Supplementary information for

Knowledge and Data-driven Two-layer Networking for Accurate Metabolite Annotation in Untargeted Metabolomics

Haosong Zhang[1,2], Xinhao Zeng[1,2], Yandong Yin[1], and Zheng-Jiang Zhu[1,3,]*

[1] Interdisciplinary Research Center on Biology and Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, 200032 P. R. China

[2] University of Chinese Academy of Sciences, Beijing, 100049 P. R. China

[3] Shanghai Key Laboratory of Aging Studies, Shanghai, 201210 P. R. China

Corresponding Author

Correspondence should be addressed to Z.J.Z (jiangzhu@sioc.ac.cn)

List for Supplementary Figures

Supplementary Figure 1. The metabolite knowledge databases of and curation of predicted metabolic reaction network.

Supplementary Figure 2. Architecture of the GNN-based model for reaction relationship prediction.

Supplementary Figure 3. Characterization of chemical space for metabolites in MRN.

Supplementary Figure 4. Two-layer interactive networking topology for recursive metabolite annotation in MetDNA3.

Supplementary Figure 5. Detailed statistics for metabolite annotations in biological samples.

Supplementary Figure 6. Feature annotation in Orbitrap and Astral datasets using MetDNA3.

Supplementary Figure 7. Correct rates for Top 3 annotations across different recursive propagation rounds.

Supplementary Figure 8. False discovery rate of Top N annotations in network-based annotation achieved by MetDNA3.

Supplementary Figure 9. Characterization of recursive-based metabolite annotation and propagation.

Supplementary Figure 10. Validations of annotated metabolites from MetDNA3 using the chemical standards.

Supplementary Figure 11. Comparison of different knowledge-driven networks for network-based annotation.

Supplementary Figure 12. Decoy MS2 spectra-based assessment of false positives in network-based metabolite annotation propagation.

Supplementary Figure 13. Evaluation of multiple in-silico tools used for cross-tool concordance.

Supplementary Figure 14. Comparison of distributions between correct annotations and incorrect candidates in the NIST urine HILIC–MS(+) Orbitrap dataset.

List for Supplementary Tables

Supplementary Table 1. Topological properties of knowledge networks.

Supplementary Table 2. Number of metabolites used for validation in Figure 3d-g from Orbitrap datasets.

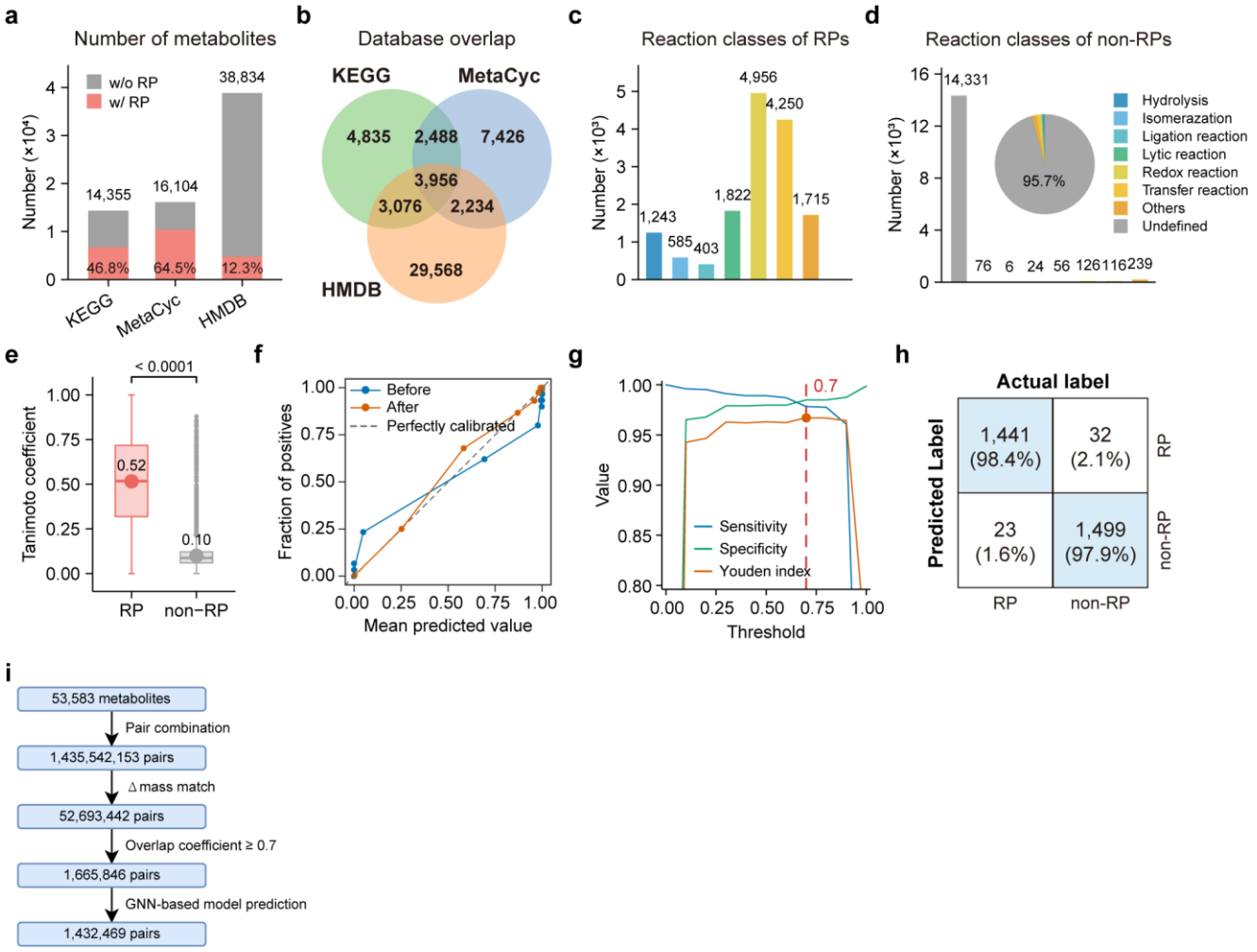
Supplementary Table 3. Parameter settings of BioTransformer.

Supplementary Table 4. Optimized parameters for graph neural network-based prediction model.

