Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Designing meaningful measures to evaluate generative graph neural networks on protein datasets.

Master Thesis

Philip Jean Hartout

July 10, 2022

Advisors: Prof. Dr. Karsten M. Borgwardt, Tim Kucera

Department of Biosystems Science and Engineering, ETH Zürich

## Abstract

This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes.

# Contents

Chapter 1

# Introduction

Chapter 2

# Background & Related Work

This chapter introduces the core concepts built upon in this thesis and surveys recent literature tackling the evaluation of generative graph neural networks and the relevance of this problem in structural biology. Section 2.1 defines core mathematical and biological concepts that will be built upon in the thesis. Section 2.2 will discuss recent advances in the design of measures used to evaluate generative graph neural networks and in structural biology.

## 2.1 Background

The set of methods investigated in this thesis lies at the interface of structural biology and machine learning. We start by defining some relevant biological properties of proteins, followed by a survey various graph theoretical abstractions derived from the protein structure. We then move on to define generative models and the various classes of measures used to evaluate them.

### 2.1.1 Proteins

Proteins are large biomolecules that are formed from a sequence of amino acids, performing their functions as determined by their three-dimensional structure, and amino acid sequence. Proteins support a vast array of functions in living organisms, such as catalysing metabolic reactions, DNA replication, providing structural support to cells, transporting molecules and sensing stimuli.

Each protein is made up of one or more chains of amino acids, each of which contain a backbone and different side chains. The atoms in the backbone include a $\alpha$-carbon, another carbon and a nitrogen atom. An overview of the peptide backbone is shown in Figure 2.1. Interestingly, a plane is forned by two alpha carbons, the carboxyl group, and the hydrogen atom attached to the nitrogen atom (see Figure 2.1), making the peptide bond between the nitrogen and carbon atom resistant to twisting. That means that the rotations enabling the 3D folding of a protein is governed by the angle of the bonds linking the nitrogen atom to the α-carbon and the other carbon atom to the α-carbon, named φ and ψ. These angles' values are frequently used to validate proteins, characterise the secondary structure of proteins (i.e. structural features observed in certain segments of proteins), etc.

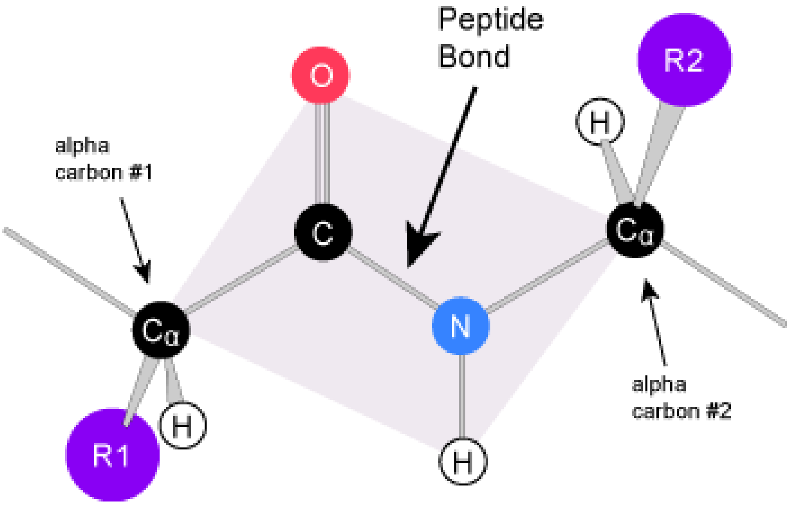

**Figure 2.1:** Schematic of the backbone of a protein. Two α-carbons are shown as well as a β-carbon in the middle. R1 and R2 represent the side chains of the amino acid.

To visualize such angles, a Ramachandran plot can be constructed for any protein. Such a plot can reveal secondary structural features such as β-sheets, α-helices, etc. An example of such a plot together with a 3D model of a protein can be found in Figure 2.2.

### 2.1.2 Graphs

Proteins are often abstracted using graphs. A graph $G$ is a pair of vertices $V$ and edges $E$ such that $G = (V, E)$, $|V| = n$ and $|E| = m$. Two vertices $i$ and $j$ are adjacent if there is an edge between them, i.e. $e_{ij} \in E$. The relationship between between edges can be represented as an $n \times n$ adjacency matrix $A$, where:

$$A_{ij} = \begin{cases} 1, & \text{if } e_{ij} \in E \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

The neighborhood of a node $v$ is the set of nodes with an edge directly to $v$, i.e. $N(v) = \{u \in V | e_{uv} \in E\}$. A graph is undirected if the edges do not contain directional information, i.e. $A_{ij} = A_{ji}$. A directed graph would result in directionality being encoded in edges, where $A_{ij}$ would not contain

