Supervised Topic Modeling for Predicting Chemical Substructure from Mass Spectrometry

Gabriel K. Reder
Bioengineering Department
Stanford University
Stanford, California 94305
gkreder@stanford.edu

Jaan Altosaar
Department of Biomedical Informatics
Columbia University
New York, New York 10027

Jakub Rajniak
Bioengineering Department
Stanford University
Stanford, California 94305

Noémie Elhadad
Department of Biomedical Informatics
Columbia University
New York, New York 10027

Susan Holmes
Department of Statistics
Stanford University
Stanford, California 94305

Michael Fischbach
Bioengineering Department
Stanford University
Stanford, California 94305

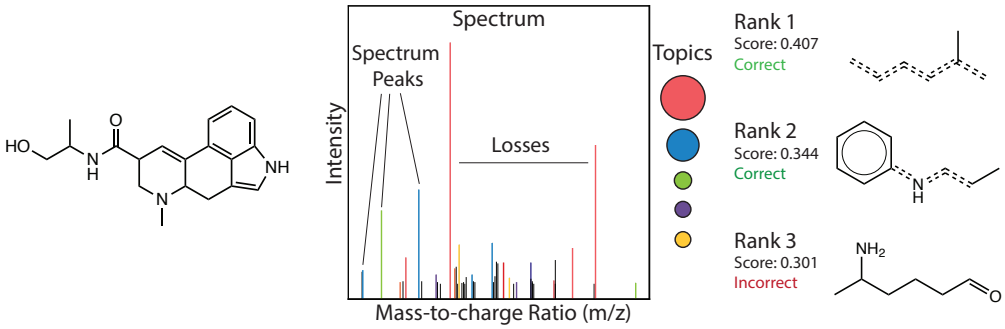
Abstract

Metabolites—organic molecules involved in or created by cellular metabolism—are ubiquitous in biological systems and are a rich source of drug candidates and disease biomarkers. Many metabolites are potentially important, but remain unidentified and unknown [7]. Mass spectrometry (MS) is a common way of identifying new metabolites en masse and is a promising option for clinical deployment. Given an unknown molecule, reliably identifying its chemical structure from its MS spectrum remains an open challenge. This is a difficult machine learning problem because the number of molecule-annotated MS spectra is low, numbering at best in the tens of thousands of human metabolites for a given instrument architecture and acquisition setup [11]. We propose a supervised topic modeling approach to identify modular groups of spectrum peaks and neutral losses in spectra corresponding to consistent chemical substructures. We use labeled latent Dirichlet allocation (LLDA) [20] to map spectrum features to known chemical structures. These structures appear in new unknown spectra, and can thus be predicted. We compare LLDA to an alternative developed in Liu et al. [14] that uses a lookup table mapping spectrum features to substructures. In an empirical study, the two approaches yield similar performance. We show that a benefit of LLDA is that it produces topics that are chemically interpretable, allowing for further model refinement.

1 Introduction

Metabolites are the bioactive small molecules that are created and used by cellular chemical processes [18]. They are ubiquitous, diverse, and central to biological systems; as such, they are important drug targets, candidates, and biomarkers of disease [26]. For example, the fungus Penicillium produces the metabolite Penicillin after metabolising amino acid dietary precursors. Many metabolites, however, remain unidentified, even those commonly found in human samples [23]. Identifying the structure of an unknown molecule or metabolite can be done via tandem mass spectrometry (MS/MS) or

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



(a) A metabolite, ergonovine (b) LC-MS/MS spectrum and LLDA topics (c) Substructure predictions

Figure 1: **Supervised topic modeling for substructure prediction**. (a) An example of a metabolite with a potentially unknown molecular structure. (b) Liquid chromatography tandem mass spectrometry (LC-MS/MS) can be used to analyze a metabolite via its spectrum. A metabolite's spectrum consists of mass-to-charge (m/z) peaks and neutral losses with associated intensities, resulting from fragmentation of the metabolite. Labeled latent Dirichlet allocation (LLDA) is used as a topic model of spectra. (c) Topics correspond to molecule substructures, and LLDA is used to predict substructures using Equation (1).

nuclear magnetic resonance spectroscopy. This work focuses on molecular identification from liquid chromatography tandem mass spectrometry (LC-MS/MS) since this technique requires low concentrations of a molecule, is cheap, fast, and has immediate clinical applications [21]. LC-MS/MS generates a spectrum for a molecule using the following process. First, a single molecule is fragmented into substructures using electric fields and collisions with an inert gas. The mass-to-charge ratio (m/z) of these fragments is measured, producing spectrum peaks. These fragments are themselves further broken apart, and the resulting smaller structures are measured or broken apart iteratively. In other words, a spectrum corresponds to one molecule, and each m/z peak in the spectrum corresponds some part of that molecule. Each peak has an associated intensity: the observed count of the number of fragments observed to have the peak's mass-to-charge value. Some substructures are lost during the fragmentation process and cannot be measured. These substructures are referred to as neutral losses, and do not appear as m/z peaks. Instead, neutral losses can be inferred by computing differences between m/z peaks ('lost' mass-to-charge) as shown in Figure 1.

