METHOD ARTICLE

Supervised topic modeling for predicting molecular substructure from mass spectrometry

[version 1; peer review: 2 approved with reservations]

Gabriel K. Reder[1], Adamo Young[2-4], Jaan Altosaar[5], Jakub Rajniak[1], Noémie Elhadad[5], Michael Fischbach[1], Susan Holmes[6]

[1]Stanford University, 443 Via Ortega, Stanford, CA, 94305, USA
[2]University of Toronto, 214 College St, Toronto, ON, M5T3A1, Canada
[3]Vector Institute, 661 University Ave Suite 710, Toronto, ON, M5G 1M1, Canada
[4]Terrence Donnelly Centre for Cellular & Biomolecular Research, 160 College St, Toronto, ON, M5S 3E1, Canada
[5]Columbia University, 622 West 168th St, New York, NY, 10032, USA
[6]Stanford University, 390 Jane Stanford Way, Stanford, CA, 94305, USA

v1 **First published:** 19 May 2021, **10**(Chem Inf Sci):403
https://doi.org/10.12688/f1000research.52549.1
Latest published: 19 May 2021, **10**(Chem Inf Sci):403
https://doi.org/10.12688/f1000research.52549.1

Abstract

Small-molecule metabolites are principal actors in myriad phenomena across biochemistry and serve as an important source of biomarkers and drug candidates. Given a sample of unknown composition, identifying the metabolites present is difficult given the large number of small molecules both known and yet to be discovered. Even for biofluids such as human blood, building reliable ways of identifying biomarkers is challenging. A workhorse method for characterizing individual molecules in such untargeted metabolomics studies is tandem mass spectrometry (MS/MS). MS/MS spectra provide rich information about chemical composition. However, structural characterization from spectra corresponding to unknown molecules remains a bottleneck in metabolomics. Current methods often rely on matching to pre-existing databases in one form or another. Here we develop a preprocessing scheme and supervised topic modeling approach to identify modular groups of spectrum fragments and neutral losses corresponding to chemical substructures using labeled latent Dirichlet allocation (LLDA) to map spectrum features to known chemical structures. These structures appear in new unknown spectra and can be predicted. We find that LLDA is an interpretable and reliable method for structure prediction from MS/MS spectra. Specifically, the LLDA approach has the following advantages: (a) molecular topics are interpretable; (b) A practitioner can select any set of chemical structure labels relevant to their problem; (c) LLDA performs well and can exceed the performance of other methods in predicting substructures in novel contexts.

Open Peer Review

Approval Status ? ?

	1	2
version 1 19 May 2021	? view	? view

1. **Joe Wandy**, University of Glasgow, Glasgow, UK

2. **Hunter N B Moseley**, University of Kentucky, Lexington, USA

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
Metabolomics, Machine Learning, Mass Spectrometry, Structure Identification, LC-MS, Tandem Mass Spectrometry, MS/MS

This article is included in the Cheminformatics gateway.

This article is included in the Artificial Intelligence and Machine Learning gateway.

Corresponding author: Gabriel K. Reder (gkreder@stanford.edu)

Author roles: Reder GK: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Young A**: Data Curation, Methodology, Writing – Review & Editing; **Altosaar J**: Data Curation, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Rajniak J**: Investigation, Supervision, Visualization, Writing – Review & Editing; **Elhadad N**: Funding Acquisition, Supervision; **Fischbach M**: Funding Acquisition, Project Administration, Supervision; **Holmes S**: Conceptualization, Formal Analysis, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This publication was made possible by a National Institutes of Health NIGMS-funded predoctoral fellowship by Grant Number T32GM008412 assigned in part to Gabriel Reder.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Reder GK, Young A, Altosaar J *et al.* **Supervised topic modeling for predicting molecular substructure from mass spectrometry [version 1; peer review: 2 approved with reservations]** F1000Research 2021, **10**(Chem Inf Sci):403 https://doi.org/10.12688/f1000research.52549.1

First published: 19 May 2021, **10**(Chem Inf Sci):403 https://doi.org/10.12688/f1000research.52549.1

Introduction

Liquid chromatography - tandem mass spectrometry (LC-MS/MS) is a powerful experimental method for identifying the small molecule metabolites in a sample of unknown composition. It provides detailed structural information from a given molecule with the only prerequisite knowledge being the parent molecule's mass-to-charge ratio. This is especially important since a vast portion of naturally occurring small molecules are believed to remain unidentified[1]. Yet identifying the structure of a molecule given its MS/MS remains challenging[2]. Traditionally, such identification is done by hand, which is difficult, time-consuming, and poses significant reproducibility issues. The repetitive nature of this task naturally lends itself well to a computational approach[3].

Existing tools generally follow one of two trends. In spectral library matching, a query MS/MS spectrum is compared to reference spectra via similarity metric such as a cosine score[4]. For example, the National Institute of Standards and Technology (NIST) Mass Spectral Library, Human Metabolome Database (HMDB), Metlin, and the Global Natural Products Social Molecular Networking (GNPS) databases all offer spectral similarity search functionalities using spectral similarity scores[5–8]. However, spectral databases tend to be sparse compared to the universe of possible molecular structures[3]. Simulated spectra can supplement databases with new structures, though matching is usually done on entire spectra[9–11]. Molecular fingerprint approaches in which vector representations of molecules are predicted directly from MS/MS spectra are increasingly popular methods[12–16].

This work focuses on the specific fingerprint task of substructure prediction in MS/MS spectra (Figure 1A). Chemical substructures are defined structural subunits that appear across

