Lightweight Alignment of Unimodal Foundation Models for Metabolite Identification

Paul Krzakala [1 2] **Gabriel Melo** [1] **Camille Lançon** [3]
Charlotte Laclau [1] **Rémi Flamary** [2] **Etienne Thévenot** [3] **Florence d'Alché-Buc** [1]

Abstract

A central challenge in building multimodal foundation models for the life sciences is the imbalance between abundant unimodal data and scarce paired observations, which limits the scalability of joint multimodal pretraining. We investigate an alternative approach based on aligning pretrained unimodal models. Focusing on metabolite identification, we introduce MSAlign, which maps a molecular transformer (ChemBERTa) and a mass spectra transformer (DreaMS) into a shared embedding space. Despite its simplicity, MSAlign substantially outperforms prior methods across benchmarks, setting a new state-of-the-art in retrieval performance. These results suggest that aligning unimodal foundation models offers an effective route to multimodal learning in biological settings where paired data remain limited.

1. Introduction

Multimodal learning has become a common strategy for combining heterogeneous data sources (Cui et al., 2025). Most existing approaches rely on joint training with paired observations across modalities. However, in many scientific settings, such paired datasets remain limited, while large collections of unimodal data are readily available. This raises a simple question: can we build strong multimodal models by aligning pretrained unimodal models rather than learning everything jointly from paired data? While a few works have recently studied this question in the context of image-text alignment (Roschmann et al., 2026), we propose to study this question by focusing on a well-known task in Metabolomics: Metabolite identification (Brown et al., 2009). Given a mass spectrum, the goal is to identify the molecule that produced it. In practice, this can be formulated as a retrieval problem: a model receives a spectrum and must rank candidate molecules according to how well they match the observed fragmentation pattern. The classical approach relies on spectral library matching, where experimental spectra are compared against reference databases. Although effective, this technique is fundamentally limited by coverage. Despite extensive curation efforts, existing spectral libraries represent only a small fraction of the known chemical space, covering less than 1% of the compounds cataloged in PubChem (Bittremieux et al., 2022). As a result, many spectra cannot be matched to a reference, motivating learning-based approaches that generalize beyond observed examples. Recent progress has been supported by standardized benchmarks and increasingly strong neural architectures (Bushuiev et al., 2024; 2025). Most existing methods learn a similarity function directly from paired spectrum–molecule data, effectively training the multimodal representation from scratch. However, this setup may be inefficient in metabolomics, where large pretrained models already exist independently for molecules and spectra.

In this work, we introduce MSAlign[1], a lightweight framework that aligns pretrained molecular and spectral encoders in a shared embedding space. The unimodal models remain frozen, and only small projection heads are learned. This makes training efficient while allowing the use of a candidate-based contrastive objective tailored to retrieval. Across several metabolite identification benchmarks, MSAlign substantially outperforms prior approaches and establishes a new state of the art. More broadly, our results suggest that aligning pretrained unimodal models may offer a simple and scalable alternative to joint multimodal pretraining in domains where paired data are scarce.

2. Problem Statement

Let $\mathcal{S}$ denote the space of mass spectra and $\mathcal{M}$ the space of molecules. Given a dataset of paired observations $(s_i, m_i)_{i=1}^N \in \mathcal{S} \times \mathcal{M}$, the goal is to learn a model

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris [2]CMAP, Ecole Polytechnique, Institut Polytechnique de Paris [3]CEA, INRAE, MetaboHUB, Université Paris-Saclay. Correspondence to: Paul Krzakala <paul.krzakala@gmail.com>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

[1]The code is available at https://github.com/KrzakalaPaul/MSAlign.

