Code available

Check out the paper

Contact us

Probabilistic framework for integration of mass spectrum and retention time information in small molecule identification

Eric Bach [1,✉], Simon Rogers [2], John Williamson [2], and Juho Rousu [1]

[1] Helsinki institute for Information Technology (HIIT), Department of Computer Science, Aalto University, Espoo, Finland
[2] School of Computing Science, University of Glasgow, Glasgow, UK

1. Small Molecule Identification in Untargeted Metabolomics

- Challenge in untargeted metabolomics studies: **Identification of the small molecules** present in a biological sample
- **LC-MS^2** widely used analysis platform: Liquid chromatography (LC) coupled with tandem mass spectrometry (MS^2) (Fig. 1)
- Most machine learning approaches for small molecule identification only utilize MS^2 information [3, 2]
- LC retention times (RT) can improve the small molecule annotation [5, 7]
- **Challenges utilizing RT information:** (1) LC-system specific RT measurements and (2) public RT databases are limited in size and coverage

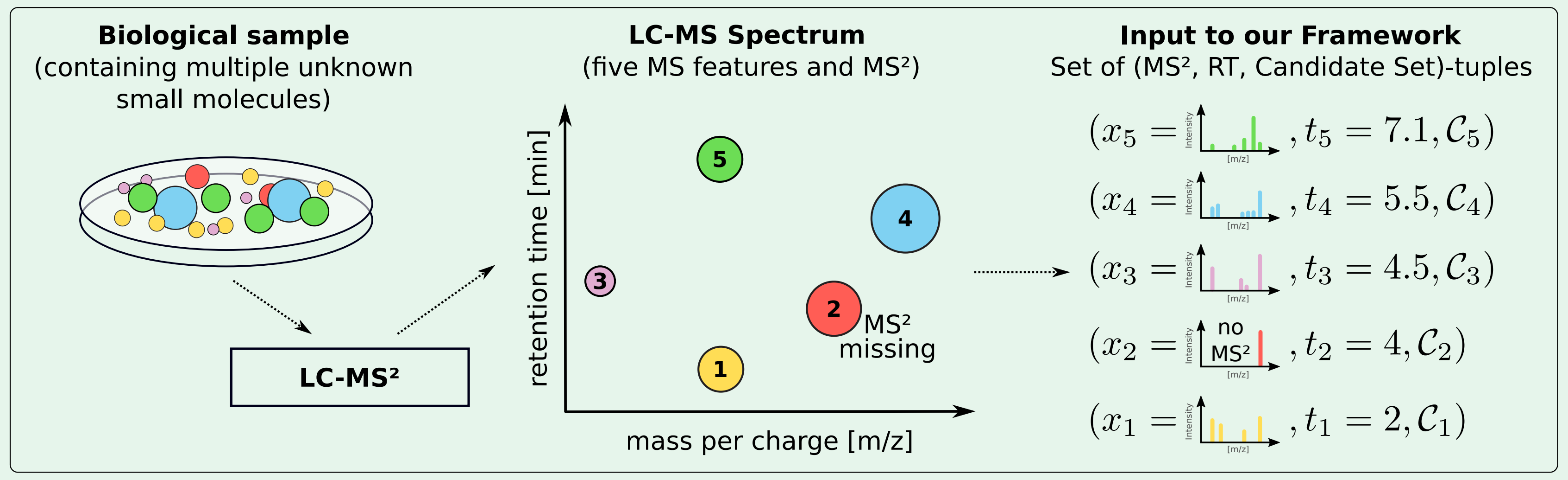

Biological sample
(containing multiple unknown small molecules)
LC-MS²
LC-MS Spectrum
(five MS features and MS²)
retention time [min]
mass per charge [m/z]
5
4
3
2
1
MS² missing
Input to our Framework
Set of (MS², RT, Candidate Set)-tuples
$(x_5 =$, $t_5 = 7.1, \mathcal{C}_5)$
$(x_4 =$, $t_4 = 5.5, \mathcal{C}_4)$
$(x_3 =$, $t_3 = 4.5, \mathcal{C}_3)$
$(x_2 =$ no MS², $t_2 = 4, \mathcal{C}_2)$
$(x_1 =$, $t_1 = 2, \mathcal{C}_1)$
Intensity
[m/z]

Fig. 1: LC-MS^2 analysis pipeline and resulting data used as input for our framework.

