

Supplementary information for

Structure-informed deep generation enables de novo metabolite annotation in untargeted metabolomics

Hongmiao Wang^{1,2}, Haosong Zhang^{1,2}, 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 China

² University of Chinese Academy of Sciences, Beijing, 100049 China

³ Shanghai Key Laboratory of Aging Studies, Shanghai, 201210 China

Corresponding Author

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

List for Supplementary Figures

Supplementary Figure 1. Workflow and parameter optimization for template search.

Supplementary Figure 2. Workflow of template encoding.

Supplementary Figure 3. Workflow of structure generation.

Supplementary Figure 4. Comparison of structural similarity and chemical space coverage between the pre-training and the fine-tuning datasets.

Supplementary Figure 5. Feature importance analysis of MetGenX.

Supplementary Figure 6. Impact of template number on generation performance in the NIST testing set.

Supplementary Figure 7. Annotation performance of MetGenX across different biological samples.

Supplementary Figure 8. Comparison of annotation coverage and annotation accuracy across different in-silico annotation tools.

Supplementary Figure 9. Standard validation of generated metabolites absent from the NIST spectral database.

Supplementary Figure 10. Metabolites discovery in mouse liver.

Supplementary Figure 11. Validation of discovered metabolites in mouse liver.

Supplementary Figure 12. MS2 spectra of adenosine 5'-diphosphate 5'-inosine (Ap2I) acquired at different collision energies.

Supplementary Figure 13. Quantification of discovered metabolites in mouse liver.

Supplementary Figure 14. Annotation performance of MetGenX using a more biologically relevant structural database.

List for Supplementary Tables

Supplementary Table 1. Database-restricted generation performance on MassSpecGym dataset

Supplementary Table 2. Database-free generation performance on MassSpecGym dataset

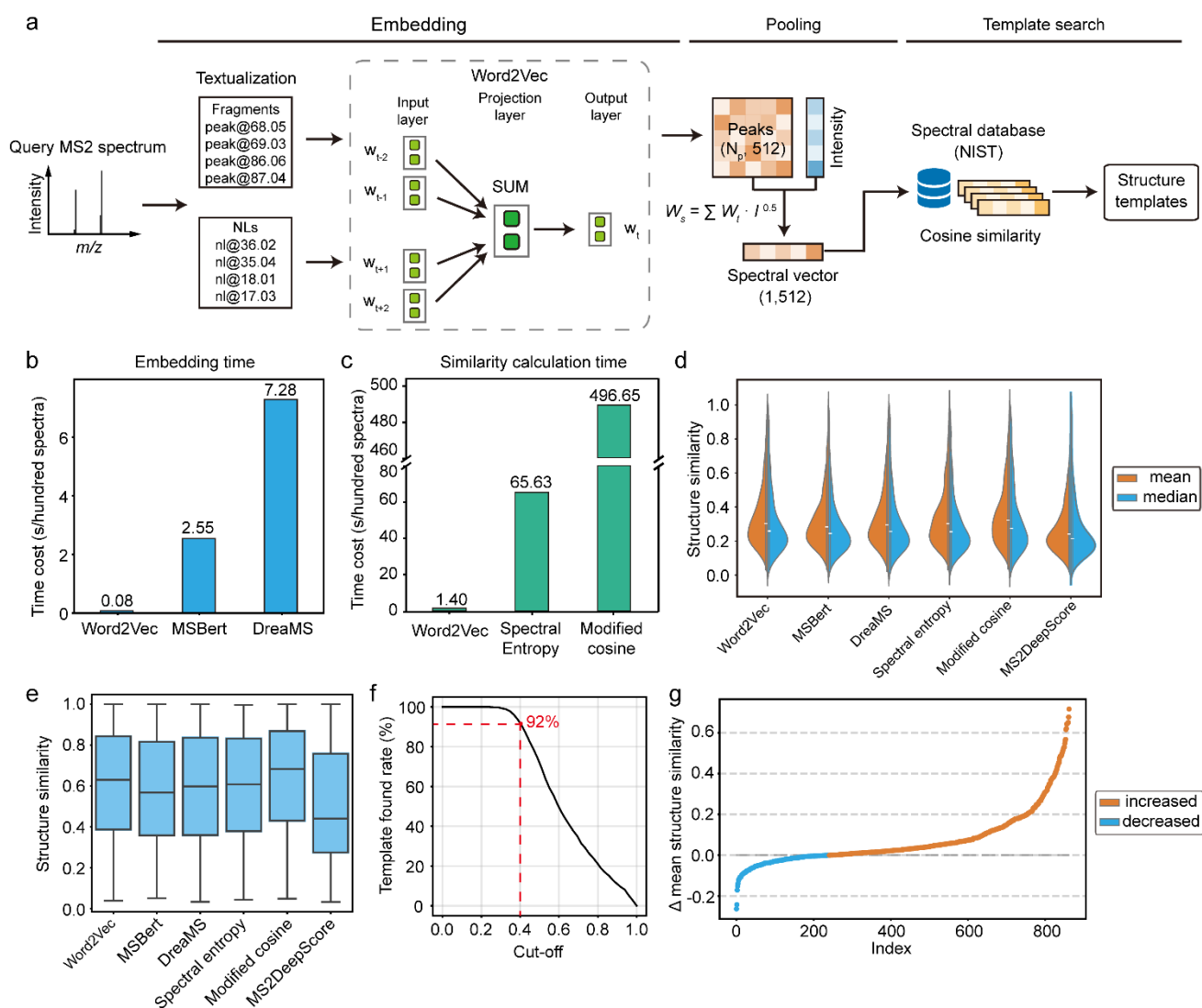
Supplementary Table 3. Adduct type and number across different spectral datasets in this work

Supplementary Table 4. Parameter settings for MetGenX

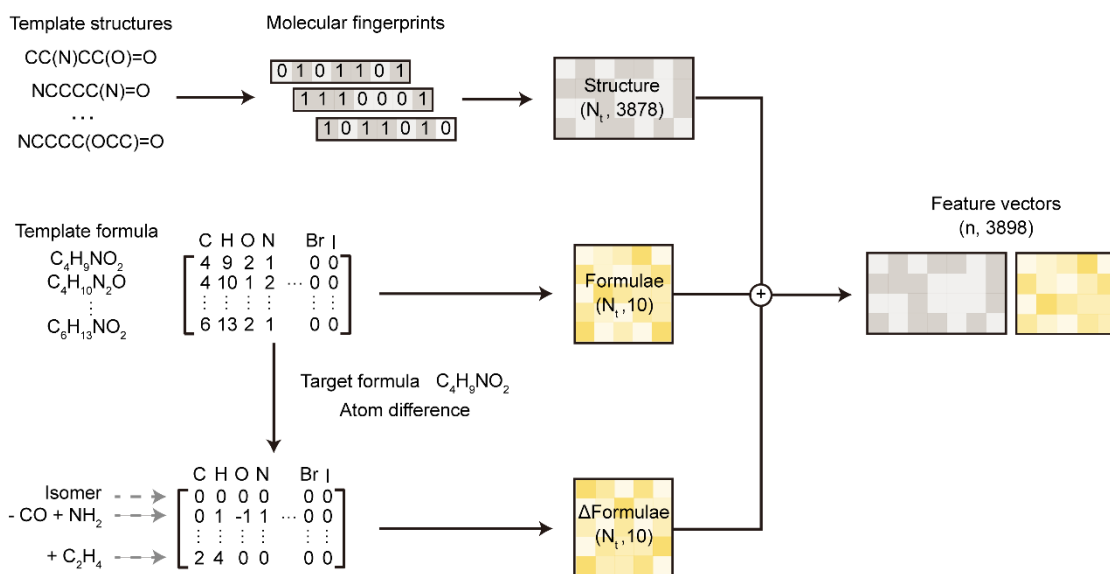
Supplementary Table 5. Parameter settings for spectral similarity approaches

