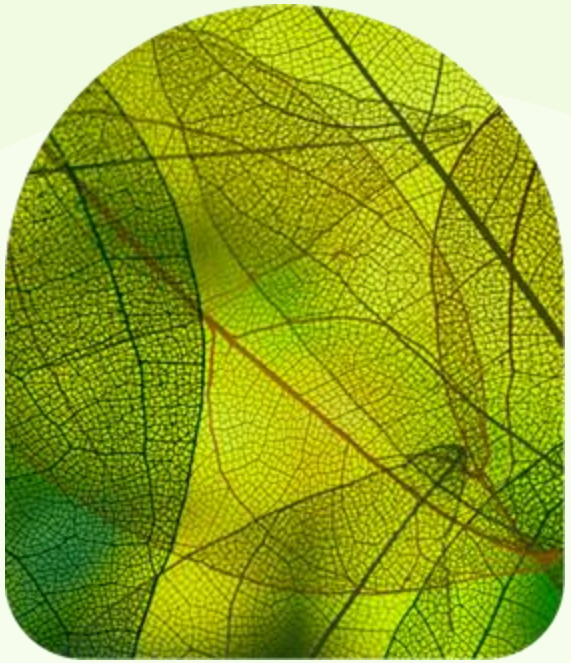
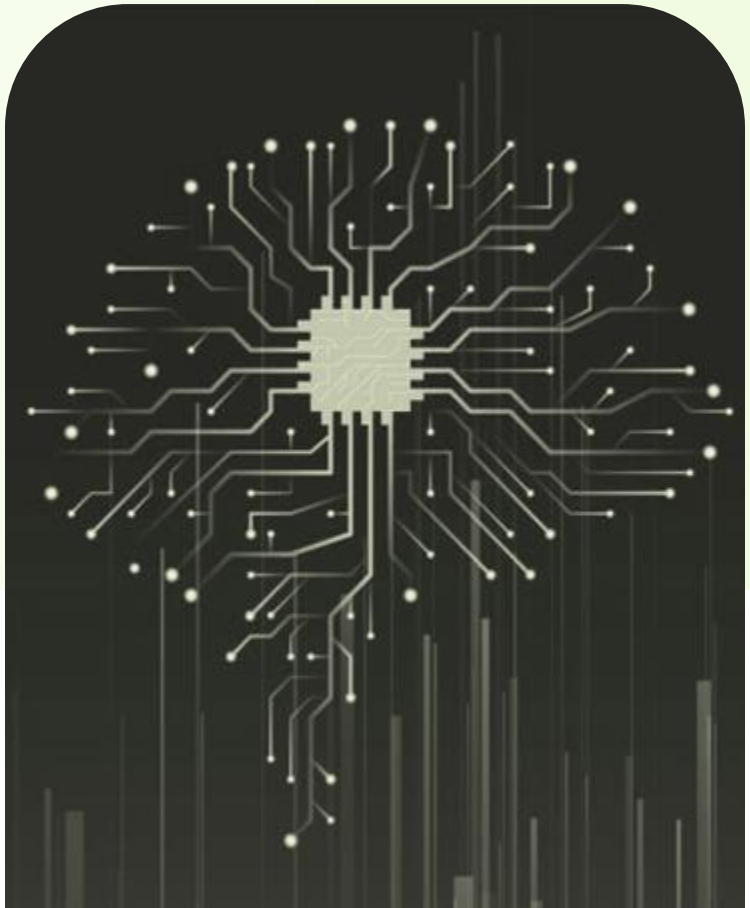


Pangea bio



SpectruMS: A Multi-modal Foundation Model for Better Generalizability on Tandem MS2 Data

Aya Abdelbaky

ICCS 2025

Outline

- Motivation
- SpectruMS base model
- Specialized models
- Results
- Learnings and Outlook

Our Workflow | PangeAI™ enables us to explore the natural world and nominate potential health products for development, at scale

PangeAI™ Discovery Workflow

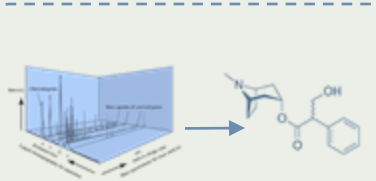
SELECT & SOURCE SPECIES

Select & source species based on therapeutic relevance, likelihood of novelty, and access to plant material

GENERATE LC-MS/MS[1] DATA

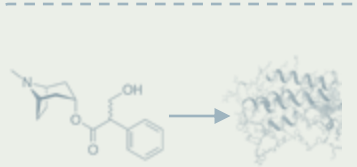
Source biomaterial and generate MS data via growing global partnership ecosystem

UNCOVER CHEMISTRY



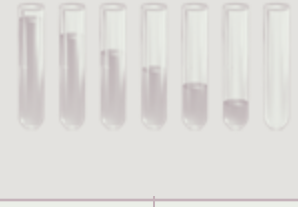
Dereplicate known and predict potential novel structures found within a species via our proprietary computational MS platform

PREDICT BIOACTIVITY



Select potential hits based on predicted activity, SAR[2], and physicochemical properties

VALIDATE EXPERIMENTALLY



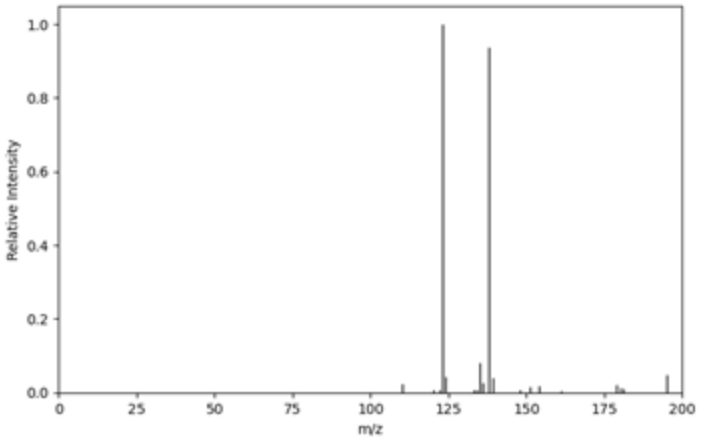
Experimentally validate activity of extract and/or compound and elucidate active chemical structure(s)

Scalable product pipeline

New bioactive compounds and extracts for further development

[1]LC-MS/MS = Liquid chromatography mass spectrometry; [2]SAR = Structure activity relationship

LC-MS/MS[1] Problem | Most found metabolites cannot be annotated



MASS SPECTRUM

Spectral reference matching

Structure retrieval

De novo generation

Spectral reference libraries (10^4)[3] (labeled data)

