Hierarchical Clustering Analysis of Spectral Fingerprints for Cheminformatics

Dylan Sam
Department of Computer Science
Brown University
Providence, RI
dylan_sam@brown.edu

Brenda M. Rubenstein
Department of Chemistry
Brown University
Providence, RI
brenda_rubenstein@brown.edu

Abstract

Many molecular computing and information processing applications rely on the identification of individual molecules within complex, potentially reacting mixtures. While identifying molecules from within such complex mixtures can in certain situations be performed with a single spectroscopic tool, such as mass spectrometry or nuclear magnetic resonance (NMR), there are many situations in which this task requires multimodal measurements using several different spectroscopic tools. To guide chemists in their selection of which measurement tools (modalities) to use to interrogate such solutions, we leverage deep learning methods and electronic structure calculations to generate mass, NMR, and infrared (IR) spectra for thousands of organic compounds from the GDB-13 database. We convert this information into spectral fingerprints that serve as addresses for where molecules lie in their high-dimensional spectral space. Using hierarchical clustering, we compare our multimodal spectral fingerprints to fingerprints generated using IR or NMR data alone, as well as to extended connectivity fingerprints to highlight the similarities and differences between our spectral space and other notions of chemical space.

1 Introduction

The need to accurately identify a multitude of species within a single solution routinely arises in the fields of molecular computing and metabolomics. Molecular computing looks to store information in and compute with vast collections of different molecules in a single solution [1, 2, 3, 4, 5], whereas metabolomics seeks to understand the relationship between the metabolites within an organism and its phenotype [6, 7, 8]. These fields require understanding the makeup of complicated solutions and the reactions that may be occurring within them. Thus, spectroscopies and spectrometries that can unambiguously and efficiently identify all of the species in these solutions are crucial.

Chemists have developed an arsenal of spectroscopic tools including infrared (IR), Raman, nuclear magentic resonance (NMR), and electron paramagnetic resonance spectroscopies, that can each identify different molecular features. To date, the selection of which tool to use to interrogate a solution has overwhelmingly been guided by personal experience, instrument availability and compatibility, and cost. However, there is little reason to believe that such *ad hoc* selections are even remotely optimal, leaving room for the development of more methodical selection techniques.

To determine which combinations of spectroscopies should be used to decipher a complex molecular mixture, one needs to understand how far apart species of interest are from one another in spectral space, the multidmensional space formed from all possible spectroscopic measurements. The closer two species are to one another in spectral space, the more difficult it is to distinguish them from one another. Moreover, two species at a fixed distance in spectral space may appear closer together or farther apart depending upon which spectroscopies are used to interrogate them (or, in other words,

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

along which axes their distances are projected). Resolving the contours of spectral space would not only inform researchers of whether molecular systems can be spectrally resolved using current tools but also of which tools would be best suited for doing so.

While there exist other methods that learn high dimensional molecular representations in an unsupervised fashion [9], little work has focused on characterizing the spectral space of a given set of molecules [10, 11, 12]. This paucity of work may in large part be attributed to the historical difficulty associated with assembling the spectral properties of a large number of species, either experimentally or computationally. However, the door to characterizing spectral space has recently been opened by automation tools that enable numerous spectra to be taken in relatively short periods of time and the development of machine learning techniques that can accurately and efficiently predict IR, Raman, NMR, and other spectra [13, 14, 15].

In this work, we leverage electronic structure simulations and machine learning algorithms to assemble the IR, NMR, and mass spectra of thousands of small molecules and use these spectra to assemble spectral fingerprints for each of these species. We then employ hierarchical clustering techniques to shed light on how our spectral fingerprints differ from popular extended connectivity fingerprints and how different spectroscopies can be combined to most effectively distinguish similar small molecule species from one another.

