BAND NN

A Deep Learning Framework for Energy Prediction and Geometry Optimization of Organic Small Molecules

Journal of Computational Chemistry, 2019

Team Number - 20

Archit Jain (2019101053)

Pulkit Gupta(2019101078)

Introduction

- Quantum Mechanical(QM) and Density Functional Theory(DFT) - Methods for calculating molecular energies and physiochemical properties.
- But these are computationally expensive and impractical for large system.
- Use Molecular Mechanics(MM) force field methods which are computationally manageable.

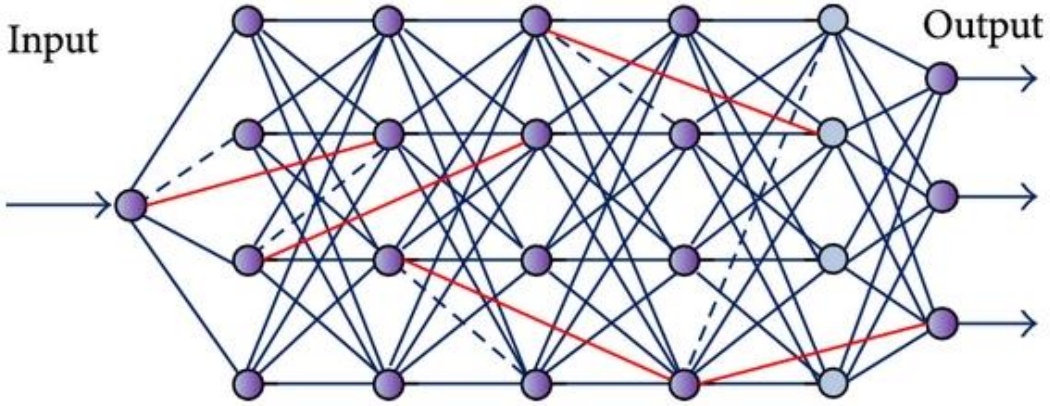
Supervised ML Algorithm

- Supervised Machine Learning Algorithm is needed – predicts energies that are of DFT level but comparable with MM with respect to computational cost.
- Accurate description of a molecule as vector is required as input to supervised algorithm.
- Different ways of generating feature vectors
  - Using local environment of each atom as input.
  - Using smooth overlap of atomic positions(SOAP).
  - Using nuclear charges(Z) and matrix of inter-atomic distances.
  - Using BAND= Bonds(B), Angle(A), Non-Bonded(N) interactions, Dihedrals(D).

Theory

Neural Network

- Used Feed-forward fully connected deep neural networks.
- Neural Networks
  - an input layer, multiple hidden layers, and an output layer
  - Input must be of fixed length



BAND Molecular Descriptor

- **Properties of a descriptor** - rotational and translational invariance, invariance with respect to the permutation of atoms, provide a unique description of the atomic positions.
- List of Bonded pairs
- List of Nonbonded pairs
- List of bond angle - identified as two consecutive bonds
- List of Dihedral angle - identified as three consecutive bonds

BAND Molecular Descriptor

- Each atom is represented by an eight-dimensional feature vector
- First four dimensions representing the atom name

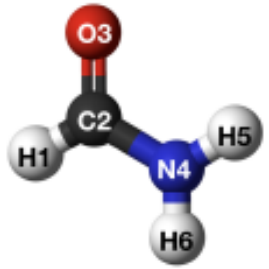
Atom identifier

H	C	O	N
1 0 0 0	0 1 0 0	0 0 1 0	0 0 0 1

- Second four dimensions representing the number of connected atom type.

