

Exploring generative modelling for structural annotation of LC/HRMS features in environmental non-targeted screening

Iris Hättestrand¹, Henrik Hupatz^{1,2}, Anneli Kruve^{1,2}

¹Department of Chemistry, Stockholm University, Svante Arrhenius väg 16, 114 18 Stockholm

²Stockholm University Center for Circular and Sustainable Systems (SUCCeSS), 106 91 Stockholm

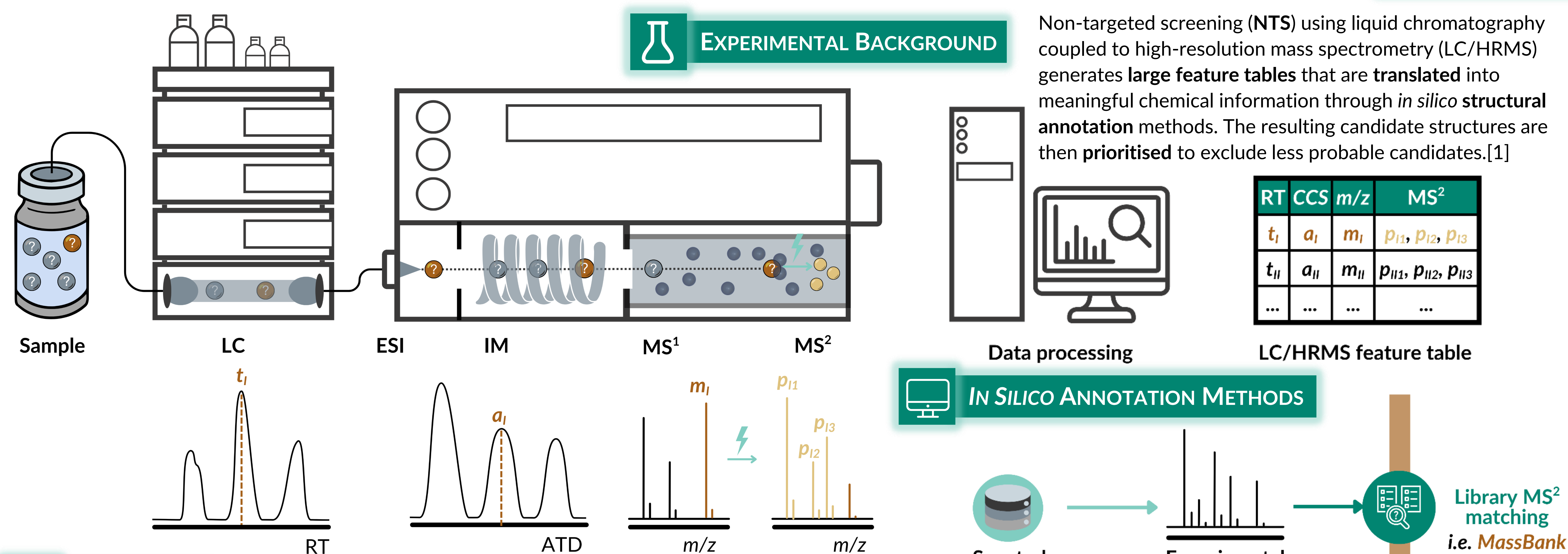


henrik.hupatz@su.se

Stockholm University
SUCCeSS
Kruve Lab

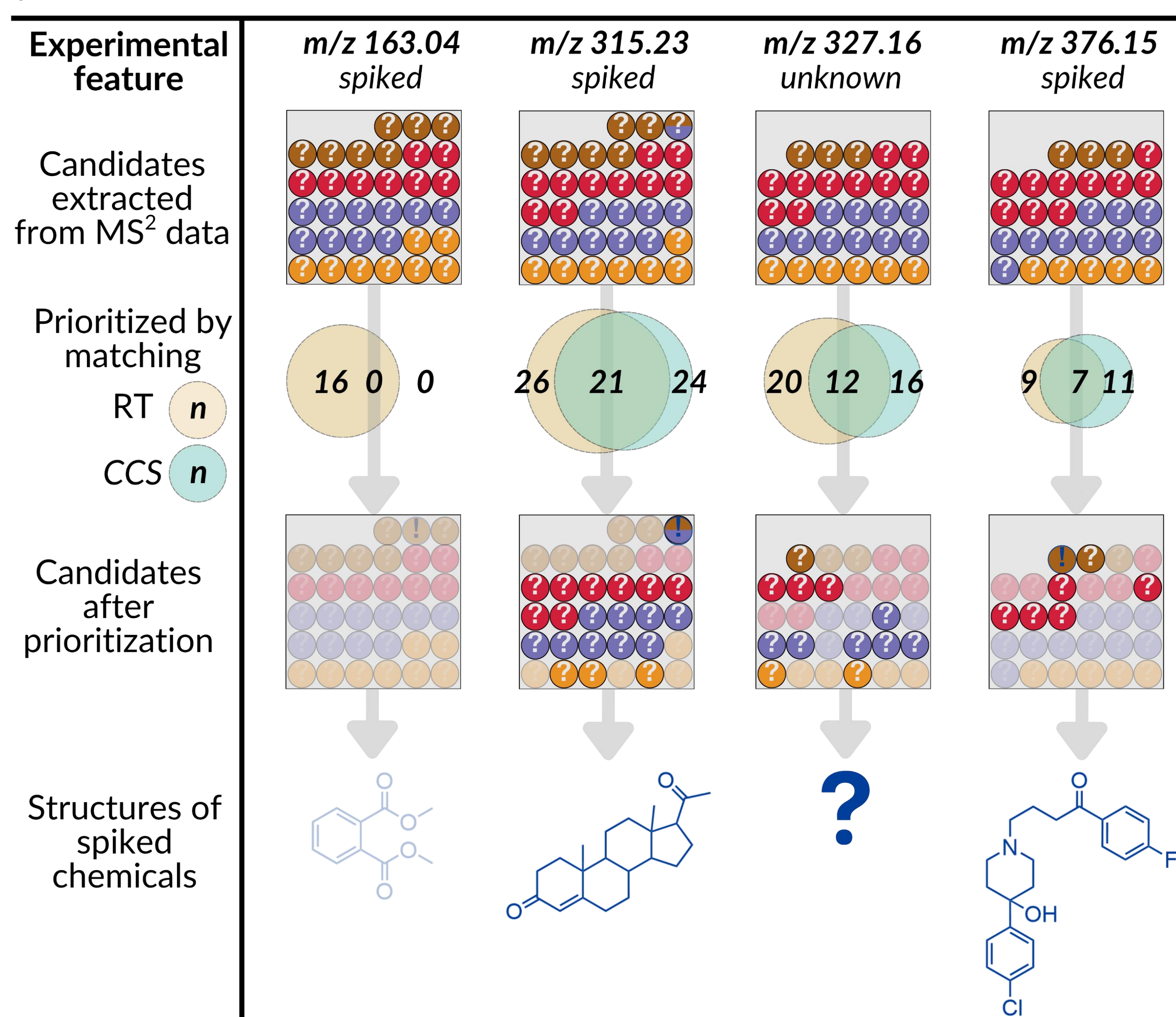


CAN GENERATIVE MODELS ANNOTATE MS² FEATURES OF NOVEL CHEMICALS?