2 Methods

2.1 Molecular Datasets

We construct our spectral fingerprints from three subsets of molecules from the GDB-13 database [16], the largest publicly-available database of chemical species. Our molecules are drawn from the GDB-13 ABCDEFG subset, which contains roughly 13 million molecules. We create two of our three datasets (MW150 and MW180) by selecting molecules with molecular weights of 150.043 and 180.137 Daltons. These datasets enable the analysis of variation solely in the IR and NMR dimensions of spectral space. We create a third dataset (Mixed MW) without any restrictions on molecular weight, allowing us to consider molecules a larger range of functional groups and ring structures. These three datasets are thus representative of a wide range of potential pharmaceutical targets in organic synthesis and small molecules used in chemical information processing, and have sizes of 1290, 10914, and 873, respectively.

The parent molecular masses are computed using RDKit [17], which is a computational chemistry package that can compute exact molecular weights with up to 11 decimal places of accuracy. We simplify our representation of these spectra to their single parent peak masses in this work.

2.2 H^1-NMR Fingerprint Generation

In order to supplement our spectral fingerprints with proton NMR information, we use deep learning methods to predict the NMR spectra of the molecules in our datasets. There are no public experimental databases that contain the NMR spectra of even a small fraction of our molecules, so we use the SPINUS algorithm [18, 19, 20] to generate predictions for our datasets. The SPINUS algorithm leverages an ensemble of 75 independently trained deep neural networks to predict chemical shifts and J-J coupling constants.

The H^1-NMR spectra are converted into binary vector representations, in which each vector bin represents a range of ppm values. The $[0, 1]$ value in a given bin quantifies if the molecule has a signal or multiple signals within that ppm range. Our NMR vectors have fixed bin widths of 0.01 ppm, which results in only a few instances of overlapping signals; bins that are too small produce vectors that are too sparse.

2.3 Infrared Fingerprint Generation

To add the IR dimensions of spectral space, theoretical predictions of normal mode frequencies are computed with the Molpro program [21]. Ground state geometries are optimized at the spin-restricted Kohn-Sham level using the B3LYP hybrid functional and the 6-31G* basis. Raw frequencies are scaled using a precomputed vibrational scaling factor for B3LYP/6-31G* (0.960 ± 0.022) [22]. The IR fingerprint vectors are also constructed using binning, with peak intensities below 5% not included.

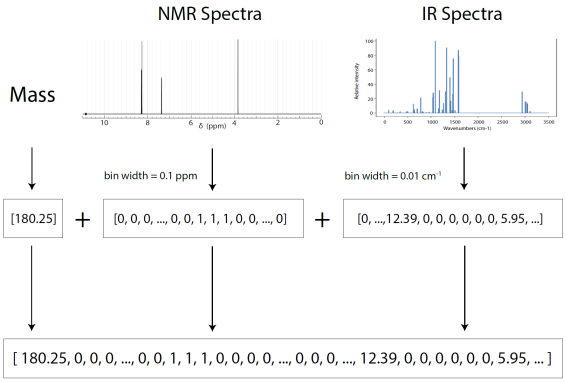

Figure 1: Creation of the spectral fingerprint of the molecule COc1nccc2ocnc12.

We use a bin width of 0.01 cm^{-1} since, as with the NMR vectors, denser vectors produce better results. The value at the index of the vector corresponding to a given bin represents the highest intensity signal in that bin's range of wavenumbers.

After creating vectors representing the H^1-NMR and IR spectra of a molecule, we concatenate them together to construct a multimodal spectral fingerprint. This process is illustrated in Figure 1 for the molecule represented by the SMILES string COc1nccc2ocnc12. We use these combined fingerprints and the individual IR and NMR vectors for each data set in our clustering analysis.

2.4 Hierarchical Clustering and Dendrogram Visualization

Hierarchical clustering is performed on our three datasets using Morgan (ECFP6, 2,048 bits) [23], IR, NMR, and IR+NMR fingerprints. We use Ward's method [24] to perform agglomerative clustering based upon the Euclidean distance between the fingerprints. To visualize our spectral space dendrograms, we use Gephi, which is a graph and network visualization platform [25]. Again, we perform hierarchical clustering on our different fingerprints to generate a set of nodes and edges using the UPGMA method. The resulting nodes and edges are input into the Gephi software to visualize our dendrograms. Within the Gephi software, we use the MultiGravity ForceAtlas2 [26] and Yifan Hu Layout [27] algorithms to orient the nodes and clusters in the visualizations.

