Supporting Information
mWISE: an algorithm for context-based annotation of LC-MS features through diffusion in graphs

Maria Barranco-Altirriba,[*,†,‡,¶,§] Pol Solà-Santos,[†,‡,¶] Sergio Picart-Armada,[†,‡,¶] Samir Kanaan-Izquierdo,[†,‡,¶] Jordi Fonollosa,[†,‡,¶] and Alexandre Perera-Lluna[†,‡,¶]

[†]*B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Av. Diagonal 647, 08028 Barcelona, Spain*

[‡]*Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), 28029 Madrid, Spain*

[¶]*Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat, 08950 Barcelona, Spain*

[§]*Department of Endocrinology and Nutrition, Hospital de la Santa Creu i Sant Pau and Institut d'Investigació Biomèdica Sant Pau (IIB Sant Pau), 08041 Barcelona, Spain*

E-mail: maria.barranco@upc.edu

Table of contents

1 Methods

1.1 Algorithm implementation

mWISE is an annotation algorithm with a modular design that provides a biological or biochemical context-based prioritized list of KEGG compounds for LC-MS peaks. It consists of three main stages. First, the LC-MS peaks are matched to KEGG database. Then, the features that are likely to come from the same metabolite are grouped and a filter based on the built clusters is applied. Finally, diffusion in biochemical or biological networks is used to provide a prioritized list based on diffusion scores.

In Figure S1, a detailed scheme of mWISE R package is provided.

