# Reactomics: Using mass spectrometry as a chemical reaction detector

Miao Yu[1], Lauren Petrick[1*]

[1] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, United States
*Corresponding author: Email: lauren.petrick@mssm.edu Phone: +1-212-241-7351. Fax: +1-646-537-9654.

## Abstract

Chemical reactions among small molecules enable untargeted metabolomics analysis, in which small molecules within tissue samples are identified through high-throughput assays. In standard mass spectrometry-based metabolomics, first significant small molecules are identified, then their biochemical relationships are probed to reveal biological fate (environmental studies) or biological impact (physiological response). However, we propose that biochemical relationships could be directly retrieved through untargeted high-resolution paired mass distance (PMD), which investigates chemical pairs in the samples without a priori knowledge of the identities of those participating compounds. We present the potential for this chemical reaction detector, or 'reactomics' approach, linking PMD from the mass spectrometer to biochemical reactions obtained via data mining of known small molecular metabolites/compounds and reaction databases. This approach encompasses both quantitative and qualitative analysis of reaction by mass spectrometry, and its potential applications include PMD network analysis, source appointment of unknown compounds, and biomarker reaction discovery instead of compound discovery. Such applications may promote novel biological discoveries that are not currently possible with classical chemical analysis.

## Introduction

Metabolomics or non-targeted analysis using high-resolution mass spectrometry is one of the most popular analytic methods for unbiased measurement of organic compounds (1, 2). A typical metabolomics sample analysis workflow follows a metabolite detection, statistical analysis, and annotation/identification of compounds using MS/MS and/or authentic standards. However, annotation or identification of unknown compounds is time consuming and sometimes impossible, which may limit biological interpretation (3). Through MS/MS, experimentally obtained fragment ions of the chemical of interest can be matched to a mass spectral database (4), but many compounds remain unreported and therefore unmatchable. Alternately, rules- or data mining-based prediction of in silico fragment ions is successful in many applications (2, 5), yet these approaches are prone to overfitting the known compounds. Finally, the final validation step requires commercially available or synthetically generated analytical standards for unequivocal identification, but such standards may not be available for all compounds. In this case, the workflow of compounds identification is always biased towards known compounds, and biological information from unknown compounds is not fully used.

Biochemical knowledge, through the integration of known relationships between biochemical reactions (e.g., pathway analysis), could also help provide potential molecular structures for annotating unknown compounds (3). Such methods are readily used to annotate compounds by chemical class. For example, the referenced Kendrick mass defect (RKMD) can be used to predict lipid class by first identifying a lipid through a specific mass distance (14.01565 Da) then identifying specific mass distances of heteroatoms to further determine lipid class (6). Similarly, isotope patterns in combination with specific mass distances characteristic of halogenated compounds (e.g., +Cl/-H, +Br/-H) can be used to screen halogenated chemical compounds in environmental samples (7). For these examples, known mass relationships among compounds are used to annotate unknown classes of compounds, providing evidence that a general relationship based annotation has the potential to uncover unknown information from samples.

The most common relationships among compounds are chemical reactions. Substrate-product pairs in a reaction form by exchanging functional groups or atoms. In fact, almost all organic compounds originate from biochemical processes, such as carbon fixation (8, 9). As in base pairing of DNA (10), organic compounds follow biochemical reaction rules that thereby result in characteristic mass differences between paired substrates and their products. Our concept, paired mass distance (PMD), reflects such rules by calculating the mass differences of two compounds or charged ions. Mass distances can also directly reveal isotopologue information (11), adducts from a single compound (12), or adducts formed via complex in-source reactions (13). High-resolution mass spectrometry (HRMS) can directly measure such paired mass distances with the mass accuracy needed to provide reaction-level specificity. Therefore, HRMS has the potential to be used as a 'reaction detector' to enable reaction-level annotations. Such reaction level information from the samples will provide an evidence-based link between protein/enzyme level changes in the samples with compounds/metabolite level changes, providing additional biological information.

Here, we use multiple databases and experimental data to provide a proof-of-concept for using mass spectrometry as a reaction detector. We discuss potential applications of this approach, such as PMD network analysis which can be used to search for biologically related metabolites to a targeted compound of interest, source appointment which can be used to characterize unknowns as endogenous or exogenous, and biomarker reaction discovery which can be used to calculate reaction level changes as a predictor of disease.

## Definition

We first define a reaction PMD ($PMD_R$) using a theoretical framework. Then we demonstrate how a $PMD_R$ can be calculated using Kyoto Encyclopedia of Genes and Genomes (KEGG) reaction R00025 as an example (see equation 1). There are three KEGG reaction classes (RC00126, RC02541, and RC02759) associated with this reaction, which is catalyzed by enzyme 1.13.12.16.

Ethylnitronate + Oxygen + Reduced FMN <=> Acetaldehyde + Nitrite + FMN + Water [1]

In general, we define a chemical reaction ($PMD_R$) as follows:

$$S_1 + S_2 + \ldots + S_n <=> P_1 + P_2 + \ldots + P_m \ (n >= 1, m >= 1) \quad [2],$$

where S means substrates and P mean products, and m and n the number of substrates and products, respectively. A PMD matrix [M1] for this reaction is generated:

| | $S_1$ | $S_2$ | ... | $S_n$ |
|---|---|---|---|---|
| $P_1$ | $\|S_1-P_1\|$ | $\|S_2-P_1\|$ | ... | $\|S_n-P_1\|$ |
| $P_2$ | $\|S_1-P_2\|$ | $\|S_2-P_2\|$ | ... | $\|S_n-P_2\|$ |
| ... | ... | ... | ... | ... |
| $P_m$ | $\|S_1-P_m\|$ | $\|S_2-P_m\|$ | ... | $\|S_n-P_m\|$ |

