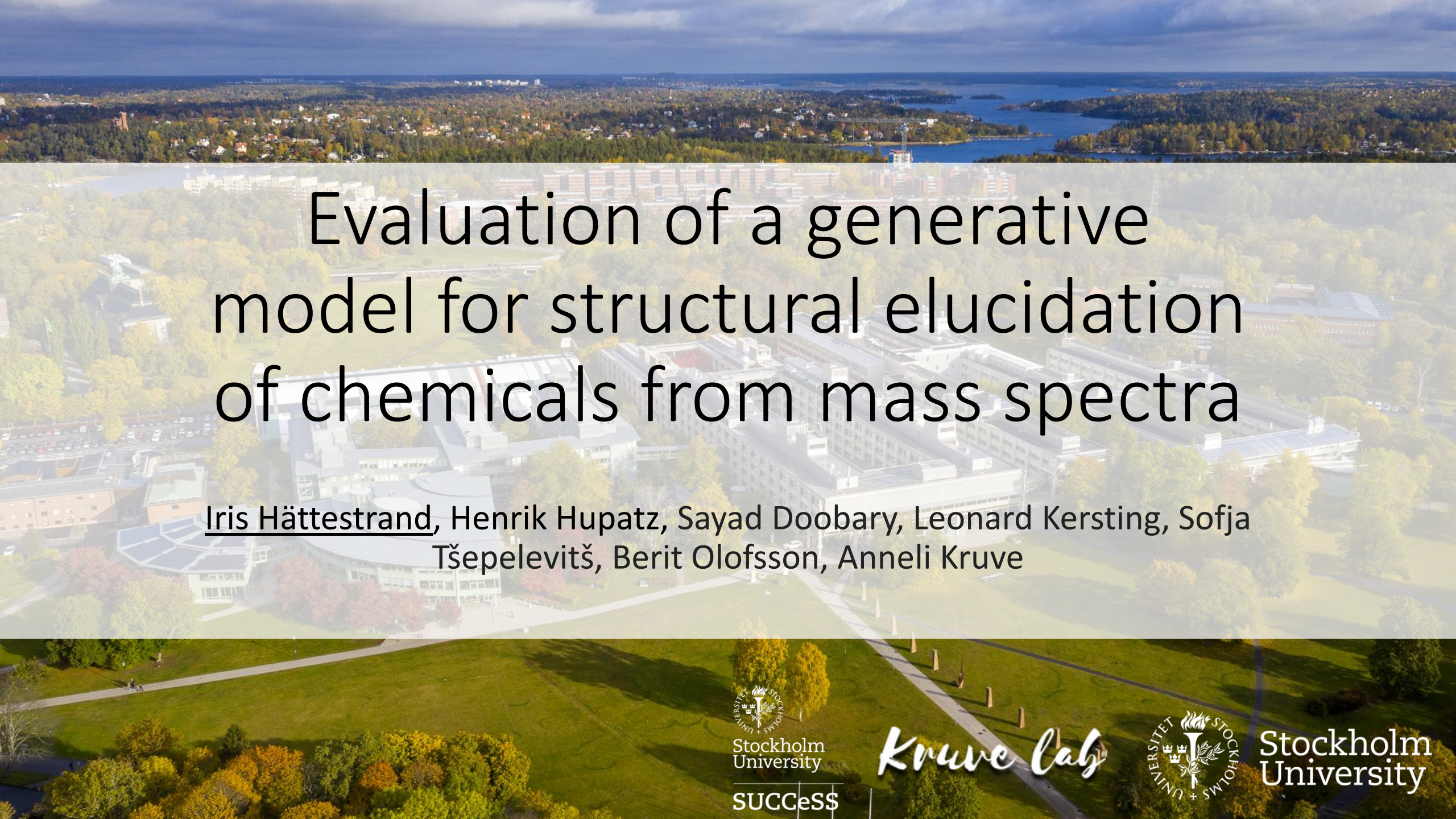


Evaluation of a generative model for structural elucidation of chemicals from mass spectra

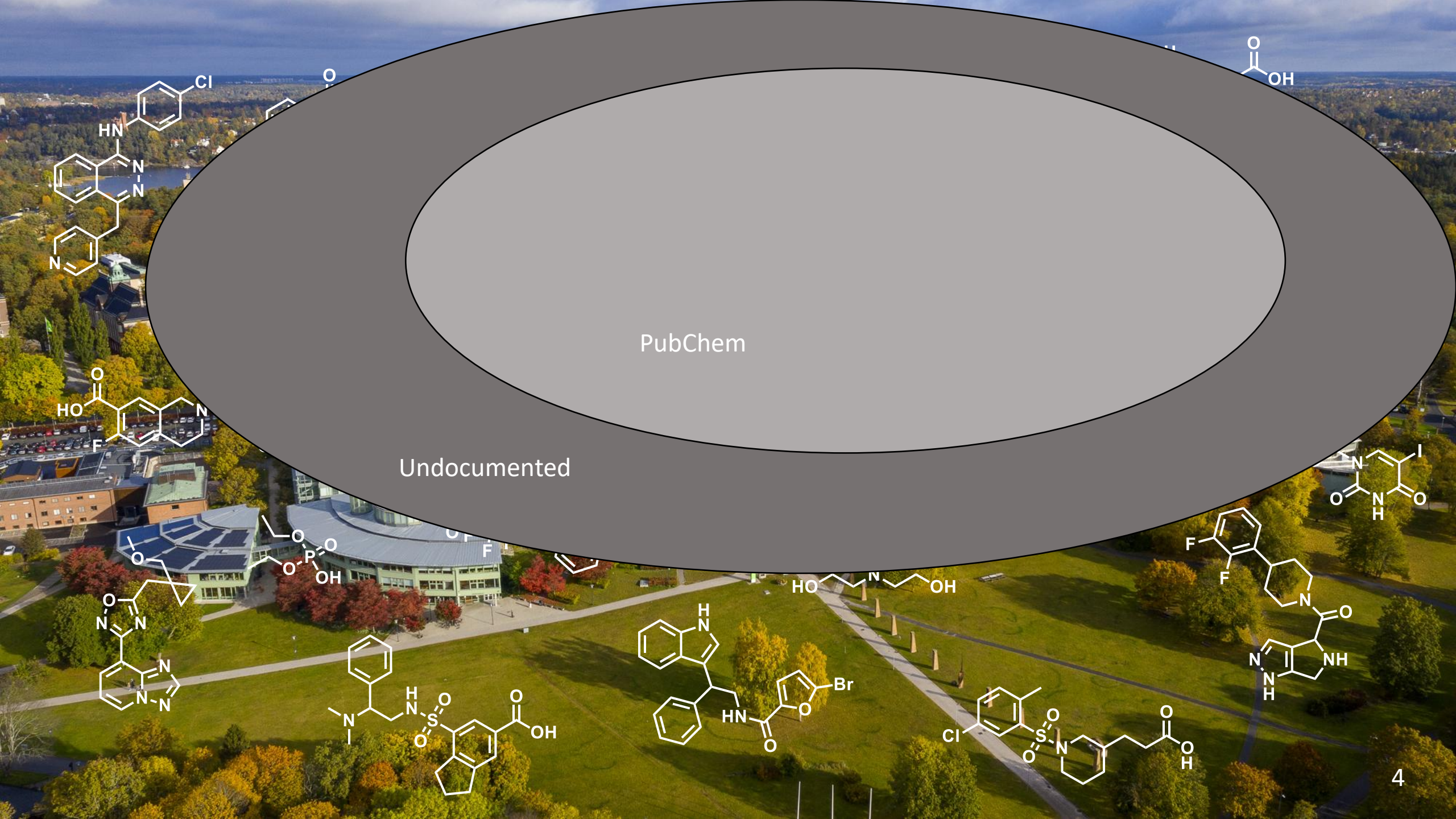
Iris Hättestrand, Henrik Hupatz, Sayad Doobary, Leonard Kersting, Sofja Tšepelevitš, Berit Olofsson, Anneli Kruve

