MS-BART: Unified Modeling of Mass Spectra and Molecules for Structure Elucidation

Yang Han[1,2], Pengyu Wang[1,2], Kai Yu[1,2], Xin Chen[1], Lu Chen[1,2]

[1] X-LANCE Lab, Shanghai Jiao Tong University, Shanghai, China
[2] Suzhou Laboratory, Suzhou, China

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

苏州实验室

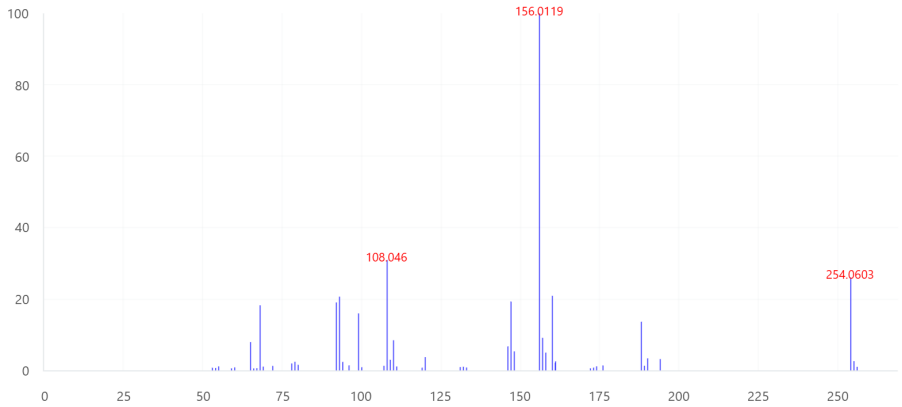
Outline

- Background
- Methodology
- Experimental Results
- Future Work

Background

LC-MS Analysis

Structure Elucidation



- Traditional method require expert knowledge and large effort.
- Language models show great potential in this task but still suffer bad performance.

Background

Based on the molecular formula $C_{10}H_{11}N_3O_3S$ and the provided mass spectrum, determine the most probable chemical structure in SMILES. Justify your reasoning by explaining key fragments in the mass spectrum.

GPT5 Thinking

To determine the most probable chemical structure of xxxx

Step 1: Analyze the Molecular Formula. xxxx

Step 2: Review the Mass Spectrum. xxxx

Step 3: Identify Key Fragments. xxxx

Step 4: Propose a Likely Structure. Xxxx

Step 5: Proposed SMILES Structure.

SMILES: C1=CC2=C(C=C1)N=C(S2)C(CN)O ❌

- Scarcity of annotated experiment spectra.
- Complexity and heterogeneity of raw spectra.

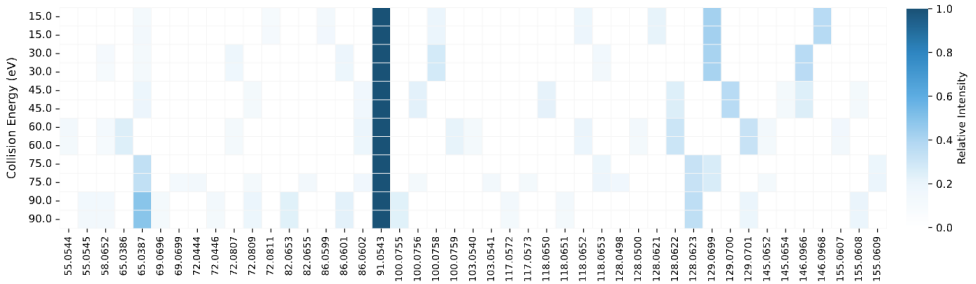


Figure 1: Randomly selected mass spectra of a molecule (SMILES: C#CCNCC1=CC=CC=C1, InChIKey: LDYBFSGEBHSTOQ) from MassSpecGym [7], acquired under varying collision energies. The x-axis shows the mass-to-charge ratio (m/z), the y-axis indicates collision energy (in eV), and color represents normalized relative intensity.

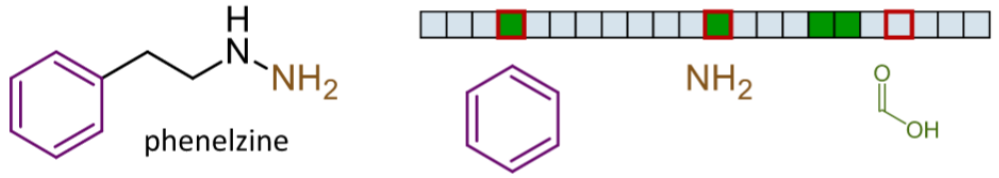
Outline

- Background
- **Methodology**
- Experimental Results
- Future Work

Methodology

The raw spectra is too difficult to model, is there any equivalent representation which is more appropriate for language modeling?