[M1]

For each substrate, $S_k$, and each product, $P_i$, we calculate a PMD.
Assuming that the minimum PMD would have a similar structure or molecular framework between substrate and products, we select the minimum numeric PMD as compound PMD ($PMD_{Sk}$), of that reaction (Eq. 3).

$$PMD_{Sk} = \min(|S_k-P_1|,|S_k-P_2|,...,|S_k-P_m|) \ (1<=k<=n) \quad [3]$$

Then, the $PMD_R$ is defined as the set of substrates' PMD(s) (Eq. 4):

$$PMD_R = \{PMD_{S1}, PMD_{S2},...,PMD_{Sn}\} \quad [4]$$

For KEGG reaction R00025, $S_1$ is ethylnitronate, $S_2$ is oxygen, $S_3$ is reduced FMN, $P_1$ is acetylaldehyde, $P_2$ is nitrite, $P_3$ is FMN, $P_4$ is water, m =3, and n=4. In the PMD matrix [M2] for this reaction, shown below, we define $PMD_{Ethylnitronate}$ = 27.023 Da, $PMD_{Oxygen}$ = 12.036 Da and $PMD_{Reduced\ FMN}$ = 2.016 Da.

| | Ethylnitronate | Oxygen | Reduced FMN |
|---|---|---|---|
| Acetaldehyde | 29.998 Da | 12.036 Da | 414.094 Da |
| Nitrite | 27.023 Da | 15.011Da | 411.120 Da |
| FMN | 382.080 Da | 424.115 Da | 2.016 Da |
| $H_2O$ | 56.014 Da | 13.979 Da | 440.110 Da |

[M2]

In our example, $PMD_R$ is 27.023 Da that is equivalent to the mass difference between two carbon atoms and three hydrogen atoms; $PMD_R$ is 12.036 Da for the additions of two carbon atoms and four hydrogen atoms and loss of one oxygen atom; and $PMD_R$ is 2.016 Da for the addition of two hydrogen atoms. One reaction can have multiple $PMD_R$, but no more than n $PMD_R$ has two notations: one is shown as an absolute difference of the substrate-product pairs' exact masses or monoisotopic masses with unit Da. Another notation is using elemental compositions. Here, we describe an elemental composition instead of chemical formula, because this notation also describes the gain and loss of elements, and therefore the neat mass change. In our example reaction, the $PMD_R$ can also be written as +2C3H, +2C4H/-O, and +2H. This elemental composition can be linked to known chemical processes retrieved from a reaction database, i.e., KEGG. For example, +2H represents the elemental composition change of a reaction involving a double bond breaking such as our KEGG example RC00126, and +2C3H indicates reaction with nitronate monooxygenase (EC:1.13.12.16) or reaction class RC02541. However, some elemental compositions, such as +2C4H/-H in our example, might not have a clear mechanism (e.g., no suggested KEGG reaction selection). By this definition, $PMD_R$ can be generated automatically in terms of elemental compositions or mass unit in Da.

## Results and discussion

### PMD analysis of a reaction database

To demonstrate the feasibility of using mass spectrometry as a chemical reaction detector, here we show that common and biologically relevant reactions can be written as $PMD_R$. KEGG, with 11262 reactions and 10213 unique formulas, was used as a reference reaction database (14). We calculated $PMD_R$ for all KEGG reactions and identified 2020 unique reactions (in Da, three decimal places). There are several common $PMD_R$ values; the 20 highest-frequency values covered 7712 KEGG reactions with frequency larger than 100. High-frequency PMDs were directly associated with similar biochemical reactions such as oxidation, breaking of double bonds, and phosphate transfer reaction (Table 1). This unique property of PMDs facilitates annotation of reaction class or reaction-associated enzymes between a pair of compounds without a priori knowledge of the identity of each compound. Furthermore, PMDs with low frequency can be used as biomarkers of unique reactions.

However, the reaction database also has limitations. For example, $PMD_R$ 17.003 Da is the mass distance between $H_2O$ and H(+), but could not be linked to any known reaction class or enzyme, making the $PMD_R$ difficult to explain biologically (Table 1). Some $PMD_R$ values involved in active phosphorus compounds such as ATP or ADP might not be detectable by mass spectrometry. Finally, the reaction database is limited to known KEGG reactions and therefore does not cover unknown reactions or $PMD_R$. In sum, however, PMD analysis of a reaction database provides inference on the reactions among compounds. Using the KEGG database as described, we generated a PMD database for reference annotation that is included in our open-source software package, pmd (https://yufree.github.io/pmd).

Table 1. Top 20 high-frequency KEGG reaction $PMD_R$ and corresponding example reaction, reaction class, and enzyme.