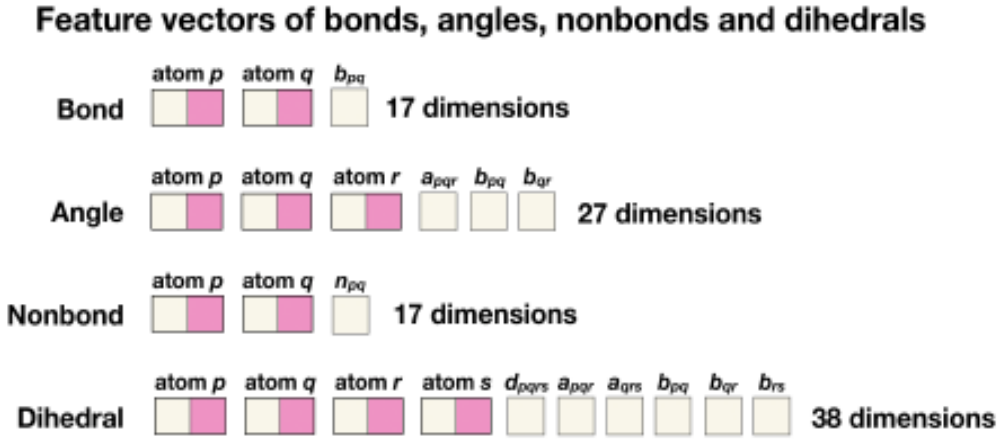
Atom identifier and atom typing

	Atom name	Atom type
H1	1 0 0 0	0 1 0 0
N4	0 0 0 1	2 1 0 0



BAND Molecular Descriptor

- **Each bond is represented by a 17-dimensional vector**
  two atoms (eight-dimensions each) + bond length
- **Each angle represented by a 27-dimensional vector**
  it is the combination of the three atomic representations (24) followed by the bond angle and two bond lengths
- **Each dihedral angle represented by 38-dimensional vector**
  four atomic representations followed by the dihedral angle, two angles and three bond lengths
- **Each Nonbond pair represented by a 17-dimensional vector**
  two atoms (eight-dimensions each) + internuclear distance.