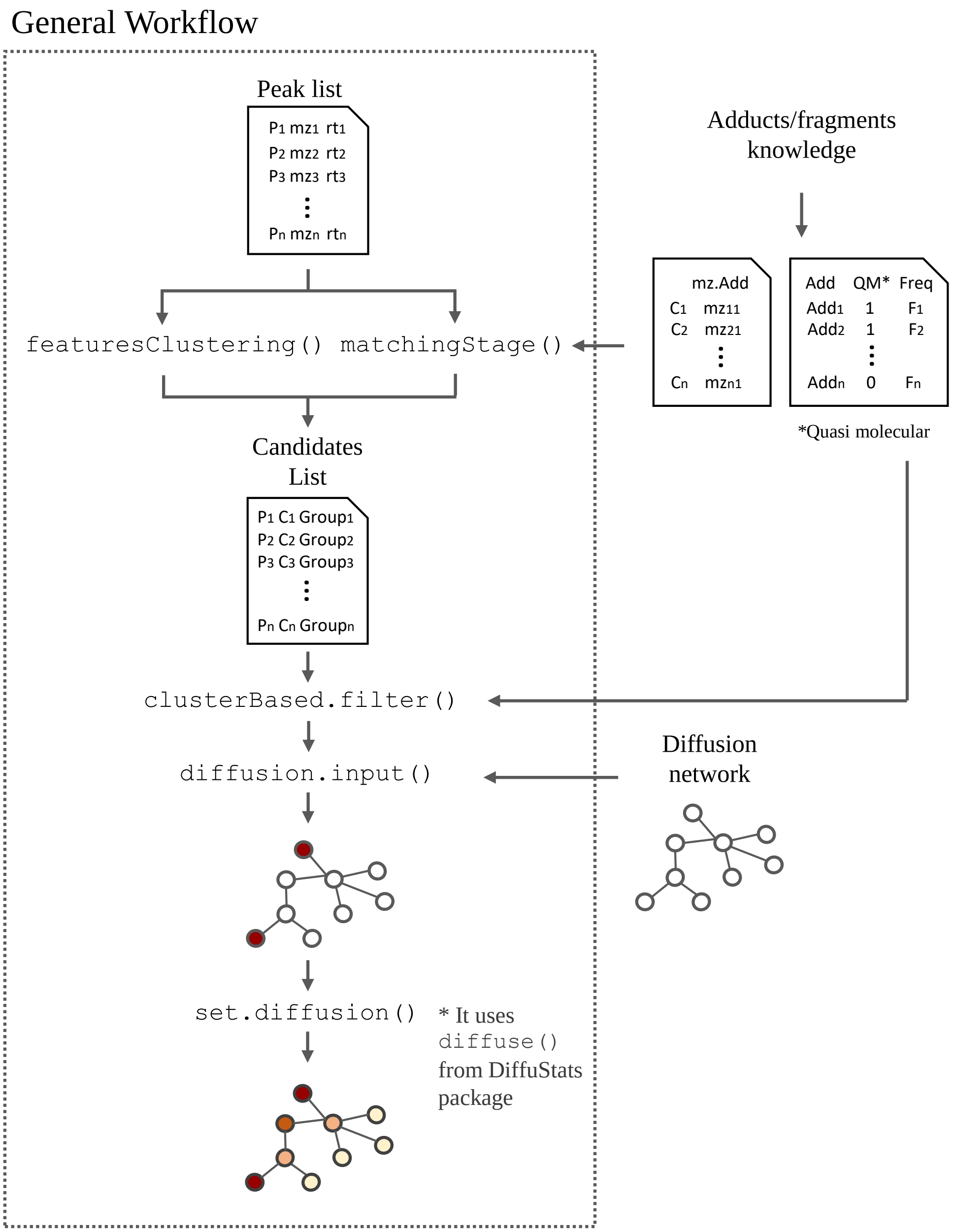

Figure S1: mWISE package scheme

The `matchingStage` command uses the GenomicRanges R package[1] to rapidly match all the LC-MS peaks to KEGG database considering a set of adducts and fragments. The adducts and fragments available for annotation in mWISE are collected from different sources and in mWISE default mode, all of them are used for annotation. However, the users can use their expertise to select the combination of adducts and fragments more appropriate for their specific LC-MS experimental setup.

The featuresClustering function applies spectral clustering in order to group those features more likely to come from the same metabolite. It uses DBSCAN algorithm and applies a set of functions to optimize the number of principal components and the epsilon parameter. In order to automatically optimize the mentioned parameters, the process of building S^{comb} is repeated setting $S_{ij}^{comb} = 0$ when $i = j$. Again, the laplacian matrix and its principal components are computed. Then, the k means algorithm is applied varying the parameter k that defines the number of clusters and also varying the number of principal components accordingly. Equations 1 and 2 are computed for each case.

$$f_1 = \sum_{c=1}^{k} \overline{S_c^I} \tag{1}$$

Where $\overline{S_c^I}$ represents the mean of the intensity similarity values of cluster c, and k represents the number of clusters in the corresponding configuration.

The next equation consists of the same computation but using the retention time similarity matrix.

$$f_2 = \sum_{c=1}^{k} \overline{S_c^{RT}} \tag{2}$$

Finally, a last equation is used (f_3), where the number of putative compound units that are positively correlated are determined. To do so, the mean of the features intensities in each cluster is computed and it must be determined which of these compound units are positively correlated. The configuration that gives a minimum value of $f_3 - (f_2 + f_1)$ is chosen, and the same procedure is repeated for ε parameter.

Then, based on the grouping of peaks performed, the clusterBased.filter command filters the data, thus reducing the number of false positive values.

Finally, the `diffusion.input` command computes the initial diffusion labels vector and the function `set.diffusion` uses DiffuStats[2] R package to diffuse the label vector in a given network.

1.2 Benchmark datasets preparation

The input peak lists were filtered. To do so, the LC-MS features without signal were removed and the 80% rule was applied in the cases where the intensity was equal to 0. The 80% rule is a widely used criterion applied when processing LC-MS data. It consists of removing those peaks that contain missing values in more than 20% of the samples.[3]

2 Results

2.1 mWISE performance and benchmark - detailed metrics

In Tables S1-S11, the NA column refers to the number of not annotated peaks, meaning a peak that has zero proposed candidates and the Ref.N column indicates the number of reference peak-to-compound assignations.

In Table S1, the entities metrics obtained in the matching stage of mWISE are shown.

Table S1: mWISE matching entities metrics

Assay	TP	FP	TN	FN	NA	Ref.N	Sens	Spec	Prec	Acc	F1
Assay 1	144	11605	0	13	0	157	0.92	0.00	0.01	0.01	0.02
Assay 2	150	17598	0	25	0	175	0.86	0.00	0.01	0.01	0.02
Assay 3	145	11037	0	16	0	161	0.90	0.00	0.01	0.01	0.03
Assay 4	91	10133	0	11	0	102	0.89	0.00	0.01	0.01	0.02
Assay 5	38	4226	0	4	0	42	0.90	0.00	0.01	0.01	0.02
Assay 6	59	7942	0	15	0	74	0.80	0.00	0.01	0.01	0.01

In Table S2, the performance of the cluster-based filter for each dataset is shown. This filter allows to reduce the number of false positives introduced in the diffusion process, thus improving the performance of the final prioritization.

Table S2: mWISE filtering entities metrics

Assay	TP	FP	TN	FN	NA	Ref.N	Sens	Spec	Prec	Acc	F1
Assay 1	128	1373	10232	29	0	157	0.82	0.88	0.09	0.88	0.15
Assay 2	126	1671	15927	49	0	175	0.72	0.91	0.07	0.90	0.13
Assay 3	131	1669	9368	30	0	161	0.81	0.85	0.07	0.85	0.13
Assay 4	87	832	9301	15	0	102	0.85	0.92	0.09	0.92	0.17
Assay 5	37	446	3780	5	0	42	0.88	0.89	0.08	0.89	0.14
Assay 6	57	739	7203	17	0	74	0.77	0.91	0.07	0.91	0.13

The diffusion-based scores are computed using the probability input type and the z normalized score and a ranked list is built. The top three candidates for each peak, if available, are selected as the final prioritized proposal. The entities metrics are shown in Tables S3 and S4 when using FELLA and RClass networks, respectively.

