MSNovelist: de novo structure generation from mass spectra

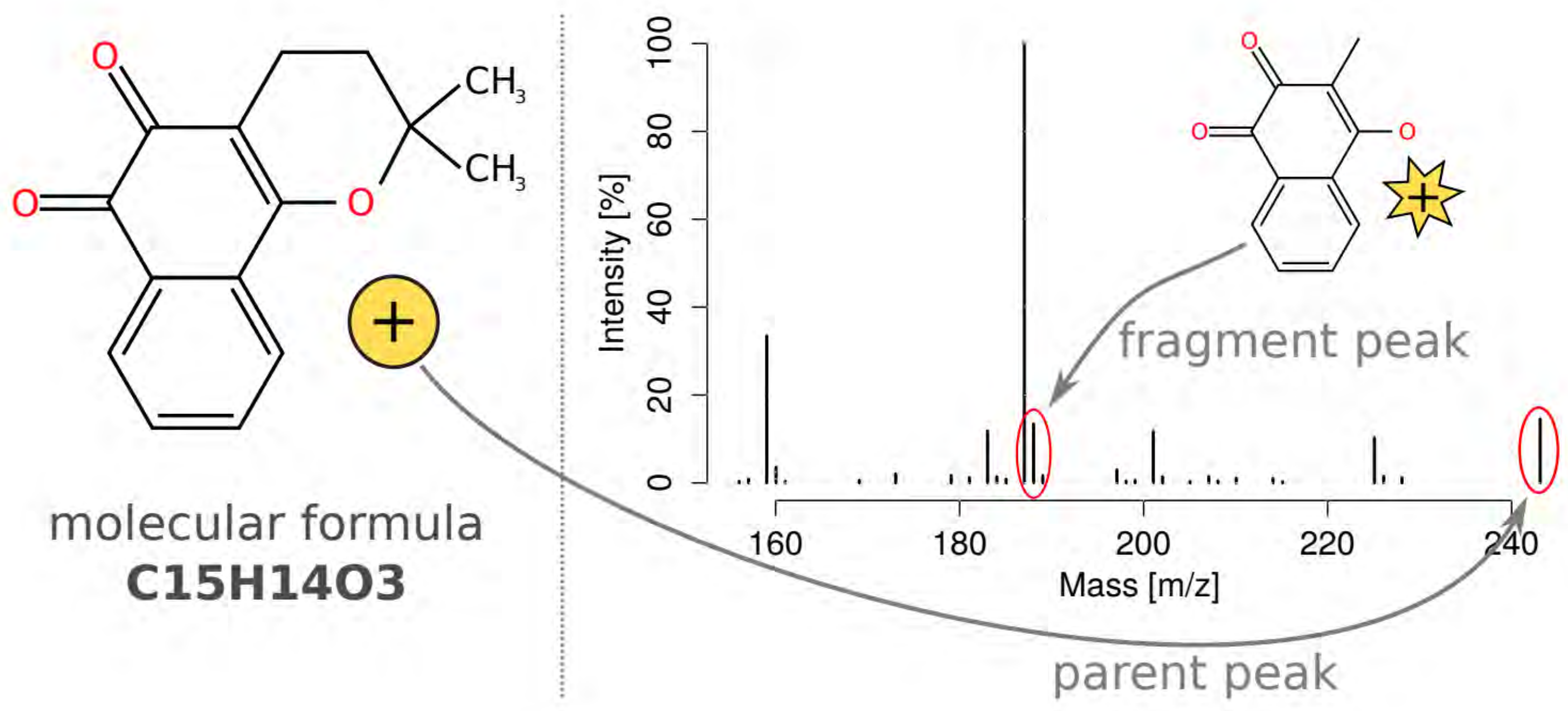
Michael A. Stravs [1,3] Kai Dührkop [2] Jonas Emmert [2] Sebastian Böcker [2] Nicola Zamboni [1]

[1]Institute of Molecular Systems Biology, Department of Biology, ETH Zürich, Zürich, Switzerland. [2]Chair for Bioinformatics, Faculty of Mathematics and Computer Science, Friedrich-Schiller-Universität Jena, Jena, Germany. [3]Present address: Eawag, Dübendorf, Switzerland.