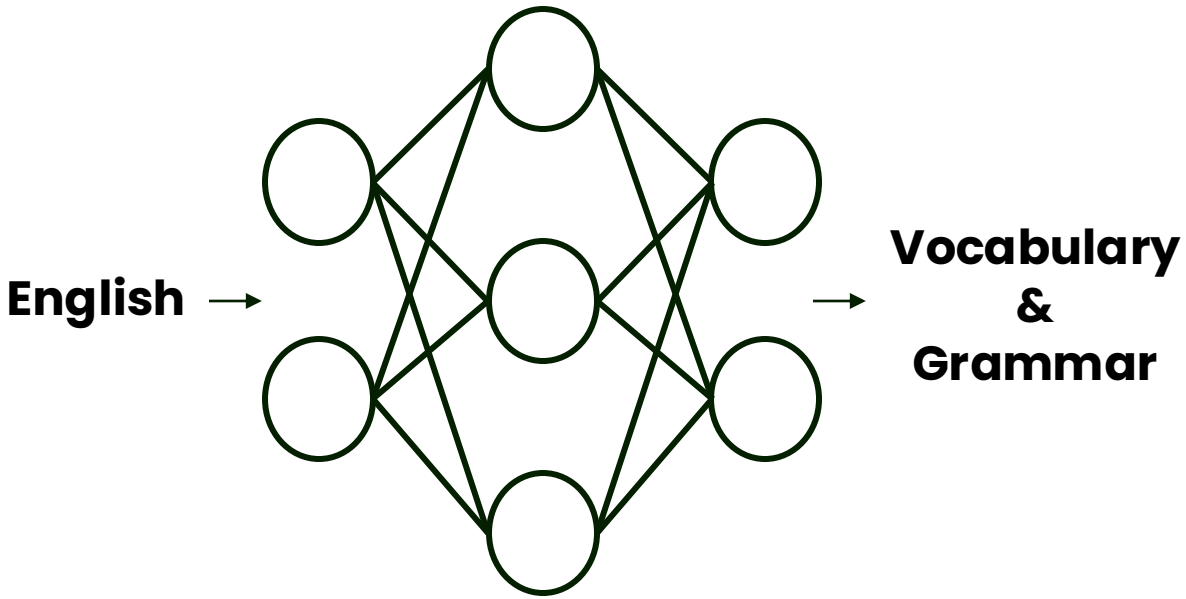
Known NPs (10^5)[4]

All NPs[2] (10^8 – 10^{10})[5]

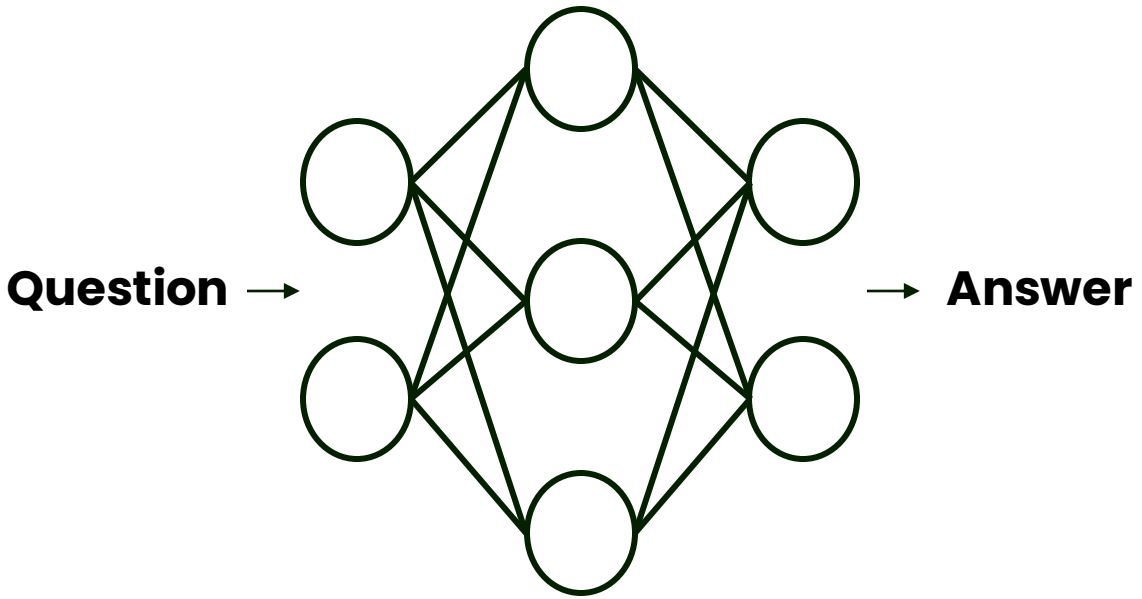
[1]LC-MS/MS = Liquid chromatography mass spectrometry
[2]NPs = Natural products
[3]NIST2023, MoNA, MassBank, GNPS
[4]Journal of Cheminformatics, 13(1):1–13, 2021
[5]Plant and Cell Physiology, 53(2):e1–e1, 2012