Classical Force Fields Equation

$$E_{total} = E_{bonded} + E_{nonbonded}$$

$$E_{bonded} = E_{bonds} + E_{angles} + E_{dihedrals}$$

$$E_{bonds} = \sum_{bonds} k_b(b - b_0)^2$$

k_b= force constant

b = bond length

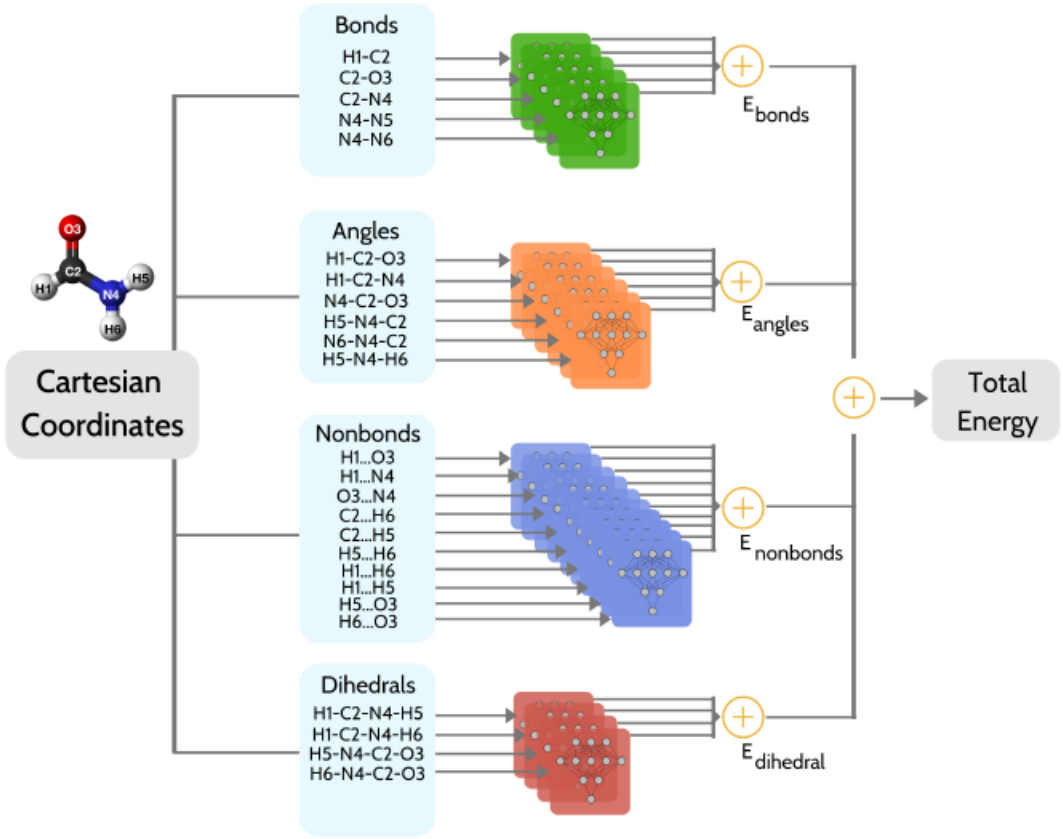
b_0 =equilibrium bond length

Model

- Atomization energy is represented as

$$E = \sum_{bonds} E_B + \sum_{angles} E_A + \sum_{nonbonds} E_N + \sum_{dihedrals} E_D$$

- Each term if estimated by a feed-forward fully connected network
- Each term have equal contribution
- Model is invariant to number of atoms as the final energy is only expressed as sum of the individual contribution from each term



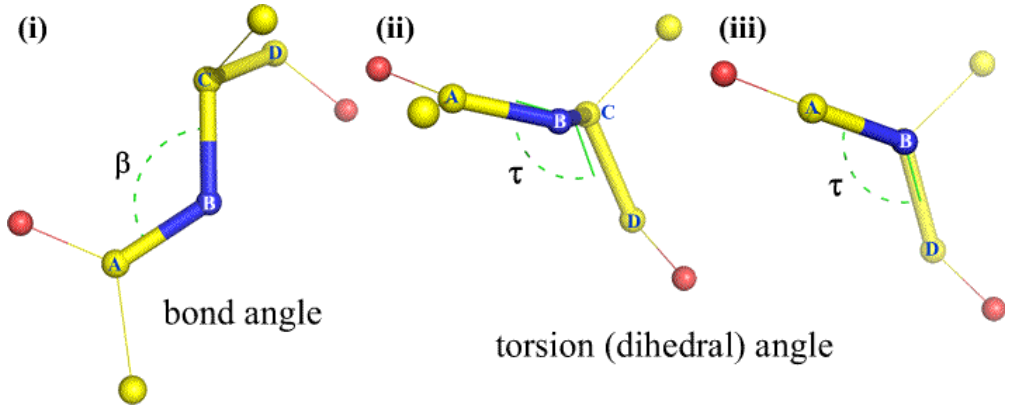
Methodology

Data Selection

- 57,462 minimum energy molecular structures was used from a non-equilibrium dataset ANI-1
- Molecules having up to 8 heavy atoms containing only H, C, N and O were picked
- All the equilibrium configurations along with each of their non-equilibrium structures whose relative energies with respect to the corresponding minimum energy structure are less than 30 kcal/mol were used for this study.

Data Preprocessing

- Used RDKit to generate list of all bonds based on the atomic coordinates
- All possible 1,3 neighbours that are connected to 2 were taken as angles
- All 1,4 neighbours where 2 and 3 are connected were taken as dihedrals.
- All pairs except 1,2 whose distances are less than 6 A in the equilibrium were taken as non-bonded pair.



Training

- Data Split
  - train-test-validation split in the ratio of 80-10-10
  - training set – 6.1 billion data points
  - test and validation set– 760000 data points
- Each network has an input layer, three hidden layers for each type and an output layer, a one-dimensional vector that predicts the energy contribution from that network. All layers were activated using ReLU activation function
- Objective minimization function – mean squared error between predicted and actual atomization energies

Type of Network	Input dimensions	Hidden Layer Dimensions
Bonds	17	128-256-128
Angles	27	128-350-128
Non-bonds	17	128-256-128
Dihedrals	38	128-512-128

Geometry Optimization

- Geometry optimization finds the least energy structure of molecule given an approximate structure over band topology.
- Used Nelder-Mead's Method that uses direct search method for nonlinear optimization.
- Initialized by construction of a simplex by randomly sampling points on the target surface and then propagates through generation of simplices by repeatedly replacing the worst point from better one.
- Terminates either when the working simplex is sufficiently small or when the differences in the function values on the vertices of the simplex is less than a threshold.