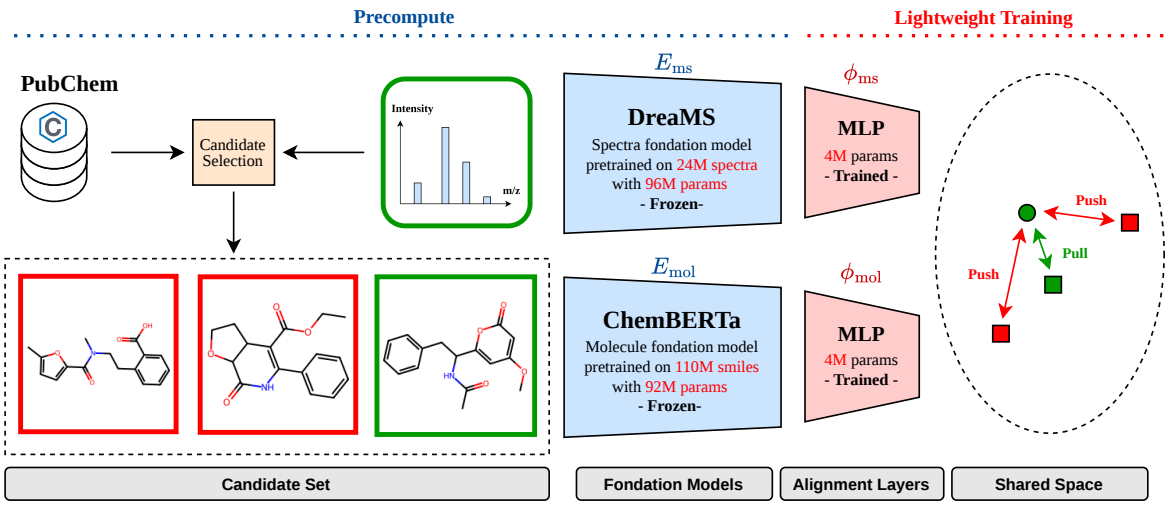

Figure 1. MSAlign framework.

$f_\theta : \mathcal{S} \to \mathcal{M}$ that predicts the molecular structure associated with a given spectrum. Direct regression in $\mathcal{M}$ is challenging due to the discrete and combinatorial nature of molecular structures (Brouard et al., 2016; Krzakala et al., 2024). An alternative is to formulate the task as conditional generation, but such *de novo* approaches remain difficult and currently achieve limited performance (Litsa et al., 2021; Stravs et al., 2022; Wang et al., 2025; Bohde et al., 2025).

A more common approach is to cast Molecule Identification as a retrieval problem. Formally the output space is restricted via a candidate set $C(s) \subset \mathcal{M}$ and f_θ is defined as

$$f_\theta(s) = \arg \max_{m' \in C(s)} \rho_\theta(s, m') \tag{1}$$

Then the goal is to learn a similarity function $\rho_\theta : \mathcal{S} \times \mathcal{M} \to \mathbb{R}$ that ranks the candidate molecules for a given spectrum. In practice such candidate sets are typically defined via a mass filter, i.e. $C(s) = \{m' \in \mathcal{M} : \text{abs}(|m'| - |s|) < \epsilon\}$ where $|\cdot|$ is the mass of the molecule/spectrum and ϵ is a small threshold and the mass of the molecule is assumed to be known. This is realistic in practice as the determination of the type of fragmented ion species (and thus the mass of the neutral molecule) is a task well handled both manually and automatically (Vaniya & Fiehn, 2022). To learn ρ_θ a natural choice is to minimize a contrastive loss i.e. a loss that encourages the similarity of the true pair (s, m) to be higher than the similarity of negative pairs (s, m'). The most popular choice being the InfoNCE loss:

$$\mathcal{L}_{\text{cand}}(\theta) = -\sum_{i=1}^{B} \log \frac{\exp(\rho_\theta(s_i, m_i))}{\sum_{m' \in C_K(s_i)} \exp(\rho_\theta(s_i, m'))} \tag{2}$$

where $C_K(s_i) \subset C(s_i)$ is a subset of K negative candidates for spectrum s_i (including the true molecule m_i). In practice, however the computation of the loss in Equation (2) is expensive as it requires computing the similarity $K \times B$ times which prevent the use of large batch size B. A common solution is to use in-batch negatives, i.e. to use the other molecules in the batch as negatives

$$\mathcal{L}_{\text{batch}}(\theta) = -\sum_{i=1}^{B} \log \frac{\exp(\rho_\theta(s_i, m_i))}{\sum_{j=1}^{B} \exp(\rho_\theta(s_i, m_j))} \tag{3}$$

This is much more computationally efficient but might make the training less informative as the negative candidates are not necessarily hard to distinguish from the true molecule.

