University of Glasgow | School of Computing Science

Knockoff-based target-decoy approach to compute similarities of metabolomic spectra and decoys using Spec2Vec embeddings

Małgorzata Kurkiewicz (2145411)

COMPSCI5082P (40 credits) — April 8, 2021

ABSTRACT

Accurate metabolite identification from liquid chromatography tandem mass spectrometry (LC–MS/MS) is a difficult problem that has been addressed by a few studies. The most common technique consists of matching fragment spectra to reference library spectra. To be able to assess the quality of these matches, the contemporary methods use target-decoy approaches that are compatible with the commonly used cosine similarity scores while keeping the False Discovery Rate (FDR) under control. Recently, there has been an advancement in definition of spectral similarity which introduced an unsupervised machine learning technique, Spec2Vec, that has proven more successful in identification of metabolites. This work evaluates some of the existing target-decoy methods using Spec2Vec and introduces a new Gaussian knockoff-based target-decoy method native to Spec2Vec embeddings.

1. INTRODUCTION

Metabolomics is a study of metabolites - small molecular end products of cellular regulatory pathways. A complete set of metabolites synthesized by a biological system such as a single organism can be referred to as its 'metabolome'[1]. Metabolites are smaller than proteins and most lipids. They complement upstream biochemical information obtained from proteins, genes and transcripts allowing us to improve our understanding of cell biology. The analytical techniques used to analyse them involve Liquid Chromatography (LC) and Gas Chromatography (GC) coupled with Mass Spectrometry (MS). These experimental techniques allow for measuring the masses of charged molecules (ions).

1.1 Metabolite Identification

The following vertical bar graph (Figure 1) represents a single mass spectrum from a mass spectrometer of one compound (trans-Zeatin) from a metabolite dataset (GNPS). When visualising mass spectra, the x-axis represents mass per charge (m/z) for each ion (most of the ions in metabolomics have a single charge, therefore in our work, the x-axis simply represents mass). The y-axis represents the intensity (the relative abundance) of a given ion. The intensity is proportional to the number of molecules present in the sample. Computationally, each spectrum is represented as a list of tuples to represent peaks, first element being the m/z of the peak and second the intensity.

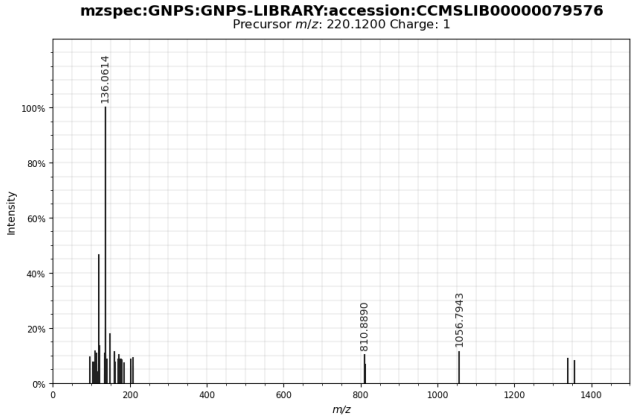

Figure 1: A representation of a single spectrum belonging to a compound trans-Zeatin collected on a Quadrupole Time-of-Flight (QTOF) LC Mass Spectrometer, present in the GNPS database and visualised in the Metabolomics Spectrum Resolver [2]. The x-axis represents mass per charge (m/z) for each ion. The y-axis represents the intensity of a given ion.

Figure 1 can be accessed via the following QR code:

1.1.1 Challenges

Mass spectrometers with high mass-accuracy have been used to assist chemical identification of metabolites. As the molecules often fragment in a way that correlates to their structure, we can try to annotate the unknown compounds via comparisons to a database [3][4][5]. One of the most basic approaches for metabolite identification is a mass match between two spectra. Given an unknown compound (query) and its spectral peaks, we search for metabolites in the database (library spectra) that are the closest to each peak considering their masses. However, as the mass of the molecule increases, the number of molecular fomulae increases exponentially[6]. Also, mass-to-charge ratio may not be sufficient enough to precisely assign a given compound to one chemical formula and even more so to a chemical structure[6]. This approach is also limited by the type of mass spectrometer as the peak masses can differ from their theoretical values due to measuring errors on different spectrometers. To this date, only a fraction of metabolites has been identified amongst many detected signals in untar-

geted metabolomic experiments[7][8]. This is not only due to the measuring errors of mass spectrometers but also because the samples contain large amounts of contaminants, chemical noise and 'degenerate features': redundant 'noise' that consists of isotopes, adducts, or in-source fragments[8][9]. This results in problems with identifying true matches and often leads to many false discoveries.

1.1.2 Scoring Systems

To help in effectiveness and validation of a true/false assignment of query-target matches, various scoring systems have been created, all of which are based on the concept of cosine similarity (See Section 3.1.1). There has been an advancement in definition of spectral similarity which introduced an unsupervised machine learning technique, Spec2Vec[10], that has proven more successful in identification of metabolites. This innovative approach to measuring spectral similarity represents related spectral fragments as vectors pointing in similar direction within latent space (See Section 3.1.2).

1.1.3 Decoy methods

All the scoring systems mentioned above do not provide sufficient power to discriminate true from false matches by themselves. Appropriate validation of metabolite assignments relies on manual inspection and comparative analyses with compound standards. This process is costly, inefficient and most importantly not very feasible for larger datasets. Therefore, there have been approaches designed to estimate the False Discover Rate (FDR), the proportion of false query-target matches amongst all matches. This allows scientists to make a conscious decision on the proportion of false positives (incorrectly identified spectra) they are willing to accept in their work. The problem of FDR estimation in metabolomics has been addressed by a few studies ([11], [12], [13]), some of which adapt their approaches from other fields such as proteomics and statistics, basing their techniques on creation of decoys[14]. Decoy libraries can be described as fake spectra that are derived from the original library, they represent metabolomic-like compounds that are not present in the target database. Therefore searching against a decoy database results in an incorrect metabolite identification. All of these target-decoy methods are compatible with the commonly used cosine similarity scores.

There is a decoy technique called the knockoff procedure[15], that has emerged in the field of statistics and has proven to accurately control the FDR. This method uses powerful statistical concepts that allow us to preserve the correlation between real features and its decoys, allowing for an appropriate imitation of the original features while keeping their statistical properties. As the knockoff procedure is flexible and can be derived for many statistical models, there has been a new contribution on generating knockoffs from a Gaussian mixture model[16]. Currently, there is no knockoff-based target-decoy approach for FDR control in metabolomic annotations in the literature.

2. AIM

The aim of this work is to implement and evaluate a new approach for creating decoy databases which allows for controlling the FDR given metabolomics data. The new technique uses a knockoff target-decoy approach and Spec2Vec embeddings for LC-MS and GC-MS metabolomic data to compute the similarities between real and decoy spectra. We examine if the statistical properties kept by creating knockoffs adhere to the methods already proposed in metabolomics. We present the comparison of the contemporary target-decoy cosine similarity-based techniques with Spec2Vec and knockoff-based approach for decoy creation. This work helps us improve current methods and introduces more efficient ways of creating decoy databases in different spaces that are not limited by their size as other metabolomic FDR-controlling methods.

3. BACKGROUND

3.1 Standard database search for metabolite identification

Due to challenges listed in Section 1.1.1, the best approach for annotating known metabolites is to use reference spectral databases which have compounds that are commercially available as pure standards. Although databases are relatively small compared to the number of unknown metabolites, a significant effort has been made to increase the number of annotated metabolites in public databases[17]. Scoring systems are the central algorithms used for database search in metabolomics. It is important to note that most of these scores alone do not indicate the statistical accuracy of the spectrum-spectrum match [18][19]. Also, since the comparison of mass spectra is not a standardized process, we can obtain different matching spectra depending on the database used. It is also worth mentioning that sorting spectral hits by spectral similarity scores is a computationally intensive task, especially given larger databases. Even when sorting all the similarity scores, there is no guarantee that the highest score would result in a true positive match as these algorithms result in a substantial overlap between true and false positive matches. The following section describes various spectral scoring systems.