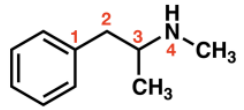
Results

Accuracy

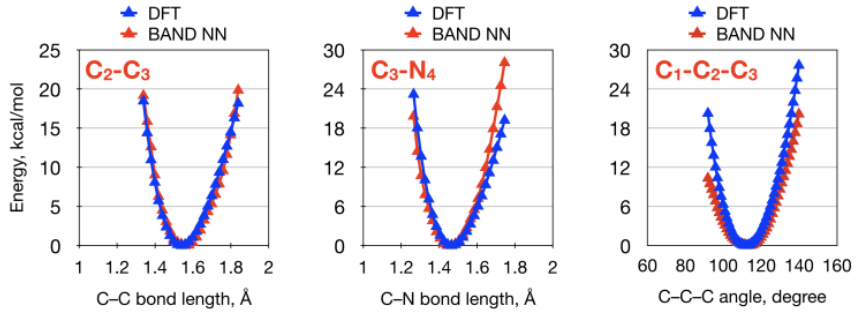
- Mean Absolute error = 1.45 kcal/mol on test set.
- Picked all structures whose relative energies are under 30 kcal/mol with respect to their corresponding minimum.
- 75% of structures in the test dataset have predicted atomization energy within 2 kcal/mol.
- High energy structures with 10 heavy atoms were calculated to test transferability of the BAND NN model to molecules with greater number of atoms than the training dataset.
- Resulted in 1500 structures and the mean absolute error of the atomization energies predicted using BAND NN for this set was found to be 2.1 kcal/mol indicating transferability.

Structural and Geometric Isomers

- Accuracy of the proposed model in satisfactorily predicting the relative energies of structural and geometric isomers.
- Quantitative agreement between the DFT and BAND NN methods is observed.
- BAND NN outperforms the semiempirical quantum mechanical AM1 method.
- Molecular size invariant ML based methods are capable of accurate modelling of molecular systems at the computational expense less than that of DFT.



