Creating Databases for Training Neural Networks to Identify Functional Groups from NMR, IR, and Raman Spectroscopic Data

Kaden M. Hubler[1], James P. Sturgill[1], Max Depperschmidt[1], Kyle V. Camarda[1]
1. Chemical and Petroleum Engineering Department, University of Kansas, Lawrence, KS.

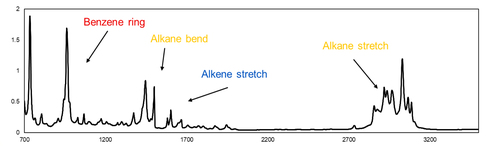
camarda Research Group

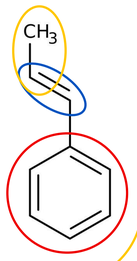
Introduction

The interpretation of spectroscopic data from chemical reaction experiments is a challenging task, usually performed by expert chemists. The goal of this work is to train artificial neural networks (ANNs) to interpret spectra, allowing for automated experimental investigation. ANNs require large datasets for training, which in this context means creating databases of chemical compounds and numerically-encoded spectral information. My focus has been on the creation and manipulation of datasets for data from three types of spectroscopic methods: Nuclear Magnetic Resonance (NMR), Infrared (IR), and Raman spectroscopy.

Spectroscopy

- NMR
  - Highest resolution, expensive
- IR
  - Medium resolution, very common
- Raman
  - Lowest resolution, cheapest





Data Sources[1-5]

hmdb The Human Metabolome Database

Ames Research Center

PURDUE UNIVERSITY

IRPA KIK

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

Common Data types

- JCAMP-DX
  - Standard data type created by IUPAC
  - Difficult to parse and each file can have small variations
  - Computationally expensive
- XML
  - Contains roots and branches
  - Easily parsed, each file is standard
  - Computationally inexpensive
- HDF5
  - Designed for large sets of data
  - Easily parsed, standard formats
  - Computationally inexpensive

Python libraries:

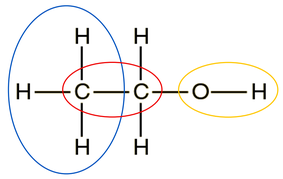
jcamp

lxml

h5py

Labels

- Compounds are stored as SMILES strings and can be obtained from CAS numbers.
- Functional groups were extracted from SMILES strings using SMARTS sub-strings. Each point was then labeled with binary variables corresponding to functional group presence.
- Ethanol example:



64-17-5 —PubChemPy→ CCO —RDKit→ Alcohol, Alkane, Methyl

Further Data Manipulation

- Normalization
  - Different sources have different scales, so normalization is necessary
  - Divide each point by the maximum to give a percentage relative to maximum instead.
  - Interpolate to achieve same resolution
- Other Important Labels
  - Solvent
  - Field strength
  - Concentration
  - Resolution
- Trimming
  - Determine which functional groups do not appear

NumPy

Data Output and Improvements

- Output
  - .csv files from DataFrames
  - Very large files, this is an area for future concern
- Improvements
  - Querying PubChem to convert CAS to SMILES is computationally expensive
  - Solution: make a dictionary of CAS to SMILES to streamline CAS conversior

Pandas

PubChem

Results

- NMR
  - >400,000 predicted points
  - Achieving ~99% overall testing accuracy
  - Lower on some groups like carboxylic acid (~95%)
- IR
  - ~375,000 predicted points, ~2000 experimental points
  - Achieving ~95% testing accuracy
- Raman
  - ~250 experimental points
  - In-progress
  - All solids

Keras

Future Work

- Find more Raman data!
- Make predicted points more realistic:
  - Add noise
  - Shift peaks left or right slightly
  - Lower the resolution
  - Add solvent fingerprints
- Advanced data manipulation:
  - Data augmentation
- Model improvements:
  - Ensemble models
  - Combine all three models, have them talk to each other, output result

SciPy

Acknowledgements/References

- This work is funded by the Department of Energy's Kansas City National Security Campus, operated by Honeywell Federal Manufacturing & Technologies, LLC. under contract number DE-NA0002839.
- The presenter would also like to acknowledge Dr. Alan Allgeier for his assistance on the project

1. Wishart, D.S., Tzur, D., Knox, C., et. al., 2022. HMDB 5.0: the Human Metabolome Database for 2022. 50 (D1), p. D622-D631.
2. Mattioda, A.L., Gavilan, L., Ricketts, C.L. Najeeb, P.K., Ricca, A., Boersma, C., 2024. The NASA Raman Spectroscopic Database: Ramdb version 1.00. 208, 115769.
3. Fremout, W. Saverwyns, S., 2012. Identification of synthetic organic pigments: the role of a comprehensive digital Raman spectra library. 43, p. 1536-1544.
4. Jin, T., Zhao, Q, Schofield, A.B., Savoie, B.M., 2024. Deductive machine learning models for product identification. 15 (3), p. 11995-125.
5. Wallace, W.E., 2024. Infrared Spectra. Nist Mass Spectrometry Data Center.