3.2 Spectral Similarity Scoring

To help us identify a true match, we obtain a range of spectra with similar mass, we assign a similarity score to each match and pick the highest one. Various flavours of similarity scoring functions exist, which are described in the subsections below. When a query and a target (library spectrum) compounds are the same, we obtain a true hit. If the query and a target compound are not the same, we obtain a false hit. The true unique identity of a compound can be described by an International Chemical Identifier (InChI) containing structured information about chemical layers (main, charge, stereochemical, isotopic, fixed-H and reconnected layer). An example of the InChI of trans-Zeatin from Figure 1 is: "1S/C10H13N5O/c1-7(4-16)2-3-11-9-8-10(13-5-12-8)15-6-14-9/h2,5-6,16H,3-4H2,1H3,(H2,11,12,13,14,15)/b7-2+". Sometimes comparisons are done via a condensed 27 character hashed version of the InChI designed for easier searches of chemical compounds, it is described as the InChIKey. The InChIKey of trans-Zeatin from Figure 1 is "UZKQTCBAMSWPJD-FARCUNLSSA-N".

3.2.1 Cosine-based Spectral Similarity Scores

Cosine-based scores are the most popular measures of spectral similarity. They assume that molecules with high

structural similarity result in spectra with a high spectral cosine similarity score.

Cosine similarity can be represented by this formula:

$$similarity(Q,T) = \frac{Q \cdot T}{||Q|| \times ||T||} = \frac{\sum_{i=1}^{n} Q_i \times \sum_{i=1}^{n} T_i}{\sqrt{\sum_{i=1}^{n_Q} Q_i^2} \times \sqrt{\sum_{i=1}^{n_T} T_i^2}}$$

where Q represents one spectrum (query), T represents another spectrum (target), n = number of matching peaks, n_Q and n_T the number of peaks for Query and Target respectively, and i a single peak.

It is computed by aligning the fragment peaks from the two spectra. The number of computations can be represented as ((no. peak fragments)^2 / 2) - no. peak fragments, which can be considered as a computationally intensive operation considering the extensive library search that is necessary for it.

For visual representation, we have presented two spectra of the same compound (tran-Zeatin) present in two different public libraries with the corresponding cosine similarity score (Figure 2).

Figure 2 can be accessed via the following QR code:

Modified cosine (normalized dot product) similarity has been created because cosine similarity measures could not fully support the structural and spectral similarity between given spectra. This method includes relative intensities of the fragment ions and the precursor m/z difference between the paired spectra[20].

These similarity measures function well with spectra that do not differ much but they are not able to appropriately adjust to the complex molecules with high structural similarity, that differ in multiple locations (molecules that contain multiple chemical modifications)[10].

In practice, cosine scores can vary based on chosen tolerance (two peaks are considered a match if their m/z ratios lie within the given tolerance), the minimum number of matching peaks and different peak weighting. It is important to search for the appropriate combination of these parameters as we may receive a high number of false positive identifications.

3.2.2 Spec2Vec Spectral Similarity Score

Spec2Vec[10] is a new approach for defining spectral similarity scores based on a popular natural language processing algorithm –Word2Vec[21]. Spec2Vec is an unsupervised machine learning technique that represents related fragments and neutral losses as vectors pointing in similar directions within latent space. It provides an ability to compute similarities in a faster and more efficient way compared to conventional cosine similarity scores. Spec2Vec embedding vectors are fixed length, which makes the similarity score computation more efficient, for instance, one can find the best match of many spectra by a single matrix multiplication. Whereas the cosine-based methods align all pairs of fragments ions that have a similar m/z, the extraction based on similar m/z is costly especially when we have a large number of matching peaks. Moreover, Spec2Vec similarity scoring method has been documented to correlate more closely with structural similarity compared to cosine-based scoring methods[10]. See Methods for the implementation of the Spec2Vec similarity score to mass spectra.

3.3 False Discovery Rate

As mentioned in the section above, when analysing tandem mass spectra, we try to compare the observed fragmentation spectrum with the spectra present in our database in order to obtain the spectrum that best matches the observed one. We have discussed various score systems that have been created to solve this problem. All of these systems attempt to identify a match (correct identification of the observed spectrum) with the library spectrum, and associate a score to that match. The appropriate follow-up question would be to identify which of the matches are correct. As discussed previously, the biggest problem with these algorithms is that there is a substantial overlap between correctly and incorrectly identified spectra. Just taking these measures into account would mean that we either need to eliminate a lot of true positive identifications in order to control the amount of false positives or accept a large amount of false positive identifications to maximise the number of true positives. Scientists from many fields have tried to come up with appropriate procedures that would help them achieve statistically significant results. One of the desirable outcomes for all researchers is a low fraction of false discoveries among all discoveries to make sure their discovery is replicable and accurate. Hence, statistical significance methods like False Discovery Rate (FDR) have also been used in the context of spectral matching. To be able to appropriately test our null hypothesis, we can search our original spectra against a decoy database. Comparing a query to a realistic decoy ensures an incorrect match hence the ability to identify the amount of false positives in our sample. We cover these techniques in detail in the following Section.

3.3.1 p-value

One of the significance measures in statistics is the p-value. The p-value is defined as the probability of obtaining a result at least as extreme as the observation at hand, assuming the null hypothesis is correct. The smaller the p-value, the stronger the evidence for rejecting the null hypothesis. In other words, the observed outcome is unlikely under the null hypothesis. Given a decoy database we can compute the p-values by taking the assigned score and calculating the number of decoy spectra that received that score or higher. Just using a p-value threshold will not be adequate enough since we perform our statistical test so many times i.e. many spectral matches, hence the need for multiple testing correction.

3.3.2 Multiple Testing Correction

FDR is a method that accounts for the proportion of type 1 errors in null hypothesis testing. Type 1 errors appear when we incorrectly reject the null hypothesis, they can be described as false positives. In the context of spectral matching, false positive is a query-target match that is incorrect, hence a false hit. Testing multiple hypotheses simultaneously using methods for testing single hypotheses results in many false positives, this phenomenon is described as multiple comparisons problem. Yoav Benjamini and Yosef Hochberg proposed a solution to that problem in 1995[22], the proposal offered controlling the FDR from data. John D. Storey improved the FDR[23] with a new metric called q-value. It allows for the FDR to be defined even when there

