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

August 23, 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

Natural Products (NPs) are chemical entities biosynthesized by living organisms. Many NPs are metabolites, which can be placed along a specialization gradient from core metabolites that play essential functions and are found in a wide range of organisms to specialized metabolites which are much more restricted across the tree of life.

Specialized metabolites display particular chemical structures and exhibit specific roles and constitute the major part of the current human therapeutic arsenal. Furthermore, their structural characterization and the elucidation of their biological roles are increasingly recognized as fundamental to understanding ecosystems functioning. Their description, however, is not an easy task. Indeed, specialized metabolites are characterized by several levels of complexity. At the scale of a single molecule, their structural complexity explains both their potent biological activities (privileged structures) and their complicated synthetic accessibility. At the organism level, specialized metabolites are found within complex mixtures of extremely diverse chemical classes spanning large dynamic ranges.

Some efforts have been made to anticipate metabolic networks or the occurrences of molecules in selected taxa. However, no model has been proposed to predict their occurrence across the tree of life. The goal of this project is to develop such a model.
TODO

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor. . . TODO

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
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]. Many NPs are metabolites, which can be positioned along a specialization gradient from core metabolites, fulfilling essential functions and found in a wide range of organisms, to specialized metabolites, much more restricted across the tree of life. Natural Products research is a transdisciplinary field with interests spanning from fundamental structural aspects of naturally occurring molecular entities to their effects on living organisms and extending to the study of chemically mediated interactions within entire ecosystems. We will use *Rutz et al.*'s in definition of Natural Product *as any chemical entity found in a living organism*, *i.e.* a structure-organism pair [2]. This pair holds an additional and fundamental element - a reference to the experimental evidence establishing the linkage between the chemical structure and the biological organism

1.2 The Importance of Natural Products in Therapeutics and Ecosystem Functioning

Unique in their chemical structures and functionalities, specialized metabolites serve as the cornerstone of today's therapeutic applications [3]. Their relevance extends beyond human and veterinary medicine, touching on areas such as agriculture [4], the food industry [5], cosmetics [6], and an array of other sectors. Intrinsically linked to renewable resources, these natural products hold substantial value within the circular economy concept.

The intricacies of their biological roles and structural characterization are crucial to fully grasp the functioning of ecosystems. These complexities influence a myriad of aspects, from the impact on individual organisms to the overall chemically mediated interactions within entire ecosystems. Despite their multifaceted nature, which presents significant challenges at the molecular level - particularly due to their powerful biological activities (privileged structures) and intricate synthetic accessibility the role of these metabolites is indispensable.

Natural products have also contributed to the development of many foundational scientific concepts, including but not limited to stereochemistry, optical activity, regioselectivity, and chirality. Their diverse and complex nature continually

inspires researchers, sparking the creation of innovative tools that can mimic natural processes to control bioregulation mechanisms and address practical challenges [7].

Even though they're 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 anticipate metabolic networks or occurrences of molecules in selected taxa. A significant resource is the LOTUS database [2] developed and maintained by the COMMONS lab, which currently lists over 750'000 occurrences of natural products. Despite this, no model has been proposed to predict their occurrence across the tree of life. The LOTUS initiative aims to address these shortcomings by consolidating and sharing structure-organism pair information via an open platform, providing transformative potential for natural products research and beyond. This process involves 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.

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.

2 Theoretical introduction

In the domain of mathematical structures, a graph represents a compilation of entities, termed as vertices or nodes, and their interconnections, designated as edges or links. Vertices act as discrete points or units, while edges denote the relationships or associations between pairs of these points. Within the context of biology, for instance, a graph can depict the intricate web of interactions among proteins in a cellular system: each protein being represented by a node, and the physical or functional interactions between them signified by an edge [8].