3. MSAlign

The core idea of MSAlign is to leverage pretrained unimodal foundation models for spectra and molecules and to learn a lightweight alignment of their embedding spaces. Since pretrained models are frozen, the embeddings can be precomputed, which considerably reduces the computational cost of training and allows using the more informative candidate-based loss Equation (2). This is directly inspired by a recent line of work in vision-language models where the alignment of pretrained unimodal models has been shown to be competitive with joint multi-modal pretraining while being computationally and data efficient (Vouitsis et al., 2024; Zhang et al., 2025; Roschmann et al., 2026).

Formally, we define the learnable similarity function as

$$\rho_\theta(s, m) = \langle g_\theta(s), h_\theta(m) \rangle \tag{4}$$

and the embedding functions f_θ and g_θ are defined as

$$g_\theta(s) = (\phi^\theta_{ms} \circ E_{ms})(s) \tag{5}$$

$$h_\theta(m) = (\phi^\theta_{mol} \circ E_{mol})(m) \tag{6}$$

Dataset	Metric	Models that don't require formula						If Formula is known		
		FFN	DeepSets	JESTR	Emb-Cos	SAIL	MSAlign	MIST	FLARE	MSAlign (+Filter)
NPLIB1	R@1	8.9	11.1	11.2	*22.5*	18.6	**31.8**	*27.2*	19.6	**34.7**
	R@5	22.0	30.4	33.1	*48.3*	46.4	**60.4**	*52.0*	50.8	**64.7**
	R@20	44.4	56.5	62.0	72.7	*75.8*	**82.9**	73.8	*76.6*	**87.9**
Spectraverse	R@1	9.5	12.7	11.3	*27.1*	14.5	**32.3**	*20.2*	12.3	**39.4**
	R@5	18.9	24.4	32.9	*51.8*	34.3	**59.1**	*40.9*	33.7	**69.0**
	R@20	34.7	44.3	60.1	*72.2*	57.4	**79.6**	*63.8*	62.5	**87.1**
MassSpecGym	R@1	2.54	5.24	11.8	*12.3*	09.5	**16.2**	14.6	*27.2*	**32.3**
	R@5	7.59	12.58	25.3	*26.2*	23.9	**35.6**	34.9	*53.8*	**61.8**
	R@20	20.00	28.21	*49.7*	46.4	45.2	**59.9**	59.1	*80.2*	**85.5**

Table 1. Retrieval performances of MSAlign and the baselines.

where E_{ms} and E_{mol} are the pretrained unimodal models for spectra and molecules and ϕ^{θ}_{ms} and ϕ^{θ}_{mol} are lightweight alignment layers. In this work unimodal models considered are DreaMS (Bushuiev et al., 2025) for spectra and ChemBERTa (Ahmad et al., 2022) for molecules. Both are transformer-based models that have been pretrained on large datasets of spectra and molecules respectively (24 Millions spectra for GeMS-A10 for DreaMS and 110 Millions molecules from PUBCHEM for ChemBERTa). Note that any pretrained unimodal model could be used in principle and we expect the proposed framework to benefit from future larger models, in line with observations in vision–language alignment that larger foundation models are easier to align (Huh et al., 2024).

4. Related Works

Using deep learning for molecular retrieval from mass spectra has been an active area of research in recent years and is rapidly gaining traction notably due to the release of large datasets and standardized benchmarks (Bushuiev et al., 2024; Gupta et al., 2026; Bushuiev et al., 2025). In particular, the introduction of DreaMS (Bushuiev et al., 2025) is fundamental to this work as it provided the first foundational encoder for mass spectra that unlock the proposed approach.