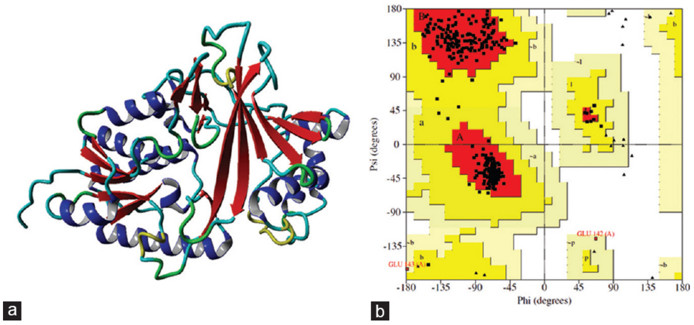

**Figure 2.2:** 3D structure of uridine diphosphogalactofuranose-galactopyranose mutase with a corresponding Ramachandran plot. The α-helices can be found on the middle left part of the Ramachandran plot, the β sheets on the upper right quadrant, and the left handed α-helices can be found in the middle upper right part of the plot. This figure is adapted from Nayak et al. [2018].

any information about $A_{ji}$. Nodes and edges in each graph can contain one or more labels. In this thesis, we will mostly deal with labeled undirected graphs, where each node will be labeled according to the amino acid type each node belongs to.

There are multiple ways of constructing graphs from proteins. First, one can extract a *contact map* of a protein by computing the (euclidean) distance between any two points belonging to each amino acid. The $\alpha$-carbon is often used for this purpose. This is a fully connected graph with continuously labeled edges representing the distance between each node. From there, it is possible to either extract a $k$-nearest neighbour graph, where $k \in \mathbb{N} > 0$ defines the amount of nodes directly connected to any given nodes; or an $\varepsilon$-graph, where each node within a given distance $\varepsilon \in \mathbb{R}^+ \setminus \{0\}$ of another node is connected. Both are graphs where each node is labeled with the residue name to which the $\alpha$-carbon belongs and the edges are unlabeled.

### 2.1.3 Topological Data Analysis

Although graphs are powerful abstractions of proteins, proteins can also be represented as *point clouds*. One powerful field of study of topological properties of point clouds is *topological data analysis*.

Topology has witnessed relentless theoretical progress since Henri Poincaré first addressed topological ideas as a distinct branch of mathematics in his 1895 publication of *Analysis Situs* [Poincaré, 1895]. Only recently, – with the advent of modern computing – has the field of computational topology and topological data analysis (TDA) gained momentum to investigate (high-dimensional) data in physics, biology, and beyond [Dey et al., 1999, Ghrist, 2008, Amézquita et al., 2020]. For material providing an extensive and formal introduction to topology and persistent homology, please refer to [Freedman and Chen, 2009, Edelsbrunner and Harer, 2010], and [Ghrist, 2008].

### 2.1.4 Generative models

### 2.1.5 Kernel methods

## 2.2 Related Work

### 2.2.1 Structural Biology

### 2.2.2 Metrics for Generative Graph Models

## 2.3 Summary

Chapter 3

# Methods

Dummy text.

## 3.1 Example Section

Dummy text.

### 3.1.1 Example Subsection

Dummy text.

#### Example Subsubsection

Dummy text.

**Example Paragraph** Dummy text.

*Example Subparagraph* Dummy text.

Chapter 4

# Results

Dummy text.

## 4.1 Example Section

Dummy text.

### 4.1.1 Example Subsection

Dummy text.

#### Example Subsubsection

Dummy text.

**Example Paragraph** Dummy text.

*Example Subparagraph* Dummy text.

Chapter 5

# Discussion

Dummy text.

## 5.1 Example Section

Dummy text.

### 5.1.1 Example Subsection

Dummy text.

#### Example Subsubsection

Dummy text.

**Example Paragraph** Dummy text.

*Example Subparagraph* Dummy text.

Chapter 6

# Conclusion

Dummy text.

## 6.1 Example Section

Dummy text.

### 6.1.1 Example Subsection

Dummy text.

#### Example Subsubsection

Dummy text.

**Example Paragraph** Dummy text.

*Example Subparagraph* Dummy text.

Appendix A

# Dummy Appendix

You can defer lengthy calculations that would otherwise only interrupt the flow of your thesis to an appendix.

# Bibliography

Erik J Amézquita, Michelle Y Quigley, Tim Ophelders, Elizabeth Munch, and Daniel H Chitwood. The shape of things to come: Topological data analysis and biology, from molecules to organisms. *Developmental Dynamics*, 2020.

Tamal K Dey, Herbert Edelsbrunner, and Sumanta Guha. Computational topology. *Contemporary mathematics*, 223:109–144, 1999.

Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

Daniel Freedman and Chao Chen. Algebraic topology for computer vision. *Computer Vision*, pages 239–268, 2009.

Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

Tapaswini Nayak, Lingaraja Jena, Pranita Waghmare, Bhaskar C Harinath, et al. Identification of potential inhibitors for mycobacterial uridine diphosphogalactofuranose-galactopyranose mutase enzyme: A novel drug target through in silico approach. *International Journal of Mycobacteriology*, 7(1):61, 2018.

Henri Poincaré. *Analysis situs*. Gauthier-Villars, 1895.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

---

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

**Title of work** (in block letters):

**Authored by** (in block letters):
*For papers written by groups the names of all authors are required.*

| **Name(s):** | **First name(s):** |
|---|---|
| | |
| | |
| | |
| | |
| | |

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

| **Place, date** | **Signature(s)** |
|---|---|
| | |
| | |
| | |
| | |
| | |

*For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.*