LLMs[1] Approach | Not limited by spectral reference libraries

Step 1
Language Learning
Pre-training with unlabeled data

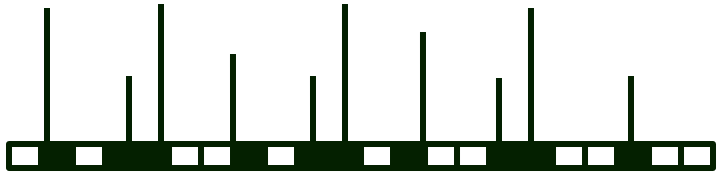


Step 2
Task Specific
Fine-tuning with labeled data

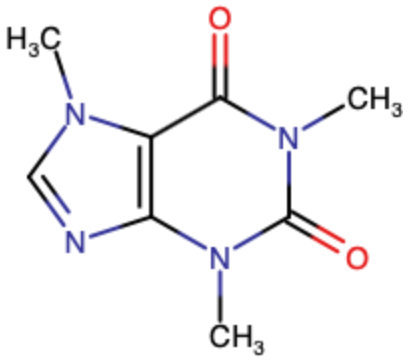


[1]LLMs = Large language models

Multi-modal SpectruMS Pre-training | Learning spectra and chemistry languages



+



→ SpectruMS →

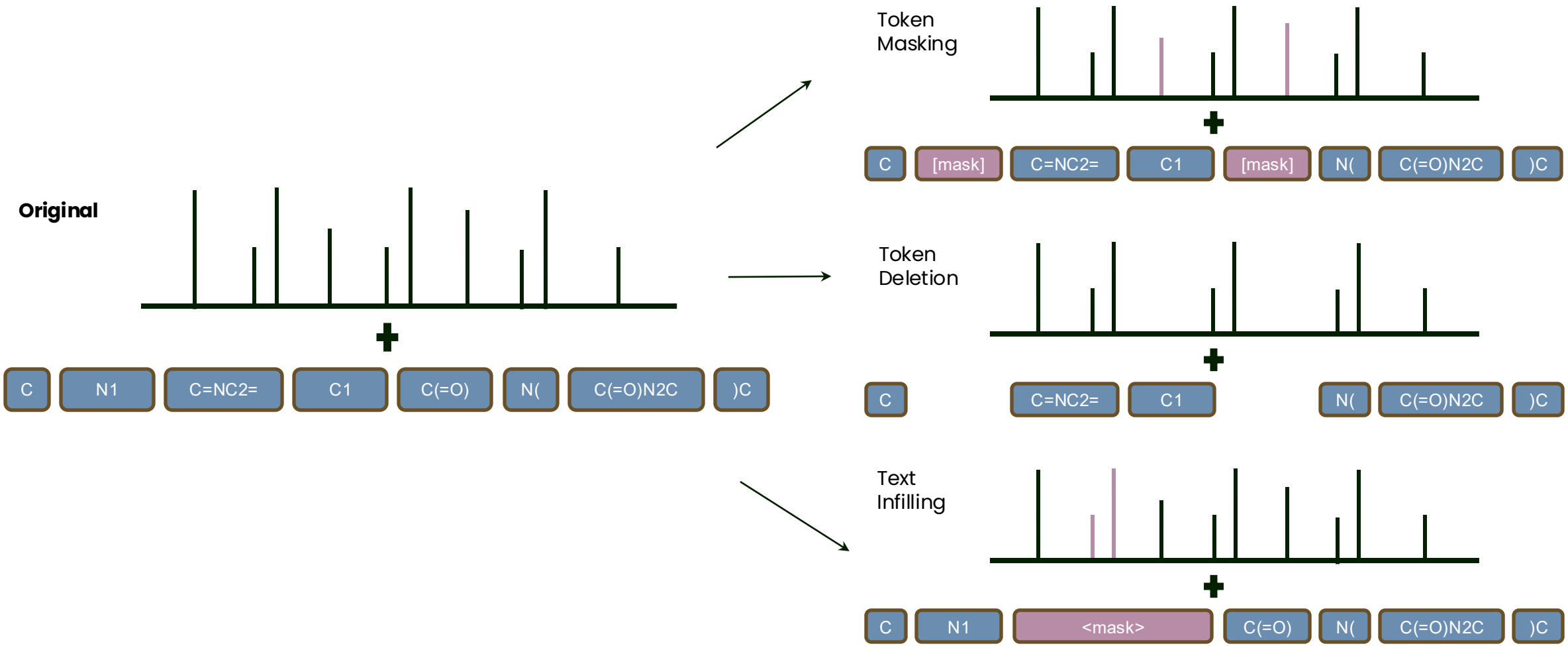
“ … 123.05 138.07 135.07 …”

+

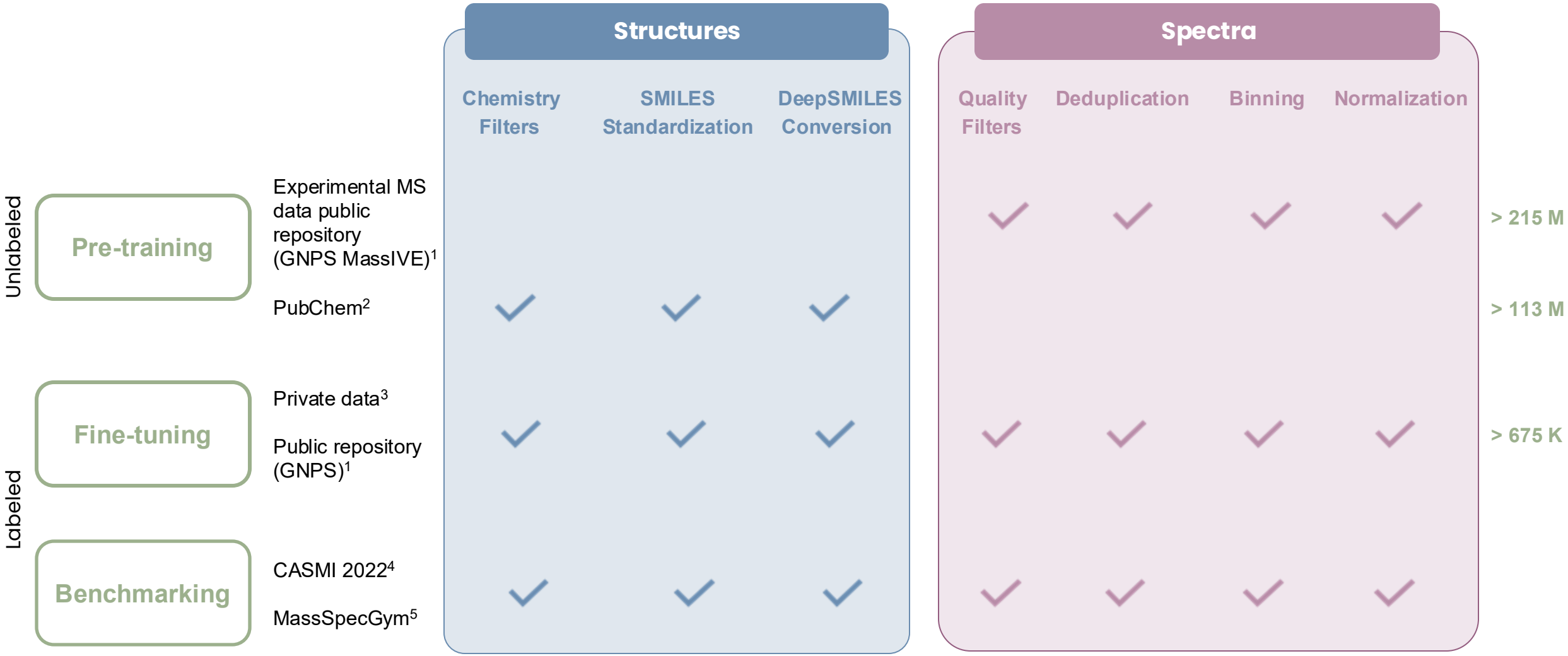
“CN1C=NC2=C1C(=O)N(C(=O)N2C)C”

