UNI
FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

UNIVERSITY OF FRIBOURG

MASTER THESIS

Anticipating the chemical compositions of organisms across the tree of life.

Author:
Marco VISANI

Supervisors:
Prof. Daniel WEGMANN
Dr. Pierre-Marie ALLARD

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science in Bioinformatics and Computational Biology

in the

Wegmann Group & COMMONS Lab
Department of Biology

October 17, 2023

UNIVERSITY OF FRIBOURG

Abstract

Faculty of Science and Medicine
Department of Biology

Master of Science in Bioinformatics and Computational Biology

Anticipating the chemical compositions of organisms across the tree of life.

by Marco VISANI

This study is centered on Natural Products (NPs) - specific chemicals synthesized by living organisms. These NPs hold significant importance in various domains, notably medicine, agriculture, and ecology. A primary resource for our research is the LOTUS database, which catalogues a vast array of NPs and their occurrence. Yet, a gap exists: there are no existing model to predict the occurrence of these NPs across different species.

In our initial strategy, the occurrence of natural products was viewed as a collection of observations and their associated variables. Although simple, this strategy immediately showed its limits when dealing with the complex nature of NPs. We switched to an advanced graph-based method after seeing the necessity for a more thorough strategy to accurately represent the intricate interactions governing NPs expression. When considering species in a phylogeny or molecular pathways, the graph-based method perceives data as a network of connected entities, offering a far more logical and natural way of thinking. By employing this better methodology, we have developed a more effective approach for investigating the intricate world of Natural Products. We hope that this strategy will open up new research directions and possibly result in ground-breaking NP-related findings.

Acknowledgements

First and foremost, I would like to extend my gratitude to my supervisors, Prof. Daniel Wegmann and Dr. Pierre-Marie Allard. Their endless support, guidance and ability to inspire enthusiasm in me, particularly when introducing new facets of the project, have been crucial to my thesis. Their ability to always find the right words to motivate me has been invaluable.

I am really grateful to Maëlle Wannier and Axel Giottonini, who have been my constant support during this journey. Your unfailing support and aid during my studies forged our friendship, securing our status as friends for life.

I would also like to thank Yves Steiner. The endless litres of free coffee and jokes you brought every day at lunch were nothing but joy. On many difficult days, your consistent good mood has been a source of happiness.

To my roommate Noémie Poli, and my ex-roommate, Gaël Vonlanthen: living with you both has been an adventure I will remember for the rest of my life. The memories we've shared in our apartment are ones that I will hold close to my heart.

To my ever-supportive girlfriend Noémie Salvador. I cannot explain how grateful I am to you. During my peaks of excitement, my moments of doubt, my never ending stress, you have been my safety net. Your immense support, patience and kindness have been invaluable.

Finally, I want to express my gratitude to my family who have been the basis of my academic journey. Their constant support throughout my education has served as the foundation for my goals in life.

Contents

List of Abbreviations

CC	**C**ollective **C**lassification
DAG	**D**irected **A**cyclic **G**raph
GCN(s)	**G**raph **C**onvolutional **N**etwork(s)
GNN(s)	**G**raph **N**eural **N**etwork(s)
(k)PCA	(kernel) **P**rincipal **C**omponent **A**nalysis
MS	**M**ass **S**pectrometry
NP(s)	**N**atural **P**roduct(s)
RMF	**R**andom **M**arkov **F**ield

1 Introduction

1.1 Natural Products: Definition and Roles

Natural Products (NPs) are chemical entities biosynthesized by living organisms [1]. NPs are metabolites, which can be arranged along a gradient of specialisation from core metabolites, which perform fundamental tasks and are present in a variety of organisms, to specialised metabolites, which are much more restricted in occurrence across the tree of life. Natural product research is interested in the underlying structural features of naturally occurring molecular entities, their effects on living organisms, and even the study of chemically mediated interactions across entire ecosystems. Through the course of this work, we will use *Rutz et al.*'s definition of Natural Product as *any chemical entity found in a living organism* [2].

1.2 The Importance of Natural Products

Because of their distinct chemical structures and activities, specialised metabolites form the basis of current therapeutic treatments [3]. They are relevant to a variety of industries, including agriculture [4], food industry [5], cosmetics [6], and other fields, in addition to human and veterinary medicine. These natural goods are inextricably related to renewable resources and add great value in our economy.

Understanding the complexities of their biological functions and structural characteristics is essential to understanding how ecosystems work. These complexities influence a variety of aspects, from the impact on individual organisms to the overall chemically mediated interactions within an entire ecosystem.

Several fundamental aspects of chemistry such as stereochemistry, optical activity, regioselectivity, and chirality, have also been advanced in the field of NP research. Their complexity has lead to the development of cutting-edge tools that can replicate natural processes to manage bioregulation mechanisms and solve practical problems [7].

Even though they are complex and challenging to describe, the role of natural products in therapeutic uses and ecosystem functionality cannot be overstated. Current developments aim to unlock this potential more efficiently, emphasizing the ongoing advancements across all sectors associated with natural products.

1.3 The LOTUS Database and Current Efforts

In recent years, efforts have been made to report the constituents of metabolic networks or occurrences of molecules in selected taxa [8]. A significant resource is the LOTUS database [2], which currently lists over 438′072 molecule-species pairs. The LOTUS initiative aims to consolidate and share structure-organism pair information via an open platform, which has the potential to make advancements in NPs research. This process involved the harmonization, curation, validation, and open dissemination of referenced structure-organism pairs. Furthermore, LOTUS data's embedding into the vast Wikidata knowledge graph facilitates new biological and chemical insights. The contemporary bioinformatic capabilities offered by the LOTUS initiative have the potential to reshape knowledge management, analysis, and interpretation of data in natural products research [2]. Despite this, no generic solution has been devised to *predict* metabolomes across the tree of life.

1.4 Project Description and Objectives

The goal of this project is to develop such a model and to train it using large-scale metabolomics and other occurrence data.

For this research, we consider two primary sources of information : 1) The LOTUS database, which lists reports of metabolites present in particular species which have been confirmed by peer-reviewed scientific literature. 2) Data derived from mass spectrometry (MS) analyses.

LOTUS's inherent value stems from its strong data integrity. Indeed, most of the time, to be present in LOTUS, a compound must have been extracted and characterized from a living organism, ensuring its data validity. However to ensure such data quality, a lot of resource and effort are required (isolation, purification, structural determination). Currently, only 0.008% of all potential occurrences in the database are represented by the 438′072 occurrences that have been archived in LOTUS.[1] A thorough knowledge of the genuine presence or absence matrix of molecules across the evolutionary spectrum is still difficult given the pace of academic research.

In contrast, mass spectrometry (MS) provides a quicker and more practical approach of molecular detection. Coupled to computational mass spectrometry tools, such untargeted approaches allows to detect and annotate thousands of NPs in a single run. However, this increased effectiveness comes with restrictions. In particular, MS shows limitations when trying to identify molecules with previously unidentified structural configurations. Moreover, the putative annotation of NPs is MS diminishes the degree of confidence in assigning a compound to a specific species, presenting a trade-off between throughput and confidence.

[1] Currently LOTUS has 36′800 species, and 148′190 molecules.

By integrating these two sources of information and using an appropriate model, we want to anticipate the complete chemical composition of organisms across the tree of life.

2 Theoretical introduction

In the early stages of our research, we did not utilize a graph-based approach. We started our work treating the data as a collection of observations and their covariates rather than recognizing the inherent graph-based structure of molecules and organisms. This initial approach, while more straightforward, failed to fully capture the intricate complexity and interconnected nature of our dataset. Recognizing the limitations of such *naive* methods, we transitioned to the application of graphical models, specifically focusing on graph neural networks and collective classification techniques (see below). By treating our data as a graph, we aim to better encapsulate the nuanced relationships and dependencies between species, molecules, and their respective attributes.

In discrete mathematics [9], a graph is a collection of elements, known as vertices or nodes, and their connections, known as edges or links. Edges represent the connections or affiliations between pairs of these points, whilst vertices serve as discrete points or units. For example, in biology, a graph can show the complex network of connections between proteins in a cellular system. Each protein is represented by a node, and the functional or physical relationships between them are denoted by edges [10]. Figure 2.1 shows an example of such graph.