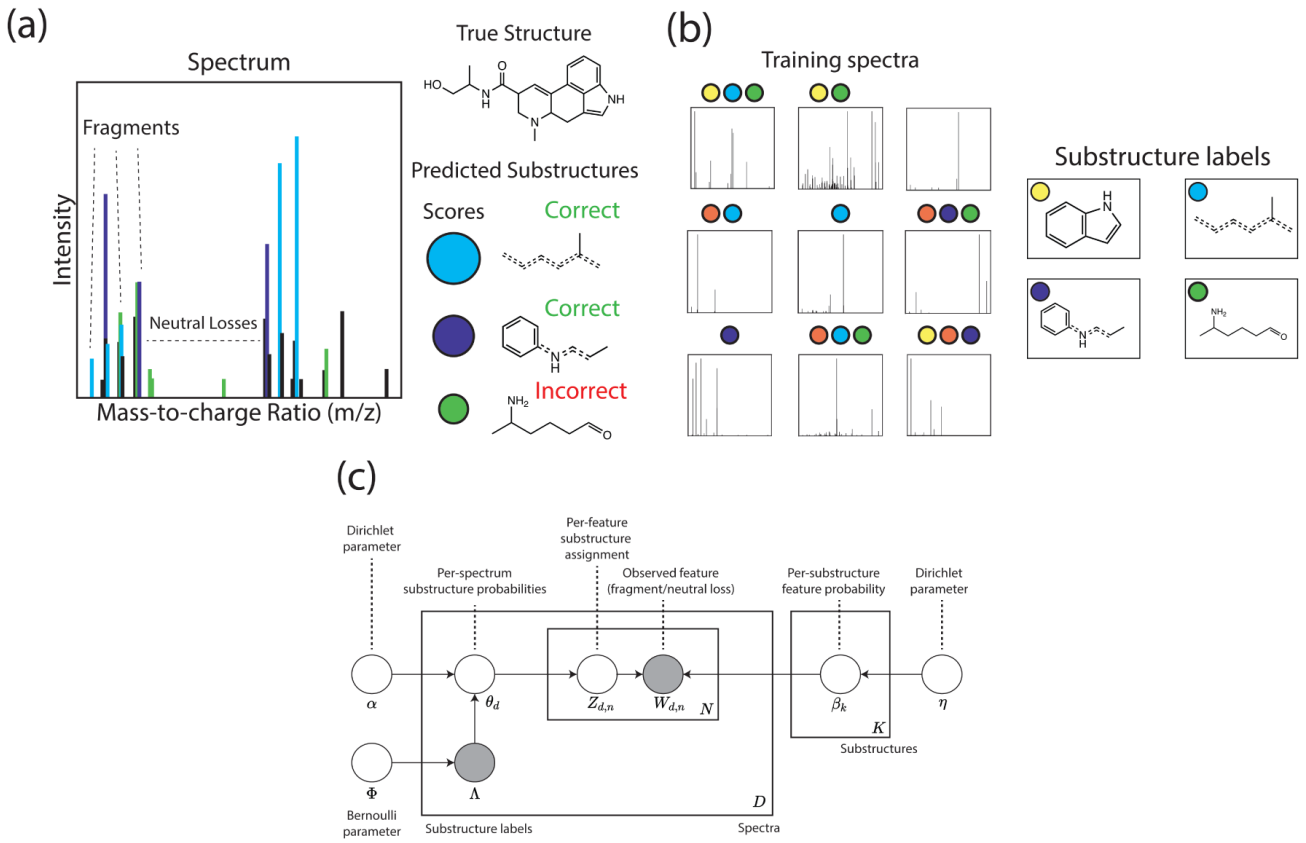

Figure 1. Labeled latent Dirichlet allocation (LLDA) for interpretably predicting structure in tandem mass spectrometry (MS/MS). (**a**) **Substructure prediction in MS/MS spectra**. A tandem (MS/MS) mass spectrum of a small molecule (ergonovine). Spectrum fragments and neutral losses provide information relevant to identifying chemical structure. The goal of the supervised topic modeling approach is to assign interpretable substructure scores for a spectrum via LLDA[17] (**b**) **Supervised learning for substructure prediction.** A training set is composed of MS/MS spectra with known chemical structures and a choice of substructure (topic) labels. Each spectrum is labeled with its parent molecule's substructures. (**c**) **The LLDA generative model.** The model is defined in 17 and labeled here as it relates to MS/MS substructure prediction. The model is trained on a corpus D of MS/MS spectra and set of molecular substructures K. Given the fragment/neutral loss (feature) composition $W_{d,n}$ of each spectrum and the substructure-spectrum label matrix Λ, the model will find the substructure-feature probability matrix β, the spectrum-substructure probability matrix θ, and the feature-substructure assignment matrix Z. α and η are Dirichlet parameters for the distributions from which θ and β are respectively assumed to have been drawn. Φ is a Bernoulli parameter for the distribution from which Λ is assumed to have been drawn. is a model. Figure (**c**) has been reproduced and modified with labels with permission from 17.

different molecules and are useful for identifying their parent molecules[13,18]. Focusing on substructures benefits practitioners because substructures have directly interpretable chemical meaning. Further, predicting chemical substructures rather than relying on spectral library matching allows for novel predictions of molecules not seen in these libraries. An existing approach, MS2LDA, uses topic modeling to find regularly co-occurring sets of MS/MS spectrum fragments and neutral losses[19]. Here we build on this work by proposing and developing a supervised topic modeling approach, using labeled latent Dirichlet allocation (LLDA)[17], to find co-occurring spectrum features associated with chemical substructures of the user's choosing (Figure 1B–C). We find that compared to alternative methods for supervised substructure identification, LLDA performs well and provides interpretable substructure predictions. Using cosine distance k-nearest neighbors (k-NN) for spectral library matching, we find that LLDA's relative performance improves as the test set becomes more chemically distinct from the training set and as the substructures being predicted appear with different frequencies between the two sets. This type of generalization is important for applications where the molecular identity of samples is unknown and thought to correspond to novel molecules.

Methods

Data sets

In order to facilitate direct comparison to an existing tool, we train and test LLDA using the same data and evaluation metrics as those used in the metabolite substructure auto-recommender (MESSAR) developed in 20. We utilize the MESSAR target training corpus of 3,146 positive mode LC-MS/MS spectra (available here) originally from the Global Natural Products Social Molecular Networking database[8]. For validation, we use the 5,164 MASSBANK[21] validation spectra and the 185 annotated test spectra as our test set both used in 20 and available here. This test dataset contains spectra for 34 drugs and 126 metabolites from MASSBANK and 25 spectra from the Critical Assessment of Small Molecule Identification 2017 contest. The 712 unique substructures provided in the rule database in 20 (available here) are used. All spectra are provided in .mgf format. Copies of the spectra and a cleaned version of the unique substructure set are included in this publication's *Underlying data*[22].

Data preprocessing

To model a mass spectrum using LLDA, it is necessary to represent a mass spectrum as a bag-of-words "document"[23]. First, any fragment having a mass-to-charge-ratio (m/z) below 30 is discarded to remove structurally uninformative fragments. Each fragment in the remaining spectrum is then assigned the molecular formula with the closest theoretical m/z and that satisfies two conditions (i) the formula's theoretical m/z is within 0.1 of the peak's m/z and (ii) the formula is a subformula of the parent spectrum's parent molecular formula (Figure 2A). We assume all MS/MS spectra have a corresponding molecular formula, a reasonable assumption for a spectrum even if its parent chemical structure is unknown[24]. This formula matching is done using the rcdk library (version 3.5.0). Peaks with no formula match are discarded. Next, all possible neutral losses are computed as pairwise differences between kept spectrum fragments. A neutral loss consists of a m/z difference between two fragments, f_a and f_b. Neutral losses in which f_a has both a greater intensity and m/z than f_b are assigned the mean of its two respective peak intensities and matched to formulas in the same manner as fragments. Neutral losses with no matching formulas are discarded. The word count Y for a given spectrum feature (fragment or neutral loss) with intensity x is calculated using a generalized logistic function[25]:

