Property Estimations with Chemical Accuracy for Industrially Relevant Molecules Using Geometric Deep Learning

Maarten R. Dobbelaere[1], István Lengyel[1,2], Christian V. Stevens[3], Kevin M. Van Geem[1]*

1 Laboratory for Chemical Technology, Ghent University, Ghent, Belgium; 2 ChemInsights LLC, Dover DE, United States of America; 3 SynBioC Research Group, Ghent University, Ghent, Belgium

**Corresponding author: Kevin.VanGeem@UGent.be*

Highlights

- Open-source property estimation tool for 2D and 3D molecular data.
- New large property databases with industrially relevant compounds and properties.
- Error below 2.5 kJ/mol for high level-of-theory enthalpy of formation predictions.
- Using the molecular geometry in machine learning is essential for high accuracy.

1. Introduction

Knowledge of physicochemical properties is a prerequisite in many chemical engineering tasks, such as developing kinetic models or optimizing industrial processes [1]. *Ab initio* modeling of thermochemical properties (*e.g.*, enthalpy of formation) is a well-established method that can reach chemical accuracy (~4 kJ/mol) for a wide range of molecules. The drawback of this method is the computational expense, which limits its use for large numbers of molecules. Faster methods are density functional theory (DFT) modeling or group contribution (GC) methods. While DFT calculations are also reliable for a vast application domain, the accuracy of the results is lower. GC approaches, on the other hand, can reach chemical accuracy but are very limited in applicability. In the past decade, machine learning (ML) models have been developed to predict thermochemical properties, aiming at chemical accuracy for a broad range of molecules at a low computation expense [2].

ML models are not yet reliable enough to replace established property estimation methods in the chemical industry. This is because they are typically trained on artificial databases with application ranges incompatible with industrial requirements [3]. To overcome this problem, we have created ThermoG3, an extensive database containing high and low-level-of-theory *ab initio* thermochemistry data of 53,000 molecules with industrial relevance. This dataset is used to train a newly developed ML model from the message-passing neural network (MPNN) family that can handle molecules in 2D and 3D format. Since molecular energetics depend on the spatial arrangement of the atoms, it is essential to include 3D information.

2. Methods

A message-passing neural network (MPNN) was constructed to create the structure-property relationships. This MPNN treats a molecule mathematically as a three-dimensional graph with nodes (atoms) and edges (bonds). The edges are directed, meaning that a chemical bond is represented by two directed edges. The MPNN is initialized with node feature vectors that describe the atoms and bonds. Two feature vectors are compared: atomic radial distribution functions (RDF) and simple atomic features (Features), such as atomic weight, number of neighbors, etc. In the message-passing phase, learned bond feature vectors are iteratively updated using messages from atoms in the neighborhood of the considered atom. This neighborhood is defined as a sphere with a cutoff radius around the starting atom of the directed edge so that messages can also be sent from atoms not covalently bonded. In the readout phase, a molecular representation is created from the learned atomic feature vectors. The learned molecular representation is used as input for a deep neural network.

A machine learning technique called Δ-learning is applied. Herein, the difference between low-level-of-theory data (B3LYP) and high-level-of-theory data (G3MP2B3) is learned with the aforementioned machine learning algorithm that uses the 3D structure of the molecule as input. ThermoG3, an extension of the earlier reported dataset by Plehiers *et al.* [4], is used for training. It consists of 53,000 organic compounds ranging in size from 1 to 23 heavy atoms and ten different types of heteroatoms (O, N, S, F, Cl, Br, Ge, B, Si, P). The standard enthalpy of formation at 298 K is calculated at B3LYP/6-31G* and G3MP2B3 levels for all compounds.

3. Results and discussion

The importance of using 3D information in the MPNN is investigated by training different model modifications on the ThermoG3 database and evaluating them on a fixed external test set. We consider two baseline models: an MPNN that only uses the 2D molecular graph and an MPNN that uses the 3D molecular graph but a different message-passing function. In that baseline model, messages can only come from atoms that are covalently bonded to the starting atom of the directed edge. The other models are as described above but differ in cutoff radius. The results of this investigation are shown in Figure 1a. It is shown that the 2D baseline model performs poorly compared to the 3D model, but also that the appropriate choice of the cutoff radius and initial features is beneficial for the model performance. Figure 1b shows the parity plot and error distribution for the 3D MPNN with simple atomic features and a cutoff radius of 2.1 Å and for the initial B3LYP calculations. The ML model can achieve chemical accuracy without any outliers, whereas the B3LYP calculations have a large spread and a tendency to overpredict the enthalpy of formation. Since this Δ-ML technique still requires DFT calculations as input (3D molecular geometries), it can be seen as a correction method that enables researchers to rapidly estimate the enthalpy of formation for a wide application range with chemical accuracy.

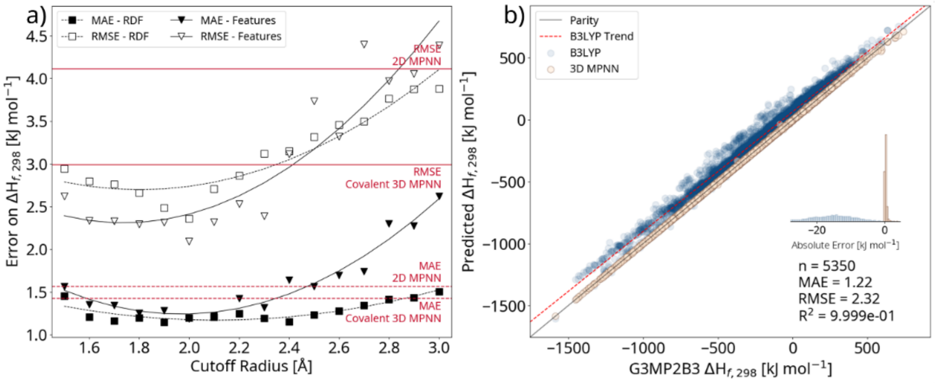

Figure 1. Performance of ML model on enthalpy predictions. (a) Effect of using geometric information on the predictive performance. (b) Parity plot and error distribution with a geometric MPNN that uses a cutoff radius of 2.1 Å compared to B3LYP calculations.

4. Conclusions

A versatile ML method for physicochemical property estimation that can effectively handle 2D and 3D molecular information is developed and validated. It was demonstrated for thermochemistry that using geometric information significantly improved the predictive performance. Using this Δ-ML method, the enthalpy of formation can be predicted with chemical accuracy for an extensive range of industrially relevant molecules at the cost of DFT calculations.

References

[1] G.M. Kontogeorgis, R. Dohrn, I.G. Economou, J.-C. de Hemptinne, A. ten Kate, S. Kuitunen, M. Mooijer, L. Fele Žilnik, and V. Vesovic, Ind. Eng. Chem. Res. 60 (2021) 4987−5013

[2] M.R. Dobbelaere, P.P. Plehiers, R. Van de Vijver, C.V. Stevens, K.M. Van Geem, Eng. 7 (2021) 1201–1211

[3] P. Bollini, M. Diwan, P. Gautam, R.L. Hartman, D.A. Hickman, M. Johnson, M. Kawase, M. Neurock, G.S. Patience, A. Stottlemyer, D.G. Vlachos, B. Wilhite, ACS Eng. Au 3 (2023) 364-390

[4] P.P Plehiers, I. Lengyel, D.H. West, C.V. Stevens, K.M. Van Geem, Chem. Eng. J. 426 (2021) 131304

Keywords

machine learning; property estimation; thermochemistry.