Table S3: Fella entities metrics using the z normalization score and the probability input

Assay	TP	FP	TN	FN	NA	Ref.N	Sens	Spec	Prec	Acc	F1
Assay 1	91	269	11336	66	2	157	0.58	0.98	0.25	0.97	0.35
Assay 2	93	281	17317	82	4	175	0.53	0.98	0.25	0.98	0.34
Assay 3	84	266	10771	77	8	161	0.52	0.98	0.24	0.97	0.33
Assay 4	57	181	9952	45	2	102	0.56	0.98	0.24	0.98	0.34
Assay 5	24	73	4153	18	1	42	0.57	0.98	0.25	0.98	0.35
Assay 6	48	95	7847	26	1	74	0.65	0.99	0.34	0.98	0.44

Table S4: RClass entities metrics using the z normalization score and the probability input

Assay	TP	FP	TN	FN	NA	Ref.N	Sens	Spec	Prec	Acc	F1
Assay 1	87	283	11322	70	2	157	0.55	0.98	0.24	0.97	0.33
Assay 2	85	316	17282	90	2	175	0.49	0.98	0.21	0.98	0.30
Assay 3	84	286	10751	77	8	161	0.52	0.97	0.23	0.97	0.32
Assay 4	56	192	9941	46	2	102	0.55	0.98	0.23	0.98	0.32
Assay 5	23	76	4150	19	1	42	0.55	0.98	0.23	0.98	0.33
Assay 6	46	115	7827	28	0	74	0.62	0.99	0.29	0.98	0.39

The same results are shown in Tables S5 and S6 but using the binary input type, the raw diffusion score and the unique annotation option for the diffusion input.

Table S5: FELLA entities metrics using the raw score, the binary input and the unique annotation option

Assay	TP	FP	TN	FN	NA	Ref.N	Sens	Spec	Prec	Acc	F1
Assay 1	103	257	11348	54	2	157	0.66	0.98	0.29	0.97	0.40
Assay 2	92	282	17316	83	4	175	0.53	0.98	0.25	0.98	0.34
Assay 3	90	260	10777	71	8	161	0.56	0.98	0.26	0.97	0.35
Assay 4	62	176	9957	40	2	102	0.61	0.98	0.26	0.98	0.36
Assay 5	29	68	4158	13	1	42	0.69	0.98	0.30	0.98	0.42
Assay 6	53	90	7852	21	1	74	0.72	0.99	0.37	0.99	0.49

Table S6: RCLASS entities metrics using the raw score, the binary input and the unique annotation option

Assay	TP	FP	TN	FN	NA	Ref.N	Sens	Spec	Prec	Acc	F1
Assay 1	85	289	11316	72	2	157	0.54	0.98	0.23	0.97	0.32
Assay 2	81	319	17279	94	2	175	0.46	0.98	0.20	0.98	0.28
Assay 3	76	295	10742	85	8	161	0.47	0.97	0.20	0.97	0.29
Assay 4	52	197	9936	50	2	102	0.51	0.98	0.21	0.98	0.30
Assay 5	27	72	4154	15	1	42	0.64	0.98	0.27	0.98	0.38
Assay 6	44	116	7826	30	0	74	0.59	0.99	0.28	0.98	0.38

In Tables S7-S11, the specific entities metrics obtained using xMSannotator[4] R package, mummichog[5] server, MI-Pack[6] algorithm, and ProbMetab[7] and CAMERA[8] R packages are shown, respectively.

Table S7: xMSannotator entities metrics

Assay	TP	FP	TN	FN	NA	Sens	Spec	Prec	Acc	F1
Assay 1	71	759	10264	76	10	0.45	0.93	0.09	0.93	0.15
Assay 2	63	924	15174	101	11	0.36	0.94	0.06	0.94	0.11
Assay 3	62	1002	9815	97	2	0.39	0.91	0.06	0.90	0.10
Assay 4	59	509	9124	42	1	0.58	0.95	0.10	0.94	0.18
Assay 5	34	305	3842	7	1	0.81	0.93	0.10	0.93	0.18
Assay 6	54	369	7138	19	1	0.73	0.95	0.13	0.95	0.22

Table S8: Mummichog entities metrics

Assay	TP	FP	TN	FN	NA	Sens	Spec	Prec	Acc	F1
Assay 1	52	451	11814	105	44	0.33	0.96	0.10	0.96	0.16
Assay 2	35	174	17923	140	90	0.20	0.99	0.17	0.98	0.18
Assay 3	19	314	11357	142	70	0.12	0.97	0.06	0.96	0.08
Assay 4	25	84	10364	77	58	0.25	0.99	0.23	0.98	0.24
Assay 5	22	154	4270	20	8	0.52	0.97	0.12	0.96	0.20
Assay 6	27	84	8160	47	34	0.36	0.99	0.24	0.98	0.29