Supplementary Table 6. Parameter settings for in silico annotation tools

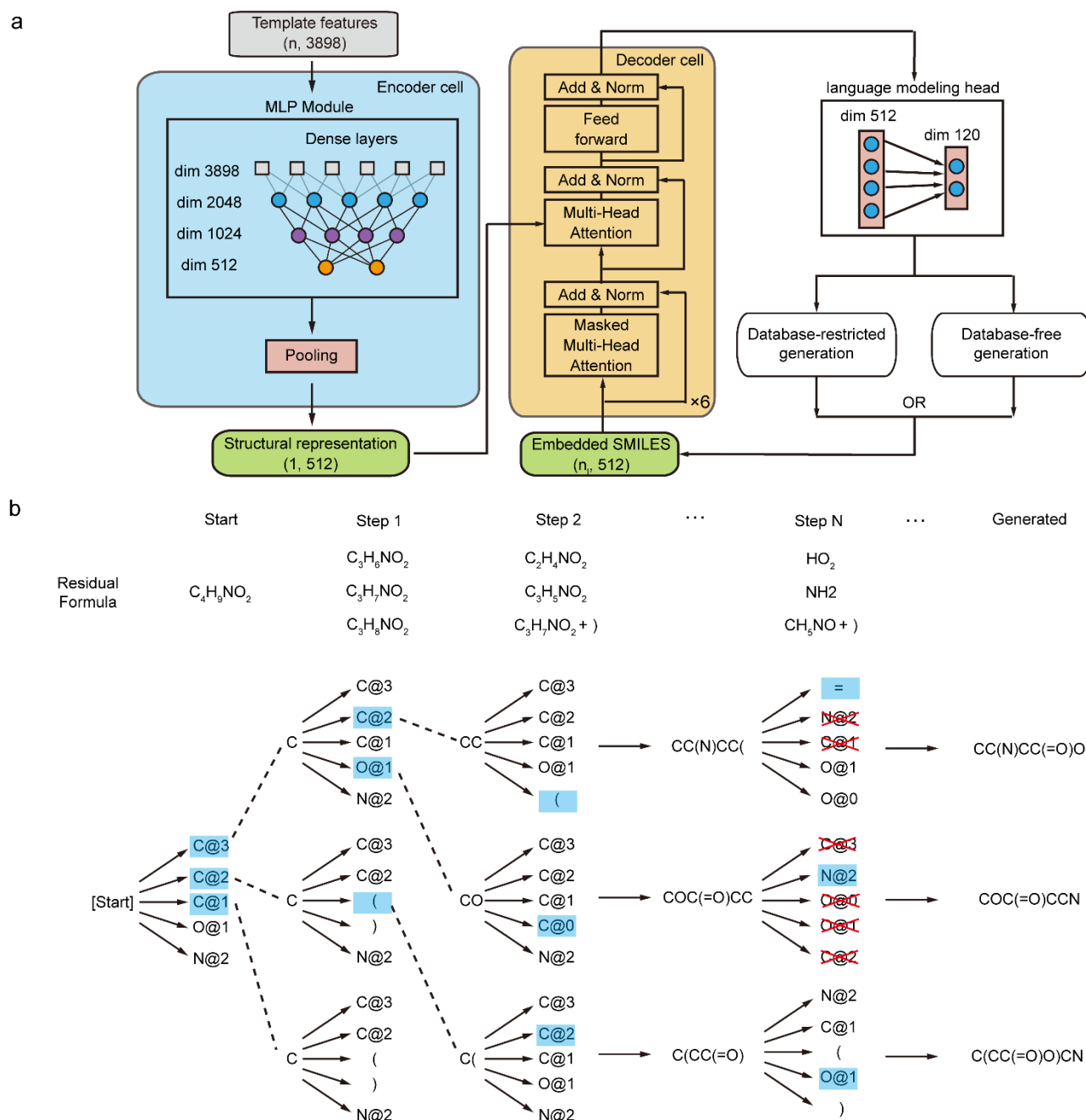
Supplementary Figures



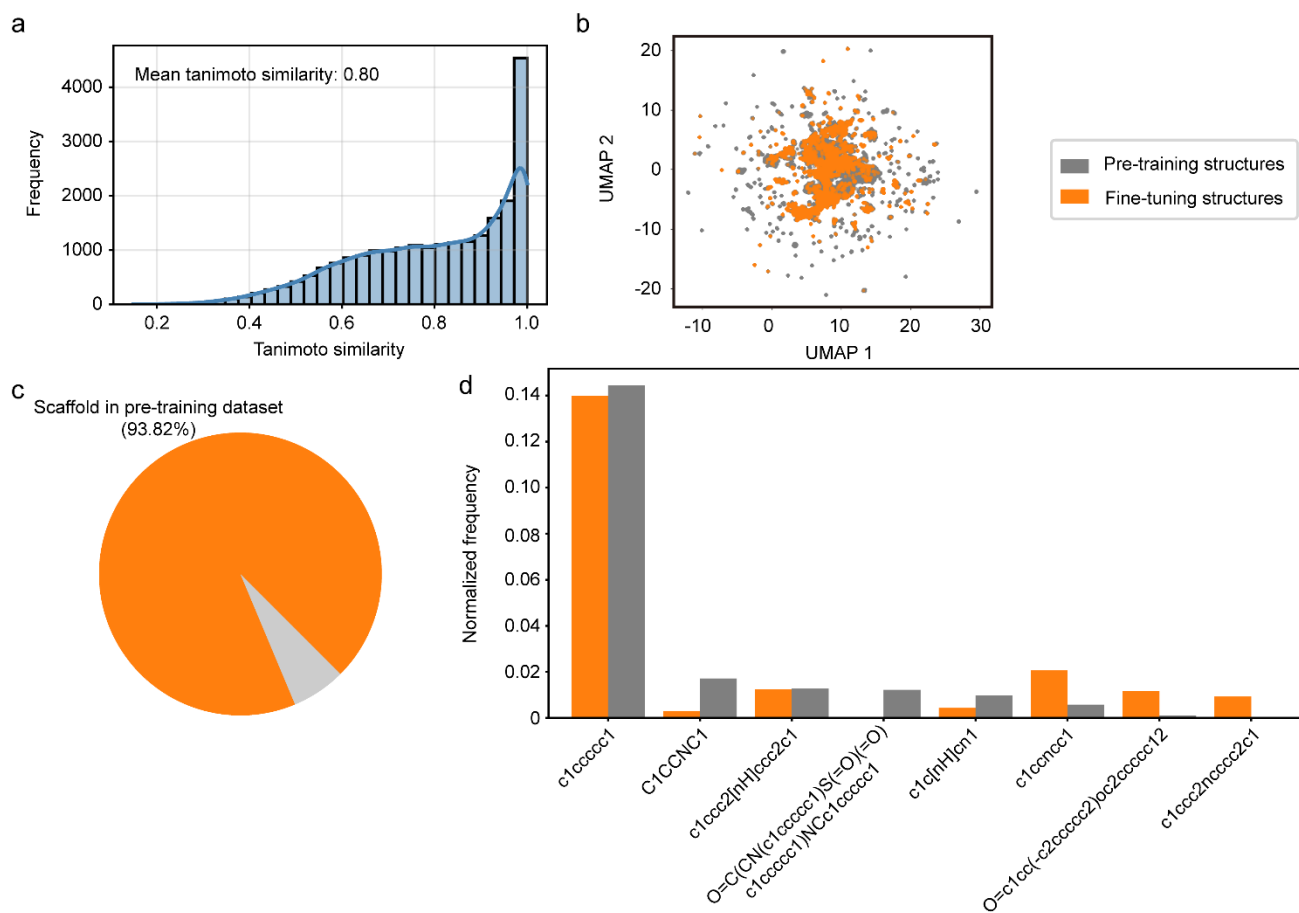
Supplementary Figure 1. Workflow and parameter optimization for template search. (a) Schematic overview of the template search workflow. The query MS2 spectrum is textualized, embedded and pooled into a spectral vector to retrieve structure templates. (b, c) Embedding time (b) and similarity calculation time (c) across different spectral embedding methods. (d) Mean and median structural similarity between retrieved templates and targeted structures for different template search methods in the NIST validation set. (e) Distribution of maximum structural similarity for different template search methods in the NIST validation set. (f) Template found rate in the NIST validation set under varying spectral similarity thresholds. (g) Change in mean structural similarity before and after applying the 0.4 spectral similarity threshold for spectra with altered templates.



