Going into the networks of Life

Project Proposal - ML in Network Science

Dwivedi Deepesh
b00792828@essec.edu
DSBA Master - CentraleSupélec

Giacopelli Nicolò
b00794705@essec.edu
DSBA Master - CentraleSupélec

Liu Dayu
b00789480@essec.edu
DSBA Master - CentraleSupélec

Remadi Adel
b00804320@essec.edu
DSBA Master - CentraleSupélec

Abstract

We present our proposal for the Machine Learning in Network Science course (CentraleSupélec). Our interest is towards bioinformatics, and the power that are being unleashed by the AI revolution for what concerns the study of the primary sources of Life. In particular, the ability to distinguish the important structures, and being able to re-generate them accurately. All the topics are subject to change, as there is still much to study in the field and something may be found along the way

Keywords: bioinformatics, proteins, enzymes, graph classification, graph generation

1 Introduction and motivation

Proteins are complex macromolecules that are essential for a wide range of biological processes. They are made up of long chains of amino acids that fold into specific three-dimensional structures, which in turn determine their biological function. Proteins play a critical role in maintaining the structure and function of cells, and they are involved in very different processes such as enzyme catalysis, signaling, defense, storage and transport. Enzymes are a specific type of protein that catalyze chemical reactions in living organisms. Enzymes are highly specific, meaning that each enzyme can catalyze a particular chemical reaction or set of reactions. Enzymes function by binding to specific substrates and facilitating their conversion to products through a series of chemical reactions. Enzymes are essential for many metabolic processes in living organisms, and they play a critical role in maintaining cellular homeostasis.

GNNs, or Graph Neural Networks, are a type of machine learning model that can be used to analyze and make predictions about data that can be represented as graphs. GNNs are particularly useful for tasks such as node classification and graph classification, and they have applications in areas such as social network analysis, chemistry, and recommendation systems. A very interesting aspect for what concerns bioinformatics is the generation of potential candidates for enzymatic structures. GANs, or Generative Adversarial Networks, are a type of machine learning model that can be used to generate new data based on patterns learned from existing data. GANs consist of two neural networks - a generator network that generates new data, and a discriminator network that evaluates the realism of the generated data. GANs have applications in areas such as image and text generation, and they are also being explored for use in generating new protein structures. Enzymes and proteins are intrinsically linked and often confused. Essentially, an enzyme is a specific type of protein that performs a very specific function. Enzymes function to regulate biochemical reactions in living things, in this sense, they operate solely as a functional protein, while a protein can be either functional or structural. Therefore, all enzymes can be adequately described as globular proteins, however, not all proteins are globular. And this small difference makes a huge impact for potential progresses in the field.

Our project proposal focuses on using machine learning techniques to classify enzymes based on their protein structures. Specifically, we use Generative Adversarial Networks (GANs) to generate graphs of protein structures and Graph Neural Networks (GNNs) to classify the enzymes. Enzymes are a specific type of protein that catalyze chemical reactions in living organisms and are highly specific in their function. By utilizing GANs and GNNs, we aim to improve the accuracy and efficiency of enzyme classification, which has important implications for drug discovery and understanding biological processes.

2 Problem definition and methodology

2.1 Problem definition

The study of proteins and enzymes is an important area of research in biochemistry and molecular biology. Understanding the structure and function of proteins and enzymes can provide insights into the fundamental processes that occur in living organisms, and it can also have practical applications in fields such as drug discovery and biotechnology. Importance of Enzyme prediction

- Enzyme prediction helps us understand enzyme function, which is important for a wide range of biological, biotechnological, and environmental applications.
- Enzyme prediction can aid in drug discovery by identifying new drug targets or designing drugs that interact

with specific enzymes in a more effective and targeted way.
- Enzyme prediction can lead to new discoveries, insights, and applications that have the potential to benefit society.

One key motivation for using graph neural networks (GNNs) for enzyme classification is that enzymes are inherently graph-like structures, composed of amino acid residues that interact with each other through various physical and chemical interactions. Traditional machine learning methods are often limited in their ability to capture these complex interactions, but GNNs have been shown to be effective at modeling graph data and extracting meaningful features. By representing enzymes as graphs and applying GNNs, we can leverage the inherent structure of enzymes to make more accurate predictions about enzyme function and classification. This can lead to new insights into enzyme activity, new drug targets, and more efficient enzyme engineering, with potential benefits for a range of fields, from biomedicine to biotechnology to environmental science.

2.2 Methodology

The general approach will be the one of benchmarking different techniques and methods to perform two different but interrelated task of classification and generation.

