Source code available

Check out our paper

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

ACADEMY OF FINLA

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 [**?** **?**]
- LC retention times (RT) can improve the small molecule annotation [**?** **?**]
- **Challenges utilizing RT information:** (1) LC-system specific RT measurements and (2) public RT databases are limited in size and coverage

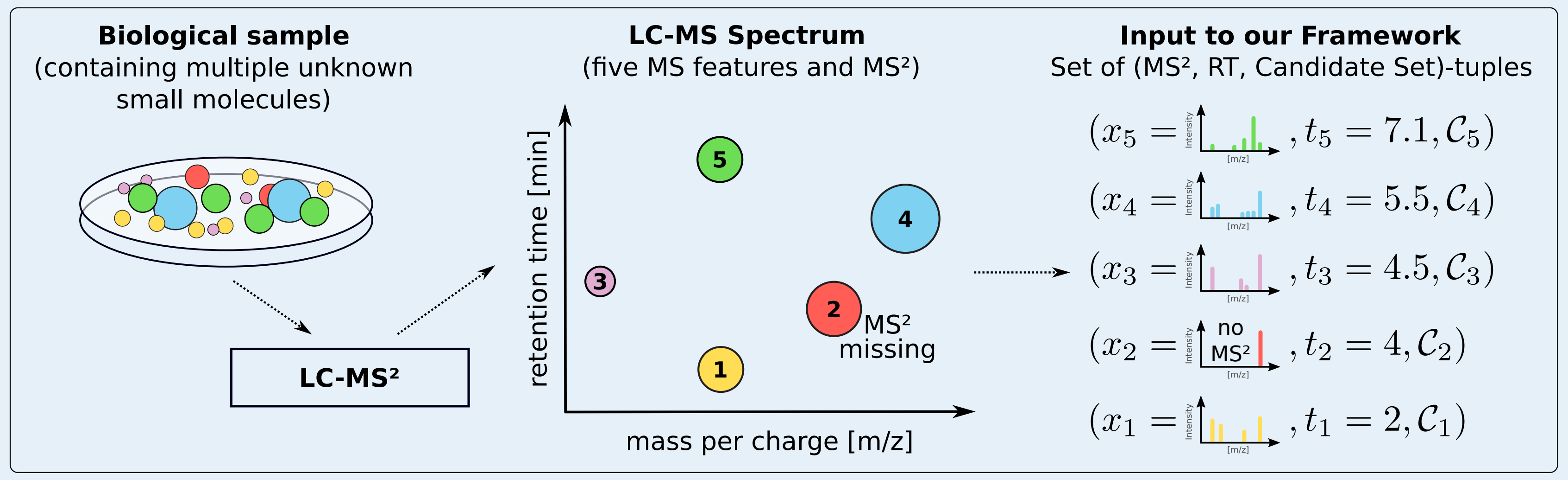

Biological sample
(containing multiple unknown small molecules)
LC-MS²
LC-MS Spectrum
(five MS features and MS²)
retention time [min]
mass per charge [m/z]
5
4
3
2
MS² missing
1
Input to our Framework
Set of (MS², RT, Candidate Set)-tuples
$(x_5 = \ , t_5 = 7.1, \mathcal{C}_5)$
$(x_4 = \ , t_4 = 5.5, \mathcal{C}_4)$
$(x_3 = \ , t_3 = 4.5, \mathcal{C}_3)$
$(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 [**?**], MetFrag [**?**] or IOKR [**?**]
- **Output:** Ranking of the molecular candidates in $\mathcal{C}_i$ for each MS feature i
- Ranking based on MS and RT information

Probabilistic Framework to integrate MS and Retention Orders

- Definition of a probabilsitic graphical model superimposed on the LC-MS data
- Let $G = (E, V)$ be a complete graph
- Nodes $i \in V$ represent the MS features, Edes $(i, j) \in E$ the feature pairs
- Association of each node with discrete random variable $z_i \in \mathcal{Z}_i = \{1, \dots, n_i\}$ ($n_i = |\mathcal{C}_i|$ number of candidates)
- Molecule annotation for complete data $\mathbf{z} = \{z_i \,|\, i \in V\} \in \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_N = \mathbf{Z}$
- Intuitively: Random variable denotes which candidate is assigned to each feature.
- Pairwise Markov Random Field as probabilsitic model [**?**]:

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

- Ranking molecular candidates via max-marginals:

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

- Intuitively, maxmimum probabilsity a candidate assignment with $z_i = r$ can achive
- Rank all candidates $r \in \{1, \dots, n_i\}$ according to there max-marginals