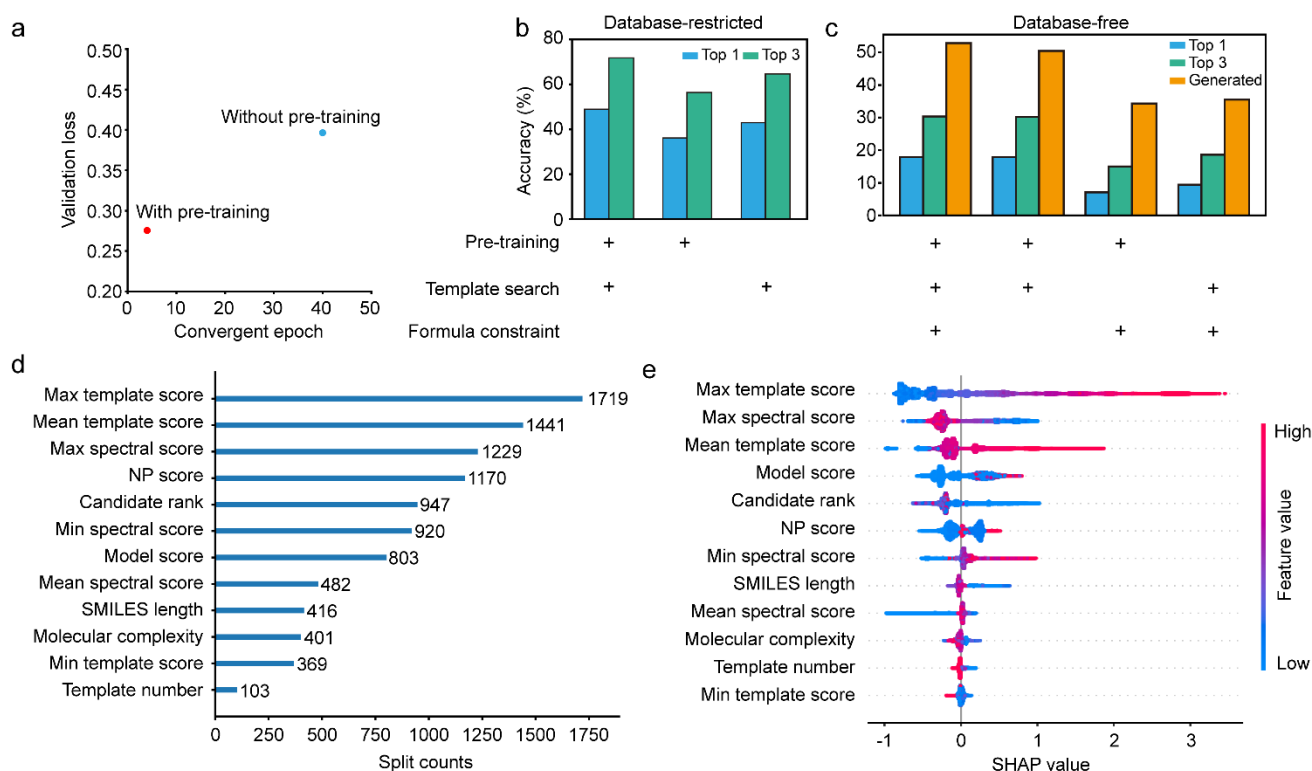
Supplementary Figure 2. Workflow of template encoding. Template structures are encoded into a 3,898-dimensional feature vector, including 3,878 bits for structural fingerprints, 10 bits for molecular formula, and 10 bits for formula differences. The vectors from all retrieved templates are concatenated as feature vectors.



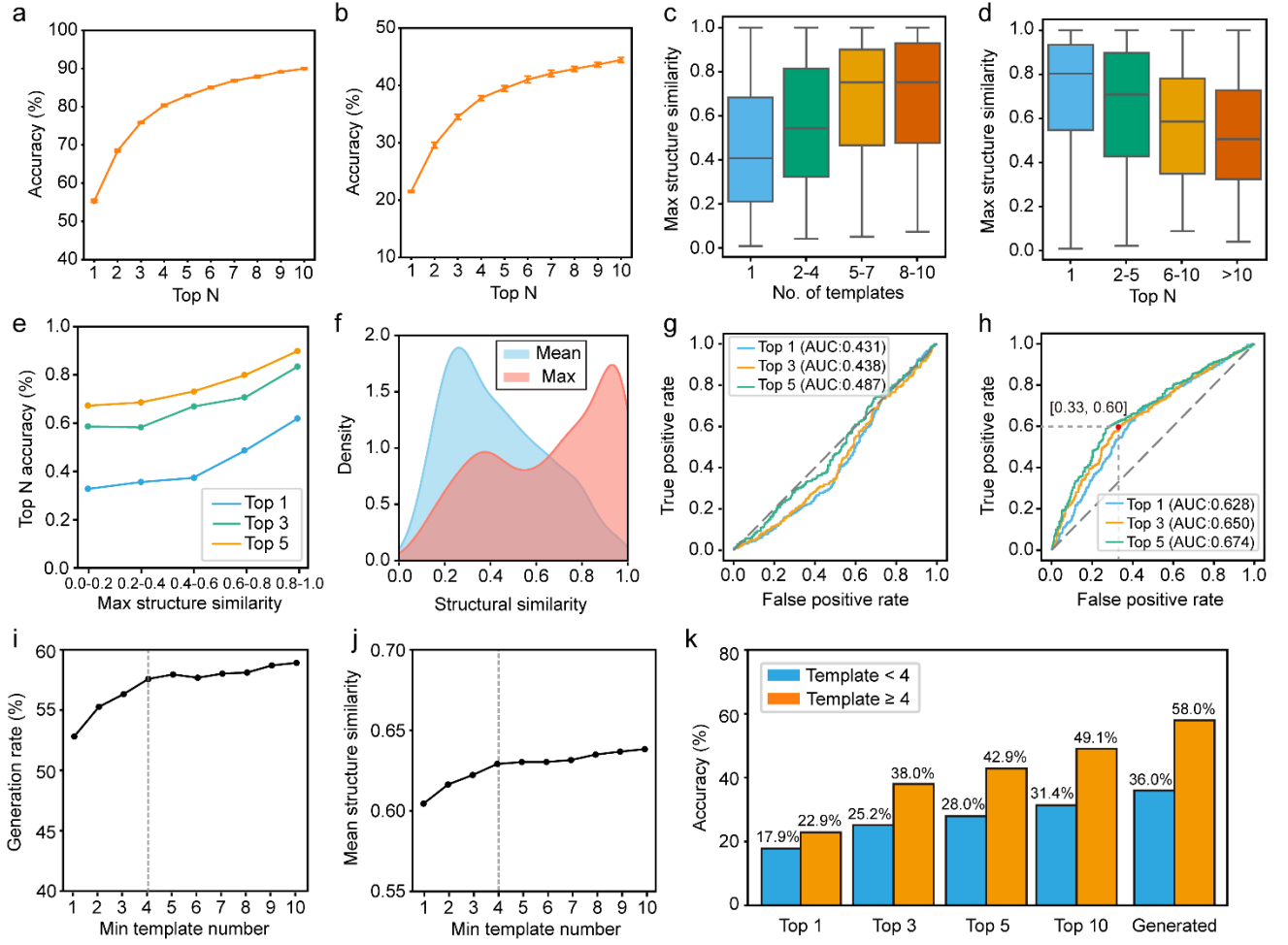
Supplementary Figure 3. Workflow of structure generation. (a) Architecture of the encoder-decoder generative model. The extracted template features are encoded into structural representation using a multilayer perceptron (MLP) encoder. These representations are then fed into a transformer decoder for structure generation. (b) Schematic illustration of formula-constrained autoregressive generation in database-free mode. Tokens beyond the allowed elemental composition are blocked from further expansion at each decoding step.