Stockholm University

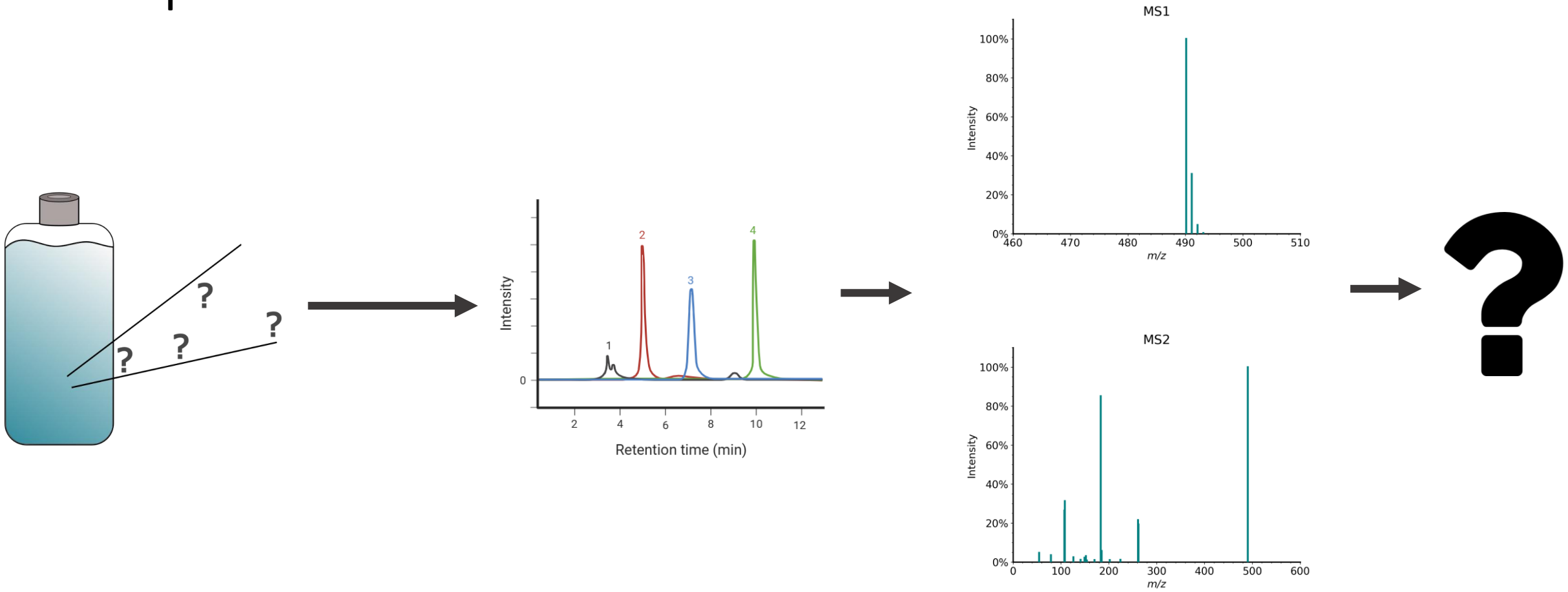
SUCCeSS

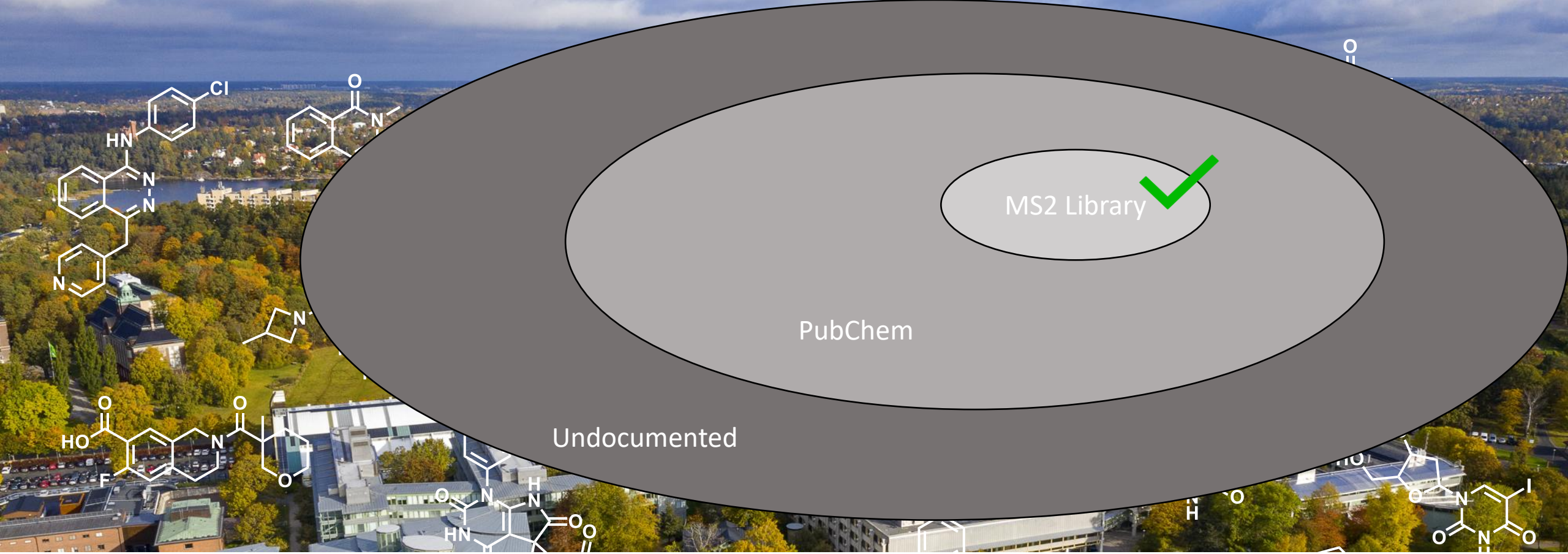
Kruve Lab

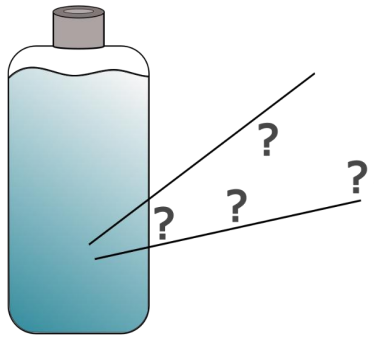
Stockholm University

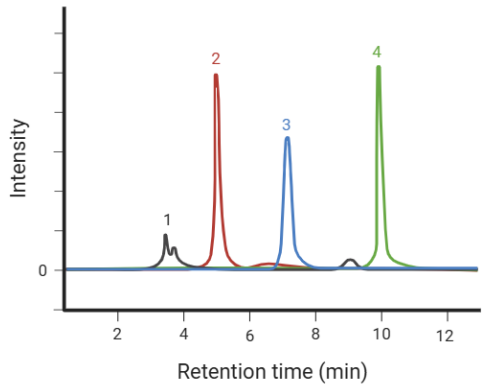

PubChem
Undocumented

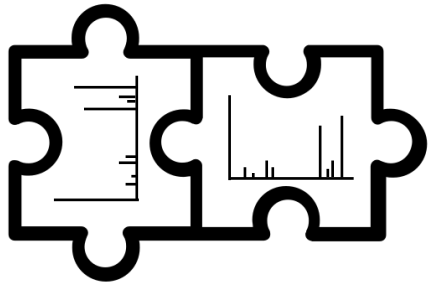
Experimental workflow

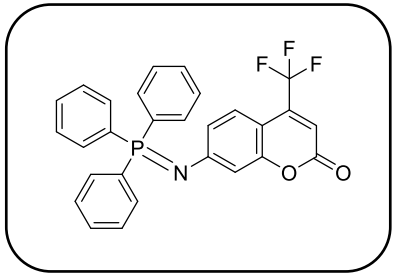


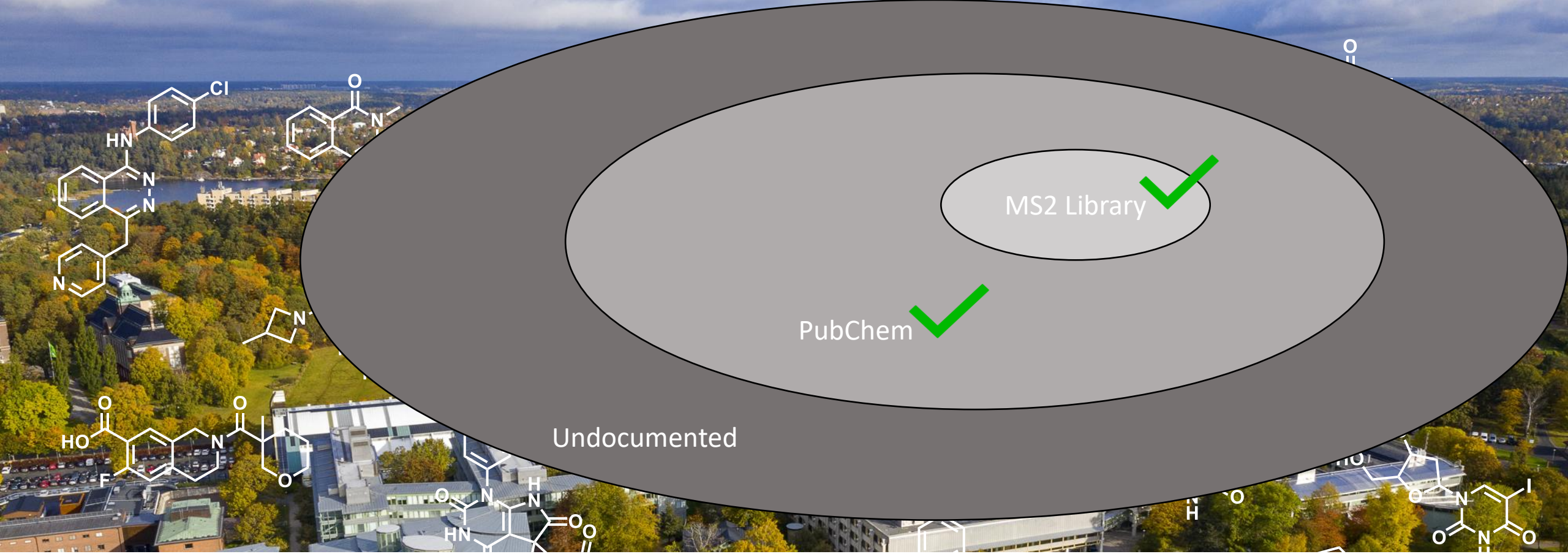