3 Results

We compare the hierarchical clustering and dendrograms of our spectral fingerprints and ECFP6 fingerprints. We only include results on the MW150 and Mixed MW datasets for brevity.

3.1 Hierarchical Clustering Panels

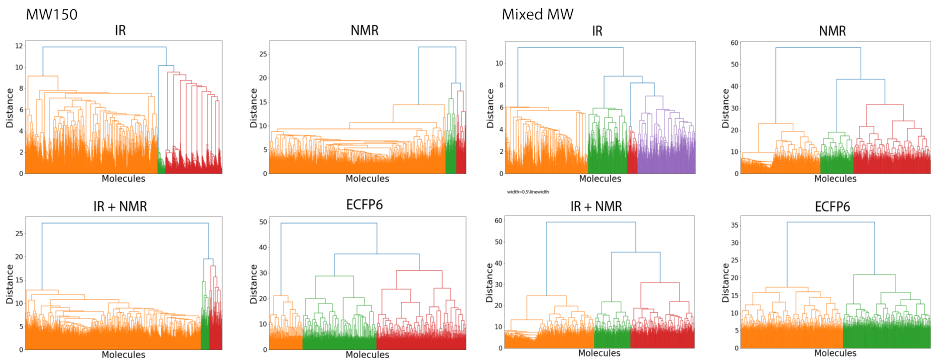

Figure 2: Clustering panels for the (left) MW150 dataset and (right) MixedMass dataset for the IR, NMR, IR + NMR, and ECFP6 fingerprints. Cluster color assignments are determined by a distance threshold that is 0.7 times the maximum distance between any two molecules.

In Figure 2, we present the results of our clustering analysis on the different fingerprints of the MW150 and Mixed MW datasets. For the MW150 panels, each fingerprint produces three primary clusters. In the IR, NMR, and IR+NMR panels, the orange cluster is by far the largest cluster, while the ECFP6 panel has clusters that are more comparable in size. This suggests that IR-based and NMR-based fingerprints differentiate MW150 dataset molecules based upon a handful of key, divergent features at large distances, but possess far less discriminatory power after these features are taken into account. The ECFP6 fingerprints are better able to discriminate features at all distances. For the Mixed MW Dataset, we see that the IR fingerprint generates four major clusters, while the NMR and IR + NMR fingerprints generate three major clusters. The ECFP6 fingerprint only generates two major equally-sized clusters. Both the IR + NMR and ECFP6 fingerprints exhibit strong discriminatory power as they create large, evenly-sized clusters. However, the fingerprints only consisting of IR and NMR spectral information lack the same ability, as they create numerous small clusters in each of their major clusters.

3.2 Dendrogram Visualizations

Figure 3 depicts our dendrogram visualizations that illustrate the clustering structure for each type of fingerprint. The ECFP6 and IR + NMR fingerprints produce more coherent and larger clusters than the IR-only and NMR-only fingerprints. Furthermore, we visualize the presence of molecules that contain a specific substructure in red; the dendrograms show that the ECFP6 and IR + NMR fingerprints cluster compounds that contain similar substructures, while the IR and NMR fingerprints group together other molecular features.