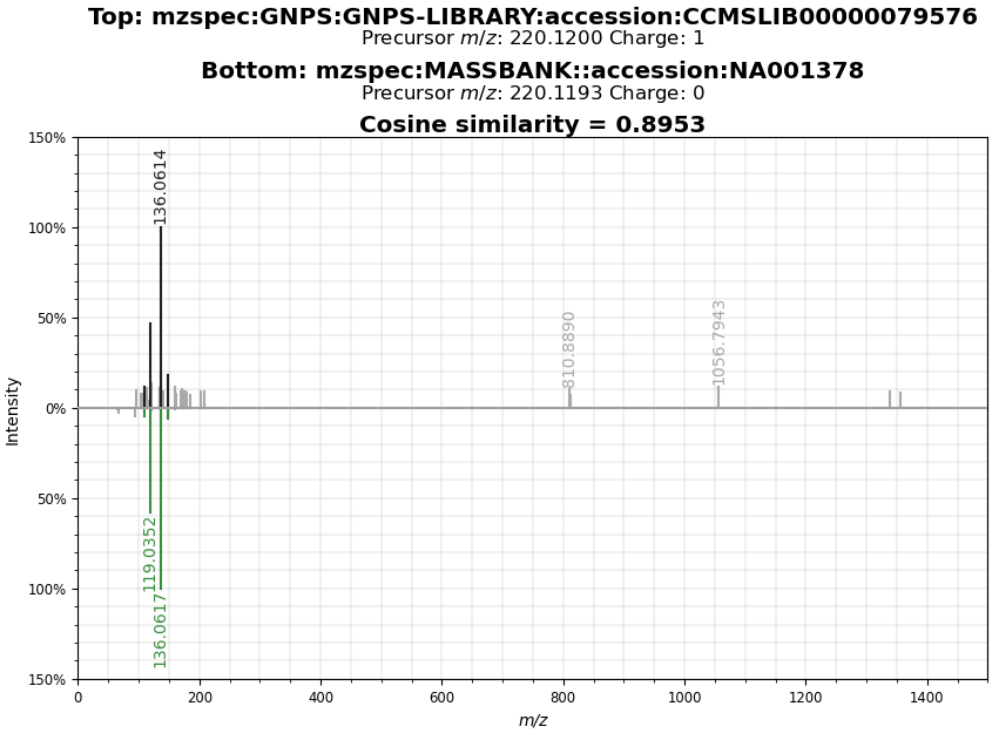

Figure 2: A comparison of two mass spectra of the same compound: trans-Zeatin, visualised in the Metabolomics Spectrum Resolver [2]. The top spectral graph represents trans-Zeatin collected on a Quadrupole Time-of-Flight (QTOF) LC Mass Spectrometer present in the GNPS database. The bottom spectral graph represents trans-Zeatin collected on a Thermo Scientific Linear Trap Quadrupole (LTQ) Orbitrap LC Mass Spectrometer present in the Massbank database. The x-axis represents the m/z and the y-axis intensity. We can see the matching peaks higlighted in black and green. The resulting cosine similarity is 0.8953.

are no positive results and helps in cases where two different scores lead to the same FDR. Therefore the current definition of FDR estimation includes the calculation of a q-value. In the simplest terms q-value is the minimal FDR threshold at which we accept the given discoveries. It is important to note that a q-value is a property of a single element of the dataset whereas the FDR is a property that describes the entire dataset. Therefore, we can associate a q-value with every similarity score resulting from a spectral match.

As the q-value is analogous to the p-value, in the estimation of the FDR we use the adjusted p-value distribution to create a q-value list. When conducting multiple hypotheses testing, for each of the tests we produce a p-value. The q-value approach finds the point where the p-value distribution for each test flattens out, that point is then used for calculation of the FDR adjusted p-values. This approach establishes the amount of significant values that are false positive. The q-values will always lie between 0 and 1.

More explicitly, if we consider the FDR value of 0.05, we mean that 5% of discoveries are expected to be false and, therefore at least 95% are true.

3.4 Target-Decoy Approaches For FDR Control in Metabolomics

3.4.1 Contemporary techniques for controlling the FDR in metabolomics

As discussed above, a problem associated with databases which are very often incomplete and contain noisy data is making match assignments prone to false positives (i.e. reported match of a spectrum with an equivalent library reference that is false). The following findings are compared to the methods described by Scheubert et al.[11]. In this section we describe the FDR estimation methods which are based on the creation of a decoy spectral library – most commonly used method in proteomics[14].

Decoy databases mimic real spectra closely but do not represent any real metabolites. This ensures that false hits in the target library are as likely as hits in the decoy database. Since metabolites are more diverse in structure the approaches for decoy creation from proteomics such as (pseudo-)reversing or shuffling the target database could not be applied. The authors have suggested three approaches to tackling the problem, however implemented only one of them. The method is described as a re-rooted fragmentation tree. The lab has created a tool called Passatutto[11] that estimates the FDR to allow for conscious selection of scoring parameters for a given MS/MS untargeted metabolomics dataset.

The following three target-decoy approaches have been tested by the authors in hopes to find the best FDR controlling method: naive, spectrum-based and fragmentation tree-based methods.

- Naive method randomly adds fragment ions from the reference library to the decoy spectrum. This process

is repeated until the decoy spectrum resembles the library spectrum with the number of fragment ions.

- Spectrum-based method adds a precursor fragment ion of the target spectrum and chooses all other spectra that contain that fragment ion (within a small mass range). The final spectrum is created by adding random fragment ions from the chosen spectra.

- Fragmentation tree-based approach creates a fragmentation tree from the real spectra. To create a decoy spectrum, an internal node of the original tree is selected as a new root with the molecular formula of the precursor ion. Molecular formulas are then calculated following the edges of the tree and subtracting losses. If this process results in impossible molecular formulas, the entire subtree is moved to another tree branch. The new root node is then calculated with relative probability 1/(no. of edges to be re-grafted+1).

All these methods use the intensities of the original fragment ion.

Given spectra were searched against a real database and decoy database. The evaluation of all methods described has been carried out by estimating q-values. In all of these estimates, it was necessary to know the true identity of all query spectra. The estimated q-values were compared with the true ones. The authors also included the impact of noise filtering on the quality of the FDR estimation.

All methods overestimated significance which means the estimated q-values were smaller than true q-values. The fragmentation tree-based approach performed best, however it is important to note that all methods required rigorous noise-filtering to be able to perform well enough to estimate the FDR. Further comparison of these techniques with the current work is discussed in Section 7.1.

Another work presented by Wang et al.[12] focused on a different approach for creating a decoy database to estimate the FDR metabolite identification. The authors claim that the existing methods cannot be directly applied to spectral library independent database search, hence the proposal of the new approach. The decoys are created through violating the octet rule by adding small odd numbers of hydrogen atoms yielding invalid formulas and structures. They mimic mass distribution of targets well enough resulting in the same possibility of a match as the compounds in the target database. The new approach is statistically validated by generating a falsified null dataset (shifting all precursor ion mass by 4.5 Da) and two widely used metabolite identification tools. The validation was followed by the FDR estimation. The results claim that this approach is effective in evaluating the confidence of metabolite identification.

3.4.2 Knockoffs

In this section we present another decoy creation approach that has not been applied to metabolomics. To be able to apply this method to mass spectra, we need to understand the main statistical concepts behind this procedure. We present its implementation under Section 3.3.3.

There have been many attempts on creating procedures that accurately control the FDR in various settings[24][25]. Procedures that could select variables corresponding to real effects of discoveries and that could be reproduced in future experiments. The empirical performance of the knockoff procedure demonstrates excellent power and effective FDR control in comparison to other methods[15].

The important concept that presents knockoffs as a powerful tool, is that they are able to preserve the same correlation between fake features ($\tilde{X}_1$ and $\tilde{X}_2$) as in the original features (X_1 and X_2). At the same time ensuring that these fake features are not identical to the original features because we could not make new discoveries otherwise. Therefore knockoffs imitate the original feature correlation structure well enough to be described as a tool for calibrating test statistics in order to control the FDR.

The (unknown) variables of interest are the ones which unambiguously affect the outcome with as few false positives as possible so that the experiments can be reproducible. The variable is of interest if it changes the distribution conditioned on what is already known. Given many variables $(X_1 ... X_n)$, we need to find the ones $P(Y|X)$ depends on.

Formally:

$$P(response|variable, others) \neq P(response|others).$$

Once we create the fake features, we can treat them as a control group to categorise the selection procedure. More explicitly, given 500 original features and 500 knockoff features, if an algorithm chooses 35 original features and 16 knockoff features, we could assume there are approximately 16 false positives amongst 35 original features.

Important characteristics of knockoffs

- Pairwise exchangeability
  Given a set of features, if we generate knockoffs (fake features), after swapping the real features with their knockoffs, we should expect the distribution to remain unchanged (i.e. the joint distribution should be invariant). Naturally, this implies that the fake features have the same distribution as their corresponding features. The $\tilde{X}$ is dependent on X. Visually:

  $$(X_1, X_2, \tilde{X_1}, \tilde{X_2}) \overset{d}{=} (X_1, \tilde{X_2}, \tilde{X_1}, X_2)$$

  Other methods, such as permutation-based methods, do not keep correlations between features and fake features[26].