Previous works typically learn the similarity ρ_θ from scratch. MIST (Goldman et al., 2023) achieve this by pretraining with fingerprint prediction, finetuning with contrastive learning and augmenting the data with a forward model. Similarly, JESTR (Kalia et al., 2025) pretrains via in-batch InfoNCE (3) and fine-tunes for 3% of the training steps with the more expensive candidates InfoNCE (2). FLARE builds upon JESTR by considering a stronger spectra encoder leveraging peak annotation and a more interpretable, richer similarity (Chen et al., 2026). Finally, Emb-Cos is the strongest of the four models proposed in (De Waele et al., 2026), similarly to MSAlign it uses (2) to learn a shared space for molecules and mass spectra but does not leverage pretrained unimodal models. Interestingly, all of these works can be described via our simple set of equations (4) to (6) which enable to summarize the different design choices of the models in a unified framework reported in Table 5.

5. Experiments

Experimental Setting We evaluate our approach on three open-source datasets of increasing scale and diversity. **NPLIB1** (Dührkop et al., 2021) is the smallest benchmark (10,633 spectrum–molecule pairs). **MassSpecGym** (Bushuiev et al., 2024) provides 231,104 curated pairs and is emerging as a community standard. **Spectraverse** (Gupta et al., 2026) is the largest collection (488,797 pairs) and the most heterogeneous, with over 50% of spectra corresponding to non-$[M+H]^+$ adducts. For NPLIB1 and Spectraverse, we use an 90/5/5 train/validation/test split while preventing data leakage by ensuring that molecules sharing the same chemical formula do not appear across splits. MassSpecGym, provides a predefined split specifically designed to be challenging, where test molecules are sampled from clusters distinct from those used for training. Following the MassSpecGym protocol (Bushuiev et al., 2024), $K = 256$ candidate molecules are retrieved from PubChem (Kim et al., 2023) (118M compounds) by selecting structures whose molecular mass matches the query spectrum within a 10 ppm tolerance, prioritizing compounds relevant to mass spectrometry.

We compare MSAlign to a range of baselines (Table 5), including recent state-of-the-art methods (MIST (Goldman et al., 2023), JESTR (Kalia et al., 2025), FLARE (Chen et al., 2026), and Emb-Cos (De Waele et al., 2026)) as well as simpler architectures (FFN and DeepSets) from MassSpecGym (Bushuiev et al., 2024). To improve reproducibility, all other baselines are re-implemented within a unified codebase, which will be released with the full paper. The only exception to this is MIST for which we follow the guidelines of (Heirman & Bittremieux, 2024). Additionally,

Table 2. Comparison of different encoders E_{mol} and E_{ms} for MSAlign. Dimensions of the embeddings are indicated in parentheses. Results are reported on Spectraverse.

Encoders		Spectraverse		
Molecules (dim)	**Spectra** (dim)	**R@1**	**R@5**	**R@20**
Fingerprint (4096)	Binned (10500)	27.1	51.8	72.2
ChemBERTa (768)	Binned (10500)	28.4	53.4	73.1
Fingerprint (4096)	DreaMS (1024)	29.8	55.2	75.1
ChemBERTa (768)	DreaMS (1024)	**32.3**	**59.1**	**79.6**

we adapt SAIL (Zhang et al., 2025) from vision–language alignment to the mass spectra domain, using ChemBERTa and DreaMS as unimodal encoders.

5.1. Retrieval Performances

Ground truth formula. Some methods assume access to the molecular formula of the target compound. MIST and FLARE use the formula to annotate spectra with subformula information, substantially enriching the model input. However, in realistic retrieval settings, the formula is unknown and must be predicted, which remains an active research problem (Hong et al., 2025; Dührkop et al., 2019; Xing et al., 2023). To enable fair comparison in this setting, we also evaluate MSAlign with a filtering step that removes candidates with incorrect formulas (MSAlign+Filter). This filtering step reduces the candidate set to $K \approx 100$ on average.

Results analysis. In Table 1 we report retrieval performance in terms of R@1, R@5, and R@20 over candidate sets of size $K = 256$ across datasets and models.

