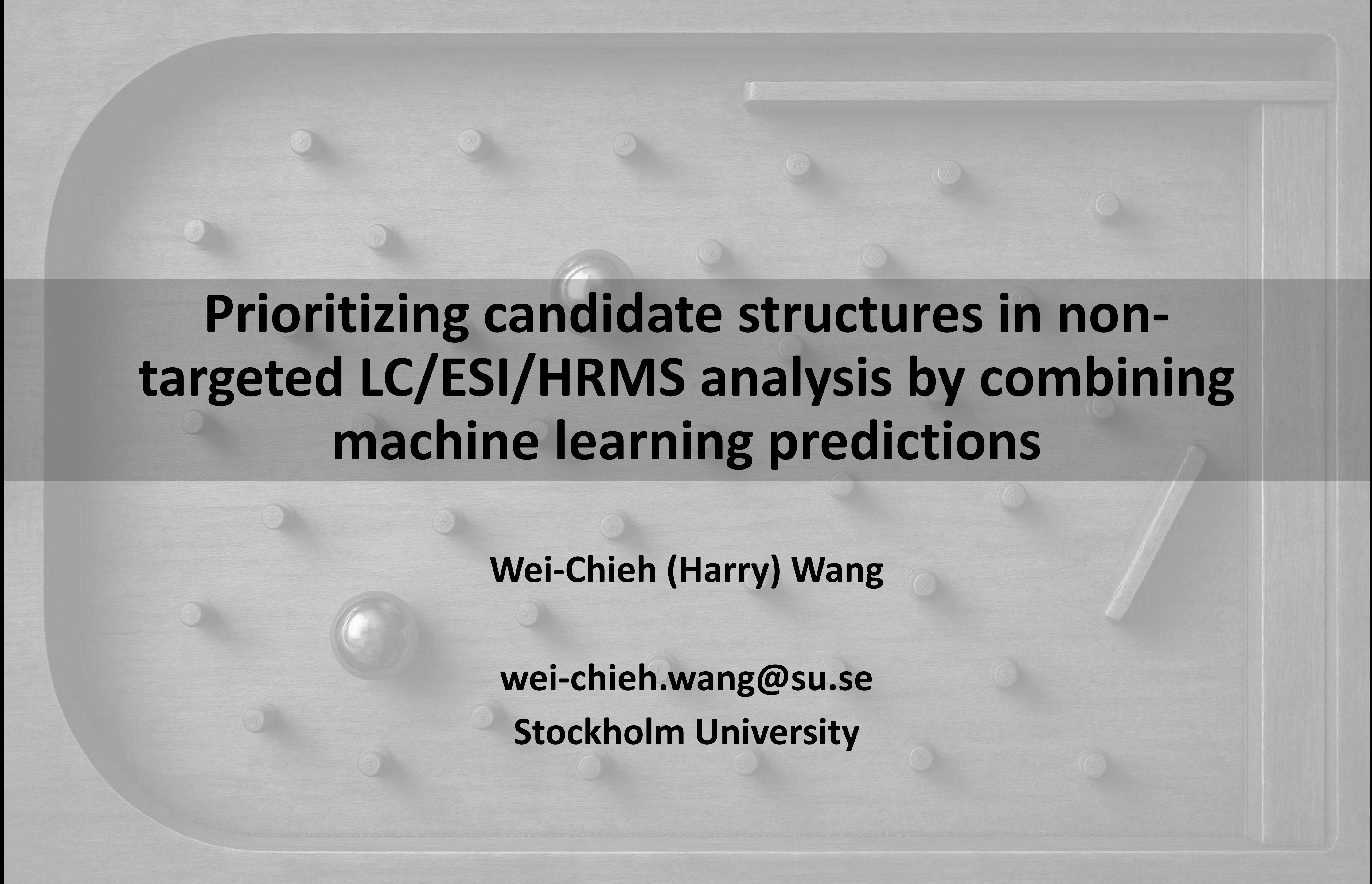


Prioritizing candidate structures in non-targeted LC/ESI/HRMS analysis by combining machine learning predictions

