Annual Report Year 1
Anticipating the Chemistry of Life (ACOL)

Marco Visani[1,2]

[1]Department of Biology, University of Fribourg, 1700 Fribourg, Switzerland
[2]Swiss Institute of Bioinformatics, 1700 Fribourg, Switzerland

2025-11-21

Contents

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 currently represents the most comprehensive collection of natural product (NP) occurrences. It, however, primarily gathers compounds physically isolated from species. While this ensures high confidence, the process of extraction and isolation is too slow to realistically document the full chemical diversity of life. In addition, most journals in the field tend to favor the publication of novel and/or bioactive compounds, leading to a strong publication bias. As a result, many known or inactive natural products remain unreported, further limiting our understanding of the chemical diversity of living systems.
To overcome these limitations, we are developing complementary strategies—both experimental and computational. On the experimental side, the Earth Metabolome Initiative (EMI) aims to build an open, linked knowledge base documenting the metabolome of all living organisms. This global effort will convert large-scale metabolomics data and associated metadata into a structured and Linked Open Data format, creating a comprehensive and accessible resource

for the scientific community. However, this project will require a substantial amount of time, coordination, and resources, as it relies on the gradual accumulation of high-quality experimental data across species.
In parallel, and to provide earlier insights while EMI is being developed, we pursue a computational approach to *anticipate the chemistry of Life* (ACOL). The goal of this project is to use existing knowledge from LOTUS together with large-scale mass spectrometry datasets to model and predict the chemical composition of species.
My PhD work focuses on developing a probabilistic model that integrates the knowledge from LOTUS with mass spectrometry datasets to predict species metabolomes and estimate the likelihood of encountering 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.

2 Markov Random Field

2.1 WP 1

Phylogenetically informed analyses have demonstrated that metabolite production is not randomly distributed but instead exhibits evolutionary patterns across taxa [2, 3, 4]. These results justify our assumption that similarities among species—and among molecules—can be modeled to predict presence/absence patterns. 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. MRFs provide a way to model joint probability distributions over many variables while focusing on local dependencies, making it possible to capture pairwise similarities between species or molecules while still maintaining a coherent global probabilistic structure.
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 C++ best practices. I also familiarized myself with the software tools available in Daniel Wegmann's lab such as *coretools* and *stattools*.
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, which led to some modifications.
In particular, we modified the rate matrix to depend on a single parameter, α_c, representing the rate of change from state 0 (*i.e.* being absent) to state 1 (*i.e.* being present), as shown in Equation 1.

$$\mathbf{\Lambda}_c = \begin{bmatrix} -\alpha_c & \alpha_c \\ 1-\alpha_c & \alpha_c - 1 \end{bmatrix} \tag{1}$$

An α_c value close to 1 indicates that if the parent node is in state 1, the child node is also very likely to be in state 1. Conversely, if the parent node is in state 0, once the child transitions to state 1, it tends to remain in that state.

Example. Suppose $\alpha_c = 0.9$. For simplicity, let us set both the branch length and scaling factor ν_c to 1. The transition probabilities between a parent node $p(n)$ and its child n are then given by:

$$\boldsymbol{P}(n) = \exp\left(\begin{bmatrix} -\alpha_c & \alpha_c \\ 1-\alpha_c & \alpha_c - 1 \end{bmatrix}\right) = \exp\left(\begin{bmatrix} -0.9 & 0.9 \\ 0.1 & -0.1 \end{bmatrix}\right) = \begin{bmatrix} 0.431 & 0.569 \\ 0.063 & 0.937 \end{bmatrix}$$

In this case, during the MCMC updates, if the parent node is 0, the child node has roughly equal probabilities of being 0 or 1. However, if the parent node is 1, the child node is also very

likely to be 1. This means that once a node transitions to state 1 during the MCMC updates, it will likely remain in that state.

The modification in Equation 1 adjusts the transition probabilities between a parent node $p(n)$ and its child n as follows:

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

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

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

Let $L_{ms} = 1$ denote a known occurrence of NP m in species s, and let $L_{ms} = 0$ denote that no evidence for such an occurrence has been reported, either because the NP m is truly absent in species s or because of a lack of research effort. Let $x(m, s)$ denote the *truth* for NP m and species s where $x(m, s) = 1$ denotes the presence of molecule m in species s and $x(m, s) = 1$ when it is absent.
To improve the robustness of the model against observational noise, we introduced an error rate parameter ϵ such that:

