MassSpecGym: A benchmark for the discovery and identification of molecules

Roman Bushuiev[1,2], Anton Bushuiev[2], Niek F. de Jonge[3], Adamo Young[4], Fleming Kretschmer[5], Raman Samusevich[1,2], Janne Heirman[6], Fei Wang[7,8], Luke Zhang[4], Kai Dührkop[5], Marcus Ludwig[9], Nils A. Haupt[5], Apurva Kalia[10], Corinna Brungs[1], Robin Schmid[1], Russell Greiner[7,8], Bo Wang[4], David S. Wishart[7], Li-Ping Liu[10], Juho Rousu[11], Wout Bittremieux[6], Hannes Rost[4], Tytus D. Mak[12], Soha Hassoun[10], Florian Huber[15], Justin J.J. van der Hooft[3,14], Michael A. Stravs[15], Sebastian Böcker[5], Josef Sivic[2], Tomáš Pluskal[1]

roman.bushuiev@uochb.cas.cz

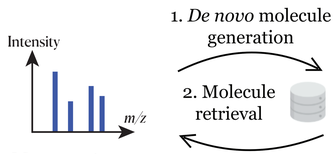
1. IOCB Prague 2. Czech Institute of Informatics, Robotics and Cybernetics CTU in Prague 3. Wageningen University & Research 4. University of Toronto 5. Friedrich-Schiller-Universität Jena 6. Universiteit Antwerpen 7. University of Alberta 8. amii 9. Giant 10. Tufts University 11. Aalto University 12. NIST 13. Hochschule Düsseldorf University of Applied Sciences 14. University of Johannesburg 15. eawag aquatic research

Neural Information Processing Systems

arXiv · GitHub · Hugging Face · Papers With Code · Polaris

The MassSpecGym benchmark

Three new challenges



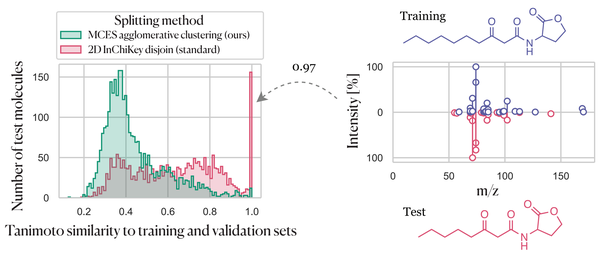
$V \in \mathbb{V}^N$, $|\mathbb{V}| = 118$ (atom/node types)

$E \in \mathbb{E}^M$, $|\mathbb{E}| = 4$ (bond/edge types)

New largest public dataset of mass spectra labeled with molecules

Dataset	Spectra	High-quality spectra	Molecules	Split
GNPS (Wang et al., 2016)	**322K**	104K	16K	✗
MoNA (Fiehn lab)	98K	62K	10K	✗
MassBank (Horai et al., 2010)	62K	58K	4K	✗
MIST CANOPUS (Goldman et al., 2023)	11K	≤ 11K	≤ 9K	✓
MassSpecGym (ours)	231K	**231K**	**29K**	✓

New generalization-demanding data split



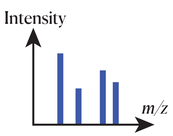
How many molecules have not yet been discovered in a randomly picked organism (e.g., this plant)?

E. coli bacteria < 19% molecules discovered

Plant kingdom < 7% molecules discovered

Human < 4% molecules discovered

Mass spectrometry (LC-MS/MS)

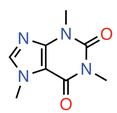


Mass spectra of molecules (MS/MS)

WANTED

New molecular generative models conditioned on mass spectra needed!

Major bottleneck!



Discovery of new molecular structures

Percentage estimates based on Alseekh et al., 2021, Nat Methods, https://doi.org/10.1038/s41592-021-01197-1

MassSpecGym baseline results

1. *De novo* molecule generation

	Top-1			Top-10		
	Accuracy ↑	MCES ↓	Tanimoto ↑	Accuracy ↑	MCES ↓	Tanimoto ↑
Random chemical generation	0.00	**28.59 (28.33-28.84)**	0.07 (0.07 - 0.07)	0.00	25.72 (25.49-25.95)	0.10 (0.10 - 0.10)
SMILES Transformer	0.00	53.80 (52.95-54.61)	0.07 (0.07 - 0.08)	0.00	21.97 (21.79-22.16)	**0.17 (0.17 - 0.17)**
SELFIES Transformer	0.00	33.28 (33.00-33.57)	**0.10 (0.10 - 0.10)**	0.00	**21.84 (21.67-22.00)**	0.15 (0.15 - 0.15)
Bonus chemical formulae challenge						
SMILES Transformer	0.00	79.39 (78.64-80.08)	0.03 (0.03 - 0.04)	0.00	52.13 (51.45-52.81)	0.10 (0.09 - 0.10)
SELFIES Transformer	0.00	38.88 (38.57-39.20)	**0.08 (0.08 - 0.08)**	0.00	26.87 (26.66-27.11)	**0.13 (0.13 - 0.13)**
Random chemical generation	0.00	**21.11 (20.97-21.26)**	**0.08 (0.08 - 0.08)**	0.00	**18.25 (18.14-18.35)**	0.11 (0.11 - 0.11)