Morgan Fingerprint



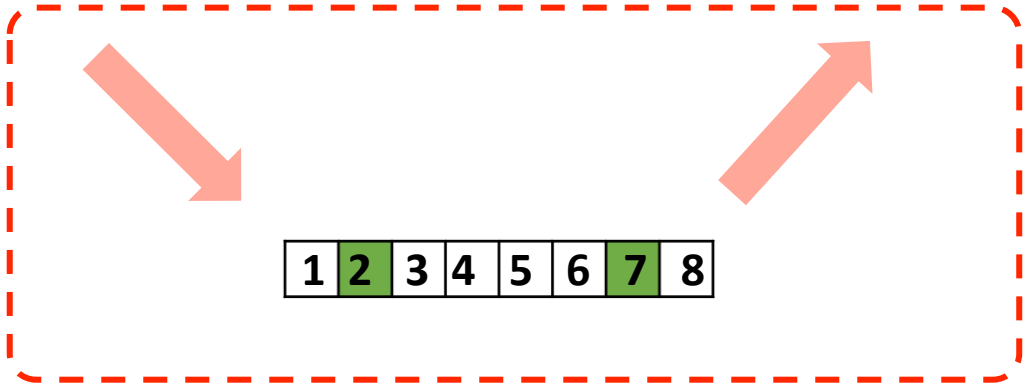
- The forward algorithm (structure -> fingerprint) is accurate calculated by RDKit and produce unlimited pair data theoretically, which alleviate the scarcity of data.
- Mogan Fingerprint is a binary vector can be easy to tokenize.

Methodology

We transfer the original problem (spectra -> structure) into an easier problem. We first transfer the spectra into fingerprint by a pretrained model (not accurate) and then predict the structure from the fingerprint.