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

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

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

Where P_m and Q_s represent the total number of relevant papers published for molecule m and species s respectively and γ_0 and γ_1 are positive scalars that are inferred by the model.
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. However, the current inference is not producing useful predictions. For instance, when running the model on a subset of LOTUS (the *Asterales* family), it ended up predicting zeros everywhere and treating all known occurrences as mistakes—most likely because of the ϵ parameter we introduced. When fixing the error rate to zero, the model only predicted the molecules already present in LOTUS and did not predict anything new.
At this point, I am not yet sure why the inference behaves this way, but one potential explanation is the extreme sparsity of the data. As a next step, I would like to enrich the dataset synthetically—for example by artificially adding occurrences in LOTUS to test whether denser data helps the model converge toward meaningful predictions. I'd also like to examine the behavior of the model in a simplified setting by inferring the presence of a single, highly abundant molecule, which should allow us to assess whether the inference is capable of predicting meaningful 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 connect computational predictions with real-world observations and to refine estimates of molecular occurrence across species. Compared to the physical isolation of compounds, MS enables much higher throughput and a greater number of annotations, while also allowing repeated acquisitions on the same extract. This makes it a powerful and complementary approach to databases like LOTUS, offering broader chemical coverage at the cost oflower annotation confidence —an aspect that we will need to take into account in the model.
We initially envisioned the use of ViMMS [5], a simulation tool 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 [6], MetFrag [7], and CFM-ID [8]. 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. The idea of remaining independent of any single annotation tool has been was shown by the rapid emergence of new models. In the past year alone, two previously closed-source approaches—ICEBERG [9] and FIORA [10]—have released open-source model weights thanks to the MassSpecGym [11] dataset. Early benchmarking suggests that both models outperform established tools such as SIRIUS, MetFrag, and CFM-ID. This rapid pace of improvement shows the importance of designing a model that can flexibly incorporate scores from multiple annotation methods, rather than relying on any fixed set of tools.
At this stage, the specific implementation strategy is still under discussion. Our plan for the coming year is to 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 Artificial Extract

As stated in the grant, we would like to model and include in the Markov Random Field, the potential *loss* of molecules during the experimental process from their presence in Nature to their identification by mass spectrometry (*i.e.* 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 an "artificial extract" that mimics the complexity of a biological sample. The goal of this is to quantify the molecular losses that occur during the analytical workflow and incorporate these biases into the MRF model. This approach will not allow us to assess losses occurring during the extraction step, since we are not re-extracting the mixture; instead, it will let us benchmark the ionization and metabolite annotation stages. In addition to molecular losses, this extract will also help us characterize the "additions," such as ions, in-source fragments, and other degenerated features introduced during MS analysis.
Furthermore, we also aim to analyze each pure compound individually using mass spectrometry. So far, Plate 1 has been processed, and Plate 3 was analyzed by Jade Dandois, a bachelor student whom I coached between February and June of this year. As shown in Figure 1, Plate 3 displays a noticeably higher number of missing molecules than Plate 1, with several compounds absent in both ionization modes. Given the high number of missing compounds in Plate 3 I plan to re-run it in the coming weeks, along with the remaining plates.

(a) Plate 1 *(b) Plate 3*

Figure 1: *Plate 1 and Plate 3 of the ApexBio Natural Products library. In Plate 1, 5 out of 88 molecules were not detected in either positive or negative ionization mode. In contrast, in Plate 3, 17 molecules were not detected.*

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. In addition, these newly generated molecules could be incorporated into the Markov Random Field to enable predictions for new and unseen compounds.

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. This expansion was carried out using the published tool MINE-Database [12], which was forked and further improved by our former Master's student, Pascal Amrein.

These generated compounds can then be fragmented to produce an *in silico* spectral database (ISDB), 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 [13],

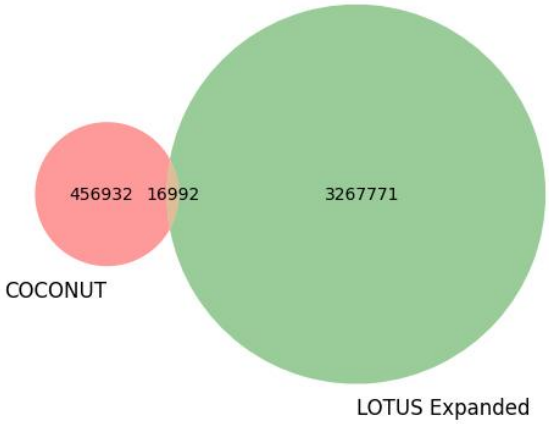

Figure 2: *Venn plot between molecules in COCONUT and the generated in silico compounds. The number represents the number of unique molecules in each dataset. 16992 molecules from the in-silico expansion can be found in COCONUT, a natural product database.*

another large natural product database that does not include species occurrence information for all its compounds. As shown in Figure 2 about 17′000 (0.57%) of the *in silico* molecules overlapped with structures in COCONUT, suggesting that the generation process can produce chemically realistic and biologically relevant compounds.
I also hope to identify some of the predicted molecules in newly published studies, such as those in the *Journal of Natural Products*, which would provide empirical support for our chemical expansion approach. However, a major limitation is that most publications in this journal do not provide molecular structures in a machine-readable format, such as SMILES. This makes it difficult to perform rapid validation or to develop automated tools capable of detecting these compounds in the literature.