Introduction

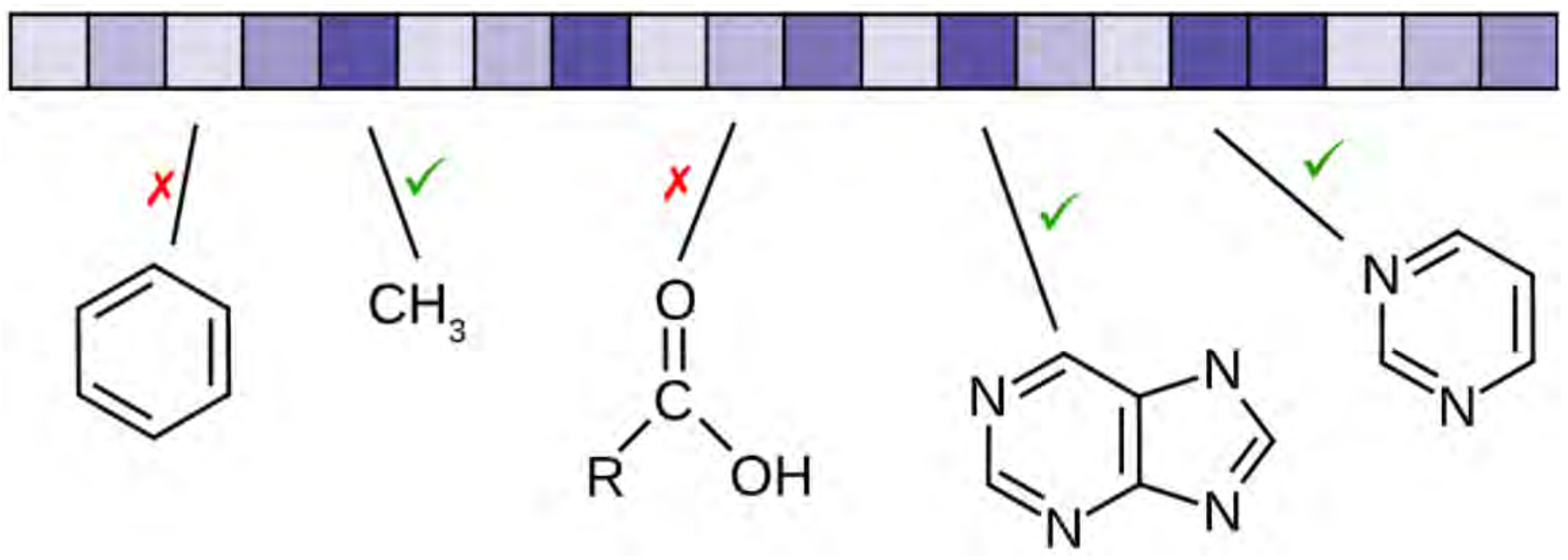
Mass spectrometry is vital in metabolomics and non-targeted analysis, providing valuable information about unknown compounds. However, current methods for structure annotation like database search have limitations in identification and representation of novel compounds and underrepresented analyte classes. To address this, MSNovelist [5] presents a novel approach that directly generates candidate structures from MS/MS spectra, bypassing the need for databases. This is accomplished by using SIRIUS and CSI:FingerID [2] results (molecular formula and fingerprint, respectively) as input to an RNN generative model.

Mass spectral data



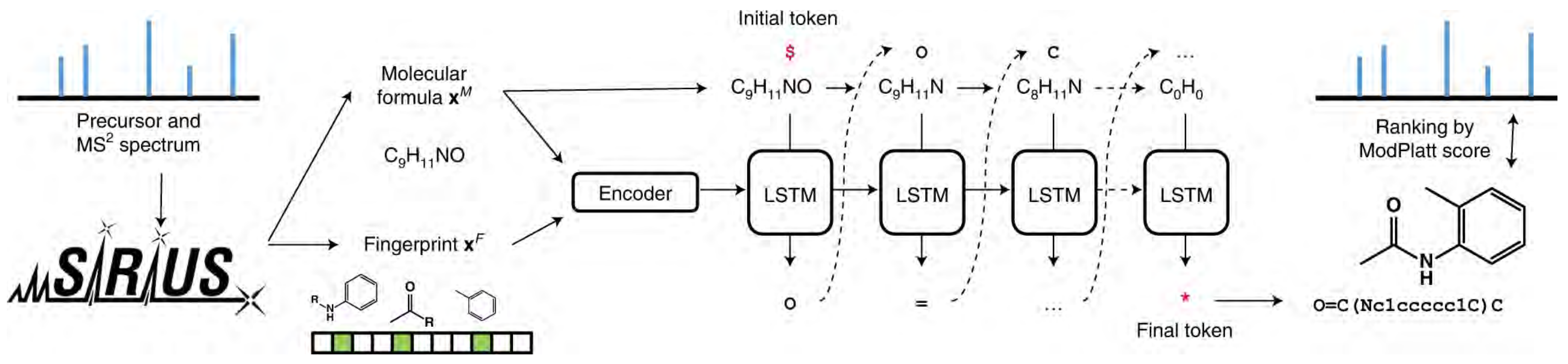
Mass spectral data represents mass to charge ratios (m/z) of molecular fragments together with their respective intensity. Figure from [1].