$$Y(x) = \frac{100}{(1+Qe^{-Bx})^2} \quad (1)$$

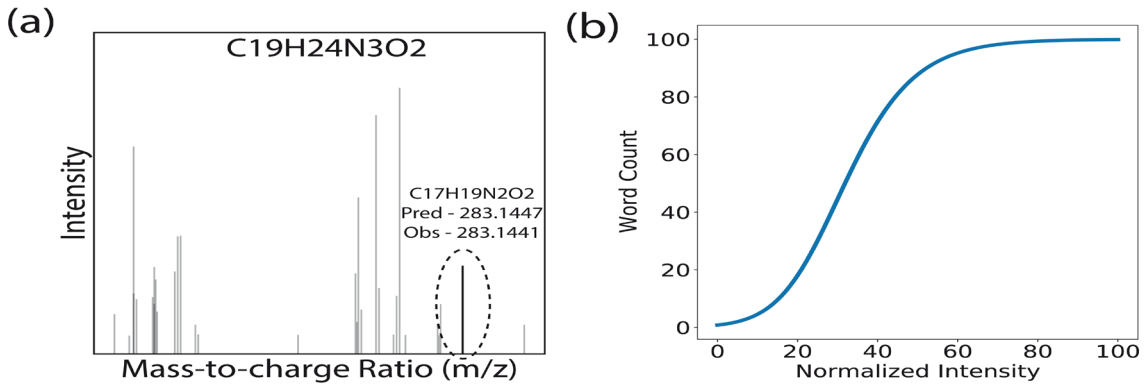

Figure 2. Tandem mass spectrometry (MS/MS) data preprocessing and bag-of-words conversion. (**a**) **Mapping spectrum features to molecular formula words.** For representing a mass spectrum as a document, each spectrum fragment corresponding to a peak is first mapped to a molecular formula by finding the formula that matches the following criteria: (i) the theoretical mass-to-charge ratio of the formula is within 0.1 m/z of the observed fragment. (ii) the formula is a subformula of the spectrum's parent formula. The closest such m/z formula is kept, the same process is repeated for neutral losses, and spectrum features with no formula matches are discarded. (**b**) **Mapping feature intensity to word counts**. Word counts are computed for each spectrum fragment and neutral loss using the feature's normalized intensity as in Equation 1. A word count response for Q = 10 and B = 0.1 is illustrated here.

where Q determines the threshold value at intensity $x = 0$ and B determines the growth rate of the word count growth rate as the intensity increases. Input intensities of spectrum fragments are normalized such that the maximum raw intensity is 100, with word counts rounded to the nearest integer. An example intensity response function for intensity values in the range [0, 100] is shown in Figure 2B.

Using the data and code provided, this document generation process including formula mapping and bag-of-words conversion can be run using the command:

```
python make_documents.py \
--in_mgf <in_mgf_file> \
--out_dir <out_documents_dir> \
--eval_peak_script evaluate_peak.R \
--n_jobs 1 \
--adduct_element H \
--adduct_element 1
```

We note that this process can be quite time consuming if run on a single core - preprocessing a single spectrum using a single core can take more than two minutes depending on the spectrum. As such, the preprocessed spectra for the specified data sets have been provided (see *Underlying data*[22]). Runtime can be decreased on a multi-core system by increasing the n_jobs input parameter.

Substructure labels for training spectra having known parent chemical structures are assigned using the RDKit library (version 2020.09.5). Training spectra are labeled with the substructures in the user-defined set that are present in the given spectrum's parent structure. Substructure labeling can be run using the provided data and code using the command:

```
python prep_data.py \
--train_mgf train.mgf \
--test_mgf test.mgf \
--validate_mgf validate.mgf \
--df_substructs df_substructs.tsv \
--embedded_spectra_out <embedded_spectra_fileName_out.npz> \
--df_labels_out <df_labels_fileName_out.tsv> \
--df_ids_out <df_ids_fileName_out.tsv>
```

Training LLDA and predicting substructures in a new spectrum

The LLDA model is implemented using the Python Tomotopy library (version 0.10.2)[26]. The model takes a set of training spectra, test spectra, and substructures as input. The spectra must be in bag-of-words format, and the training spectra must have been labeled with ground truth substructures as described above. The original LLDA model is described in full in 17. We note that every component of LLDA for modeling text documents has an analog useful for modeling a MS/MS spectrum. A document is an MS/MS spectrum, words are spectrum features (fragments and neutral losses), topics are commonly co-occurring spectrum features, and tags are chemical substructures (Figure 1). LLDA was trained for 2,000 iterations, since model perplexity tended to converge at around 2,000 iterations across various data sets generated using the preprocessing scheme. We note that the number of necessary iterations may differ across data sets. In addition to including a required input argument for the number of iterations in the LLDA model, we also include an optional flag to record and plot model perplexity.

Substructure predictions in test spectra are calculated using cosine similarity. The cosine similarity between a new spectrum i and substructure j is calculated as:

$$sim(k,d) = \frac{v_k^\top v_d}{\left\|v_k\right\|\left\|v_d\right\|} \quad (2)$$

Here v_k is the word distribution for topic k and v_d is the word count in document d that appears in the training corpus. Both vectors inherit their lengths from the number of words present across all documents. In this manner, every substructure is assigned a score for presence/absence in a test spectrum. After preprocessing, the LLDA model can be trained and tested using the commands

```
python run_llda.py \
--Q <Q> \
--B <B> \
--out_dir <llda_out_directory> \
--train_mgf train.mgf \
--test_mgf test.mgf \
--documents_dir documents \
--df_substructs df_substructs.tsv \
--df_labels df_labels.tsv \
--num_iterations <number_iterations>
```

K-nearest neighbors

For comparison to the practice of spectral library matching, we also implement a k-nearest neighbors method for substructure prediction. Spectra are represented as vectors by assigning fragments to m/z bins of width 0.1. Only fragments with mass-to-charge ratio in the range [0, 1000] are kept; the vectors representing the spectra are thus of length 10,000. To make a prediction for substructures associated with a new spectrum using the k-nearest neighbors algorithm, the k nearest neighbors in the training set of spectra are computed using the cosine similarity between the vectors corresponding to spectra in the training set. The score for each substructure in the new spectrum is calculated as the mean of the substructure presence and absence labels in the k nearest neighbors' spectra. For example, for a single substructure s and test spectrum d, if 4 of the 5 nearest neighbor spectra contain s, the predicted score for s in d will be 4/5 = 0.8. After data preprocessing, k-NN can be run using the command:

```
python run_knn.py \
--df_ids <df_ids_filename.tsv>
--df_labels <df_labels_filename.tsv> \
--k 10 \
--embedded_spectra <embedded_spectra_fileName.npz> \
--out_dir <knn_out_directory>
```

Further code documentation including required libraries and optional arguments can be found in the README file included in this publication's *Underlying data*[22]. A Docker image containing all necessary software and library requirements is also available at Docker Hub (see *Extended data*).