| **PMD (Da)** | Freq | Example reaction class | Example enzyme | Example reaction |
|---|---|---|---|---|
| 1.008 | 2029 | RC00001 | 1.1.1.42 | NAD(+) + succinate <=> fumarate + H(+) + NADH |
| 2.016 | 1732 | RC00095 | 1.3.1.84 | NAD(+) + propanoyl-CoA <=> acryloyl-CoA + H(+) + NADH |
| 15.995 | 1169 | RC00002 | 2.7.4.6 | ATP + GDP <=> ADP + GTP |
| 13.979 | 1128 | RC01658 | 1.13.12.22 | deoxynogalonate + $O_2$ <=> H(+) + $H_2O$ + nogalonate |
| 17.003 | 936 | NA[a] | NA | H2O + hypotaurine + NAD(+) <=> H(+) + NADH + taurine |

| | | | | |
|---|---|---|---|---|
| 79.966 | 729 | RC00002 | 3.6.1.3 | ATP + $H_2O$ <=> ADP + H(+) + phosphate |
| 14.016 | 594 | RC00060 | 1.5.3.1 | S-Adenosyl-L-methionine + Glycine <=> S-Adenosyl-L-homocysteine + Sarcosine |
| 0 | 532 | RC00302 | 5.1.1.3 | L-glutamate <=> D-glutamate |
| 162.053 | 365 | RC00049 | 3.2.1.23 | $H_2O$ + lactose <=> D-galactose + D-glucose |
| 18.011 | 359 | RC00331 | 4.3.1.17 | L-serine <=> 2-aminoprop-2-enoate + $H_2O$ |
| 0.984 | 344 | RC00477 | 3.5.4.17 | ATP + $H_2O$ <=> ITP + Ammonia |
| 1.032 | 262 | RC00006 | 1.4.1.2 | L-Glutamate + NAD+ + $H_2O$ <=> 2-Oxoglutarate + Ammonia + NADH + H+ |
| 159.933 | 243 | RC00634 | 4.2.3.10 | Geranyl diphosphate + $H_2O$ <=> (-)-endo-Fenchol + Diphosphate |
| 42.011 | 237 | RC00004 | 2.3.1.54 | Acetyl-CoA + Formate <=> CoA + Pyruvate |
| 12 | 228 | RC00738 | 1.13.11.27 | 3-(4-Hydroxyphenyl)pyruvate + Oxygen <=> Homogentisate + $CO_2$ |
| 27.995 | 176 | RC00292 | 3.5.4.16 | GTP + $H_2O$ <=> 7,8-Dihydroneopterin 3'-triphosphate + Formate |
| 43.99 | 168 | RC00626 | 1.1.1.42 | Oxalosuccinate <=> 2-Oxoglutarate + $CO_2$ |
| 31.99 | 136 | RC00388 | 1.13.11.1 | Catechol + Oxygen <=> cis,cis-Muconate |
| 177.943 | 131 | RC00637 | 4.2.3.15 | Geranyl diphosphate <=> Myrcene + Diphosphate |
| 42.01 | 106 | RC00070 | 3.6.1.20 | Acetyl adenylate + H2O <=> AMP + Acetate |

[a]: no associated reaction class to this PMD.

## Mass spectrometry as a reaction detector

### *Qualitative PMD analysis*

We propose that PMD analysis can be applied using mass spectrometry. Mathematically, a PMD of uncharged compounds is equivalent to the PMD of their charged species observed during mass spectrometry, as long as both compounds share the same adducts, neutral losses, and charges. In our example reaction, reduced FMN has a monoisotopic mass of 458.120265 Da while FMN has a monoisotopic mass of 456.104615 Da. Spectra from the human metabolome database (HMDB) (15) showed that common ions for reduced FMN and FMN using liquid chromatography (LC)-HRMS in negative mode are typically $[M-H]^-$ with m/z 457.1124 and 455.0968, respectively. The mass distance of the monoisotopic masses is 2.016 Da, as is the mass distance of the observed adducts. In cases such as this, mass spectrometry can be used to detect the PMD of paired compounds.

A challenge with HRMS is that for each analyte there is usually not a single ion, but redundant peaks that include various adducts, in-source fragments, neutral losses, and isotopes that are generated from the same analyte. For PMD analysis we assumed that compounds involved in paired biological reactions will generate the same type of redundant peaks in the mass spectra. To perform PMD analysis, we must first reduce the number of redundant peaks into independent peaks. Using the GlobalStd algorithm (12) or psuedu-spectra from annotation tools such as CAMERA (16) or RAMclust (17), a single peak representing the same type (adduct, neutral loss, or isotope) between paired analytes is selected for each cluster of redundant peaks. When the resulting filtered peaks are used for PMD analysis, they can then be linked to a specific biological reaction ($PMD_R$).

Once PMDs are calculated, linking the $PMD_R$ to specific elemental compositions will provide valuable biological context. However, annotation of the elemental compositions of certain PMD is dependent on high-resolution mass spectrometers. Low-resolution instruments that only measure nominal mass may not be specific enough to distinguish elemental compositions. For example, PMD 14 Da could be the addition or loss of a nitrogen atom or the addition of one oxygen atom and loss of two hydrogen atoms.

Here, we use HMDB (15) to compare low-resolution versus high-resolution measurements in determining elemental compositions. HMDB contains 114,100 compounds with 11,523 unique chemical formulas with known elemental compositions. PMD, as well as the elemental composition, was computed for the unique chemical formulas rounded to one, two, or three decimal places. Higher frequencies of a PMD are observed when rounding to fewer digits, suggesting the presence of false positives (Table 2). As confirmation of the annotation accuracy, we determined how many of the PMDs in Table 2 resulted from a change in chemical formula linked with the appropriate PMD for the range of reported decimal places (Table 3). For example, of the 4934 ion pairs with a PMD of 14.016 Da, > 97% of the pairs included an elemental change of +C2H. However, when two decimal places were reported, e.g., PMD of 14.02 Da, only 60% of the 8003 ion pairs included an elemental change of +C2H. For the top 10 PMDs, accuracy > 94% was observed when the PMDs were rounded to three decimals, only ≥ 51% when rounded to two decimal places, and < 10% when only 1 or 0 decimals were used (see

Table 3), confirming that high-resolution mass spectrometry is required for qualitative PMD analysis and elemental composition annotation.