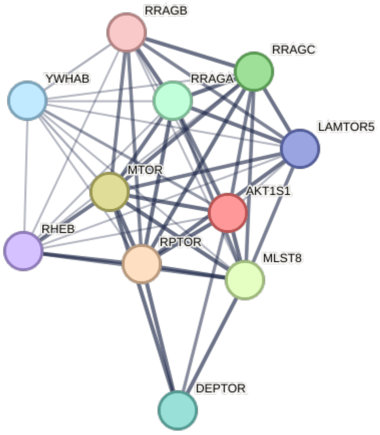

Figure 2.1: Graph representation of the interactions of AKT1S1, a subunit of mTORC1 in *Homo sapiens*. Each protein is represented by a node, and the physical relationships between them are denoted by edges. The edges' thickness indicates the confidence that two proteins are linked. The image was downloaded from STRING [11] queried for mTORC1 in *Homo sapiens*.

Often, these nodes carry specific attributes (also known as features). For example, nodes in a graph could represent cities around the world. The attributes of these cities might include population size, geographic location, or average temperature. However, obtaining these features can be challenging due to obstacles in data collection, inherent complexity, or, in some cases, privacy concerns [12]. To mitigate this, graph-based semi-supervised learning, also known as node classification, is employed to predict missing labels for some nodes given known attributes (*i.e.* features). This strategy has been effective in a myriad of applications, including predicting molecular functions and categorization of substances [13, 14].

Link prediction is also a fundamental task in graph theory, aiming to forecast the likelihood of a potential relationship or edge between two nodes within a network. In the context of social networks, as highlighted by Liben-Nowell and Kleinberg [15], the challenge is to determine which interactions are likely to emerge in the future based on the current network topology. The underlying hypothesis is that the inherent structure of the network contains valuable information about future interactions. Various measures of node "proximity" or similarity within the network can be employed to make these predictions, and some nuanced measures (*e.g.* SEAL developed by [16]) have been found to outperform more direct ones (*e.g.* Adamic–Adar index [17]).

Graph neural networks (GNNs) [18] have been frequently employed for semi-supervised learning or for link predictions, also in the context of molecular networks [19]. Initially, GNNs synthesize the features and graph structure in the vicinity of each node into a single vector representation. Then, this representation is individually used for the classification of each node. The benefits of using GNNs include automatic differentiation, enabling end-to-end training and straightforward sub-sampling schemes for handling extensive networks. However, the use of GNNs hinges on the assumption that node labels are conditionally independent given all features. Moreover, these networks do not leverage correlations between training and testing labels during inference, and due to the complexities of their transformations and aggregation functions, the derived models can be challenging to interpret.

Alternatively, collective classification (CC) [20] provides an interpretable approach, utilizing graphical models that directly exploit label correlation for prediction. One such model used within our research is Markov networks also known as Markov Random Field (MRF). MRFs model the joint distribution of all node labels within a conditional random field and predict an unknown label with its marginal probabilities. This method makes use of label correlation during inference, which involves conditioning on the training labels. However, the increased interpretability and convenience of collective classification comes at a price. The models are learned by maximizing the joint likelihood, rendering end-to-end training extremely difficult. This, in turn, restricts the capacity and versatility of the model [12].

Below, we introduce the three models we have developed. Components of the naive model were retained for modelling the probability of the data in the RMF model (refer to Sections 2.1.1 and 2.2.1. We believe that the RMF model holds the most promise among our developments; however, implementation and testing are still underway and constitute our current work in progress.

2.1 Naive approach

Our objective is to infer the presence or absence of metabolites across a collection of samples, which are differentiated by T discrete dimensions such as species, tissue type, and environmental conditions or any other arbitrary dimension. For any compartment c, let $\tau_t(c) = 1, \ldots, n_t$ indicate the compartment index along axis $t = 1, \ldots, T$. For convenience, let us further denote by $\tau_{\mathcal{M}}(c)$ and $\tau_S(c)$ the metabolite and species of that compartment.

We denote x_c the presence ($x_c = 1$) or absence ($x_c = 0$) of a metabolite $\tau_{\mathcal{M}}(c)$ in compartment c and let $\boldsymbol{x} = (x_1, \ldots, x_C)$ be the full vector x_c across all compartments $c = 1, \ldots, C$ with $C = \prod_t n_t$.

We will assume that similarities across any of the axes of compartmentalization is reflected in the patterns of presences and absences in $\boldsymbol{x}$. For instance, closely related species may share a similar set of metabolites and metabolites related in their synthesis may share a similar distribution across species. To model such similarities, we assume that the probability $\mathbb{P}(x_c = 1 | \boldsymbol{\mu}_c, \epsilon_c)$ with which metabolite $\tau_{\mathcal{M}}(c)$ is present in compartment c is given by

$$\operatorname{logit} \mathbb{P}(x_c = 1 | \boldsymbol{\mu}_c, \epsilon_c) = \sum_{t=1}^{T} \mu^{(t)}_{\tau_t(c)} + \epsilon_c \tag{2.1}$$

where $\boldsymbol{\mu}_c = (\mu^{(1)}_{\tau_1(c)}, \ldots, \mu^{(T)}_{\tau_T(c)})$ is a vector of axis specific intercepts and ϵ_c is normally distributed with mean 0 and co-variance

$$\operatorname{cov}(\epsilon_c, \epsilon_{c'}) = \sum_t \beta^{(t)}_{\tau_t(c)} + \sum_t \beta^{(t)}_{\tau_t(c')} + \sum_t \sum_{f=1}^{F_t} \alpha_{tf} \sigma_{tf} \Big(\tau_t(c), \tau_t(c') \Big). \tag{2.2}$$

Here, the $\beta^{(t)}_{\tau_t(c)}$ are positive intercepts specific for the compartment index $\tau_t(c)$ along axis t, the $\sigma_{tf}, f = 1, \ldots, F_t$, are the F_t known covariance matrices between entries along axis t, and the α_{tf} are positive scalars.

In Figure 2.2, a graphical representation illustrates the probabilistic association of a metabolite's presence within a particular species, predicated upon its mean prevalence across a phylogeny (denoted by the black vertical line at $x = 2.5$). Consider, for example, a metabolite ubiquitously observed across various taxa. If a distinct clade within the phylogenetic tree lacks this metabolite, the likelihood of its occurrence in species phylogenetically proximate to this clade diminishes — as indicated by the dashed leftmost red line.

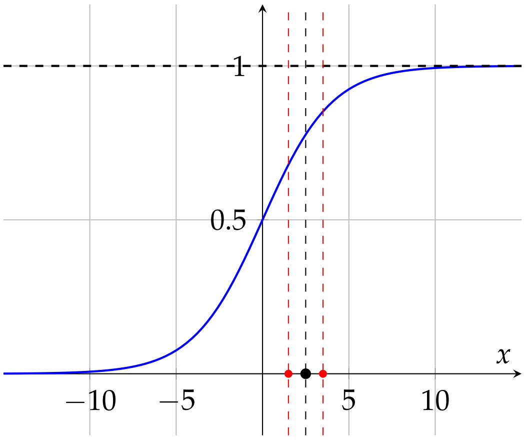

Figure 2.2: Illustration of the proposed model on a logistic function given by Equation (2.1). The black vertical line at $x = 2.5$ denotes the mean presence of a metabolite in the system. Flanking this, the two red dashed lines depict potential shifts in this average presence due to the influence of covariates, denoted as ϵ_c in Equation (2.1).

2.1.1 Data sources

We consider two types of data informative about $\mathcal{X}$: i) presence-only reports of specific NPs in specific species as available through the LOTUS database and ii) presence-absence data obtained with mass-spectrometry (LC-MSMS).

LOTUS

As previously stated, LOTUS database [2] lists known occurrences of metabolites in species. Let $L_{ms} = 1$ denote a known occurrence of metabolite m in species s, while $L_{ms} = 0$ denotes that no evidence for such an occurrence has been reported, either because the metabolite m is truly absent in species s or because of a lack of research effort. For a particular data set $d = 1, \ldots, D$, let $\boldsymbol{\xi}_d = \{\xi_{d1}, \ldots, \xi_{du}\}$ denote the sets of distinguished compartments. We then define the presence of ($\boldsymbol{x}(\xi_{du}) = 1$) or absence ($\boldsymbol{x}(\xi_{du}) = 0$) in set $\xi_{du}, u = 1 \ldots, U$, as

$$\boldsymbol{x}(\xi_{du}) = \min\left(1, \sum_{c \in \xi_{du}} x_c\right). \tag{2.3}$$

Let us further denote by R_{sm} the probability of discovery of metabolite m in species s such that

$$\mathbb{P}(L_{ms} | \boldsymbol{x}(\xi(m,s)), R_{ms}) = \begin{cases} 0 & \text{if } \boldsymbol{x}(\xi(m,s)) = 0 \text{ and } L_{ms} = 1, \\ 1 & \text{if } \boldsymbol{x}(\xi(m,s)) = 0 \text{ and } L_{ms} = 0, \\ R_{ms} & \text{if } \boldsymbol{x}(\xi(m,s)) = 1 \text{ and } L_{ms} = 1, \\ 1 - R_{ms} & \text{if } \boldsymbol{x}(\xi(m,s)) = 1 \text{ and } L_{ms} = 0, \end{cases} \tag{2.4}$$

where $\xi(m,s)$ is the set of compartments relevant for metabolite m and species s, i.e. all compartments c for which $\tau_{\mathcal{M}}(c) = m$ and $\tau_{S}(c) = s$.