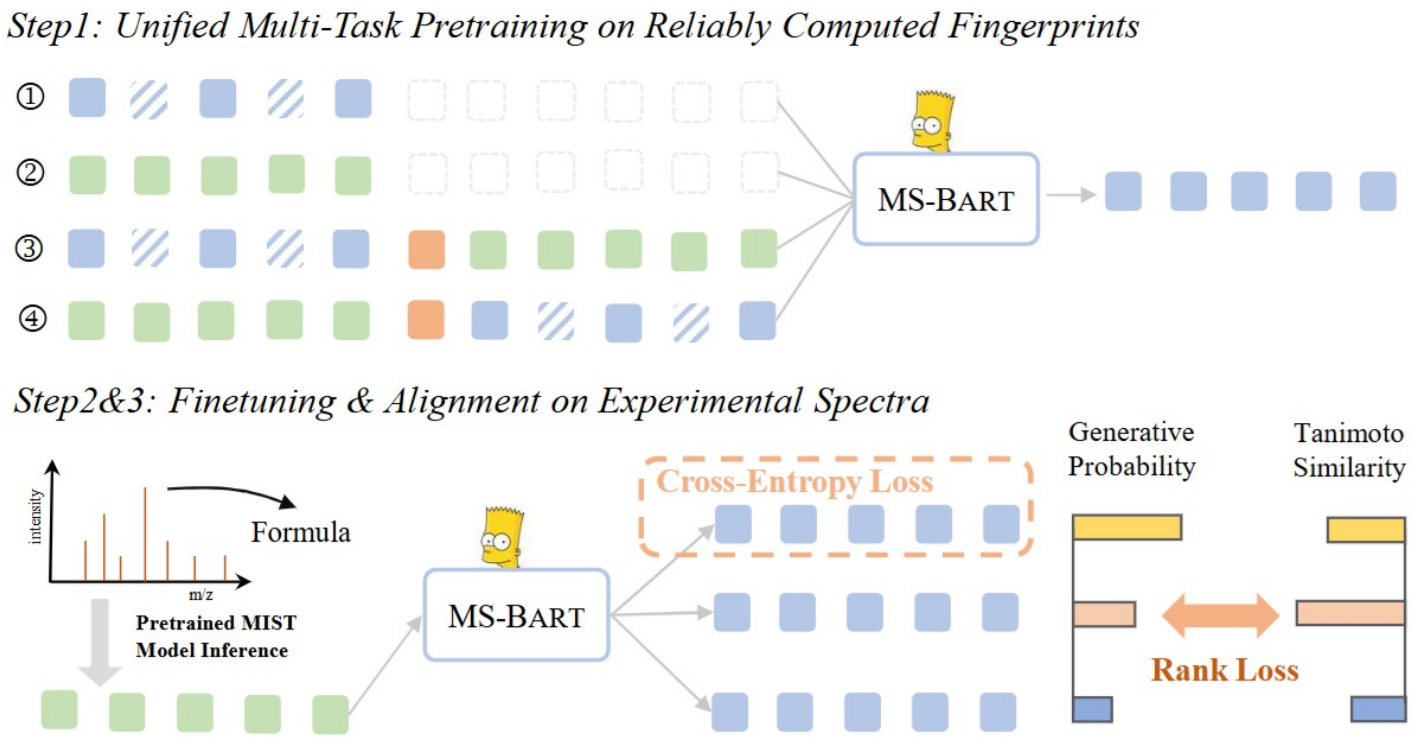
Mass Spectra
Structure Elucidation

Our Method (MS-BART)



Methodology

Three stages (pretraining–finetuning–alignment) end-to-end framework for structure elucidation.



- ✓ *Stage1: Fundamental understanding of molecular structure and fingerprints.*
- ✓ *Stage2: Alleviate the computational and real-world data distribution gap.*
- ✓ *Stage3: Align the model's probabilistic rankings of generated molecules with their Tanimoto similarity to the true structure.*

Outline

- Background
- Methodology
- **Experimental Results**
- Future Work

Experimental Results

Quantitative comparison on two public benchmarks show MS-BART achieves SOTA performance across 5/12 key metrics.

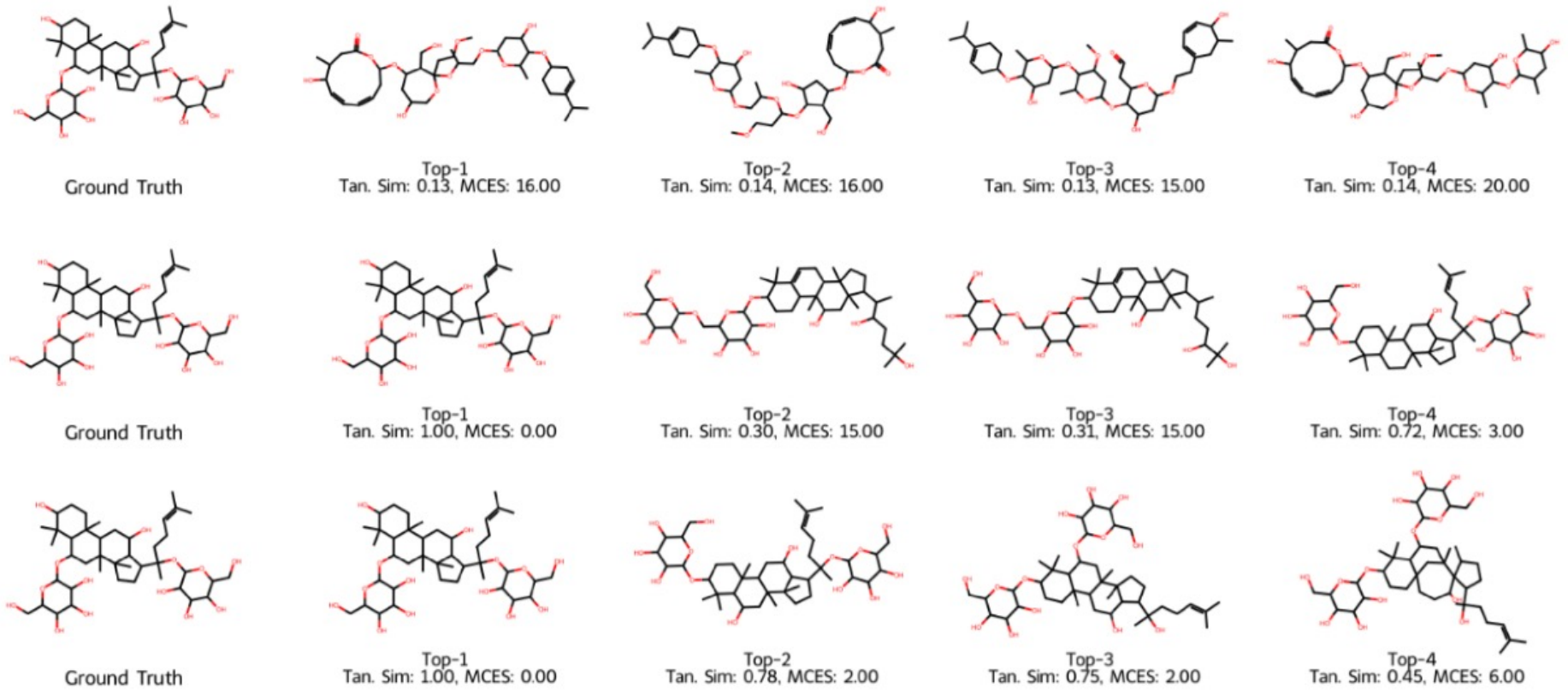
Table 1: Performance comparison of MS-BART and baseline methods on the NPLIB1 [10] and MassSpecGym [7]. Results marked with * are reproduced from MassSpecGym and DIFFMS. **Bold** denotes the best performance, underlined indicates the second-best.