Table 2: Frequency of PMDs calculated from compounds in HMDB with decreasing mass accuracy.

| PMD(digits = 3)* | Frequency | PMD(digits = 2) | Frequency | PMD(digits = 1) | Frequency | PMD(unit) | Frequency |
|---|---|---|---|---|---|---|---|
| 14.016 | 4934 | 14.02 | 8003 | 14.0 | 50419 | 14 | 156245 |
| 2.016 | 4909 | 2.02 | 7959 | 2.0 | 50467 | 2 | 156260 |
| 28.031 | 4878 | 28.03 | 7799 | 28.0 | 50797 | 28 | 155410 |
| 26.016 | 4229 | 26.02 | 7343 | 26.0 | 48517 | 26 | 154346 |
| 15.995 | 4214 | 15.99 | 7731 | 16.0 | 51278 | 16 | 155811 |
| 12.000 | 3861 | 12.00 | 7145 | 12.0 | 49335 | 12 | 155339 |
| 56.063 | 3861 | 56.06 | 6699 | 56.1 | 36417 | 56 | 151894 |
| 42.047 | 3771 | 42.05 | 6558 | 42.0 | 49808 | 42 | 153764 |
| 30.011 | 3698 | 30.01 | 6761 | 30.0 | 51241 | 30 | 154369 |
| 24.000 | 3689 | 24.00 | 6963 | 24.0 | 48099 | 24 | 154278 |

* Ten selected PMDs that occurred most frequently based on three decimal places.

Table 3: Effect of mass accuracy on elemental composition annotation accuracy of top-ten selected PMDs from Table 2.

| | PMD(digits = 3) | PMD(digits = 2) | PMD(digits = 1) | PMD(unit) |
|---|---|---|---|---|
| +C2H | 0.9755 | 0.6014 | 0.0955 | 0.0354 |
| +2H | 0.9703 | 0.5984 | 0.0944 | 0.0352 |
| +2C4H | 0.9783 | 0.6119 | 0.0939 | 0.0356 |
| +2C2H | 0.9775 | 0.5630 | 0.0852 | 0.0309 |
| +O | 0.9808 | 0.5346 | 0.0806 | 0.0307 |
| +C | 0.9826 | 0.5310 | 0.0769 | 0.0283 |
| +4C8H | 0.9653 | 0.5564 | 0.1026 | 0.0286 |
| +3C6H | 0.9737 | 0.5599 | 0.0737 | 0.0275 |
| +C2HO | 0.9440 | 0.5163 | 0.0681 | 0.0260 |
| +2C | 0.9810 | 0.5197 | 0.0752 | 0.0273 |

### *Quantitative PMD analysis*

In addition to qualitative analysis, peaks that share the same PMD can be summed and used as a quantitative group measure of that specific 'reaction' in the sample, thereby, providing a description of chemical reaction level changes in a sample without annotating individual compounds. There were two types of PMD across samples: static PMD in which intensity ratios between the pairs were stable across samples, and dynamic PMD in which the intensity ratios between pairs changed across samples. Only static PMDs, those with similar instrument response, can be used for quantitative analysis to avoid the complexity of changes from multiple peaks(see Table 4 for theoretical example). Similar to another non-targeted analysis (18), we suggest an RSD between quantitative pair ratios < 30% and a high correlation between the paired peaks' intensity ( > 0.6) to be considered a static PMD. We provide functions in the pmd package to determine static PMD.

Table 4. Demonstration of the selection of quantitative PMD pairs. Mass pair [A, B], [C,D], and [E,F] were involved in the same PMD. Only [A, B] and [E, F] are considered static PMD and suitable for quantitative analysis since their intensity ratios were stable across sample 1 and sample 2.

| | A[a] | B | Intensity ratio | C | D | Intensity ratio | E | F | Intensity ratio |
|---|---|---|---|---|---|---|---|---|---|
| Sample 1 | 100 | 50 | 2:1 | 100 | 50 | 2:1 | 30 | 40 | 3:4 |
| Sample 2 | 1000 | 500 | 2:1 | 10 | 95 | 2:19 | 120 | 160 | 3:4 |

[a] peak intensity of theoretical m/z.

While in-source reactions, mass accuracy, and stable paired mass intensity ratio are three important considerations for reaction-level qualitative and quantitative analysis via PMDs, the described tools and methodologies, namely, removal of redundant peaks, use of HRMS, and static PMD selection, can overcome these challenges to enable use of HRMS as a reaction detector.

**Reactomics**

We suggest 'reactomics' as a new approach to investigate reaction-level changes in biological and environmental samples and to link untargeted metabolomics data with biological processes. While we envision multiple possible applications for Reactomics, here we describe three examples. These applications are facilitated by our free and open-source software (pmd package,

version 0.1.6, https://yufree.github.io/pmd) for performing this analysis with annotation using both KEGG and HMDB databases.

***PMD network analysis***

PMD network analysis enables identification of metabolites associated with a known biomarker of interest. For example, links between high-frequency PMDs from the KEGG reaction database and a target analyte can be determined. Selected metabolites caffeine, glucose, bromophenol, and 5-cholestene were paired with other metabolites in the KEGG reaction database using the top-20 high-frequency PMDs from Table 1. Different topological properties (e.g., number of nodes, communities, etc.) of compounds' PMD network were observed for each selected target metabolite (Figure 1). Comparing these networks with known pathways may allow tentative annotation of unknown pathways. For example, an unknown compound with a similar topological structure as caffeine (see Figure 1), might have similar biological activity to caffeine.