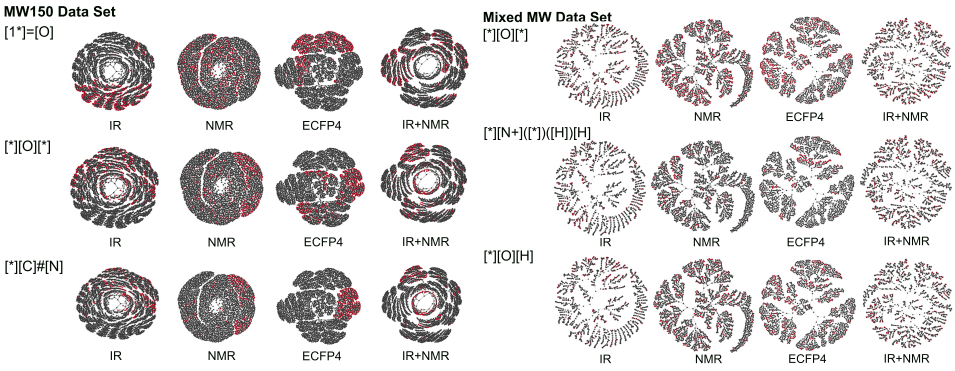

Figure 3: Dendogram visualization of the MW150 and Mixed MW clusters. Red points denote compounds that contain the substructure denoted by the SMARTS formula.

4 Conclusions

In this work, we employ modern computational chemistry and deep learning techniques to assemble and analyze the mass spectral, NMR, and IR fingerprints of thousands of organic compounds from the GDB-13 database. We perform hierarchical clustering and visualize the resulting dendrograms to illustrate the clustering properties of our spectral fingerprints compared to those of ECFP6 fingerprints. Our spectral fingerprints display discriminatory power similar to that of the topological fingerprints, and illustrate that clustering can be successfully performed along key spectral dimensions. Our novel procedures for generating large spectral datasets and our hierarchical clustering analyses facilitate an improved understanding of spectral space, which will guide chemists seeking to interrogate the complex solutions that typify molecular computation applications and metabolomics.

Acknowledgement

This research was supported by funding from the Defense Advanced Research Projects Agency (DARPA W911NF-18-2-0031). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

[1] Eamonn Kennedy, Christopher E. Arcadia, Joseph Geiser, Peter M. Weber, Christopher Rose, Brenda M. Rubenstein, and Jacob K. Rosenstein. Encoding information in synthetic metabolomes. *PLOS ONE*, 14(7):1–12, 07 2019.

[2] Jacob K. Rosenstein, Christopher Rose, Sherief Reda, Peter M. Weber, Eunsuk Kim, Jason Sello, Joseph Geiser, Eamonn Kennedy, Christopher Arcadia, Amanda Dombroski, Kady Oakley, Shui Ling Chen, Hokchhay Tann, and Brenda M. Rubenstein. Principles of Information Storage in Small-Molecule Mixtures. *arXiv e-prints*, May 2019.

[3] C. E. Arcadia, H. Tann, A. Dombroski, K. Ferguson, S. L. Chen, E. Kim, C. Rose, B. M. Rubenstein, S. Reda, and J. K. Rosenstein. Parallelized linear classification with volumetric chemical perceptrons. In *2018 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–9, Nov 2018.

[4] Brian J. Cafferty, Alexei S. Ten, Michael J. Fink, Scott Morey, Daniel J. Preston, Milan Mrksich, and George M. Whitesides. Storage of information using small organic molecules. *ACS Central Science*, 5(5):911–916, 2019.

[5] Lee Organick, Siena D. Ang, Yuan-Jyue Chen, Yekhanin Sergey Lopez, Randolph, Konstantin Makarychev, Miklos Z. Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Christopher N. Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Luiz Ceze, and Karin Strauss. Random access in large-scale dna data storage. *Nature Biotechnology*, 36:242–248, 2018.

[6] Katja Dettmer, Pavel A. Aronov, and Bruce D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1):51–78, 2007.

[7] Lee D. Roberts, Amanda L. Souza, Robert E. Gerszten, and Clary B. Clish. Targeted metabolomics. *Current Protocols in Molecular Biology*, 98(1):30.2.1–30.2.24, 2012.

[8] Abdul-Hamid Emwas, Raja Roy, Ryan T. McKay, Leonardo Tenori, Edoardo Saccenti, G. A. Nagana Gowda, Daniel Raftery, Fatimah Alahmari, Lukasz Jaremko, Mariusz Jaremko, and et al. Nmr spectroscopy for metabolomics research. *Metabolites*, 9(7):123, Jun 2019.