MSAlign consistently outperforms all baselines across datasets and metrics. Emb-Cos is the second-best method, which we attribute to its direct optimisation of a candidate-level InfoNCE objective (2). MSAlign further improves upon this approach by leveraging pretrained foundation models. Finally, MSAlign with formula-based filtering (MSAlign+Filter) outperforms both MSAlign and FLARE. This suggests that restricting the candidate space using the correct formula can provide stronger gains than subformula-based peak annotation alone, and may indicate that the latter has been overestimated in prior work.

5.2. Ablation Studies

Molecule and spectra encoding. A key design choice in MSAlign is the selection of encoders for spectra and molecules. To the best of our knowledge, MSAlign is the first approach to leverage pretrained foundation models for both modalities. We evaluate this choice by comparing MSAlign to variants using non-trainable encoders, namely Morgan fingerprints ($d = 4096$) for molecules and binned representations for spectra (width 0.1), which are commonly used in prior work. Results in Table 2 show that both pretrained encoders contribute to performance, with DreaMS providing the largest gain.

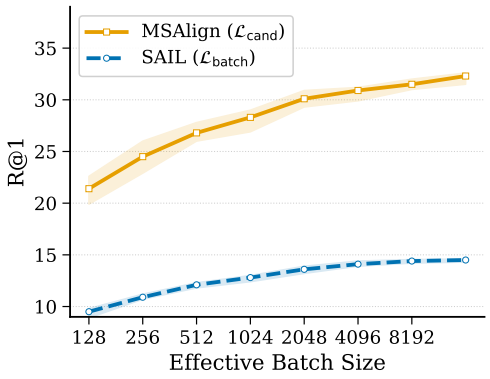

Figure 2. Effect of scaling the effective batch size on Spectraverse performances. For MSAlign the effective batchsize is $B \times K$, we fix $B = 128$ and scale the number of negatives K. For SAIL we directly scale the batchsize B.

These results highlight the importance of pretrained foundation models in achieving strong performance. Moreover, we expect that further improvements may come from scaling such models, which would be consistent with observations in vision–language alignment, where larger unimodal foundation models are easier to align (Huh et al., 2024).

InfoNCE variants. Contrastive approaches typically rely on one of two variants of the InfoNCE objective, which differ in how negative samples are constructed. In the *in-batch* variant, negatives are the other molecules within the same minibatch (Eq. 3). In the *candidate-based* variant, each spectrum is paired with K candidate molecules (Eq. 2), requiring encoding $B \times K$ molecules per batch. While this formulation is more closely aligned with the retrieval setting, it is also K times more computationally expensive. Recent work has shown that in-batch InfoNCE can achieve strong performance in vision–language alignment when the batch size is scaled to large values (Zhang et al., 2025). To fairly compare both formulations, we evaluate them as a function of an *effective batch size*, defined as B for the in-batch and $B \times K$ for the candidate-based variant (Figure 2).

Overall, both methods benefit from increasing *effective batch size*, but the candidate-based formulation remains consistently superior. It goes from 21.4% R@1 at $B \times K = 128$ to 32.3% at $B \times K = 16$k, while the in-batch variant saturates at 14.5% even when scaling to $B = 16$k.

6. Conclusion

We presented MSAlign, a lightweight approach that aligns molecular and mass spectra foundation models in a shared

space. Despite its simplicity, it achieves state-of-the-art performance on metabolite identification benchmarks. More broadly, our findings suggest that aligning pretrained unimodal models is a promising direction for multimodal learning in life sciences.

Acknowledgements

The study was funded by French National Research Agency (ANR) through the following projects: PEPR IA FOUNDRY (ANR-23-PEIA-0003), e-Lucid (ANR-25-TSIA-0002-01) and MetaboHUB (ANR-11-INBS-0010). It also received funding from the European Union's Horizon Europe research and innovation programme under grant agreement 101120237 (ELIAS). The first and second authors respectively received PhD scholarships from Institut Polytechnique de Paris.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., and Ramsundar, B. ChemBERTa-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

Bittremieux, W., Wang, M., and Dorrestein, P. C. The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics*, 18 (12):94, November 2022. ISSN 1573-3890. doi: 10.1007/s11306-022-01947-y.

