

Annual Report Year 1

Anticipating the Chemistry of Life (ACOL)

Marco Visani^{1,2}

¹Department of Biology, University of Fribourg, 1700 Fribourg, Switzerland

²Swiss Institute of Bioinformatics, 1700 Fribourg, Switzerland

November 7, 2025

Contents

1	Introduction	1
2	Markov Random Field	2
2.1	WP 1	2
2.2	WP 2	3
2.3	Fake Extract	3
2.4	Fake extract	3
3	LOTUS Expanded	4
4	LOTUS Expanded	4
5	EMI-Monorepo	5
6	Perspectives	5

1 Introduction

Public databases compiling genes (e.g., GenBank) and proteins (e.g., UniProt) have long been essential resources for the life sciences. Until recently, no equivalent resource existed for small molecules and their producing organisms. The LOTUS[1] project was designed to address this issue, and forms currently the most comprehensive collection of natural products (NP) occurrences. It however gathers compounds physically isolated from species. While this ensures high confidence, the process of extraction and isolation is too slow to realistically document the chemical diversity of life. To address this issue, we explore two directions: experimental and computational. On one side, the Earth Metabolome Initiative will establish an open, linked knowledge base documenting the metabolome of all living organisms. In parallel, we develop computational strategies to anticipate the chemical constitution of species based on prior knowledge.

My PhD work focuses on developing a probabilistic model that integrates the knowledge from LOTUS with large-scale mass spectrometry datasets to predict species metabolomes and estimate the likelihood of encountering specific natural products across the tree of life. After

training and validation, the model should be able to predict the probability of finding a given molecule in a given species, even if that compound has never been experimentally detected in it before. I am hoping that my PhD will contribute to a better understanding of the distribution of specialized metabolites across the tree of life.

2 Markov Random Field

This constitutes the main project of my PhD.

2.1 WP 1

We assume that the pattern of presence and absence is described by similarities within each dimension. For instance, closely related species may share a similar set of NPs and NPs related in their synthesis may share a similar distribution across species. To model such similarities, we adopt a **Markov Random Field** (MRF) approach.

The first months of my PhD were dedicated to developing the technical skills to implement this model. To develop my C++ programming skills, I started by creating basic model components such as bitwise operations, class design and best practices. I also familiarized myself with the software tools available in Daniel Wegmann’s lab.

Once the first version of the model was implemented, we performed some simulations to evaluate whether the inference procedure could accurately recover known parameters from synthetic data. These simulations revealed some issues in the inference process, which led to refinements in the model’s formulation. In particular, we changed the rate matrix to depend on a single parameter α_c , representing the rate of change:

$$\mathbf{\Lambda}_c = \begin{pmatrix} -\alpha_c & \alpha_c \\ 1 - \alpha_c & \alpha_c - 1 \end{pmatrix} \quad (1)$$

This modification adjusts the transition probabilities between a parent node $p(n)$ and its child n as follows:

$$\mathbf{P}(n) = \exp(\mathbf{\Lambda}_c \nu_c b(n)). \quad (2)$$

From this, the stationary distribution of the root state is derived as:

$$\mathbf{P}_\infty = (1 - \alpha_c, \alpha_c). \quad (3)$$

To improve the robustness of the model against observational noise, we also introduced an error rate parameter ϵ such that:

$$\mathbb{P}(L_{ms} | \mathbf{x}(m, s), R_{ms}) = \begin{cases} \epsilon & \text{if } \mathbf{x}(m, s) = 0, L_{ms} = 1, \\ 1 - \epsilon & \text{if } \mathbf{x}(m, s) = 0, L_{ms} = 0, \\ R_{ms} & \text{if } \mathbf{x}(m, s) = 1, L_{ms} = 1, \\ 1 - R_{ms} & \text{if } \mathbf{x}(m, s) = 1, L_{ms} = 0. \end{cases} \quad (4)$$

Finally, the probability of discovery R_{ms} was redefined as:

$$R_{ms} = (1 - e^{-\gamma_0 P_m})(1 - e^{-\gamma_1 Q_s}) \quad (5)$$

Overall, the objectives defined in the first work package (WP1) were achieved within the expected time frame. We now have a first working implementation of the MRF model that can simulate and infer molecule-species occurrence patterns.