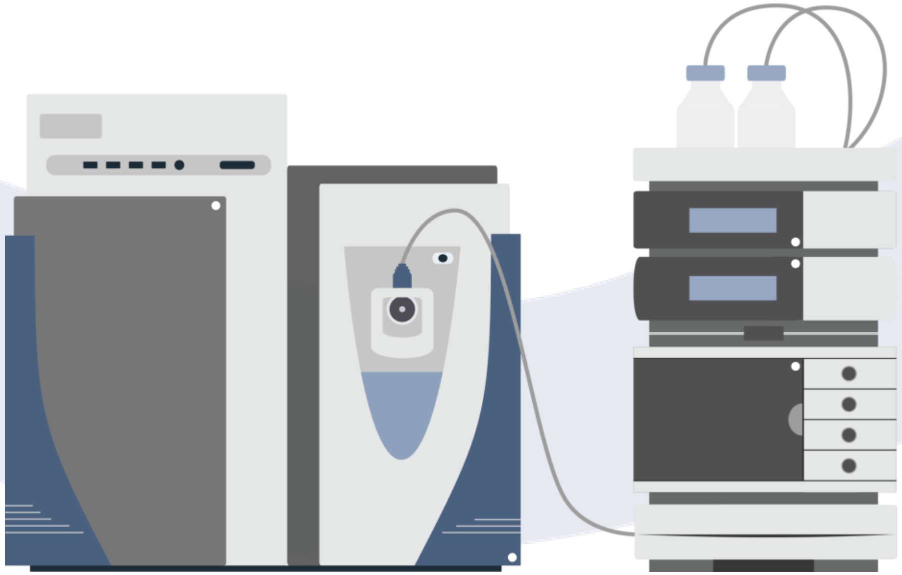
Wei-Chieh (Harry) Wang

wei-chieh.wang@su.se

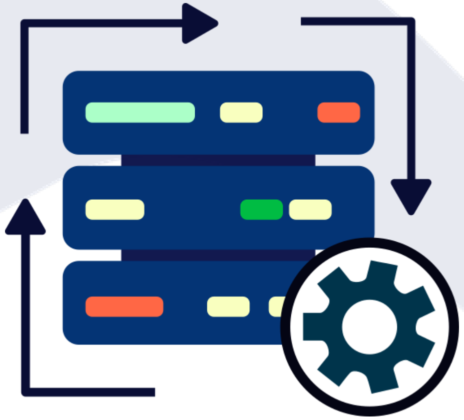
Stockholm University

Non-targeted screening (NTS)