Bohde, M., Manjrekar, M., Wang, R., Ji, S., and Coley, C. W. DiffMS: diffusion generation of molecules conditioned on mass spectra. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Brouard, C., Shen, H., Dührkop, K., d'Alché Buc, F., Böcker, S., and Rousu, J. Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32 (12):i28–i36, 2016.

Brown, M., Dunn, W. B., Dobson, P., Patel, Y., Winder, C., Francis-McIntyre, S., Begley, P., Carroll, K., Broadhurst, D., Tseng, A., et al. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst*, 134(7):1322–1332, 2009.

Bushuiev, R., Bushuiev, A., de Jonge, N. F., Young, A., Kretschmer, F., Samusevich, R., Heirman, J., Wang, F., Zhang, L., Dührkop, K., et al. MassSpecGym: A benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37: 110010–110027, 2024.

Bushuiev, R., Bushuiev, A., Samusevich, R., Brungs, C., Sivic, J., and Pluskal, T. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nature Biotechnology*, pp. 1–11, 2025.

Chen, Y. Z., Rushing, B., and Hassoun, S. FLARE: Fine-grained learning for alignment of spectra-molecule representation enhances metabolite annotation. *bioRxiv*, pp. 2026–01, 2026.

Cui, H., Tejada-Lapuerta, A., Brbić, M., Saez-Rodriguez, J., Cristea, S., Goodarzi, H., Lotfollahi, M., Theis, F. J., and Wang, B. Towards multimodal foundation models in molecular cell biology. *Nature*, 640(8059):623–633, 2025.

De Waele, G., Wydmuch, M., Waegeman, W., et al. Small molecule retrieval from tandem mass spectrometry: what are we optimizing for? *arXiv preprint arXiv:2602.16507*, 2026.

Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M., Dorrestein, P. C., Rousu, J., and Böcker, S. SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16(4):299–302, April 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0344-8.

Dührkop, K., Nothias, L.-F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M. A., Petras, D., Gerwick, W. H., Rousu, J., Dorrestein, P. C., et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature biotechnology*, 39(4): 462–471, 2021.

Goldman, S., Xin, J., Provenzano, J., and Coley, C. W. MIST-CF: Chemical formula inference from tandem mass spectra. *Journal of Chemical Information and Modeling*, 64(7):2421–2431, 2023.

Gupta, V., Qiang, H., Chung, H.-H., Herbst, E., and Skinnider, M. A. Comprehensive curation and harmonization of small-molecule MS/MS libraries in Spectraverse. *Analytical Chemistry*, 98(5):3934–3943, 2026.

Heirman, J. and Bittremieux, W. Reusability report: annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 6(11):1296–1302, 2024.

Hong, Y., Li, S., Ye, Y., and Tang, H. FIDDLE: a deep learning method for chemical formulas prediction from tandem mass spectra. *Nature Communications*, 16(1): 11102, 2025.

Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The platonic representation hypothesis. In *ICML*, pp. 20617–20642, 2024.

Kalia, A., Zhou Chen, Y., Krishnan, D., and Hassoun, S. JESTR: Joint embedding space technique for ranking candidate molecules for the annotation of untargeted metabolomics data. *Bioinformatics*, 41(7):btaf354, 2025.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. PubChem 2023 update. *Nucleic acids research*, 51(D1): D1373–D1380, 2023.

Krzakala, P., Yang, J., Flamary, R., d'Alché Buc, F., Laclau, C., and Labeau, M. Any2graph: Deep end-to-end supervised graph prediction with an optimal transport loss. *Advances in Neural Information Processing Systems*, 37: 101552–101588, 2024.

Landrum, G. et al. Rdkit documentation. *Release*, 1(1-79): 4, 2013.

Litsa, E., Chenthamarakshan, V., Das, P., and Kavraki, L. Spec2Mol: An end-to-end deep learning framework for translating ms/ms spectra to de-novo molecules. *ChemRxiv*, 2021.

Roschmann, S., Krzakala, P., Mazelet, S., Bouniot, Q., and Akata, Z. SOTAlign: Semi-supervised alignment of unimodal vision and language models via optimal transport. *arXiv preprint arXiv:2602.23353*, 2026.