2. Retention Time (RT) Utilization for Small Molecule Identification

- Different approaches proposed in the literature
- Multiple approaches to utilize RT for molecule annotation exist
- (utilization of RT information, scalable, cross laboratories (LC-systems), RT reference free)

1) Compare measured RTs with in-house reference RTs ✓, ✗, ✗, ✗
2) Compare measured RTs with projected reference RTs ✓, ✗, ●, ✗
3) Compare measured RTs with predicted RTs ✓, ✓, ●, ●
4) Compare measured RTs with predicted RTs proxies, e.g. LogP ✓, ✓, ✓, ✗
5) Compare measured retention orders with predicted ones ●, ✓, ✓, ✓

- Fully supported: ✓, Partially supported: ●, Not supported: ✗
- RT comparison to prune candidate lists or (re)ranking **[CITATION]**

3. LC-MS^2 Experiment Data: Input and Output of our Framework

- **Input:** Preprocessed LC-MS^2 data, i.e. after peak-picking and alignment (Fig.1):

$$\mathcal{D} = \{(x_i, t_i, \mathcal{C}_i)\}_{i=1}^{N}$$

x_i : MS Information; MS^2 or MS^1 (precursor m/z), if no fragmentation available
t_i : Measured RT
$\mathcal{C}_i$: Molecular candidate sets, e.g. molecular structures found in PubChem by exact mass search
N : Number of MS features

- **Precomputed MS scoring assumed:** MS^1 deviation of candidate and precursor mass or MS^2 scores, e.g. by CSI:FingerID [3], MetFrag [5] or IOKR [2]
- **Output:** Ranking of the molecular candidates in $m_{ir} \in \mathcal{C}_i$ for each MS feature i
- Ranking based on MS and RT information

4. Probabilistic Framework to integrate MS and RT Information

- **Graphical model** G superimposed on the LC-MS^2 data (Fig. 2)
- Let $G = (V, E)$ be complete graph with a **node** $i \in V$ for each MS feature, and an **edge** $(i, j) \in E$ for each feature pair
- Discrete random variable $z_i \in \mathcal{Z}_i = \{1, \dots, n_i\}$ associated with each node ($n_i = |\mathcal{C}_i|$)
- Candidate annotation for the complete data $\mathbf{z} = \{z_i \,|\, i \in V\} \in \mathcal{Z}_1 \times \dots \times \mathcal{Z}_N = \mathcal{Z}$
- Intuitively: Random variable z_i denotes the candidate $m_{ir} \in \mathcal{C}_i$ assigned to feature i.
- Pairwise **Markov Random Field** as probabilistic model [4]:

$$p(\mathbf{z}) = \frac{1}{Z} \prod_{i \in V} \psi_i(z_i) \prod_{(i,j) \in E} \psi_{ij}(z_i, z_j)$$

- Potential functions: $\psi_i(z_i)$ MS score and $\psi_{ij}(z_i, z_i)$ match of observed and **predicted retention order**
- Molecular **candidates ranked** based on max-marginals [4] (Fig. 2):

$$p_{\max}(z_i = r) = \max_{\{\mathbf{z}' \in \mathcal{Z} \,|\, z'_i = r\}} p(\mathbf{z}')$$

- Intuitively: Maximum marginal probability of a candidate assignment with $z_i = r$.