- $\tilde{X} \perp\!\!\!\perp Y|X$ if there is a response Y.
  There is no assumption on the relation between the response Y and variables X, these variables do not provide any information about Y beyond what is already known. Other methods make assumptions on $P(Y|X)$, the distribution of Y being conditional on X, but knockoffs are generated independently of Y. However, we still need to be able to specify the distribution of the feature variables[27] (denoted as P_X).

3.4.3 Gaussian knockoffs and mass spectra

A Gaussian mixture is a function comprised of several Gaussians. These Gaussians are called components. They can be defined as the number of clusters for our dataset. Each Gaussian component has its mean μ defining its centre, a covariance Σ defining its width and a mixing probability (π) defining the size of the Gaussian function (must be always between 0 and 1). More formally a Gaussian density function can be represented as:

$$N(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Note the above formula only represents one Gaussian density function. In terms of a Gaussian mixture, we have a number of Gaussian density function that are fitted based on the number of components chosen.

Given a spectrum represented as one of the points in our Gaussian mixture, to be able to create knockoffs, we need to calculate the probability of that point coming from each of the Gaussian density functions.

Implementation of Gaussian knockoffs for mass spectra

To be able to create Gaussian knockoffs, we need embedded spectra so that they are represented as n-dimensional vectors. We need to be able to choose an appropriate number of Gaussian mixture components. This choice will depend on the number of spectra and how variable they are (i.e. how far from one another they are in latent space). Each component should have its own mean and diagonal variance matrix defined. For our future knockoff of each embedded spectrum, we use the mean and the variance of the component according to the posterior probability for the given spectrum. We also need to choose the diagonal matrix (D) in order to ensure the joint covariance matrix $\begin{pmatrix} \Sigma_k & \Sigma_k - D \\ \Sigma_k - D & \Sigma_k \end{pmatrix}$ is positive definite. We then calculate the knockoff's mean and variance following the Gimenez et al. approach[16], and take a single sample from the resulting distribution[28].

Figure 3 visualises the representation of spectra in 2 dimensions with 3 Gaussian density functions fitted.

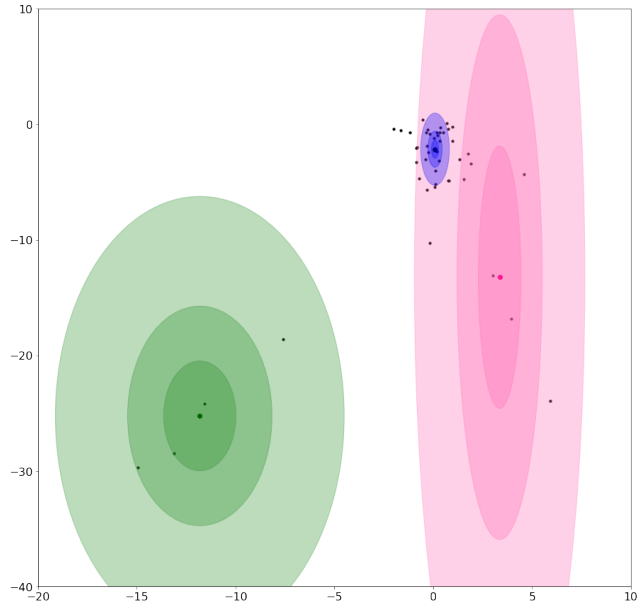

Figure 3: A Gaussian mixture model with 3 components fitted to a subset of a LC-MS GNPS dataset (described in Section 4.3) representing 46 spectra each embedded in 2 dimensions (vectors of size 2). Sparsely distributed points represent mass spectra. Each Gaussian density function (green, blue, pink ellipses) has its mean highlighted by a bigger point in the colour of the corresponding density function. The ellipses represent the multiples of the respective covariances ($\Sigma/2$, Σ, 2Σ – ordered from darkest to lightest shade).

4. METHODS

4.1 Conventional decoy methods

To recreate the study by Scheubert et al.[11], we used the MassBank dataset collected on an Orbitrap instrument for our query provided by the authors and the unfiltered version of the GNPS library. The MassBank dataset consisted of 458 spectra, the GNPS unfiltered dataset consisted of 4,138 spectra. To be able to assess the technique appropriately, we also used two decoy datasets provided by the authors: "GnpsDecoyRandomPeaks" and "GnpsDecoyConditionalPeaks" representing the naive and spectrum-based decoy methods respectively. These decoy datasets were created using the same GNPS library, hence also containing 4,138 spectra.

To be able to process the ".ms" files, a separate parser was created, which is available here: `https://github.com/nitrozyna/FDR-Metabolomics/blob/main/src/passatutto_parser.py`. We used the `matchms` Python package[29] to convert all the JSON objects to Spectrum objects to ease our downstream analysis. To be able to identify matching query and library spectra, we used the `matchms.similarity.CosineGreedy` method with tolerance of 0.005. This method uses a 'greedy' approach to quantify the similarity between two mass spectra. Two peaks were considered a potential match if their m/z ratios lied within the given tolerance. The highest score was then recognised as a miss/false hit, when the first InChI layer did not match (i.e. main layer: chemical formula, atom connections and hydrogen atoms) and as a true hit when the main layer matched. Computationally, as InChI is just a string with the delimiter "/", we split the InChIs on the first four parts of the string for both spectra and compared them. As an optimisation for the process of obtaining true/false hits, we only searched for queries that are within ±3 ppm of the precursor mass knowing we do not need to search the entire library dataset to match our query as differing precursor masses always result in different compounds.

Using the same parameters for the cosine scoring function and the same optimisation technique, we defined the hits for naive and spectrum-based decoys. The only difference being the identification of a hit as true/false/decoy and not true/false.

Having all three lists of hits, we then calculated true q-value scores for the query and library hits, estimated q-value scores for the naive decoy hits and query and estimated q-value scores for the spectrum-based decoy hits and query. The important part to note in the q-value calculation is that the true q-value is only increasing when we obtain a false hit. Whereas for the estimated q-value calculation (decoys), the q-value keeps increasing only when we obtain a decoy hit (i.e. not when we obtain a false hit).

We then plotted the resulting estimated vs true q-values for each decoy method visible in Section 7.1 (Figure 5).

4.2 Knockoffs in comparison with conventional decoys

In this experiment we used the same datasets as above. The two main differences in this experiment are the representation of decoys as embedded vectors and the comparison of knockoff-based approach with the embedded naive decoy and spectrum-based decoy methods.

The only additional pre-processing that needed to be applied in order to represent all the spectra appropriately in the latent space was normalization. We achieved it by using the `normalize_intensities` method from `matchms.filtering`. All the library, query and decoy spectra were then converted to documents using `spec2vec.SpectrumDocument`. Every peak was converted into a string "word" in the form `'peak@XXX.XX'` where `XXX.XX` represents the m/z value of that peak. For each spectrum peak, there was also a loss "word" calculated ($precursor_{m/z} - peak_{m/z}$). Combined list of peaks and losses results in a document. All the documents were converted into vectors represented in 75 dimensions using `spec2vec.calc_vector` function. As `calc_vector` requires a trained Word2Vec model (Spec2Vec similarity scores are derived on the basis of a pre-trained Word2Vec model), we used a model that was trained on a larger LC-MS dataset (See Section 6). We then looked for spectra that were within ±3 ppm of the precursor mass to keep the conditions identical to the previous experiment. Vector embeddings for query vs library, query vs naive decoys and query vs spectrum-based decoys were then compared using Spec2Vec `cosine_similarity_matrix` which is an efficient implementation of computing the cosine similarity with multiple vectors simultaneously. The most important part of this implementation is precomputing the query embeddings, since computing them in place every time when comparing to a library/decoy takes a significant amount of time. The highest score was then recognised as a true/false hit following the same logic as described in the previous experiment. We created the true q-value list and two estimated q-value lists for the two decoy methods using the same procedure.