Often, these nodes carry specific attributes. In our work, these might include attributes such as the molecule's atomic structure, size, or charge. However, the acquisition of comprehensive attribute data can be challenging due to obstacles in data collection, inherent complexity, or, in some cases, privacy concerns. To mitigate this, graph-based semi-supervised learning, also known as node classification, is employed to predict missing attributes (i.e., labels y) for some nodes given known attributes (i.e., features x). This strategy has been effective in a myriad of applications, including predicting molecular functions and categorization of substances [9, 10].

Link prediction is 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 [11], 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 have been found to outperform more direct ones [12].

Graph neural networks (GNNs) [13] have been frequently employed for semi-supervised learning or for link predictions, also in the context of molecular networks [14]. Initially, GNNs synthesize the features and graph structure in the vicinity of each node into a single vector representation. Subsequently, 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) [15] 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 (RMF). RMFs model the joint distribution of all node labels within a conditional random field and predict an unknown label with its marginal probabilities. This method allows for the leveraging of label correlation during inference, which involves conditioning on the training labels. However, the increased interpretability and convenience of collective classification come 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 [16].

Our research did not initially utilize a graph-based approach. We started our work using more traditional, naive methods that treated our data as a simple collection of independent observations rather than recognizing the inherent interconnectedness of the molecular structures and interactions within organisms. This initial approach, while more straightforward, failed to fully capture the intricate complexity and interconnected nature of our dataset, a key characteristic of biological systems.

However, recognizing the limitations of such naive methods, we transitioned to the use of graphical models, specifically graph neural networks and collective classification techniques. This shift in our modelling paradigm has significantly enhanced our ability to predict the molecules present in any organism on Earth, allowing for the more effective use of both the Markov random field approach and GNNs. By treating our data as a graph, we've been better equipped to capture the nuanced relationships and dependencies between different molecular entities and their attributes.

2.1 Naive approach

Our objective is to infer the occurrence 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.

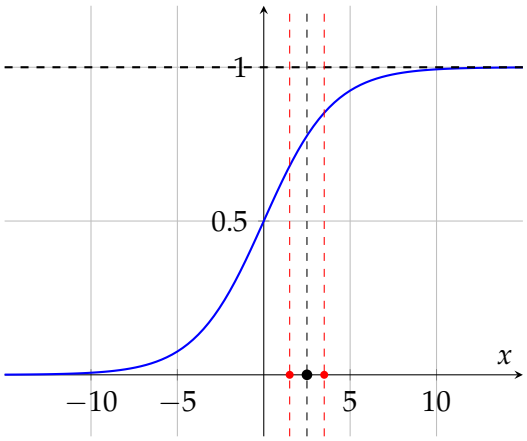


Figure 2.1: Illustration of the proposed model on a logistic function given by Equation (2.1). The black vertical line at $x = 2.5$ denotes the mean presence of a metabolite in the system. Flanking this, the two red dashed lines depict potential shifts in this average presence due to the influence of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

In Figure 2.1, a graphical representation illustrates the probabilistic association of a metabolite's presence within a particular species, predicated upon its mean prevalence across a taxonomic spectrum (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 conspicuously lacks this metabolite, the likelihood of its occurrence in species phylogenetically proximate to this clade diminishes — as indicated by the leftmost red line.

2.1.1 Emission probabilities

We consider several different types of data to inform about $\boldsymbol{x}$. This data may be of different dimensionality, e.g. may only discriminate along a subset of the axes

or at a higher scale along some axes. 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

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

2.1.2 Data sources

We consider two sets of data informative about $\boldsymbol{x}$: i) Presence-absence data obtained with mass-spectrometry and ii) presence-only reports of specific metabolites in specific species.

Mass spectrometry

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

$$\mathbb{P}(\boldsymbol{d}_{sj}|\boldsymbol{x}, \epsilon_{01}, \epsilon_{10}) = \prod_m \left[x_{sm}\left(\epsilon_{10}^{1-d_{sjm}}(1-\epsilon_{10})^{d_{sjm}}\right) + (1-x_{sm})\left(\epsilon_{01}^{d_{sjm}}(1-\epsilon_{01})^{1-d_{sjm}}\right)\right] \tag{2.4}$$

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.

