



Predicting HFE using Deep Learning

Keshavan Seshadri (20171051)

Nikhil Jakhar (20171186)



Introduction

Hydration Free Energy - The amount of energy released when one mole of molecules is dissolved in water. How soluble a substance is in water?

Current Methods Used are Experiment based and Calculations Based.

Calculation based Methods are either:

- Quantum Mechanics Based (more accurate, but slow)
- Molecular Dynamics Based (less accurate, but fast)
- Hybrid (QM/MM)

Many calculation based methods are efficient(fast), but are not accurate enough.

Therefore, there is a need for accurate and efficient method. This will help in predicting the hydrophobicity and help in drugs research.



DATASETS

FreeSolv Dataset

```
iupac,smiles,expt,calc
"4-methoxy-N,N-dimethyl-benzamide",CN(C)C(=O)c1ccc(cc1)OC,-11.01,-9.625
methanesulfonyl chloride,CS(=O)(=O)Cl,-4.87,-6.219
3-methylbut-1-ene,CC(C)C=C,1.83,2.452
2-ethylpyrazine,CCc1cnccn1,-5.45,-5.809
```

Datasets usually contain from 10^3 to 10^7 molecules. But there are about 10^{60} drug molecules.

=> HFEs for most molecules are unknown.

The datasets contain a smile string and the experimental and calculated values of HFE.

FreeSolv Size of the Dataset - 643 small molecules

QM9 - 134k molecules

eMolecules - 20 million molecules with around 34k molecules with only C,O and N



Related Work in the Area

Kernel based Machine Learning Method to predict HFE: Clemens Rauer, Tristan Bereau, Hydration free energies from kernel-based machine learning: Compound-database bias, July 2020, The Journal of Chemical Physics 153(1):014101, DOI: 10.1063/5.0012230

Predicts solvation energy(any solvent) uses simple RNN based model for solute and solvent separately: Delfos: deep learning model for prediction of solvation free energies in generic organic solvents†, Hyuntae Lim*and Yong Joon Jung, Chem. Sci.,2019,10,8306

Increase Latent Space for better QSAR predictions: Improving Chemical Autoencoder Latent Space and Molecular De novo Generation Diversity with Heteroencoders, Esben Jannik Bjerrum, Boris Sattarov, Biomolecules 2018, 8(4), 131

Input Preparation

The dataset contains a SMILE string which represents the molecule.

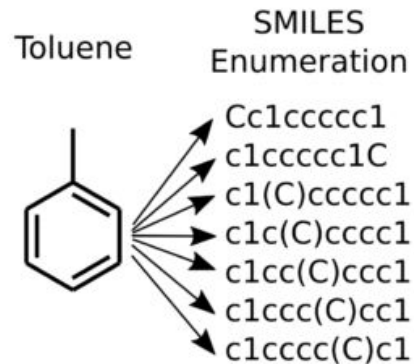
From the Dataset, we include atoms that are common in most molecules:

* Atoms Included -> C, H, O and N

Taking Inspiration from NLP(encoding - mol2vec) and CV(data augmentation),

SMILE String(one) -> Enumerate(many) -> Feature Matrix (Morgan Algorithm)

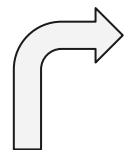
We enumerate the smile strings. This is used to slightly broaden the latent space, which in-turn helps in better predictions.

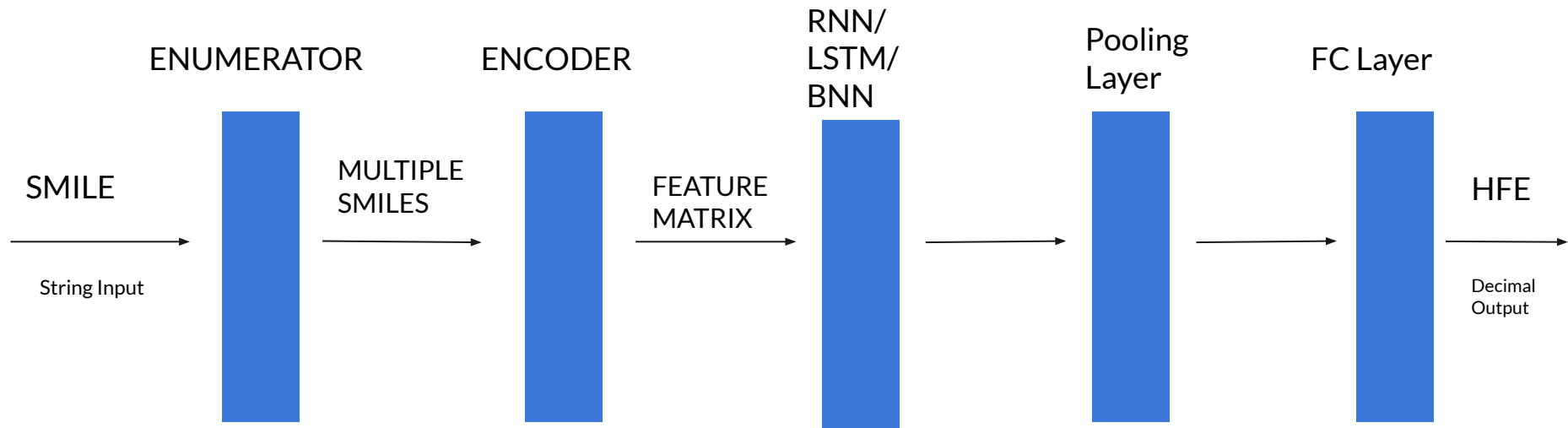


SMILES enumeration of toluene, the topmost is the canonical SMILES



MODEL

 Comparison





NOVELTY

- None of the current approaches use enumeration of SMILES to predict HFE which opens up new possibilities for prediction.
- This is expected to improve the predictions significantly when compared to results without enumeration.
- Current methods have trained only on a subset of a large dataset or only few molecules(10^3).
- Though LSTMs have been used in other predictions, not many have used it specifically for HFE prediction.
- The method(using SMILES) is also simpler compared to other methods that require the relative atom coordinates.



Aim

- Provide a simple and accurate method to predict HFE for new molecules.
- We will compare the results from the RNN/LSTM model with Bayesian Neural Network, 3D CNN and 2D CNN based models.
- Compare predictions from different NN model w.r.t experimental and calculated values(Accuracy, F1-score, ROC-AUC).
- Compare w.r.t previous studies(Kernel machine learning, Delfos).
- Compare results with non-enumerated input.



Timeline and Feasibility

It is possible to finish the work in 1.5 month time.

Timeline:

Input Preparation(Dataset preparation, Augmentation, Encoding) => 1 week.

Comparing different models => 3 weeks

Cumulate Results and Analysis => 2 week.