Sample preparation

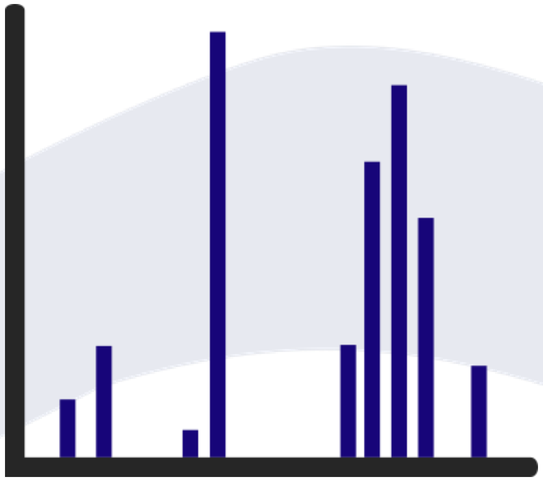


LC/ESI/IM/HRMS measurement

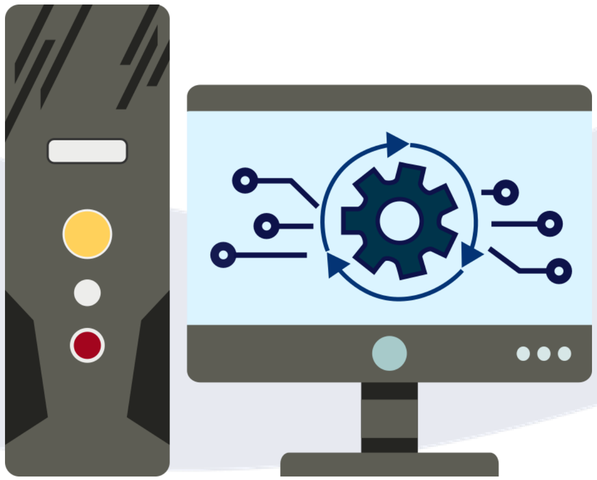


Data processing

Data processing



MS spectra



Annotation

Candidate lists

Candidate validation

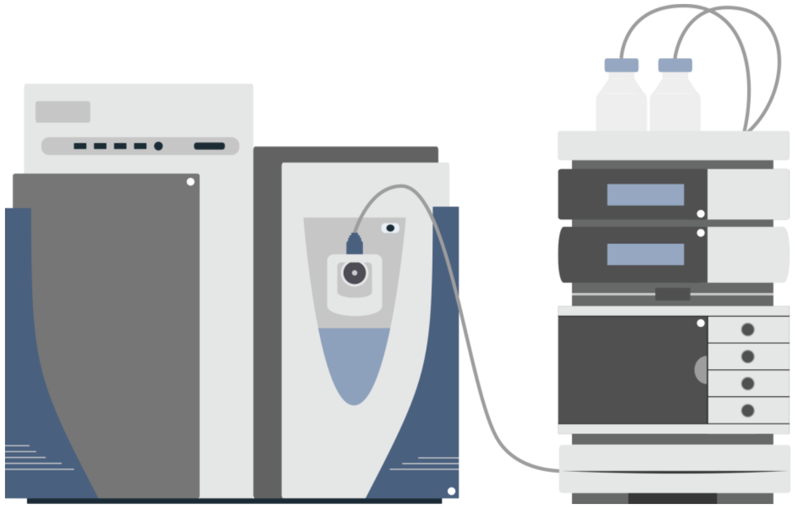
Candidate validation

Lack of chemical standards

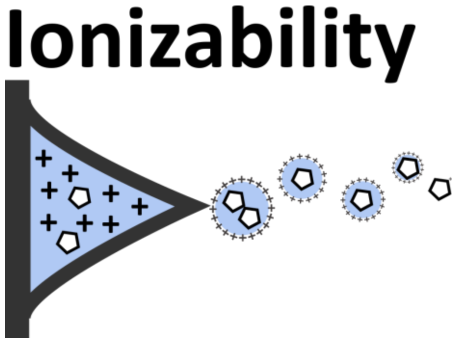
Costly

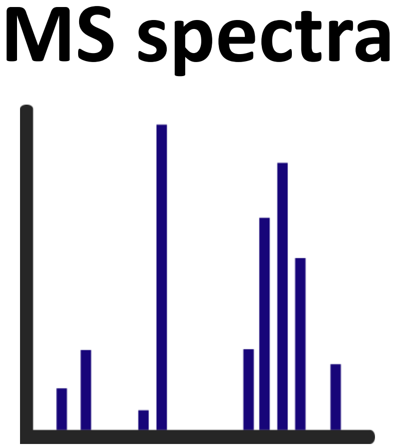
Time-consuming

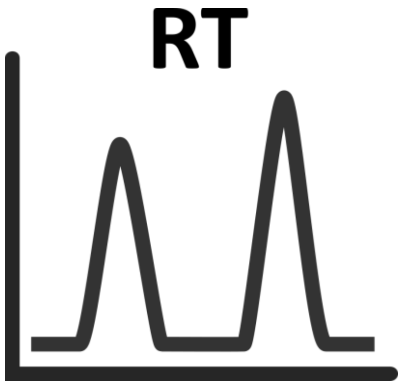
Information from measurements

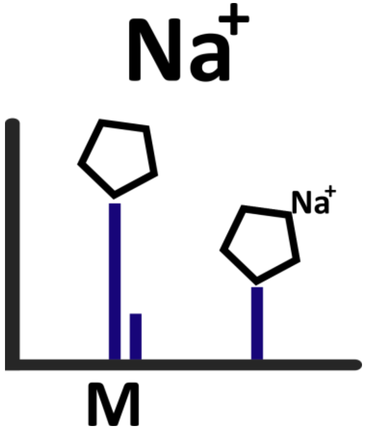


LC/ESI/IM/HRMS measurement



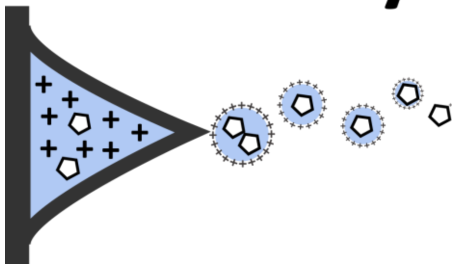


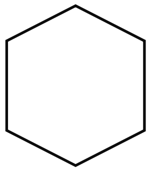




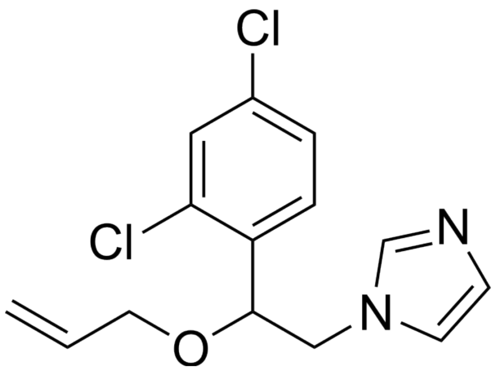
Prioritization

Ionizability

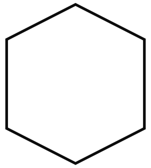




Cyclohexane



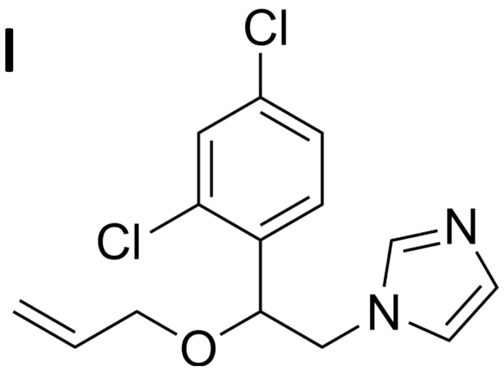
Imazalil



Cyclohexane

Properties of the candidates ✗

✓ **Observed experimental properties**

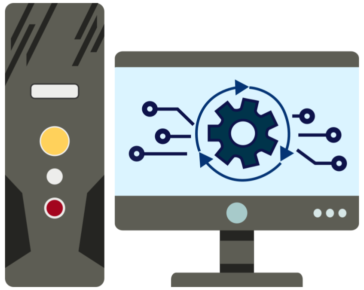


Imazalil

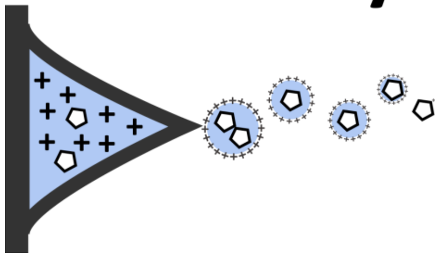
De-prioritized

Prioritized

Machine learning (ML) models



Ionizability

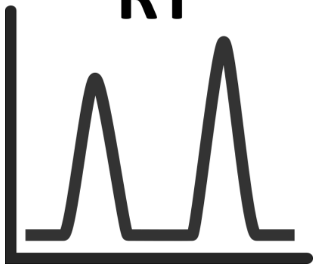


Binary behavior

Continuous predictions

Threshold

RT



Continuous values

Continuous predictions

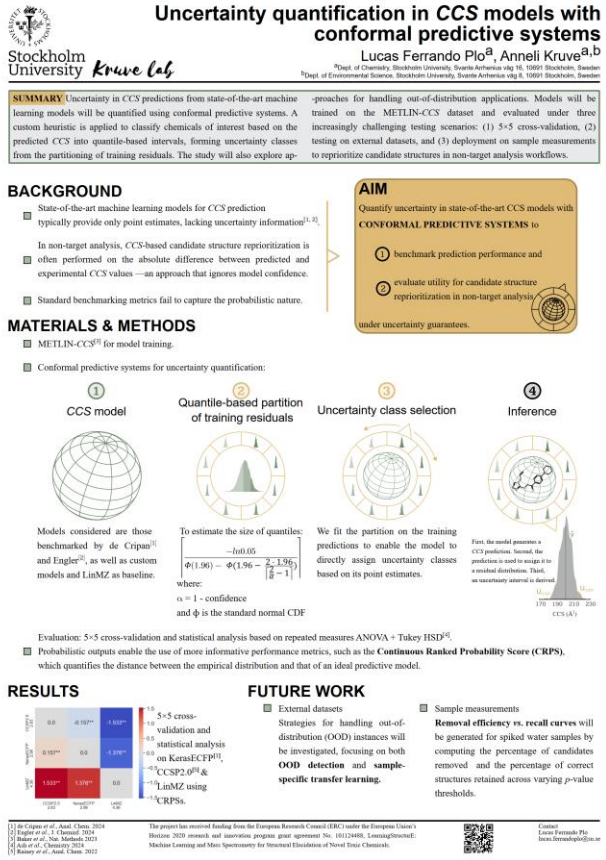
Error ranges (RMSE)

Uncertainty

- Model-based uncertainty
  - Model prediction errors (RMSE)

- Compound-based uncertainty
  - Conformal prediction system (CPS)

Poster: 3.11.P-Th187



Uncertainty quantification in *CCS* models with conformal predictive systems

Lucas Ferrando Plo[a], Anneli Kruve[a,b]
[a]Dept. of Chemistry, Stockholm University, Svante Arrhenius väg 16, 10691 Stockholm, Sweden