Learnt Context = Generalizability

Token Masking | Representations are learnt from unlabelled data



Datasets | Pre-processing pipeline for both labeled and unlabeled data

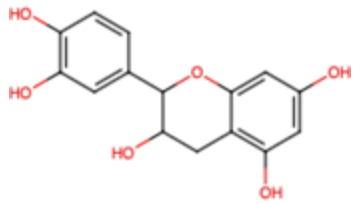


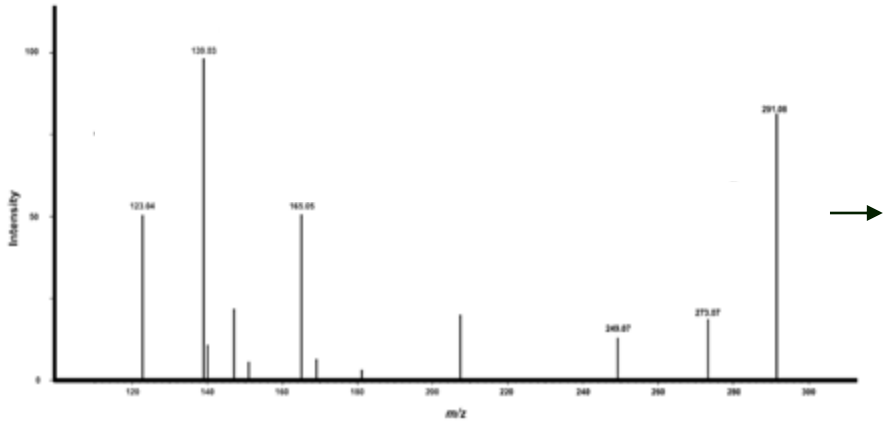
[1]Nature biotechnology 34, no. 8 (2016): 828. PMID: 27504778
[2]Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., & Bolton, E. E. (2023). PubChem in 2023: new data content and improved web interfaces. Nucleic Acids Research, 51 (D1), D1373–D1380. https://doi.org/10.1093/nar/gkac956
[3]Johnson, R. D. (Ed.). (2023). NIST Chemistry WebBook, NIST Standard Reference Database Number 69. National Institute of Standards and Technology, Gaithersburg MD. https://webbook.nist.gov
[4]Fiehn, O. (2022). CASMI 2022: Critical Assessment of Small Molecule Identification. Fiehn Lab, UC Davis.
[5]Bushuiev, R., Bushuiev, A., de Jonge, N., Young, A., Kretschmer, F., Samusevich, R., ... & Pluskal, T. (2024). MassSpecGym: A benchmark for the discovery and identification of molecules. Advances in Neural Information Processing Systems, 37, 110010-110027.

Multi-modal SpectruMS Fine-tuning | Task specialization

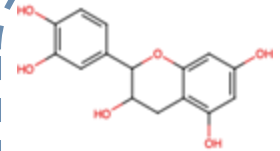
Structure to Spectra

- **In-silico spectra generation**

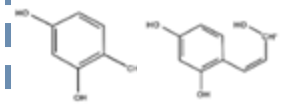




Spectra to Structure



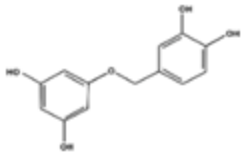
- **De novo structure generation**



- **Molecular feature prediction**

Flavonoids

- **Chemical class prediction**

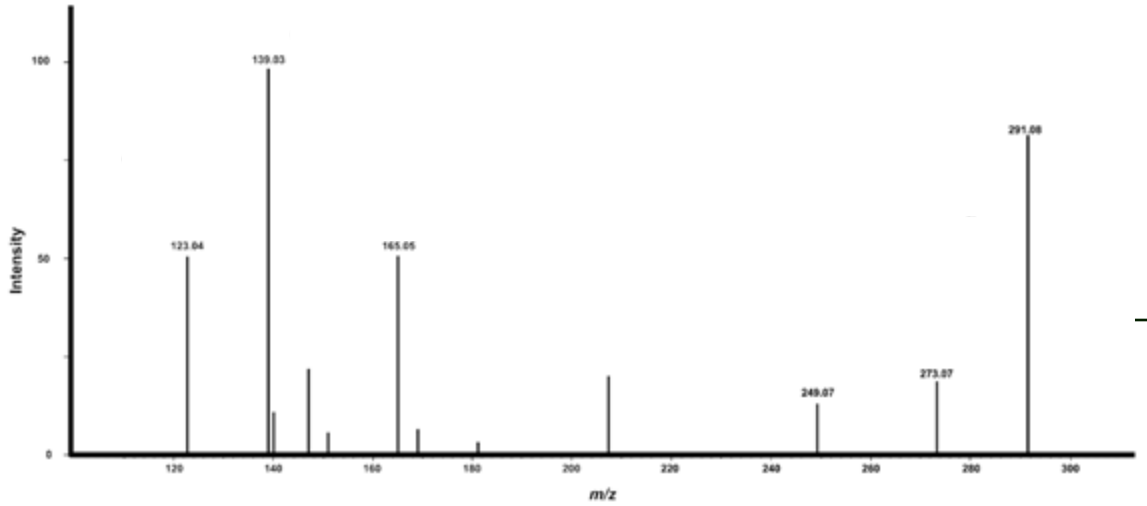


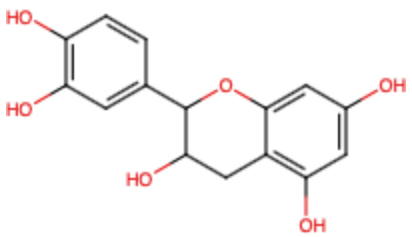
- **Substructure search**

Spectrum figure adapted from Wolf, S., Schmidt, S., Müller-Hannemann, M. et al. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics 11, 148 (2010). https://doi.org/10.1186/1471-2105-11-148

Task 1 | De novo structure generation from spectra

Epicatechin

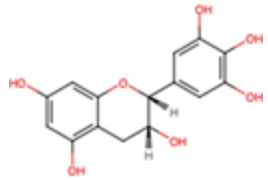




C1C(C(OC2=CC(=CC(=C21)O)O)C3=CC(=C(C=C3)O)O)O

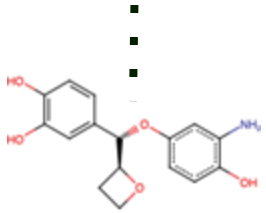
Generative Model

1



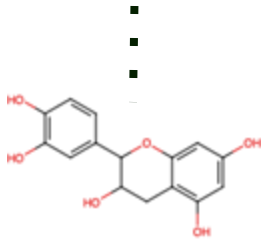
Epigallocatechin (EGC)