Supplementary Figure 4. Comparison of structural similarity and chemical space coverage between the pre-training and the fine-tuning datasets. (a) Maximum Tanimoto similarity for fine-tuning structures compared with pre-training structures. The identical structures in pre-training structures were excluded. (b) UMAP projection of fingerprints in fine-tuning datasets covered the chemical space of the pre-training dataset (sampled 100,000 molecules) (c) Proportion of scaffolds in the fine-tuning dataset that are present in the pre-training dataset. (d) Normalized frequency of the top-5 scaffolds across two datasets.

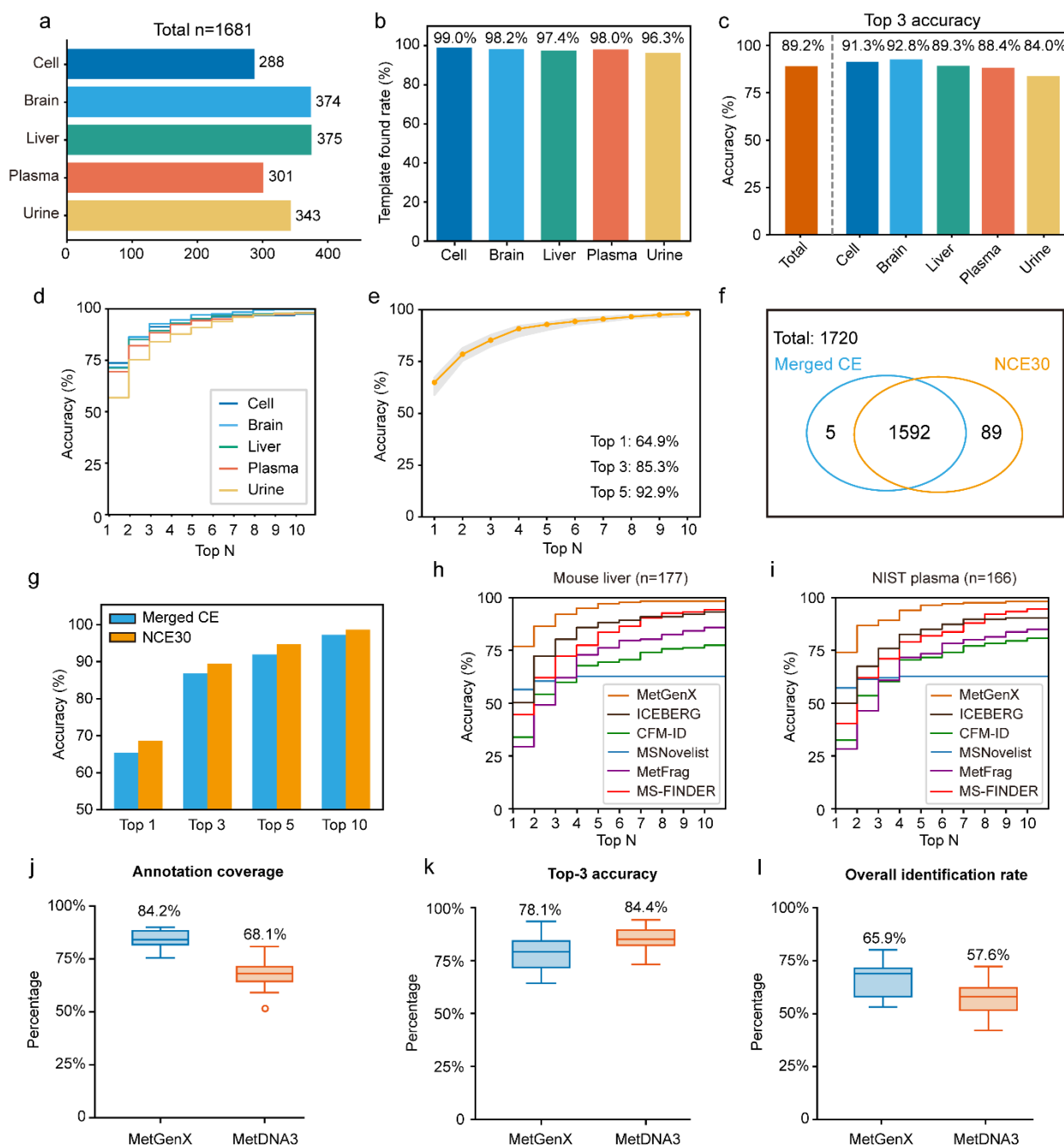


Supplementary Figure 5. Feature importance analysis of MetGenX. (a) Comparison of convergence epochs and validation loss for the generative model with and without pre-training. The generated rate means the percentage of true structures can be found across all candidates. (b, c) Top N accuracy on the NIST testing set under different ablation settings for database-restricted (b) and database-free (c) modes. (d, e) Feature importance of the re-ranking model, including feature split counts (d) and SHAP values (e).

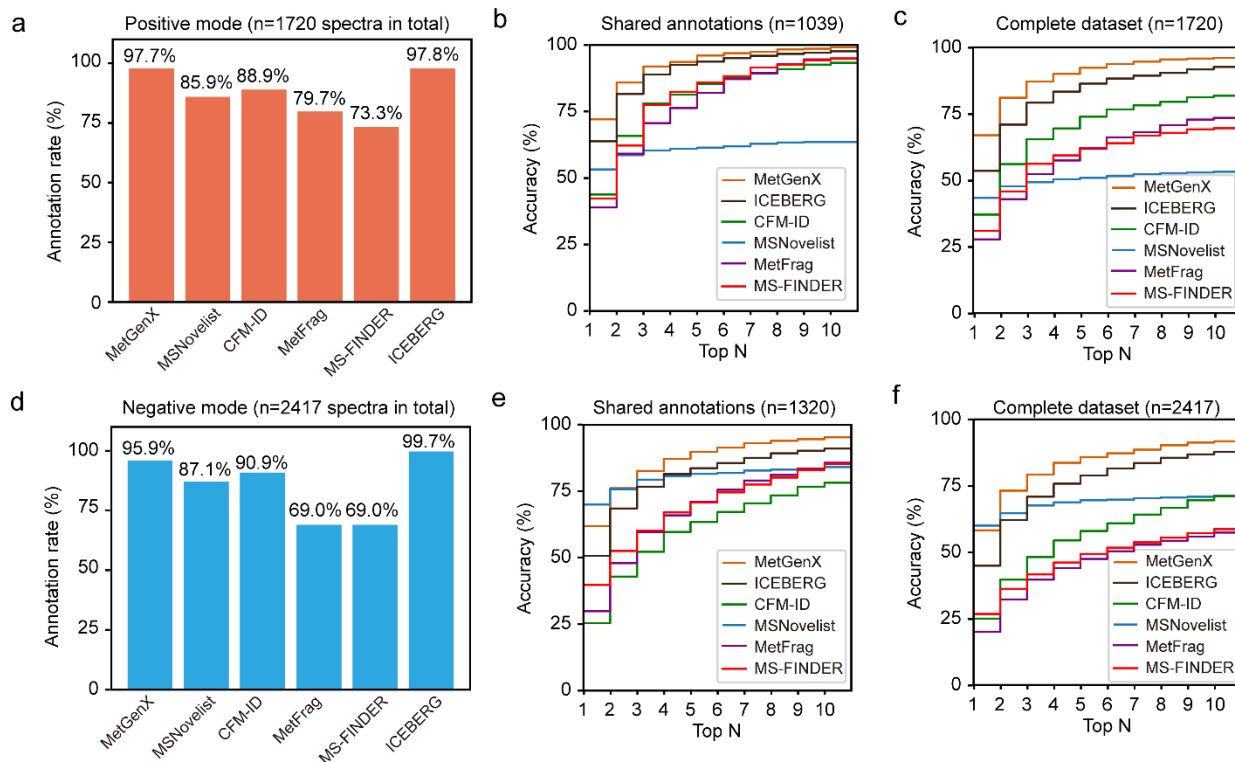


Supplementary Figure 6. Impact of template number on generation performance in the NIST testing set.