Table S9: MI-Pack entities metrics

Assay	TP	FP	TN	FN	NA	Sens	Spec	Prec	Acc	F1
Assay 1	0	20	11764	157	151	0.00	1.00	0.00	0.99	0.00
Assay 2	14	84	13663	161	144	0.08	0.99	0.14	0.98	0.10
Assay 3	16	284	8402	145	128	0.10	0.97	0.05	0.95	0.07
Assay 4	10	30	9118	92	90	0.10	1.00	0.25	0.99	0.14
Assay 5	12	70	2723	30	26	0.29	0.97	0.15	0.96	0.19
Assay 6	20	72	5278	54	46	0.27	0.99	0.22	0.98	0.24

Table S10: ProbMetab entities metrics

Assay	TP	FP	TN	FN	NA	Sens	Spec	Prec	Acc	F1
Assay 5	33	85	4298	9	0	0.79	0.98	0.28	0.98	0.41
Assay 6	62	125	8139	12	0	0.84	0.98	0.33	0.98	0.48

Table S11: CAMERA metrics

Assay	TP	FP	FN	N.A	Ref.N	Sens.	Prec.	F1	time (min)
Assay 5	10	8	32	27	42	0.24	0.56	0.33	4.86
Assay 6	8	10	66	60	74	0.11	0.44	0.17	10.74

The computation time of each algorithm is shown in Table S12 for each assay.

Table S12: Computation time for each algorithm and dataset in minutes.

Algorithm	Assay 1	Assay 2	Assay 3	Assay 4	Assay 5	Assay 6
mWISE - Fella-z score	15.70	54.20	45.20	26.35	18.10	35.41
mWISE - Fella-raw score	15.70	53.38	43.40	26.45	17.84	35.07
mWISE - RClass-z score	15.66	54.68	43.86	26.30	18.16	34.84
mWISE - RClass-raw score	15.52	53.06	42.74	26.06	17.77	34.96
xMSannotator	106.34	318.04	192.81	201.52	127.20	245.25
Mummichog	1.55	1.45	1.30	1.05	1.20	1.13
MI-Pack	1892.27	1971.13	5547.60	3376.95	2130.65	4243.38
CAMERA	-	-	-	-	10.74	4.86

Table S13 shows the characteristics of each algorithm. mWISE, mummichog, ProbMetab, MI-Pack and xMSannotator provide biological knowledge to the annotation process, as well as the proposal of specific metabolites, while CAMERA process ends with the adducts and fragments annotation. An important limitation of mWISE is the databases offered, since mWISE only offers the data to annotate in KEGG database. This is an important limitation with respect to xMSannotator that should be addressed in future versions of mWISE.

Table S13: Algorithms' characteristics

Algorithm	Grouping features	Adducts/fragments information	Biological knowledge	Metabolite annotation	Databases available
CAMERA	Yes	Yes	No	No	None
mWISE	Yes	Yes	Yes	Yes	KEGG
mummichog	Yes	Yes	Yes	Yes	KEGG
ProbMetab	Yes	Yes	Yes	Yes	KEGG
MI-Pack	Yes	Yes	Yes	Yes	KEGG
xMSannotator	Yes	Yes	Yes	Yes	KEGG, HMDB, LipidMaps, T3DB

In Table S14, the input objects required by each algorithm are shown. mWISE, mummichog, MI-Pack and xMSannotator are more flexible than ProbMetab and CAMERA, since a peak-list data frame is required as input. This input can be obtained using any LC-MS pre-processing software.

Table S14: Input objects type

Algorithm	Input
mWISE	Peak-intensity matrix
mummichog	LC-MS features (m/z and rt)
CAMERA	xcms object
xMSannotator	Peak-intensity matrix
ProbMetab	CAMERA/mzMatch object
MI-Pack	Peak-intensity matrix

2.2 Tanimoto similarity - detailed metrics

In Tables S15-S18, the number of peaks in which mWISE proposes at least a compound with a chemical structure identical to the correct compound are shown in the column named Tanimoto Hits. In order to determine which compounds are identical, the Tanimoto similarity coefficient is computed between the proposed compounds and the correct ones, and those cases with a Tanimoto measure equal to 1 are considered as equal. The ratio of these peaks (Tanimoto ratio) with respect to the number of reference peak-to-compound assignations is also shown. These results show that in a considerable proportion of peaks considered as false positives in the entities metrics, mWISE proposes a compound that probably shares several properties and biological reactions with the correct one.