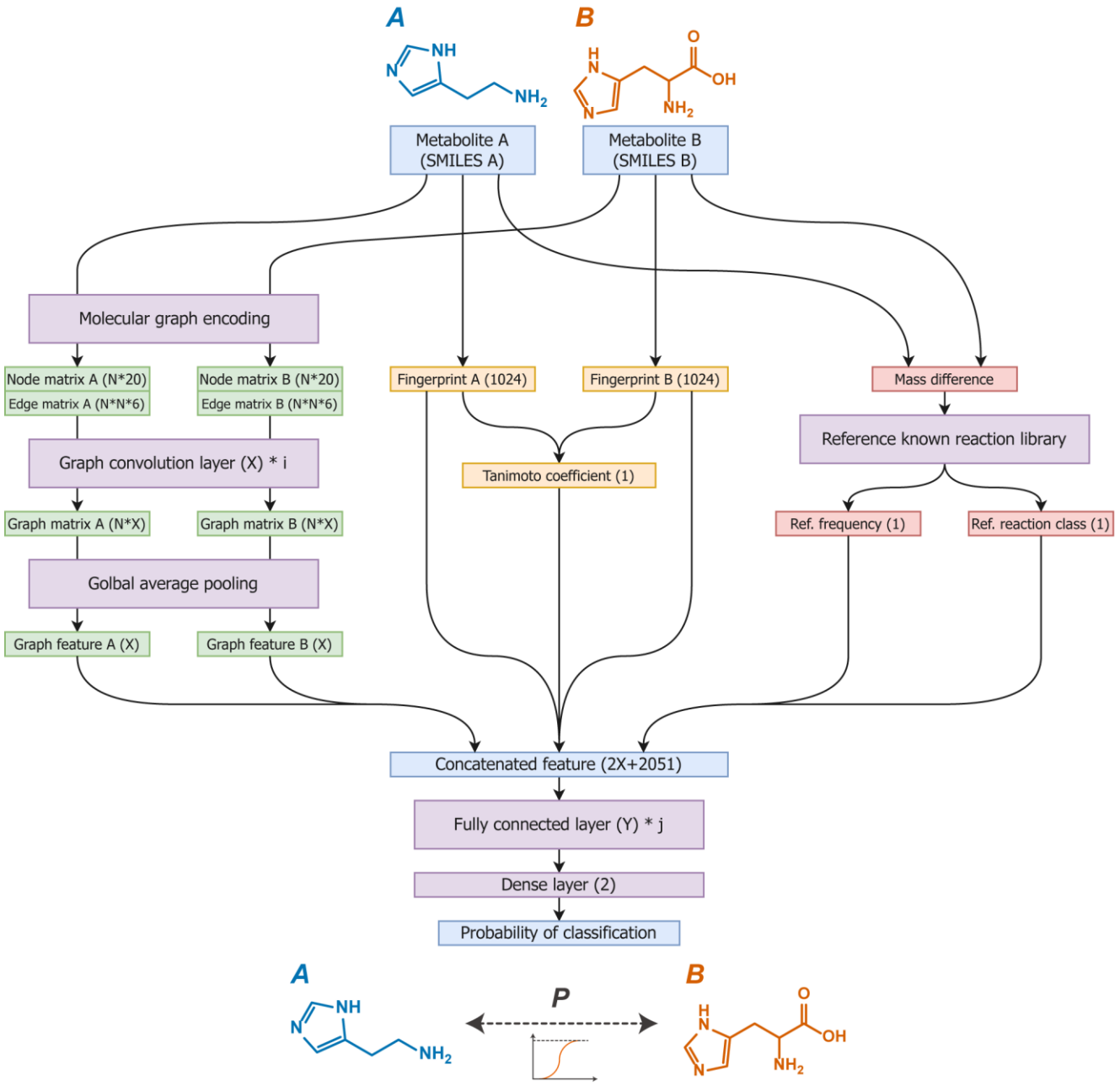
Supplementary Table 5. Parameter settings of four bioinformatic tools.

Supplementary Table 6. Five putative unknown metabolites filtered for validation.