2.2 WP 2

The second work package focuses on modeling and integrating mass spectrometry (MS) data into the model developed in WP1. Mass spectrometry is a key source of experimental evidence for chemical composition and metabolite identification. Integrating such data into the model is essential to link computational predictions with real-world observations and to refine estimates of molecular occurrence across species.

We initially explored the use of ViMMS [2], a simulation framework for MS data, as a potential way to model experimental variability and generate realistic spectra. However, after several discussions with Daniel and Pierre-Marie, we concluded that this approach would not fully meet our objectives. Specifically, ViMMS focuses on simulating acquisition processes, whereas our goal is to incorporate *interpreted* spectral information—i.e., molecular annotations—into the probabilistic model.

We therefore decided to shift direction toward directly integrating the outputs and scores produced by existing annotation tools such as SIRIUS [3], MetFrag [4], and CFM-ID [5]. These tools provide complementary estimates of molecular identity and structural similarity, which can be combined probabilistically to represent confidence in compound-spectrum assignments. Using only the scores from these methods, rather than relying on any single tool, would make the approach more flexible and robust to future methodological improvements.

At this stage, the specific implementation strategy is still under discussion. Our plan for the coming year is to design a framework that can integrate and weight the outputs of multiple annotation tools, allowing the MRF model to incorporate MS-based evidence in a consistent and scalable way.

2.3 Fake Extract

An important aspect of this project is to account for the potential *loss* of molecules during the experimental process from their natural occurrence in biological samples to their detection by mass spectrometry. Modeling this loss is essential to better interpret observed absence patterns and to refine the link between experimental data and the true chemical composition of species. As stated in the grant, our goal is to characterize the classes of analytes that are lost at different stages of the analytical workflow, including extraction, liquid chromatography, and mass spectrometry fragmentation. To achieve this, we have acquired a library of 550 pure natural products, which we plan to mix together to create a “fake extract” that mimics the complexity of a biological sample. This controlled setup will allow us to quantify molecular losses occurring throughout the experimental pipeline, ultimately enabling us to incorporate these biases into the probabilistic model.

In addition to the mixed extract experiment, we aim to analyze each pure compound individually using mass spectrometry. So far, plate 1 has been successfully processed. Plate 3, previously analyzed by a bachelor student, produced inconsistent results and will therefore be reanalyzed. The remaining plates will also be measured and analyzed in the coming weeks or months. The resulting dataset will form the basis for estimating compound-specific detection probabilities and integrating these estimates into the overall modeling framework.

2.4 Fake extract

We would like to model the *loss* of molecules during the extraction process from Nature to the data analysis.

As stated in the grant, we want to model the classes of analytes lost at the stages of extraction, liquid chromatography and mass spectrometry fragmentation. To do so, we have bought a

library of 550 pure natural products that we are going to mix together to simulate a biological extract. This will allow us to estimate the loss of molecules between the Nature and the mass spectrometry analysis. Ideally we would like to integrate this loss into the model.

Additionally we want to pass each pure compound in the mass spec. So far plate 3 was performed by a bachelor student, but results are weird, so we'll pass everything again. Plate 1 is done. Will need to do the rest and analysis the results. I am planning on doing this in the following weeks/months.

3 LOTUS Expanded

One of the side projects of my PhD involves the *in silico* expansion of the molecular space covered by LOTUS. The LOTUS database contains experimentally verified natural products, but it only represents a fraction of the chemical diversity that likely exists in nature. To better capture this diversity, we aim to generate new, hypothetical molecules that are chemically plausible yet not currently documented in LOTUS. The goal of this task is to extend the known chemical space and thereby improve our capacity to annotate real mass spectrometry (MS) data.