Table S15: Tanimoto metrics for Fella graph and z score

Dataset	TP	Tanimoto hits	Ref N	Tanimoto ratio
Assay 1	91	97	157	0.62
Assay 2	93	116	175	0.66
Assay 3	84	95	161	0.59
Assay 4	57	69	102	0.68
Assay 5	24	30	42	0.71
Assay 6	48	50	74	0.68

Table S16: Tanimoto metrics for Fella graph and raw score

Dataset	TP	Tanimoto hits	Ref N	Tanimoto proportion
Assay 1	103	109	157	0.69
Assay 2	92	115	175	0.66
Assay 3	90	98	161	0.61
Assay 3	62	77	102	0.75
Assay 4	29	32	42	0.76
Assay 5	53	53	74	0.72

Table S17: Tanimoto metrics for RClass graph and z score

Dataset	TP	Tanimoto hits	Ref N	Tanimoto proportion
Assay 1	87	96	157	0.61
Assay 2	85	107	175	0.61
Assay 3	84	89	161	0.55
Assay 4	56	64	102	0.63
Assay 5	23	26	42	0.62
Assay 6	46	48	74	0.65

Table S18: Tanimoto metrics for RClass graph and raw score

Dataset	TP	Tanimoto hits	Ref N	Tanimoto proportion
Assay 1	85	92	157	0.59
Assay 2	81	103	175	0.59
Assay 3	76	90	161	0.56
Assay 4	52	64	102	0.63
Assay 5	27	30	42	0.71
Assay 6	44	48	74	0.65

Hereafter, the Tanimoto coefficients computed between the correct peak-to-compound assignations and the compounds proposed by mWISE and xMSannotator are plotted against peak's degree. Peak's degree is defined as the number of proposed compounds for each peak. Only the non-correct assignations are considered. The p-values obtained when comparing the Tanimoto coefficients between mWISE and xMSannotator using a Brunner-Munzel test are seen in Figure S2.

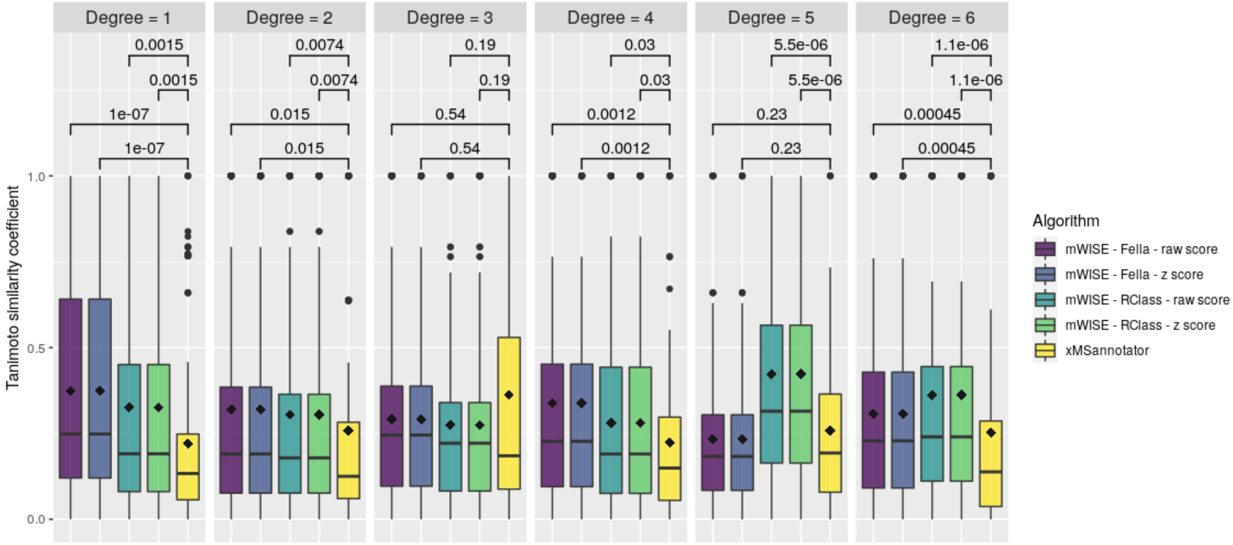

Figure S2: Comparison of xMSannotator and mWISE Tanimoto coefficients between the proposed compounds and the correct ones. The compounds correctly proposed, and therefore considered as true positives have been discarded. A Brunner-Munzel test has been applied in each comparison and the mean value is plotted with a central point. The six panels indicate the number of proposed compounds for each peak (degree of each peak).

2.3 Diffusion prioritization analysis - detailed metrics

As explained in the paper, the diffusion prioritization of randomly arranged graphs has been compared to the results obtained when using the real networks. To do so, the diffusion-based ranking of both real and surrogate cases (results obtained when permuting the graphs) have been plotted against the degree of each peak, defining degree as the number of possible candidates for a peak. In here, four different plots for each diffusion configuration are shown.