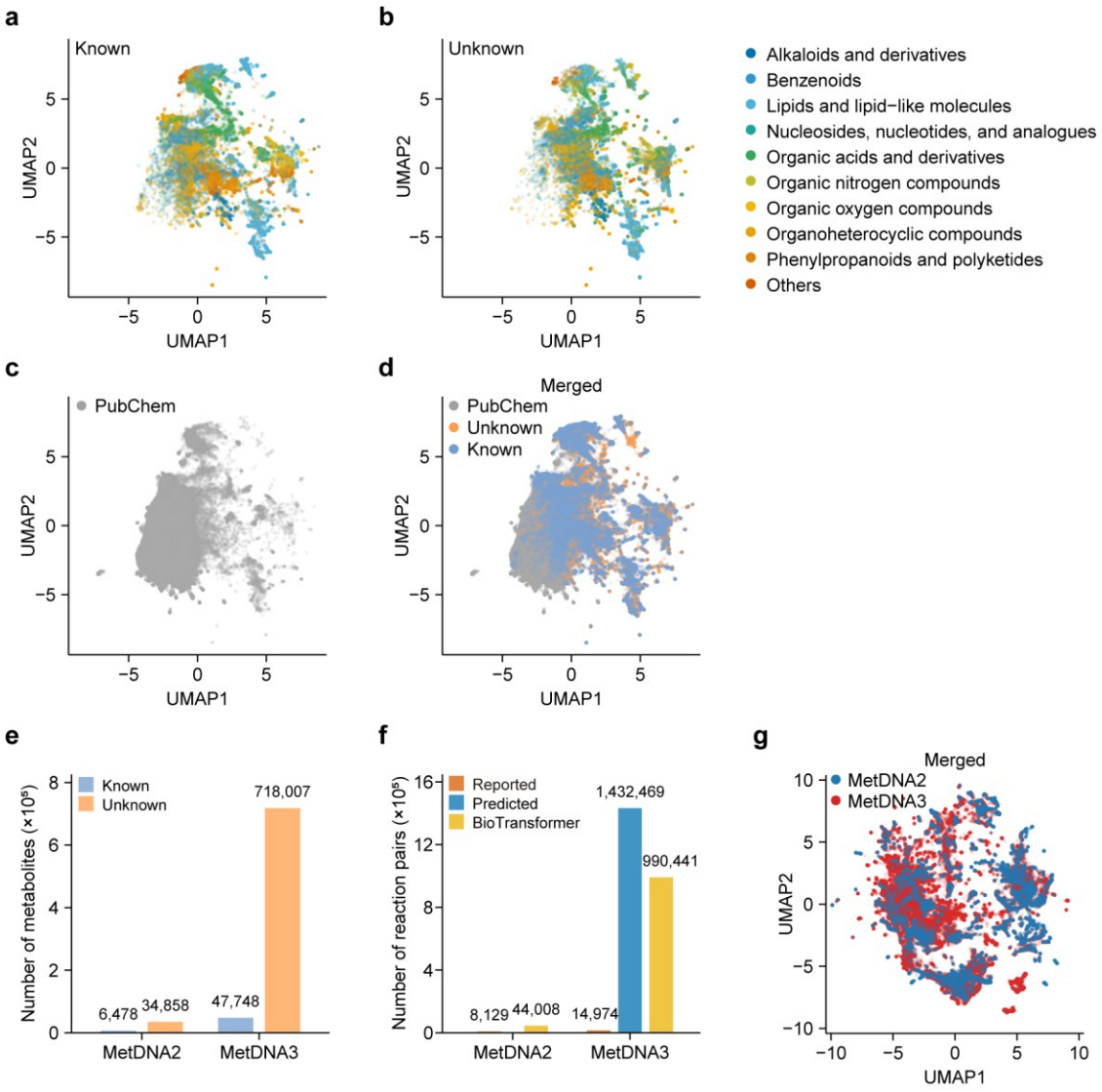
Supplementary Figures



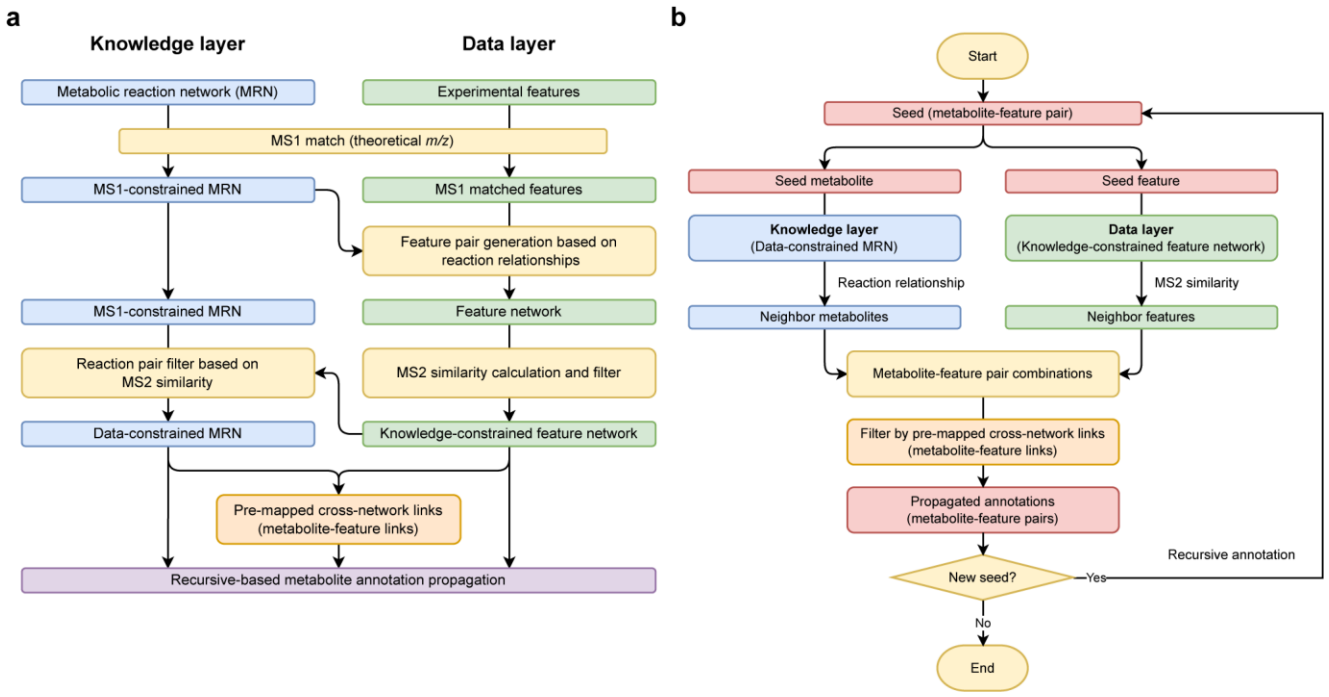
Supplementary Figure 1. The metabolite knowledge databases of and curation of predicted metabolic reaction network. (**a**) Number of metabolites and proportion of metabolites with reaction pairs (RPs) in metabolite knowledge databases. (**b**) Overlap of metabolites across different metabolite knowledge databases. (**c,d**) Distributions of reaction classes of RPs (c) and non-RPs (d). (**e**) Distributions of Tanimoto structural similarity for RPs and non-RPs (n=14,974 per group; two-tailed Student's *t*-test). (**f-h**) Probability calibration (f), threshold determination (g), and prediction performance (h) of GNN-based prediction of reaction relationships in the testing dataset (n=2,995). (**i**) Two-step pre-screening workflow to reduce metabolite pairs input into the GNN-based prediction.



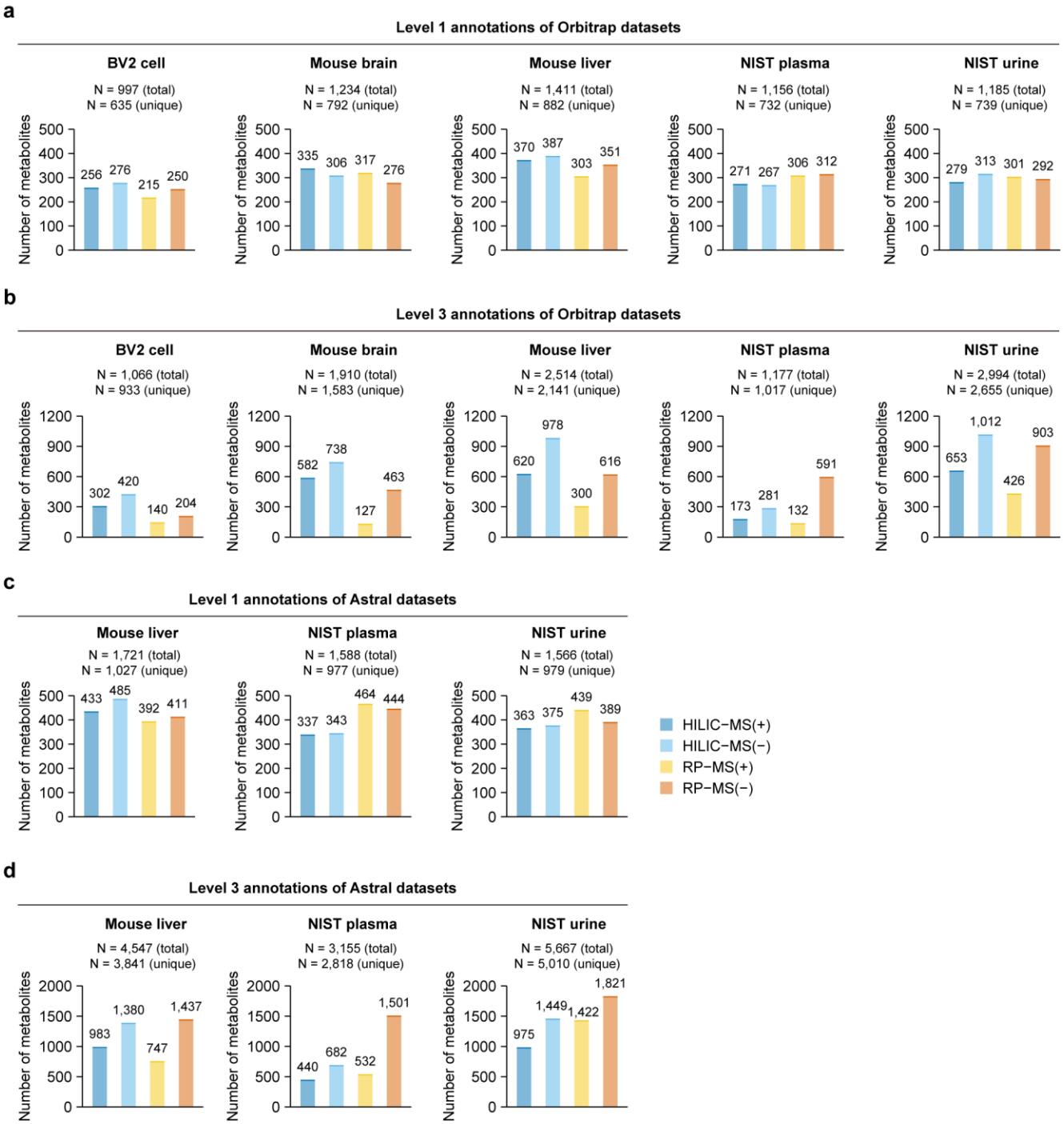
Supplementary Figure 2. Architecture of the GNN-based model for reaction relationship prediction.