MS2 Library
PubChem
Undocumented

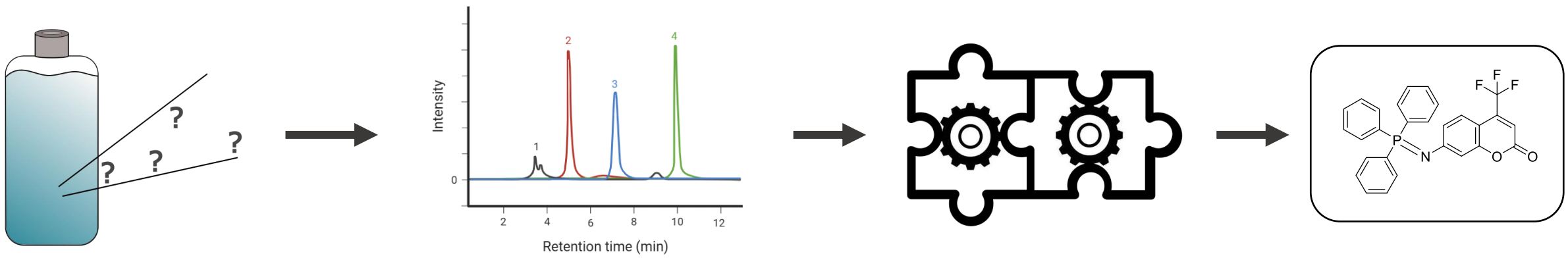



Intensity
Retention time (min)
1
2
3
4

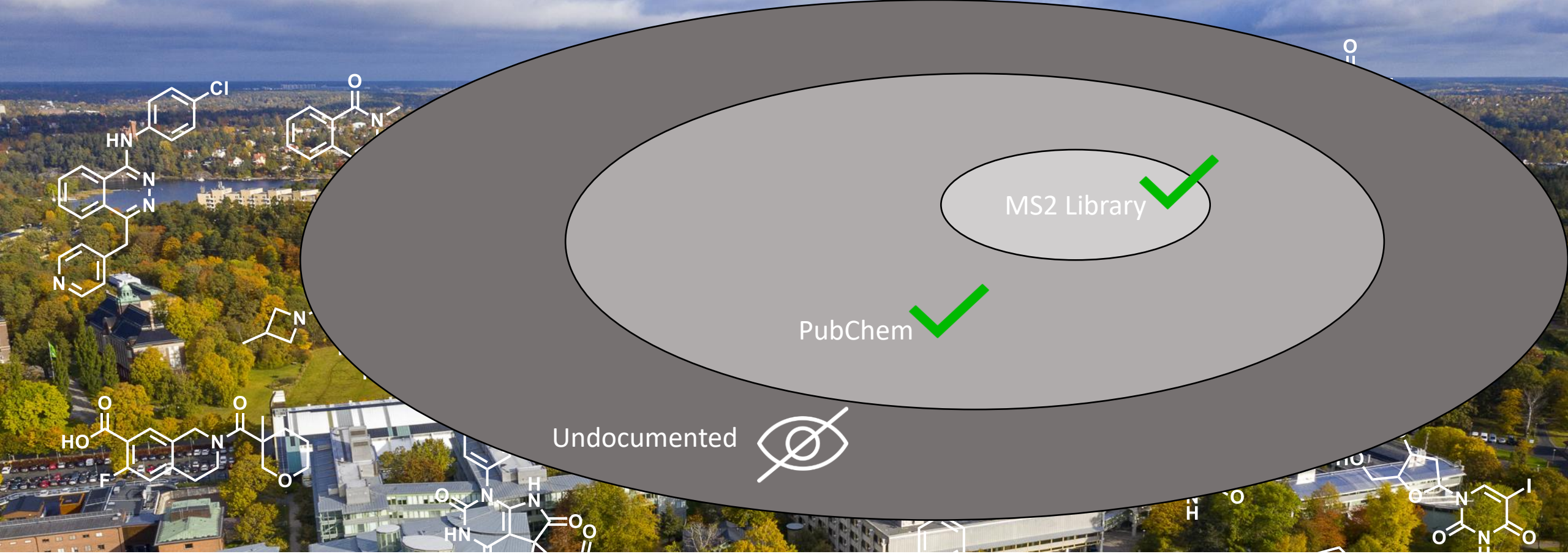


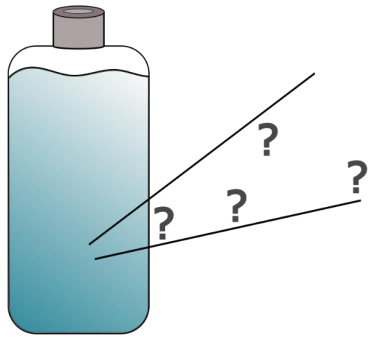


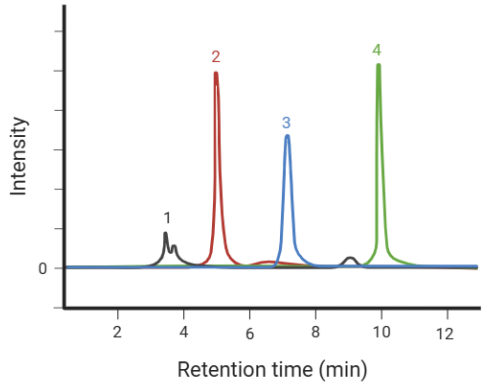




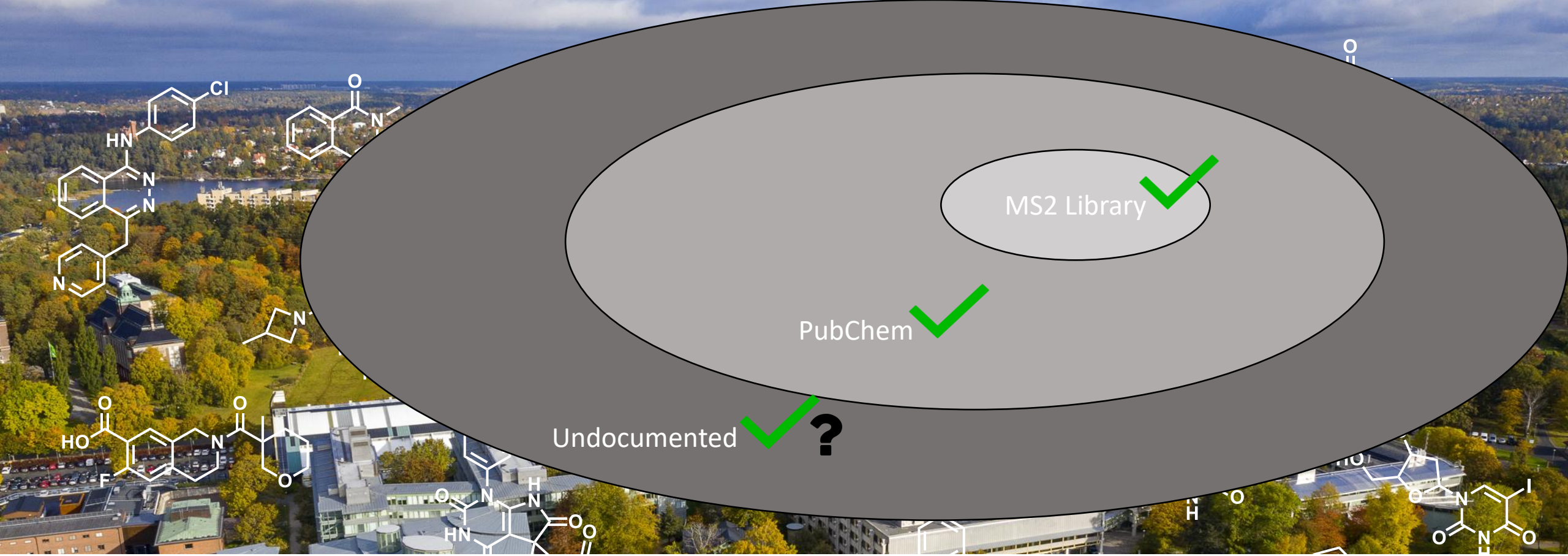
Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; et al., Nat. Methods, 2019, 299
Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Proc. Natl. Acad. Sci, 2015, 12580

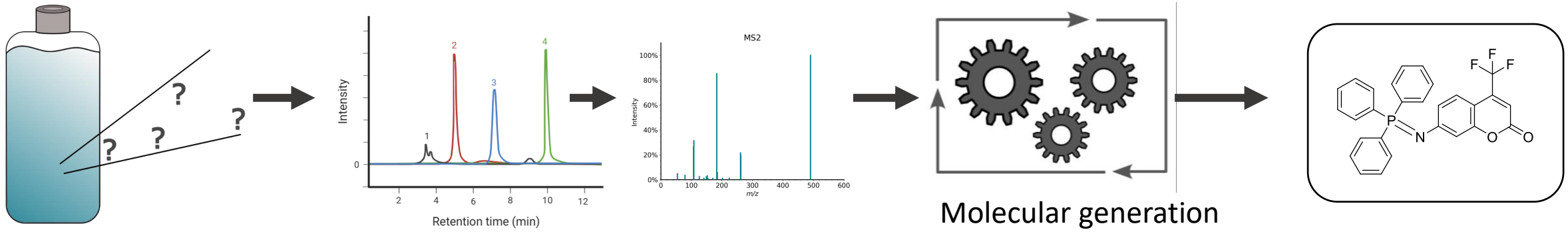

MS2 Library
PubChem
Undocumented



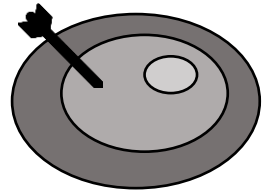

Intensity
Retention time (min)
1
2
3
4



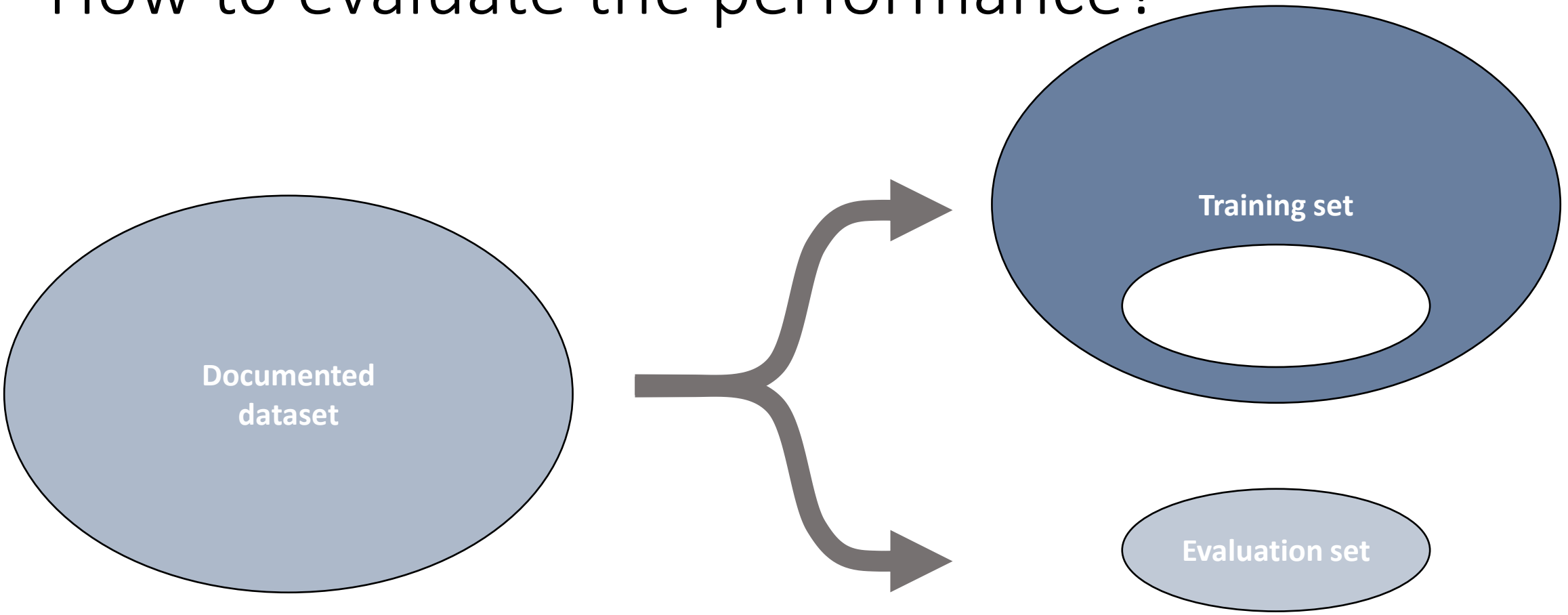




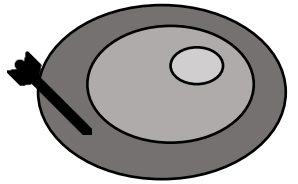
Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. Nat. Methods, 2022, 865