5



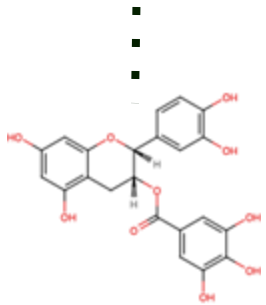
New

10



Epicatechin

25

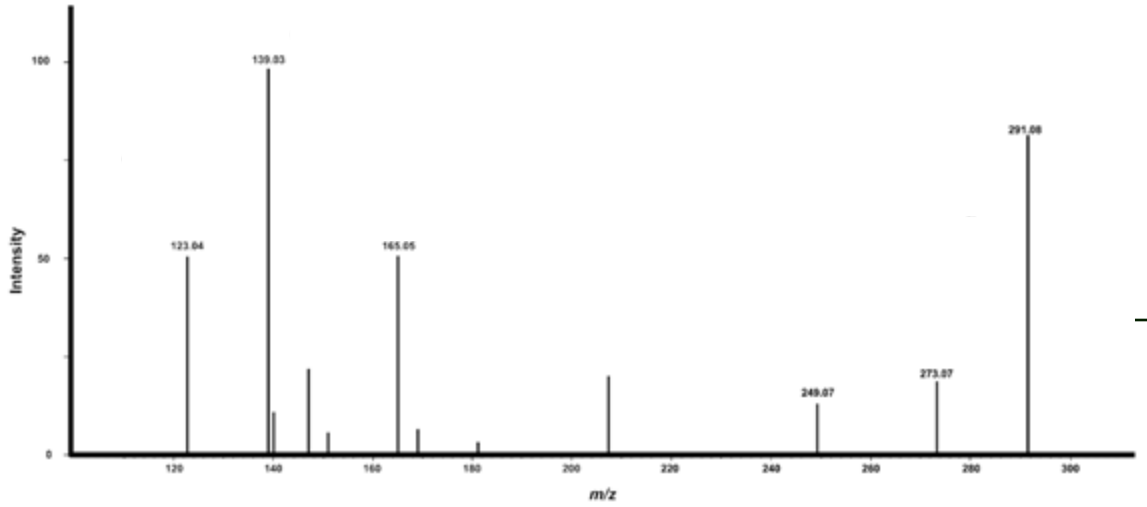


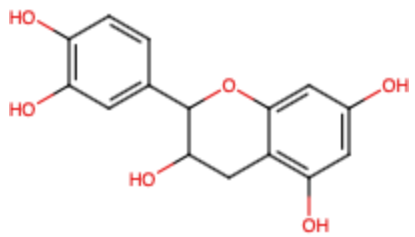
Epicatechin gallate (ECG)

Spectrum figure adapted from Wolf, S., Schmidt, S., Müller-Hannemann, M. et al. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics 11, 148 (2010). https://doi.org/10.1186/1471-2105-11-148

Task 2 | Molecular features prediction from spectra

Epicatechin

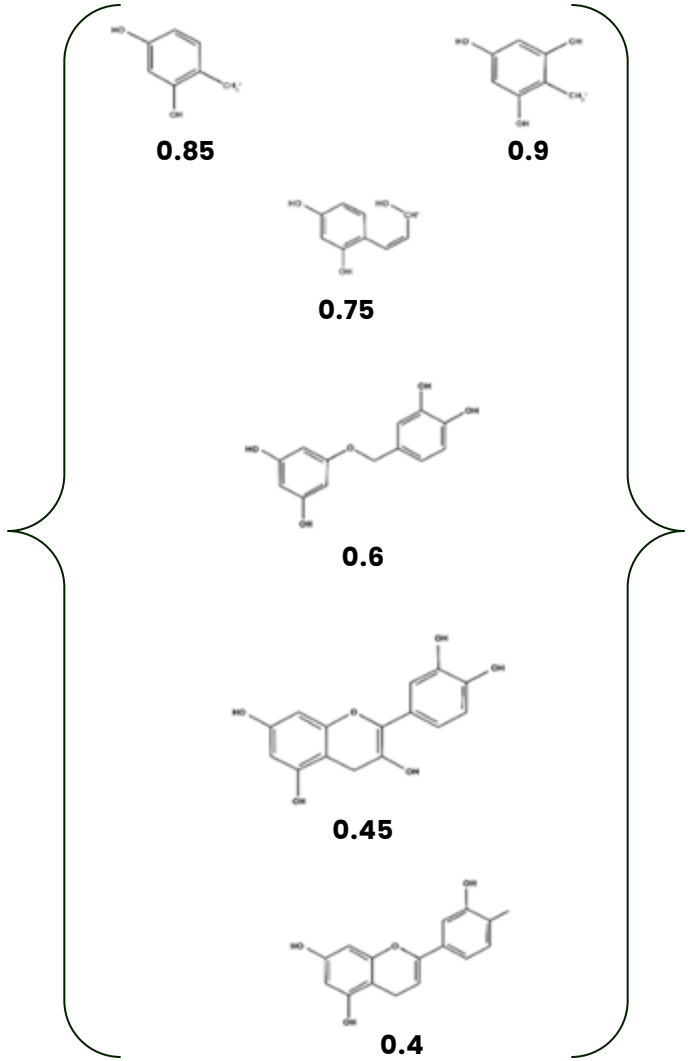




C1C(C(OC2=CC(=CC(=C21)O)O)C3=CC(=C(C=C3)O)O)O

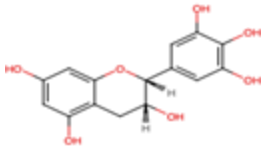
Molecular Feature Model

Molecular Features

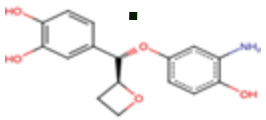


Spectrum figure adapted from Wolf, S., Schmidt, S., Müller-Hannemann, M. et al. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics 11, 148 (2010). https://doi.org/10.1186/1471-2105-11-148

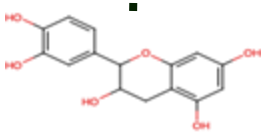
Optimized Ranking | Re-ranking the generated list of candidates

1 

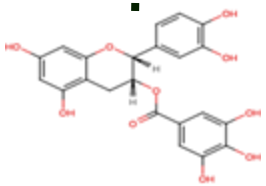
Epigallocatechin (EGC)

5 

New

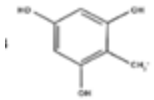
10 

Epicatechin

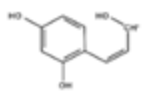
25 

Epicatechin gallate (ECG)

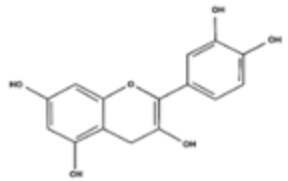
Weighted Ranking



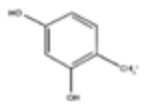
0.9



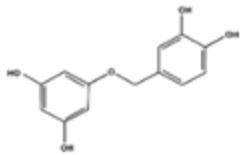
0.75



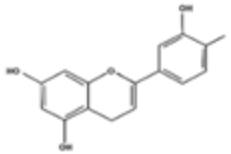
0.45



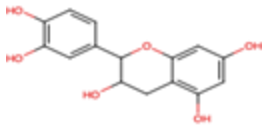
0.85



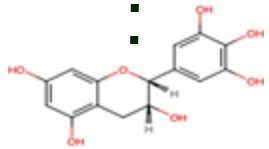
0.6



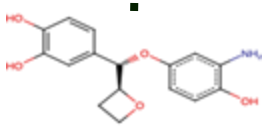
0.4

1 

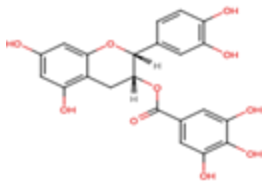
Epicatechin

5 

Epigallocatechin (EGC)