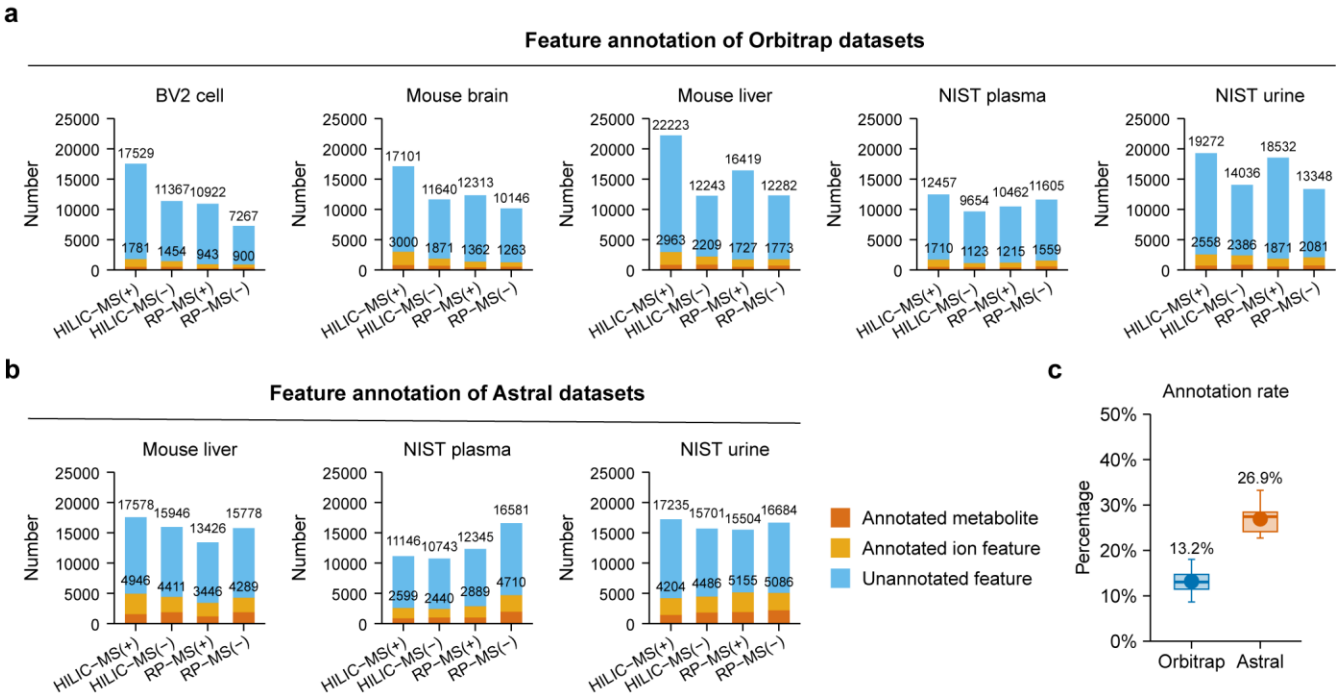
Supplementary Figure 3. Characterization of chemical space for metabolites in MRN. (**a,b**) UMAP visualization of chemical space distributions of known (n=47,748) (a) and random sampled unknown (n=50,000) (b) metabolites in MRN, with colors indicating different metabolite superclasses. (**c,d**) UMAP visualization of of chemical space distributions of random sampled PubChem compounds (n=100,000) (c) and merged distributions of MRN and PubChem (d). (**e**) Numbers of known and unknown metabolites in MRNs of MetDNA2 and MetDNA3. (**f**) Numbers of reaction pairs in MRNs of MetDNA2 and MetDNA3, categorized into reported, predicted, and BioTransformer-generated pairs. (**g**) UMAP visualization of chemical space distributions of metabolites in MRNs of MetDNA2 (n=41336, 6,478 knowns, 34,858 unknowns) and MetDNA3 (n=97,748, 47,748 knowns, 50,000 random sampled unknowns). Metabolite embeddings were generated using continuous and data-driven descriptors (CDDD)[1] for UMAP visualization, with molecules containing missing values excluded. UMAP projection was performed using the uwot package in R with the following parameters: n_neighbors=30, min_dist=0.1, n_threads=6, batch=10000, and approx_pow=FALSE.



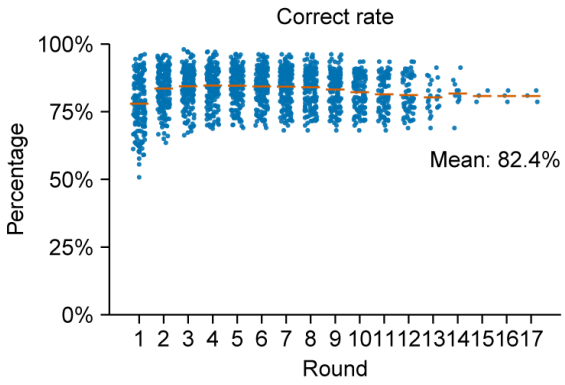
Supplementary Figure 4. Two-layer interactive networking topology for recursive metabolite annotation in MetDNA3. (**a**) Flowchart of the two-layer network topology construction. (**b**) Flowchart of the recursive-based metabolite annotation propagation.