4 EMI-Monorepo

The EMI-Monorepo is a Rust-based monolithic repository that hosts all software tools developed for the Earth Metabolome Initiative. The project is led by our postdoctoral researcher, Luca, who supervises the development of the entire platform—from backend and frontend components to data analysis modules. Building the infrastructure in Rust ensures high performance, reliability, and scalability, which are essential for handling large-scale metabolomics data.
Given the complexity of the portal, Luca occasionally assigns me specific tasks that serve both the project and my own training in Rust programming. This has allowed me to contribute to several (*crates*) within the repository and to better understand the structure and design principles of the software. Over the past months, however, my involvement has decreased as I focused on other aspects of my PhD, and I have not coded in Rust recently.
In the coming year, I aim to re-engage with the EMI-Monorepo and further develop my expertise in Rust. Mastering this language is one of my personal objectives for the PhD. I believe that developing a dedicated chemical library—similar to RDKit but implemented in Rust—would be a valuable contribution to the scientific community and a meaningful way to advance my skills. This idea aligns with a proposal from our colleague Daniel Probst, who suggested developing such a Rust-based chemoinformatics library and invited us to collaborate on it.

5 Frag-Graph

Another side project I am exploring is what I call the *fragmentation graph*. Existing tools such as SIRIUS use fragmentation trees to annotate molecules from mass spectra. These trees represent how a single molecule can fragment under specific conditions. However, I wondered whether it would be possible to generalize this concept by constructing a single, large *fragmentation graph* that connects all molecules together through their potential fragmentation relationships.

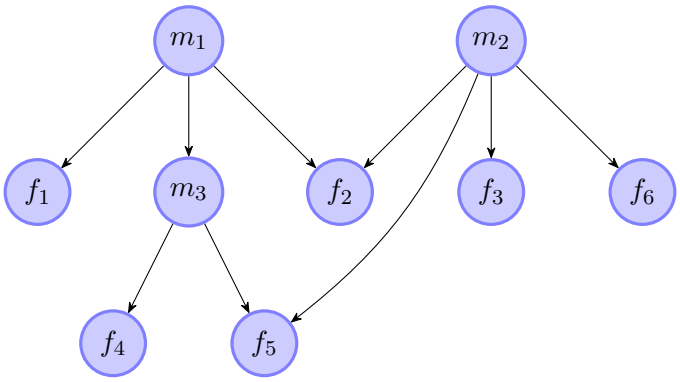

Figure 3: *Example of a fragmentation graph limited to a single generation (i.e., fragments generated from the starting molecules are not further fragmented). Nodes labeled m represent starting molecules, and nodes labeled f represent fragments. A starting molecule may produce another starting molecule as a fragment (e.g., m_1 produces m_3), and different molecules may generate identical fragments (e.g., both m_1 and m_2 generate f_2).*

In this approach, each input molecule is fragmented by breaking all its bonds. The resulting fragments are considered as the molecule's children, and the full collection of molecules and fragments forms one single graph.
When analyzing a mass spectrum, we can then iterate over all nodes in the graph and check whether a fragment mass matches any of the observed peaks (within a defined tolerance). If a match is found, the corresponding intensity is assigned to that fragment and propagated to all its parent molecules in the graph.

Example. Let us use Figure 3 as an example. Suppose we iterate over the masses in a mass spectrum and that the true molecule is m_1. If one of the observed masses matches fragment f_5, then the parent molecule could be m_1, m_2 or m_3. If we then observe a mass matching f_4 the most likely parents narrow to m_1 or m_3. Finally, if a mass corresponding to f_1 is detected, m_1 becomes the most plausible candidate, as it explains the largest number of observed fragments. This approach relies on the assumption that a spectrum originates from a single molecule, and that enough of its fragments will be detected to uniquely identify it. In practice, different molecules generate distinct—though sometimes partially overlapping fragments, so accumulating multiple matching peaks helps narrow down the candidates until only the correct molecule remains.