CASE STUDY

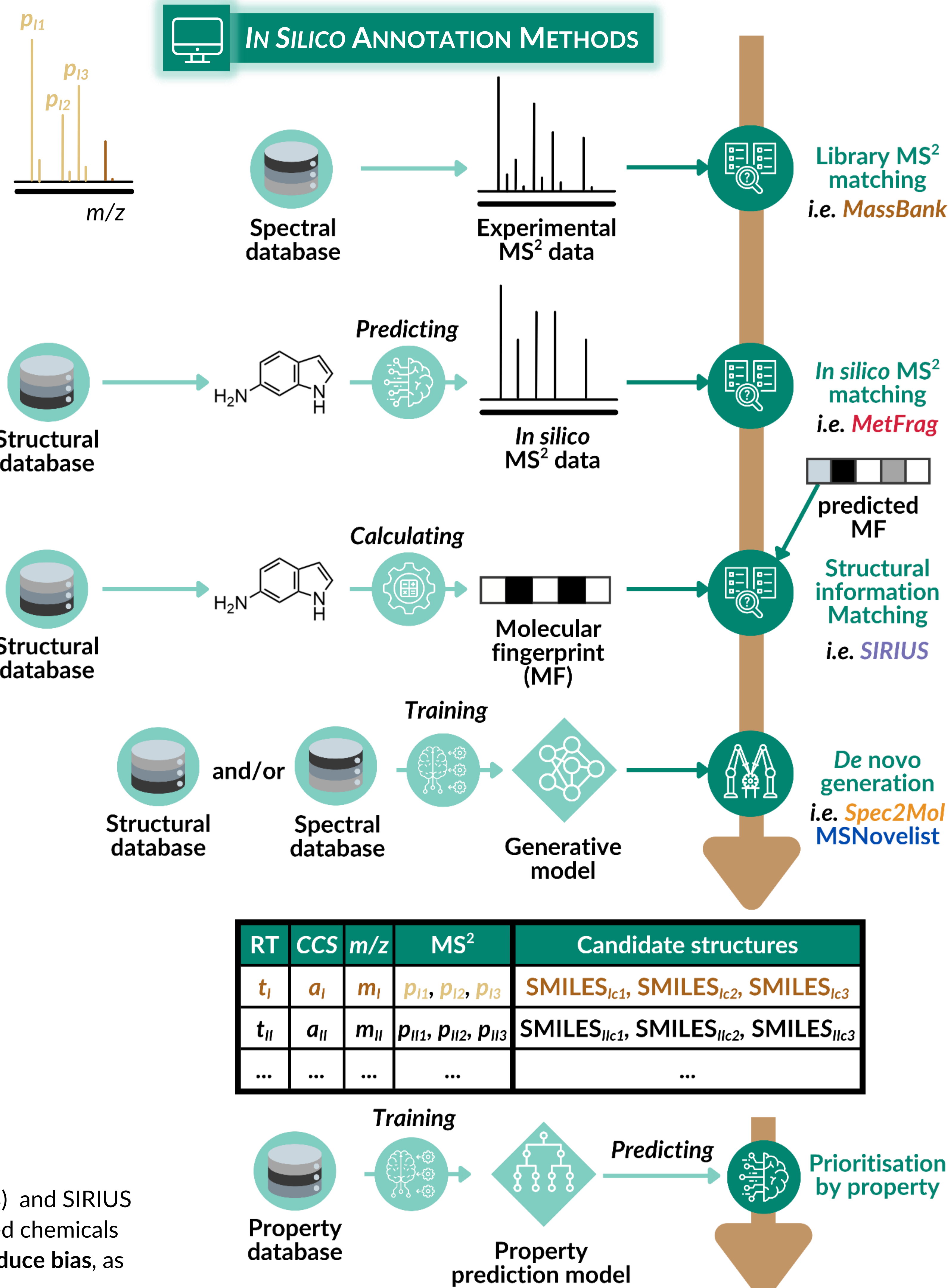
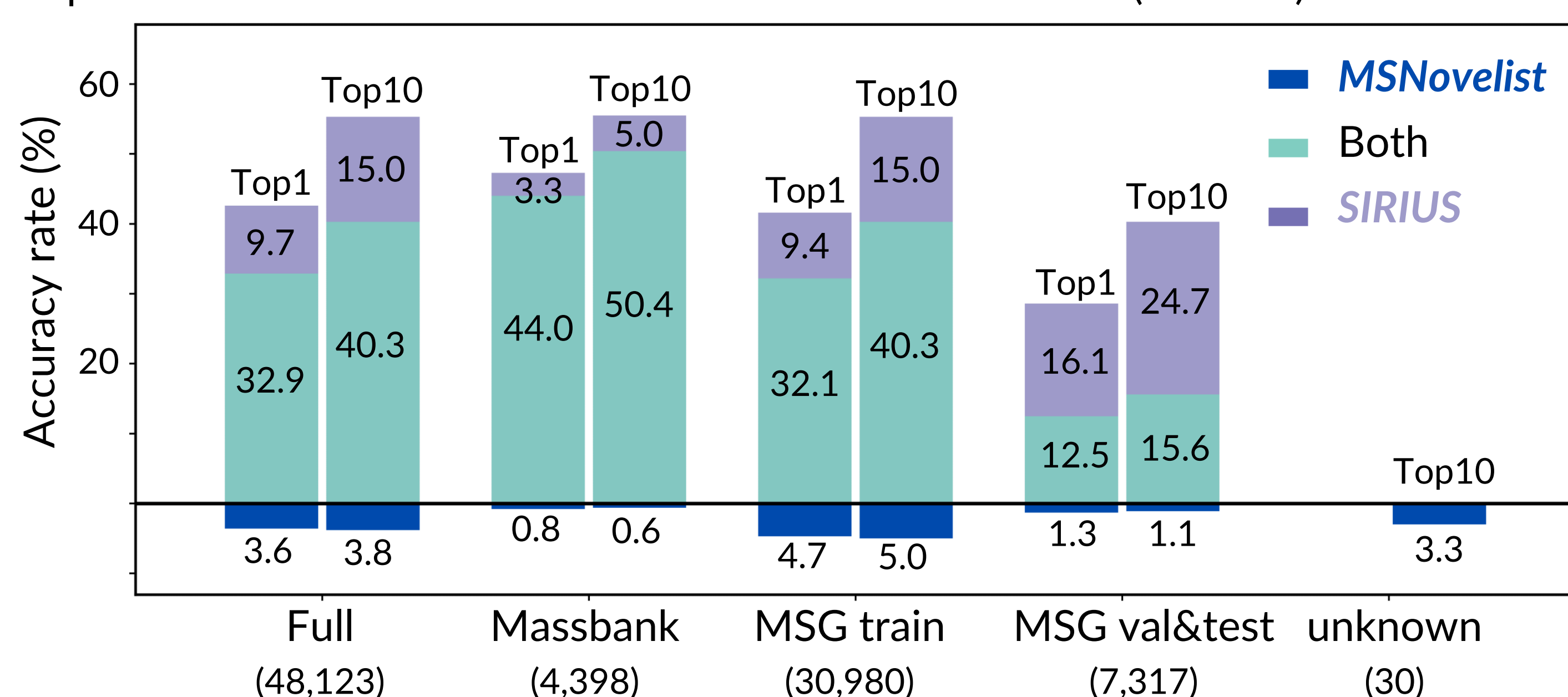
For illustration of *in silico* annotation workflows in NTS, we examined 10 spiked and 10 unknown LC/HRMS features from a wastewater sample, showcasing one example per annotation method (MassBank EU, MetFrag, SIRIUS+CSI:FingerID[2], and Spec2Mol), as well as prioritisation using predicted retention time (RT) from the RTI model[3] and predicted collision cross section (CCS) from the CCSBase[4] model.