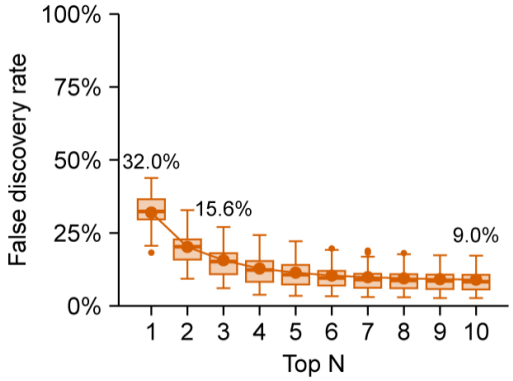
Supplementary Figure 5. Detailed statistics for metabolite annotations in biological samples. (**a**) Numbers of level 1 metabolite annotations for Orbitrap datasets. (**b**) Numbers of level 3 annotations for Orbitrap datasets. (**c**) Numbers of level 1 annotations for Astral datasets. (**d**) Numbers of level 3 annotations of Astral datasets. HILIC–MS(+): positive MS mode with HILIC separation; HILIC–MS(-): negative MS mode with HILIC separation; RP–MS(+): positive MS mode with RP separation; RP–MS(-): negative MS mode with RP separation. “Total” means the sum of all annotation results; “unique” means removing duplicate metabolite annotations.



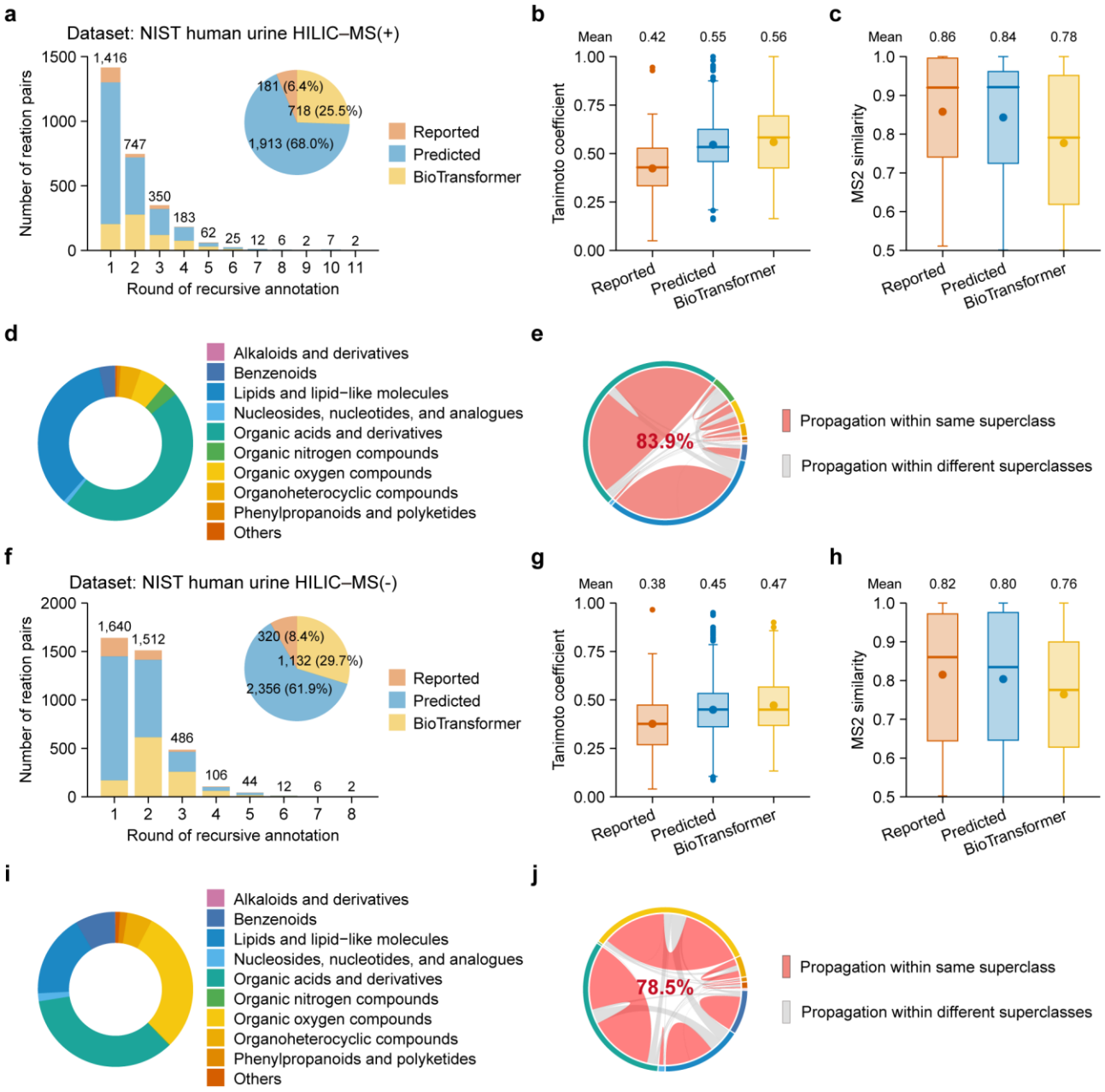
Supplementary Figure 6. Feature annotation in Orbitrap and Astral datasets using MetDNA3. (**a**) Feature annotations for Orbitrap datasets. (**b**) Feature annotations for Astral datasets. (**c**) Feature annotation rate, defined as the percentage of annotated features (including annotated metabolites and ion features) relative to the total number of detected features (20 Orbitrap datasets and 12 Astral datasets). HILIC–MS(+): positive MS mode with HILIC separation; HILIC–MS(-): negative MS mode with HILIC separation; RP–MS(+): positive MS mode with RP separation; RP–MS(-): negative MS mode with RP separation.