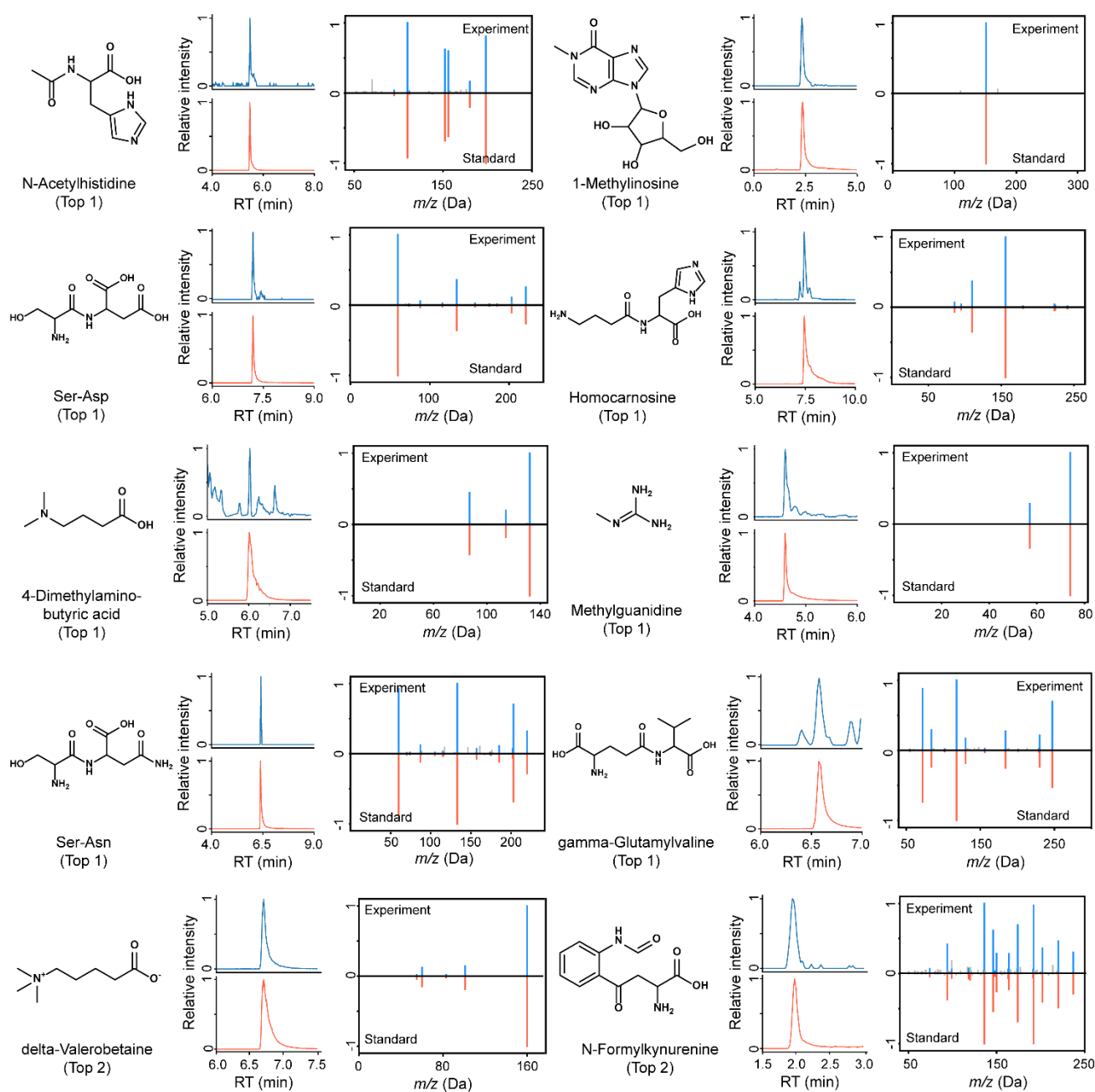
(a, b) Annotation accuracies of the database-restricted generation (a) and database-free generation (b) after re-ranking in different random seeds ($n=5$). (c, d) Maximum structural similarity between templates and target structures across different numbers of templates (c) and across different candidate ranks (d) in database-restricted mode. (e) Top-N accuracy in database-restricted mode stratified by the maximum structural similarity between templates and target structures. Only spectra yielding at least two candidates were included. (f) Structural similarity between the retrieved templates and the targeted structures. (g, h) ROC curves of maximum spectral similarity (g) and maximum structure similarity (h) for templates to predict correct generation (top-1, top-3, and top-5). Red dots mark optimal thresholds of correct generation across top-3 candidates. Only spectra yielding at least two candidates were included. (i, j) Generation rate (i) and mean structural similarity between the top-1 generated candidates and target structures (j) across different template number thresholds for database-free generation. (k) Annotation accuracy in database-free mode when using more than or fewer than four templates.



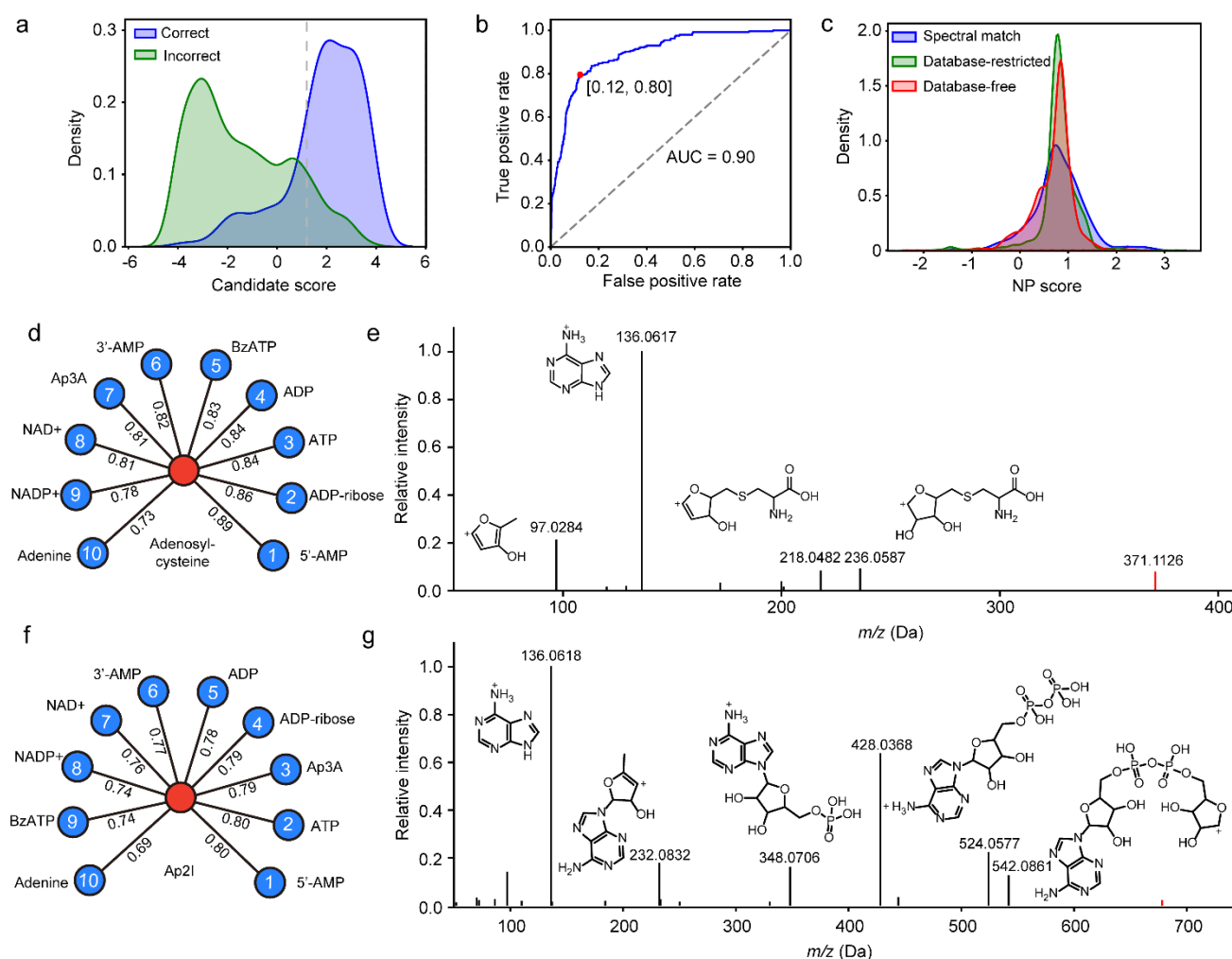
Supplementary Figure 7. Annotation performance of MetGenX across different biological samples. (a, b) Number of level 1 annotated spectra generated by MetGenX (a) and template found rate (b) across different sample types. (c) Top-3 accuracy for all spectra and each sample type individually. (d) Top-N accuracy of MetGenX stratified by sample types. (e) Top-N accuracy of MetGenX stratified by metabolites. The orange line with scatter points represents the accuracy of mean rank, while the shaded region represents the range between the accuracy of minimum and maximum. (f, g) Number of annotated spectra (f) and corresponding annotation accuracy (g) when using merged spectra versus spectra acquired at NCE 30% as reference spectral databases. (h, i) Annotation accuracy for spike-in metabolites in mouse liver (h, n=177) and NIST plasma (i, n=166) samples across different in-silico annotation tools. (j, k, l) Comparative evaluations of annotation coverage (j), top-3 accuracy (k) and top-3 overall identification rate (l) between MetGenX and MetDNA3 (20 datasets per group). The mean of each group is shown above each box. The overall identification rate means percentage of total correctly identified spectra across all input spectra.