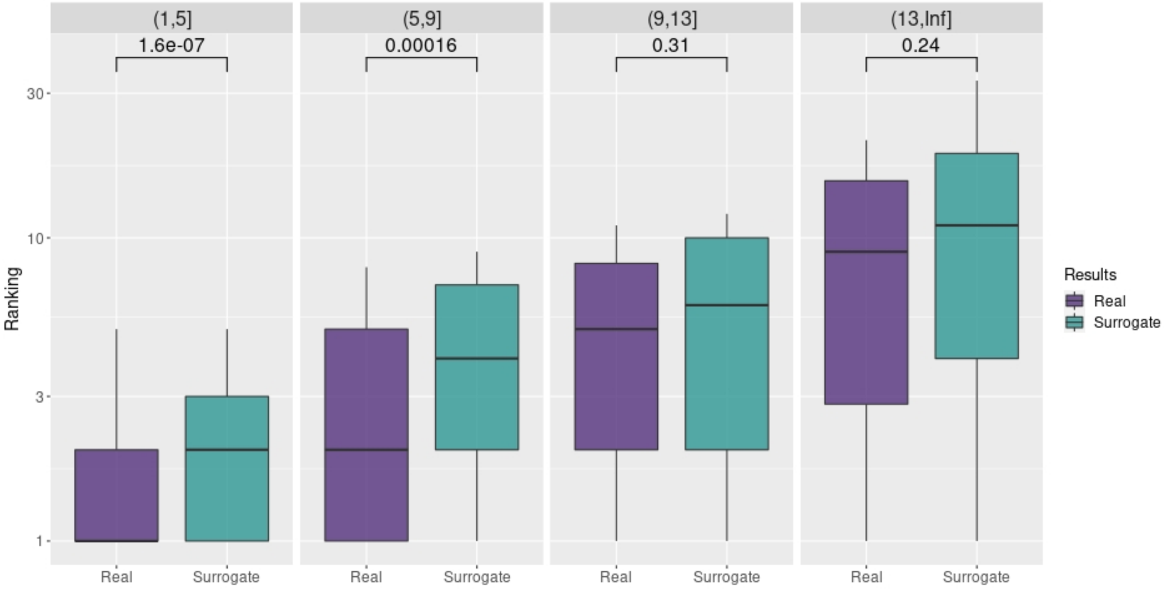

Figure S3: Diffusion-based ranking of both real and surrogate results divided in ranges of degrees, defining degree as the number of possible compounds proposed for a peak. The p-values of a Brunner-munzel test are shown on the top of the plot. The alternative hypothesis of the tests are that the ranking of the real cases is lower than the surrogate cases. The results are obtained using the Fella graph and the z score.

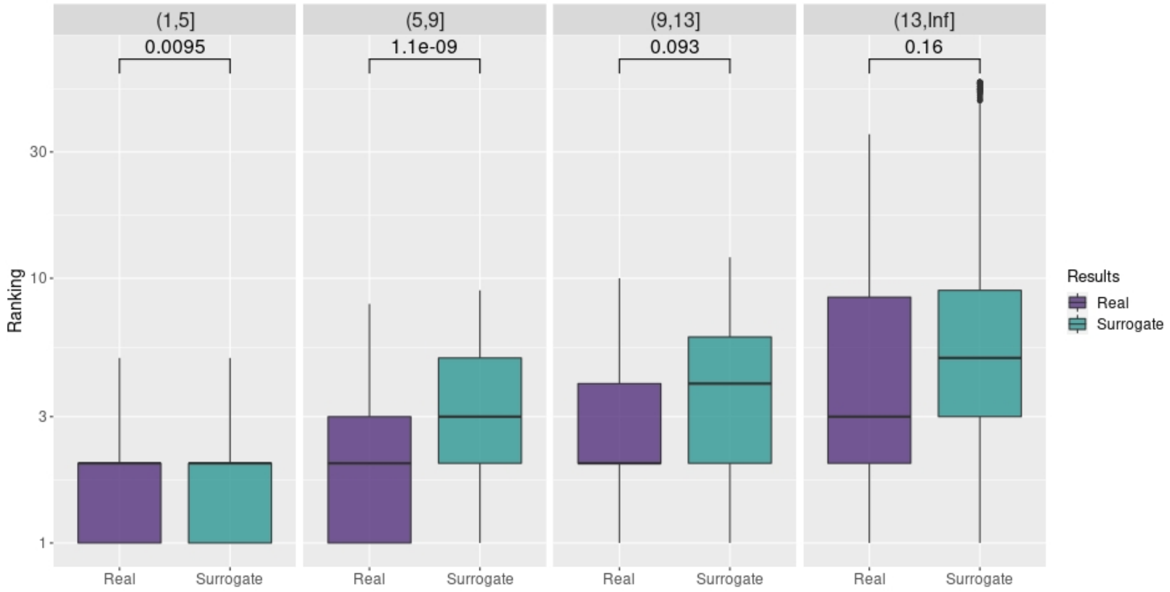

Figure S4: Diffusion-based ranking of both real and surrogate results divided in ranges of degrees, defining degree as the number of possible compounds proposed for a peak. The p-values of a Brunner-munzel test are shown on the top of the plot. The alternative hypothesis of the tests are that the ranking of the real cases is lower than the surrogate cases. The results are obtained using the Fella graph and the raw score.