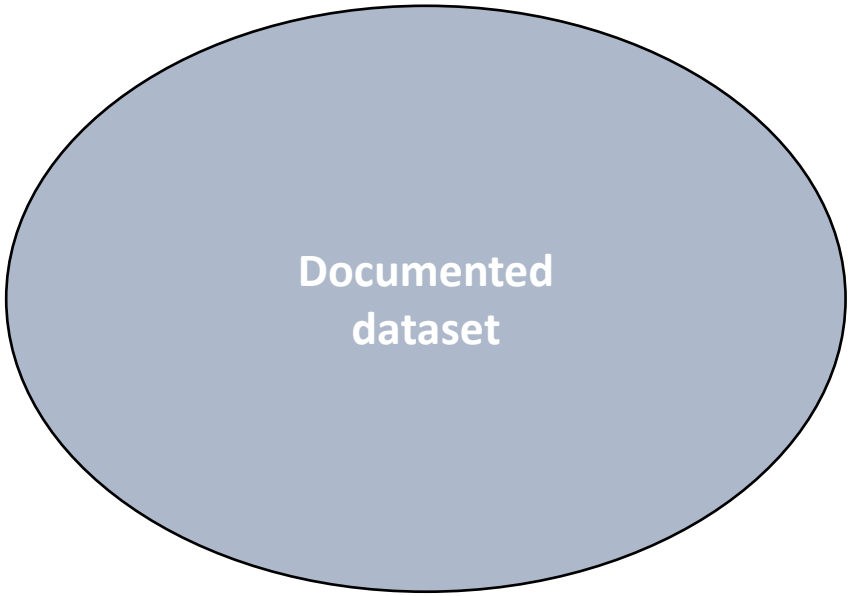


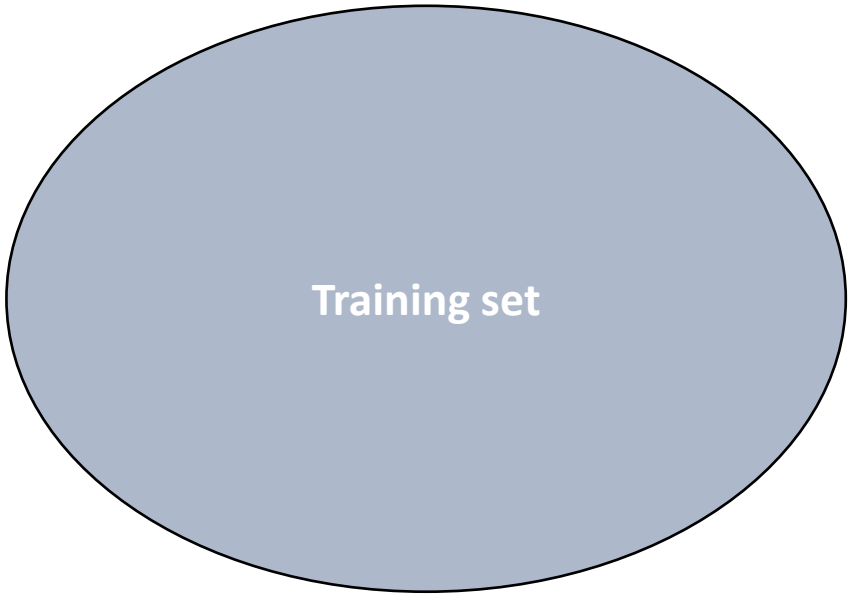
How to evaluate the performance?

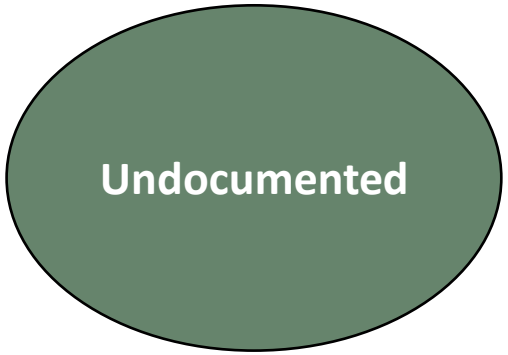


Evaluating the performance

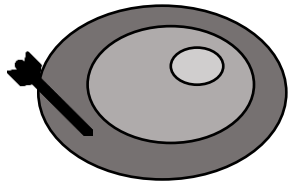






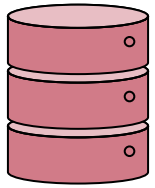


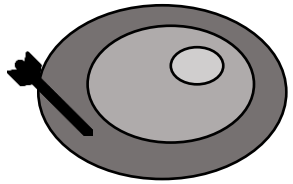
Datasets



1. In-house dataset

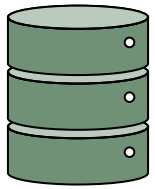
- Undocumented
- In-house measurements
- 50 chemicals

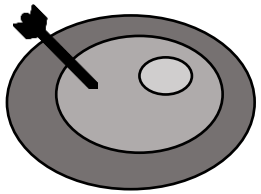




2. MSnLib undocumented dataset

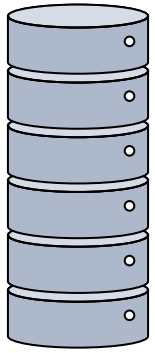
- Undocumented
- MSnLib
- 58 chemicals





3. Documented literature dataset

- Documented
- MassBank EU, MSnLib and MassSpecGym
- 35,895 chemicals

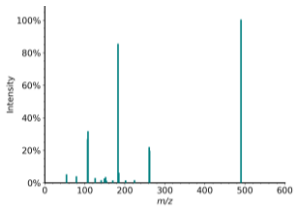




Brungs, C.; Schmid, R.; Heuckeroth, S.; Mazumdar, A.; et al., Nat. Methods 2025, 2028
Bushuiev, R.; Bushuiev, A.; Jonge, N. F. de; Young, A.; et al., arXiv 2025

Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. Nat. Methods, 2022, 865
Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; et al., Nat. Methods, 2019, 299; Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Proc. Natl. Acad. Sci, 2015, 12580

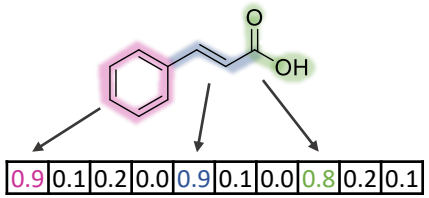
Computational workflow



MS²

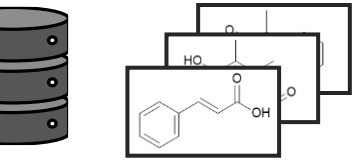
Translation
of spectra to predicted
structural features

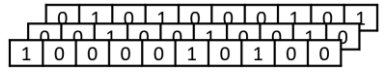
Rank	Formula	Adduct
1	$C_{10}H_{12}O$	$[M+H]^+$
2	$C_{10}H_{12}O$	$[M+H]^+$
...	...	...



Predicted fingerprint

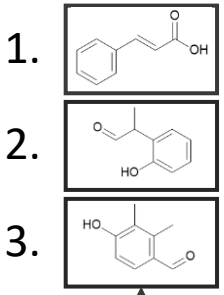
Databases
E.g., PubChem



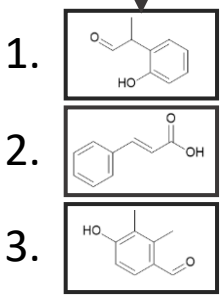


Database
matching

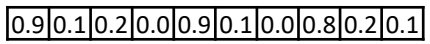
Ranking based on
CSI:FingerID score

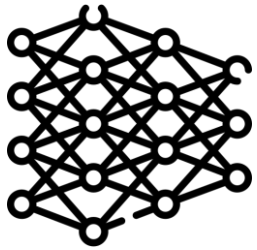


Compare predicted
and true structure

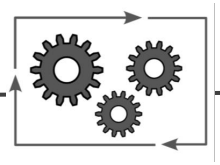


$C_{10}H_{12}O$
+





RNN



De novo
molecular generation

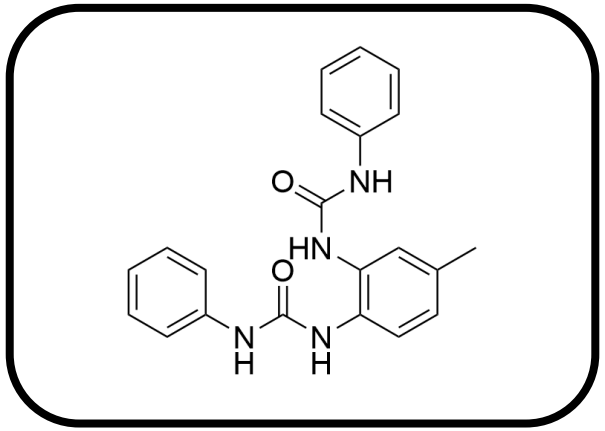
Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; et al., Nat. Methods, 2019, 299
Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Proc. Natl. Acad. Sci, 2015, 12580
Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. Nat. Methods, 2022, 865