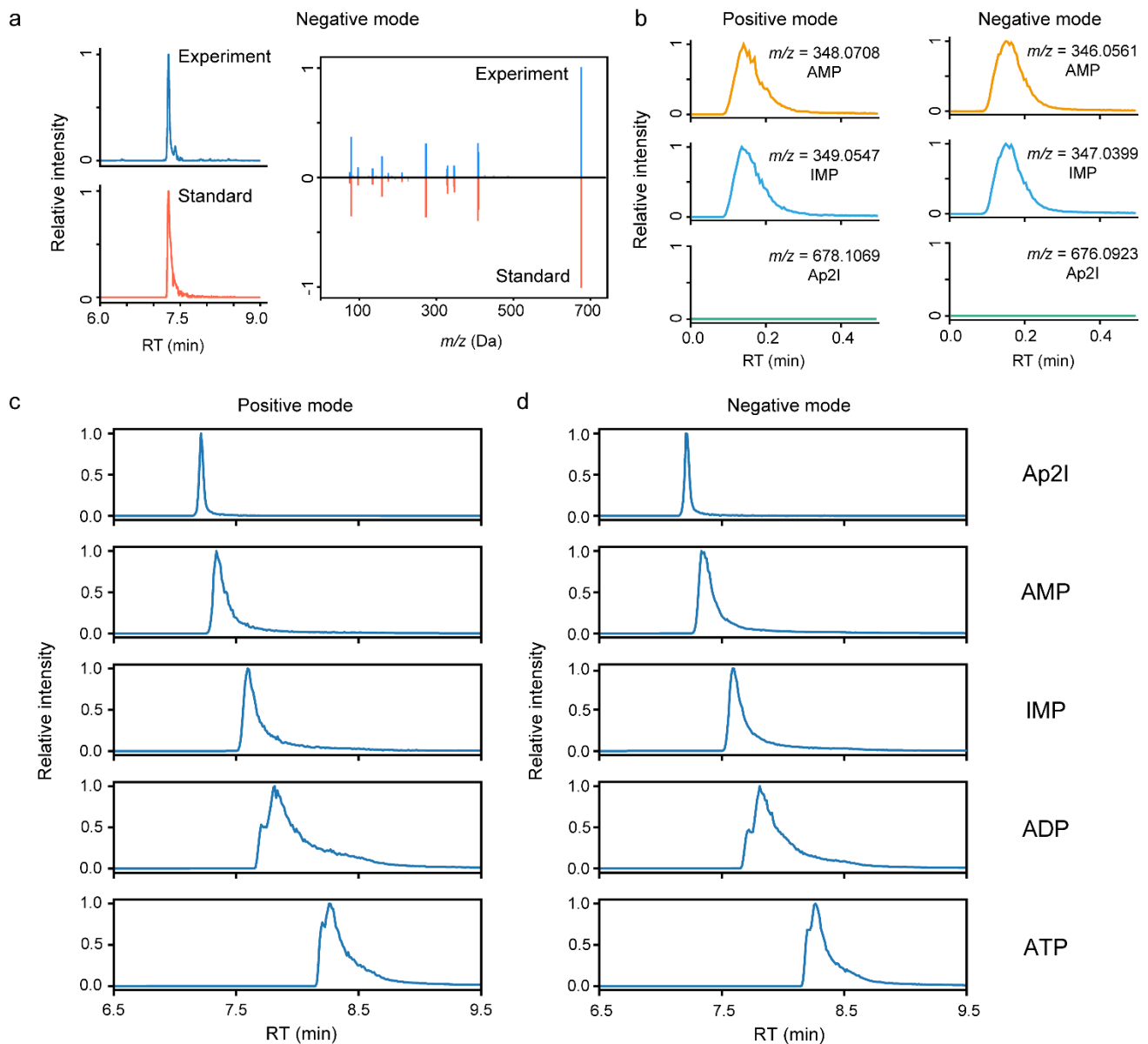
Supplementary Figure 8. Comparison of annotation coverage and annotation accuracy across different in-silico annotation tools. (a, d) Annotation rate across different in-silico annotation tools in positive (a) and negative (d) ionization modes. (b, e) Annotation accuracy for shared annotation across different in-silico annotation tools in positive (a) and negative (d) ionization modes. (c, f) Annotation accuracy for complete dataset across different in-silico annotation tools in positive (c) and negative (f) ionization modes.