In fact, PMD network analysis can also be used in combination with classic identification techniques, to enhance associated networks with targeted biomarkers. As proof-of-concept, we re-analyzed data from a published study to find the biological metabolites of exposure to tetrabromobisphenol A (TBBPA) in pumpkin (19, 20) using a local, recursive search strategy (see Figure 2). Using TBBPA as a target of interest, we searched for PMDs linked with debromination process, glycosylation, malonylation, methylation and hydroxylation, which are phase II reactions (e.g., primary metabolites) found in the original paper (19). The identified peaks with these PMDs were added to the network as secondary metabolites, and the process repeated until all PMDs and extensions were exhausted. Using PMD network analysis, we identified 22 unique m/z ions of potential TBBPA metabolites; 15 of these were unique ions not described in the original study. Most of the potential metabolites of TBBPA were found as higher-generation TBBPA metabolites (Figure 2), which are too computationally intensive to be identified using *in silico* prediction.

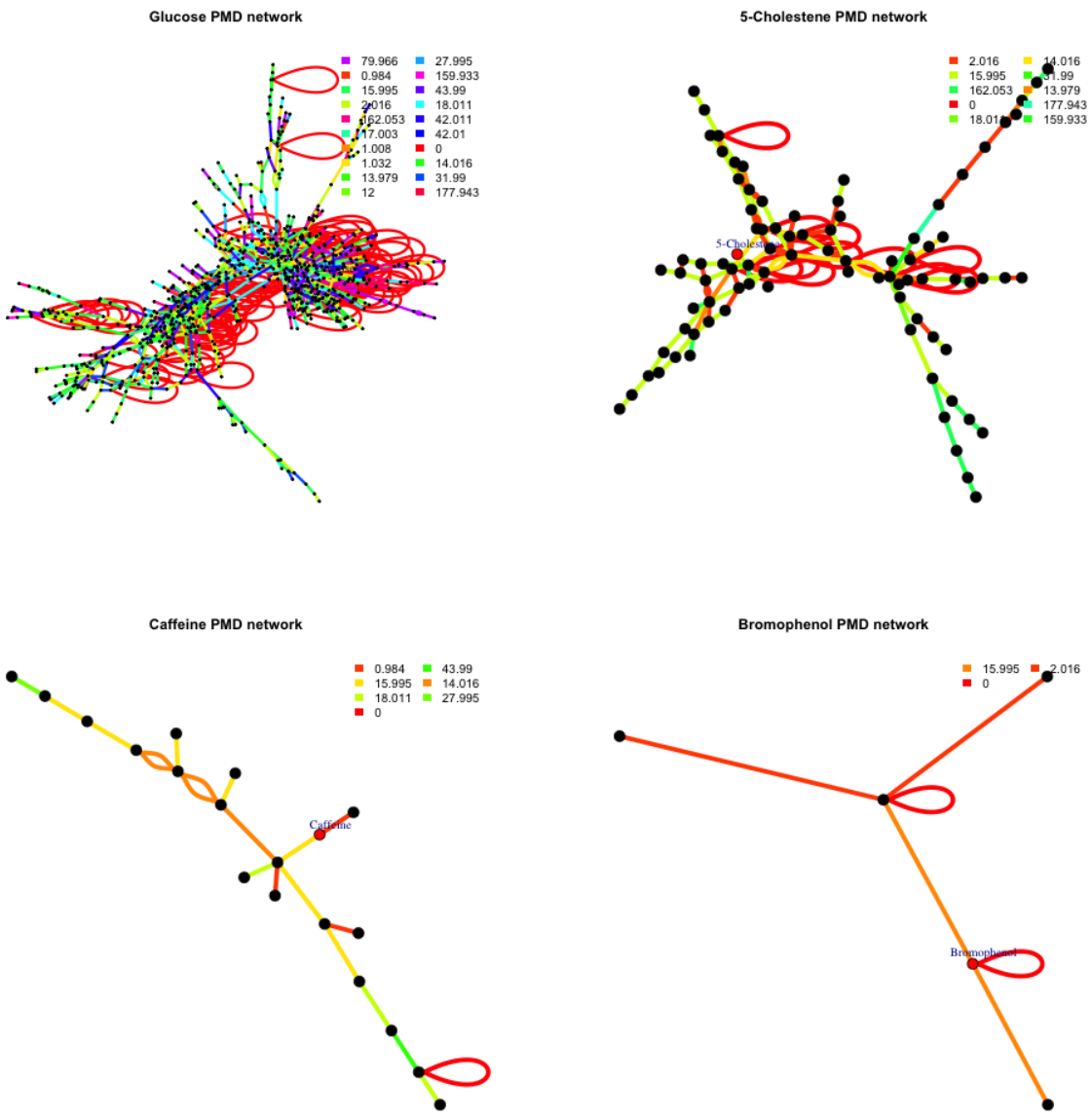


Figure 1. PMD networks for four selected compounds from the KEGG reaction database. Networks are limited to relationships with the 20 top-frequency PMDs.