Molecular fingerprints



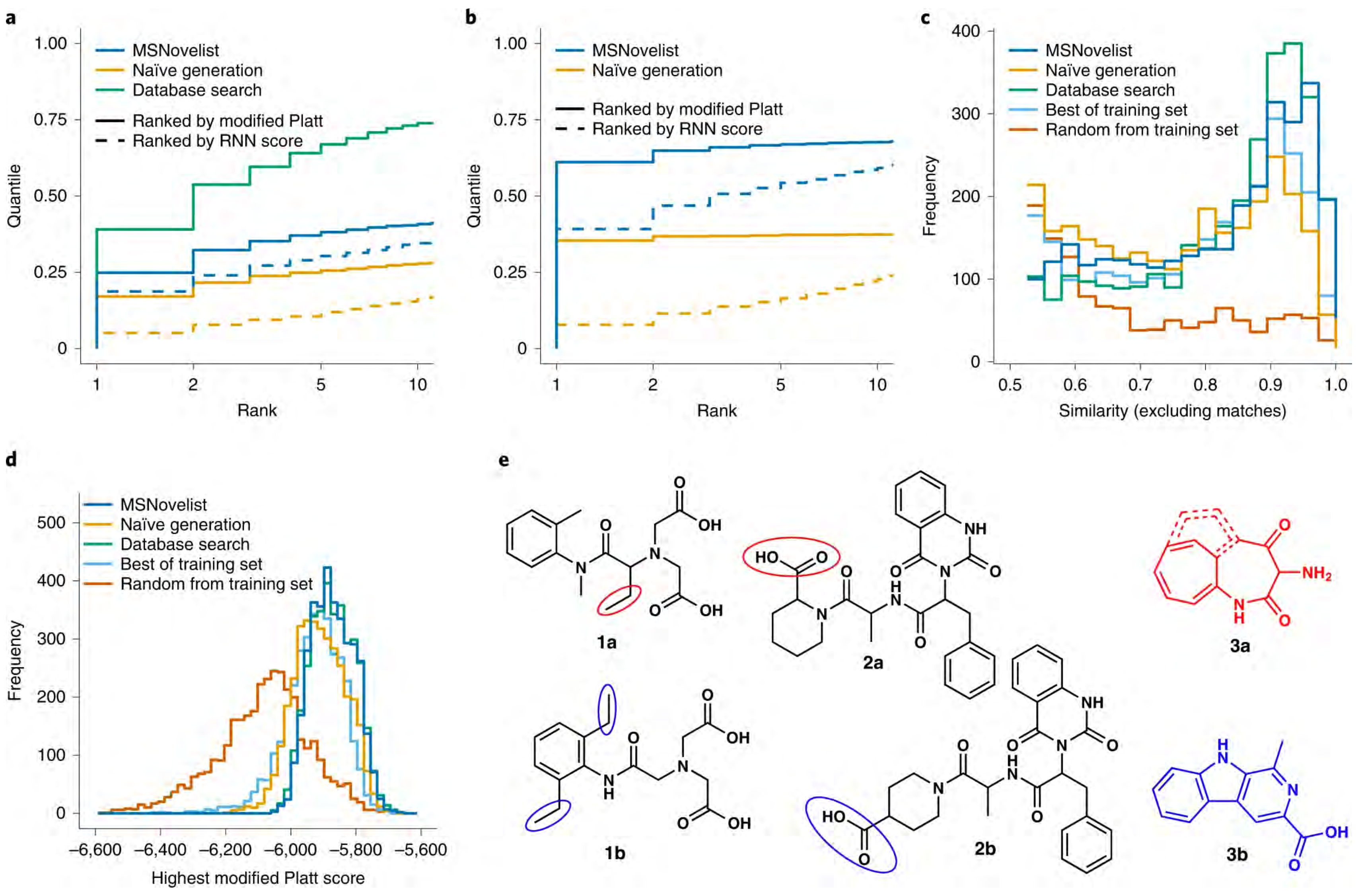
Molecular fingerprints from CSI:FingerID are represented by a vector of thousands of values ranging from 0 to 1, indicating the likelihood of specific structural characteristics. Figure from [3].