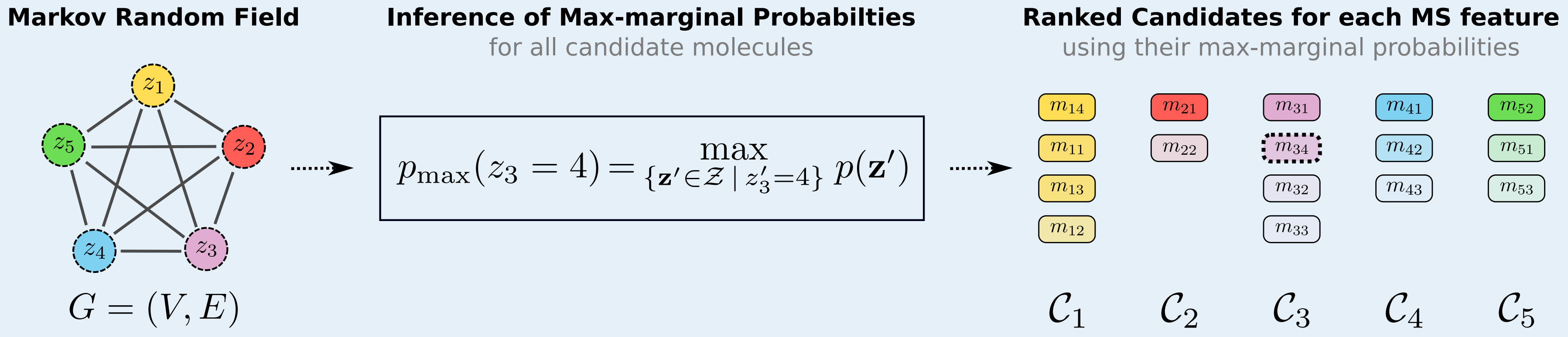
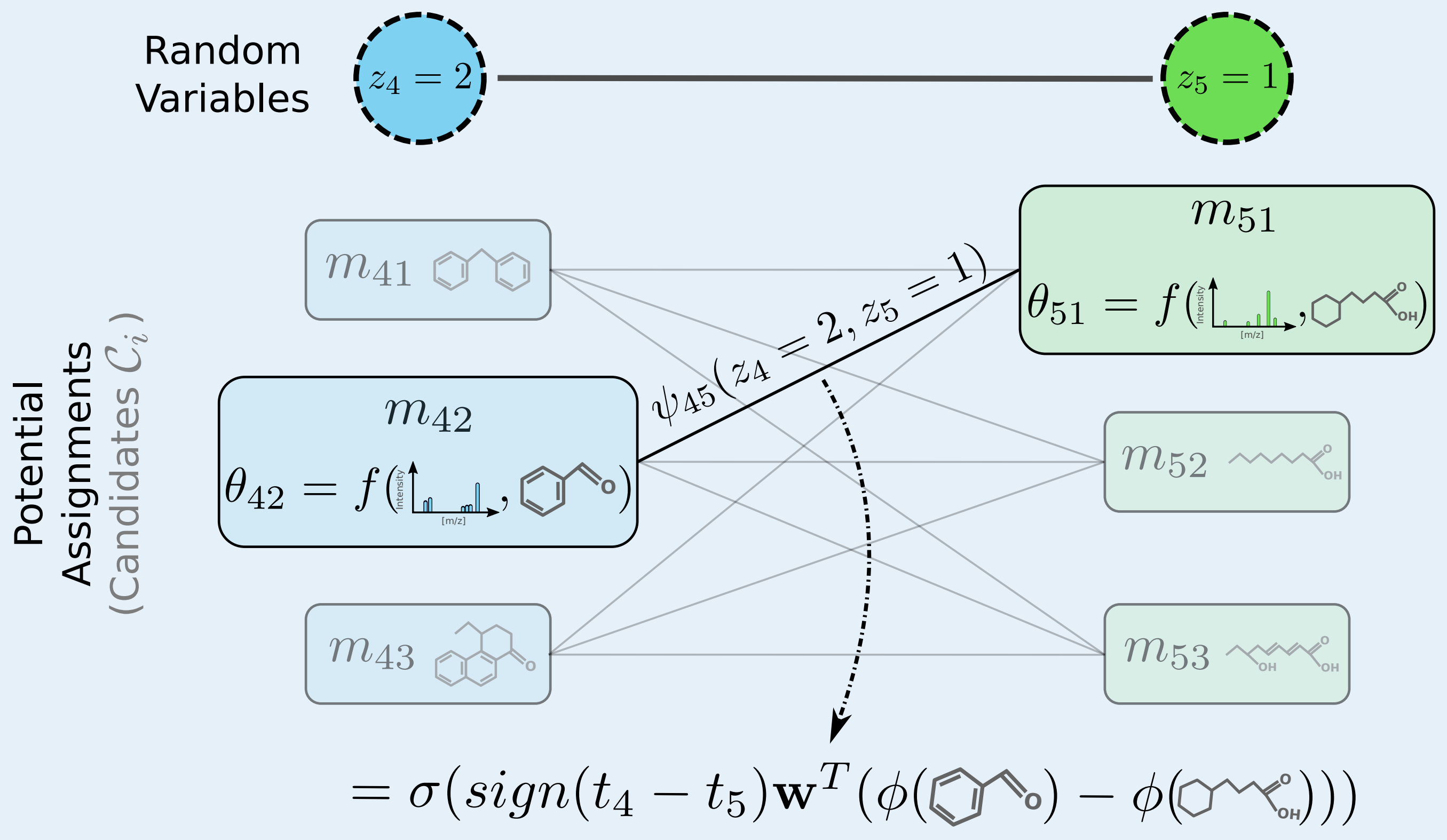

