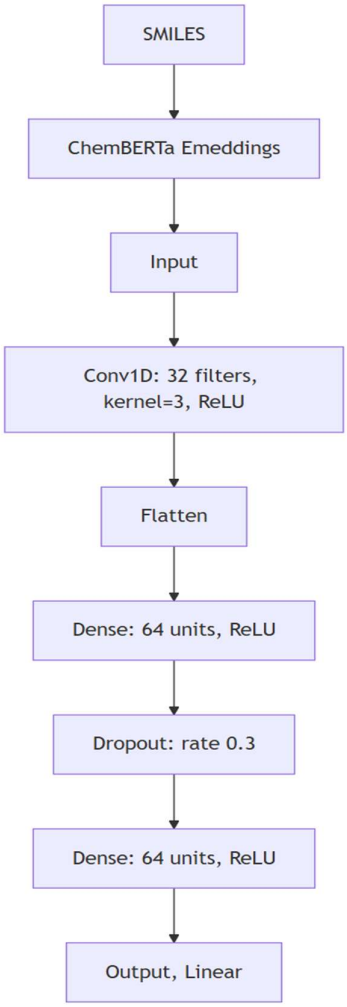

Figure 2 - Full Data Path

V. Outcomes

The primary goal of this project was to develop a machine learning model capable of accurately predicting the heat of combustion for any molecule provided in the Simplified Molecular Input Line Entry System (SMILES) format. This would reduce the need for bomb calorimetry experiments and expedite research involving the study of bond energies. A neural network was created and trained on as much experimental data as could be reasonably attained. The model architecture is documented in figure 4 at $1/8^{th}$ scale.

After training, the neural network reported an R^2 value of 0.522 and an average mean squared error of 9,970,441. These results are less accurate than we had hoped for. One source of trouble could be that the dataset is too small. The experimental results for heat of combustion are not recorded for many molecules and were often reported in non-standard units which greatly reduced the size of the dataset.

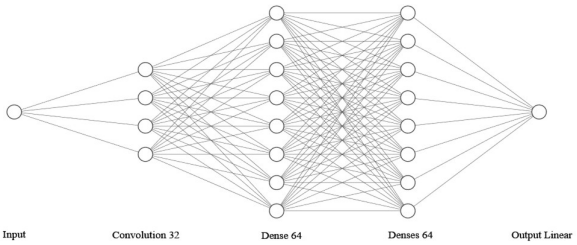

Figure 4 -Neural Net Architecture

There was a pattern of smaller molecules being more accurate as shown in table 2. It is unclear of this is due to the dataset having more small molecules or if the model has a higher capability for predictions on smaller molecules.

SMILES	Prediction	Target	Diff.
Most Accurate			
NC	-1088.39	-1086.81	1.58
N#CC#N	-1093.25	-1097.07	3.81
C(CN)C=O	4164.06	4159.50	4.56
Least Accurate			
CC1C2CCC1CC(C2)…	1067.86	4863.73	3795.86
C1=NC(=C(N1)C(=O…	1041.53	-1910.62	2952.15
C1=CC=C2C(=C1)C3…	7133.93	10040.60	2906.6

Table 2 - Most/Least Accurate Predictions

The neural network was retrained on an artificially expanded synthetic dataset, resulting in a significant improvement in training performance. Table 3 shows the statistical improvements the synthetic dataset provided. These enhancements highlight the model's potential for accurately predicting heat of combustion. Notably, the largest prediction error decreased from 30,807 kJ/mol to 4,156 kJ/mol, demonstrating a substantial reduction in outlier deviations. With the increase in the size of the dataset the training time increased to 2 minutes and 40 seconds. This is a roughly linear increase showing promising potential for the scalability of the model. The training was also done on a CPU, which isn't optimal for machine learning. The use of a GPU would accelerate the speed of training and allow for quick processing of massive datasets.