[b]Dept. of Environmental Science, Stockholm University, Svante Arrhenius väg 8, 10691 Stockholm, Sweden

SUMMARY Uncertainty in *CCS* predictions from state-of-the-art machine learning models will be quantified using conformal predictive systems. A custom heuristic is applied to classify chemicals of interest based on the predicted *CCS* into quantile-based intervals, forming uncertainty classes from the partitioning of training residuals. The study will also explore approaches for handling out-of-distribution applications. Models will be trained on the METLIN-*CCS* dataset and evaluated under three increasingly challenging testing scenarios: (1) 5×5 cross-validation, (2) testing on external datasets, and (3) deployment on sample measurements to reprioritize candidate structures in non-target analysis workflows.

BACKGROUND

- State-of-the-art machine learning models for *CCS* prediction typically provide only point estimates, lacking uncertainty information[1, 2].
- In non-target analysis, *CCS*-based candidate structure reprioritization is often performed on the absolute difference between predicted and experimental *CCS* values —an approach that ignores model confidence.
- Standard benchmarking metrics fail to capture the probabilistic nature.

AIM

Quantify uncertainty in state-of-the-art CCS models with **CONFORMAL PREDICTIVE SYSTEMS** to

① benchmark prediction performance and

② evaluate utility for candidate structure reprioritization in non-target analysis

under uncertainty guarantees.

MATERIALS & METHODS

- METLIN-*CCS*[3] for model training.
- Conformal predictive systems for uncertainty quantification:

① *CCS* model

Models considered are those benchmarked by de Cripan[1] and Engler[2], as well as custom models and LinMZ as baseline.