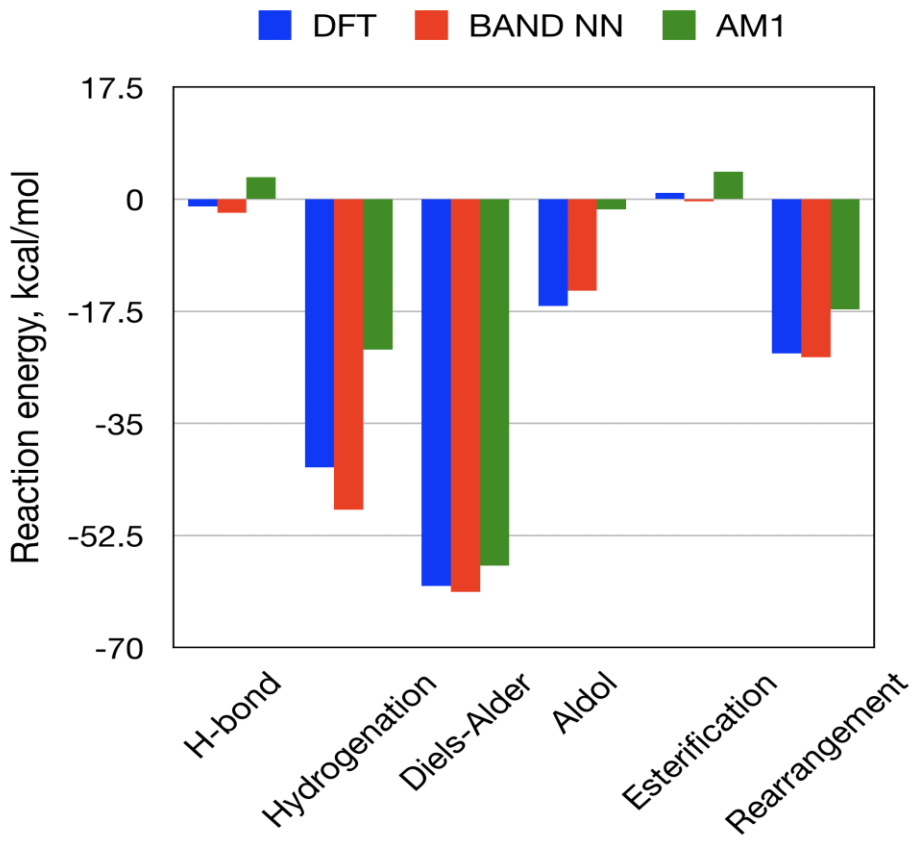
Potential Energy Surfaces



- BAND NN model is capable of prediction atomization energies of small organic molecules
- Potential energy scans with respect to bonds and angles were performed on molecules that are significantly larger than those in the training set.
- Comparing BAND NN to DFT over different bonds
  - For C-C and C-N bonds the positions of the minima are predicted accurately .
  - For C-C-C angle indicates very good agreement
  - curves maintain a smooth curvature
- Mean absolute error is only about 0.6 kcal/mol for the energies of different conformers and transition states.

Reaction Energies

- BAND NN model is compared with DFT and AM1 methods on reaction energy in different reactions
- Largest difference in energies are observed in hydrogenation reaction since absence of H2 in training dataset.
- BAND NN outperforms AM1 and was comparable to DFT method



Impact of Bond angle and Dihedral terms

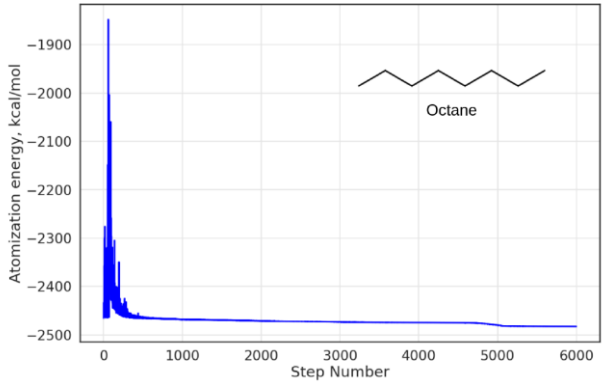
- In previous studies all the angles and dihedrals in these molecule are in their equilibrium values and hence the variances of the angles and dihedrals in the dataset are not large.
- BAND NN is well suited for handling non-equilibrium structures compared to those that include only 2-body terms.

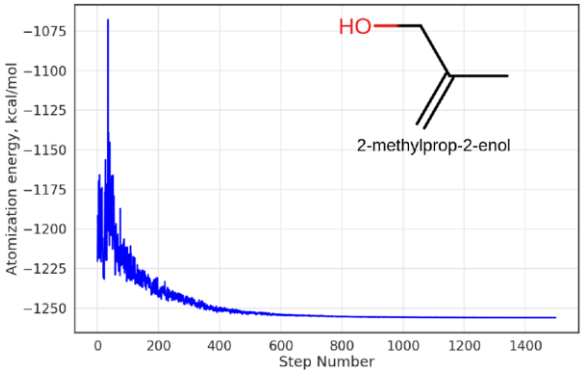
Model	Mean Absolute Error in kcal/mol
BAND NN	1.45
BAN NN	2.4
BN NN	2.7

Geometry Optimization

- Shortfalls of ML models to predict atomization energies
  - Cannot be applied on molecules larger than present in training set.
  - Cannot be applied to structures that are not in their minima on the potential energy surface and they have not been used for geometry optimizations.

- BAND NN model has been trained on high energy structures with explicit topology of the molecule.
- Starting from a reasonable guess structure of octane and 2-methylprop-2-enol, geometry optimization was performed.
- The energies of the two molecules gradually decrease with respect to the optimization step and reaches convergence.
- The optimizer converged the molecules to structures whose energies are significantly lower than those of the initial structure.





Conclusion

BAND NN model proposed not only predicts the atomization energy for equilibrium and off-equilibrium structures but also can be used to perform geometry optimization.

THANK YOU