Structure generating network



The predicted molecular formula from SIRIUS and the predicted molecular fingerprint from CSI:FingerID are used as input to an encoder-decoder model. The decoding LSTM [4] generates SMILES sequences that can be translated to structures. After exclusion of invalid and duplicate structures, resulting candidates are re-ranked based on the modified Platt score, measuring their match to the query fingerprint.

Model evaluation for GNPS dataset



a Rank of correct structure in results for the GNPS dataset (n = 3863).

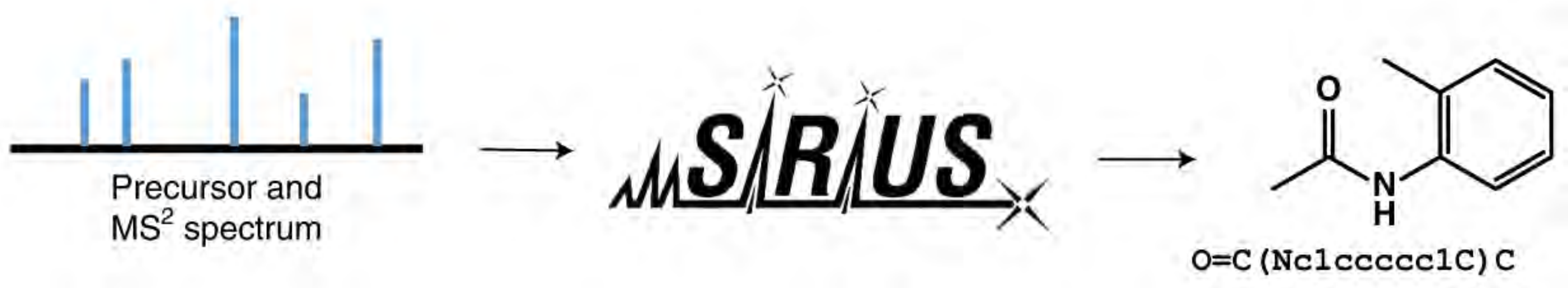
b Rank of correct structure in results for GNPS-OK dataset, containing only correct structural annotations from database search (n = 1507).

c Tanimoto similarity of best incorrect candidate to correct structure, measuring the proportion of common fingerprint bits.

d Modified Platt score of top candidates.

e Three randomly chosen examples of incorrect predictions. Structures 1a, 2a and 3a represent de novo prediction; structures 1b, 2b and 3b represent a correct result. Red marks sites predicted incorrectly by the model (or the entire molecule if the prediction was completely wrong), and blue marks the corresponding correct alternative.

Ongoing work



MSNovelist is currently being integrated into SIRIUS, making the method easier to use as well as allowing computation on external resources. Simultaneously, the model is re-trained on new and improved fingerprints from CSI:FingerID, enabling prediction for both positive and negative mode spectral data. In combination with having more training data at hand, a boost in performance would be expected.

References

[1] Céline Brouard, Eric Bach, Sebastian Böcker, and Juho Rousu. Magnitude-preserving ranking for structured outputs. In *Asian Conference on Machine Learning*, pages 407–422. PMLR, 2017.

[2] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A Aksenov, Alexey V Melnik, Marvin Meusel, Pieter C Dorrestein, Juho Rousu, and Sebastian Böcker. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature methods*, 16(4):299–302, 2019.

[3] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Martin A. Hoffmann, Fleming Kretschmer, and Sebastian Böcker. SIRIUS 6 - MS/MS-centric untargeted metabolite structure annotation. ASMS Poster, 2023.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] Michael A. Stravs, Kai Dührkop, Sebastian Böcker and Nicola Zamboni. MSNovelist: de novo structure generation from mass spectra. *Nature Methods*, 19(7):865–870, 2022.