4.2.1 Creation of knockoffs

To be able to create knockoffs, we used the embedded library spectra and fitted N Gaussian mixture components with each component having its own mean and diagonal variance matrix. We used grid-search to find the optimal number of components. In this experiment N=3 and corresponding diagonal matrix D=110. After obtaining knockoffs, we calculated the estimated q-value list in the same manner as with other decoy methods.

We then plotted the resulting estimated vs true q-values for each decoy method adding to the results from the previous experiment (Figure 6).

4.3 Knockoff technique for LC-MS datasets

To present the general Gaussian knockoff creation, we chose a subset of a large LC-MS dataset provided on GNPS, available at:
https://gnps-external.ucsd.edu/gnpslibrary/ALL_GNPS.json. The subset has been created by Huber et al. All of the spectra present in the subset are with positive ionization mode and have been cleaned and processed using `matchms`[10].

The data contains 112,956 spectra, out of which 92,954 have InChIKey annotations, we filtered the dataset based on these. We then removed spectra with duplicate peaks (spectra that have exactly the same peaks when rounded to 2 decimal points). These would always give us a 100% match, which would only pollute the results, and not allow for any new discoveries. This filtering resulted in 89,517 spectra. InChIKey strings are 27 characters long, divided into three parts connected by hyphens. We created a mapping of the first 14 characters (the connectivity and the proton layers) to spectra, in order to identify matching spectra. This procedure let us divide the dataset into the library and query. The mapping resulted in 81,392 single spectra and 8,125 multiple spectra assigned to the first 14 characters of the InChIKeys. To ensure a possible match from each of the multiple spectra assigned, we randomly chose one for the query, therefore allowing for at least one possible match in the library. The chosen size of the query was 1,000 spectra. All the other single spectra have been added randomly as noise to fill in the rest of the library size (88,517 spectra). We embedded all the spectra appropriately, normalizing the intensities. We then looked for spectra that were within ±3 ppm of the precursor mass of each query and calculated the similarity scores. The highest score was then recognised as a true/false hit following the same logic as described in the previous experiments. We created the true q-value list.

4.3.1 Creation of knockoffs

We followed the same procedure for creating knockoffs as in the previous experiment, but our grid-search found a different optimal number of components and corresponding diagonal matrix, equal to 5 and 45 respectively.

After obtaining knockoffs, we calculated the estimated q-value list.

We then plotted the resulting estimated vs true q-values for our knockoff method (Figure 7).

4.4 Knockoff technique for GC-MS datasets

To present knockoffs for GC-MS metabolomic datasets, we used the "MoNA-export-GC-MS_Spectra.msp" file available under: https://mona.fiehnlab.ucdavis.edu/downloads. The dataset contains 18,898 spectra. After filtering (choice of minimum number of 5 peaks per spectrum, removal of spectra with duplicate peaks), we obtained 14,658 spectra. An additional preprocessing included normalization of intensities and rounding the spectra peaks to the nearest integer since these spectra are low resolution.

We created a mapping of the first 14 characters of the InChIKey to spectra, in order to identify matching spectra. The mapping resulted in 12,311 single spectra and 2,347 multiple spectra assigned to the first 14 characters of the InChIKeys. To ensure a possible match, from each of the multiple spectra assigned we randomly chose one for the query. The chosen size of the query was 1,000 spectra. All the other single spectra have been added randomly as noise to fill in the rest of the library size (13,658 spectra). We created documents for all the spectra and trained our model on all of them using 30 iterations and 75 dimensions following a similar grid-search procedure to the optimization experiment described in Section 5. We embedded all the spectra appropriately. We then looked for spectra that were within ±3 ppm of the 'ExactMass' parameter of each query and calculated the similarity scores. The highest score was then recognised as a true/false hit. We created the true q-value list.

4.4.1 Creation of knockoffs

We followed the same procedure for creating knockoffs as in the previous experiment, but our grid-search found a different optimal number of components and corresponding diagonal matrix, equal to 5 and 6 respectively.

After obtaining knockoffs, we calculated the estimated q-

value list.

We then plotted the resulting estimated vs true q-values for our knockoff method (Figure 8).

5. OPTIMIZATION OF SPEC2VEC DIMENSIONS

The default size parameter for training a model in Spec2Vec is 300. This means the model will be trained by embedding all spectra in 300 dimensions. One of our goals in this work was reducing the computational complexity for creating knockoffs and calculating similarity scores. We performed a set of experiments with ranging dimensions to be able to find the limit at which Spec2Vec Cosine Similarity function performs well in comparison with the `CosineGreedy` method from `matchms`. We tested a number of dimensions ranging from 5 to 300 and trained models according to parameters described in Section 6. We then repeated the process with a progressively narrowing range.

5.1 Methodology

1,000 randomly selected spectra that had at least 1 more corresponding InChIKey were removed from the GNPS dataset described in Section 4.3. They were then matched to the remaining 88,517 spectra. For both Spec2Vec Cosine Similarity and `CosineGreedy` from `matchms`, a spectrum was considered a candidate match, when the precursor m/z was lying within tolerance of ±1 ppm. We then picked the highest score from all candidates and defined if it is a true match or a false match by comparing the first 14 characters of the InChIKeys. It is important to note that cosine tolerance was chosen at the level of 0.005 and Spec2Vec tolerance (number of decimal places at which spectra are embedded) was equivalent to 2 for a fair comparison. We also tuned the `min_match` parameter for `CosineGreedy` to 6 peaks, which means we assigned a score of 0 to each pair of spectra that match on fewer than 6 peaks.

The resulting comparison of true-positives vs false-positives is present on Figure 4.

We can notice the performance of Spec2Vec similarity score to be noticeably better in all cosine thresholds compared to the conventional cosine similarity. Please note that under the same parameters, apart from dimensionality parameter, Spec2Vec performed better for Huber et al.[10] but `CosineGreedy` performed similarly. This change is understandable as the dimensions for our experiment were significantly reduced (from 300 to 75).

Further details are present in the `dimension_optim notebook` available here: `https://github.com/nitrozyna/FDR-Metabolomics/blob/main/src/dimension_optim.py`.

6. TRAINED MODELS

Spec2Vec is an unsupervised method, hence it is possible to train a model on the same data the model is later applied to. The model for GNPS LC-MS dataset was trained with 75 dimensions following the dimensionality experiment (See Section 5). The training has been done using `spec2vec.model_building.train_new_word2vec_model`.
The model was trained with a maximum of 30 iterations (upon investigation, the results plateau from around 10 to around 50 iterations and from then onwards the results of false/true positive hits worsen). We also used the default initial learning rate = 0.025 and the default learning rate decay = 0.00025 (the rate at which the learning rate is lowered with every iteration).

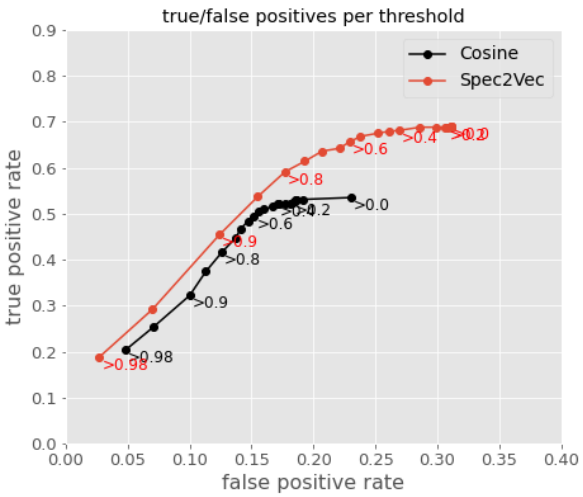

Figure 4: True/False positive rate per threshold of two similarity scoring functions: `matchms Cosine Greedy` and Spec2Vec `cosine_similarity_matrix`. For the Spec2Vec similarity all the spectra were embedded in 75 dimensions. The comparison was done on a subset of LC-MS GNPS library. The subset was divided into 1,000 query spectra and 88,517 library spectra and ran independently on two of the similarity scoring function.