To test this idea, I generated a fragmentation graph from the molecules in *ISDB: In Silico DataBase (isdb_lotus_pos_energySum.mgf)*. Starting from 141,514 molecules, the procedure produced a graph containing 1,656,090 nodes (molecules or fragments) and 3,038,519 edges. I then downloaded the spectra from MassSpecGym and randomly sampled 500 of them for evaluation. As shown in Table 1, even with a very naive fragmentation scheme—simply breaking each bond of the parent molecule and recording the two resulting fragments—the correct molecule appears

within the top 10 candidates in more than 50% of the test cases. For 36 spectra, the correct molecule was not retrieved within the ranked results.
These results are far from optimal, but this is expected given how rudimentary the current fragmentation strategy is. The procedure can (and likely should) be improved by incorporating more realistic fragmentation rules or by using better scoring functions than peak intensity alone. The simple bond-breaking approach was inspired by FIORA, an *in silico* fragmentation algorithm designed to predict MS/MS spectra with high accuracy.

	rank
count	464
mean	30.06
std	79.41
min	1
25%	3
50%	10
75%	32
max	1390

Table 1: *Evaluation of 500 randomly sampled spectra from MassSpecGym. Out of 500 spectra, the true inchikey was not found in 36 of them. In half of the spectra, the true inchikey was in the top 10 or better.*

Given these interesting preliminary results, I am motivated to continue developing this approach and explore its potential as a tool for metabolite annotation. One of its main advantages is that it does not rely on machine learning; instead, the fragmentation rules can be iteratively refined should we ever want to. In the coming months, I plan to continue developing the tool and assess its performance against existing annotation tools.

6 Perspectives

In the coming year, I intend to focus and continue to work on the following aspects. First, I will integrate MS/MS (MS2) data into the Markov Random Field (MRF) model and design a method for converting the outputs of various annotation tools (such as SIRIUS, MetFrag, and CFM-ID) into a format compatible with the MRF. This step is essential for linking experimental mass spectrometry evidence with computational predictions and for improving the model's accuracy in estimating molecular occurrences across species.
Regarding WP2, I plan to implement an efficient system for storing and accessing the annotation scores used by the MRF. Each MS run corresponds to a species and can contain thousands of spectra, with each spectrum potentially assigned to multiple candidate molecules. As a result, the volume of data can become extremely large, and we need to determine an effective strategy for managing and querying it. In addition, the scores produced by some annotation tools do not directly represent probabilities, so we must develop a method to appropriately normalize or convert them before integration into the model.
I also plan to complete and publish a short paper describing the *LOTUS Expanded* project. Additional validation of the generated compounds is still required, but this work could provide valuable insights into the unexplored regions of natural product chemical space.
In parallel, I intend to continue developing my programming skills in Rust through contributions to the EMI-Monorepo. My goal is to progressively build small, modular cheminformatics libraries

(*crates*) that could later be integrated into the EMI platform, contributing to the broader open-source metabolomics community.
Finally, I plan to further investigate the *Frag-Graph* concept, which has shown promising initial results for metabolite annotation. I will continue testing this approach, with the hopes of developing it into a functional annotation tool.

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.

[2] M. Adamek, M. Alanjary, and N. Ziemert. "Applied Evolution: Phylogeny-Based Approaches in Natural Products Research". In: *Natural Product Reports* 36.9 (2019). DOI: 10.1039/C9NP00027E.

[3] N. Rønsted et al. "Can Phylogeny Predict Chemical Diversity and Potential Medicinal Activity of Plants? A Case Study of Amaryllidaceae". In: *BMC Evolutionary Biology* 12.1 (2012). DOI: 10.1186/1471-2148-12-182.

[4] Y. Zhang et al. "Phylogenetic Patterns Suggest Frequent Multiple Origins of Secondary Metabolites across the Seed-Plant 'Tree of Life'". In: *National Science Review* 8.4 (24, 2021). DOI: 10.1093/nsr/nwaa105.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] R. Wang et al. *Neural Spectral Prediction for Structure Elucidation with Tandem Mass Spectrometry*. 1, 2025. DOI: 10.1101/2025.05.28.656653. URL: http://biorxiv.org/lookup/doi/10.1101/2025.05.28.656653 (visited on 11/20/2025). Pre-published.

[10] Y. Nowatzky et al. "FIORA: Local Neighborhood-Based Prediction of Compound Mass Spectra from Single Fragmentation Events". In: *Nature Communications* 16.1 (7, 2025). DOI: 10.1038/s41467-025-57422-4.

[11] R. Bushuiev et al. *MassSpecGym: A Benchmark for the Discovery and Identification of Molecules*. Version 3. 2024. DOI: 10.48550/ARXIV.2410.23326. URL: https://arxiv.org/abs/2410.23326 (visited on 11/20/2025). Pre-published.

[12] J. Strutz et al. "MINE 2.0: Enhanced Biochemical Coverage for Peak Identification in Untargeted Metabolomics". In: *Bioinformatics* 38.13 (27, 2022). Ed. by Z. Lu. DOI: 10.1093/bioinformatics/btac331.

[13] M. Sorokina et al. "COCONUT Online: Collection of Open Natural Products Database". In: *Journal of Cheminformatics* 13.1 (2021). DOI: 10.1186/s13321-020-00478-9.