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

$$\mathbb{P}(L_{ms}|\boldsymbol{x}(\xi(m,s)), R_{ms}) = \begin{cases} 0 & \text{if } \boldsymbol{x}(\xi(m,s)) = 0 \text{ and } L_{ms} = 1, \\ 1 & \text{if } \boldsymbol{x}(\xi(m,s)) = 0 \text{ and } L_{ms} = 0, \\ R_{ms} & \text{if } \boldsymbol{x}(\xi(m,s)) = 1 \text{ and } L_{ms} = 1, \\ 1 - R_{ms} & \text{if } \boldsymbol{x}(\xi(m,s)) = 1 \text{ and } L_{ms} = 0, \end{cases} \tag{2.5}$$

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

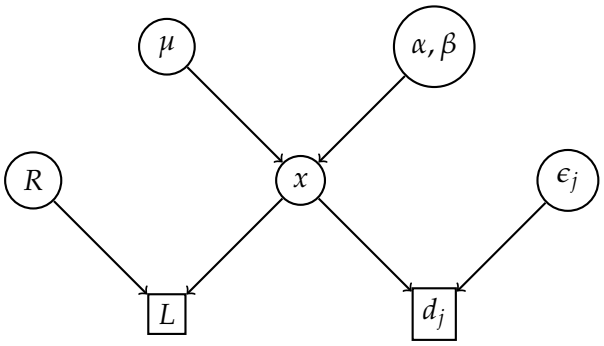


Figure 2.2: In the Directed Acyclic Graph (DAG) representing the *naive model*, the variable x denotes the binary state of a molecule's presence or absence within a designated species. Herein, μ signifies the mean presence along a specified axis. Both α and β denote the axis-specific intercept and a positive scalar constant, respectively. The parameter R stands for the dedicated research effort given to find x while L is indicative of the presence/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 (2.4).

$$R_{ms} = 1 - e^{-\gamma P_m - \delta Q_s} \tag{2.6}$$

with positives scalars γ and δ. In Figure 2.2 we show a Directed Acyclic Graph (DAG) of the proposed model.

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 [17, 18].

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

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

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

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

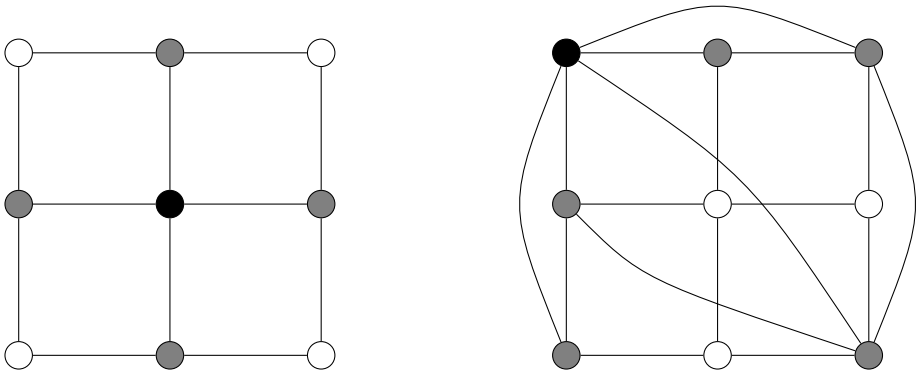
We assume the root state probabilities are given by the stationary distribution of the Markov chain:

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

The clique function $\phi(c)$

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

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 $\delta_{\nu(c)}(z) = p(\delta_{\nu(c)}(x))$.



(a) **(b)**

Figure 2.3: Representation of a Markov random field through an undirected graphical model where vertices depict arbitrary values and edges denote inter-nodes connectivity. **(a)** The Ising model [19] 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. Analogously in our proposed model, interactions extend beyond immediate neighbours, encompassing higher-order relationships as described by the trees $\mathcal{T}$. The illustration was conceptualized based on insights and frameworks derived from [20] and [21].