To quantify the research effort R_{ms} of a particular entry L_{ms}, we will rely on two measures, the total number of relevant papers published for metabolite m (P_m) and for species s (Q_s), such that

$$R_{ms} = 1 - e^{-\gamma P_m - \delta Q_s} \tag{2.5}$$

with positives scalars γ and δ. In Figure 2.3 we show a Directed Acyclic Graph (DAG) of the proposed model.

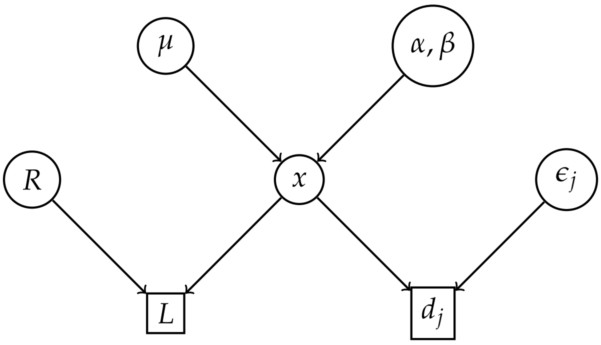

Figure 2.3: In the Directed Acyclic Graph (DAG) representing the *naive model*, the variable x denotes the binary state of a molecule's presence or absence within a designated species and, μ signifies the mean presence of the metabolite along a specified axis. Both α and β denote the axis-specific intercept and a positive scalar constant, respectively. The parameter R stands for the dedicated research effort given to find x while L indicates the presence or absence of x within the LOTUS database. The term d_j characterizes the j^{th} iteration of mass spectrometry executed for the particular species. Finally, ϵ_j quantifies the affiliated error rates, as elaborated in Equation (2.6).

Mass spectrometry

Let $\boldsymbol{d}_{sj} = (d_{sj1}, \dots, d_{sjM})$ be the presence-absence vector of each metabolite m obtained with mass-spectrometry run $j = 1, \dots, J_s$ performed on species s. Assuming a false-positive and false-negative error rates ϵ_{01} and ϵ_{10}, respectively, we have

$$\mathbb{P}(\boldsymbol{d}_{sj} | \boldsymbol{x}, \epsilon_{01}, \epsilon_{10}) = \prod_m \left[x_{sm} \left(\epsilon_{10}^{1-d_{sjm}} (1-\epsilon_{10})^{d_{sjm}} \right) + (1 - x_{sm}) \left(\epsilon_{01}^{d_{sjm}} (1-\epsilon_{01})^{1-d_{sjm}} \right) \right] \tag{2.6}$$

Equation (2.6) has not been developed further. For our most recent approach of modelling MS data, refer to Section 2.2.1.

2.2 Random Markov Field

As previously stated, our objective is to infer the occurrence or absence of metabolites across a collection of samples, which are differentiated by discrete dimensions such as species, tissue type, and environmental conditions or any other arbitrary dimension. We hypothesize that the distribution pattern of these metabolites is moderated by shared characteristics within each dimension. For instance, metabolites can exhibit a similar distribution across phylogenetically close species, or if their synthesis pathways are interrelated. To quantitatively represent such similarities, we use a Markov random field approach [21,22].

Let D denote the total number of dimensions. Without any loss of generality, we assume the first dimension corresponds to the metabolite. Each dimension, denoted by $d = 1, \ldots, D$, consists of a set $\mathcal{E}_d$ of discrete entities (e.g., individual species along the species dimension). We model similarities between entries of dimension d using a Markov process along a known tree $\mathcal{T}_d$ consisting of $\mathcal{N}_d = \mathcal{E}_d \cup \mathcal{R}_d \cup \mathcal{I}_d$ nodes, of which the entries $\mathcal{E}_d$ are leaves, connected to the set of roots $\mathcal{R}_d$ through a set $\mathcal{I}_d$ of internal nodes. We thus have $\mathcal{E}_d \cap \mathcal{R}_d = \varnothing$, $\mathcal{E}_d \cap \mathcal{I}_d = \varnothing$ and $\mathcal{R}_d \cap \mathcal{I}_d = \varnothing$. For every node $n \in \mathcal{N}_d, n \notin \mathcal{R}_d$ that is not a root, we denote $p(n) \in \mathcal{N}_d$ its parent node and $b(n) \geq 0$ the length of the branch connecting it to its parent.

We denote $\mathcal{X}$ a Markov Random Field of which each variable $x \in \mathcal{X}$ represents a unique combination of nodes from each dimension D, indicating the presence ($x = 1$) or absence ($x = 0$) of a metabolite. Let $\delta_d(x) \in \mathcal{N}_d$ reflect the node of x in dimension d with $\delta_1(x)$ indicating the metabolite of x, and let $\delta(x) = (\delta_1(x), \ldots, \delta_D(x))$. We only consider two sets of variables: 1) the set $\mathcal{Y}$ of variables representing an entry in each dimension such that for a variable $y \in \mathcal{Y}$, $\delta_d(y) \in \mathcal{E}_d$ for all $d = 1, \ldots, D$, and 2) the set $\mathcal{Z}$ of variables representing leaves in all dimensions except one such that for a variable $z \in \mathcal{Z}$, $\delta_k(z) \in \mathcal{I}_k$ and $\delta_d(z) \in \mathcal{E}_d$ for all $d \neq k$. We then have $\mathcal{X} = \mathcal{Y} \cup \mathcal{Z}$ and $\mathcal{Y} \cap \mathcal{Z} = \varnothing$.

We suppose that the joint density of $\mathcal{X}$ can be factorized over a set of cliques $\mathcal{C}$. Each clique $c \in \mathcal{C}$ consist of a set of variables $x_1, x_2, \ldots \in \mathcal{X}$ that represent the same leaves in all but one dimension k. Specifically, for all $x \in c$, $\delta_d(x) \in \mathcal{E}_d$ for all $d \neq k$ and $\delta_k(x) \in \mathcal{N}_k$, and for all $x_i, x_j \in c$, $\delta_{-k}(x_i) = \delta_{-k}(x_j)$, where $\delta_{-k}(x)$ denotes the vector of nodes of x in all dimensions but k. For such a clique, we will refer to the dimension $\nu(c) = k$ as its *variable* dimension and will denote by $\delta_{-\nu(c)}(c)$ the vector of nodes in the *fixed* dimensions. By definition, $\delta_{-\nu(c)}(c) = \delta_{-\nu(c)}(x)$ for every $x \in c$.

We will further denote by $\mathcal{C}_k \subset \mathcal{C}$ the subset of cliques that share the variable dimension k, i.e. $\nu(c) = k$ for all $c \in \mathcal{C}_k$. Note that each clique is in exactly one subset ($\mathcal{C}_k \cap \mathcal{C}_d = \varnothing$ for all $k \neq d$) and cliques of the same subset do not share any variables ($c_1 \cap c_2 = \varnothing$ for all $c_1, c_2 \in C_k$). However, each variable $x \in \mathcal{Y}$ will be part of exactly one clique from each subset: the clique $c \in \mathcal{C}_k$ for which $\delta_{-k}(c) = \delta_{-k}(x)$. In contrast, each variable $x \in \mathcal{Z}$ will be part of exactly one clique: the clique $c \in \mathcal{C}$ for which $\delta_{-\nu(c)}(c) = \delta_{-\nu(c)}(x)$ and $\delta_{\nu(c)}(x) \in \mathcal{I}_{\nu(c)}$.