Graph Classification The first problem we plan to tackle is a Graph Classification one. In this context, we want to learn a function $f : \mathcal{G} \rightarrow \mathcal{Y}$ predicting a property of the graph. For instance, if $\mathcal{G}$ is the set of molecular graphs, it may be interesting to be accurate in predicting the absence or presence of toxicity. In this field, the main strand of research have passed from Graph kernels to Graph Neural Networks around the year 2016. Our plan is then to benchmark techniques of supervised graph classification with GCN ([6]), and Deep Graph CNN ([11]) andn GraphSAGE ([5]). Different architectures will be considered, as well as specific modules, so that different aggregator functions like Top-K and hierarchical pooling module ([12]) will be tried out, as well as the Attention mechanism ([10]). At the same time, another benchmark will be constituted by unsupervised embedding techniques, such as graph2vec ([8]) and Unsupervised Graph-level embedding (from [1]), which can incorporate a supervised loss for classification. Comparing different embedding techniques will prove to be extremely interesting in relation to the second task of graph generation.

Graph Generation A fundamental problem that rises from bioinformatics is the problem of drug discovery and synthesis of molecules with beneficial properties. Many advances are currently being made in the field of graph generation, based on the tradition of GANs applied to structured data, like images. The general structures is in fact the same, composed by a generator and a discriminator. The former takes as input a random variable $z \sim p(z)$ and outputs the node feature matrix and the adjacency matrix A, which characterize structurally and semantically the graph structure. The discriminator, on the other hand, is a classifier trained adversarily with the generator to distinguish between the real images and the fake generated ones. The general approach in the field of biotechnology is the one of MolGAN ([3]), which adds a reward functions from the reinforcement learning world, encouraging desirable structures of the generated protein. In our case, the reward function is very difficult to obtain, but we will work with the architecture beneath.
A potential added value of our research will be the comparison that we could be able to trace between the GAN architecture and the encoder-decoder schemes used for efficient graph embedding, that can be used for downstream machine learning tasks or for reconstruction of the original protein structure.

2.3 Evaluation

The dataset considered comes from the TUDataset open-source project, started in 2016 ([7]). TUDataset project has the aim of providing a collection of graph datasets for a uniform benchmarking across researchers, that was found to be lacking previously. The project has made available 120 graph datasets of various sizes, spanning from social neworks to bioinformatics and computer vision. They are all accessible directly from APIs of popular frameworks like PyTorch Geometric and Spektral (the one chosen). In particular, for the moment we are considering the PROTEINS and ENZYMES datasets. The first comes from the Protein Data Bank (gathered by [4]), while the second was derived from the Brenda database ([9]). Each graph represents a macromolecule, where each node is an amino acid. Following the model introduced in [2], an edge connects two nodes if the nodes are neighbors along the sequence of amino acids or if they are among three-nearest neighbors in space (secondary-structure content). The PROTEINS dataset concerns a binary classification problem of proteins into enzyme/non-enzyme, since proteins can have different functions (support, defense, transport, support...) alongside the enzymatic one. The ENZYMES dataset gives the possibility of benchmarking for a multi-class classification problem, assigning enzymes to one of the 6 EC functional classes, reflecting the catalyzed chemical reaction. The first dataset contains 1113 graphs (avg n. of nodes $n = 40$, avg n. of edges $m = 73$), the second 600 ($n = 33$, $m = 62$). For what concerns model benchmarking, the structure will depend on the two problems considered. For the classification problem, there are clear metrics such as accuracy and F1 score. A potential added value of our research could be using the classifier trained independently to test the efficiency of our graph generation technique.

References

[1] Yunsheng Bai, Hao Ding, Yang Qiao, Agustin Marinovic, Ken Gu, Ting Chen, Yizhou Sun, and Wei Wang. 2019. Unsupervised inductive graph-level representation learning via graph-graph proximity. *arXiv preprint arXiv:1904.01098* (2019).

[2] Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. 2005. Protein function prediction via graph kernels. *Bioinformatics* 21, suppl_1 (2005), i47–i56.

[3] Nicola De Cao and Thomas Kipf. 2018. MolGAN: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973* (2018).

[4] Paul D Dobson and Andrew J Doig. 2003. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of molecular biology* 330, 4 (2003), 771–783.

[5] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).

[6] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[7] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. 2020. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663* (2020).

[8] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005* (2017).

[9] Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. 2004. BRENDA, the enzyme database: updates and major new developments. *Nucleic acids research* 32, suppl_1 (2004), D431–D433.

[10] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat* 1050, 20 (2017), 10–48550.

[11] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. 2018. An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[12] Zhen Zhang, Jiajun Bu, Martin Ester, Jianfeng Zhang, Chengwei Yao, Zhi Yu, and Can Wang. 2019. Hierarchical graph pooling with structure learning. *arXiv preprint arXiv:1911.05954* (2019).