7. RESULTS

7.1 Conventional decoy methods

This experiment intended to repcreate the current state-of-art target-decoy methods for controlling FDR in LC-MS metabolomic datasets. As visible on Figure 5, our estimated and true q-values for both naive and spectrum-based approaches allow for the FDR estimation where spectrum-based method would be preferable to choose. All methods tend to overestimate significance (estimated q-values are smaller than the true q-values), there is a significant shift towards overestimation for q-values greater than 0.2. It is worth noting that these datasets are unfiltered. These results significantly outperform results by Scheubert et al., however it may be attributed to a different cosine similarity implementation and an incomplete dataset (the authors mention averaging over 10 decoy libraries whereas only one was made available to us). The Boecker laboratory only obtains accurate FDR estimates after noise-filtering. It is important to note that we have not included the fragmentation-tree based decoy approach in our experiment which is the decoy method that performed best for the Boecker lab and was implemented in the Passatutto tool. This method applies noise-filtering by design, hence the appropriate comparison would be using the noise-filtered GNPS library. We do not recreate any noise-filtered results as the purpose of this work is to present the advantage of obtaining performant results with Spec2Vec even without rigorous filtering.

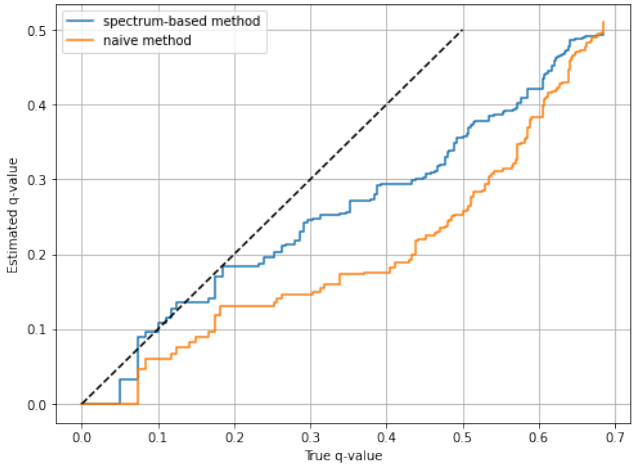

Figure 5: Recreation of two decoy methods presented by Scheubert et al.: naive and spectrum-based. Estimated (y-axis) vs true q-values (x-axis) using the Massbank dataset as query and the unfiltered GNPS dataset as library for calculation of the true q-values. The spectral similarity scoring function used was `CosineGreedy` from `matchms`.

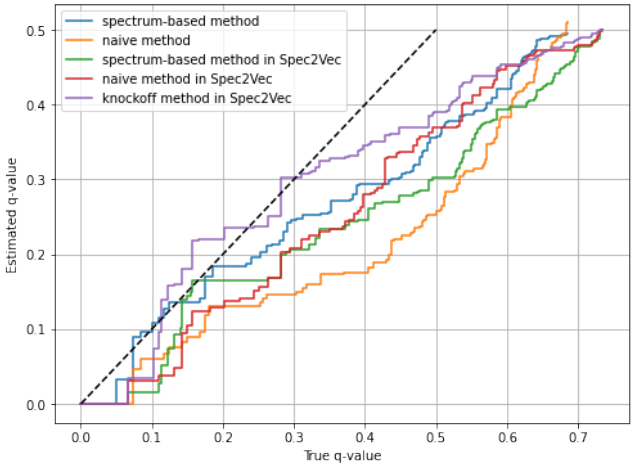

Figure 6: Estimated (y-axis) vs true q-values (x-axis) using the Massbank dataset as query and the unfiltered GNPS dataset as library for calculation of the true q-values. This Figure adds two decoy methods recreated from the Scheubert et al. study to Figure 5: naive and spectrum-based are embedded in 75 dimensions using Spec2Vec. Additionally, a new Gaussian knockoff-based decoy approach with a diagonal matrix D=110 and 3 mixture components is added. The spectral scoring function used was `cosine_similarity_matrix` for the three Spec2Vec decoy methods.

7.2 Knockoffs in comparison with conventional decoys

The results from this experiment (Figure 6) let us see the impact of spectral embeddings on the contemporary target-decoy approaches and a comparison with a knockoff-based technique. We can notice that just embedding the naive decoy spectra using Spec2Vec has resulted in a better performance to the previous experiment (Mean Absolute Error between estimated q-value and true q-value lowered from 0.170 to 0.148, resulting in a 13% improvement). Although they still tend to overestimate significance, this result is better than any of the unfiltered results presented by the Boecker group. This may suggest that presenting decoy techniques in latent space may bring benefits in terms of FDR control.

The cosine similarity for true q-value = 0.5 was 0.29. In practice, we would never be interested in the q-values above 0.5 (the number of false positives would exceed the number of true positives), as we do not know if the query compound is even present in the database. When filtering the results to cosine similarity score > 0.29, all methods result in closer similarity of estimated q-values to true q-values. Under these conditions, the knockoff-based approach outperforms any of the contemporary decoy techniques (See Table 1) for unfiltered datasets.

Table 1 shows the Mean Absolute Error and the MAE for true q-values < 0.5 for each decoy method used in this experiment. The lowest MAE was obtained by the spectrum-based method while the knockoff-based method followed closely after. We included the more realistic scenario where true q-values are capped at 0.5 which shows that the knockoff-based approach outperforms other decoy methods.

method	MAE	MAE (q_{true} < 0.5)
naive	0.170	0.119
spectrum-based	0.115	0.062
naive (S2V)	0.148	0.090
spectrum-based (S2V)	0.177	0.096
knockoff	0.118	0.038

Table 1: The Mean Absolute Error (MAE) and MAE for true q-values < 0.5 for recreated conventional decoy methods, conventional decoy methods embedded in Spec2Vec and knockoff-based approach. The following results were obtained using the Massbank dataset as query and the unfiltered GNPS dataset as library for calculation of the true q-values.

7.3 Knockoff technique for LC-MS datasets

From Figure 7 we can notice that the knockoffs for this dataset did not perform similarly to the previous experiment. There is a heavy overestimation of significance especially until true and estimated q-values reach 0.28. Although other choices for components and diagonal matrix performed better before reaching q-values of 0.28, in this scenario, it is important to consider the distribution of our assignment of cosine scores to spectra. Most of the spectra (75%) lie within the cosine range of 0.2 to 0.9 (Figure 9) - the sparse points representing individual spectra. The Mean Absolute Error between estimated and true q-values for the spectra that lied within the 75%, was 0.005 and the difference between the estimated and true q-values never exceeded 0.03. The reason for overestimation before reaching the q-values of 0.28 is that the true q-values increase drastically within that interval, meaning many false hits (24%) for cosine similarity scores being higher than 0.98.