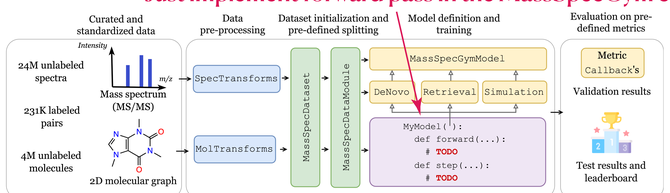
2. Molecule retrieval

The performance of current methods is limited!

	Hit rate @ 1 ↑	Hit rate @ 5 ↑	Hit rate @ 20 ↑	MCES @ 1 ↓
Random	0.37 (0.24-0.54)	2.01 (1.68-2.39)	8.22 (7.53-8.89)	30.81 (30.40-31.21)
DeepSets	1.47 (1.18-1.77)	6.21 (5.64-6.82)	19.23 (18.24-20.26)	25.11 (24.84-25.39)
Fingerprint FFN	2.54 (2.17-2.99)	7.59 (6.96-8.28)	20.00 (19.01-20.98)	24.66 (24.38-24.94)
DeepSets + Fourier features	5.24 (4.71-5.83)	12.58 (11.80-13.42)	28.21 (27.10-29.36)	22.13 (21.85-22.43)
MIST	**14.64 (13.82-15.54)**	**34.87 (33.69-36.10)**	**59.15 (57.89-60.39)**	**15.37 (15.12-15.62)**
Bonus chemical formulae challenge				
Random	3.06 (2.64-3.52)	11.35 (10.60-12.12)	27.74 (26.52-28.84)	13.87 (13.70-14.03)
DeepSets	4.42 (3.92-4.97)	14.46 (13.58-15.36)	30.76 (29.67-31.93)	15.04 (14.89-15.19)
Fingerprint FFN	5.09 (4.57-5.66)	14.69 (13.83-15.56)	31.97 (30.86-33.10)	14.94 (14.79-15.09)
DeepSets + Fourier features	6.56 (5.95-7.16)	16.46 (15.58-17.35)	33.46 (32.39-34.59)	14.14 (13.98-14.31)
MIST	**9.57 (8.88-10.30)**	**22.11 (21.10-23.13)**	**41.12 (39.98-42.34)**	**12.75 (12.59-12.91)**

3. Spectrum simulation

	Cosine Similarity ↑	Jensen-Shannon Similarity ↑	Hit Rate @ 1 ↑	Hit Rate @ 5 ↑	Hit Rate @ 20 ↑
Precursor m/z	0.15 (0.14-0.17)	0.41 (0.40-0.42)	0.38 (0.21-0.62)	1.72 (1.32-2.18)	7.17 (6.32-8.04)
FFN Fingerprint	0.25 (0.24-0.26)	0.48 (0.47-0.48)	8.44 (7.56-9.34)	21.43 (20.10-22.79)	38.57 (36.99-40.23)
GNN	0.19 (0.18-0.20)	0.44 (0.44-0.45)	3.95 (3.37-4.62)	11.92 (10.87-13.00)	26.27 (24.83-27.82)
FraGNNet	**0.52 (0.51-0.53)**	**0.63 (0.63-0.64)**	**46.64 (44.98-48.26)**	**72.56 (71.18-74.00)**	**83.58 (82.34-84.75)**
Bonus chemical formulae challenge					
Precursor m/z	-	-	2.09 (1.66-2.59)	8.52 (7.65-9.53)	22.65 (21.26-24.01)
FFN Fingerprint	-	-	7.62 (6.77-8.54)	22.70 (21.32-24.12)	44.12 (42.51-45.75)
GNN	-	-	3.63 (3.05-4.29)	13.55 (12.46-14.68)	33.77 (32.26-35.37)
FraGNNet	-	-	**31.93 (30.40-33.50)**	**63.20 (61.64-64.76)**	**82.70 (81.45-83.93)**

Evaluating new methods is easy

Just implement forward pass in the MassSpecGym codebase!