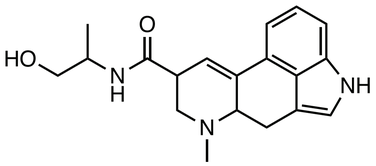
Supervised Topic Modeling for Predicting Chemical Substructure from Mass Spectrometry

*gkreder@stanford.edu

Gabriel Reder[1*], Adamo Young[2], Jaan Altosaar[3], Jakub Rajniak[1], Noémie Elhadad[3], Susan Holmes[4], Michael Fischbach[1]

[1] Stanford University Bioengineering Department
[2] University of Toronto Department of Computer Science
[3]Columbia University Department of Biomedical Informatics
[4] Stanford University Statistics Department

Metabolomics

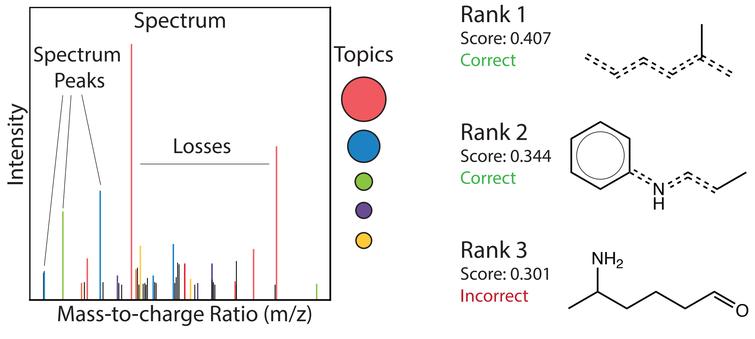


An example small molecule metabolite (ergonovine)

Metabolites are the small-molecule organic products of cellular metabolism. They are crucial in biological systems but many remain unidentified

Liquid chromatography – mass spectrometry (LC-MS) is often the method of choice for analyzing and discovering new metabolites, however chemical structure prediction from MS2 spectra is difficult

Topic modeling for MS2 substructure prediction



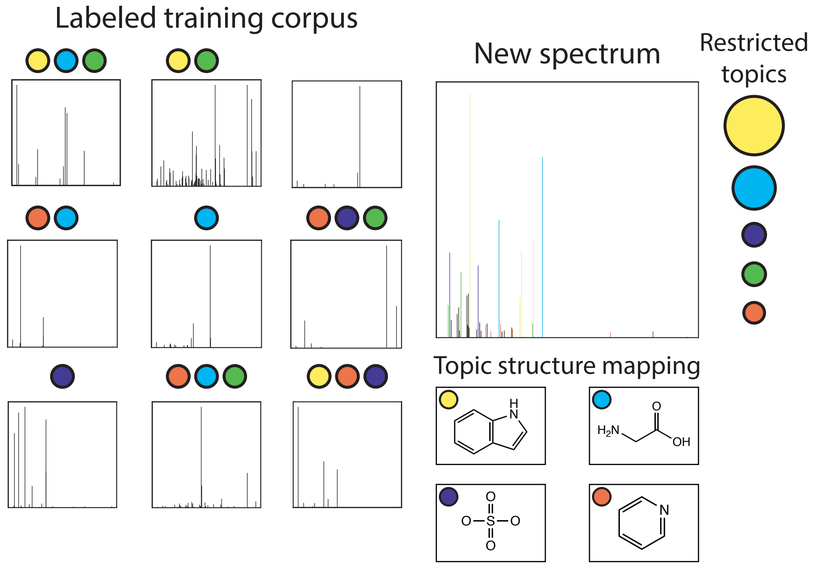
Topic modeling decomposes MS2 spectra into probabilistic topics using spectrum features (e.g. peaks and neutral losses)

See MS2LDA [1] for original implementation of topic modeling for MS2 substructure predictions

References

[1] J. J. J. v. d. Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess, and S. Rogers. "Topic modeling for untargeted substructure exploration in metabolomics". Proceedings of the National Academy of Sciences 48 (2016).
[2] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. "Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora". Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 - EMNLP '09. 2009
[3] Y. Liu, A. Mrzic, P. Meysman, T. D. Vijlder, E. P. Romijn, D. Valkenborg, W. Bittremieux, and K. Laukens. "MESSAR: Automated recommendation of metabolite substructures from tandem mass spectra". PLOS ONE 1 (2020)

Labeled latent Dirichlet allocation for MS2 spectra



Labeled Latent Dirichlet Allocation (LLDA) [2] is a supervised topic model that restricts topics to predefined user-specified tags

The labels are linked to specific chemical substructures during preprocessing before training. Spectrum features are thus fit to the substructures themselves

Spectrum features are also mapped to constrained molecular formulas in order to make resulting model topics chemically interpretable

Post-training: for a spectrum (d), the **similarity score** for a given substructure (k) is calculated using the following formula:

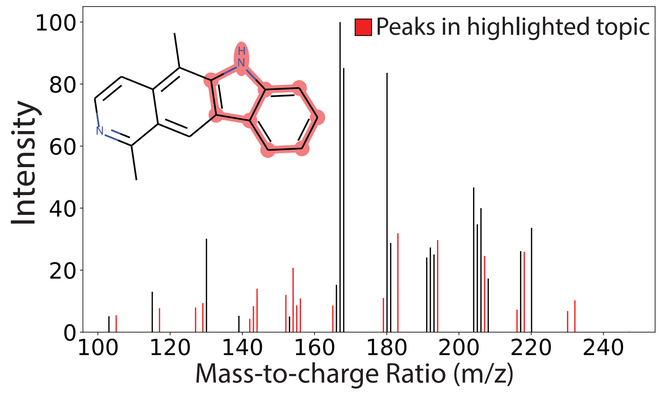
$$sim(k,d) = \frac{v_d^T v_k}{\|v_d\|\|v_k\|}$$

Where v_k is the word distribution for substructure k and v_k is the count in document d for every word in the training corpus

Comparison to alternatives

Methods	Spectra with ≥1 hit in top 3	Spectra with ≥ 2 hit in top 3
LLDA	125/185	82/185
MESSAR	79/185	40/185

Comparison to alternatives (continued)



Topic C1C[NH]C(C)C1CCC Composition

Comparison of our LLDA substructure prediction model to MESSAR [3], using the same train/test spectra and substructures labels. An example test spectrum is shown with its top (correctly) labeled substructure and interpretable topic

Peak Words	Neutral Loss Words
C12H8	loss_C2N
C15H13N	loss_C5N
C10H7	loss_C7H5N
C13H9N	loss_C8H4N
C10H10N	loss_C4H6
C11H8N	loss_C2HN
C14H12N	loss_C8H5N
C11H9N	loss_C7H2N
C12H11N2	loss_C6H5
C16H12N	loss_C4H4N

LLDA was also compared to a k-nearest neighbors (kNN) approach in increasingly difficult subsets of the test data in terms of chemical similarity (right to left) and substructure subset (blue to yellow). LLDA's relative performance increases in more difficult tests sets in both cases