② Quantile-based partition of training residuals

To estimate the size of quantiles:

$$\left\lceil \frac{-ln0.05}{\Phi(1.96) - \Phi(1.96 - \frac{2\cdot 1.96}{\lceil \frac{2}{\alpha} - 1 \rceil})} \right\rceil$$

where:
α = 1 - confidence
and ϕ is the standard normal CDF

③ Uncertainty class selection

We fit the partition on the training predictions to enable the model to directly assign uncertainty classes based on its point estimates.

④ Inference

First, the model generates a CCS prediction. Second, the prediction is used to assign it to a residual distribution. Third, an uncertainty interval is derived.

Evaluation: 5×5 cross-validation and statistical analysis based on repeated measures ANOVA + Tukey HSD[4].

- Probabilistic outputs enable the use of more informative performance metrics, such as the **Continuous Ranked Probability Score (CRPS)**, which quantifies the distance between the empirical distribution and that of an ideal predictive model.

RESULTS

5×5 cross-validation and statistical analysis on KerasECFP[3], CCSP2.0[5] & LinMZ using CRPSs.

FUTURE WORK

- External datasets
  Strategies for handling out-of-distribution (OOD) instances will be investigated, focusing on both **OOD detection** and **sample-specific transfer learning.**
- Sample measurements
  Removal efficiency *vs.* recall curves will be generated for spiked water samples by computing the percentage of candidates removed and the percentage of correct structures retained across varying *p*-value thresholds.

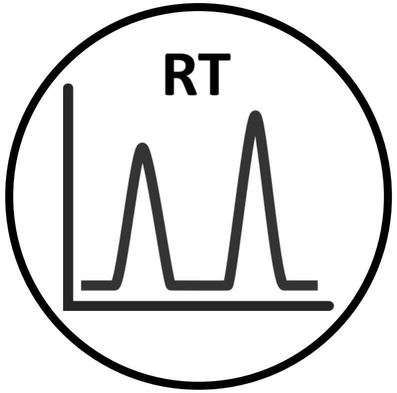
[1] de Cripan *et al.*, Anal. Chem. 2024
[2] Engler *et al.*, J. Cheminf. 2024
[3] Baker *et al.*, Nat. Methods 2023
[4] Ash *et al.*, Chemistry 2024
[5] Rainey *et al.*, Anal. Chem. 2022

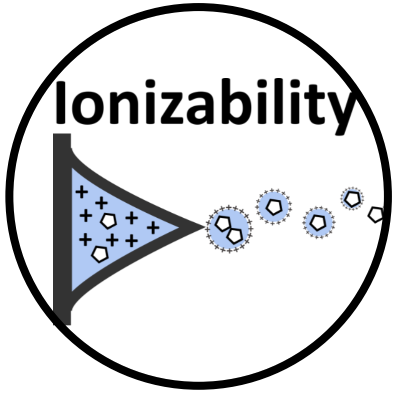
The project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement No. 101124488, LearningStructurE: Machine Learning and Mass Spectrometry for Structural Elucidation of Novel Toxic Chemicals.

Contact
Lucas Ferrando Plo
lucas.ferrandoplo@su.se

G. Shafer, V. Vovk. J. Machine Learning Research 9, 371-421 (2008).

ML-supported prioritization



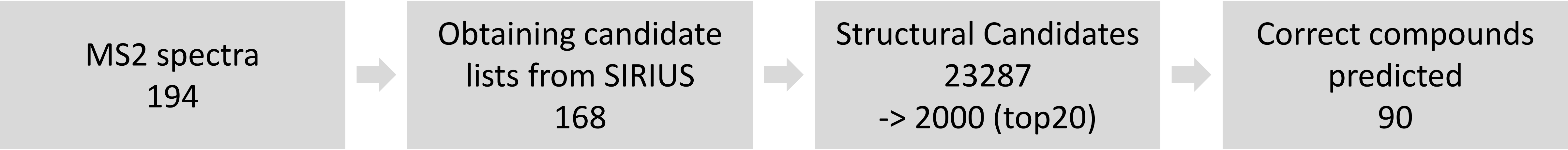


$$Accuracy = \frac{Number\ of\ true\ positive}{Numbre\ of\ the\ LC/HRMS\ features}$$

$$Efficiency = \frac{Number\ of\ the\ eliminated\ candidates}{Number\ of\ the\ total\ candidates}$$