The joint density of $\mathcal{X}$ factorizes as

$$\mathbb{P}(\mathcal{X}) = \prod_{d=1}^{D} \prod_{c \in \mathcal{C}_d} \phi(c), \tag{2.7}$$

where we model the clique functions $\phi(c)$ using a Markov model along tree $\mathcal{T}_d$. Let

$$\boldsymbol{\Lambda}_c = \begin{pmatrix} -\mu_{c1} & \mu_{c1} \\ \mu_{c0} & -\mu_{c0} \end{pmatrix} \tag{2.8}$$

be the rate matrix for changes between states 0 and 1 along the tree. For each node $n \in \mathcal{N}_d, n \notin \mathcal{R}_d$ that is not a root, the transition probabilities between parent node $p(n)$ and n are then given by

$$\boldsymbol{P}(n) = \exp(\boldsymbol{\Lambda}_c b(n)). \tag{2.9}$$

We assume the root state probabilities are given by the stationary distribution of the Markov chain:

$$\boldsymbol{P}_\infty = \left(\frac{\mu_{c0}}{\mu_{c0} + \mu_{c1}}, \frac{\mu_{c1}}{\mu_{c0} + \mu_{c1}} \right). \tag{2.10}$$

The clique function $\phi(c)$

$$\phi(c) = \prod_{x \in c,} \Big(\mathrm{I}(x \in \mathcal{R}_{\nu(c)})[\boldsymbol{P}_\infty]_x + \mathrm{I}(x \notin \mathcal{R}_{\nu(c)})[\boldsymbol{P}(\delta_{\nu(c)}(x))]_{p_c(x),x} \Big) \tag{2.11}$$

where we used the shorthand $x \in \mathcal{R}_{\nu(c)}$ for $\delta_{\nu(c)}(x) \in \mathcal{R}_{\nu(c)}$ to indicate whether the node in the variable dimension of c of x is a root and $p_c(x)$ to identify the variable $z \in c$ for which $\delta_{\nu(c)}(z) = p(\delta_{\nu(c)}(x))$.

Figure 2.4 shows two examples of simple Markov Random Field. The Ising model, as shown in Figure 2.4a is based on the idea that a node's state depends only on its close neighbours *i.e.* it is conditionally independent given the neighbours' states. On the other hand, our suggested model which resembles more to the structure show in Figure 2.4b, includes more complex connections within a graph. Indeed, we hypothesize that a node's state not only depends on its direct neighbours but also by remote connections and the architectural configurations of the trees $\mathcal{T}$.

2.2.1 Data sources

We model the data sources similarly as in the *naive model*.

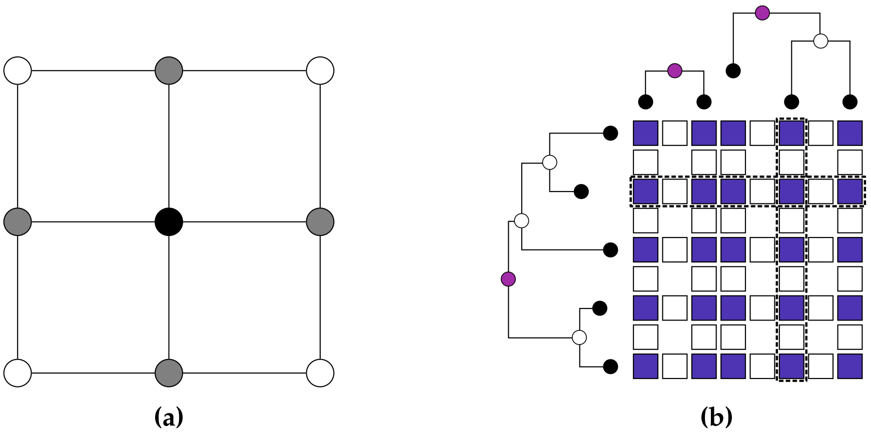

Figure 2.4: Two examples of a Markov Random Field model. **(a)** The Ising model [23] characterizes nearest-neighbour interactions. In this model, given the state of the grey nodes, the central black node becomes conditionally independent of all external nodes. **(b)** An *advanced* structural representation, incorporating higher-order interactions beyond the traditional Ising model. Interactions extend beyond immediate neighbours, encompassing higher-order relationships as described by the trees $\mathcal{T}$. Here, pink dots represent the roots $\mathcal{R}$, in white the internal nodes $\mathcal{I}$ and in black the leaves $\mathcal{E}$. The purple squares represent the $\mathcal{Y}$ and white squares the $\mathcal{Z}$ variables of our model. Dashed rectangles represent examples of cliques. This illustration was conceptualized based on insights and frameworks derived from [24] and [25].

LOTUS

We model the probability of LOTUS similarly as in the *naive model*. However we adapt Equation (2.3) as follows.

Let $x(m,s)$ denote the variable in $\mathcal{X}$ for NP m and species s, which, in case $\mathcal{X}$ contains additional dimensions, is obtained by collapsing:

$$x(m,s) = \min\left(1, \sum_{x\in\mathcal{X}} \mathrm{I}(\delta_1(x) = m)\,\mathrm{I}(\delta_2(x) = s)x\right). \tag{2.12}$$

Probabilities of LOTUS and research effort remain the same as in Equations (2.4) and (2.5).

Mass spectrometry

Ultra High Performance Liquid Chromatography coupled to fragmentation Mass Spectrometry (LC-MSMS) is the analytical workhorse for the molecular characterization of complex biological matrices. Coupled to computational mass spectrometry tools, such untargeted approaches allows to detect and putatively annotate thousands of NPs in a single run. These analysis are fundamental in our project as they will allow us to complete the currently patchy overview offered by LOTUS. However, the resulting data is complex and requires careful processing to be informative as it is high-dimensional, noisy and incomplete. We here propose a model that builds on previous work regarding the Liquid Chromatography (LC) process [26, 27], the establishment of virtual mass spectrometers [28–30] and the integration of LC and MS dimensions in the NP annotation process [31,32], yet is simpler and more streamlined to render it computationally feasible for the large scales considered here.

Suppose $\mathcal{D}_i$ is an LC-MSMS profile obtained from a sample representing a specific vector $\boldsymbol{\xi} = (\xi_{i2}, \ldots, \xi_i D), \xi_{id} \in \mathcal{E}_d$ of leaves in all dimensions except NPs one, such as, for instance, a sample representing a specific tissue of a specific species. We will calculate the probability of the LC-MSMS data $\mathcal{D}_i$ given $\mathcal{X}$. Let $\boldsymbol{x}(\boldsymbol{\xi}_i) \subset \mathcal{X}$ denote a slice through $\mathcal{X}$ relevant for $\mathcal{D}_i$, *i.e.* consisting of all variables that represent a leave in the metabolite dimension and the specific leaves $\boldsymbol{\xi}$ in each other dimension such that for all $x, x' \in \boldsymbol{x}(\boldsymbol{\xi}_i)$, $\delta_1(x) \in \mathcal{N}_1$, $\delta_1(x) \neq \delta_1(x')$ and $\delta_{d \neq 1}(x) = \delta_{d \neq 1}(x') = \xi_{id} \in \mathcal{E}_d$. We will develop an NPs-flavored **V**irtual **M**etabolomics **M**ass **S**pectrometer (**ViMMS**) building on the original implementation [28]. It will be fed by an in silico spectral database of the last LOTUS contents (ISDB-LOTUS) and informed by prior expert knowledge regarding the classes of analytes lost in the reductionist and stochastic metabolomics approach here formed by *extraction*, *liquid chromatography* and *mass spectrometry fragmentation* stages. The NPs-ViMMS will allow the generation of theoretical metabolomics datasets for any given input ($\mathcal{D}'_i$), these will be then be compared to experimental results ($\mathcal{D}_i$).

To compare the resulting LC-MSMS profile $\mathcal{D}'_i$ to $\mathcal{D}_i$, we will then take advantage of **MEMO** (MS2 BasEd SaMple VectOrization), a method we recently established for the computationally efficient comparison of large sets of samples based on their LC-MSMS profiles [33]. The first step is to extract fragment ions and neutral losses from each MSMS spectrum binned from the detected features in so-called “documents” using Spec2Vec [34]. Then, for a given sample, all the documents created are aggregated based on word occurrences to form a fingerprint (a MEMO vector). The MEMO strategy exploits the advantages of LC, namely its separation power (thus simplifying the chemical complexity of the sample being analyzed and allowing resolution of isomerisms) while avoiding the disadvantages of RT-based alignment since MEMO vectors contain only mass spectrometry information. Here we'll implement a stochastic comparison of MEMO vectors $\mathcal{M}(\mathcal{D}'_i)$ and $\mathcal{M}(\mathcal{D}_i)$ using a per entry error rate ϵ to be inferred from the data.