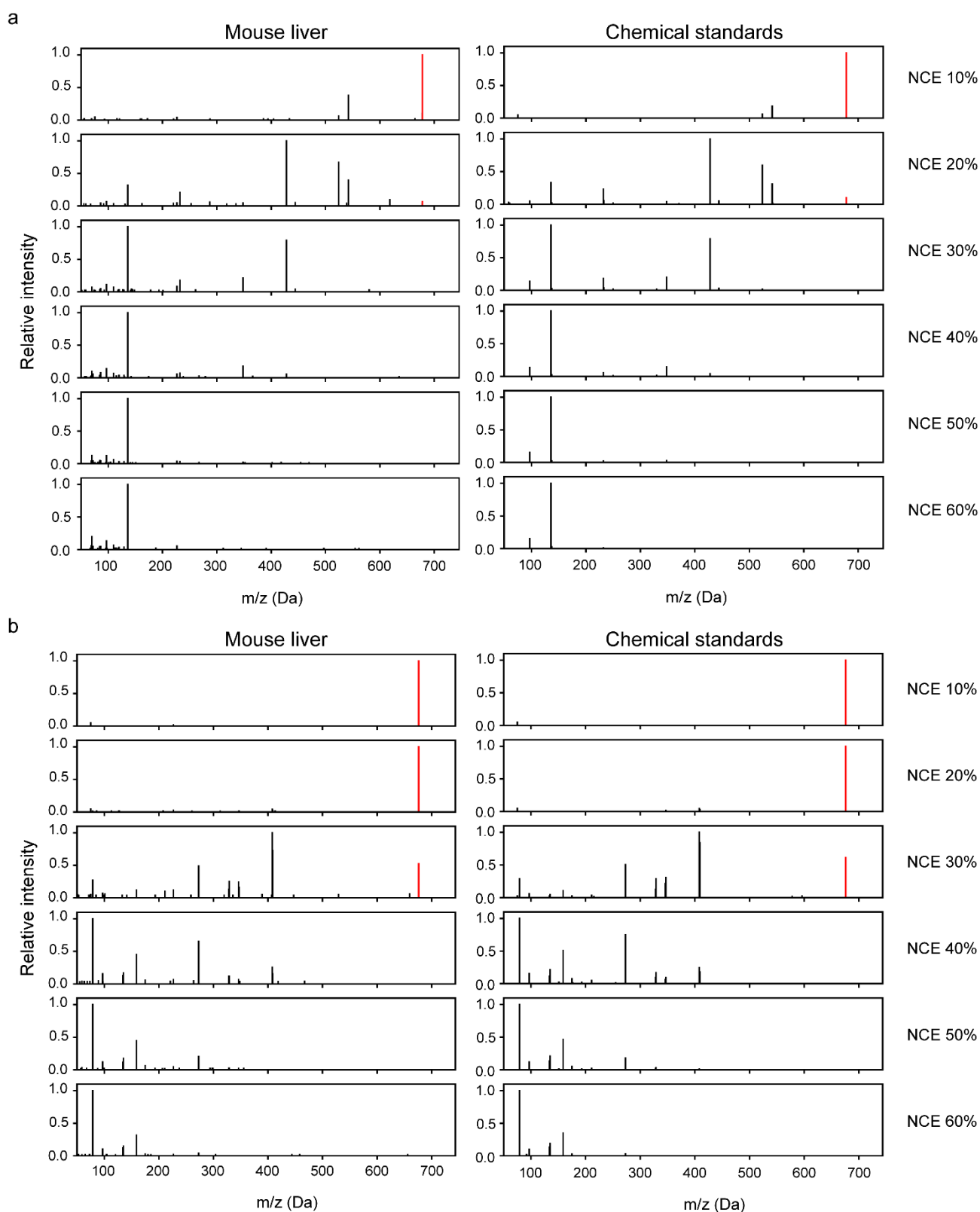
Supplementary Figure 9. Standard validation of generated metabolites absent from the NIST spectral database.



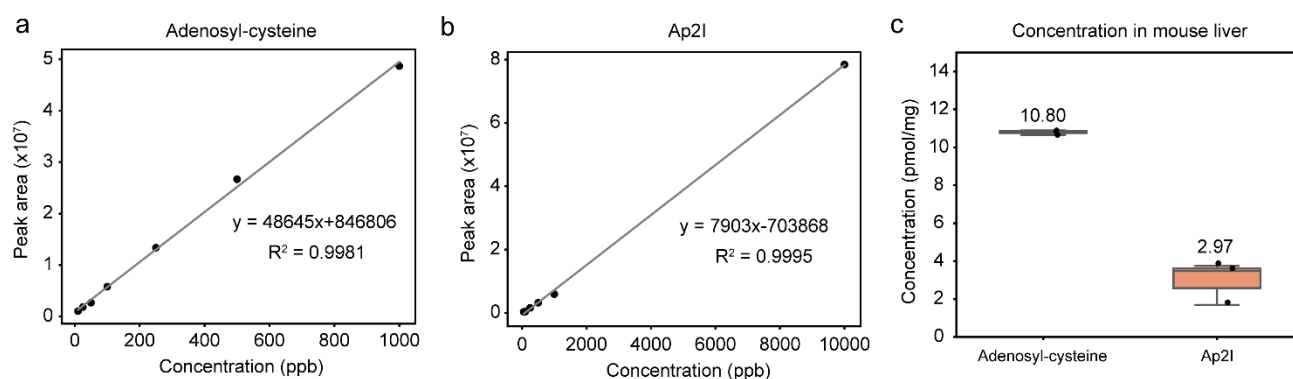
Supplementary Figure 10. Metabolites discovery in mouse liver. (a, b) Density distribution (a) and ROC curve (b) of candidate scores for correctly (blue) and incorrectly (green) annotated metabolites, based on generated results for identified metabolites in the mouse liver dataset. (c) Natural product (NP) score distributions for library-based spectral matching (blue), and high-quality candidate structures ($S > 1.2$) generated from database-restricted (green) and database-free (red) modes. (d) Templates used for generating adenosyl-cysteine. Values indicate the spectral similarity between each template and the query MS2 spectrum. (e) Manual interpretation of MS2 fragment ions for adenosyl-cysteine. (f) Templates used for generating adenosine 5'-diphosphate 5'-inosine (Ap2l). Values indicate the spectral similarity between each template and the query MS2 spectrum. (g) Manual interpretation of MS2 fragment ions for Ap2l.



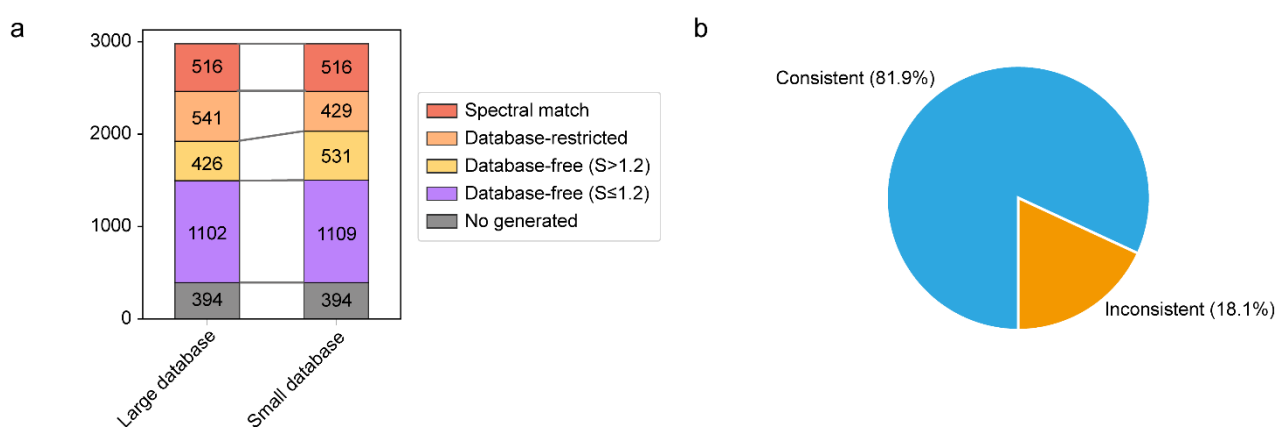
Supplementary Figure 11. Validation of discovered metabolites in mouse liver tissue. (a) Standard validation of adenosine 5'-diphosphate 5'-inosine (Ap2I) using a chemical standard by comparing the retention time and MS2 spectral matches in negative ionization mode. (b) Extracted ion chromatograms of the mixture of nucleotide precursors (AMP and IMP) for Ap2I acquired in both positive (left) and negative (right) ionization modes. (c) Extracted ion chromatograms of Ap2I, AMP, IMP, ADP and ATP in positive (left) and negative (right) ionization modes.



Supplementary Figure 12. MS2 spectra of adenosine 5'-diphosphate 5'-inosine (Ap2I) acquired at different collision energies. (a, b) MS2 spectra of Ap2I acquired in mouse liver (left) and from the chemical standard (right) under positive (a) and negative (b) ionization modes at different collision energies (NCE 10%, 20%, 30%, 40%, 50%, and 60%).



Supplementary Figure 13. Quantification of discovered metabolites in mouse liver. (a, b) standard curves of synthetic standards for Adenosyl-cysteine (a) and Ap2I (b). (c) concentrations of two metabolites in mouse liver tissue (n=3).