Inferring chemical structure from a MS/MS spectrum is an open problem [4]. When done by hand, this process is time-consuming, laborious, and potentially unreliable [2]. Current computational methods to tackle this problem are described in Section 2. Here, we develop a supervised topic modeling approach to address the problem of labeling chemical substructures to help identify the structure of unknown metabolites. We use labeled latent Dirichlet allocation (LLDA) [20] to decompose MS/MS spectra into constituent chemical topics. Given a set of metabolites whose structure is known, alongside their MS/MS spectra, every spectrum can be labeled with a mixture of topics, or the substructures in the metabolite. LLDA learns patterns in these labels. Further, LLDA can predict the labels (substructures) of new spectra of molecules whose structure is unknown. We conduct an empirical study to compare LLDA to the closest existing alternative developed by Liu et al. [14]. LLDA performs as well as this competitor, and yields chemically interpretable topics that can help improve model fitting. As a probabilistic topic model, LLDA may help correct for label noise, arising from redundancy and ambiguity in computing which substructures occur in a spectrum.

2 Related Work

Spectral Library Matching One approach to molecular identification from MS/MS spectra is to search for a spectral match in libraries of publicly-available spectra. Such methods rely on computing a similarity score between two spectra and ranking library spectra matches in relation to an input query spectrum [2]. This approach is effective if the query spectrum corresponds to a previously-identified metabolite with an available spectrum in the library. However, spectral libraries tend to be sparse. The largest database of MS/MS spectra, Metlin, contains roughly 13,000 spectra from human metabolites [17]. Additionally, this matching is often done on the entire spectrum and can miss partial structure

matches. Simulated fragmentation approaches in which a predicted spectrum is generated from an input chemical structure offers a promising method of supplementing spectral libraries and has been used to expand the search space in spectral library matching [1, 25]. A popular method for generating these simulated spectra is Competitive Fragmentation Modeling-ID (CFM-ID) [5].

Fingerprint Prediction Another approach to molecular identification is to predict a molecular fingerprint from a spectrum, rather than the entire molecular structure. A molecular fingerprint is a vector representation of a molecule's structure [22]. The values in these fingerprint vectors may represent chemical substructures and properties as in Klekota and Roth [12] or may result from a learned embedding of molecular structure [27]. Notable current methods include CSI:FingerID [6], SIMPLE [16], and DeepEI [9]. A fingerprint-based approach requires mapping both molecular structures and spectra to fingerprints. However, mapping a molecule to a fingerprint can add noise [22].

Topic Modeling and Substructure Prediction Instead of identifying the structure of an entire molecule, chemical substructures can be identified from a spectrum. Chemical substructures are structural subunits that appear consistently across different molecules and are useful in identification [12, 15]. In tandem mass spectrometry, substructures often fragment consistently, independent of the rest of the molecule, and can therefore produce a recognizable set of mass-to-charge peaks in a spectrum. Methods to identify chemical substructures in spectra include the metabolite substructure auto-recommender (MESSAR) [14] and MS2LDA [8]. MS2LDA uses latent Dirichlet allocation [3] to associate groups of m/z peaks and neutral losses with chemical substructures [8]. While a promising approach, this method is unsupervised and requires manual labeling of the resulting topics produced by the model. MESSAR is a supervised method that uses association rule mining rather than topic modeling; we build on MS2LDA by using a supervised topic model.

3 Method

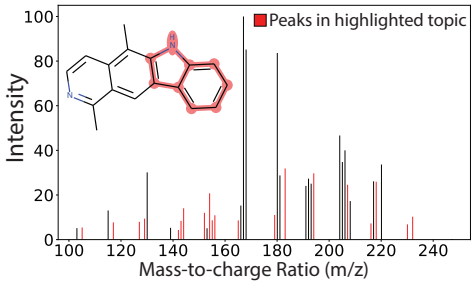
Supervised Topic Modeling We use labeled latent Dirichlet allocation (LLDA) [20] to model mass spectra and predict chemical substructure. LLDA is a supervised variant of latent Dirichlet allocation [3], which treats each document in a corpus as composed of words that come from a mixture of topics. LLDA assumes that every document in a corpus has been tagged or labeled with a subset of a known collection of topics; every word in a document is sampled from one of these topics' distributions over words. The model is described in full in Ramage et al. [20]. Instead of modeling documents, we model mass spectra corresponding to molecules. We note that every component of LLDA for modeling a document has an analog useful for modeling a spectrum. A document is an MS/MS spectrum. The words are both observed mass-to-charge peaks and neutral losses. Topics in the model are co-occurring spectrum fragments and differences. A visualization of these components is in Figure 1.