2.3 Graph Neural Network

The low-dimensional representation of nodes within large graphs plays a critical role in various domains of scientific research and industrial applications, such as bioinformatics, social networks, and content recommendation systems. The utilization of these embeddings has proven effective in diverse prediction tasks, including clustering, node classification, and link prediction. However, traditional methods for generating these embeddings have predominantly focused on the *transductive* setting, requiring all nodes to be present during training and thus limiting generalization to unseen nodes or entirely new subgraphs [35,36].

GraphSAGE (SAmple and aggreGatE) [37] was presented as a solution to this challenge, offering a general *inductive* framework that leverages both node feature information and topological structure. Unlike transductive approaches, which rely on matrix factorization and are constrained to fixed graphs, GraphSAGE is designed to efficiently generate embeddings for previously unseen nodes.

The novelty of GraphSAGE lies in its ability to learn a function that generates embeddings through the sampling and aggregation of features from a node's local neighbourhood. It utilizes a set of trainable aggregator functions that encapsulate information from different search depths (hops), away from a given node. By simultaneously learning the topological structure and distribution of node features in the neighbourhood, GraphSAGE accommodates feature-rich graphs as well as graphs lacking specific node features.

The applicability of GraphSAGE extends beyond simple convolutions, embracing a framework that generalizes Graph Convolutional Networks (GCNs) for the task of inductive unsupervised learning [38]. Unlike traditional methods that optimize embeddings for each node, GraphSAGE's inductive approach promotes efficiency and adaptability, allowing for an easy alignment of newly observed subgraphs with pre-existing node embeddings.

GraphSAGE is particularly well-suited for the task of predicting which molecule is present in which species due to its robust inductive learning framework that generalizes to *unseen* nodes and subgraphs. In the context of biological data, such as molecular structures and species interactions, GraphSAGE's ability to leverage both the topological structure and node feature information offers a powerful means to understand the complex relationships within a graph. Its approach of sampling and aggregating features from a node's local neighbourhood enables the capture of intricate patterns and structural properties that can be essential in identifying molecular presence across species. Furthermore, the inductive nature of GraphSAGE allows for the efficient generalization across different phylogenies, facilitating the prediction in entirely new or evolving graphs.

HinSAGE [39], a derivative of GraphSAGE, has been specifically designed to handle heterogeneous graphs, where nodes and edges can be of various types. Developed by CSIRO's Data61, HinSAGE adeptly extends the foundational principles of GraphSAGE to contexts where the graph's heterogeneity introduces additional complexities. Unlike homogeneous graphs where the relation between nodes is more uniform, heterogeneous graphs present varying relationships and patterns, which HinSAGE is explicitly tailored to capture. By learning distinct embeddings for different types of nodes and relations, HinSAGE can uncover nuanced relationships within complex networks. This makes HinSAGE especially valuable for tasks such as predicting links within a bipartite graph, where one set of nodes represents species and another molecules.

3 Methods

3.1 Naive model

Prior to the application of our actual data, a series of simulations were executed to evaluate the feasibility of estimating the entire set of parameters from the information contained within our dataset. Specifically, the variables μ were generated by sampling from a normal distribution with mean value of 0 and variance of 1. Meanwhile, the parameters α, β, γ, and δ were each modelled using distinct exponential distributions, where individual values for the rate parameter λ were attributed to each. In order to replicate the observed phenomenon, the number of papers per entry in x were synthetically constructed by drawing from a Poisson distribution. Additionally, the variables σ were simulated by drawing from a Wishart distribution [40].

As elaborated in Section 2.1, the simulation process was initiated by drawing probabilities that $x = 1$ from the *expit* function, as defined in Equation (2.1). Samples were then drawn from a Bernoulli distribution, where the probability parameter was informed by the previous *expit* function. The probabilities of LOTUS were constructed in accordance with Equations (2.4) and (2.5). A condition was imposed such that if x for any given pair was 0, then the corresponding probability was explicitly set to 0. This removed the possibility of having papers in the simulated LOTUS when the actual value of x was 0.

The probabilities of LOTUS were then employed as parameters for another Bernoulli sampling, generating binary outcomes that determined the number of papers associated with each pair. Specifically, if the result was 0, the number of papers for that particular pair was set to 0. Conversely, if the result was 1, the number of papers for that pair was assigned based on random Poisson values that had been drawn previously in the simulation process.

This systematic approach resulted in the production of a simulated x and a corresponding simulated LOTUS. This led to the occurrence of certain pairs that appearing empty, even though the molecule was indeed present within the species. All codes are available on GitHub.

3.2 Random Markov Field

Due to time constraints of the thesis, test and simulation for this model were not performed. The beginning of the implementation is available both on Bitbucket and GitHub.

3.3 Graph Neural Network

The LOTUS database was aggregated to include only unique pairs of molecules and species. Once aggregated, the data was randomly partitioned into two subsets: 70% allocated for training and 30% for testing.

Graphs were systematically constructed for both the training and testing subsets using the software library NetworkX v3.1 [41]. In these graphs, individual nodes were designated to represent each molecule and species. When a specific species-molecule pair was identified in the LOTUS database, a directed edge was drawn between the two corresponding nodes. This procedure led to the creation of a bipartite graph, with directed edges labeled as "has" from species to molecules and "present in" from molecules to species.

The species' features were defined by extracting their phylogenetic information through the GBIF API [42, 43]. Molecules' features were composed of the combinaiton of their classification data from Classyfire [44] and their Morgan Fingerprint [45], encoded using a 128-bit representation and a radius of 2.

In the preprocessing stage, features corresponding to both species characteristics and molecules' Classyfire properties were transformed through binary encoding. This transformation was essential to represent these categorical attributes as numerical values, thus making them suitable as features for the nodes within the graph. In the case of the molecules, the two different sets of features, namely the binary-encoded Classyfire attributes and the Morgan Fingerprint, were concatenated to form a unified feature vector.

The model training was carried out using the Stellargraph library [39]. Two distinct models were trained to handle different relationships within the graph; one was tailored to the edges labeled "has" and the other to the edges described as "present in".

The HinSAGE models were configured with two layers, comprising 1024 neurons in each layer. The first layer was designed with a neighborhood sampling size of 3, enabling the model to encapsulate the local structural information, while the second layer utilized a sampling size of 1, thus focusing on immediate neighbors. By employing this hierarchical structure, the models could capture different scales of locality in the graph.

Furthermore, the HinSAGE models were implemented with a mean aggregator function, which served to combine the features of the neighboring nodes, thus generating a representative feature vector for each target node. A dropout rate of 0.3 was applied to mitigate the risk of overfitting, and "elu" and "selu" activation functions were utilized in the respective layers.The number of layers, neurons per layer, hops, dropout rate, and activation functions were all chosen based on a comprehensive grid search for optimal parameters.

We moved forward by attempting to anticipate every potential species-molecule combination found in LOTUS. This analysis resulted in a total of $5.45 \cdot 10^9$ pairs. All codes are available on Github.

4 Results and Discussion

4.1 Naive model

In the preliminary stages of our research, we recognized the necessity to understand the behaviour of our model prior to applying it to the actual dataset. To achieve this, we carried out a series of simulations to assess if the model's parameters could be accurately estimated based on these synthetic data.

Specifically, we simulated 100 molecules and 10 species in alignment with the theoretical framework described by Equations 2.4 and 2.5.

For the parameter estimation process, we employed Markov Chain Monte Carlo (MCMC) techniques to accurately estimate the parameters γ and δ. The results of this estimation process were consistent and close to our simulated values, demonstrating the effectiveness of the approach.

Furthermore, we utilized Gibbs sampling to estimate the variable x. This method too yielded satisfactory results, corroborating the validity of our model in this aspect.

However, the challenges encountered during the modelling and simulation process were primarily centred around the convergence of the axis-specific intercepts μ. This crucial component, detailed in Equation 2.1, resisted precise estimation through our initially chosen techniques. During this period of reevaluation, the idea of seeing our data as a graph emerged, adding a new dimension to our perspective. Persisting with our original data treatment no longer seemed intuitive. Given the inherent structure and relationships in our dataset, transitioning to a graph-based approach felt more logical and natural. As a result, the inability to accurately estimate μ and the allure of a graph-centric methodology prompted us to explore alternative models and techniques, seeking a better alignment with the intrinsic characteristics of our data. To reproduce our simulations, codes are available on GitHub.

4.2 Random Markov Field

Due to the time constraints of the thesis, we were unable to implement, test, and simulate this model. However, we have designed the model (see Section 2.2), and its implementation is our current focus. We would like to emphasize that so far, this is the best model we currently possess.