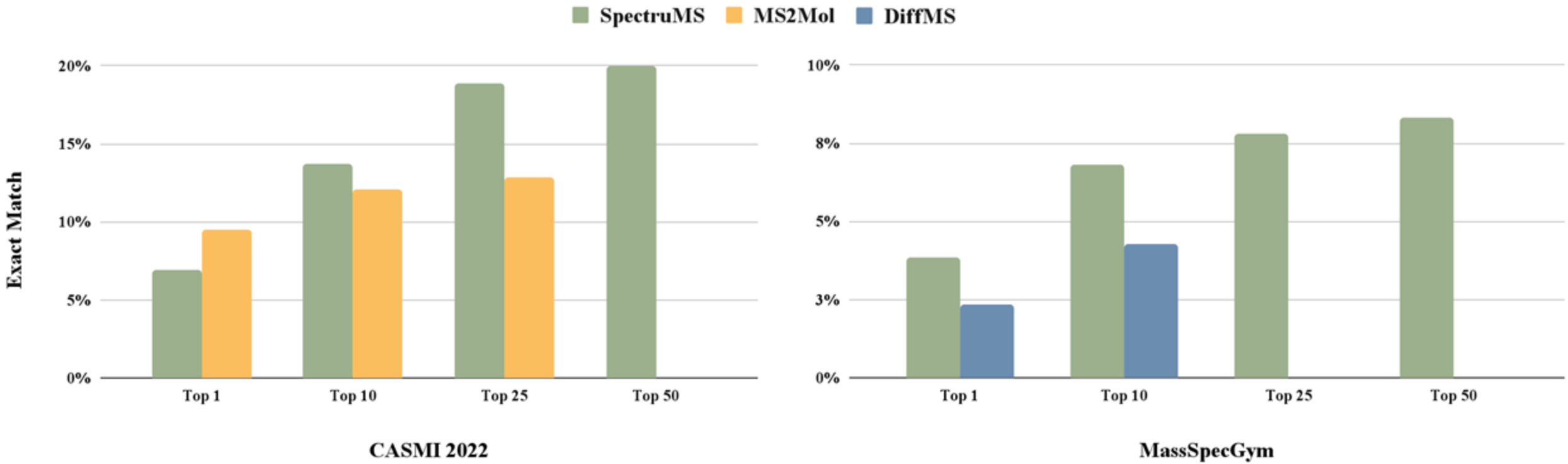
10 

New

25 

Epicatechin gallate (ECG)

Results | SpectruMS outperforms *state-of-the-art* in generating the correct structure in the *top-k* candidates



*Benchmarking is unidentical

[1]Exact Match = Correct prediction having tanimoto coefficient of 1.0 with label

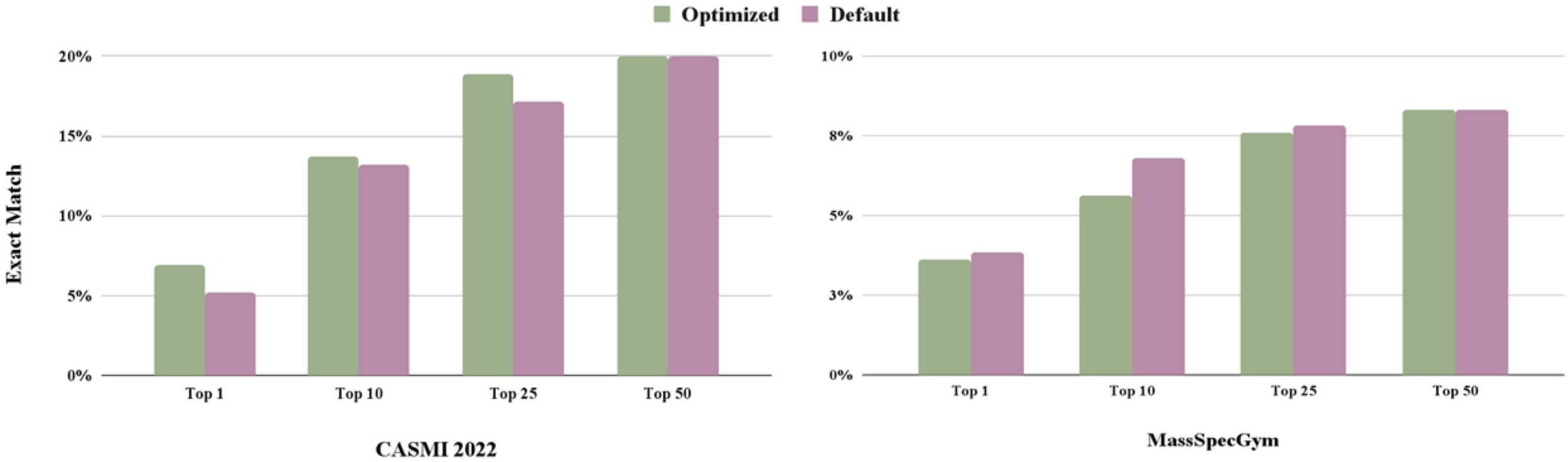
[2] Butler T, Frandsen A, Lightheart R, Bargh B, Taylor J, Bollerman T, et al. MS2Mol: A transformer model for illuminating dark chemical space from mass spectra. ChemRxiv. 2023; doi:10.26434/chemrxiv-2023-vsmpx-v3 This content is a preprint and has not been peer-reviewed.

[3] Fiehn, O. (2022). CASMI 2022: Critical Assessment of Small Molecule Identification. Fiehn Lab, UC Davis.

[4] Bohde, M, Manjrekar, M., Wang, R., Ji, S., & Coley, C. W. (2025). DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra. arXiv preprint arXiv:2502.09571.