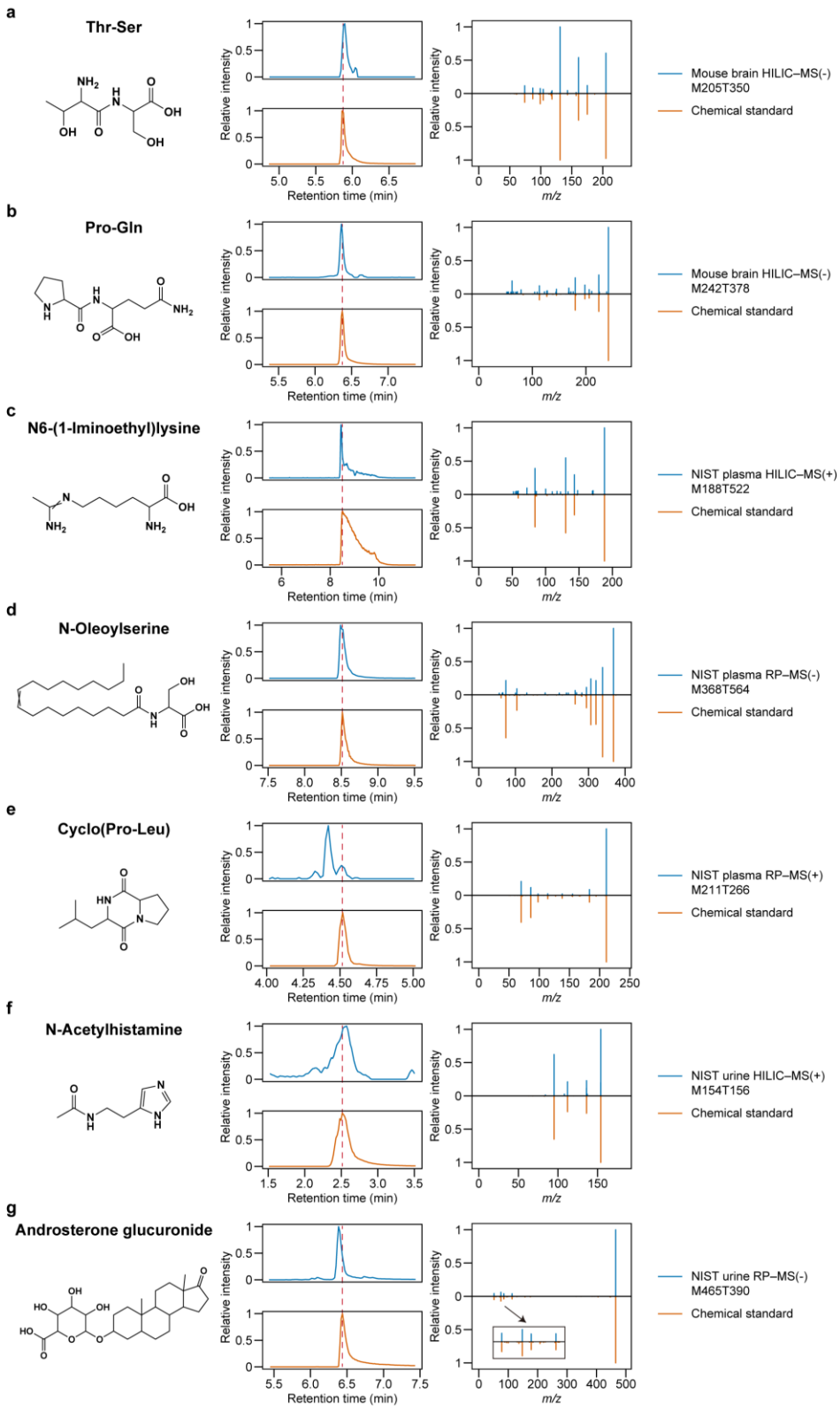
Supplementary Figure 7. Correct rates for Top 3 annotations across different recursive propagation rounds.



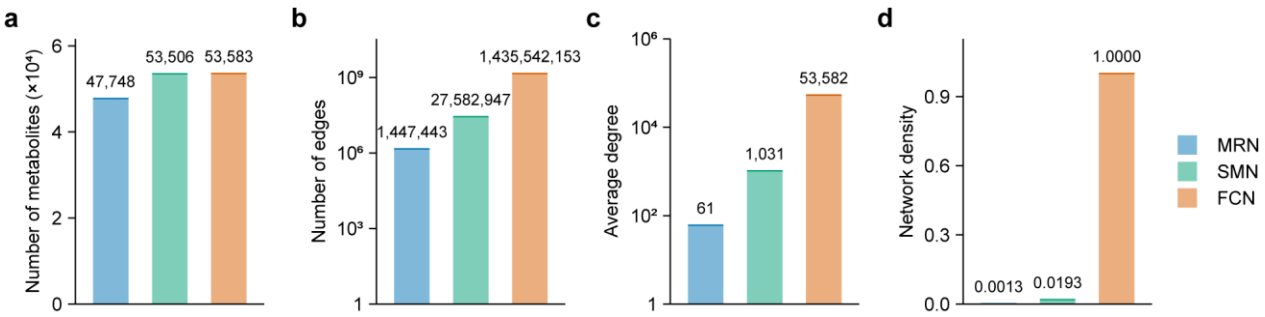
Supplementary Figure 8. False discovery rate of Top N annotations in network-based annotation achieved by MetDNA3.



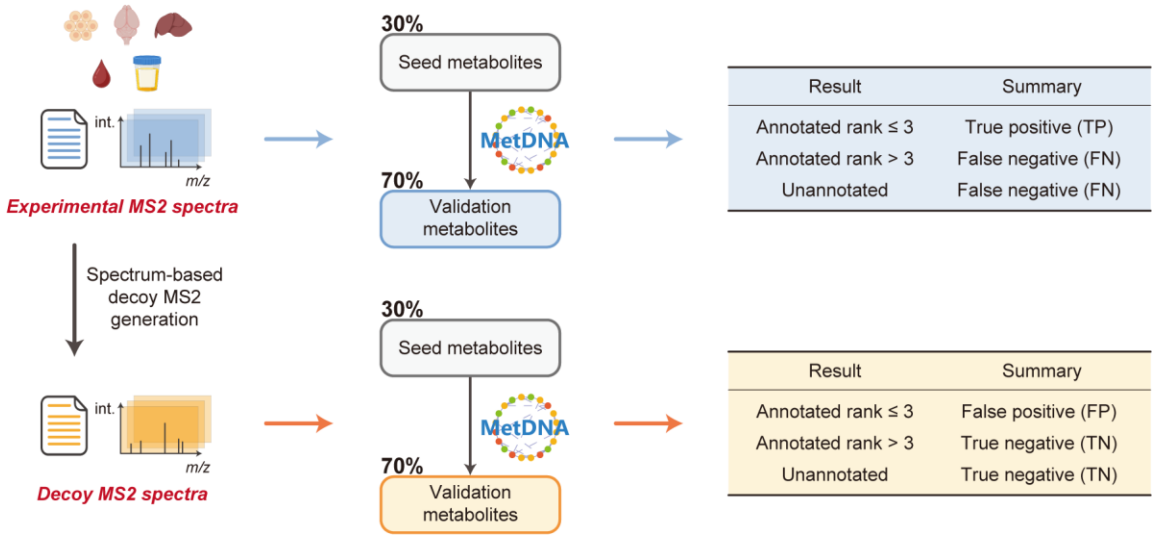
Supplementary Figure 9. Characterization of recursive-based metabolite annotation and propagation. (**a, f**) Distributions and proportions of reported, predicted, and BioTransformer-generated reaction pairs during annotation propagation in HILIC–MS(+) (a) and HILIC–MS(-) (f). Human urine samples were tested (n=6 technical replicates). (**b, g**) Distributions of Tanimoto structural similarity for different types of reaction pairs during annotation propagation in HILIC–MS(+) (b) and HILIC–MS(-) (g). (**c, h**) Distributions of MS2 spectral similarity for different types of reaction pairs during annotation propagation in HILIC–MS(+) (c) and HILIC–MS(-) (h). (**d, i**) Superclass distribution of annotated metabolites in HILIC–MS(+) (d) and HILIC–MS(-) (i). Superclass taxonomy was calculated using ClassyFire[2]. (**e, j**) Chord diagrams illustrating the proportion of reaction pair propagation within and across metabolite superclasses in HILIC–MS(+) (e) and HILIC–MS(-) (j).