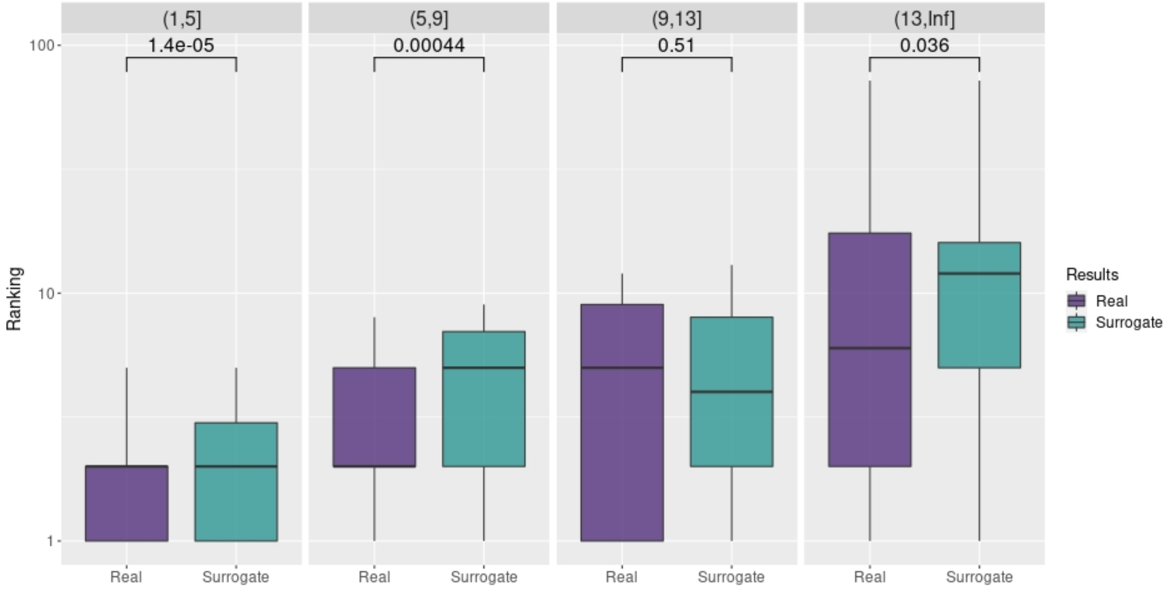

Figure S5: Diffusion-based ranking of both real and surrogate results divided in ranges of degrees, defining degree as the number of possible compounds proposed for a peak. The p-values of a Brunner-munzel test are shown on the top of the plot. The alternative hypothesis of the tests are that the ranking of the real cases is lower than the surrogate cases. The results are obtained using the RClass graph and the z score.

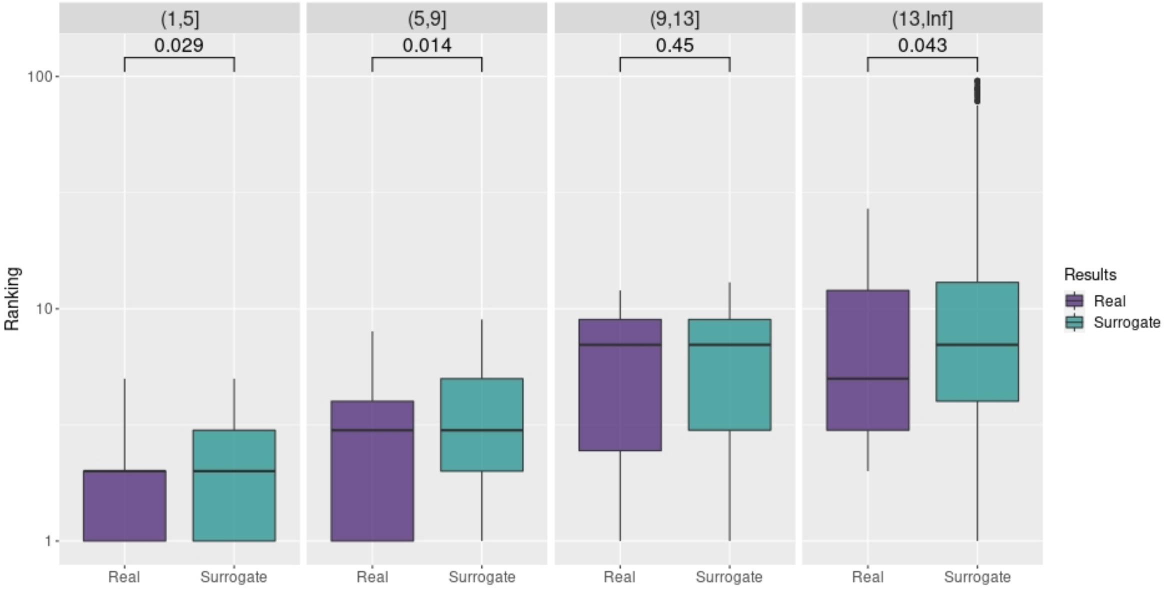

Figure S6: Diffusion-based ranking of both real and surrogate results divided in ranges of degrees, defining degree as the number of possible compounds proposed for a peak. The p-values of a Brunner-munzel test are shown on the top of the plot. The alternative hypothesis of the tests are that the ranking of the real cases is lower than the surrogate cases. The results are obtained using the RClass graph and the raw score.