Spiked features were correctly annotated (as blue !) by matching with MassBank (100%) and SIRIUS (50%). Predicted RTs and CCSs deprioritised the correct structures for 20% of the spiked chemicals each. Evaluating NTS workflows on spiked chemicals from known databases may introduce bias, as these chemicals are often included in the training data of *in silico* methods.

PRELIMINARY RESULTS

We investigate if a generative model (MSNovelist[5]) adds information beyond the gold-standard SIRIUS on literature data (MassBank and MassSpecGym[6] (MSG)), and if it can generate novel rather than close analogues of known chemicals on experimental datasets of chemicals absent from PubChem (unknown).



RT	CCS	m/z	MS ²	Candidate structures	RTpred	CCSpred
t_i	a_i	m_i	p_{i1}, p_{i2}, p_{i3}	SMILES _{ic1} , SMILES _{ic2} , SMILES _{ic3}	$t_{ic1}, t_{ic2}, t_{ic3}$	$a_{ic1}, a_{ic2}, a_{ic3}$
t_{ii}	a_{ii}	m_{ii}	$p_{ii1}, p_{ii2}, p_{ii3}$	SMILES _{ilc1} , SMILES _{ilc2} , SMILES _{ilc3t}	$t_{ilc1}, t_{ilc2}, t_{ilc3}$	$a_{ilc1}, a_{ilc2}, a_{ilc3}$
...

On literature data, MSNovelist added the correct structure in 3.6% (3.8%) in Top 1 (Top 10) compared to SIRIUS, but also missed three times as many correct structures, highlighting that *de novo* generation can complement but not replace established methods.

MSNovelist correctly annotated one structure of the unknown dataset in the Top 10. Comparison of the mean maximum common edge subgraph (MCES) distance and Tanimoto distance on ECFP (r = 2) fingerprints to the correct structures showed that SIRIUS annotations are slightly more similar, indicating that the generative Model MSNovelist generates structures similar to known chemicals for LC/HRMS feature of unknown chemicals.