4.3 Graph Neural Network

Using unseen edge data to assess the models, the performance measures showed that each model had varying degrees of accuracy. Particularly, the 0.92 accuracy was demonstrated by the model that was trained to predict the "present in" relationships. The model that attempted to predict the "has" associations, in contrast, had an accuracy of 0.8.

For evaluating the accuracy of the models, a probability value of 0.5 was used as the cutoff for assessing whether metabolites were present or absent. As a result, probability above this cutoff were classed as a presence, denoted as $x = 1$, and values below this cutoff were classified as an absence, denoted by $x = 0$.

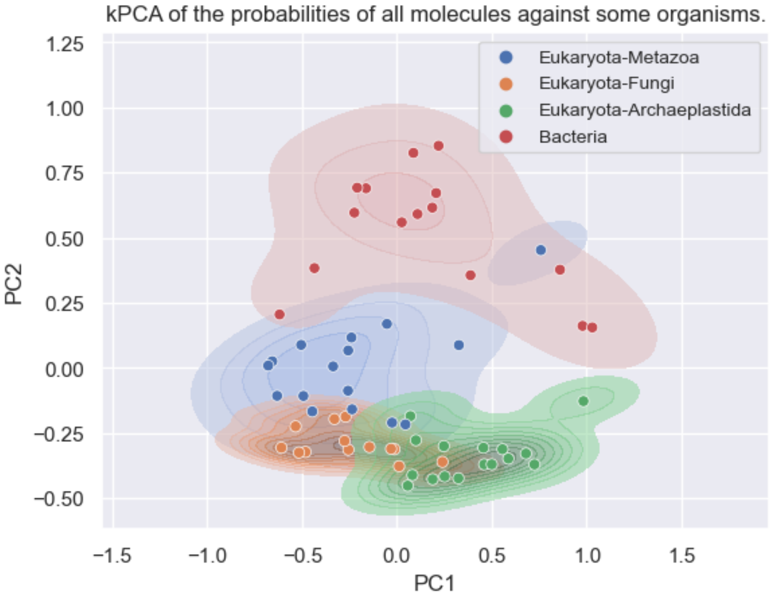

Figure 4.1: kPCA visualization of sampled species across primary biological domains. Archaea were not sampled due to lack of data.

After predicting the entirety of the LOTUS database, we randomly sampled a dozen species from each primary biological domain, excluding Archaea [1]. Each of these species was associated with 148′190 probability values. A kernel Principal Component Analysis (kPCA) was then carried out to identify potential variance between the domains. Kernel Principal Component Analysis (kPCA), as opposed to classic Principal Component Analysis (PCA), became the preferred approach given the high-dimensionality of our dataset. The advantages of kPCA over PCA are its skill at handling large data dimensions and its ability to recognise non-linear correlations between features that PCA's linear assumptions might miss. This innate ability makes sure that kPCA keeps the most important and complex relationships within the data as well as reducing the dimensions of the data.

[1] Due to an insufficient number of species.

A clear distinction between the biological domains can be seen by looking at Figure 4.1. To explain these observed variances, two hypotheses have been postulated. First, our Graph Neural Network (GNN) may be capable of distinguishing the molecular signatures that are unique to each domain. Alternatively, it is possible that the kPCA is mostly displaying the chemical biases included in the LOTUS database. Given that the LOTUS database organises its data according to triples of molecules-species-papers, it is reasonable to think that a majority of researched molecules, meriting scholarly publication, are specific to particular domains i.e are specialized metabolites. This claim is supported by *Rutz et al.* [2], who note that a significant majority (more than 90%) of the compounds included in LOTUS display domain specificity. The patterns in the kPCA results that have been found may be explained by such innate biases. Either of the previous hypothesis has to be supported by a more in-depth examination.

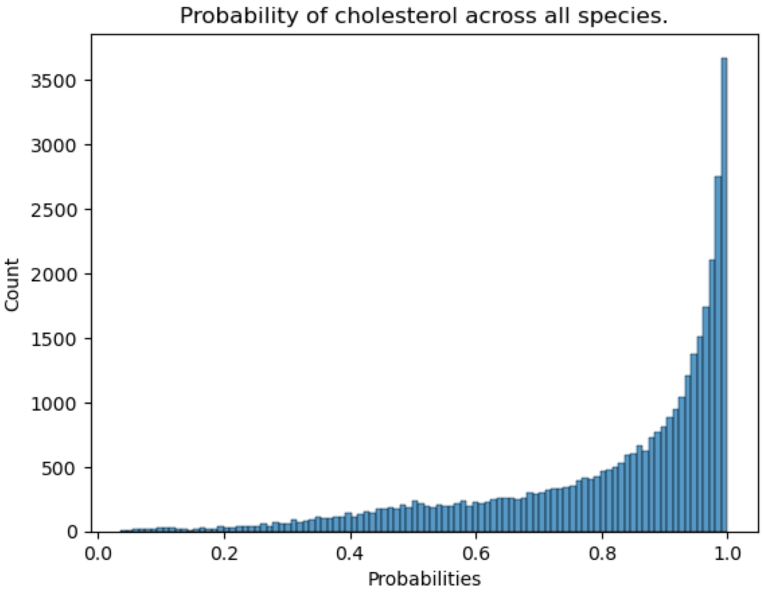

Figure 4.2: Probability distribution of cholesterol across species in the LOTUS database. The species already associated with cholesterol were removed before the calculations.

In Figures 4.2 and 4.3, we present the probability distributions of cholesterol and erythromycin across the species present in LOTUS. Based on the literature [46, 47], cholesterol is ubiquitously found. The predictions of our algorithm align with this notion, indicating a prevalent cholesterol presence in most species. This predictive consistency is attributed to the mechanics of GraphSAGE. Given that cholesterol is related to 522 distinct species in the LOTUS database, the algorithm is expected to recognise the broad distribution of cholesterol across a diverse range of species, leading it to predict its prevalence in the majority of the additional species.

Erythromycin, on the other hand, has a fairly low representation in the database, being linked to only eight species. In particular, these links primarily concern bacteria from the Actinobacteria phylum. As a result, our model tends to predict a low occurrence of erythromycin across species, as visualized in Figure 4.3. However, the accuracy with which it anticipates its presence is noteworthy. Key genera such as Streptomyces are correctly identified, with minor discrepancies like Micromonospora, as detailed in Table 4.1. This highlights GraphSAGE's capability in discerning potential associations between molecules and species, even in scenarios of sparse data.

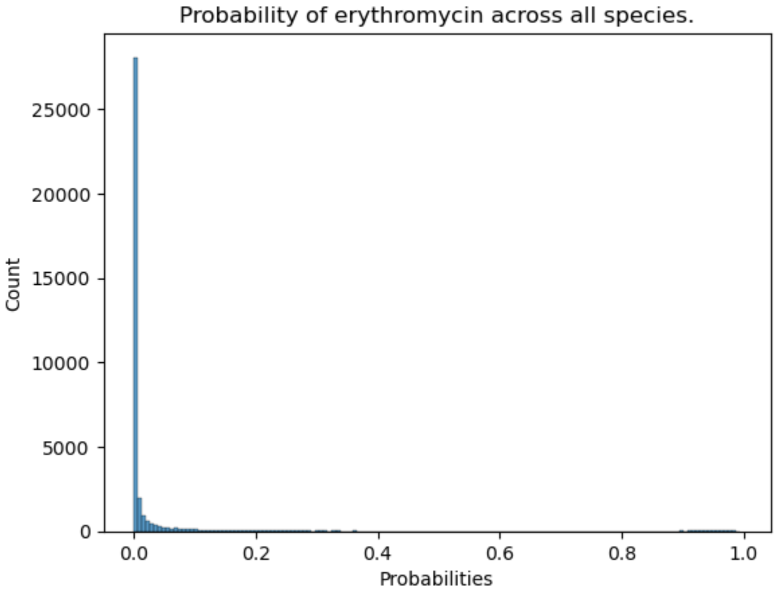

Figure 4.3: Probability distribution of erythromycin across species in the LOTUS database. The species already associated with erythromycin were removed before the calculations.

While GraphSAGE has shown notable abilities in making nuanced predictions,our ability to verify the occurrence of molecules in species has limitations. For instance, in the case of *Streptomyces diastaticus* presented in Table 4.1, a lack of documentation regarding its erythromycin production exists in the current literature. On the other hand, the potential presence of this molecule in *Streptomyces achromogenes* is reported in the research presented by [48].

A thorough analysis using mass spectrometry of the highlighted species becomes necessary to confirm the reliability and effectiveness of our approach. This could reveal new molecular species relationships and provide empirical support for the hypothesised associations presented in Table 4.1.