Knockoffs started converging to real q-values around the cosine similarity score of 0.89 and kept matching the real q-values for decreasing scores. Upon closer investigation the compounds that obtain higher scores (e.g. 0.99996) and false identifications, are usually isomeric (e.g. Benzo[c]acridine,

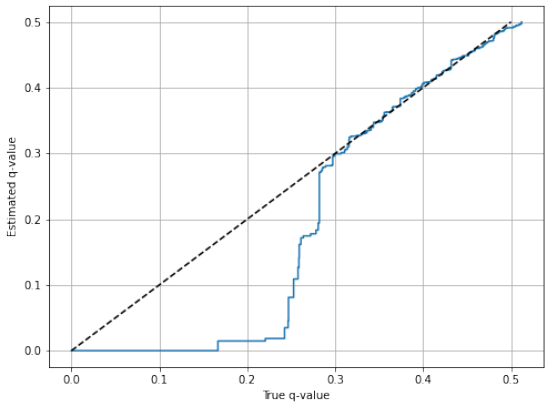

Figure 7: Estimated (y-axis) vs true q-values (x-axis) in the subset of the LC-MS GNPS dataset. The subset was divided into 1,000 query spectra and 88,517 library spectra. For the estimated q-values, the Gaussian knockoff-based decoys were created using 88,517 library spectra, diagonal matrix D=45 and 5 mixture components. All the spectra were embedded in 75 dimensions using Spec2Vec. The spectral scoring function used was `cosine_similarity_matrix` from Spec2Vec.

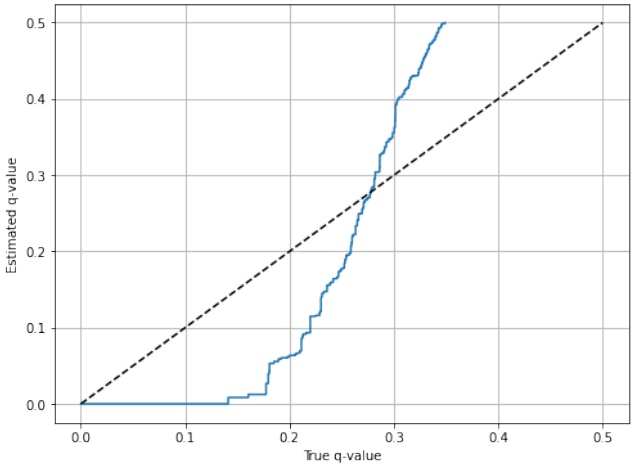

Figure 8: Estimated (y-axis) vs true q-values (x-axis) in the GC-MS Mona library. The dataset was divided into 1,000 query spectra and 13,658 library spectra. For the estimated q-values, the Gaussian knockoff-based decoys were created using all the library spectra, diagonal matrix D=6 and 5 mixture components. All the spectra were embedded in 75 dimensions using Spec2Vec. The spectral scoring function used was `cosine_similarity_matrix` from Spec2Vec.

Benz[a]acridine). It is also important to note that this dataset is quite noisy as it contains 5081 pairs of spectra that have the same peaks (i.e. cosine score of 1.0) and 823 of these pairs have been assigned different InChIKeys. Based on this outcome, we can assume the knockoff-based technique is a reasonable approach for the FDR-control.

7.4 Knockoff technique for GC-MS datasets

Our knockoff technique did not perform well enough to estimate the FDR for the GC-MS dataset (Figure 8). The true q-value list increases rapidly with 14% false hits starting at cosine similarity score above 0.98. Upon closer investigation the compounds that obtain higher scores and false identifications, are usually isomeric (e.g. 2,6-Dimethylphenol and 2,5-Dimethylphenol resulting in similarity score 0.996). We obtained the first decoy match at a cosine score of 0.96. Our method overestimated significance until q-values reached 0.28 and underestimated from that point onwards.

7.5 Knockoff decoy variability

As visible on Figure 10, the knockoff procedure follows the real FDR quite closely from below cosine similarity score of 0.89. The variability of running the Gaussian knockoff-based technique 10 times independently was negligible, hence we can assume the technique to be relatively stable. The true FDR has been obtained by knowing the identity of all compounds, hence this is the FDR we would most likely obtain using Spec2Vec similarity scoring function. In reality we would never know the true identity of our compound, therefore it is preferable for us to have a method that would let us imitate it. The knockoff-based FDR closely follows the true FDR, hence it can be recognised as a method that is able to accurately estimate it.

8. CONCLUSIONS

We presented three experiments in this study. The first one comparing the conventional decoy methods by Scheubert et al. with a knockoff-based approach on the same dataset. The second one, with a larger LC-MS dataset to present the general Gaussian knockoff-based approach. The third testing the Gaussian knockoff-based approach with a GC-MS metabolomic dataset.

We presented an improvement to the current target-decoy naive method by embedding the decoys in latent space using Spec2Vec. In addition, we presented a new target-decoy knockoff-based method that is able to estimate the FDR in LC-MS metabolomic datasets. We also included a possibility of creating knockoffs for GC-MS datasets, which given these settings does not allow for FDR estimation. However, we should investigate this method further, as this change may be attributed to the quality of the given dataset. Moreover, we have optimised the knockoff technique by calibrating the dimensions in which the Spec2Vec embedded spectra should be presented while keeping the quality of the true/-false assignments similar to the contemporary cosine similarity techniques. All of this work opens a new outlook at assigning significance to metabolomic datasets.

9. FUTURE WORK

Unfortunately, knockoffs are not fully deterministic as running the procedure several times may result in selecting different features and possibly leading to different conclusions. Although rerunning the same procedure several times for our experiment did not affect the final q-value plots significantly (Section 7.5, Figure 10), addressing this instability may be important to be able to reproduce the results. There have been some advancements in this area of research with an algorithm based on entropy maximization for generating Gaussian multi-knockoffs which improves the stability and has been shown to be more powerful compared to single

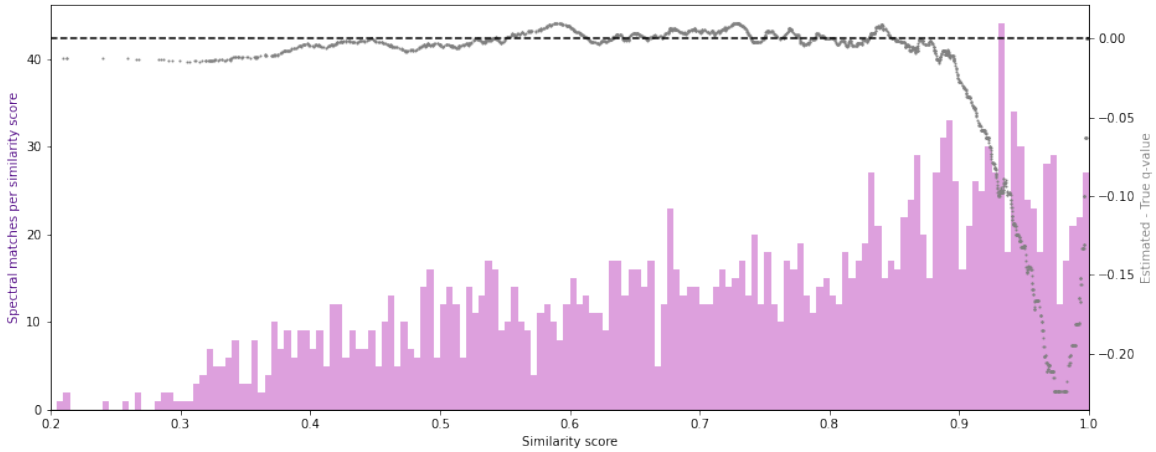

Figure 9: Difference between the estimated and true q-values (y-axis right side), each spectrum represented as a gray point. Number of spectral matches per similarity score represented as a vertical bar graph (y-axis left side) for the subset of LC-MS GNPS library using Spec2Vec similarity threshold. The subset was divided into 1,000 query spectra and 88,517 library spectra. For the estimated q-values, the Gaussian knockoff-based decoys were created using 88,517 library spectra, diagonal matrix D=45 and 5 mixture components. All the spectra were embedded in 75 dimensions using Spec2Vec. The spectral scoring function used was `cosine_similarity_matrix` from Spec2Vec.