In Figure 2.3, a prototypical representation of a Markov Random Field is depicted. The Ising model, as illustrated in Figure 2.3a, operates under the premise that a node's state is solely contingent upon its immediate neighbours. Contrarily, our proposed model, akin to the structure presented in Figure 2.3b, embodies intricate relationships reminiscent of phylogenetic interconnections. Within our framework, the state of a given node is not merely influenced by its proximal counterparts within the phylogeny, but also by the relational distances encompassing those neighbours.

2.2.1 Data sources

The probabilities of the data given x were formulated employing the same model as delineated in Section 2.1.2. In Figure 2.4, we illustrate the Directed Acyclic Graph

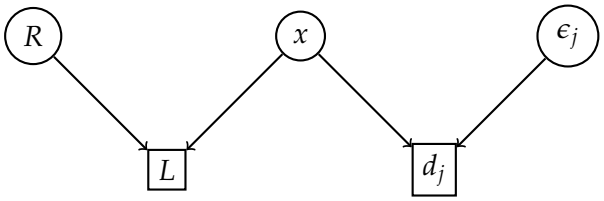


Figure 2.4: In the Directed Acyclic Graph (DAG) for the RMF model, x denotes the binary state of a molecule's presence or absence in a designated species. R signifies the research effort allocated towards the determination of x. L embodies the incidence of xx within the LOTUS database. The term d_j is representative of the j^{th} iteration of mass spectrometry performed on the specific species in question. Lastly, ϵ_j quantifies the associated error rates, as detailed in Equation (2.4).

(DAG) representation of the proposed model.

2.3 Graph Convolution Neural Network (GraphSAGE)

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 [22, 23].

GraphSAGE (SAmple and aggreGatE) [24] 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), or 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 [25]. Unlike traditional methods that optimize embeddings for each node, GraphSAGE's inductive approach promotes efficiency and adaptability, allowing for a seamless 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 and across molecular graphs. Its novel 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 organisms, facilitating the prediction in entirely new or evolving graphs.

HinSAGE [26], 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. HinSAGE's capability to seamlessly navigate the intricacies of such heterogeneous structures ensures a richer and more accurate representation of the connections, fostering improved predictive accuracy for link prediction tasks in biological contexts where species-molecule interactions need to be discerned.

3 Methods

3.1 Naive model

Prior to the application of actual experimental 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 variable μ was 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 species Q_s and per molecule P_m, were synthetically constructed by drawing from a Poisson distribution. Additionally, the variable σ was simulated by drawing from a Wishart distribution [27].

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). To emulate this binary characteristic, samples were subsequently drawn from a Bernoulli distribution, where the probability parameter was informed by the previous *expit* function. Building upon this stochastic framework, the probabilities of LOTUS were constructed in accordance with Equations (2.5) and (2.6). A condition was imposed such that if x for any given pair was 0, then the corresponding probability was explicitly set to 0.

These tailored probabilities were then employed as parameters for another Bernoulli distribution, 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. Code is available both on Bitbucket and GitHub.

3.3 HinSAGE

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, adhering to common practices in machine learning and data analysis.

Graphs were systematically constructed for both the training and testing subsets using the software library NetworkX v3.1 [28]. 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 [29, 30]. Molecules' features were composed of their classification data from Classyfire [31] and their Morgan Fingerprint [32], 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, 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 [26]. 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. These activation functions were chosen for their properties in mitigating vanishing gradient problems, thereby aiding the convergence of the model during the training process.

We moved forward with an objective to predict the entire scope of the LOTUS database, specifically targeting the probability for every possible species-molecule combination. This analysis covered a total of $5.45 \cdot 10^9$ pairs. All codes are available on Github.

4 Results and Discussion

4.1 Naive approach

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 2.5 and 2.6.

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 proposition of interpreting our data as a graph surfaced, 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 time constraints of the thesis, test and simulation for this model were not performed.

However this could work since paper TODO

4.3 HinSAGE

After evaluating the models using unseen edge data, the performance metrics revealed differing accuracy levels for each model. Specifically, the model trained to predict the "present in" relationships exhibited an accuracy of 0.92. In contrast, the model aimed at predicting the "has" relationships demonstrated an accuracy of 0.8.