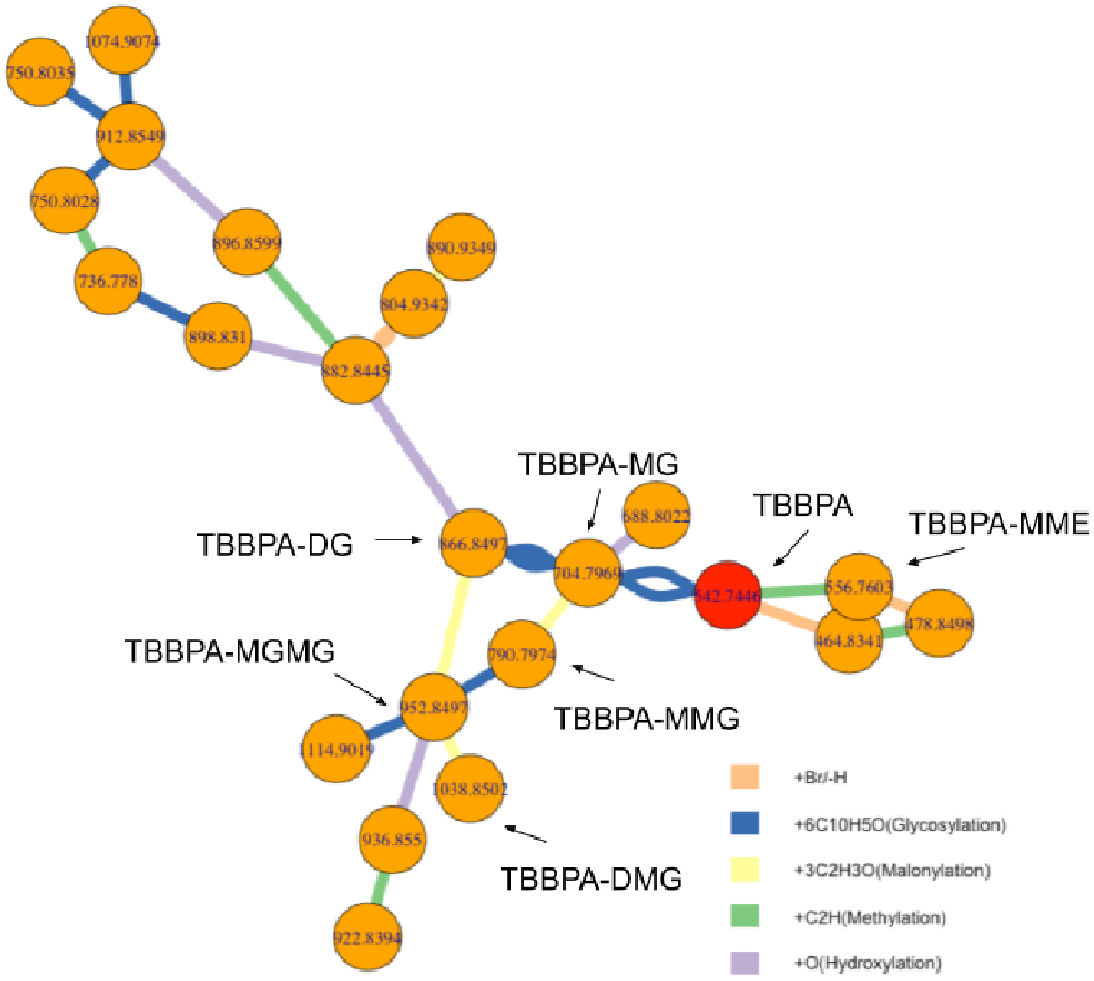


Figure 2. Metabolites of TBBPA in pumpkin seedlings' root samples. Edges between two nodes were defined as Pearson's correlation coefficients > 0.6 and shared reactions related PMDs including 77.91 Da for debromination, 162.05 Da for glycosylation, 86 Da for malonylation, 14.02 Da for methylation, and 15.99 Da for hydroxylation. Previously-reported metabolites are labeled.

***Source appointment of unknown compounds***

When an unknown compound is identified as a potential biomarker, determining whether it is associated with endogenous biochemical pathways or exogenous exposures can provide important information toward identification. We found that high-frequency PMDs from HMDB are dominated by reactions with carbon, hydrogen, and oxygen (Table 3), suggesting links to metabolism pathways. Therefore, if an unknown biomarker is mapped using a PMD network, connection to these high-frequency PMDs would suggest an endogenous link. However,

separation from this network is expected for an exogenous biomarker in which the reactive enzyme is not in the database, the exogenous compound is secreted in the parent form, or if it undergoes changes in functional groups such as during phase I and phase II xenobiotic metabolism processes. In this case, endogenous and exogenous compounds should be separated by their PMD network in samples.

Topological differences in PMD networks for endogenous and exogenous metabolites were explored using compounds from The Toxin-Toxin-Target Database, T3DB, which annotates their entries as endogenous or exogenous origin, and carcinogenic classifications (20). T3DB contains 3673 compounds with 2686 unique formulas and 255 endogenous compounds. To avoid too many random PMDs with high frequency from exogenous compounds and to focus on those with known adverse health associations, we calculated the PMDs between all of the 255 endogenous compounds and 705 exogenous compounds with carcinogenic 1, 2A, or 2B classifications and constructed a PMD network from the top-20 high frequency PMDs. A majority of the endogenous compounds were connected into a large network, while the exogenous compounds' networks were much smaller (Figure 3). Interestingly, carcinogenic compounds were not connected by high-frequency PMDs. In fact, the average degree of connection with other nodes was 8.1 for endogenous compounds and 2.3 for carcinogenic compounds. In this case, metabolites with very high or very low degrees of connectivity would suggest whether that compound is endogenous or exogenous, respectively.