Stravs, M. A., Dührkop, K., Böcker, S., and Zamboni, N. MSNovelist: de novo structure generation from mass spectra. *Nature Methods*, 19(7):865–870, 2022.

Vaniya, A. and Fiehn, O. Revisiting CASMI: Compound ID for 500 new unknowns, using LC-MS/MS data, 2022.

Vouitsis, N., Liu, Z., Gorti, S. K., Villecroze, V., Cresswell, J. C., Yu, G., Loaiza-Ganem, G., and Volkovs, M. Data-efficient multimodal fusion on a single GPU. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27239–27251, 2024.

Wang, Y., Chen, X., Liu, L., and Hassoun, S. MADGEN: Mass-spec attends to de novo molecular generation. *arXiv preprint arXiv:2501.01950*, 2025.

Xing, S., Shen, S., Xu, B., Li, X., and Huan, T. BUDDY: Molecular formula discovery via bottom-up MS/MS interrogation. *Nature Methods*, 20(6):881–890, June 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01850-x.

Zhang, L., Yang, Q., and Agrawal, A. Assessing and learning alignment of unimodal vision and language models. In *CVPR*, pp. 14604–14614, 2025.

A. Implementation details

All SMILES strings are canonicalized and sanitized using RDKit (Landrum et al., 2013). Molecular formulas and weights are computed by explicitly accounting for implicit hydrogen atoms. All baselines are trained using the hyperparameters reported in their respective original works. MSAlign is trained on a single NVIDIA V100 GPU using the AdamW optimizer with linear warmup followed by cosine decay. Additional hyperparameters are provided in Table 3.

Table 3. MSAlign Hyperparameters.

Parameter	**MassSpecGym**	**Spectraverse**	**NPLIB**
MLP			
# hidden layers	2	2	1
# hidden dim	2048	2048	2048
# shared space dim	1024	1024	1024
dropout	0.2	0.2	0.2
layer norm	Yes	Yes	Yes
Optimization			
learning rate	1e−4	1e−4	1e−4
# max steps	24000	24000	16000
# warmup steps	4000	4000	4000
Loss			
# batch size	128	128	128
# candidates per MS	128	128	128

B. Additional information.

Table 4. Dataset statistics.

	NPLIB	**MassSpecGym**	**Spectraverse**
# Mol/MS pairs	10,633	231,104	488,797
# Unique molecules	8,488	28,936	44,307
Non [M+H]+	25%	15%	52%

Table 5. Comparison of baselines methods. The SAIL methodology was adapted from (Zhang et al., 2025) to the mass spectra domain.

	Molecule encoder		**Spectra encoder**		**Training objective**		**Unsup. data**		**Require Formula**
Method	**Fixed** (E_{mol})	**Trained** (ϕ_{mol})	**Fixed** (E_{ms})	**Trained** (ϕ_{ms})	**Pretrain**	**Finetune**	**Spectra**	**Mol.**	
FFN	Fingerprint	—	Binarize Spectra	—	Cosine Sim.	—	×	×	No
DeepSets	Fingerprint	—	Fourier Features	DeepSets	Cosine Sim.	—	×	×	No
MIST	Fingerprint	MLP	Peaks Annotation	Transformer	Cosine Sim.	InfoNCE (cand)	Forward Model	Negatives	Yes
Emb-Cos	Fingerprint	MLP	Binarize Spectra	MLP	InfoNCE (cand)	—	×	Negatives	No
JESTR	Graph	GNN	Binarize Spectra	MLP	InfoNCE (batch)	InfoNCE (cand)	×	Negatives	No
FLARE	Graph	GNN	Peaks Annotation	Transformer	InfoNCE (batch)	InfoNCE (cand)	×	Negatives	Yes
SAIL	ChemBERTa	MLP	DreaMS	MLP	InfoNCE (batch)	—	ChemBERTa	DreaMS	No
MSAlign	ChemBERTa	MLP	DreaMS	MLP	InfoNCE (cand)	—	ChemBERTa	DreaMS	No