Supplementary Figure 14. Annotation performance of MetGenX using a more biologically relevant structural database. (a) Spectral annotated for mouse liver tissue through a large structural database (0.52 M structures) and more biological relevant small database (66,874 structures). (b) The consistent annotation and inconsistent annotation result across top-3 annotation in database-restricted mode after changed database.

Supplementary Table 1. Performances for database-restricted generation on the MassSpecGym testing dataset. The best performing model for each metric is highlighted as bold.

Model	Hit rate @ 1 ↑	Hit rate @ 5 ↑	Hit rate @ 20 ↑	MCES @ 1 ↓	Reference
Random	3.06%	11.35%	27.74%	13.87	Bushuiev et al., 2024 ¹
DeepSets	4.42%	14.46%	30.76%	15.04	
Fingerprint FFN	5.09%	14.69%	31.97%	14.94	
DeepSets + Fourier features	6.56%	16.46%	33.46%	14.14	
MIST	9.57%	22.11%	41.12%	12.75	
MetGenX	18.34%	31.39%	48.43%	11.89	This work
MetGenX-Full	17.85%	30.30%	48.31%	11.98	

Supplementary Table 2. Performances for database-free generation on the MassSpecGym testing dataset. The best performing model for each metric is highlighted as bold.

Model	Top-1			Top-10			Reference
	Accuracy ↑	MCES ↓	Tanimoto ↑	Accuracy ↑	MCES ↓	Tanimoto ↑	
SMILES Transformer	0.00%	79.39	0.03	0.00%	52.13	0.10	Bushuiev et al., 2024 ¹
MIST + MSNovelist	0.00%	45.55	0.06	0.00%	30.13	0.15	
SELFIES Transformer	0.00%	38.88	0.08	0.00%	26.87	0.13	
Spec2Mol	0.00%	37.76	0.12	0.00%	29.40	0.16	
MIST + Neuraldecipher	0.00%	33.19	0.14	0.00%	31.89	0.16	Bohde MM et al., 2025 ²
Random Generation	0.00%	21.11	0.08	0.00%	18.26	0.11	
MADGEN	1.31%	27.47	0.20	1.54%	16.84	0.26	This work
DiffMS	2.30%	18.45	0.28	4.25%	14.73	0.39	
MetGenX	2.50%	20.92	0.22	4.62%	17.33	0.28	
MetGenX-Full	2.22%	20.65	0.22	4.25%	17.18	0.28	

Note: For MetGenX-Full, the similarity threshold of 0.4 in the template search of MetGenX was removed and ten structure templates were selected with the highest similarity scores as inputs for MetGenX. With this revision, all input MS2 spectra could be annotated by MetGenX.

Supplementary Table 3. Adduct type and number across different spectral datasets in this work

Adduct type	NIST20	MassSpecGym
[M+H] ⁺	20578	195237
[M+Na] ⁺	1344	35867
[M+K] ⁺	9	/
[M] ⁺	148	/
[M-H ₂ O+H] ⁺	1095	/
[M+2Na-H] ⁺	35	/
[M+2K-H] ⁺	4	/

Supplementary Table 4. Parameter settings for MetGenX

Module	Parameters
Template search-spectral embedding	sg = 0
	negative = 5
	vector_size = 512
	window = 600
	min_count = 10
	alpha = 0.025
	min_alpha = 0.0125
	epochs = 50
	TopN = 100
	min_int = 0.01
	NL_range = [0.5, 200]
	Index = IndexFlatIP
Template encoding	L2_normalize = True
	n_elements = 10
	n_fingerprint = 3878
Structure generation	d_model = 512
	dense_layer = [2048, 1024, 512]
	decoder_ffn_dim = 1024
	decoder_layers = 6
	decoder_attention_heads = 8
	dropout = 0.1
	activation_function = relu
	lr = 0.01 (pre-training) / 5e-5 (fine-tuning)
	num_warmup = 400 (pre-training) / 0 (fine-tuning)
	batch_size = 32
Re-ranking	weight_decay = 1e-6
	boosting type = gbrt
	objective = lambdarank
	metric = ndcg at 10
	max_bin = 255
	num_iterations = 1000
	learning_rate = 0.01
	num_leaves = 11

```
tree_learner = serial  
min_data_in_leaf = 30
```

Supplementary Table 5. Parameter settings for spectral similarity approaches

Tool	Parameter
Modified cosine score	tolerance=0.01
	max_charge_consideration=1
DreaMS (version 1.0.0)	batch_size = 32
	Other parameters: default
Spectral entropy (version 1.3.3)	ms2_tolerance_in_da=0.01
	Other parameters: default
MSbert (1.0)	batch_size = 16
	Other parameters: default
MS2DeepScore (2.0.0)	All parameters: default

Supplementary Table 6. Parameter settings for in silico annotation tools

Tool	Parameter
MetGenX	DB_cutoff = 0.4 rerank = True mode = "Restricted"
	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
CFM-ID (version 4.0)	PrecursorIonMode = 1/-1 IsPositiveIonMode = True/False FragmentPeakMatchAbsoluteMassDeviation = 0.01 FragmentPeakMatchRelativeMassDeviation = 15 MaximumTreeDepth = 2 NumberThreads = 4 useFormula = True MetFragCandidateWriter = CSV MetFragDatabaseType = LocalCSV MetFragPreProcessingCandidateFilter = UnconnectedCompoundFilter, IsotopeFilter MetFragScoreWeights = 1.0 MetFragScoreTypes = FragmenterScore
	IsUserDefinedDB=True AllProcess=True FormulaFinder=True StructureFinder=True TryTopNMolecularFormulaSearch=100 Other parameters: default
MetFrag (version 2.4.6-CL)	
MS-FINDER (version 3.61)	
MSNovelist (SIRIUS version 6.0.6)	MS2 mass accuracy = 10 ppm Other parameters: default

ICEBERG
(<https://github.com/coleygroup/ms-pred>)

DAG Model

batch-size: 32
learning-rate: 0.000996
lr-decay-rate: 0.7214
dropout: 0.2
mpnn-type: GGNN
pe-embed-k: 14
pool-op: avg
set-layers: 0
hidden-size: 512
weight-decay: 0.0
layers: 6
root-encode: gnn
inject-early: False
encode-forms: True
embed-adduct: True
embed-collision: False
embed-elem-group: True
add-hs: true
max-nodes: 100

Intensity model

batch-size: 32
learning-rate: 0.000736
lr-decay-rate: 0.825
dropout: 0.1
mpnn-type: GGNN
pe-embed-k: 0
pool-op: avg
hidden-size: 256
weight-decay: 1e-7
set-layers: 0
frag-set-layers: 3
mlp-layers: 1
gnn-layers: 4
loss-fn: entropy
root-encode: gnn
inject-early: false
include-unshifted-mz: true
binned-targs: true
embed-adduct: true
embed-collision: false
embed-elem-group: true
grad-accumulate: 1
encode-forms: true

add-hs: true
sk-tau: 0.05
ppm-tol: 20

Contrastive finetuning

batch-size: 32
learning-rate: 0.0005
lr-decay-rate: 0.825
contr-weight: 1
dropout: 0.1
mpnn-type: GGNN
pe-embed-k: 0
pool-op: avg
hidden-size: 256
weight-decay: 1e-7
set-layers: 0
frag-set-layers: 3
mlp-layers: 1
gnn-layers: 4
loss-fn: entropy
root-encode: gnn
inject-early: false
include-unshifted-mz: true
binned-targs: true
embed-adduct: true
embed-collision: false
embed-elem-group: true
grad-accumulate: 1
encode-forms: true
add-hs: true
sk-tau: 0.05

Candidates retrieval

batch_size: 8
max_nodes: 100
num_bins: 15000
ppm: 20
sparse_k: 100
threshold: 0.0

References

- 1 Bushuiev R, *et al.* MassSpecGym: A benchmark for the discovery and identification of molecules. Preprint at <https://doi.org/10.48550/arXiv.2410.23326> (2024).
2. Bohde MM, Wang R, Ji S, Coley CW. DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra. Preprint at <https://doi.org/10.48550/arXiv.2502.09571>, (2025).