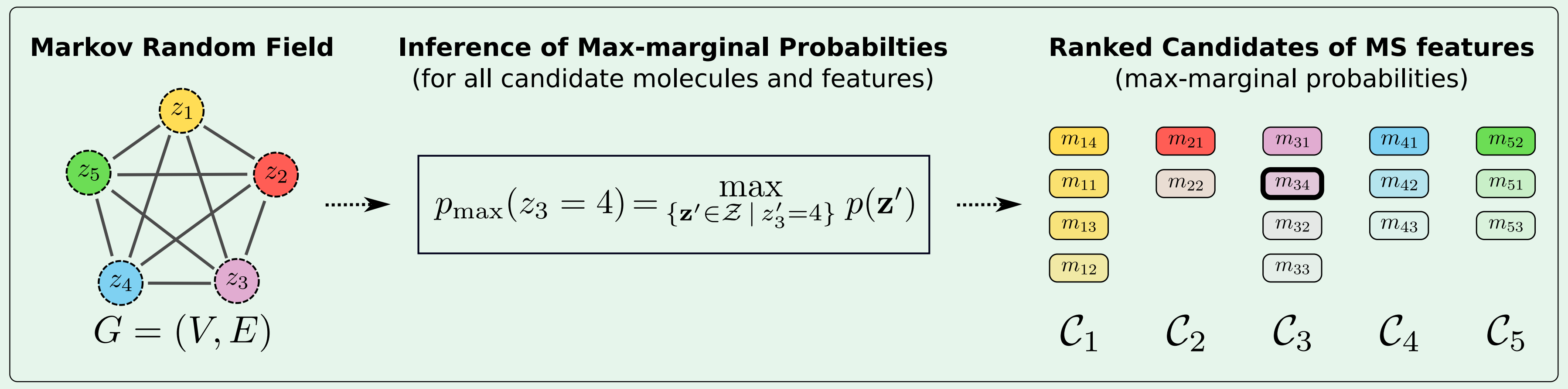

Markov Random Field
z_1
z_2
z_3
z_4
z_5
$G = (V, E)$
Inference of Max-marginal Probabilties
(for all candidate molecules and features)
$p_{\max}(z_3 = 4) = \max_{\{\mathbf{z}' \in \mathcal{Z} \,|\, z'_3 = 4\}} p(\mathbf{z}')$
Ranked Candidates of MS features
(max-marginal probabilities)
m_{14} m_{11} m_{13} m_{12}
m_{21} m_{22}
m_{31} m_{34} m_{32} m_{33}
m_{41} m_{42} m_{43}
m_{52} m_{51} m_{53}
$\mathcal{C}_1$ $\mathcal{C}_2$ $\mathcal{C}_3$ $\mathcal{C}_4$ $\mathcal{C}_5$

Fig. 2: MRF probability distribution and candidate ranking, e.g. MS feature $i = 3$ and candidate 4 (m_{34}).

5. Encoding MS and Retention Order Information: ψ_i and ψ_{ij}

- Node potential $\psi_i : \mathcal{Z}_i \rightarrow \mathbb{R}_{>0}$: $\psi_i(z_i = r) = f(x_i, m_{ir})$
- f returns the MS matching score $\in (0, 1]$ of spectrum x_i and candidate m_{ir}
- Edge potential $\psi_{ij} : \mathcal{Z}_i \times \mathcal{Z}_j \rightarrow \mathbb{R}_{>0}$, with σ being the sigmoid function:

$$\psi_{ij}(z_i = r, z_j = s) = \sigma(\underbrace{\text{sign}(t_i - t_j)}_{\text{observed retention order}} \cdot \underbrace{\langle \mathbf{w}, \phi(m_{ir}) - \phi(m_{js}) \rangle}_{\text{predicted retention order}})$$

- Intuitively: Matching observed and predicted retention orders receive high scores.
- **Retention order prediction** using Ranking Support Vector Machine (RankSVM) $\mathbf{w}$ [1]
- Candidate molecules m_{ir} representation using non-linear features ϕ

