ETH
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 [**REVISE**] defines core mathematical and biological concepts that will be built upon in the thesis. Section [**REVISE**] will discuss recent advances in the design of measures used to evaluate generative graph neural networks and in structural biology.

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 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 α-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.

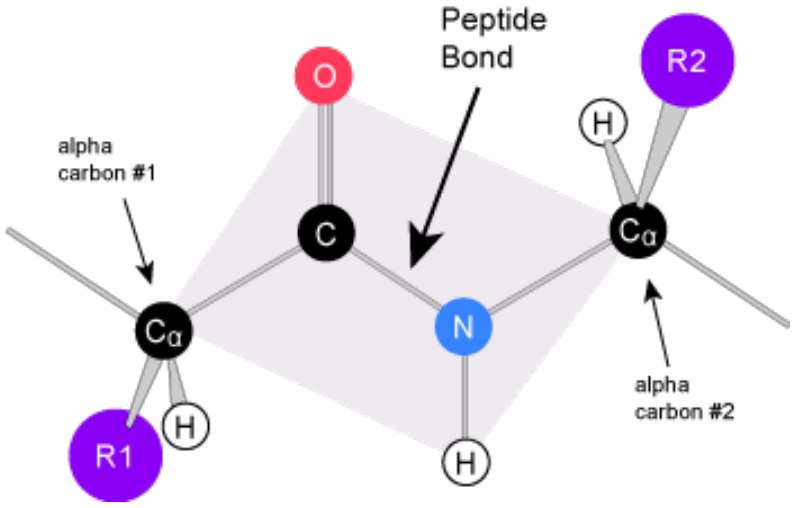

Figure 2.1: Schematic of the backbone of a protein. Two α-carbons are shown as well as a β-carbon in the middle. R1 and R2 represent the side chains of the amino acid.

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.

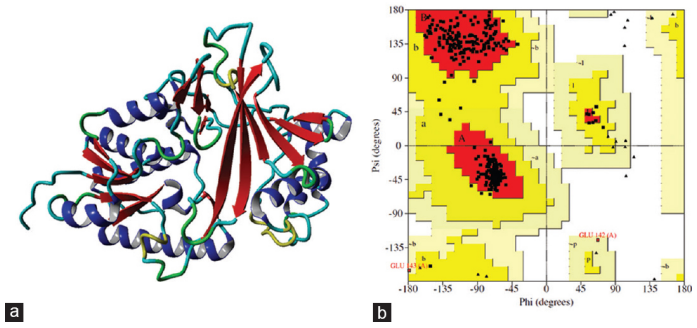

Figure 2.2: 3D structure of uridine diphosphogalactofuranose-galactopyranose mutase with a corresponding Ramachandran plot. The α-helices can be found on the middle left part of the Ramachandran plot, the β sheets on the upper right quadrant, and the left handed α-helices can be found in the middle upper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

2.2 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 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 α-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 ε-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 α-carbon belongs and the edges are unlabeled.

2.3 Topological Data Analysis

Although graphs are powerful representation of proteins, the latter 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].

A powerful computational technique to analyse topological properties of point clouds is *persistent homology*, which first requires us to define simplicial homology. Simplicial homology refers to a way of assigning connectivity information to topological objects, such as point clouds, which are represented by simplicial complexes. A simplicial complex K is a set of simplices which correspond to vertices in dimension 0, edges in dimension 1 and triangles in dimension 2. The subsets of a simplex $\sigma \in K$ are referred to as its faces, and every face $\tau \in K$. Moreover, any non-empty intersection of two simplices also needs to be part of the simplicial complex, i.e. $\sigma \cap \sigma' \neq \emptyset$ for $\sigma, \sigma' \in K$ implies $\sigma \cap \sigma' \in K$, meaning that K is closed under calculating the faces of a simplex.

Persistent homology extends simplicial homology by employing filtrations to imbue K with scale information. This process captures rich, mutli-scale topological information related to K in a principled way. The filtration process is generally defined by a function $f : K \to \mathbb{R}$ satisfying some finite number of values m and $f^0 \leq f^1 \leq \cdots \leq f^{m-1} \leq f^m$. This allows us to sort K using f, for instance by extending f linearly to higher-dimensional simplices via $f(\sigma) := \max_{v \in \sigma} f(v)$, leading to a nested sequence of simplicial complexes like so:

$$\emptyset = K^{(0)} \subseteq K^{(1)} \subseteq \cdots \subseteq K^{(m-1)} \subseteq K^{(m)}, \tag{2.2}$$

where $K^{(i)} := \{\sigma \in K \mid f(\sigma) \leq f^{(i)}\}$. This relationship enables tracking the appearance (i.e. a connected component arising) and the dissapearance (i.e. two connected components into one) of topological features across scales as one transitions from $K^{(i)}$ to $K^{(i+1)}$. The birth (i.e. appearance) and death (i.e. dissapearance) of topological features for different values of f are usually summarized in a *persistence diagram*, which is a multiset of tuples, each of which contains the values at which each features is born or dies.