Results and discussion

Preprocessing hyperparameter search

As described in the preprocessing pipeline, the generalized logistic function (GLF) is used to convert normalized feature intensity values into word counts to represent spectra as documents. To evaluate the hyperparameter choices in the GLF, LLDA was evaluated using a range of Q and B values corresponding to a wide range of intensity response functions. This performance was measured using the same metric as in 20. Specifically, the metric, denoted as t3≥2, measures the number of spectra in which at least two of the top-3 scoring substructures are true positives. This is similar to the standard metric of recall used in recommender systems[27]. In this experiment, LLDA predicted substructures most sensitively for Q and B values corresponding to binarized word counts. In other words, for a given document, words that appear in the given spectrum receive a word count of 1 and all other words receive a word count of 0; this corresponds to Q = 10.0 and B = 0.001 (Figure 3A). This finding deserves further research and means that in this formulation, intensity information may harm model performance. This could be for a number of reasons. One possibility is that the GLF formulated in Equation 1

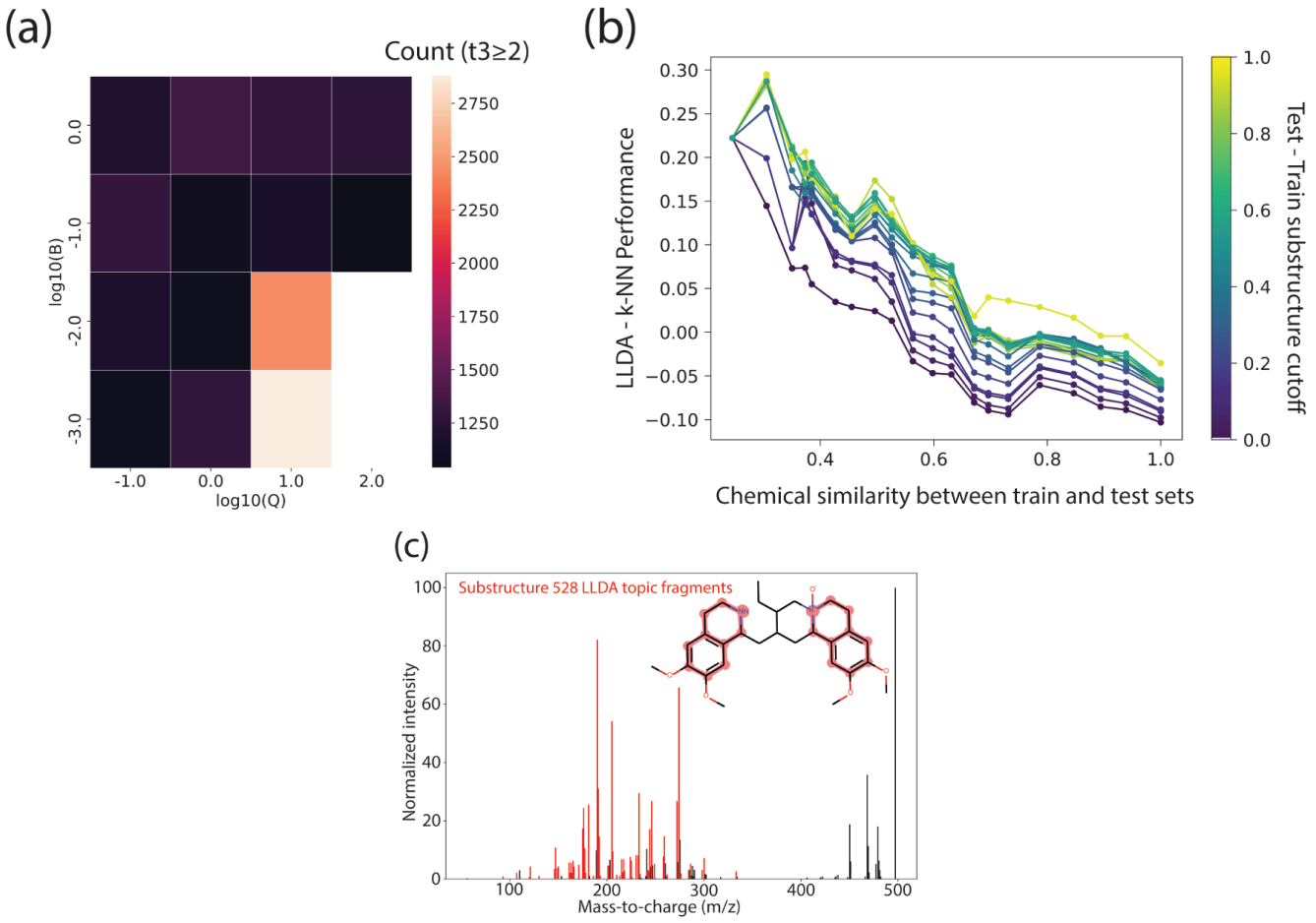

Figure 3. Preprocessing-optimized labeled latent Dirichlet allocation (LLDA) performs competitively in substructure identification and generalizes in difficult test sets. (**a**) **Empirical study of normalized intensity response function.** The performance of the LLDA model using the intensity response shown in Figure 2b with various values of Q and B on the validation set of 5,164 MASSBANK spectra used in 20. Performance is measured by counting the number of spectra in which at least 2 of the top-3 predicted substructures are correct (t3≥2). The binarized response function performs best (Q = 10, B = 0.001), meaning that a generative process that does not include normalized intensity information outperforms other choices in this formulation. (**b**) **LLDA outperforms k-nearest neighbors (k-NN) in generalizing to evaluation sets with less overlap in substructure or chemical similarity.** The difference between LLDA performance (measured by the area under the receiver operating characteristic or AUC) and k-NN performance is computed. The average chemical similarity is computed between the train and test sets using the RDKit fingerprint. Colors represent the substructure appearance difference between the training and test sets. Increasing the difficulty of generalization along either axis of molecular similarity or substructure overlap increases LLDA performance relative to the k-NN model. (**c**) **LLDA can recognize substructures outside train context.** A test spectrum is shown in which substructure 528 appears. The parent structure (Pubchem ID 57509371) is displayed with the portions of the structure containing substructure 528 highlighted. Fragments in red are in the 95th percentile of words in the topic for substructure 528 in the LLDA. The k-NN model performs poorly on such substructures that appear out of context in the test set, while LLDA maintains predictive performance in this case.

is not an optimal choice of function to map intensities to word counts. There are many options of function for this conversion, the GLF was chosen because of its flexibility to produce a diverse set of intensity-word count relationships. Another possibility is that neutral loss intensity encoding must be revisited. There may be better methods of representing a neutral loss's intensity other than taking the mean of its two constituent fragments. Additionally, valuable information might potentially be destroyed when raw ion counts are converted to normalized intensities as is standard practice[14,18].