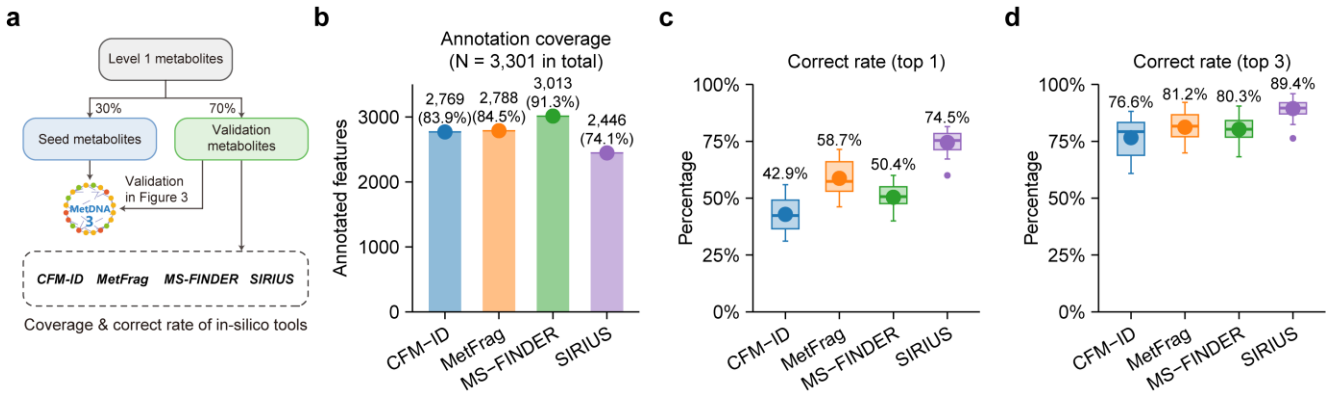
Supplementary Figure 10. Validations of annotated metabolites from MetDNA3 using the chemical standards.



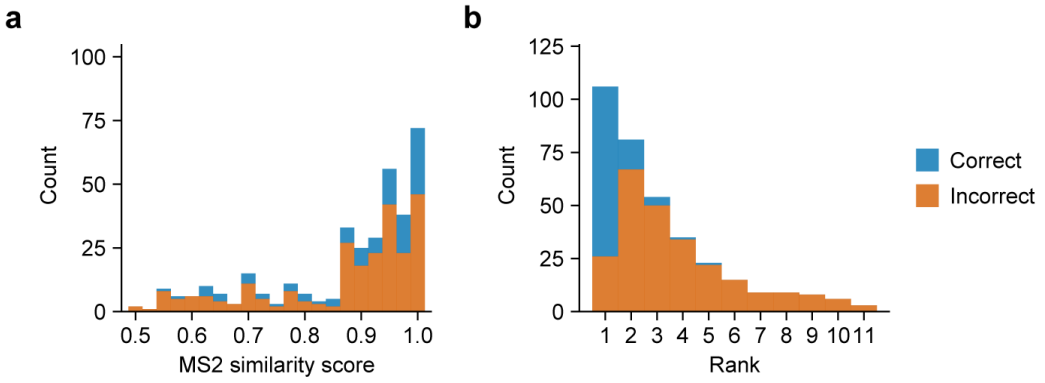
Supplementary Figure 11. Comparison of different knowledge-driven networks for network-based annotation. (**a-d**) Numbers of metabolites (a), numbers of edges (b), average degrees (c) and network densities (d) of different knowledge-driven networks. MRN, metabolic reaction network; SMN, structure-guided molecular network; FCN, fully connected network.



Supplementary Figure 12. Decoy MS2 spectra-based assessment of false positives in network-based metabolite annotation propagation. The workflow of experimental and decoy MS2 spectra-based annotation propagation using MetDNA3. Decoy MS2 spectra were generated by spectrum-based method proposed by Scheubert et al.[3]. Details were described in the Methods section. Illustrations created in BioRender. Zhu, Z. (2025) https://BioRender.com/aeyidih.



Supplementary Figure 13. Evaluation of multiple in-silico tools used for cross-tool concordance. (**a**) Workflow for performance validation. (**b**) Annotation coverage. (**c**) Correct rate for Top 1 annotations. (**d**) Correct rate for Top 3 annotations.



Supplementary Figure 14. Comparison of distributions between correct annotations and incorrect candidates in the NIST urine HILIC–MS(+) Orbitrap dataset. (**a**) Distribution of MS2 similarity scores. (**b**) Distribution of candidate ranks. A total of 100 correct annotations and 249 incorrect candidates were included in the comparison.

Supplementary Tables

Supplementary Table 1. Topological properties of knowledge networks.