[9] Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8466–8478. Curran Associates, Inc., 2019.

[10] E. L. Willighagen, H. M. G. W. Denissen, R. Wehrens, and L. M. C. Buydens. On the use of 1h and 13c 1d nmr spectra as qspr descriptors. *Journal of Chemical Information and Modeling*, 46(2):487–494, 2006.

[11] Romualdo Benigni, Laura Passerini, David J. Livingstone, Mark A. Johnson, and Alessandro Giuliani. Infrared spectra information and their correlation with qsar descriptors. *Journal of Chemical Information and Computer Sciences*, 39(3):558–562, 1999.

[12] Padmakar V. Khadikar, Vimukta Sharma, and R.G. Varma. Novel estimation of lipophilicity using 13c nmr chemical shifts as molecular descriptor. *Bioorganic Medicinal Chemistry Letters*, 15(2):421 – 425, 2005.

[13] Jyrki Taskinen and Jouko Yliruusi. Prediction of physicochemical properties based on neural network modelling. *Advanced Drug Delivery Reviews*, 55(9):1163 – 1183, 2003. Artificial neural network modeling for pharmaceutical research.

[14] Michael W. Lodewyk, Matthew R. Siebert, and Dean J. Tantillo. Computational prediction of 1h and 13c chemical shifts: A useful tool for natural product, mechanistic, and synthetic organic chemistry. *Chemical Reviews*, 112(3):1839–1862, 2012.

[15] Mojtaba Haghighatlari, Jie Li, Farnaz Heidar-Zadeh, Yuchen Liu, Xingyi Guan, and Teresa Head-Gordon. Learning to make chemical predictions: The interplay of feature representation, data, and machine learning methods. *Chem*, 6(7):1527 – 1542, 2020.

[16] Lorenz C. Blum and Jean-Louis Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *Journal of the American Chemical Society*, 131(25):8732–8733, 2009.

[17] Greg Landrum, Paolo Tosco, Brian Kelley, sriniker, gedeck, NadineSchneider, Riccardo Vianello, Ric, Andrew Dalke, Brian Cole, AlexanderSavelyev, Samo Turk, Matt Swain, Dan N, Alain Vaucher, Maciej Wójcikowski, Eisuke Kawashima, Axel Pahl, JP, Francois Berenger, strets123, JLVarjo, Noel O'Boyle, David Cosgrove, Patrick Fuller, Jan Holst Jensen, Gianluca Sforna, DoliathGavid, Karl Leswing, and Jeff van Santen. rdkit/rdkit: 2019 09 2 (q3 2019) release, 2019.

[18] Joao Aires-de Sousa, Markus C. Hemmer, and Johann Gasteiger. Prediction of 1h nmr chemical shifts using neural networks. *Analytical Chemistry*, 74(1):80–90, 2002.

[19] Yuri Binev, Maria M. B. Marques, and João Aires-de Sousa. Prediction of 1h nmr coupling constants with associative neural networks trained for chemical shifts. *Journal of Chemical Information and Modeling*, 47(6):2089–2097, 2007.

[20] Andres M. Castillo, Luc Patiny, and Julien Wist. Fast and accurate algorithm for the simulation of nmr spectra of large spin systems. *Journal of Magnetic Resonance*, 209(2):123 – 130, 2011.

[21] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schütz. Molpro: a general-purpose quantum chemistry program package. *WIREs Comput Mol Sci*, 2:242–253, 2012.

[22] NIST Computational Chemistry Comparison and Benchmark Database. Nist standard reference database number 101. 20, 2019.

[23] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[24] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[25] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[26] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOS ONE*, 9(6):1–12, 06 2014.

[27] Yifan Hu. Efficient and high quality force-directed graph drawing. *Mathematica Journal*, 10:37–71, 01 2005.