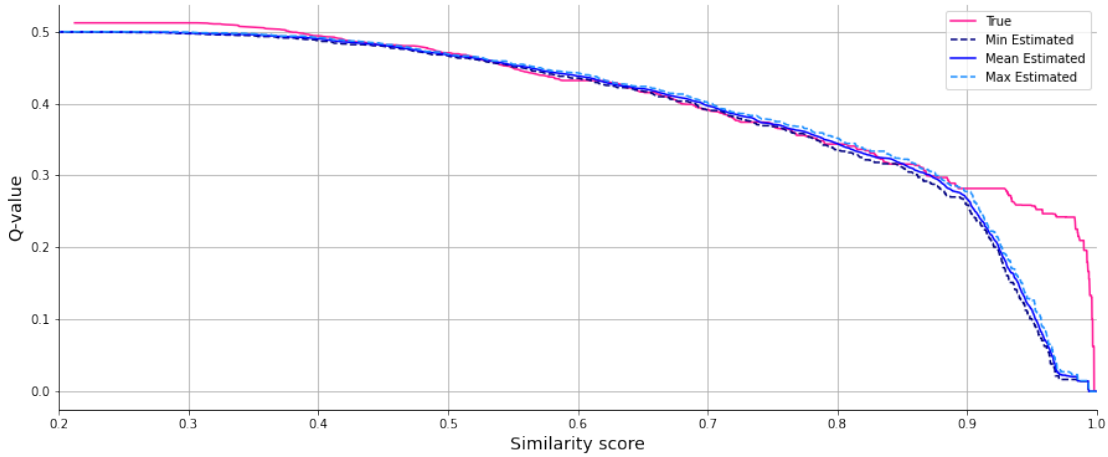

Figure 10: Representation of the Gaussian knockoff-based procedure variability. The Gaussian knockoff-based decoys were created using 88,517 library spectra, diagonal matrix D=45 and 5 mixture components. All the spectra were embedded in 75 dimensions using Spec2Vec. The spectral scoring function used was `cosine_similarity_matrix` from Spec2Vec. The 'Mean Estimated' line presents the average of running the procedure 10 times, the dashed lines below and above the average line present the minimum and maximum estimated q-values obtained.

knockoffs[30]. Implementing this method for knockoff creation for mass spectra could improve the current technique and allow for better reproducibility.

10. CODE AVAILABILITY

All the experiments from this paper can be replicated using the following notebooks in this repository:
`https://github.com/nitrozyna/FDR-Metabolomics`
Conventional decoy methods
`/notebooks/passatutto.ipynb`
Knockoffs in comparison with conventional decoys
`/notebooks/knockoffs_passatutto.ipynb`
Knockoff technique for LC-MS datasets
`/notebooks/knockoffs_lcms.ipynb`
Knockoff technique for GC-MS datasets
`/notebooks/knockoffs_gcms.ipynb`

11. REFERENCES

[1] O. Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2):155–171, Jan 2002.

[2] Wout Bittremieux, Christopher Chen, Pieter C. Dorrestein, Emma L. Schymanski, Tobias Schulze, Steffen Neumann, Rene Meier, Simon Rogers, and Mingxun Wang. Universal MS/MS visualization and retrieval with the metabolomics spectrum resolver web service. May 2020.

[3] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, et al. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, July 2010.

[4] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology*, 34(8):828–837, August 2016.

[5] Carlos Guijas, J. Rafael Montenegro-Burke, Xavier Domingo-Almenara, et al. METLIN: A technology platform for identifying knowns and unknowns. *Analytical Chemistry*, 90(5):3156–3164, January 2018.

[6] Tobias Kind and Oliver Fiehn. *BMC Bioinformatics*, 7(1):234, April 2006.

[7] Maureen Kachman, Hani Habra, William Duren, Janis Wigginton, Peter Sajjakulnukit, George Michailidis, Charles Burant, and Alla Karnovsky. Deep annotation of untargeted LC-MS metabolomics data with binner. *Bioinformatics*, 36(6):1801–1806, October 2019.

[8] Nathaniel G. Mahieu and Gary J. Patti. Systems-level annotation of a metabolomics data set reduces 25000 features to fewer than 1000 unique metabolites. *Analytical Chemistry*, 89(19):10397–10406, September 2017.

[9] Ke-Shiuan Lynn, Mei-Ling Cheng, Yet-Ran Chen, Chin Hsu, Ann Chen, T. Mamie Lih, Hui-Yin Chang, Ching jang Huang, Ming-Shi Shiao, Wen-Harn Pan, Ting-Yi Sung, and Wen-Lian Hsu. Metabolite identification for mass spectrometry-based metabolomics using multiple types of correlated ion information. *Analytical Chemistry*, 87(4):2143–2151, January 2015.

[10] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. van der Hooft. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):e1008724, February 2021.

[11] Kerstin Scheubert, Franziska Hufsky, Daniel Petras, Mingxun Wang, Louis-Félix Nothias, Kai Dührkop, Nuno Bandeira, Pieter C. Dorrestein, and Sebastian Böcker. Significance estimation for large scale metabolomics annotations by spectral matching. *Nature Communications*, 8(1), November 2017.

[12] Xusheng Wang, Drew R. Jones, Timothy I. Shaw, Ji-Hoon Cho, Yuanyuan Wang, Haiyan Tan, Boer Xie, Suiping Zhou, Yuxin Li, and Junmin Peng. Target-decoy-based false discovery rate estimation for large-scale metabolite identification. *Journal of Proteome Research*, 17(7):2328–2334, May 2018.

[13] Andrew Palmer, Prasad Phapale, Ilya Chernyavsky, Regis Lavigne, Dominik Fay, Artem Tarasov, Vitaly Kovalev, Jens Fuchser, Sergey Nikolenko, Charles Pineau, Michael Becker, and Theodore Alexandrov. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods*, 14(1):57–60, November 2016.

[14] Lukas Käll, John D. Storey, Michael J. MacCoss, and William Stafford Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research*, 7(1):29–34, January 2008.

[15] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection, December 2017.

[16] Jaime Roquero Gimenez, Amirata Ghorbani, and James Zou. Knockoffs for the mass: new feature importance statistics with false discovery guarantees, May 2019.

[17]

[18] Jennifer E. Schollée, Emma L. Schymanski, Michael A. Stravs, Rebekka Gulde, Nikolaos S. Thomaidis, and Juliane Hollender. Similarity of high-resolution tandem mass spectrometry spectra of structurally related micropollutants and transformation products. *Journal of The American Society for Mass Spectrometry*, 28(12):2692–2704, September 2017.

[19] Antoni Aguilar-Mogas, Marta Sales-Pardo, Miriam Navarro, Roger Guimerà, and Oscar Yanes. iMet: A network-based computational tool to assist in the annotation of metabolites from tandem mass spectra. *Analytical Chemistry*, 89(6):3474–3482, March 2017.

[20] J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira, and P. C. Dorrestein. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences*, 109(26):E1743–E1752, May 2012.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*,

volume 26. Curran Associates, Inc., December 2013.

[22] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, January 1995.

[23] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, August 2002.

[24] Xie Y, Pan W, and Khodursky Arkady B. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, 21(23):4280–4288, September 2005.

[25] Winston Haynes. Benjamini–hochberg method. *Encyclopedia of Systems Biology*, 2013.

[26] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5), October 2015.

[27] Rina Foygel Barber, Emmanuel J. Candès, and Richard J. Samworth. Robust inference with knockoffs, February 2019.

[28] Małgorzata Kurkiewicz. FDR-Metabolomics. `https://github.com/nitrozyna/FDR-Metabolomics/blob/main/src/knockoffs.py`, 2021.

[29] Florian Huber, Stefan Verhoeven, Christiaan Meijer, Hanno Spreeuw, Efraín Castilla, Cunliang Geng, Justin van der Hooft, Simon Rogers, Adam Belloum, Faruk Diblen, and Jurriaan Spaaks. matchms - processing and similarity evaluation of mass spectrometry data. *Journal of Open Source Software*, 5(52):2411, August 2020.

[30] Jaime Roquero Gimenez and James Zou. Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization, May 2019.