It is important to note that the threshold for determining the presence or absence of metabolites was set at a probability value of 0.5. Consequently, probabilities exceeding this threshold were categorized as a presence, denoted as $x = 1$, whereas values below this threshold were classified as an absence, represented as $x = 0$.

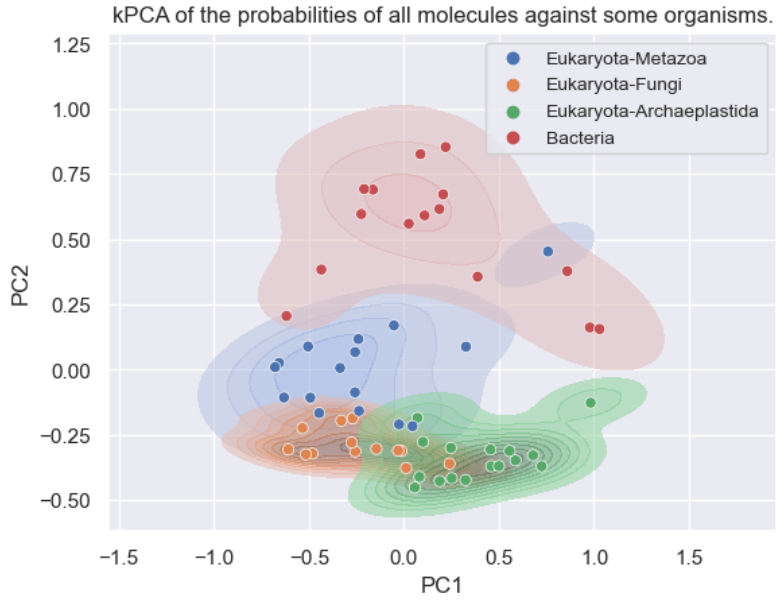


Figure 4.1: kPCA visualization of sampled species across primary biological domains. The second principal component distinguishes between domains, with potential influences from our GNN's molecular pattern recognition or biases within the LOTUS database.

Upon completion of the predictive analysis for the entirety of the LOTUS database, we sampled a dozen species from each primary biological domain, excluding Archaea. Each of these species was associated with 148′190 probability values. Subsequent to this, a kernel Principal Component Analysis (kPCA) was executed to discern potential variance across the domains. Given the high-dimensional nature of our dataset, kernel Principal Component Analysis (kPCA) emerged as the preferred method over traditional PCA. kPCA's strength lies in its adeptness at managing vast data dimensions and its capacity to capture non-linear relationships between features, something the linear assumptions of PCA might overlook. This inherent capability ensures that kPCA not only reduces data dimensions but also preserves the most salient and intricate relationships within the data.

Upon examination of Figure 4.1, a discernible differentiation between the biological domains is evident, notably facilitated by the second principal component.

Two hypotheses have been posited to elucidate these observed variances. Firstly, our Graph Neural Network (GNN) might be adept at distinguishing between the domains, thereby faithfully rendering the molecular signatures inherent to each. Alternatively, the kPCA may predominantly be reflecting the molecular biases inherent within the LOTUS database. Given that the LOTUS database structures its data in a molecule-species-paper triad, it stands to reason that a preponderance of extensively researched molecules, meriting scholarly publication, are intrinsically specific to particular domains. This contention finds resonance in *Rutz et al.* [2], which delineates that a substantive majority (exceeding 90%) of molecules catalogued in LOTUS exhibit domain specificity. Such inherent biases could feasibly account for the observed patterns in the kPCA results. A more granular investigation is requisite to substantiate either of the aforementioned hypotheses.

Figures 4.2 and 4.3 show examples of distributions of probabilities of all the species against cholesterol and erythromycin respectively.