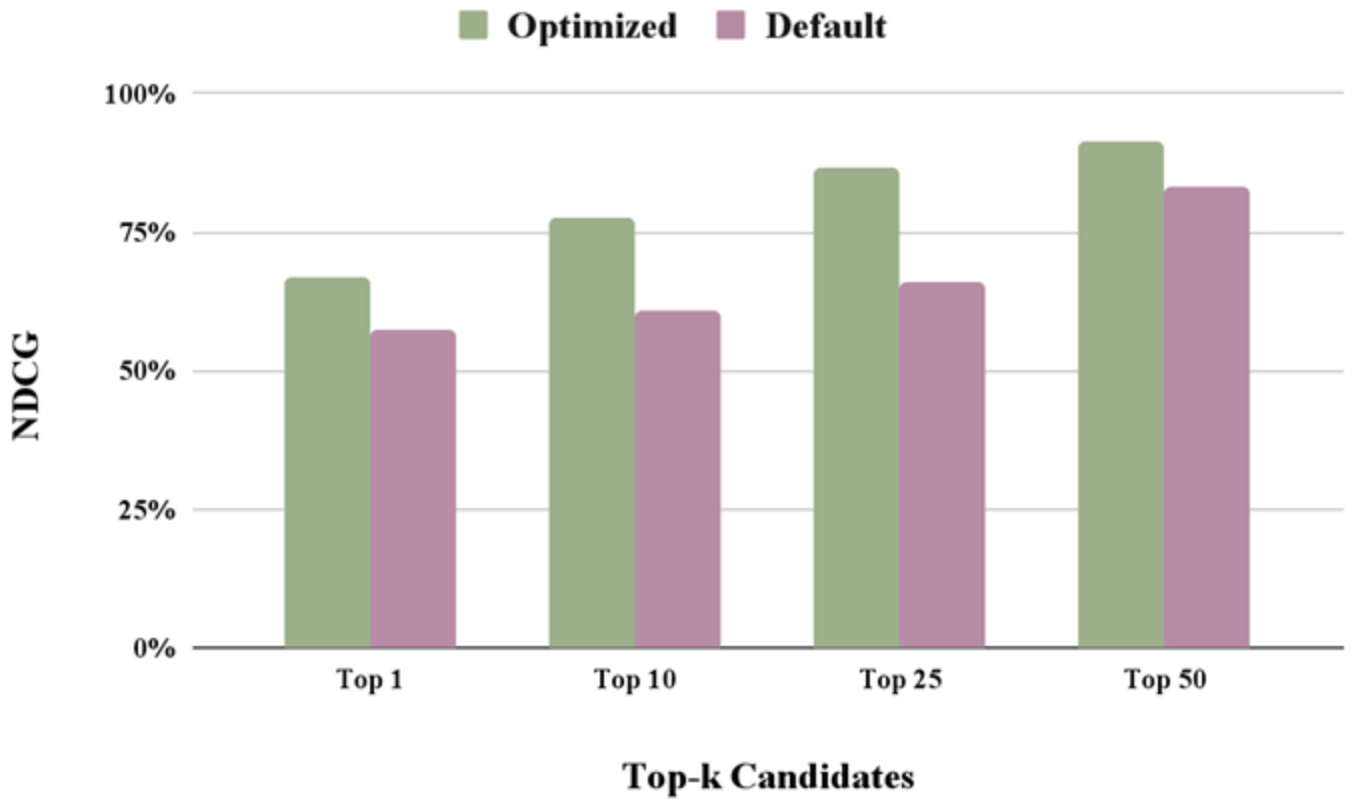
[5] Bushuiev, R., Bushuiev, A., de Jonge, N., Young, A., Kretschmer, F., Samusevich, R., ... & Pluskal, T. (2024). MassSpecGym: A benchmark for the discovery and identification of molecules. Advances in Neural Information Processing Systems, 37, 110010-110027.

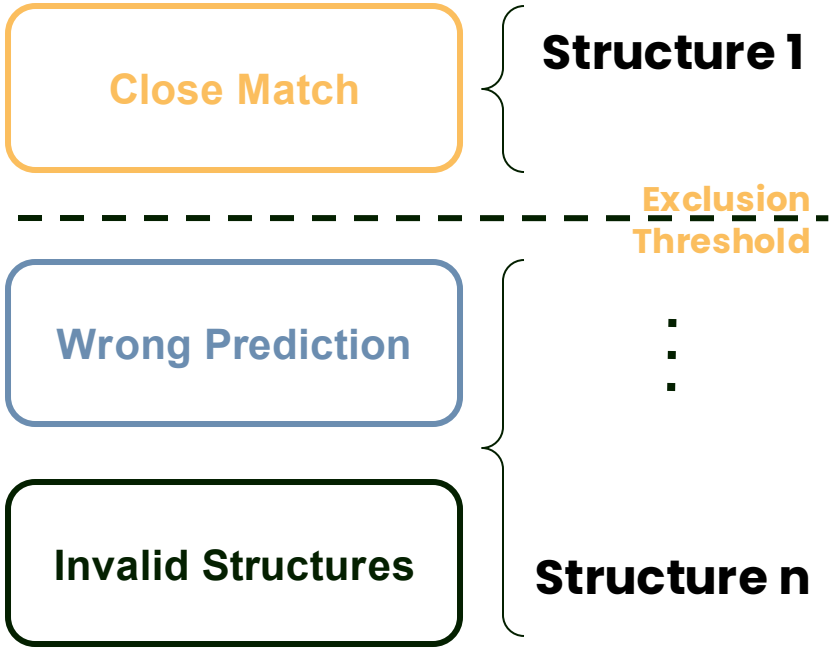
Results | Knowledge of the molecular features improves predicting the correct structure in the *top-k* candidates



[1]Exact Match = Correct prediction having tanimoto coefficient of 1.0 with label
[2] Optimized = Weighted ranking based on predicted molecular features
[3] Default = Ranking of predictions by the generative model
[4] Fiehn, O. (2022). CASMI 2022: Critical Assessment of Small Molecule Identification. Fiehn Lab, UC Davis.
[5] Bushuiev, R., Bushuiev, A., de Jonge, N., Young, A., Kretschmer, F., Samusevich, R., ... & Pluskal, T. (2024). MassSpecGym: A benchmark for the discovery and identification of molecules. Advances in Neural Information Processing Systems, 37, 110010-110027.