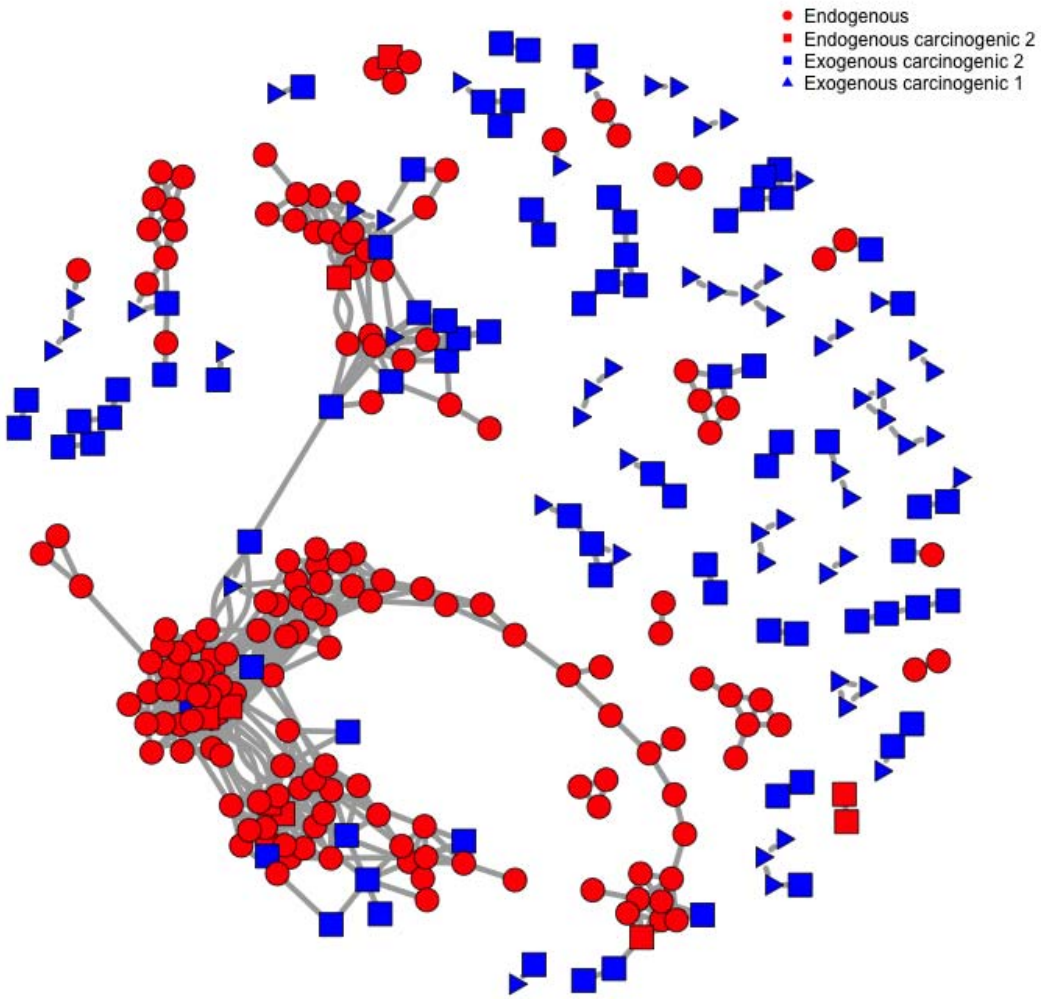


Figure 3. PMD network of 255 endogenous compounds and 705 exogenous carcinogens from T3DB database.

***Biomarker reactions***

Reactomics can also be used to discover biomarker 'reactions' instead of biomarker 'compounds'. Unlike typical biomarkers that are a specific chemical compound, biomarker reactions contain all peaks within a fixed PMD relationship and correlation coefficients cutoff. Thus, quantitative PMD analysis can be used to determine if there are differences between

groups (e.g., control or treatment, exposed or not-exposed, etc.) on a reaction level. Such differences would be described as a biomarker 'reaction'.

In the publicly available dataset MTBLS28 (21), four independent peaks from 1807 samples generated the quantitative responses of PMD 2.02 Da. This biomarker reaction (e.g., +2H from our annotated database) was significantly decreased in case samples compared with the control group (t-test, $p < 0.05$; see Figure 4). The original publication associated with this dataset did not report any molecular biomarker associated with this reaction (21). Thus, quantitative PMD analysis can offer additional information on biological differences between groups at the reaction level that may be lost when focused on analysis at the chemical level. Furthermore, these results suggest that follow-up analysis in this population should include targeted analysis of proteins or enzymes linked with +2H changes.