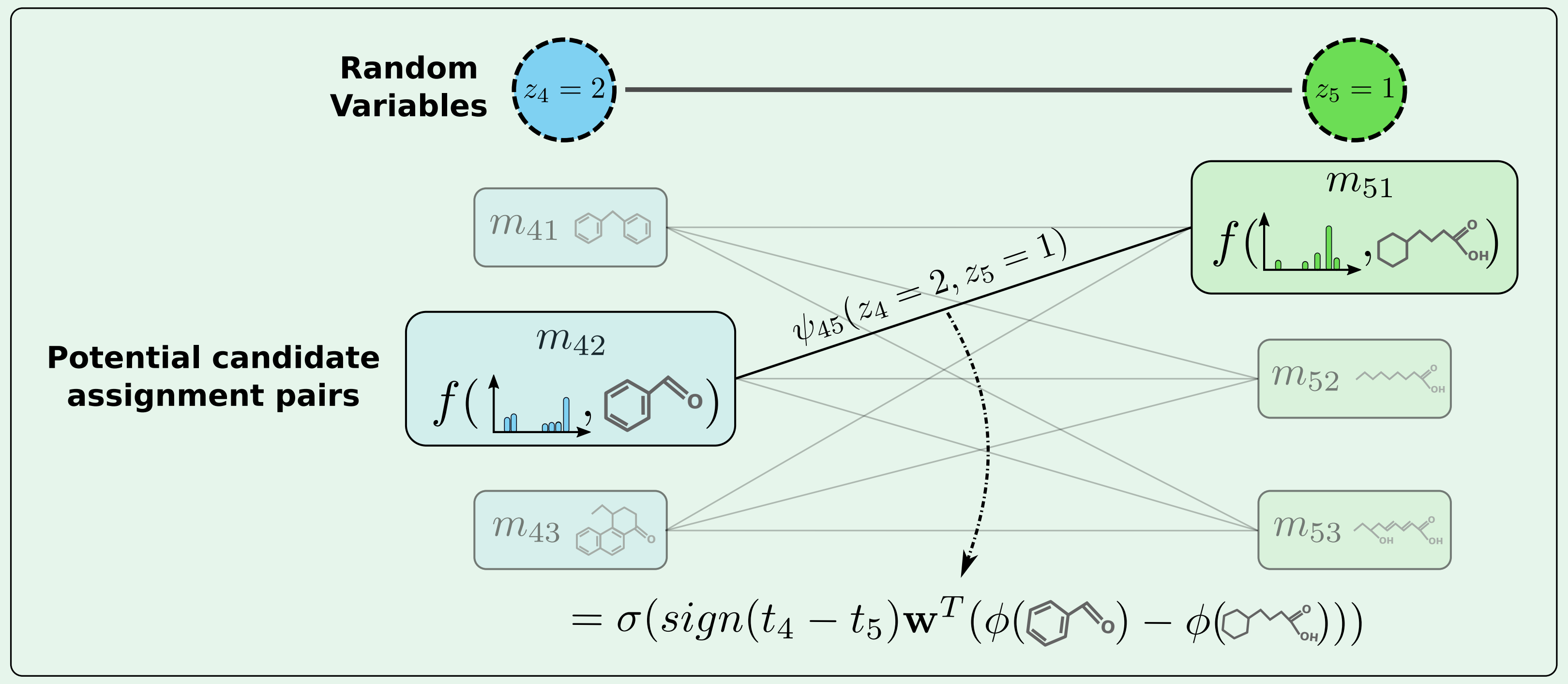

Random Variables
$z_4 = 2$
$z_5 = 1$
Potential candidate assignment pairs
m_{41}
m_{42}
$f($, $)$
m_{43}
m_{51}
$f($, $)$
m_{52}
m_{53}
$\psi_{45}(z_4 = 2, z_5 = 1)$
$= \sigma(sign(t_4 - t_5)\mathbf{w}^T(\phi($ $) - \phi($ $)))$

Fig. 3: Example: Node and edge score for all candidate pairs of feature $i = 4$ and $j = 5$.

6. Experiments and Results

- **Evaluation datasets:** CASMI 2016 [6], EA subset from MassBank used by [5]
  - 681 (MS^2, RT)-tuples with median number of candidates between 120 and 919
- **RankSVM training data:** 1248 RTs from PredRed [7] and CASMI 2016 training
  - No evaluation set molecule in RankSVM tranining set
- **Performance measure:** Top-k accuracy, percentage of correct molecular candidates at rank $\leq k$

Experiment 1: Comparison to MetFrag + LogP Prediction

- MetFrag relaunched [5]: Prediction of LogP values for candidates, linear model mapping measured RTs to LogPs, candidate re-ranking based on LogP deviation

Method	Top-1	Top-5	Top-10	Top-20
MS^2 + RT (Our)	21.3	52.9	64.0	74.3
MS^2 + RT (MetFrag & LogP)	20.5	49.1	61.2	72.6
Only MS^2 (baseline)	16.7	49.5	60.4	70.6