Substructure identification benchmark

LLDA, with the best-performing values of Q and B (Q = 10.0, B = 0.001, corresponding to binary word counts) was then run on the benchmark test dataset from 19, consisting of 185 spectra. On this data, the authors of 20 report the following results for MESSAR on the top 3 recommended substructures for each spectrum: 79 spectra in which at least 1 recommendation is correct (t3≥1) and 40 spectra in which at least 2 recommendations are correct (t3≥2). The LLDA model yields 125 cases for t3≥1 and 82 cases for t3≥2.

The k-nearest neighbors method was studied using k = [1, 5, 10] on the same set of evaluation data. For these experiments, we find that the t3≥1 metric is [111, 128, 130] respectively and that t3≥2 is [74, 117, 121]. These results indicate that the k-nearest neighbors method with k=10 outperforms all other methods on this set of train/test data according to these metrics. These results are shown in Table 1. However, it is unclear whether it is possible to extract an analogue to 'topics' from the k-nearest neighbors algorithm, as is common in LLDA and topic models. Further research is needed to develop methods that perform as well as k-nearest neighbors while retaining the interpretability and modularity of topic modeling approaches to substructure identification, such as LLDA. As the output of machine learning methods for mass spectrometry data is typically inspected by a human in an experimental procedure, developing interpretable methods remains important. We note that the same train spectra, test spectra, and evaluation metrics as used in 20 were used in this work. This was done to maximize the comparability between methods. Future development of community-wide standards for benchmark datasets and evaluation metrics will greatly facilitate development of new methods.

Table 1. Performance of labeled latent Dirichlet allocation (LLDA) and cosine distance k-nearest neighbors on test data from 20.

Method	t3≥1	t3≥2
MESSAR[20]	79	40
LLDA	125	82
K-nearest neighbors (k=1)	111	74
K-nearest neighbors (k=5)	128	117
K-nearest neighbors (k=10)	130	121

Ablation study

To study how well k-nearest neighbors performs compared to LLDA on novel test molecules poorly represented in the training set, increasingly difficult subsets of the test data were constructed using two notions of how data might be limited in a real-world experimental setting (i) chemical similarity between training and test molecules and (ii) differential train-test appearance for substructures.

For chemical similarity, the RDKit fingerprint similarity was calculated between each test set of spectrum parent structure and all structures in the training set. These similarity scores were then used to set similarity thresholds for selecting increasingly difficult subsets of the test data. For example, a maximum similarity threshold of 0.4 produced a subset of the test spectra in which the maximum pairwise structure fingerprint similarity to the train structures is at most 0.4. For substructure appearance, the number of times a substructure appears in the training set is counted and normalized according to the number of spectra. The same is done for the test set. Next, substructures are ordered by these differences and percentile cutoffs are used to produce subsets of test spectra of increasing difficulty in terms of test - train appearance. For example, a 60th percentile cutoff means testing only substructures whose normalized test minus train appearance values are above the 60th percentile of all such values across all substructures. The performance of the LLDA model compared to the k-NN model was calculated as the LLDA area under the receiver operating characteristic curve (AUC) averaged across the given test set minus the average AUC for the k-NN model.

The results of this study are shown in Figure 3B. Increasing difficulty along either axis of chemical similarity or substructure overlap improved the performance of the LLDA model relative to the k-NN model. This effect was especially pronounced as test-train molecular similarity became more distant. These results indicate use cases in which an approach such as LLDA may be especially useful compared to spectral library searching. LLDA may be able to better recognize novel chemical configurations of substructures in new spectra. As such it may be a better-suited model for characterizing spectra measuring molecules coming from new and understudied areas of chemical space not well represented in spectral libraries.

Conclusions

Metabolites are central in biology, yet the vast majority are likely unidentified[28]. Untargeted metabolomics via LC-MS/MS is a promising option for identifying new metabolites in a high throughput pipeline. Improved computational methods for identifying chemical structure from MS/MS spectra are needed for this promise to become a reality. With this in mind, we developed, described, and open-sourced a supervised topic modeling method for identifying chemical substructures in tandem mass spectrometry data via LLDA[17]. In a series of

empirical studies, this supervised topic model was trained and tested on publicly available benchmark data and substructures, and LLDA was compared to an alternative method, MESSAR[20]. A k-nearest neighbors (k-NN) was also implemented as a means of testing spectral library matching to predict substructure labels based on neighbor averages.

We report several benefits of the LLDA method. First, when trained and tested using the same spectra, LLDA provides interpretable, probabilistic substructure topics and performs well using the same metrics as in 20. These topics can incorporate a large number of spectrum fragments and neutral losses, so the patterns of spectrum fragments and neutral losses associated with substructures can be as complex as necessary for good predictive performance. LLDA is a probabilistic model that can compensate for ambiguity, redundancy, and other noise that arises from computing substructure labels. The advantage of such a probabilistic method is that substructure labels often have significant overlap[12]. Finally, the LLDA model offers a flexible supervised framework. A practitioner may choose any set of substructure labels on which to train the LLDA model, allowing the user to tailor the output of the model to a specific application requiring accurate and interpretable substructure identification. LLDA offers a supervised topic modeling approach that complements both the benefits of MS2LDA[19], circumventing the need to pick an arbitrary number of unsupervised topics or to map output topics to substructures - this relationship is predetermined by the user.

By systematically exploring the preprocessing pipeline used to map spectrum features to a representation of spectra as documents, we find that this LLDA model performs best when intensity information is hidden from the model and binary word counts are used to represent MS/MS spectra (Figure 3A). This aligns with existing work in probabilistic topic models used in recommender systems, such as collaborative topic Poisson factorization, where binarized ratings lead to improved predictive performance[29]. However, we also note that this effect may arise from the choice of train, test, and validation sets that were taken from publicly available spectra with potential heterogeneity in collision energies. We further note the possibility that a different function for translating intensities to word counts, especially for neutral losses, may result in better performance. Neutral losses may be more robust against systemic uniform measurement error in a spectrum and as such could represent more stable sources of information. However, the manner in which intensities are assigned to neutral losses and which neutral losses are kept will heavily affect model performance.

The k-nearest neighbors (k-NN) model with k=10 performs very well using the same data and evaluation metrics, raising important considerations about trade-offs between predictive performance of substructure identification and interpretability of results. We find that the k-NN model may perform well for the task of substructure identification in situations in which substructures appear in similar patterns between train and test sets. Similarly, k-NN may perform well when a test spectrum corresponds to a molecule that is similar to those in the train set. But as explored in Figure 3B, the k-NN model suffers when these similarities are reduced, and the test spectra correspond to increasingly different molecules from those in the train set or when the substructure appearance varies between train and test sets. This observation is important to consider in the development and evaluation of machine learning methods for substructure identification; assessing generalization performance in real-world settings should reflect cases in which a new spectrum is coming from an underexplored area of chemical space that is not well represented by existing spectral libraries. We leave to future work the problem of including such quantities into evaluation metrics to more accurately assess generalization. Ablation studies such as the one presented here can provide the foundation for better metrics and ways of construct evaluation sets relevant to a substructure identification task in a specific problem setting.