A. Souihi, M.P. Mohai, E. Palm et al. J. Chroma 1666, 462867 (2022).

Annotation performance from SIRIUS

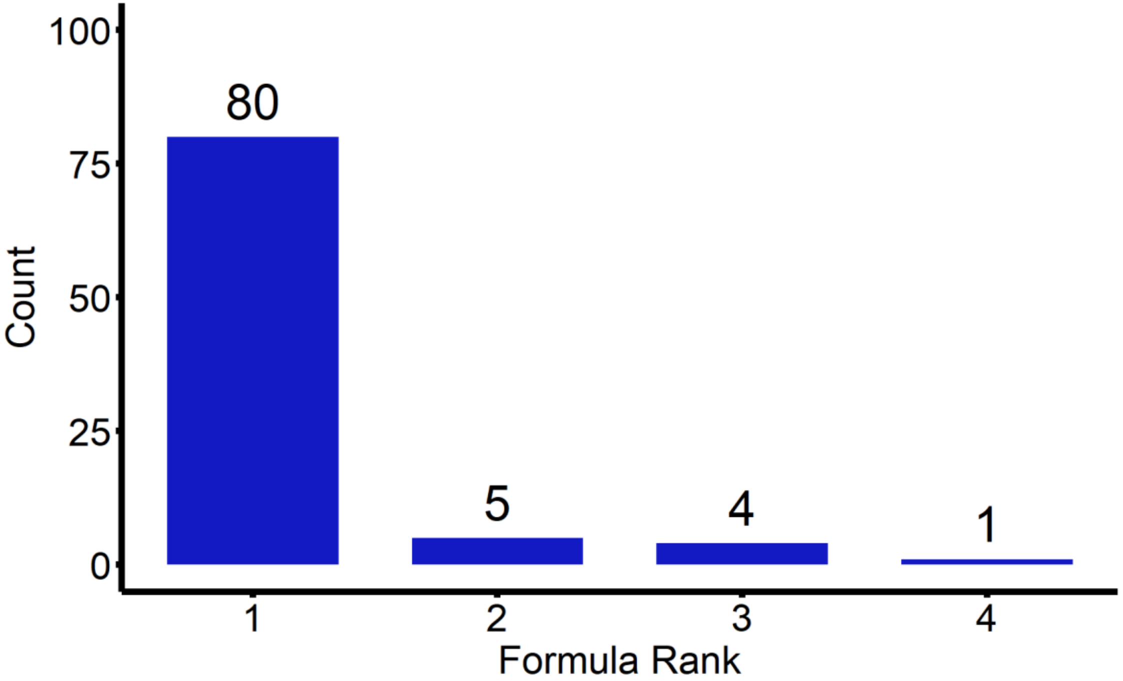


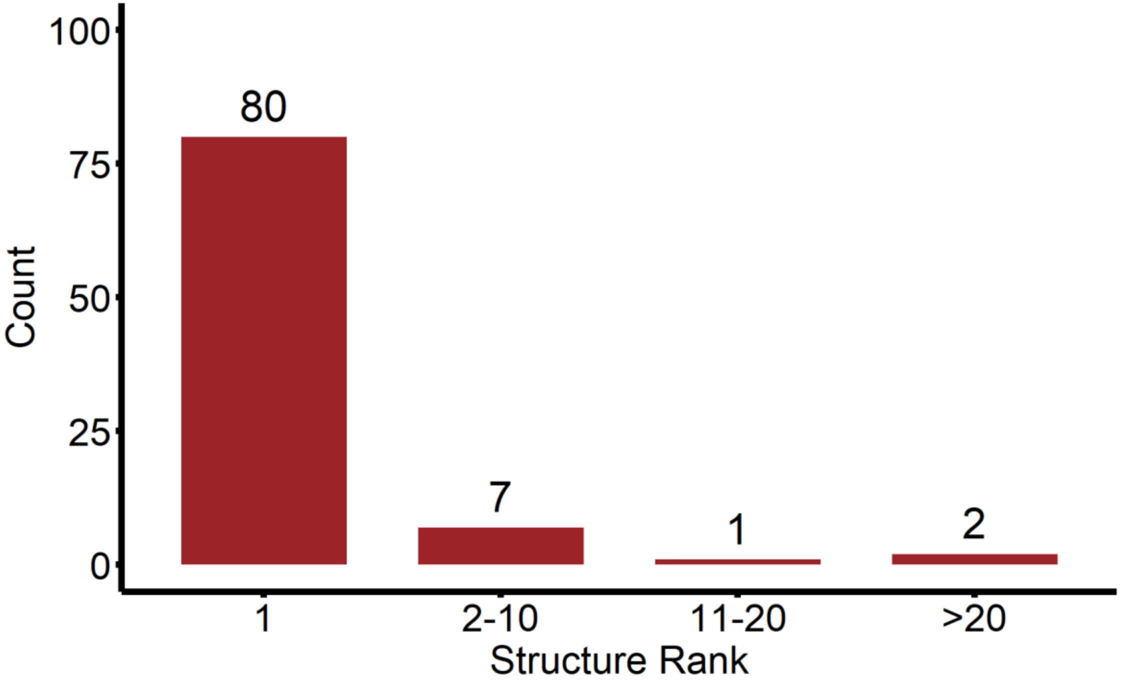
MS2 spectra
194

Obtaining candidate lists from SIRIUS
168

Structural Candidates
23287
-> 2000 (top20)