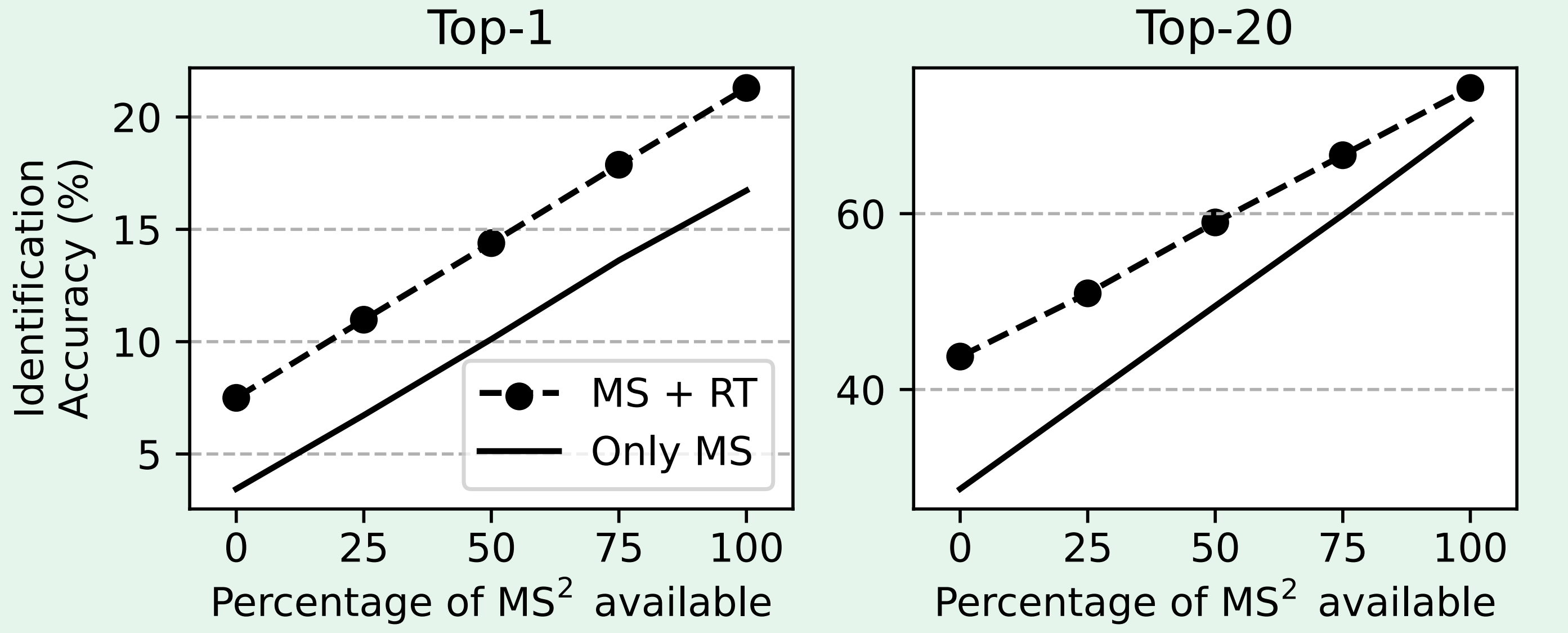
Experiment 2: Performance with different MS^2-Scoring Methods

- MetFrag (in-silico fragmenter scores) and IOKR [2] as MS^2-scoring methods

MS^2-Scorer	Method	Top-1	Top-5	Top-10	Top-20
MetFrag	MS^2 + RT (our)	21.3	52.9	64.0	74.3
	Only MS^2 (baseline)	16.7	49.5	60.4	70.6
IOKR	MS^2 + RT (our)	26.7	52.1	62.5	70.3
	Only MS^2 (baseline)	25.1	49.5	60.3	67.6

Experiment 3: Missing MS^2 Spectra

- Simulating missing MS^2 information: Varying from 0% (MS^2) to 100%
- If only MS^1: Use mass deviation between precursor and candidate molecule



References

[1] E. Bach, S. Szedmak, C. Brouard, S. Böcker, and J. Rousu. Liquid-chromatography retention order prediction for metabolite identification. *Bioinformatics*, 34(17):i875–i883, 2018.

[2] C. Brouard, H. Shen, K. Dührkop, F. d'Alché-Buc, S. Böcker, and J. Rousu. Fast metabolite identification with Input Output Kernel Regression. *Bioinformatics*, 32(12):i28–i36, 2016.

[3] K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksenov, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu, and S. Böcker. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods*, 2019. Doi 10.1038/s41592-019-0344-8.

[4] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2005.

[5] C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, and S. Neumann. Metfrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*, 8(1):3, Jan 2016.

[6] E. L. Schymanski, C. Ruttkies, M. Krauss, C. Brouard, T. Kind, K. Dührkop, F. Allen, A. Vaniya, D. Verdegem, S. Böcker, J. Rousu, H. Shen, H. Tsugawa, T. Sajed, O. Fiehn, B. Ghesquière, and S. Neumann. Critical assessment of small molecule identification 2016: automated methods. *Journal of Cheminformatics*, 9(1):22, Mar 2017.

[7] J. Stanstrup, S. Neumann, and U. Vrhovsek. Predret: Prediction of retention time by direct mapping between multiple chromatographic systems. *Analytical Chemistry*, 87(18):9421–9428, 2015. PMID: 26289378.

Acknowledgements: This work has been supported by the Academy of Finland grant 310107 (MACOME); and the Aalto Science-IT infrastructure. SR and JHW were supported by EPSRC (EP/R018634/1). This work started during a visit by JR to SR, funded through the Scottish Informatics and Computing Science Alliance (SICSA) distinguished visiting fellow scheme.