An example case study: the substructure with identifier 528 in the unfiltered train/test set corresponds to SMARTS string C1Cc2ccccc2CN1 and appears in one test spectrum. In this test spectrum, substructure 528 appears without many of its co-substructures from the training set, and these co-substructures from the training set appear in the test set without substructure 528. As such, the k-NN (k = 10) model produces false positive predictions for substructure 528 in the test set, resulting in an AUC of 0.43. But the LLDA model picks up on this substructure's presence in the test set without suffering from false positives, achieving an AUC of 0.99. The spectrum (corresponding to spectrum 128 in the test set) containing substructure 528, its structure, and spectrum fragments in the top 0.95 percentile of substructure 528's LLDA topic are shown in Figure 3C.

Further work is required to better optimize LLDA. Additional preprocessing steps can be explored, such as keeping spectrum fragments that do not map to a child molecular formula but appear consistently across spectra (rather than discarding them as described in the Method). The inclusion of such orphaned fragments could improve downstream ranking of substructures. A similar preprocessing step is effective in processing amplicon sequencing data[30]. A number of limitations remain in LC-MS/MS metabolite identification including a paucity of training data[31], difficulties in selecting a vocabulary of substructures[12], and the heterogeneity involved in instrument choice, acquisition method, and sample conditions. Performance drops resulting from these issues are often difficult to disentangle from features of the data that result solely from chemical structure. Future work includes optimization of substructure label sets, incorporation of prior knowledge such as ionization mode or instrument type, and testing these models on a larger dataset. The work presented here highlights the increasing interest and relevance of representing MS/MS spectra in manners that can capture more information than a mass-to-charge ratio binning scheme and applying a metric for predictions such as cosine similarity. Another recent example of such an effort can be found in 32.

By releasing all code and preprocessed training data for the methods developed here, we encourage reproducibility

efforts alongside the development of machine learning methods to solve *de novo* identification of unknown metabolites in LC-MS/MS data as computational metabolomics stands to benefit from community engagement, consistent public evaluation methods, and open-source models.

List of abbreviations

MS/MS: Tandem mass spectrometry

LC-MS/MS: Liquid chromatography - tandem mass spectrometry

m/z: mass-to-charge ratio

LLDA: Labeled latent Dirichlet allocation[17]

MESSAR: Metabolite substructure auto-recommender[20]

k-NN: K-nearest neighbors

AUC: Area under the receiver operating characteristic curve

GLF: Generalized logistic function

Data availability

Underlying data

Zenodo: MS2 LLDA Topic Model.

https://doi.org/10.5281/zenodo.4589653[22].

This project contains the following underlying data:

- readme.md (file containing descriptions, specifications, and instructions for included data and using code).
- requirements.txt (exported conda environment containing the necessary Python dependencies – R decencies must be installed separately. See readme.md).
- data.zip (data used in paper's empirical study, see readme.md).
- code.zip (code for preparing data, running LLDA, and running kNN. See readme.md).

Extended data

Docker image: https://hub.docker.com/r/gkreder/ms2-topic-model

Analysis code available from Github: https://github.com/gkreder/ms2-topic-model

Archived code at time of publication: https://doi.org/10.5281/zenodo.4589653[22].

References

1. Viant MR, Kurland IJ, Jones MR, *et al.*: **How close are we to complete annotation of metabolomes?** *Curr Opin Chem Biol.* 2017; **36**: 64–9.
PubMed Abstract | Publisher Full Text | Free Full Text
2. de Vijlder T, Valkenborg D, Lemière F, *et al.*: **A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation.** *Mass Spectrom Rev.* 2018; **37**(5): 607–29.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Nguyen DH, Nguyen CH, Mamitsuka H: **Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches.** *Brief Bioinform.* 2019; **20**(6): 2028–43.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Blaženović I, Kind T, Ji J, *et al.*: **Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics.** *Metabolites.* 2018; **8**(2): 31.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Stein S: **Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification.** *Anal Chem.* 2012; **84**(17): 7274–82.
PubMed Abstract | Publisher Full Text
6. Wishart DS, Feunang YD, Marcu A, *et al.*: **HMDB 4.0: the human metabolome database for 2018.** *Nucleic Acids Res.* 2018; **46**(D1): D608–17.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Guijas C, Montenegro-Burke JR, Domingo-Almenara X, *et al.*: **METLIN: A Technology Platform for Identifying Knowns and Unknowns.** *Anal Chem.* 2018; **90**(5): 3156–64.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Wang M, Carver JJ, Phelan VV, *et al.*: **Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking.** *Nat Biotechnol.* 2016; **34**(8): 828–37.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Allen F, Greiner R, Wishart D: **Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification.** *Metabolomics.* 2015; **11**(1): 98–110.
Publisher Full Text
10. Wei JN, Belanger D, Adams RP, *et al.*: **Rapid Prediction of Electron-Ionization Mass Spectrometry Using Neural Networks.** *ACS Cent Sci.* 2019; **5**(4): 700–8.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Djoumbou-Feunang Y, Pon A, Karu N, *et al.*: **CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification.** *Metabolites.* 2019; **9**(4): 72.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Skinnider MA, Dejong CA, Franczak BC, *et al.*: **Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm.** *J Cheminformatics.* 2017; **9**(1): 46.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Klekota J, Roth FP: **Chemical substructures that enrich for biological activity.** *Bioinformatics.* 2008; **24**(21): 2518–25.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Dührkop K, Shen H, Meusel M, *et al.*: **Searching molecular structure databases with tandem mass spectra using CSI:FingerID.** *Proc Natl Acad Sci U S A.* 2015; **112**(41): 12580–5.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Nguyen DH, Nguyen CH, Mamitsuka H: **SIMPLE: Sparse Interaction Model over Peaks of moLEcules for fast, interpretable metabolite identification from tandem mass spectra.** *Bioinformatics.* 2018; **34**(13): i323–32.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Ji H, Deng H, Lu H, *et al.*: **Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks.** *Anal Chem.* 2020; **92**(13): 8649–53.
PubMed Abstract | Publisher Full Text
17. Ramage D, Hall D, Nallapati R, *et al.*: **Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora.** In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing - EMNLP '09*. Singapore: Association for Computational Linguistics; 2009; **1**: 248.
Reference Source
18. Ma Y, Kind T, Yang D, *et al.*: **MS2Analyzer: A Software for Small Molecule Substructure Annotations from Accurate Tandem Mass Spectra.** *Anal Chem.* 2014; **86**(21): 10724–31.
PubMed Abstract | Publisher Full Text | Free Full Text
19. van der Hooft JJJ, Wandy J, Barrett MP, *et al.*: **Topic modeling for untargeted substructure exploration in metabolomics.** *Proc Natl Acad Sci U S A.* 2016; **113**(48): 13738–43.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Liu Y, Mrzic A, Meysman P, *et al.*: **MESSAR: Automated recommendation of metabolite substructures from tandem mass spectra.** *PLoS One.* 2020; **15**(1): e0226770.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Horai H, Arita M, Kanaya S, *et al.*: **MassBank: a public repository for sharing mass spectral data for life sciences.** *J Mass Spectrom.* 2010; **45**(7): 703–14.
PubMed Abstract | Publisher Full Text