Correct compounds predicted
90





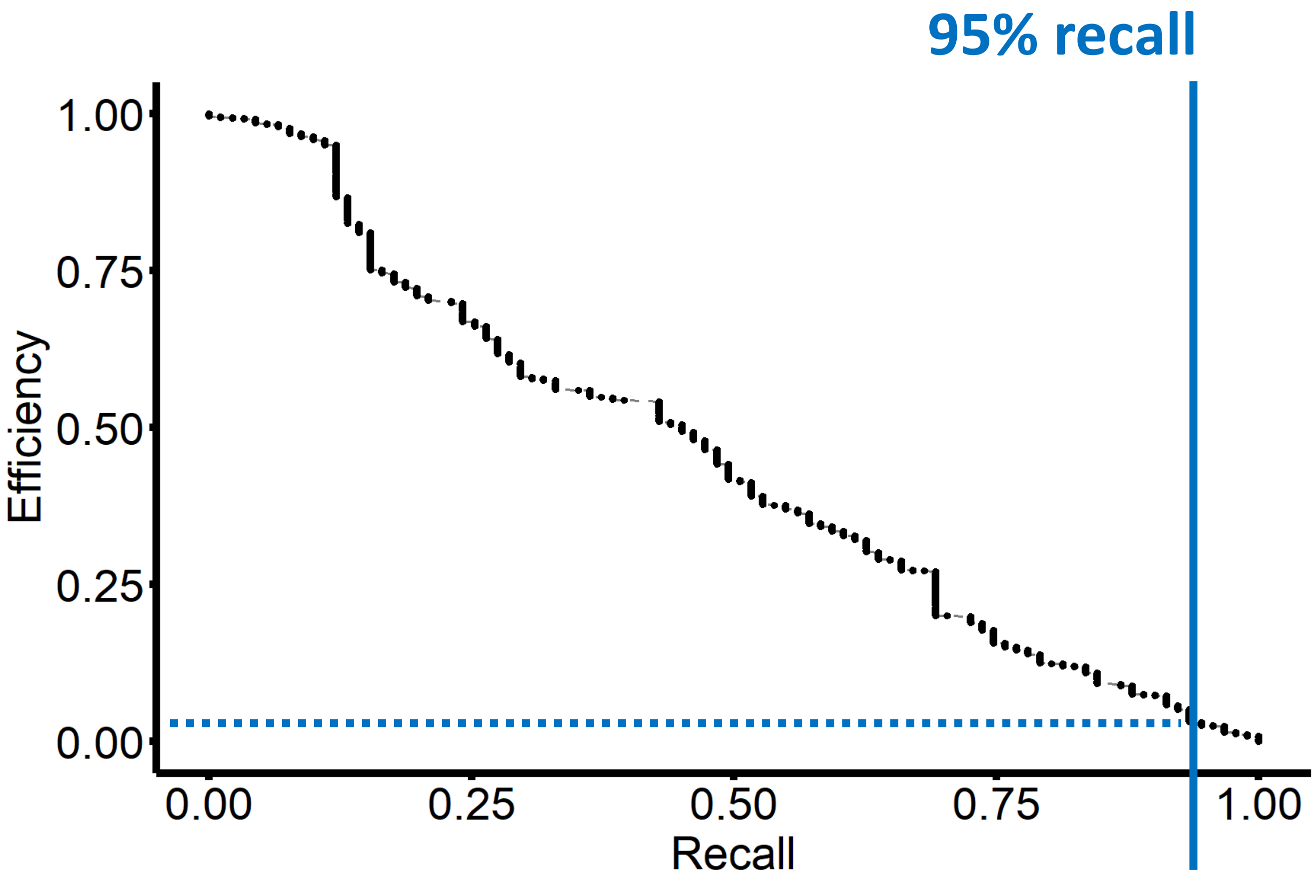
Dührkop, K., Fleischauer, M., Ludwig, M. et al. Nat Methods 16, 299–302 (2019).

Results for prioritization

- RT prediction:
  - Efficiency: 84.02%
  - Accuracy: 13.64%

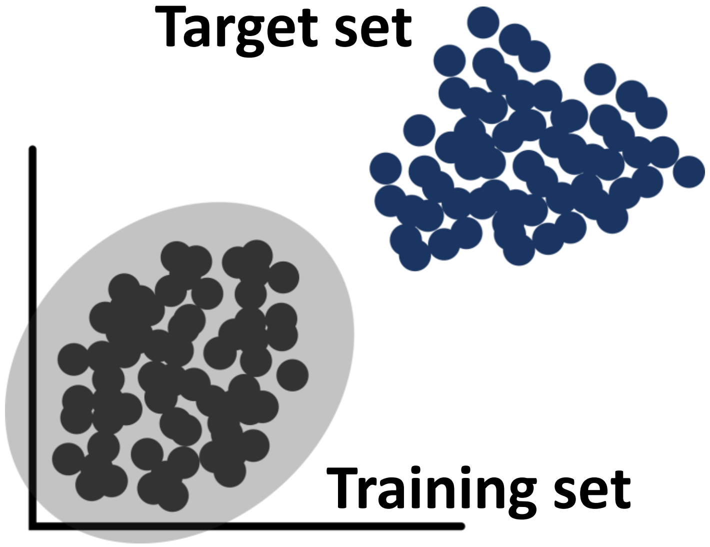
- Ionizability prediction
  - Recall-Efficiency curve

- Combining two models
  - Efficiency: 99.55%
  - Accuracy: 5.36 %



Current challenges for combining predictions from various ML models

- Different application domains
  - The model was trained in different chemical spaces.



