POLITECNICO
MILANO 1863

Data Science in Chemical Engineering

Prediction of thermodynamic properties based on functional groups

In 1958, Benson and Buss [1] introduced the foundational concept that thermochemical properties, such as enthalpy of formation and heat capacity, can be effectively estimated by summing the contributions of a molecule's constituent functional groups. These groups have proven to be relatively transferable between different molecules, meaning, for example, that a methyl group's contribution remains consistent whether it is part of methane or propane.

Today, extensive thermodynamic data exists for numerous organic molecules. The QM9 dataset (http://quantum-machine.org/datasets/) comprises geometric, energetic, electronic, and thermodynamic properties for approximately 134,000 stable, small organic molecules composed of carbon (C), hydrogen (H), oxygen (O), nitrogen (N), and fluorine (F). Each entry provides molecular structure details and includes 15 calculated molecular properties.

Objective: Your task is to develop two predictive models—a classical machine learning (ML) model and a neural network (NN) model—to estimate molecular properties based on identified functional groups. Functional groups for each molecule can be extracted using the RDKit library [2–4] (https://app.readthedocs.org/projects/rdkit/downloads/pdf/latest) by utilizing the molecule's SMILES representation.

Suggested workflow:

- **Dataset preprocessing**:
  - Download a part of the dataset molecules (≈ 10k molecules are sufficient)
  - Understand how to parse a .xyz file and extract the SMILE string and the target properties. Prediction of Rotational Constant A, B and C can be skipped.
    - **Major Hint**: https://figshare.com/articles/dataset/Readme_file_Data_description_for_Quantum_chemistry_structures_and_properties_of_134_kilo_molecules_/1057641?backTo=%2Fcollections%2F_%2F978904&file=3195392
  - Create a matrix/object/dataframe to store the data for all the molecules
  - Using rdkit.Chem.FragmentCatalog, calculate the number of functional groups.
    - **Major Hint**: https://www.rdkit.org/docs/GettingStartedInPython.html#molecular-fragments,
    - **Major Hint 2**: use "fparams = FragmentCatalog.FragCatParams(1, 1, fName)"
  - Possible suggestion: also count the number of atoms for each element in the molecule using RDkit.
- **Dataset visualization**
- **Training**:
  - The training features are the type and the number of functional groups (and possibly the number and type of atoms in a molecule).
- **Validation:**
  - As per seen in class. You also are encouraged to present results in an original and creative way.

- **Discussion:**
    - Visualize the dataset in an appropriate way.
    - Compare the performance of the ML model against the NN model using appropriate performance metrics and validation methods.
    - Analyze the impact of model complexity, training duration, and potential overfitting.

Submission:

You are required to submit the following:

- A detailed report containing your methodology, analysis of model performances, key findings, and insights from your work.
- Clearly commented code files demonstrating your data preprocessing steps and implementations of both ML and NN models.

References

[1] S.W. Benson, J.H. Buss, Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties, *The Journal of Chemical Physics*. 29 (1958) 546–572.

[2] G. Landrum, Rdkit documentation, *Release*. 1 (2013) 4.

[3] G. Landrum, others, RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling, *Greg Landrum*. 8 (2013) 5281.

[4] M. Lovrić, J.M. Molero, R. Kern, PySpark and RDKit: moving towards big data in cheminformatics, *Molecular Informatics*. 38 (2019) 1800082.