Model	Top-1			Top-10		
	Accuracy ↑	MCES ↓	Tanimoto ↑	Accuracy ↑	MCES ↓	Tanimoto ↑
NPLIB1						
SPEC2MOL*	0.00%	27.82	0.12	0.00%	23.13	0.16
MIST + NEURALDECIPHER*	2.32%	12.11	0.35	6.11%	9.91	0.43
MIST + MSNOVELIST*	5.40%	14.52	0.34	11.04%	10.23	0.44
MADGEN	2.10%	20.56	0.22	2.39%	12.69	0.27
DIFFMS	**8.34%**	11.95	0.35	**15.44%**	9.23	0.47
MS-BART	7.45%	**9.66**	**0.44**	10.99%	**8.31**	**0.51**
MS-BART(Gold Fingerprint)	73.50%	2.14	0.90	79.12%	1.60	0.94
MASSSPECGYM						
SMILES TRANSFORMER*	0.00%	79.39	0.03	0.00%	52.13	0.10
SELFIES TRANSFORMER*	0.00%	38.88	0.08	0.00%	26.87	0.13
RANDOM GENERATION*	0.00%	21.11	0.08	0.00%	18.26	0.11
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
MIST + MSNOVELIST*	0.00%	45.55	0.06	0.00%	30.13	0.15
MADGEN	1.31%	27.47	0.20	1.54%	16.84	0.26
DIFFMS	**2.30%**	18.45	**0.28**	**4.25%**	**14.73**	**0.39**
MS-BART	1.07%	**16.47**	0.23	1.11%	15.12	0.28
MS-BART(Gold Fingerprint)	47.56%	3.26	0.85	64.62%	2.02	0.93

MS-BART can almost find the exact match or extremely similar candidates if the fingerprint predicted from spectra is accurate and indicating that further work can be devoted to improving the performance of the fingerprint prediction model.

Experimental Results

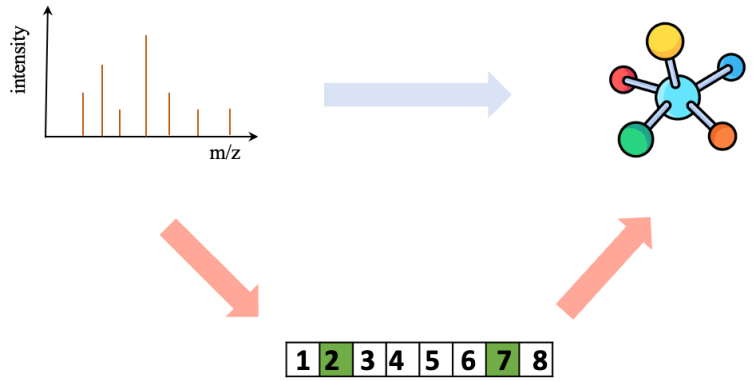
Qualitative comparison between three stages to show the benefit of each stage.



Outline

- Background
- Methodology
- Experimental Results
- **Future Work**

Future Work



- Future work can be devoted to improve the fingerprint prediction model.
- Base on fingerprint vocabulary, many NLP algorithm can be applied to improve the final performance.

Paper

Code

Weight

Demo

Thank you !