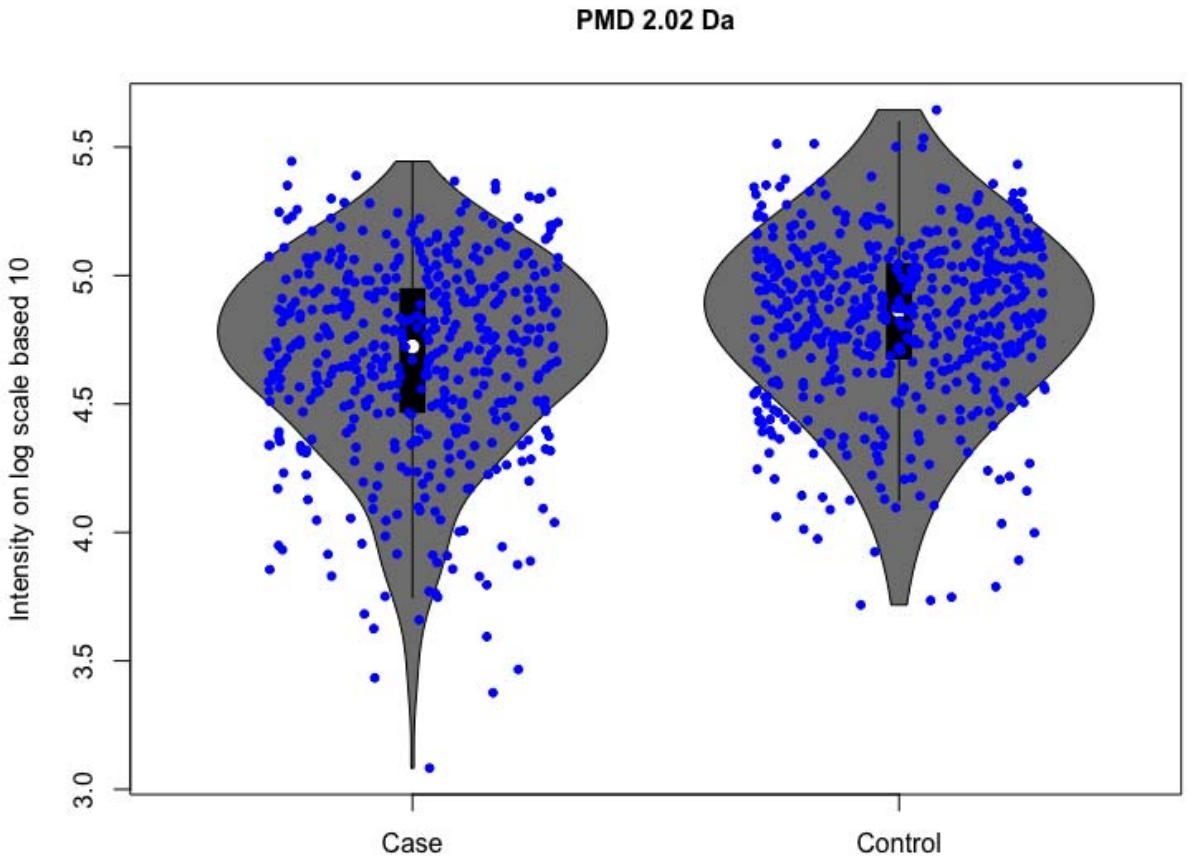

Figure 4. Quantitative PMD analysis identifies PMD 2.02 Da as a potential biomarker reaction for lung cancer (MTBLS28 dataset).

## Conclusion

We provide the theoretical basis and empirical evidence that high-resolution mass spectrometers can be used as reaction detectors through calculation of high-resolution paired mass distances and linkage to reaction databases such as KEGG. Reactomics, as a new concept in bioinformatics, can be used to find biomarker reactions or develop PMD networks. These techniques can provide new information on biological changes, to ultimately promote novel biological inferences that may not be observed through classic chemical biomarker discovery strategies.

**Acknowledgments**
This research was supported by NIEHS grants P30ES23515 and 1U2CES030859.

## References


1. A. Zhang, H. Sun, P. Wang, Y. Han, X. Wang, Modern analytical techniques in metabolomics analysis. *The Analyst* **137**, 293–300 (2012).
2. J. J. J. van der Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess, S. Rogers, Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci.* **113**, 13738–13743 (2016).
3. X. Domingo-Almenara, J. R. Montenegro-Burke, H. P. Benton, G. Siuzdak, Annotation: A Computational Solution for Streamlining Metabolomics Analysis. *Anal. Chem.* **90**, 480–489 (2018).
4. C. Guijas, *et al.*, METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Anal. Chem.* **90**, 3156–3164 (2018).
5. S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann, In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **11**, 148 (2010).
6. L. A. Lerno, J. B. German, C. B. Lebrilla, Method for the Identification of Lipid Classes Based on Referenced Kendrick Mass Analysis. *Anal. Chem.* **82**, 4236–4245 (2010).
7. A. Léon, *et al.*, HaloSeeker 1.0, a user-friendly software to highlight halogenated chemicals in non-targeted high resolution mass spectrometry datasets. *Anal. Chem.* (2019) https:/doi.org/10.1021/acs.analchem.8b05103 (February 19, 2019).
8. A. Bar-Even, E. Noor, N. E. Lewis, R. Milo, Design and analysis of synthetic carbon fixation pathways. *Proc. Natl. Acad. Sci.* **107**, 8889–8894 (2010).
9. D. Normile, Round and Round: A Guide to the Carbon Cycle. *Science* **325**, 1642–1643 (2009).
10. J. Donohue, K. N. Trueblood, Base pairing in DNA. *J. Mol. Biol.* **2**, 363–371 (1960).
11. A. Chokkathukalam, *et al.*, mzMatch–ISO: an R tool for the annotation and relative quantification of isotope-labelled mass spectrometry data. *Bioinformatics* **29**, 281–283 (2013).
12. M. Yu, M. Olkowicz, J. Pawliszyn, Structure/reaction directed analysis for LC-MS based untargeted analysis. *Anal. Chim. Acta* **1050**, 16–24 (2019).
13. N. G. Mahieu, G. J. Patti, Systems-Level Annotation of a Metabolomics Data Set Reduces 25□000 Features to Fewer than 1000 Unique Metabolites. *Anal. Chem.* **89**, 10397–10406 (2017).
14. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
15. D. S. Wishart, *et al.*, HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
16. C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson, S. Neumann, CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **84**, 283–289 (2012).
17. C. D. Broeckling, F. A. Afsar, S. Neumann, A. Ben-Hur, J. E. Prenni, RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Anal. Chem.* **86**, 6812–6817 (2014).
18. W. B. Dunn, I. D. Wilson, A. W. Nicholls, D. Broadhurst, The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **4**, 2249–2264 (2012).
19. X. Hou, *et al.*, Glycosylation of Tetrabromobisphenol A in Pumpkin. *Environ. Sci. Technol.* (2019) https:/doi.org/10.1021/acs.est.9b02122 (July 26, 2019).

20. M. Yu, *et al.*, Evaluation and reduction of the analytical uncertainties in GC-MS analysis using a boundary regression model. *Talanta* **164**, 141–147 (2017).
21. E. A. Mathé, *et al.*, Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer Res.* **74**, 3259–3270 (2014).