Results undocumented datasets

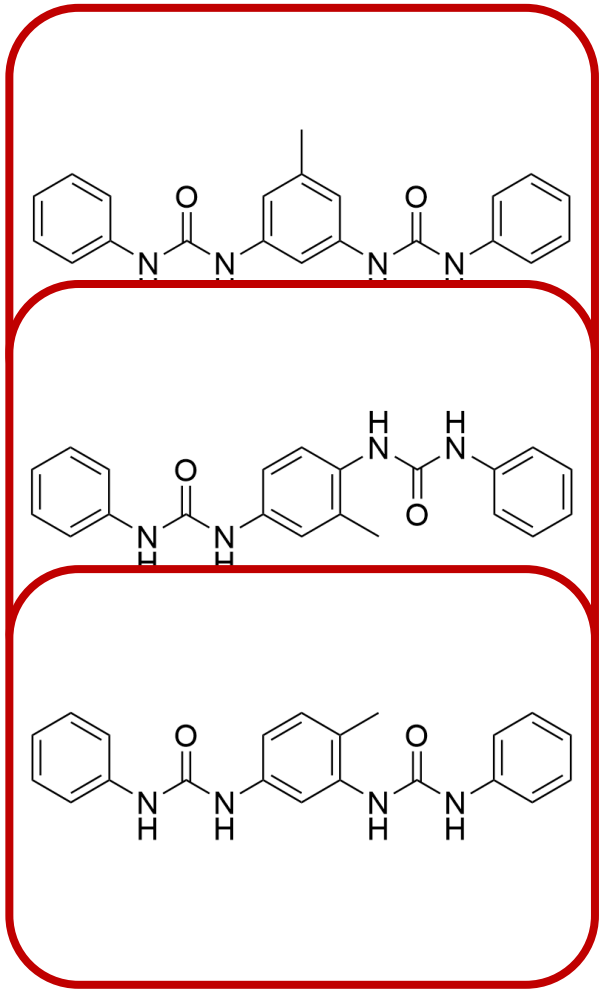
Dataset	n chemicals	n features	Correct annotations top-10	Correct annotation top-10 (%)
In-house	50	90	1	1.1
MSnLib	58	58	1	1.7