Markov Random Field
z_1
z_2
z_3
z_4
z_5
$G = (V, E)$
Inference of Max-marginal Probabilties
for all candidate molecules
$p_{\max}(z_3 = 4) = \max_{\{\mathbf{z}' \in \mathcal{Z} \,|\, z'_3 = 4\}} p(\mathbf{z}')$
Ranked Candidates for each MS feature
using their max-marginal probabilities
m_{14} m_{11} 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: From the MRF probability distribution to the candidate ranking: MS feature $i = 3$ and candidate 4 (m_{34}).

Node and Edge Potentials

- Node potential function $\psi_i : \mathcal{Z}_i \rightarrow \mathbb{R}_{>0}$: goodness of the match between measured spectrum x_i and candidates of feature i
- Edge potential function $\psi_{ij} : \mathcal{Z}_i \times \mathcal{Z}_j \rightarrow \mathbb{R}_{>0}$: consistency between the observed retention order of feature i and j with the predicted retention order of the candidates z_i and z_j

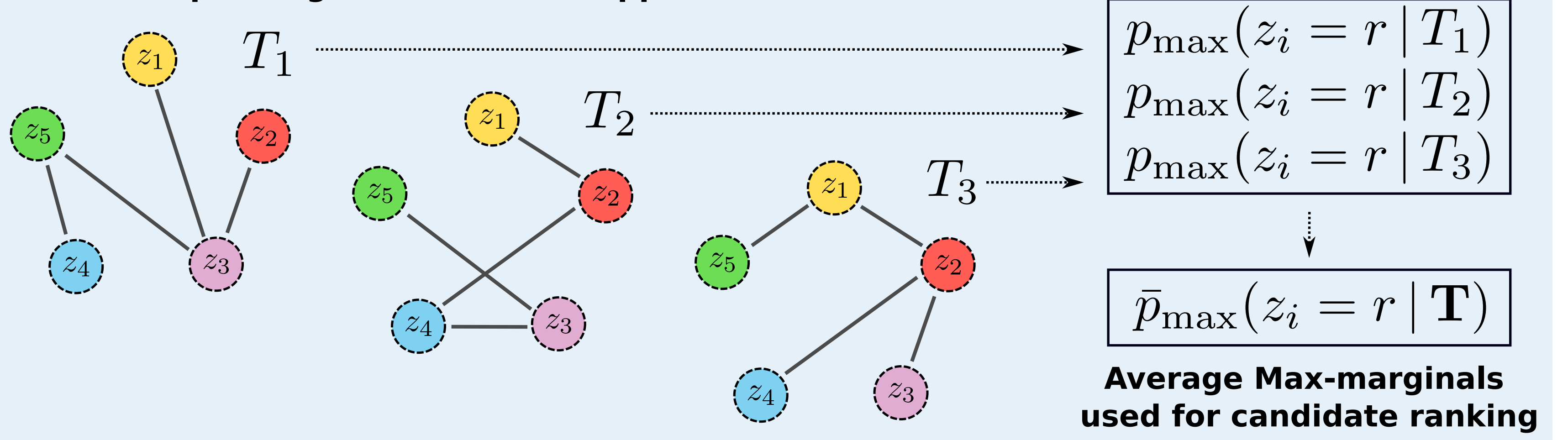
Node and Edge Potential Calculation



Spanning Tree Approximation

- Marginal inference intractable in practice due to exponentail sized candidate assignment space $\mathcal{Z}$
- Exact inference is feasible if G is tree-like **[CITATION]**
- Resort to infer the max-marginals a set of trees $\mathbf{T} = \{T_t\}_{t=1}^{L}$ sampled from G
- Each tree $T_t = (V, E_t)$ is connected graph with all nodes of G but reduces edges set $E_t \subseteq E$
- Avergaged marginals used for ranking

$$\bar{p}_{\max}(z_i = r \,|\, \mathbf{T}) = \frac{1}{L} \sum_{t=1}^{L} p_{\max}(z_i = r \,|\, T_t)$$


Random Spanning Trees are used approximate the MRF

T_1
T_2
T_3
z_1
z_2
z_3
z_4
z_5
$p_{\max}(z_i = r \,|\, T_1)$
$p_{\max}(z_i = r \,|\, T_2)$
$p_{\max}(z_i = r \,|\, T_3)$
$\bar{p}_{\max}(z_i = r \,|\, \mathbf{T})$

Average Max-marginals used for candidate ranking

Fig. 3: TODO

5. Experiments and Results

- Dataset description
- Show table 4 from the paper
- Show figure 3 from the paper

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.