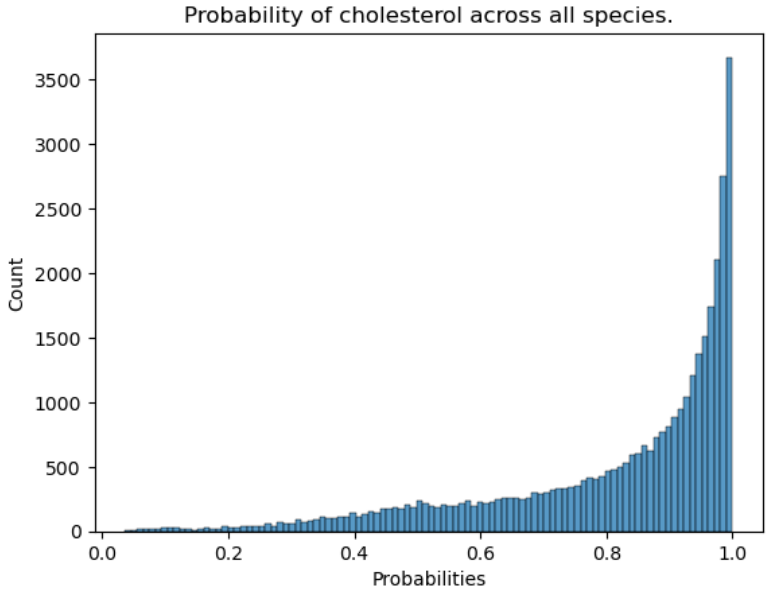


Figure 4.2: TODO

According to [33, 34], cholesterol is present everywhere. Our algorithm clearly shows that the majority of species should contain cholesterol. The fact that this works is due to the inherent functioning of GraphSAGE. Indeed since cholesterol is present a lot in the LOTUS database (522 different species) this means that in our graph, cholesterol has many edges, then this means that the algorithm is able to "understand" the context of cholesterol and have a good predictive value.

However, the algorithm does not need to have many papers. For example erythromycin has only 8 edges. However those edges are extremely local : they are all linked to bacteria and those bacteria are from the Actinobacteria phylum. The algorithm will then predict that most of the species don't have erythromycin (as seen in

Figure 4.3) but the ones it predicts to have are all very close to the Actinobacteria phylum (as we can see in Table 4.1).

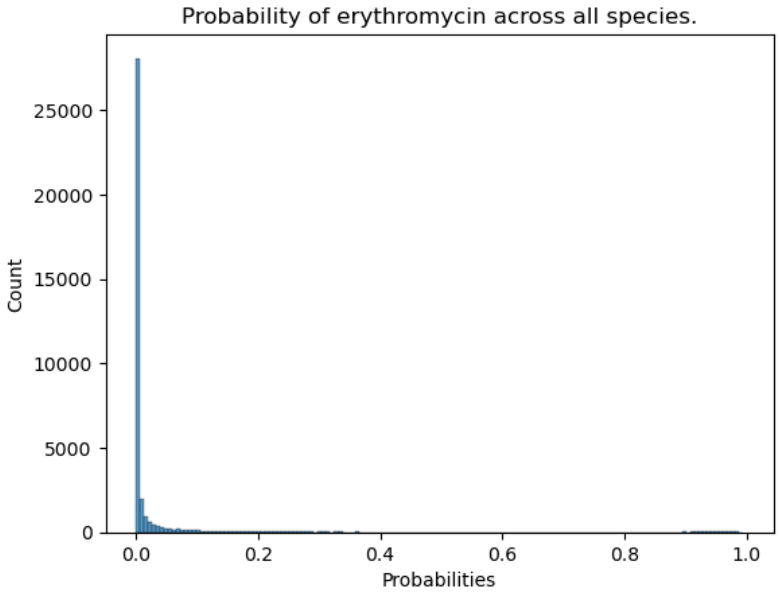


Figure 4.3: TODO

On the other hand, TODO: show image of probability of water across all species.