Results undocumented datasets

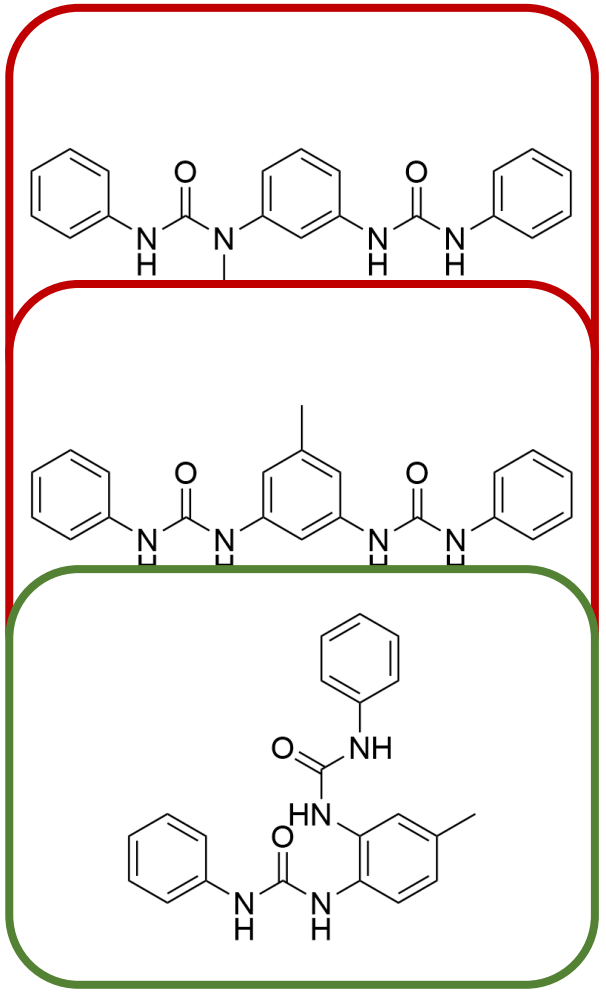
True structure



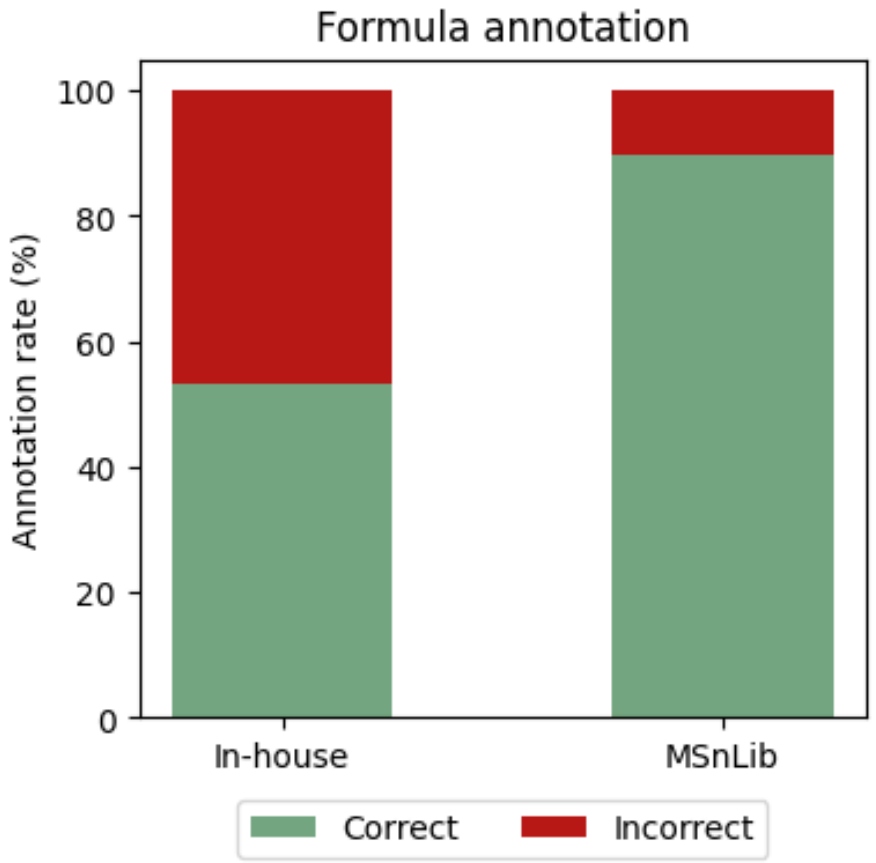
Structural database matching



De novo



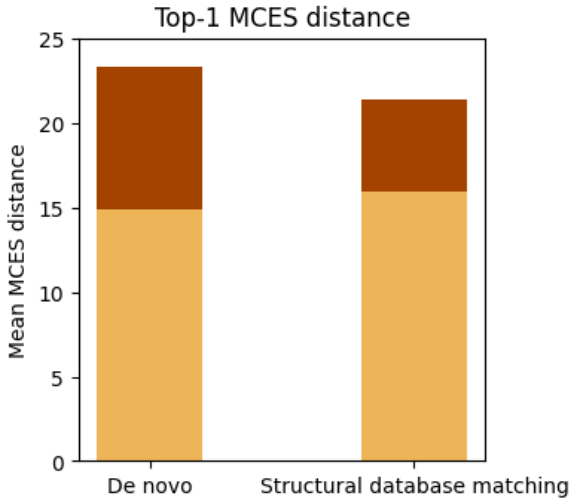
Results undocumented datasets

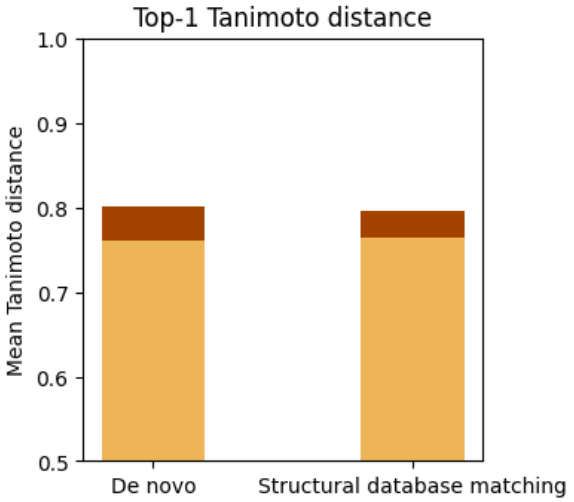


Results undocumented datasets

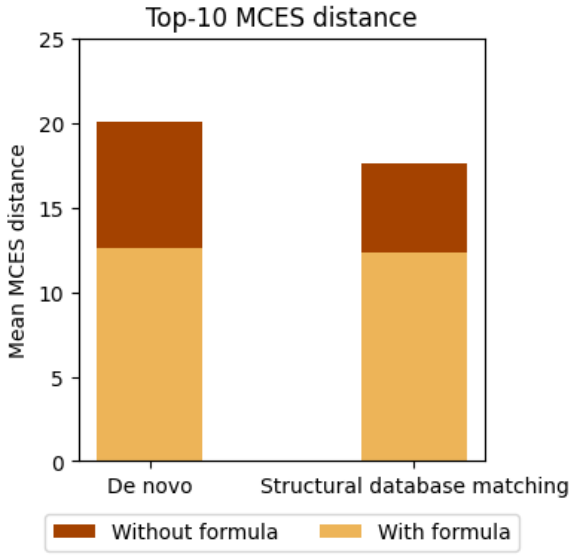
Dataset	n chemicals	n features	Correct annotations top-10	Correct annotation %
In-house	50	90	1	1.1
MSnLib	58	58	1	1.7

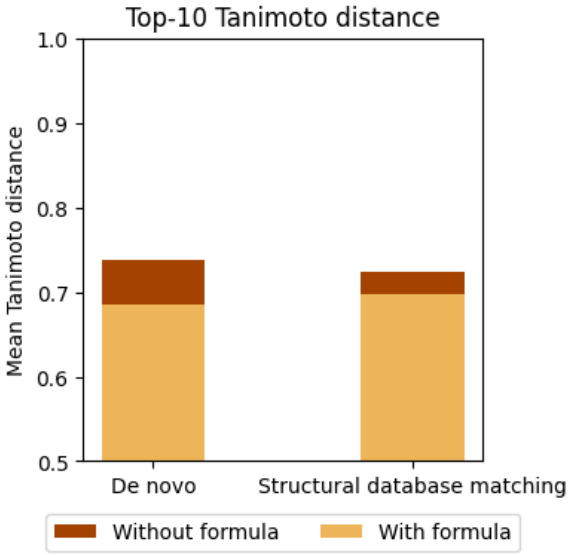
Results undocumented datasets



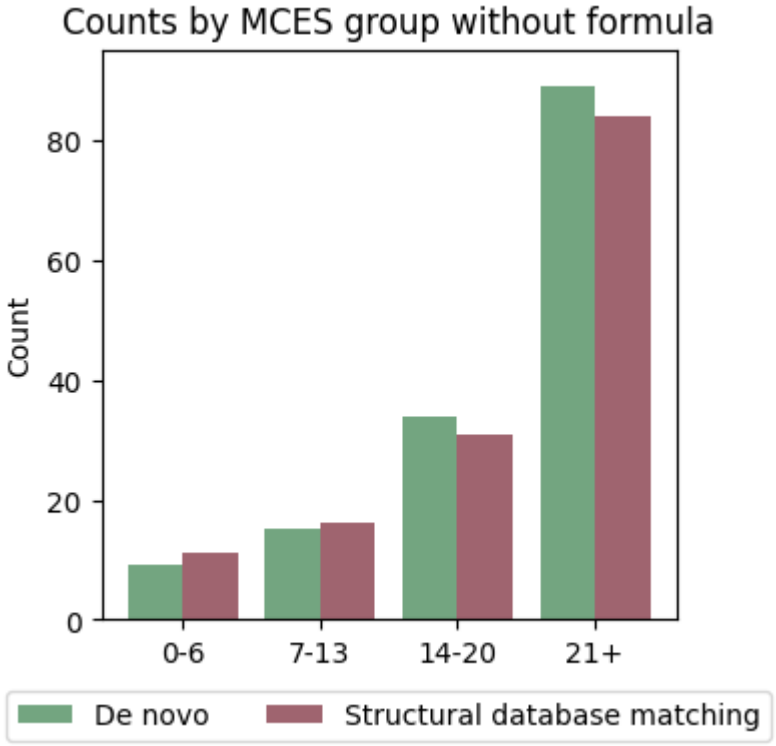


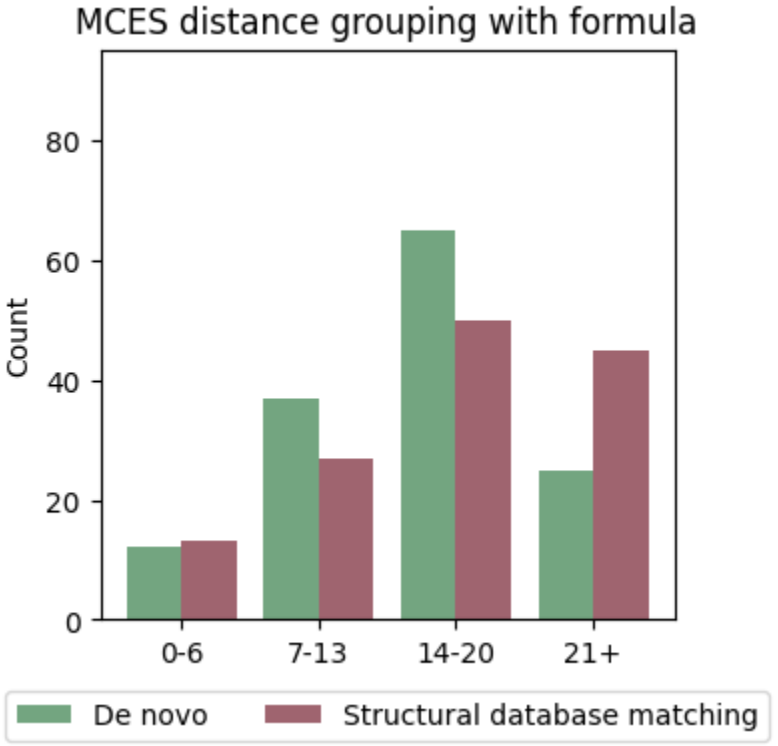
ECFP-4 (2048 bits)