To achieve this, we performed an *in silico* expansion starting from approximately 100'000 compounds in LOTUS, generating a set of about $3 \cdot 10^6$ new molecular structures. These generated compounds can then be fragmented *in silico* to produce predicted MS/MS spectra, which can be compared with experimental data. This process should increase the number of possible annotations in real datasets and help us identify potential natural products that have not yet been reported.

As an initial validation, we compared the generated compounds with entries from COCONUT, another large natural product database that does not include species-level occurrence information. About 17'000 of the *in silico* molecules overlapped with structures in COCONUT, suggesting that the generation process produces chemically realistic and biologically relevant compounds rather than random ones.

In the coming months, we plan to use these *in silico* molecules for MS/MS annotation, leveraging the taxonomic information already available in LOTUS to assign potential biological contexts to the newly generated compounds. We also hope that some of these predicted molecules will later appear in newly published studies, for example in the *Journal of Natural Products*, providing empirical support for our computational predictions.

4 LOTUS Expanded

One of my side projects is to perform the *in silico* expansion the molecules present in LOTUS. This allows to create molecules that are not present in LOTUS but that could be potential existing natural products. We then want to fragment *in silico* these molecules. We hope that this would allow to increase the number of annotations we are able to make on real data. We hope that we will find some of the molecules we have generated *in silico* in brand-new research publication such as in the *Journal of Natural Products*. So far, starting from 100'000 compounds in LOTUS we generated $3 \cdot 10^6$ *in silico* compounds. As a first proof that the molecules generated are not completely random, there are about 17'000 of the compounds that are found in COCONUT which is another natural products databases that doesn't report the biological occurrence of the structure.

Using the *in silico* molecules for MS2 annotation using the molecules from LOTUS would also allow to a certain taxonomic information about this *in silico* molecule.

5 EMI-Monorepo

EMI-Monorepo is our rust coded monolith were we are building all the tools for the Earth Metabolome Initiative. Our post-doc Luca is leading the software development of the project. We are building everything in Rust from the backend to the frontend, to the data analysis part. Given the difficulty of the development of the portal, when Luca feels like a project could be a good training for me to learn Rust, he asks me to do it.

I helped Luca for a couple of packages (crates) but in the past months I've lost a bit track of what he is doing and I haven't coded in a while in Rust now.

I'd like to get back onto coding in Rust because it is a skill that I want to expertise by the end of my PhD. I believe more and more that a chemical library similar to RDKit in Rust could be extremely useful for the community.

6 Perspectives

I'd like to continue working with Luca in Rust as it is a skill that I want to master by the end of the PhD.

References

- [1] A. Rutz et al. "The LOTUS Initiative for Open Knowledge Management in Natural Products Research". In: *eLife* 11 (26, 2022). DOI: [10.7554/eLife.70780](https://doi.org/10.7554/eLife.70780).
- [2] J. Wandy et al. "ViMMS 2.0: A Framework to Develop, Test and Optimise fragmentation Strategies in LC-MS Metabolomics". In: *Journal of Open Source Software* 7.71 (30, 2022). DOI: [10.21105/joss.03990](https://doi.org/10.21105/joss.03990).
- [3] K. Dührkop et al. "SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information". In: *Nature Methods* 16.4 (2019). DOI: [10.1038/s41592-019-0344-8](https://doi.org/10.1038/s41592-019-0344-8).
- [4] C. Ruttkies et al. "MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation". In: *Journal of Cheminformatics* 8.1 (2016). DOI: [10.1186/s13321-016-0115-9](https://doi.org/10.1186/s13321-016-0115-9).
- [5] F. Wang et al. "CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification". In: *Analytical Chemistry* 93.34 (31, 2021). DOI: [10.1021/acs.analchem.1c01465](https://doi.org/10.1021/acs.analchem.1c01465).