Evaluating a machine learning approach to molecular representation

Luke McGovern
Mentored by Ian Michel-Tyler

Introduction

Artificial intelligence is one of the most rapidly developing scientific fields. Contemporary examples of the capabilities of artificial intelligence are the images, art, and articulate thoughts that programs such as DALL-E and ChatGPT generate. These programs utilize machine learning and natural language processing (NLP), two subfields of artificial intelligence. Machine learning is comprised of programs which improve their performance on a task given repeated trials. This is useful for NLP tasks, in which a computer attempts to interpret human language. An example of an NLP application exists in computational chemistry is molecular representation (the way in which a molecule's identity is expressed). By representing molecules through a concise text string, the amount of computer processing required is reduced. Several different machine learning approaches have been used for molecular representation, the most successful utilizes a variational autoencoder (VAE). The success of a VAE can be attributed to its ability for continuous representation, ensuring unique outputs which improve efficiency (Wigh et al., 2022). The junction tree variational autoencoder (JTVAE), an algorithm which produces a scaffold over chemical substructures and combines them into one string utilizing a VAE, was investigated for this reason (Jin et al., 2019). The purpose of the project was to evaluate the accuracy of the JTVAE in molecular representation.

The JTVAE model was constructed over two iterations on a limited biochemical data set. The goal was to improve the model's predictive accuracy over multiple iterations. The model was evaluated on its ability to predict molecular density (g/mol) and heat of formation (kJ/mol).

Materials and Methods

Google Colab was selected as the medium for coding, which utilizes Python. For this reason, prerequisite training was completed through the "Machine Learning Crash Course", an online machine learning tutorial for Google Colab. Further training from the data science website Kaggle provided experience in producing predictive algorithms. This served as a basis for use of predictive properties to be utilized in JTVAE molecular representation. With preliminary training codes completed, work began on producing a functional, first program for molecular representation. A set of 1,305 molecules was obtained through Springer API as a sample database. Molecules were selected by using keyword query searches for: "amino", "methyl", "phosphate", "hydroxyl", "carboxyl", and "carbonyl". The article abstracts for each molecule in the set were downloaded as .json files containing information on molecular geometry and features, as well as their SMILES strings (a text string representing a molecule). A literature alignment model was created in Python to align the text of the named entities (important features). The named entities were uploaded to the molecular representation model. The first model was trained with molecular graph encodings (Balakrishnan et al., 2021). Once the code could be executed without syntax errors (the first iteration of the program), predicted values for density (g/mol) and heat of formation (kJ/mol) were collected, to determine the model's predictive accuracy (Tables 1 and 2). This was repeated in the second iteration where the model was trained with a combination of graph encodings and semantic information.

Materials and Methods (continued)

SMILES string	Iteration 1 percent error	Iteration 2 percent error
c1(cc(c(c(c1)N(=O)=O)N(=O)=O)N(=O)=O)C	6.260	4.795
O=N(=O)C1=C(C(=NN1)N(=O)=O)N(=O)=O	7.905	2.523
n1c(nn(c1N)/C(=N/N(=O)=O)/N)N(=O)=O	1.331	0.210

Table 1 (above): The difference between algorithm-predicted and actual values for density (g/mol) of three specific molecule over the two iterations of the model.

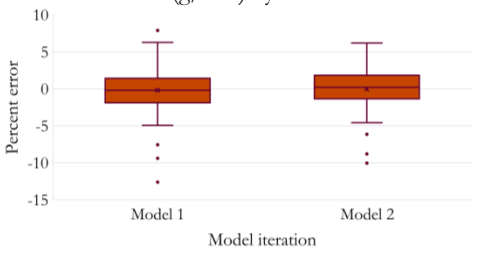
SMILES string	Iteration 1 percent error	Iteration 2 percent error
c1(cc(c(c(c1)N(=O)=O)N(=O)=O)N(=O)=O)C	−193.57	−166.45
O=N(=O)C1=C(C(=NN1)N(=O)=O)N(=O)=O	214.60	124.40
n1c(nn(c1N)/C(=N/N(=O)=O)/N)N(=O)=O	−36.73	−29.84

Table 2 (above): The difference between algorithm-predicted and actual values for heat of formation (kJ/mol) of three specific molecules over the two iterations of the model.

Results

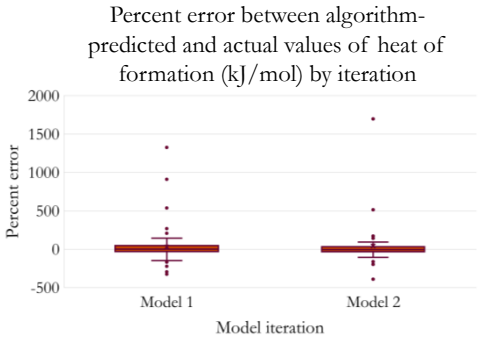
The percent error was tracked from iteration one to two (Graphs 1 and 2). Descriptive statistics were used to determine the difference between predicted and actual values for both molecular density (g/mol) and heat of formation (kJ/mol) across both iterations of the model.

Graph 1 (right): A box-and-whisker plot of the percent error between predicted and actual values for molecular density (g/mol) of machine learning model one (M = −0.183, SD = 3.495) and model two (M = −0.015, SD = 2.964). A value of zero indicates a perfect prediction (n = 81).



Results (continued)

Graph 2 (right): A box-and-whisker plot of the percent error between predicted and actual values for heat of formation (kJ/mol) of machine learning model one (M = −30.41, SD = 209.24) and model two (M = 13.55, SD = 216.16). A value of zero indicates a perfect prediction (n = 81).



Conclusions

The purpose of this project, to develop a machine learning approach to molecular representation utilizing a JTVAE framework, was successfully completed. The percentage change (Table 3) was calculated from iteration one to iteration two to confirm model improvement across these iterations.

Molecular trait	Percent change in mean error	Percent change in standard deviation
Molecular density	91.80%	15.20%
HOF	55.44%	−3.30%

Table 3 (above): The percent change in mean error and standard deviation for molecular density and heat of formation over the two model iterations.

Analysis of this project was limited to organic molecules. In further studies, inorganic molecules should be used to further determine accuracy. Predictive accuracy should also improve with a larger quantity of molecules in a data set. This project supports JTVAE as an effective pathway for the development of molecular representation. An application of this molecular representation is in the search for molecules in the pharmaceutical, energetic, and materials industries.

References

Balakrishnan, S., VanGessel, F. G., Boukouvalas, Z., Barnes, B. C., Fuge, M. D., & Chung, P. W. (2021). Locally optimizable joint embedding framework to design nitrogen-rich molecules that are similar but improved. *Molecular Informatics, 40*(7), 1–17. https://doi.org/10.1002/minf.202100011

Jin, W., Barzilay, R., & Jaakkola, T. (2019). Junction tree variational autoencoder for molecular graph generation. 1–17. https://doi.org/10.48550/arXiv.1802.04364

Wigh, D. S., Goodman, J. M., & Lapkin, A. A. (2022). A review of molecular representation in the age of machine learning. *WIREs Computational Molecular Science, 12*(5), 1–8. https://doi.org/10.1002/wcms.1603