An inherent bias in favour of specialised metabolites represents another important constraint. Understanding the existence of basic metabolites like water frequently does not attract academic curiosity and suffers from publication biases. Predictions are then difficult since the research output rarely associates such a chemical with its related species. This is demonstrated by the LOTUS database, which only

Table 4.1: Erythromycin was originally discovered in *Saccharopolyspora erythraea* [49]. Here we show the top 10 species that are associated to erythromycin according to our model. In most new predictions, no evidence of erythromycin has been found in the literature. This highlights the importance of MS analysis in order to quickly verify their presence.

Species	Probability	Presence
Streptomyces diastaticus	0.9944	Needs investigation
Streptomyces drozdowiczii	0.9939	Needs investigation
Streptomyces antibioticus	0.9928	Needs investigation
Micromonospora	0.9927	Needs investigation
Streptomyces achromogenes	0.9924	Potential presence [48]
Streptomyces griseus	0.9919	Needs investigation
Streptomyces griseosporeus	0.9912	Needs investigation
Streptomyces albogriseolus	0.9906	Needs investigation
Streptomyces varsoviensis	0.9899	Needs investigation
Streptomyces ansochromogenes	0.9898	Needs investigation

records six instances of water, leading to poor predicative results as seen in Figure 4.4.

Additional complexities exist that limit the effectiveness of our model. In the training stage of GraphSAGE, the algorithm only samples negative edges from edges that do not exist in the graph. This methodology could introduce potential anomalies; certain molecules may in fact be associated with distinct species, but the algorithm is predisposed to miss such associations. The algorithm's indifference for phylogenetic distances is a serious matter as well. Although evolutionary information is encoded in the node features, the general architecture of the graph lacks it, which unintentionally leaves out important information. This shortcoming might affect the precision of inferring occurrences in particular phylogenies. Excluding phylogenies from the graph presents an other challenge. As previously mentioned, one of GraphSAGE's strengths is that it doesn't necessitate the inclusion of all nodes during the training process, thereby facilitating the addition of new nodes and theoretically maintaining prediction quality. However, in our specific application, since the context of each node is encoded within its features, introducing a new node will result in a node with node edges (singleton). While this new node might share features with existing nodes, its lack of connections diminishes the quality of edge predictions. Such a node will be missing the crucial contextual information that the algorithm relies upon for accurate predictions.

Lastly, the current model configuration does not factor in the research effort, as delineated in Equation (2.4). This omission could lead to biases, potentially comparing molecules with considerable documentation in a given species to ones with less characterization in a relatively unknown species.

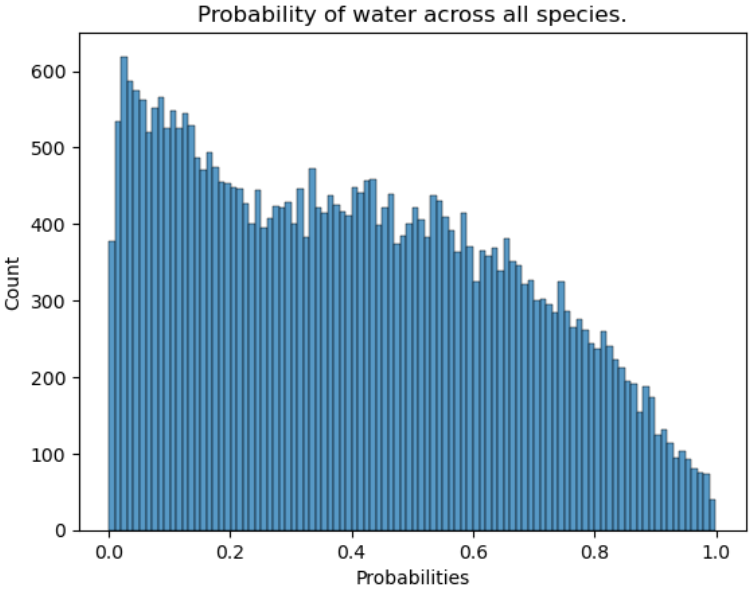

Figure 4.4: Probability distribution of water across species in the LOTUS database. The species already associated with water were removed before the calculations. This visualization sheds light on a limitation of our model. It's a reasonable expectation that water is a ubiquitous component in a vast majority of species. However, with only six mentions of water in LOTUS, our predictions seem to fall short.

5 Conclusion and Future Work

In the course of this work, we have demonstrated the viability of developing a model to predict the chemical compositions of organisms across the tree of life. By employing an approach that structures molecules and species as graphs, we have been able to make better predictions compared to the more basic model.

The application of a Graph Neural Network (GNN) employing the GraphSAGE technique has proven effective as an initial strategy. Notably, this modest implementation already yields reliable results for well-characterized metabolites. Nonetheless, several challenges emerged. Foremost among these is the technical issue of ignoring phylogenies, which consequently ignores useful and available data. Additionally, the methodology employed for encoding molecular features, specifically the straightforward concatenation of Classyfire classification and Morgan fingerprint, may not be the ideal representation of the molecule's information.

Another significant constraint of the GNN is its inability to generalize its predictions. Indeed, given the sparse nature of the existing data, the algorithm might struggle to accurately represent the true intricate relationships and dependencies that exist between nodes within the graph. Its predictions may then lack substantial significance. Simply extrapolating LOTUS with this methodology would result in predictions of "obvious" edges without providing further insights. We thus believe that this does not address the question at stake, *i.e.* predicting a species' metabolome. Additionally, the GNN currently cannot incorporate mass spectrometry data or factor in research efforts. Recognizing and reconciling those discrepancies are crucial considerations in order to improve the GNN's accuracy.

In contrast, given that both GNN and Markov Random Field address analogous issues [12], our hypothesis is that MRF could yield more robust outcomes. The flexibility of our model, allowing for multi-dimensional data integration, could potentially ameliorate the molecule encoding challenge we encountered with GNN. Indeed, given the current configuration of our model, the species phylogenies, the existing Classyfire classification tree, and a tree for the Morgan fingerprint could be integrated, forming three distinct trees. Furthermore, the proposed MRF model could easily include mass spectrometry data, a task that the GNN is now unable to handle. In addition, the model would interpret both LOTUS and mass spectrometry data as noisy observations of $\mathcal{X}$, which the current GNN is also incapable of doing.

To conclude, we believe that the MRF model has the potential to outperform the current existing GNN. We hope that the development of such a model will catalyse advancements in metabolomics, ecology, and drug discovery. More work is required but we want to believe that through this thesis, we have laid the foundations for future development of such a model.

References

[1] All natural. *Nature Chemical Biology*, 3(7):351–351, July 2007. doi:10.1038/nchembio0707-351.

[2] Adriano Rutz, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G Graham, Ralf Stephan, Roderic Page, Jiří Vondrášek, Christoph Steinbeck, Guido F Pauli, Jean-Luc Wolfender, Jonathan Bisson, and Pierre-Marie Allard. The LOTUS initiative for open knowledge management in natural products research. *eLife*, 11:e70780, May 2022. doi:10.7554/eLife.70780.

[3] Alan L. Harvey, RuAngelie Edrada-Ebel, and Ronald J. Quinn. The reemergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery*, 14(2):111–129, February 2015. doi:10.1038/nrd4510.

[4] Yan Yan, Qikun Liu, Steven E Jacobsen, and Yi Tang. The impact and prospect of natural product discovery in agriculture: New technologies to explore the diversity of secondary metabolites in plants and microorganisms for applications in agriculture. *EMBO reports*, 19(11):e46824, November 2018. doi:10.15252/embr.201846824.

[5] Susana González-Manzano and Montserrat Dueñas. Applications of Natural Products in Food. *Foods*, 10(2):300, February 2021. doi:10.3390/foods10020300.

[6] Ji-Kai Liu. Natural products in cosmetics. *Natural Products and Bioprospecting*, 12(1):40, December 2022. doi:10.1007/s13659-022-00363-y.

[7] Pavel B. Drasar and Vladimir A. Khripach. Growing Importance of Natural Products Research. *Molecules*, 25(1):6, December 2019. doi:10.3390/molecules25010006.

[8] Carlos E. Rodríguez-López, Yindi Jiang, Mohamed O. Kamileen, Benjamin R. Lichman, Benke Hong, Brieanne Vaillancourt, C. Robin Buell, and Sarah E. O'Connor. Phylogeny-Aware Chemoinformatic Analysis of Chemical Diversity in Lamiaceae Enables Iridoid Pathway Assembly and Discovery of Aucubin Synthase. *Molecular Biology and Evolution*, 39(4):msac057, April 2022. doi:10.1093/molbev/msac057.