Results | Knowledge of the molecular features improves candidates ranking

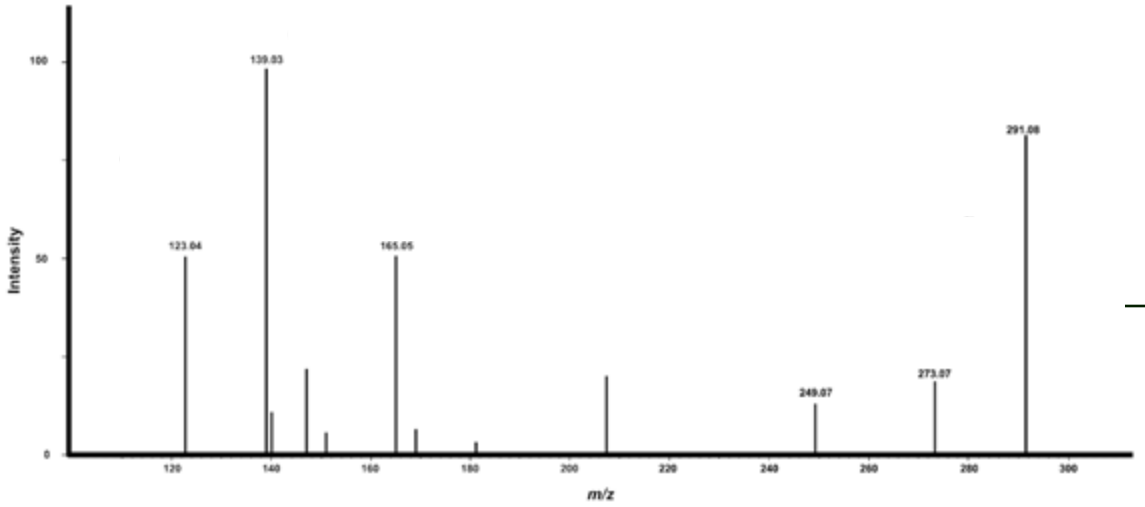


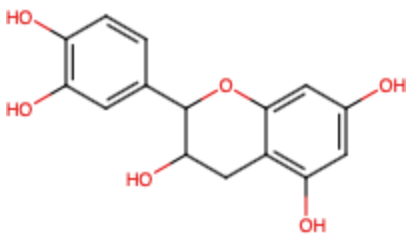


1. NDCG = Normalized Discounted Cumulative Gain
2. Optimized = Weighted ranking based on predicted molecular features
3. Default = Ranking of predictions by the generative model

Task 3 | Chemical class prediction from spectra

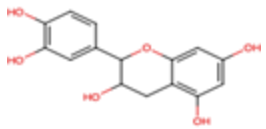
Epicatechin





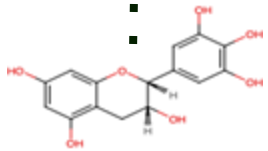
C1C(C(OC2=CC(=CC(=C21)O)O)C3=CC(=C(C=C3)O)O)O

1



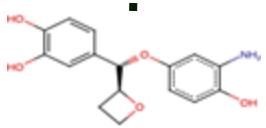
Epicatechin

5



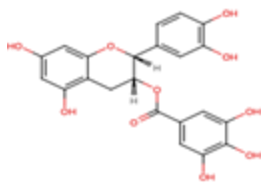
Epigallocatechin (EGC)

10



New

25



Epicatechin gallate (ECG)

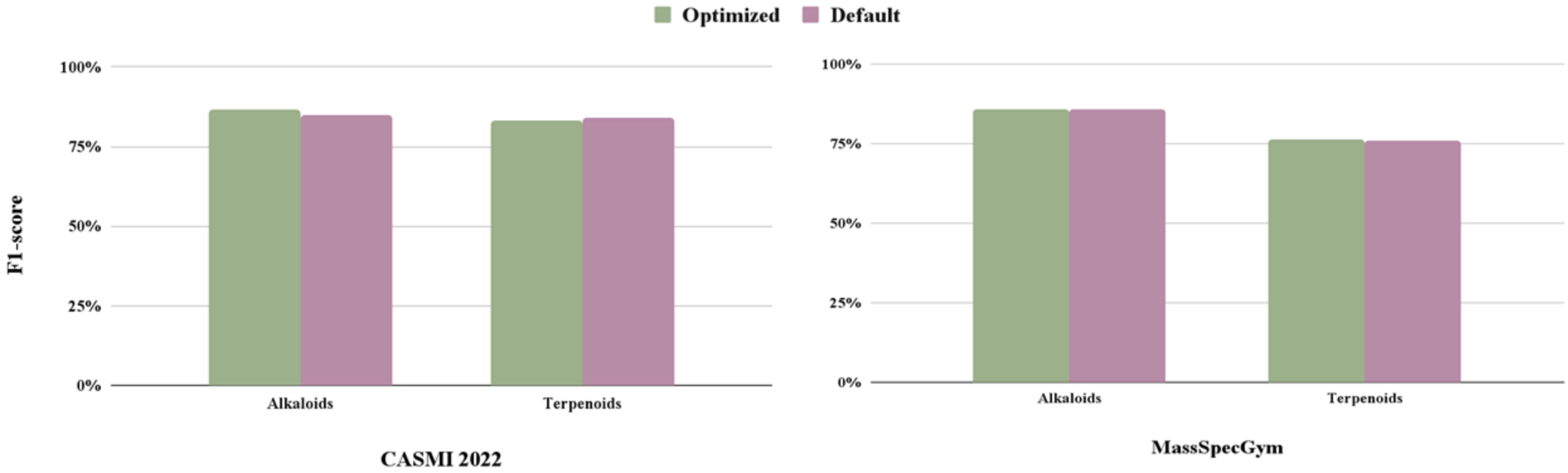
NP Classifier

Flavonoids

1.Spectrum figure adapted from Wolf, S., Schmidt, S., Müller-Hannemann, M. et al. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics 11, 148 (2010). https://doi.org/10.1186/1471-2105-11-148

2. Kim HW, Wang M, Leber CA, Nothias LF, Reher R, Kang KB, van der Hooft JJJ, Dorrestein PC, Gerwick WH, Cottrell GW. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. J Nat Prod. 2021 Oct 18. doi: 10.1021/acs.jnatprod.1c00399. Epub ahead of print. PMID: 34662515.

Results | Chemical class prediction from spectra has F1-score > 75%



[1] F1-score = harmonic mean of precision and recall
[2] Optimized = Weighted ranking based on predicted molecular features
[3] Default = Ranking of predictions by the generative model
[4] Fiehn, O. (2022). CASMI 2022: Critical Assessment of Small Molecule Identification. Fiehn Lab, UC Davis.
[5] Bushuiev, R., Bushuiev, A., de Jonge, N., Young, A., Kretschmer, F., Samusevich, R., ... & Pluskal, T. (2024). MassSpecGym: A benchmark for the discovery and identification of molecules. Advances in Neural Information Processing Systems, 37, 110010-110027.

Conclusion | SpectruMS pushes the boundaries of known MS2 chemistry

Some learnings:

- Model generalizability is as good as its understanding of the language represented as unlabeled data
- Specialized models performance is highly dependent on the base model and pre-training data sizes
- Knowledge of applicability domain of a model is more important than model performance

Potential Avenues for Improvement:

- Train a larger base model with an extended MS2 data corpus including spectra with lower quality
- Incorporate molecular fragment-based and physics-informed training
- Build a comparable model such as Graph Transformers

Thank you!

- **Daniel Crusius**
- **Tornike Onoprishvili**
- Kamen Petrov
- Klaudia Caba
- Jui-Hung Yuan
- Sona Chandra
- Andreas Bender
- Yoann Gloaguen