22. Reder G: **MS2 LLDA Topic Model.** *Zenodo*. 2021.
http://www.doi.org/10.5281/zenodo.4655149

23. HaCohen-Kerner Y, Miller D, Yigal Y: **The influence of preprocessing on text classification using a bag-of-words representation.** *PLoS One.* 2020; **15**(5): e0232525.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Kind T, Fiehn O: **Advances in structure elucidation of small molecules using mass spectrometry.** *Bioanal Rev.* 2010; **2**(1–4): 23–60.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Richards FJ: **A Flexible Growth Function for Empirical Use.** *J Exp Bot.* 1959; **10**(2): 290–301.
Publisher Full Text

26. bab2min, Fenstermacher D: **bab2min/tomotopy: 0.10.0**. *Zenodo.* 2020; [cited 2021 Feb 1].
Reference Source

27. Dacrema MF, Cremonesi P, Jannach D: **Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches.** *Proc 13th ACM Conf Recomm Syst.* 2019; 101–9.
Publisher Full Text

28. da Silva RR, Dorrestein PC, Quinn RA: **Illuminating the dark matter in metabolomics.** *Proc Natl Acad Sci U S A.* 2015; **112**(41): 12549–50.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Gopalan P, Charlin L, Blei DM: **Content-based recommendations with Poisson factorization.** In: *Proceedings of the 27th International Conference on Neural Information Processing Systems.* Cambridge, MA, USA: MIT Press; 2014; **2**: 3176–84. (NIPS' 14).
Reference Source

30. Callahan BJ, McMurdie PJ, Rosen MJ, *et al.*: **DADA2: High-resolution sample inference from Illumina amplicon data.** *Nat Methods.* 2016; **13**(7): 581–3.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Kind T, Tsugawa H, Cajka T, *et al.*: **Identification of small molecules using accurate mass MS/MS search.** *Mass Spectrom Rev.* 2018; **37**(4): 513–32.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Huber F, Ridder L, Verhoeven S, *et al.*: **Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships.** *PLoS Comput Biol.* 2021; **17**(2): e1008724.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 06 September 2021

https://doi.org/10.5256/f1000research.55846.r92355

?

Hunter N B Moseley
Department of Molecular & Cellular Biochemistry, University of Kentucky, Lexington, KY, USA

This is a paper that describes the use of labeled latent Dirichlet allocation (LLDA) in combination with k-nearest neighbor selection of library spectra to develop a set of molecular substructure classifiers for MS/MS spectra. While the paper is reasonably written and the results are interesting and useful for further methods development, there are the following major issues that should be addressed.

Major Issues:

1. While Figure 1 is a very helpful introduction to how LLDA is being used, the combination of this figure and the main text is still difficult to discern the full mathematical form being used. As this reviewer (who has degrees in mathematics and computer science, but also teaches statistics) can discern, there is a deconvolution of per-substructure spectral feature probabilities from a training set of substructure-labeled spectra. This is done in terms of pairs of co-occurring spectral features per substructure. This reviewer suspects that a set of 2-dimensional Dirichlet's are being deconvoluted (i.e., derived) during the supervised training. The authors should improve their description of the supervised learning so that these details are clear. Providing mathematical definitions of alpha, theta, beta, Z, W, Bernouli, and lambda would greatly aid this fundamental understanding of the methodology.

2. The authors point about k-NN for spectral library matching improving LLDA relative performance as the test set becomes more chemically distinct from the training set is important. However, it may highlight a more fundamental problem with utilizing k-NN to help identify all substructures for a given MS/MS spectrum. The fundamental issue may be missing the combination of spectral library spectra that provides complete coverage of spectral features in the test spectrum. If this is the case, then k-NN should not be blindly applied. This reviewer wants the authors to consider this possibility in their discussion.

3. The authors should indicate how heterogeneous in terms of analytical platforms and spectral resolution the LC-MS/MS spectra in the GNPS and MassBank spectra used in

training and testing. Were any filtering criteria utilized in selecting these sets of spectra? There appears to be a de facto filtering of spectra which do not contain any peaks with associated molecular formulas.

4. Coding style of the codebase is at the typical research programming level, which needs improvement. Here is a non-exhaustive list of needed improvements. There is no use of python docstrings. Many variable names are not descriptive. There is no main function defined and utilized in the accepted pythonic manner. There is global namespace pollution with some of the import statements. Codebase is also not designed as a full python package that allows both command line execution or use as a library. These comments come from an expectation for industrial standards of coding, even in scientific programming projects.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: metabolomics data analysis, metabolic network analysis, bioinformatics, systems biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 11 August 2021

https://doi.org/10.5256/f1000research.55846.r88758

?

Joe Wandy
Glasgow Polyomics, University of Glasgow, Glasgow, UK

Predicting substructural information from LC-MS/MS fragmentation data is an important problem in untargeted metabolomics. In this work, the authors introduced a novel workflow to preprocess fragmentation data and perform supervised topic modelling approach using labeled latent Dirichlet allocation (LLDA). The LDDA model finds co-occurring fragment and loss features that are associated with chemical substructures provided by users.

The main benefit of incorporating label information through LLDA is that it allows for a list of substructures of interest to be inferred. Topics in LLDA are associated with meaningful substructure labels, aiding in model interpretation and improving from standard LDA where users are required to specify the number of topics in advance and manually annotate topics to provide chemical meaning - a time consuming and laborious process.

In this study, LLDA was evaluated on a benchmark dataset of 185 chemicals having known substructural annotations against a rule-based alternative (MESSAR) and a baseline k-nearest neighbour method, and it was found to perform competitively in retrieving substructural hits particularly when applied to new and unseen test data.

Overall I think the work done in this manuscript is interesting and advances the field of substructural discovery in metabolomics. I am happy to recommend this for indexing as long as my comments below are addressed. The comments are separated into major and minor categories, and further divided by the corresponding sections of the manuscript.