A common construction for obtaining such features is the Vietoris-Rips complex [Vietoris, 1927]. It requires a distance threshold ε and a metric $\mathrm{d}(\cdot, \cdot)$ (usually, the Euclidean distance, as we will use in this thesis). The Vietoris-Rips complex at scale ε of an input protein point cloud is defined as $\mathcal{V}_\varepsilon(X) := \{\sigma \subseteq X | \mathrm{d}(x(i), x(j)) \leq \varepsilon\}$, $\forall x(i), x(j) \in \sigma$, i.e. $\mathcal{V}_\varepsilon$ contains all subsets of the input space whose pairwise distances are less than or equal to ε. $\mathcal{V}_\varepsilon$ is conceptually very similar to the ε-graphs discussed in section 2.2, except that ε here ranges over the entire space of possible distance values, and $\mathcal{V}_\varepsilon$ also tracks topological features over all three dimensions, instead of

only connected nodes[1].

Note that the multiplicity of the persistence diagram corresponds to the number of homology dimensions under study. In this thesis, given proteins are represented as three-dimensional point clouds, we choose to track topological features across three homology dimension: 0,1 and 2. Effectively, this tracks connected components in dimension 0, circular holes in dimension 1 and two dimensional voids or cavities in dimension 2 as the filtration function is applied. For a more thorough introduction to homology and homology groups, please refer to Edelsbrunner and Harer [2010].

2.4 Generative models

This thesis deals with measures to assess generative model performance, so we define generative models here. While discriminative machine learning techniques aim to learn some dependent variable $\mathcal{Y}$ from a set of (independent) features $\mathcal{X}$, generative machine learning models generate synthetic samples $\mathcal{X}'$ following the distribution of $\mathcal{X}$. Computing such probabilistic distributions through maximum likelihood estimation and related methods is intractable in many cases; as such, new learning paradigms were established to enable the modeling of complex, real-world distributions through gradient-based methods.

One such seminal method was that of generative adversarial learning, pioneered by Goodfellow et al. [2014], where a (deep) generator is pitted against a (deep) discriminator. The former's goal is to generate samples identical to the training distribution, while the latter is to classify whether or not the sample originated from the generator or the training distribution. Simultaneously developed methods by Kingma and Welling [2013] generalized this idea further and introduced the variational auto-encoder, where instead of a discriminator, the second network leverages the representation of the generator to perform approximate inference. In both cases, the two networks (i.e. the generator and the discriminator/inference network) are jointly trained using backpropagation to minimize some appropriate loss function.

[ADD LOSS FUNCTIONS]

A recent review of the existing landscape of generative modelling methods has been provided by Bond-Taylor et al. [2021].

These techniques have been particularly successful in the image domain, where modern GANs have been able to tackle multiple practical challenges such as mode collapse and convergence failure to produce realistic images,

[1] Technically, since our input consists of points (a.k.a. 0-simplicial complexes) exclusively (and not 1- or 2- simplicial complexes), we are actually primarily concerned with a subcomplex of $\mathcal{V}_\varepsilon(X)$ called the Čech complex.

Figure 2.3: Sample images generated by StyleGAN-XL, the state-of-the-art GAN by Sauer et al. [2022] at the time of writing.

such as the sample seen in Figure 2.3. More pertinent to this thesis is the application of generative models to graphs. The application domain has been reviewed by Zhou et al. [2020]. In short, graph generative networks are capable of operating on the highly versatile and extensible graph domain. It has been shown that they can produce small molecules, generate social networks, knowledge graphs, among many other real-world tasks. Generative networks can be grouped into two categories: those that generate nodes in each graph sequentially, such as GraphRNN by You et al. [2018], and those that generate graphs from some latent distribution directly, such as MolGAN by De Cao and Kipf [2018].

Operating in the graph domain incurs some unique challenges. From a modelling standpoint, dealing with graphs means dealing with a much larger and variable output space. Specifically, when dealing when an undirected graph of n nodes, at least $n^2/2$ values need to be specified. Additionally, the number of edges and nodes vary from sample to sample, which also needs to be accounted for in the model structure. Additionally, building a generative model generating graphs of up to n nodes, $n!$ possible adjacency matrices can be generated. Such a high representation complexity is challenging to model, expensive to compute, and difficult for objective functions to optimize. The last modelling-related issue when dealing with graphs is that the presence of one edge is not independent from another, i.e. real-world graphs often exhibit patterns of local connectedness which need to be accounted for in the model.

But perhaps the most significant problem plaguing all generative models is the evaluation problem. Concretely, this problem can be phrased into the following question: how does the practitioner go about evaluating the

quality of the set of samples generated? While sidestepping the problem is possible in the image domain by manually inspecting generated samples, a practice that might reveal interesting modelling pathologies, this cannot be done at scale, and for generative models operating in the graph domain.

2.5 Maximum Mean Discrepancy & Kernel methods

2.6 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.

Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *arXiv preprint arXiv:2103.04922*, 2021.

Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. *arXiv preprint arXiv:2202.00273*, 2022.

Leopold Vietoris. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Mathematische Annalen*, 97(1):454–472, 1927.

Jiaxuan You, Rex Ying, Xiang Ren, William Hamilton, and Jure Leskovec. GraphRNN: Generating realistic graphs with deep auto-regressive models. In *International conference on machine learning*, pages 5708–5717. PMLR, 2018.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

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.