Table 4.1: Top 10 probabilities of Erythromycin. TODO

Species	Probability
Streptomyces diastaticus	0.9896
Streptomyces	0.9934
Micromonospora	0.9896
Streptomyces argillaceus	0.9926
Streptomyces fradiae	0.9928
Streptomyces antibioticus	0.9917
Streptomyces setonii	0.9897
Streptomyces halstedii	0.9891
Streptomyces ansochromogenes	0.9898
Streptomyces diastatochromogenes	0.9900

TODO Limitations: Sampling negatives edges is sampling only edges that do not exist yet. However this may lead to problems since by sampling negative edges, we might sample an edge that we may have but has not been discovered yet.

5 Conclusion and Future Work

Question 1 : Do you want LOTUS anticipated ?

Question 2 : Do you want the true presence/absence ?

Future work : Continue developing RMF. As shown in paper [16], both model can work. Restart HinSAGE and include all phylogenies. Include mass spectrometry runs in all models. Knowledge graph completion. Add more data coming from knowledge graph completion. explain what it is. Insist on continuing on MRF since it has great potential.

5.1 Knowledge Graph Completion

Knowledge Graphs (KGs) provide a robust technique for the consolidation of diverse data and the modelling of complex interactions, crucial for areas like forecasting the natural products present in various species [35].

In its simplest form, a graph is a data structure that designates items (or "nodes") and the links (or "edges") between them. In the context of our discussion, nodes might symbolize different species, while edges could denote a variety of relationships such as common ecosystems, shared characteristics, or the potential to yield similar natural products. Depending on the kind of the relationships, the graph can either be directed (representing asymmetric relations) or undirected (symbolizing symmetric relations).

Two primary categories of graphs exist: homogeneous and heterogeneous. A homogeneous graph possesses nodes and edges of the same category. In this scenario, a homogeneous graph might be comprised of species nodes interconnected by edges representing a particular connection, like a mutual ecosystem. Conversely, a heterogeneous graph contains nodes and edges of diverse types. For instance, a heterogeneous graph in this case could encompass nodes representing species, ecosystems, and natural products, linked through various relationships like "inhabits", "generates", or "has common traits with".

A specific kind of graph known as multigraphs permits multiple edges between the same pair of nodes and can also accommodate loops. This is advantageous when there are various types of relations between the same pair of nodes. Multigraphs are predominantly heterogeneous.

A Knowledge Graph (KG) is a distinct type of graph used to encode significant knowledge concerning a specific domain. It is a directed, heterogeneous multigraph

with domain-specific semantics for its node and relation types. In this setting, a KG could be employed to encode knowledge about species and their capability to produce natural products. The nodes in the KG, also known as entities, could symbolize different species, ecosystems, or specific natural products. The directed edges, usually represented as triples (head, relation, tail), encapsulate the relationships between these entities.

Knowledge graph embeddings are techniques to convert the discrete entities and relations in a KG into continuous vectors in a high-dimensional space, while preserving the original relationships from the KG [36]. This conversion to a continuous vector space is especially valuable in predicting missing information, like forecasting a species' potential to produce a specific natural product based on its relations with other entities in the KG.

In conclusion, KGs and their corresponding embeddings act as an essential instrument for encoding and analysing multifaceted, relational data. Within the context of predicting natural products in species, they can encapsulate and represent complex relationships between species, ecosystems, and the natural products themselves, potentially contributing to the discovery and comprehension of new natural products.

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 re-emergence 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] Richard J. Trudeau and Richard J. Trudeau. *Introduction to Graph Theory*. Dover Books on Advanced Mathematics. Dover Pub, New York, 1993.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

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

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

[21] 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.

[22] 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.

[23] 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.

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

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

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

[27] 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.

[28] 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.

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

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

[31] 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.

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

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

[34] International Agency for Research on Cancer et al. IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans. Supplement. *(No Title)*, 1987.

[35] Lisa Ehrlinger and Wolfram Wöß. Towards a Definition of Knowledge Graphs. September 2016.

[36] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, December 2017. doi:10.1109/TKDE.2017.2754499.