Deep Learning for Properties prediction based on 3D properties with

Aymeric Hernandez Nay Chi Hnin Htut Yuto Tsuruta

May 2, 2025

Abstract

The accurate prediction of molecular properties from structural information is essential for accelerating discovery in chemistry and materials science. While Density Functional Theory (DFT) provides reliable quantum mechanical predictions, its high computational cost limits its applicability in large-scale screening. In this study, we develop a neural network-based regression model to predict molecular properties—specifically total energy—directly from three-dimensional atomic coordinates. Using the DFT_all.npz dataset available from Zenodo, which contains a variety of DFT-computed properties for small organic molecules, we train the model in a supervised manner to learn the structure–property relationship. Our results demonstrate that neural networks can effectively approximate DFT-level accuracy while significantly reducing computation time. This work highlights the potential of machine learning as a scalable alternative to traditional quantum chemical simulations, enabling faster exploration of chemical space for materials and drug design.

1 Introduction

Predicting molecular properties directly from structural information is a fundamental task in computational chemistry and materials science. Traditionally, this is achieved through quantum mechanical methods such as Density Functional Theory (DFT), which provide accurate predictions but are computationally expensive and limited in scalability. As the demand grows for rapid property evaluation in high-throughput screening and molecular design, data-driven alternatives have gained significant attention.

Recent advances in machine learning, particularly neural networks, have opened new pathways for modeling the complex relationship between a molecule's structure and its physicochemical properties. These models can learn from large datasets of precomputed molecular structures and properties to make fast, accurate predictions without relying on costly simulations.

In this project, we focus on **predicting molecular properties from 3D molecular structures** using supervised learning with neural networks. We use the **QM24** dataset, derived from DFT calculations and available through [Zenodo](https://zenodo.org/records/11164951), which contains atomic coordinates and quantum-level properties for a variety of small organic molecules.

Our goal is to **train a neural network to accurately predict key molecular properties—such as total energy—from 3D atomic coordinates**, thereby capturing the structure–property relationship encoded in quantum mechanical simulations. This approach aims to demonstrate how machine learning models can serve as efficient surrogates for DFT, accelerating materials discovery and molecular design through predictive modeling.

2 Data Science Method

We decided to realize a Property prediction of the elements from their 3D molecular structure. We will use Supervising training on Neural Network. To do so we will use the 3D properties of the molecules as training inputs, then we will train the network with the use of the desired Property (Atomization Energy for example) as a label.

The input data should need no pretreatment. The output label should be a continuous value defining the property of the element.

The dataset is composed of 784875 element which is a quite huge amount of data (to compare, MNIST which is a basic digit recognition dataset contains 70000 elements) so the split between training subset and test subset should be relevant.

The main limit of this method is for each property we would like to predict, we would have to entirely redo the training with a different label.

3 Exploratory Data Analysis

3.1 Explanation of our data set

Variable name	Python representation format
compounds	array
atoms	array
freqs	array
vibmodes	array
zpves	float64
U0	float64
U298	float64
H	float64
S	float64
G	float64
Cv	float64
Cp	float64
coordinates	array
Vesp	array
Qmulliken	array
dipole	array
quadrupole	array
octupole	array
hexadecapole	array
rots	array
gap	float64
Eee	float64
Exc	float64
Edisp	float64
Etot	float64
Eatomization	float64

variables	units	discreption
compounds		Stoichiometric formulas of the molecules
atoms		Atomic numbers in the molecule
freqs	cm^{-1}	Vibrational frequencies obtained from harmonic frequency analysis.
vibmodes	$\AA$	Normal modes of vibration represented as displacement vectors.
U0	Ha	Internal energy at 0 K
U298	Ha	Internal energy at 298 K
H	Ha	Enthalpy
S		Entropy
G	Ha	Gibbs free energy
Cv		Heat capacity at constant volume
Cp		Heat capacity at constant pressure
coordinates		coordinates (XYZ) of atoms in the molecule.
Vesp		Electrostatic potential
Qmulliken		Mulliken atomic charges
dipole	a.u.	Dipole moment
quadrupole	a.u.	Quadrupole moment
octupole	a.u.	Octupole moment
hexadecapole	a.u.	Hexadecapole moment
rots	MHz	Rotational constants of the molecule.
gaps	Ha	HOMO-LUMO energy gap
Eee	Ha	Electron-electron repulsion energy
Exc	Ha	Exchange-correlation energy
Edisp	Ha	Dispersion correction energy
Etot	Ha	Total electronic energy
Eatomization	Ha	Atomization energy