Preprocessing a Spectrum To convert a spectrum to a document, each peak is assigned the closest m/z-matched molecular formula such that (1) the formula's theoretical m/z is within 0.1 of the peak's m/z and (2) the formula is a subformula of the spectrum's parent molecular formula. (Even for unknown spectra, we assume that a molecular formula can be found since this problem has been addressed to a larger degree than structural prediction [10].) Peaks with no such formula match are discarded. Next, all possible neutral losses are computed as pairwise differences between peaks. A neutral loss is kept only if it corresponds to a valid molecular formula (the difference between the parent peak formula and the child peak formula) and the parent intensity of the neutral loss is greater than the child intensity. Each neutral loss is assigned the mean of its two respective peak intensities. Next, this set of peaks and neutral losses is converted to a document by setting each formula equal to a word, prepending `loss` to formulas of neutral losses, and setting word counts to rounded integer values of their associated intensities (spectrum-normalized to 100). Spectra are labeled with topics using the Python RDKit library [13]. As every spectrum corresponds to a molecular structure, the set of labels for a spectrum is a list of SMILES/SMARTS strings computed to be substructures of the parent molecule.

Identifying Substructures in a New Spectrum Training LLDA on a corpus of spectra yields a word distribution for every topic (substructure). To predict which substructures are likely to be

Method	# with $\geq$ 1 hit in top 3	# with $\geq$ 2 hits in top 3
LLDA	84/185	25/185
MESSAR [14]	79/185	40/185

(a) Substructures predicted in test set



(b) Example spectrum from test set

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

(c) Top peak words and neutral loss words in substructure topic highlighted in (b)

Figure 2: **LLDA performs similarly to competing methods, and provides interpretable topics.** (a) LLDA and MESSAR can both predict correct substructures. (b) An example spectrum from the test set, with a correctly-labeled substructure highlighted (SMILES string `C1C[NH]C(C)C1CCC`). Spectrum peaks corresponding to the most probable spectrum peak words for the topic are shown in red. (c) The highest-probability topic peak and neutral loss words are also shown. LLDA allows inspection of the model in terms of molecular formulas and probabilities in substructure topics.

present in a new, unlabeled spectrum, the cosine similarity between a new spectrum (document) d and substructure (topic) k is calculated:

$$\text{sim}(k, d) = \frac{v_d^\top v_k}{\|v_d\| \|v_k\|} . \tag{1}$$

Here v_k is the word distribution for topic k and v_d is the word count for every word in document d that appears in the training corpus. Ranking all substructures according to this cosine similarity results in a list of likely substructures in a new spectrum. We also tested a collapsed Gibbs sampler approach to inferring the topic distribution of a held-out spectrum; this was outperformed by Equation (1).

4 Experiments

Data To compare LLDA to the association rule mining approach presented in MESSAR [14], the same data from Liu et al. [14] is used. The training corpus is 3,146 positive mode LC-MS/MS spectra from a spectral library [24]. The topic labels for LLDA are the substructures in file S1 of Liu et al. [14]. For testing purposes, we used the same labeled 185 CASMI spectra used by the MESSAR authors (S2 data [14]). This dataset contains MASSBANK Q-TOF spectra for 34 drugs and 126 metabolites combined with 25 spectra from the CASMI 2017 contest (`http://casmi-contest.org/2017`).

Implementation The Python Tomotopy library [19] is used to train LLDA on documents generated as described in Section 3. Spectra are labeled with topics using the Python RDKit library [13]. LLDA is trained for 2,000 iterations and Equation (1) is used to predict substructures.

Results On the test dataset of 185 spectra, Liu et al. [14] report the following results for MESSAR on the top 3 recommended substructures for each spectrum: 79 cases in which at least 1 recommendation is correct and 40 cases in which at least 2 recommendations are correct. LLDA yields 84 cases in which at least 1 recommendation is correct and 25 cases in which at least 2 recommendations are correct. These results and an example spectrum and topic are shown in Figure 2.