Metric	Standard Dataset	Synthetic Dataset	Difference
MSE	9,970,441	2,104,165	↓ 80%
R^2 Value	0.522	0.896	↑ 75%
Largest Error	30,807	4,156	↓ 86.5%

Table 3 - Synthetic Data improvement

To finalize results, prediction was done on the testing data. These results are recorded in table 4. The accuracy of of the model for the unseen testing data was not the most promising. The trend continued of small molecules being accurate but the model does not have the capacity to give valid results when molecules get above double digit atoms. The results for small molecules are useful but the model is currently too limited for general use.

MAE	R^2	MSE
1558	0.241	4,913,290

Table 4 - Test Data Validation

V. Future Work

To enhance the accuracy of predictions, a key focus will be expanding the dataset to include a broader range of molecular structures. This expansion should incorporate molecules of varying sizes, from small organic compounds to larger, more complex structures, as well as a diverse set of functional groups. A more comprehensive dataset will improve the model's ability to generalize, enabling more accurate predictions for novel compounds. This can be accomplished by standardizing the data points that were removed from the data set due to inconsistent units of measurement, allowing them to be utilized in the training data. Access to other chemical databases is another method to expand the dataset and increase the variety of compounds accessible. The last way to increase and diversify the dataset is through self-experimentation. By performing our own bomb calorimetry experiments, we could consistently add data points and control the variety of the types of molecules.

Beyond dataset improvements, further experimentation is required to fine-tune the model architecture. This includes optimizing hyperparameters such as learning rate, activation functions, and layer configurations to maximize predictive accuracy while preventing overfitting. The impact of additional hidden layers, dropout regularization, and alternative loss functions should also be explored to determine the most effective neural network configuration. By systematically evaluating these factors, the model can be further refined to achieve greater accuracy and robustness in predicting heat of combustion values.

References

[1] Skinner, H. A. (1965). The strengths of metal-to-carbon bonds. In Advances in Organometallic Chemistry (Vol. 2, pp. 49-114). Academic Press.

[2] Sunner, S., & Månsson, M. (Eds.). (2016). Combustion Calorimetry: Experimental Chemical Thermodynamics. Elsevier.

[3] Šegota, S. B., Anđelić, N., Lorencin, I., Musulin, J., Štifanić, D., & Car, Z. (2021, October). Preparation of simplified molecular input line entry system notation datasets for use in convolutional neural networks. In 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 1-6). IEEE.

[4] TheDevastator. (2020). Wikipedia molecules properties dataset [Data set].Kaggle. https://www.kaggle.com/datasets/thedevastator/wikipedia-molecules-properties-dataset

[5] Nakajima, M., & Nemoto, T. (2021). Machine learning enabling prediction of the bond dissociation enthalpy of hypervalent iodine from SMILES. *Scientific Reports*, *11*(1), 20207.

[6] Ioelovich, M. (2020). Short Overview of Methods for Calculation of Combustion Heat.

[7] Yalamanchi, K. K., Van Oudenhoven, V. C., Tutino, F., Monge-Palacios, M., Alshehri, A., Gao, X., & Sarathy, S. M. (2019). Machine learning to predict standard enthalpy of formation of hydrocarbons. *The Journal of Physical Chemistry A*, *123*(38), 8305-8313.

[8] Burgess, Jr.(Donald R.), Burgess Jr, D. R., & Hamins, A. P. (2023). *Heats of Combustion and Related Properties of Pure Substances*. US Department of Commerce, National Institute of Standards and Technology.

[9] PubChem. (n.d.). *PubChem*. http://Pubchem.com/

[10] Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. arXiv. https://arxiv.org/abs/2010.09885

[11] ChemAxon. (n.d.). Fingerprints: Extended Connectivity Fingerprint (ECFP). Retrieved April 30, 2025, from https://docs.chemaxon.com/display/docs/fingerprints_extended-connectivity-fingerprint-ecfp.md

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[13] Chollet, F., & others. (2015). *Keras* [Computer software]. GitHub. https://github.com/keras-team/keras