- No compound-based uncertainty available

Conclusions & Future perspectives

- A strict combination of machine learning models led to an undesired removal of true positives.

- Incorporate additional machine learning prediction models.
- Retrain models using data from the same chemical space.
- Estimate compound-based uncertainty using a conformal prediction system.

Acknowledgement

Kruve lab

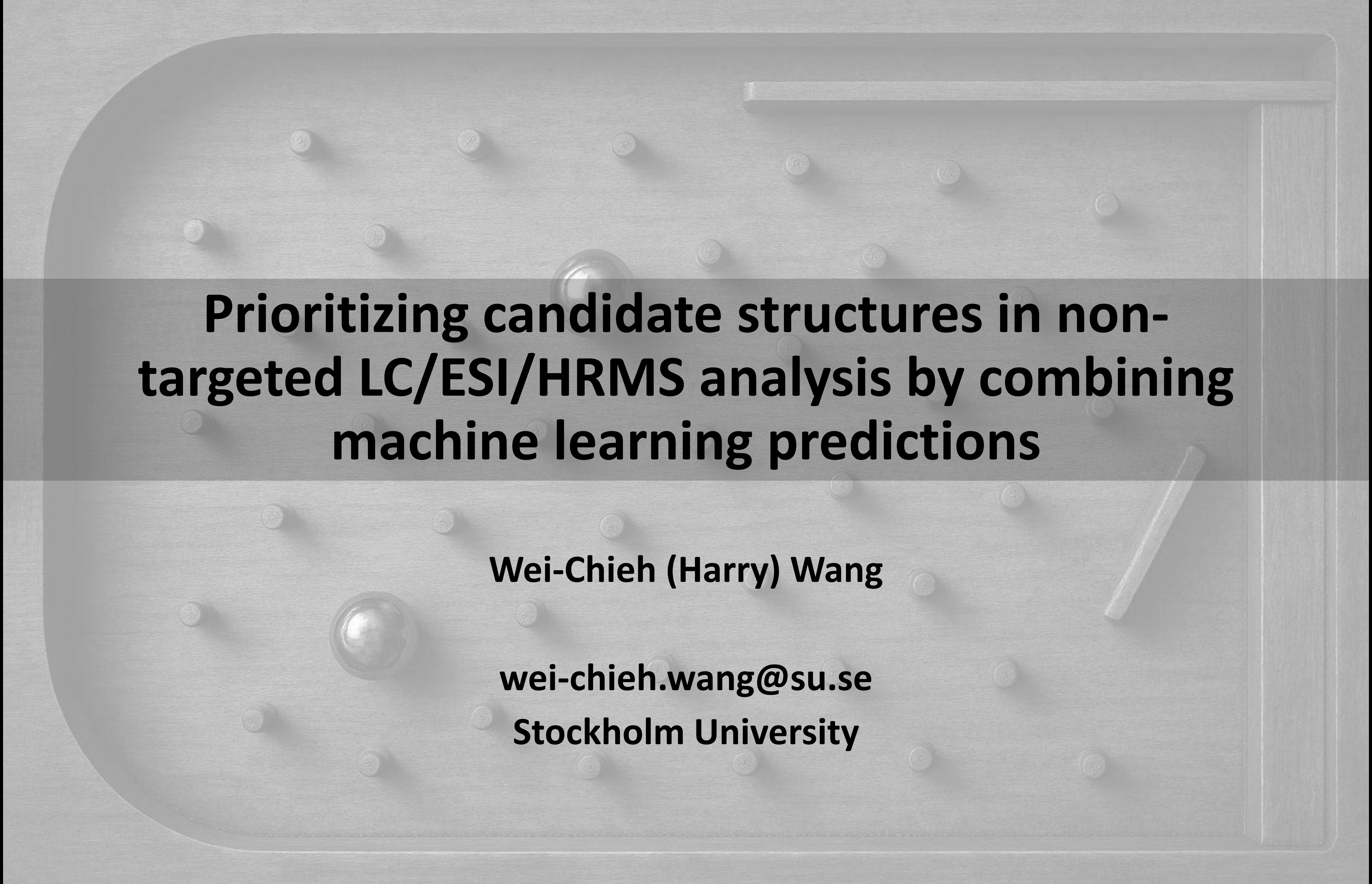
Stockholm University

erc

European Research Council

Established by the European Commission

Swedish Research Council



Prioritizing candidate structures in non-targeted LC/ESI/HRMS analysis by combining machine learning predictions

Wei-Chieh (Harry) Wang

wei-chieh.wang@su.se

Stockholm University