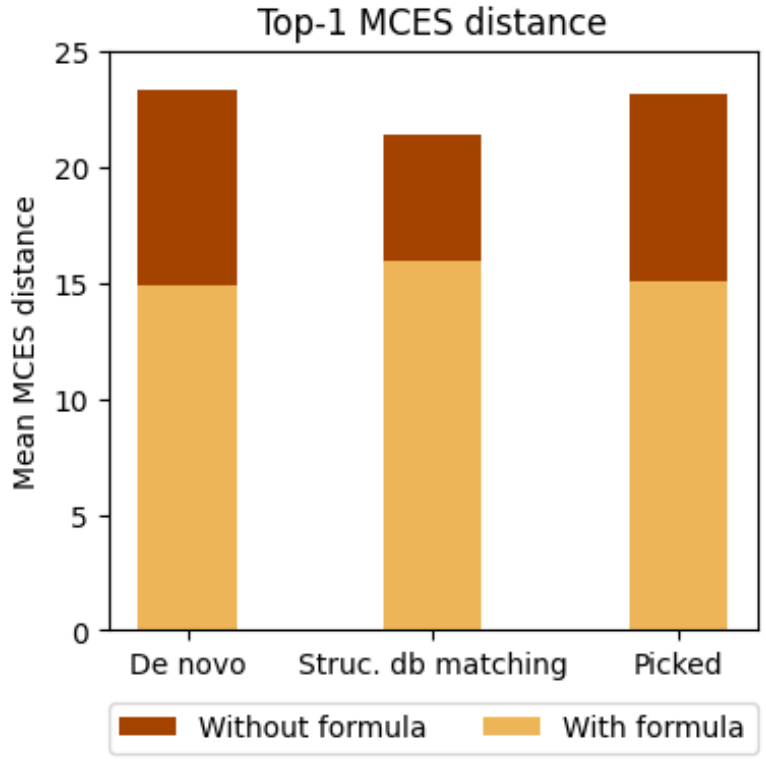
Results undocumented datasets

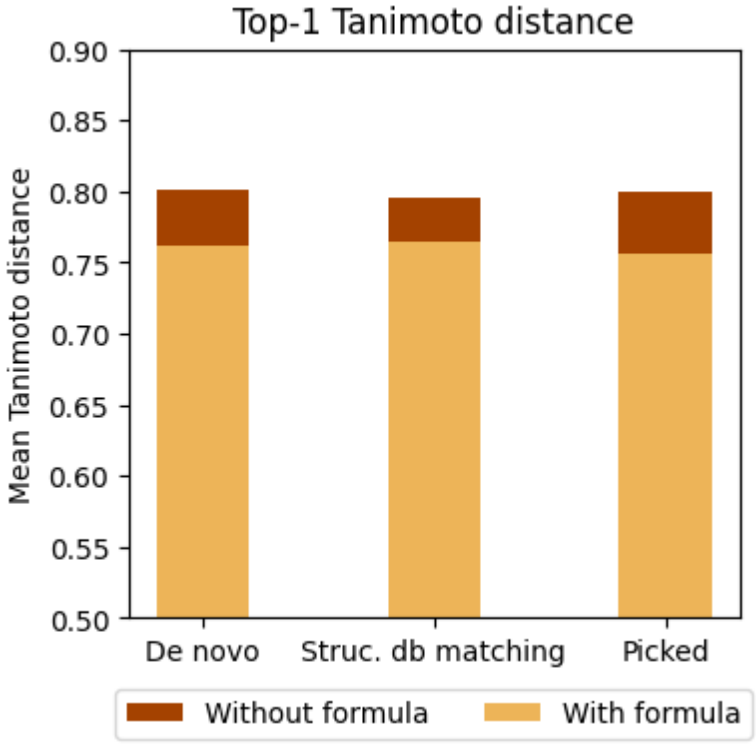




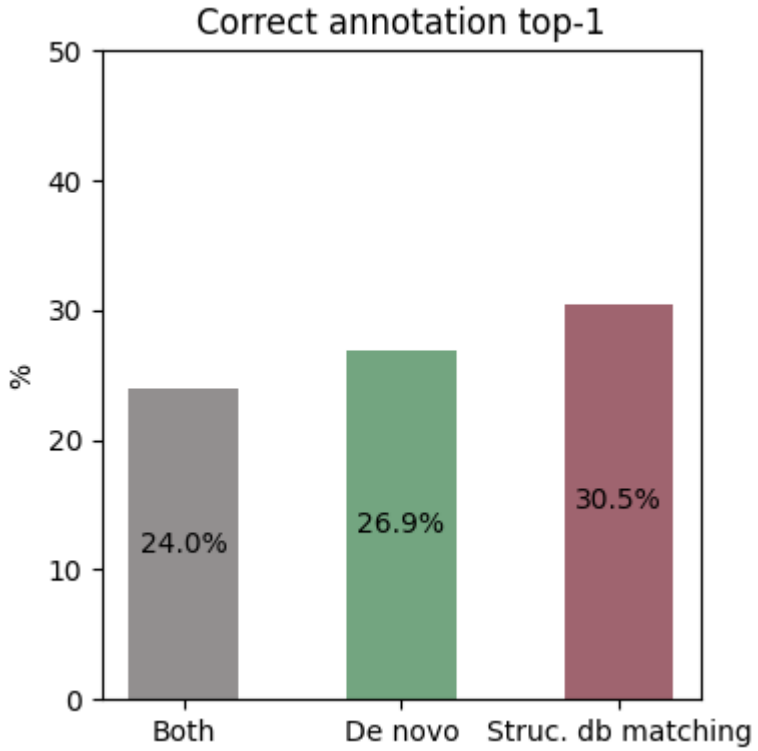
0-6	High similarity
7-13	Medium similarity
14-20	Low similarity
21+	Not similar

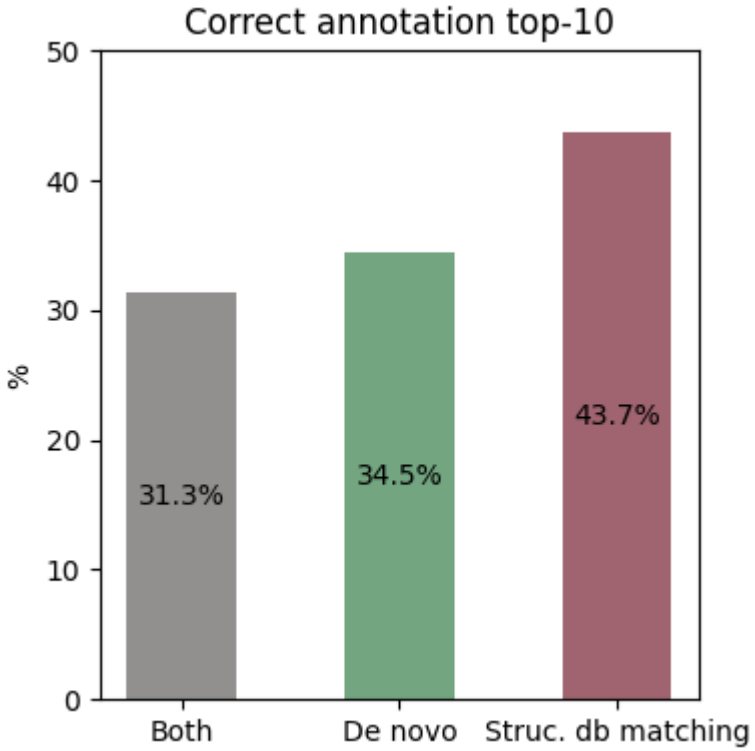
Results undocumented datasets





Results literature dataset



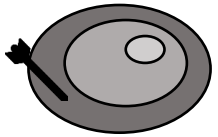
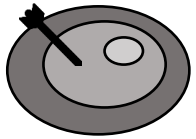


Conclusions

Literature results

Model	Evaluation dataset from	*De novo* Top-1	*De novo* Top-10
MSNovelist[1]	GNPS	25 %	
MIST + MSNovelist[2]	MassSpecGym	0 %	0 %
MassSpecGym[3]	MassSpecGym	0 %	0 %
DiffMS[2]	MassSpecGym	2.3 %	4.3 %

Our results

	De novo Top-1	*De novo* Top-10
	0 %	1.4 %
	26.9 %	34.5 %

- Undocumented chemicals give challenging datasets for generative models
- Undocumented structures are possible to annotate with a generative model

[1] Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. Nat. Methods, 2022, 865; [2] Bushuiev, R.; Bushuiev, A.; Jonge, N. F. de; Young, A.; et al., arXiv 2025
[3] Bohde, M.; Manjrekar, M.; Wang, R.; Ji, S.; Coley, C. W, arXiv, 2025

Evaluation of a generative model for structural elucidation of chemicals from mass spectra

Iris Hättestrand, Henrik Hupatz, Sayad Doobary, Leonard Kersting, Sofja Tšepelevitš, Berit Olofsson, Anneli Kruve



Stockholm University

Kruve lab

European Research Council
Established by the European Commission

Contact: iris.hattestrand@su.se