2.4 TPs comparison between algorithms

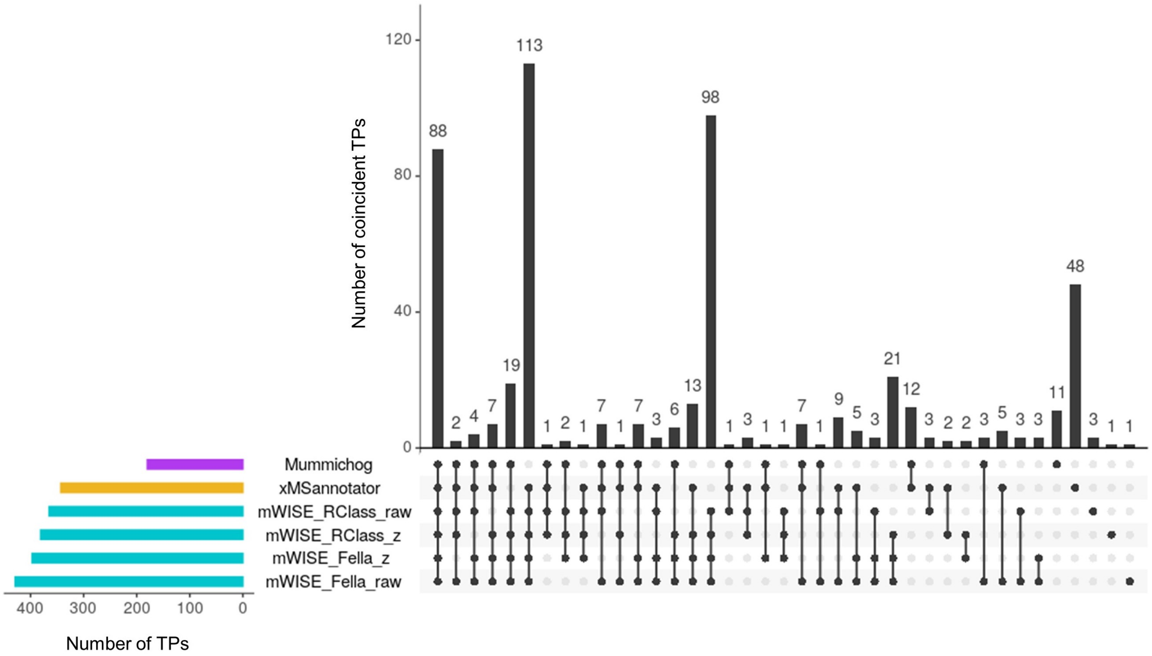

Figure S7: Upset plot showing the comparison of true positives (TPs) between different algorithms across all datasets. The bars in the left show the total number of true positives for each algorithm. The top bars show the number of coincident TPs of the intersections indicated in the matrix below. The first column indicates that 88 peaks are correctly annotated by all the algorithms. Similarly, 113 peaks are correctly annotated by all algorithms except mummichog.

References

(1) Lawrence, M.; Huber, W.; Pagès, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M. T.; Carey, V. J. Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **2013**, *9*, 1–10.

(2) Picart-Armada, S.; Thompson, W. K.; Buil, A.; Perera-Lluna, A. DiffuStats: An R package to compute diffusion-based scores on biological networks. *Bioinformatics* **2018**, *34*, 533–534.

(3) Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; Ni, Y. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **2018**, *8*, 1–10.

(4) Uppal, K.; Walker, D. I.; Jones, D. P. xMSannotator: An R package for network-based annotation of high-resolution metabolomics data. *Anal. Chem.* **2017**, *89*, 1063–1067.

(5) Li, S.; Park, Y.; Duraisingham, S.; Strobel, F. H.; Khan, N.; Soltow, Q. A.; Jones, D. P.; Pulendran, B.; Ouzounis, C. A. Predicting Network Activity from High Throughput Metabolomics. *PLoS Comput Biol* **2013**, *9*, 1–11.

(6) Weber, R. J.; Viant, M. R. MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 75–82.

(7) Silva, R. R.; Jourdan, F.; Salvanha, D. M.; Letisse, F.; Jamin, E. L.; Guidetti-Gonzalez, S.; Labate, C. A.; Vêncio, R. Z. ProbMetab: An R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics* **2014**, *30*, 1336–1337.

(8) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **2012**, *84*, 283–289.