Major comments

Section: Data sets

1. Please make it clear that the '712 unique substructures provided in the rule database in [20]' are obtained by selecting unique substructures entries from the final MESSAR database of 8378 association rules after filtering on FDR and recall. It's also worth emphasising in the manuscript that these 712 unique substructures are mined from the GNPS dataset in [20]. Additionally please indicate in the manuscript how users could also generate their own list of substructures if they have unique/special datasets where the pre-generated GNPS list of substructures from [20] don't apply.

Section: Data Processing

1. The authors remarked that the process of extracting a document from a spectra is a computationally expensive process.- *We note that this process can be quite time consuming if run on a single core - preprocessing a single spectrum using a single core can take more than two minutes depending on the spectrum.* - Please provide more details what's the breakdown of computational cost in the feature extraction script. Which part is the slowest? As I understand it, there are two steps to document creation: formula mapping and bag-of-words conversion. It seems that the largest computational cost would come from elemental

formula enumeration via cdk. A typical fragmentation data set easily consists of a few thousand fragmentation spectra, and at two minutes each, the total feature extraction process could easily run up to days (for a single core) or many hours (for multicores). If true, please state this clearly in the manuscript so the reader is aware of it.

2. Is performing elemental formula mapping actually necessary? The author comments that '*keeping spectrum fragments that do not map to a child molecular formula but appear consistently across spectra (rather than discarding them as described in the Method)*' could be beneficial. This raises the question, is it necessary to perform such expensive procedure to map formulae to fragment and loss features, while the LDA (and LLDA model) could also work on the binned fragment and loss features that appear consistently across spectra, e.g. using 'fragment_132.0152' and 'loss_100.9203' as words of the document, following MS2LDA [19]. Filtering features by formula could also potentially remove frequently occurring fragment or loss features that are important to substructural identification. It would make the manuscript stronger if the authors could investigate how an alternative and simpler feature extraction scheme performs in comparison to what's been done in the manuscript.

Section: Training LLDA and predicting substructures in a new spectrum

1. Please expand the description of LLDA below. - *The original LLDA model is described in full in 17. We note that every component of LLDA for modeling text documents has an analog useful for modeling a MS/MS spectrum. A document is an MS/MS spectrum, words are spectrum features (fragments and neutral losses), topics are commonly co-occurring spectrum features, and tags are chemical substructures (Figure 1).* - I would like to see some textual elaboration of the plate diagram in Figure 1C, in particular the generative process of LLDA (using mass spec vocabularies) and how it differs from standard LDA. What are the advantages of LLDA vs LDA in this application? I had to refer to the original LLDA paper [17] and deduced all these things myself, but really it should be made explicit in the manuscript.

2. What's the perplexity/log likelihood of LLDA compared to standard LDA on the training and test data sets? Also please report the wall clock of running 2000 iterations of LLDA. And what are the hyperparameters used for alpha and eta in the experiments? Were they using the defaults in tomotopy or were they optimised? How sensitive are the results in Table 1 to the choice of alpha and eta?

3. Is there any particular reason why substructure predictions in test spectra are performed using the cosine similarity in equation 2, instead of directly inferring topic distributions for the new document using the trained model -- as what's commonly done with LDA? It seems that in tomotopy this can be easily done using the make_doc and infer commands, so I don't understand why cosine similarity is used instead.

Section: Preprocessing hyperparameter search

1. Please elaborate on how 't3>2' is computed, specifically what are the meaning of the entries in the confusion matrix (TP, FP, TN, FN), and how precision and recall are defined in this context. I can see this is already defined in the MESSAR paper [20], but it would be useful to have it self-contained in this manuscript too (even if it's in the supplementary) -- particularly since the terminology used here ('t3>2') is slightly different from [20].

2. Please include the results of $t5 \geq 3$ and $t10 \geq 5$ for completeness in the comparison to

MESSAR? Also how's a 'hit' calculated in this manuscript? In MESSAR: - *Small substructures with fewer than 5 non-hydrogen atoms were discarded. Based on the output of both tools, we retrieved the number of hits (when the predicted substructure is an exact match of ground-truth) among top 3, 5 and 10 recommended substructures.*- Is it also done the same when computing the results in Table 1? If yes, please mention it here..

Section: Conclusions

1. Is there any downsides of using LLDA vs standard LDA?

Minor comments

Section: Data Processing

1. It should be made clearer that the 'words' used as input to the model are the elemental formulae, rather than the binned fragment and loss features - unless my understanding is incorrect?

2. In the code block of *make_documents.py*, the purpose of passing an R script (*evaluate_peak.R*) to *--eval_peak_script* is to be able to perform elemental formula annotation via rcdk, right? If yes, the manuscript should make it clearer. Also this is only my preference, but I prefer not to mix languages (Python and R) in a script, so would it be better to run the R script separately and generate an output file that can be read by the Python script later on?

3. The author notes that - *"Substructure labels for training spectra having known parent chemical structures are assigned using the RDKit library (version 2020.09.5). Training spectra are labeled with the substructures in the user-defined set that are present in the given spectrum's parent structure."* - I think this paragraph can be made clearer as to what RDKit is used for. My understanding is: using RDKit, the SMILES of the parent structure of the training spectra are used to check for the presence of substructures from the list of 712 unique substructures. Present substructures are then used as labels of the training spectrum. Is this correct? If yes, please make it clearer in the manuscript.

Section: Training LLDA and predicting substructures in a new spectrum

1. The text above equation 2 says that 'the cosine similarity between a new spectrum i and substructure j' but in the equation 2 below it, it shows the similarity between k and d instead. Please clarify.

Section: K-nearest neighbors

1. I found the description in this paragraph (below) to be somewhat confusing with regard to which spectrum is in the test, training and validation sets, although I could try to guess. '*To make a prediction for substructures associated with a new spectrum using the k-nearest neighbors algorithm, the k nearest neighbors in the training set of spectra are computed using the cosine similarity between the vectors corresponding to spectra in the training set. The score for each substructure in the new spectrum is calculated as the mean of the substructure presence and absence labels in the k nearest neighbors' spectra.*'

Section: Preprocessing hyperparameter search

1. The figure caption mentions that the grid search on Q and B are done using the validation data set (MassBank), but the manuscript doesn't seem to mention this. Please mention it

too in the main text.

Section: Conclusions

1. The paragraph that begins with '*An example case study:* ...' it seems to me it should be in the results rather than in the conclusion.

Code availability

1. Please include the requirements file (or the conda equivalent) in the project github.

2. Could also include a link to the data at https://doi.org/10.5281/zenodo.4589653 in the github, since the readme file mentioned 'data' but it isn't found anywhere.

3. The script `*run_baseline.py*` is also mentioned in the readme but can't be found in the github.

4. The manuscript should mention where the codes are that were used to calculate the performance in Table 1 (Substructure identification benchmark) and also for the Ablation study so it can be used for benchmarking and reproduction by the community. If they were not included in the code repository, perhaps they should be there.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Metabolomics, machine learning, topic modelling.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.