		Network of knowledge databases	MRN (known+unknown)	MRN (known)	SMN (known)	FCN (known)
Source	Metabolite database	KEGG; MetaCyc	KEGG; MetaCyc; HMDB	KEGG; MetaCyc; HMDB	KEGG; MetaCyc; HMDB	KEGG; MetaCyc; HMDB
	Edge relationship	Reported RP	Reported/predicted/ BioTransformer RP	Predicted RP	Structural similarity	Fully connected
Information	Node	11532	765755	47748	53506	53583
	Edge	14974	2437884	1447443	27582947	1435542153
	# connected components	528	959	1090	8	1
	Node in the largest connected component	9795	734870	44213	53491	53583
Connectivity	Avg. degree	2.60	6.37	60.63	1031.02	53582
	Density	2.25E-04	8.31E-06	1.27E-03	1.93E-02	1
	Global clustering coefficient	3.33E-02	4.30E-01	4.54E-01	5.77E-01	1
	Avg. path length	11.09	9.14	6.19	3.85	1
	Diameter	42	24	21	12	1
Community	# communities	601	5536	1181	22	1
	Min community size	2	2	2	2	53583
	Max community size	791	69036	9696	23820	53583
	Avg. community size	19.19	155.77	40.43	2432.09	53583
	Imbalance	39.55	442.54	236.49	9.79	0
	Edge cut	1384	133319	117497	3084120	0
	Edge cut (portion)	9.24%	5.47%	8.12%	11.18%	0
	Modularity	0.88	0.87	0.75	0.63	0

Supplementary Table 2. Number of metabolites used for validation in Figure 3d-g from Orbitrap datasets.

Sample	Data acquisition	# annotated metabolites[a]	# Seed metabolites (30%)	# Validation metabolites (70%)
BV2 cell	HILIC–MS(+)	205	61	144
BV2 cell	HILIC–MS(-)	229	68	161
BV2 cell	RP–MS(+)	124	37	87
BV2 cell	RP–MS(-)	201	60	141
Mouse brain	HILIC–MS(+)	287	86	201
Mouse brain	HILIC–MS(-)	273	81	192
Mouse brain	RP–MS(+)	171	51	120
Mouse brain	RP–MS(-)	247	74	173
Mouse liver	HILIC–MS(+)	304	91	213
Mouse liver	HILIC–MS(-)	351	105	246
Mouse liver	RP–MS(+)	150	45	105
Mouse liver	RP–MS(-)	304	91	213
NIST plasma	HILIC–MS(+)	201	60	141
NIST plasma	HILIC–MS(-)	236	70	166
NIST plasma	RP–MS(+)	153	45	108
NIST plasma	RP–MS(-)	249	74	175
NIST urine	HILIC–MS(+)	238	71	167
NIST urine	HILIC–MS(-)	281	84	197
NIST urine	RP–MS(+)	234	70	164
NIST urine	RP–MS(-)	267	80	187
Sum		4705	1404	3301
Unique[b]		1019	671	931

Note: [a] We selected metabolite annotations corresponding to the $[M+H]^+$ or $[M-H]^-$ adduct forms. [b] Duplicated metabolite structures and stereoisomers were removed when counting unique metabolites.

Supplementary Table 3. Parameter settings of BioTransformer.

Parameter	Value	Description
numOfThreads	8	
timeLimit	36000	
Cyp450Mode	3	CYP450 prediction mode. "3" means "Combined": CypReact + BioTransformer rules + CyProducts
NumIterations	2	The number of steps for the prediction.
useDatabase	true	

Supplementary Table 4. Optimized parameters for graph neural network-based prediction model.

Parameter	Range	Optimized value
No. of epochs	(5, 25)	22
Learning rate	(1e-5, 1e-3)	3.4658e-4
Batch size	32, 48, 64, 80, 96, 112	32
No. of convolution layers	(2, 4)	3
Unit of convolution layers	(128, 257)	144
No. of dense layers	(2, 4)	2
Unit of dense layers	(128, 513)	512

Supplementary Table 5. Parameter settings of four bioinformatic tools.

Tool	Parameter
CFM-ID[4] (version 4.0)	num_highest = -1
	ppm_mass_tol = 10
	abs_mass_tol = 0.01
	prob_thresh = 0.001
	param_file = ' param_output0.log'
	config_file = 'param_config.txt'
	score_type = 'Jaccard'
	apply_postprocessing = 1
MetFrag[5] (version 2.4.6-CL)	PrecursorIonMode = 1/-1
	IsPositiveIonMode = True/False
	DatabaseSearchRelativeMassDeviation = 15
	FragmentPeakMatchAbsoluteMassDeviation = 0.01
	FragmentPeakMatchRelativeMassDeviation = 15
	MaximumTreeDepth = 2
	NumberThreads = 4
	MetFragCandidateWriter = CSV
	MetFragDatabaseType = LocalCSV
	MetFragPreProcessingCandidateFilter =
	UnconnectedCompoundFilter,IsotopeFilter
	MetFragScoreWeights = 1.0
	MetFragScoreTypes = FragmenterScore
MS-FINDER[6] (version 3.61)	IsUserDefinedDB=True
	AllProcess=True
	FormulaFinder=True
	StructureFinder=True
	TryTopNMolecularFormulaSearch=5
	Other parameters: default
SIRIUS[7] (version 6.0.6)	Adduct: $[M+H]^+$/$[M-H]^-$
	Other parameters: default

Supplementary Table 6. Five putative unknown metabolites filtered for validation.

Sample and dataset	Feature	Putative annotation	MetFrag score	MS-FINDER score	SIRIUS score
BV2 cell HILIC–MS(+)	M168T345	4-(2-aminoallyl)-1H-imidazole-1-carboxylic acid	1	6.9866	-144.362
Mouse liver HILIC–MS(+)	M168T351		1	7.0026	-161.606
Mouse brain HILIC–MS(-)	M182T222	N-glycolyltaurine	1	7.4138	-95.755
Mouse liver HILIC–MS(-)	M182T223		1	6.1993	-84.234
NIST urine HILIC–MS(-)	M182T223		0.9975	6.2455	-84.334
Mouse brain HILIC–MS(+)	M306T438	γ-Glu-Thr-Gly	1	5.8417	-61.298
Mouse liver HILIC–MS(+)	M306T438		1	6.8446	-46.619
BV2 cell HILIC–MS(+)	M313T406	N-acetyl-Asp-His	1	5.2773	-119.134
Mouse brain HILIC–MS(+)	M313T407		1	6.868	-69.778
BV2 cell HILIC–MS(-)	M610T486	N5-(3-((2-(4-carboxy-4-oxobutanamido)-3-((carboxymethyl)amino)-3-oxopropyl)disulfaneyl)-1-((carboxymethyl)amino)-1-oxopropan-2-yl)glutamine	1	5.8591	-80.571
Mouse liver HILIC–MS(-)	M610T487_1		1	5.9215	-81.546

References

1. Winter, R., Montanari, F., Noé, F. & Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **10**, 1692–1701 (2019).

2. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminformatics* **8**, 61 (2016).

3. Scheubert, K. *et al.* Significance estimation for large scale metabolomics annotations by spectral matching. *Nat. Commun.* **8**, 1494 (2017).

4. Wang, F. *et al.* CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal. Chem.* **93**, 11692–11700 (2021).

5. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminformatics* **8**, 3 (2016).

6. Lai, Z. *et al.* Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat. Methods* **15**, 53–56 (2018).

7. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).