5 Discussion

We developed a supervised topic model approach to identify molecular substructures in LC-MS/MS data. We identify three benefits of our approach: (1) topics from LLDA topics are interpretable in terms of molecular formulas, allowing for further model refinement and incorporation of prior knowledge (2) scalability: we place no restrictions on the number of spectrum peaks and neutral losses that may be associated with chemical substructures (unlike other approaches such as MESSAR, where the number is capped) (3) LLDA is a probabilistic topic model, and can thus help compensate for ambiguity, redundancy, or other noise from computing substructure labels. A number of limitations remain in LC-MS/MS metabolite identification including limited availability of training data [11] and difficulty of choosing a substructure set [22]. These are often difficult to disentangle from data features that are purely based on chemical structure. Future work includes incorporating prior knowledge such as ionization mode or instrument type, and testing LLDA on a larger dataset to study how much prior knowledge corrects for a lack of data. We nevertheless believe that our approach presents a promising way forward for de novo identification of unknown metabolites in LC-MS/MS data.

References

[1] F. Allen, R. Greiner, and D. Wishart. "Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification". *Metabolomics* 1 (2015).

[2] I. Blaženović, T. Kind, J. Ji, and O. Fiehn. "Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics". *Metabolites* 2 (2018).

[3] D. M. Blei. "Latent Dirichlet Allocation" ().

[4] T. De Vijlder, D. Valkenborg, F. Lemière, E. P. Romijn, K. Laukens, and F. Cuyckens. "A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation". *Mass Spectrometry Reviews* 5 (2018).

[5] Y. Djoumbou-Feunang, A. Pon, N. Karu, J. Zheng, C. Li, D. Arndt, M. Gautam, F. Allen, and D. S. Wishart. "CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification". *Metabolites* 4 (2019). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[6] K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker. "Searching molecular structure databases with tandem mass spectra using CSI:FingerID". *Proceedings of the National Academy of Sciences* 41 (2015).

[7] I. Gertsman and B. A. Barshop. "Promises and Pitfalls of Untargeted Metabolomics" (2019).

[8] 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).

[9] H. Ji, H. Deng, H. Lu, and Z. Zhang. "Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks". *Analytical Chemistry* 13 (2020). Publisher: American Chemical Society.

[10] T. Kind and O. Fiehn. "Advances in structure elucidation of small molecules using mass spectrometry". *Bioanalytical Reviews* 1-4 (2010).

[11] T. Kind, H. Tsugawa, T. Cajka, Y. Ma, Z. Lai, S. S. Mehta, G. Wohlgemuth, D. K. Barupal, M. R. Showalter, M. Arita, and O. Fiehn. "Identification of small molecules using accurate mass MS/MS search". *Mass Spectrometry Reviews* 4 (2018).

[12] J. Klekota and F. P. Roth. "Chemical substructures that enrich for biological activity". *Bioinformatics* 21 (2008).

[13] G. Landrum. *RDKit: Open-Source Cheminformatics Software*. URL: `https://www.rdkit.org/` (visited on 10/05/2020).

[14] 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).

[15] Y. Ma, T. Kind, D. Yang, C. Leon, and O. Fiehn. "MS2Analyzer: A Software for Small Molecule Substructure Annotations from Accurate Tandem Mass Spectra". *Analytical Chemistry* 21 (2014).

[16] D. H. Nguyen, C. H. Nguyen, and H. Mamitsuka. "SIMPLE: Sparse Interaction Model over Peaks of moLEcules for fast, interpretable metabolite identification from tandem mass spectra". *Bioinformatics* 13 (2018).

[17] D. H. Nguyen, C. H. Nguyen, and H. Mamitsuka. "Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches". *Briefings in Bioinformatics* 6 (2019).

[18] G. J. Patti, O. Yanes, and G. Siuzdak. "Innovation: Metabolomics: the apogee of the omics trilogy". *Nature Reviews Molecular Cell Biology* 4 (2012).

[19] *Python package of Tomoto, the Topic Modeling Tool*. URL: `https://bab2min.github.io/tomotopy/` (visited on 10/05/2020).

[20] 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.

[21] C. Seger and L. Salzmann. "After another decade: LC–MS/MS became routine in clinical diagnostics". *Clinical Biochemistry* (2020).

[22] M. A. Skinnider, C. A. Dejong, B. C. Franczak, P. D. McNicholas, and N. A. Magarvey. "Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm". *Journal of Cheminformatics* (2017).

[23] M. R. Viant, I. J. Kurland, M. R. Jones, and W. B. Dunn. "How close are we to complete annotation of metabolomes?" *Current Opinion in Chemical Biology* (2017).

[24] M. Wang et al. "Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking". *Nature Biotechnology* 8 (2016).

[25] J. N. Wei, D. Belanger, R. P. Adams, and D. Sculley. "Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks". *ACS Central Science* 4 (2019).

[26] D. S. Wishart. "Current Progress in computational metabolomics". *Briefings in Bioinformatics* 5 (2007).

[27] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay. "Analyzing Learned Molecular Representations for Property Prediction". *Journal of Chemical Information and Modeling* 8 (2019).