[9] Richard Johnsonbaugh. *Discrete Mathematics*. Pearson, New York, NY, eighth edition edition, 2018.

[10] Richard J. Trudeau. *Introduction to Graph Theory*. Dover Books on Advanced Mathematics. Dover Pub, New York, 1993.

[11] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, January 2015. doi:10.1093/nar/gku1003.

[12] Junteng Jia, Cenk Baykal, Vamsi K. Potluru, and Austin R. Benson. Graph Belief Propagation Networks. 2021. doi:10.48550/ARXIV.2106.03033.

[13] James Atwood and Don Towsley. Diffusion-convolutional neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/390e982518a50e280d8e2b535462ec1f-Paper.pdf.

[14] Qimai Li, Zhichao Han, and Xiao-ming Wu. Deeper Insights Into Graph Convolutional Networks for Semi-Supervised Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. doi:10.1609/aaai.v32i1.11604.

[15] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 556–559, New Orleans LA USA, November 2003. ACM. doi:10.1145/956863.956972.

[16] Muhan Zhang and Yixin Chen. Link Prediction Based on Graph Neural Networks. 2018. doi:10.48550/ARXIV.1802.09691.

[17] Lada A Adamic and Eytan Adar. Friends and neighbors on the Web. *Social Networks*, 25(3):211–230, July 2003. doi:10.1016/S0378-8733(03)00009-1.

[18] Lingfei Wu, Peng Cui, Jian Pei, and Liang Zhao, editors. *Graph Neural Networks: Foundations, Frontiers, and Applications*. Springer, Singapore, 2022. doi:10.1007/978-981-16-6054-2.

[19] Tolutola Oyetunde, Muhan Zhang, Yixin Chen, Yinjie Tang, and Cynthia Lo. BoostGAPFILL: Improving the fidelity of metabolic network reconstructions through integrated constraint and pattern-based methods. *Bioinformatics*, 33(4):608–611, February 2017. doi:10.1093/bioinformatics/btw684.

[20] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective Classification in Network Data. *AI Magazine*, 29(3):93–106, September 2008. doi:10.1609/aimag.v29i3.2157.

[21] David Sherrington and Scott Kirkpatrick. Solvable Model of a Spin-Glass. *Physical Review Letters*, 35(26):1792–1796, December 1975. doi:10.1103/PhysRevLett.35.1792.

[22] Ross Kindermann and J. Laurie Snell. *Markov Random Fields and Their Applications*. Contemporary Mathematics ; v. 1. American Mathematical Society, Providence, R.I, 1980.

[23] Ernst Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, February 1925. doi:10.1007/BF02980577.

[24] Peter Orchard. Markov Random Field Optimisation. URL: https://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV0809/ORCHARD/.

[25] Pinar Acar and Veera Sundararaghavan. A Markov random field approach for modeling spatio-temporal evolution of microstructures. *Modelling and Simulation in Materials Science and Engineering*, 24(7):075005, October 2016. doi:10.1088/0965-0393/24/7/075005.

[26] William Heymann, Juliane Glaser, Fabrice Schlegel, Will Johnson, Pablo Rolandi, and Eric Von Lieres. Advanced error modeling and Bayesian uncertainty quantification in mechanistic liquid chromatography modeling. *Journal of Chromatography A*, 1708:464329, October 2023. doi:10.1016/j.chroma.2023.464329.

[27] Paweł Wiczling, Agnieszka Kamedulska, and Łukasz Kubik. Application of Bayesian Multilevel Modeling in the Quantitative Structure–Retention Relationship Studies of Heterogeneous Compounds. *Analytical Chemistry*, 93(18):6961–6971, May 2021. doi:10.1021/acs.analchem.0c05227.

[28] Joe Wandy, Vinny Davies, Justin J. J. van der Hooft, Stefan Weidt, Rónán Daly, and Simon Rogers. In Silico Optimization of Mass Spectrometry Fragmentation Strategies in Metabolomics. *Metabolites*, 9(10):219, October 2019. doi:10.3390/metabo9100219.

[29] Joe Wandy, Vinny Davies, Ross McBride, Stefan Weidt, Simon Rogers, and Rónán Daly. ViMMS 2.0: A framework to develop, test and optimisefragmentation strategies in LC-MS metabolomics. *Journal of Open Source Software*, 7(71):3990, March 2022. doi:10.21105/joss.03990.

[30] Joe Wandy, Ross McBride, Simon Rogers, Nikolaos Terzis, Stefan Weidt, Justin J. J. Van Der Hooft, Kevin Bryson, Rónán Daly, and Vinny Davies. Simulated-to-real benchmarking of acquisition methods in untargeted metabolomics. *Frontiers in Molecular Biosciences*, 10:1130781, March 2023. doi:10.3389/fmolb.2023.1130781.

[31] Eric Bach, Simon Rogers, John Williamson, and Juho Rousu. Probabilistic framework for integration of mass spectrum and retention time information in small molecule identification. *Bioinformatics*, 37(12):1724–1731, July 2021. doi:10.1093/bioinformatics/btaa998.

[32] Eric Bach, Emma L. Schymanski, and Juho Rousu. Joint structural annotation of small molecules using liquid chromatography retention order and tandem mass spectrometry data. *Nature Machine Intelligence*, 4(12):1224–1237, December 2022. doi:10.1038/s42256-022-00577-2.

[33] Arnaud Gaudry, Florian Huber, Louis-Félix Nothias, Sylvian Cretton, Marcel Kaiser, Jean-Luc Wolfender, and Pierre-Marie Allard. MEMO: Mass Spectrometry-Based Sample Vectorization to Explore Chemodiverse Datasets. *Frontiers in Bioinformatics*, 2:842964, April 2022. doi:10.3389/fbinf.2022.842964.

[34] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H. Spaaks, Faruk Diblen, Simon Rogers, and Justin J. J. Van Der Hooft. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2):e1008724, February 2021. doi:10.1371/journal.pcbi.1008724.

[35] Aditya Grover and Jure Leskovec. Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA, August 2016. Association for Computing Machinery. doi:10.1145/2939672.2939754.

[36] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, New York New York USA, August 2014. ACM. doi:10.1145/2623330.2623732.

[37] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. 2017. doi:10.48550/ARXIV.1706.02216.

[38] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. 2016. doi:10.48550/ARXIV.1609.02907.

[39] StellarGraph Machine Learning Library. CSIRO's Data61, 2018. URL: https://github.com/stellargraph/stellargraph.

[40] John Wishart. THE GENERALISED PRODUCT MOMENT DISTRIBUTION IN SAMPLES FROM A NORMAL MULTIVARIATE POPULATION. *Biometrika*, 20A(1-2):32–52, 1928. doi:10.1093/biomet/20A.1-2.32.

[41] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11–15, Pasadena, CA USA, 2008.

[42] GBIF.org. GBIF Home Page, July 2023. URL: https://www.gbif.org/.

[43] Gbif/pygbif. Global Biodiversity Information Facility, July 2023. URL: https://github.com/gbif/pygbif.

[44] Yannick Djoumbou Feunang, Roman Eisner, Craig Knox, Leonid Chepelev, Janna Hastings, Gareth Owen, Eoin Fahy, Christoph Steinbeck, Shankar Subramanian, Evan Bolton, Russell Greiner, and David S. Wishart. ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics*, 8(1):61, December 2016. doi:10.1186/s13321-016-0174-y.

[45] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. doi:10.1021/ci100050t.

[46] Maryadele J O'Neil. *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals*. RSC Publishing, 2013.

[47] International Agency for Research on Cancer. IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans.(20, Suppl. 1) Chemicals and industrial processes associated with cancer in humans. (20, Suppl. 1), 1979.

[48] Fatemeh Moosawi, Hassan Mohabatkar, and Sasan Mohsenzadeh. Computational Prediction of Properties and Analysis of Molecular Phylogenetics of Polyketide Synthases in Three Species of Actinomycetes. *Medicinal Chemistry*, 6(2):100–107, March 2010. doi:10.2174/157340610791321497.

[49] M. Beran, V. Přikrylová, P. Sedmera, J. Novák, J. Zima, M. Blumauerová, and T.Kh. Todorov. Isolation of erythromycin A N-oxide and pseudoerythromycin A hemiketal from fermentation broth of Saccharopolyspora erythraea by thin-layer and high-performance liquid chromatography. *Journal of Chromatography A*, 558(1):265–272, September 1991. doi:10.1016/0021-9673(91)80132-Z.