

# Software Requirements Specification for Re-ProtGNN: Re-implementation of the ProtGNN model

Yuanqi Xue

March 27, 2025

# Contents

|          |  |              |
|----------|--|--------------|
| <b>1</b> | <b>Reference Material</b>                        | <b>iv</b>    |
| 1.1      | Table of Units . . . . .                         | iv           |
| 1.2      | Table of Symbols . . . . .                       | iv           |
| 1.3      | Abbreviations and Acronyms . . . . .             | v            |
| 1.4      | Mathematical Notation . . . . .                  | v            |
| <b>2</b> | <b>Introduction</b>                              | <b>vi</b>    |
| 2.1      | Purpose of Document . . . . .                    | vi           |
| 2.2      | Scope of Requirements . . . . .                  | vi           |
| 2.3      | Characteristics of Intended Reader . . . . .     | vi           |
| 2.4      | Organization of Document . . . . .               | vii          |
| <b>3</b> | <b>General System Description</b>                | <b>vii</b>   |
| 3.1      | System Context . . . . .                         | vii          |
| 3.2      | User Characteristics . . . . .                   | viii         |
| 3.3      | System Constraints . . . . .                     | viii         |
| <b>4</b> | <b>Specific System Description</b>               | <b>viii</b>  |
| 4.1      | Problem Description . . . . .                    | viii         |
| 4.1.1    | Terminology and Definitions . . . . .            | ix           |
| 4.1.2    | Physical System Description . . . . .            | ix           |
| 4.1.3    | Goal Statements . . . . .                        | x            |
| 4.2      | Solution Characteristics Specification . . . . . | x            |
| 4.2.1    | Types . . . . .                                  | x            |
| 4.3      | Scope Decisions . . . . .                        | xi           |
| 4.3.1    | Modelling Decisions . . . . .                    | xi           |
| 4.3.2    | Assumptions . . . . .                            | xi           |
| 4.3.3    | Theoretical Models . . . . .                     | xi           |
| 4.3.4    | General Definitions . . . . .                    | xiv          |
| 4.3.5    | Data Definitions . . . . .                       | xx           |
| 4.3.6    | Data Types . . . . .                             | xxi          |
| 4.3.7    | Instance Models . . . . .                        | xxi          |
| 4.3.8    | Input Data Constraints . . . . .                 | xxii         |
| 4.3.9    | Properties of a Correct Solution . . . . .       | xxiii        |
| <b>5</b> | <b>Requirements</b>                              | <b>xxiii</b> |
| 5.1      | Functional Requirements . . . . .                | xxiv         |
| 5.2      | Nonfunctional Requirements . . . . .             | xxiv         |
| 5.3      | Rationale . . . . .                              | xxiv         |
| <b>6</b> | <b>Likely Changes</b>                            | <b>xxiv</b>  |

|    |                                  |      |
|----|----------------------------------|------|
| 7  | Unlikely Changes                 | xxv  |
| 8  | Traceability Matrices and Graphs | xxv  |
| 9  | Development Plan                 | xxv  |
| 10 | Values of Auxiliary Constants    | xxvi |

## Revision History

| Date        | Version | Notes       |
|-------------|---------|-------------|
| Feb 6, 2024 | 1.0     | First Draft |

# 1 Reference Material

This section records information for easy reference.

## 1.1 Table of Units

Not applicable, as there is no unit in this software.

## 1.2 Table of Symbols

The table that follows summarizes the symbols used in this document along with their units. The choice of symbols was made to be consistent with the heat transfer literature and with existing documentation for solar water heating systems. The symbols are listed in alphabetical order.

| symbol                            | unit     | description                                    |
|-----------------------------------|----------|--|
| $X$                               | Unitless | Node feature matrix                            |
| $X_p$                             | Unitless | Prototype node feature matrix                  |
| $A$                               | Unitless | Adjacency matrix                               |
| $A_p$                             | Unitless | Prototype adjacency matrix                     |
| $y$                               | Unitless | True class label                               |
| $\hat{y}$                         | Unitless | Predicted class label                          |
| $W$                               | Unitless | Weight matrix for the fully connected layer    |
| $b$                               | Unitless | Bias term                                      |
| $W_g$                             | Unitless | Weight matrix for GNN encoder                  |
| $P$                               | Unitless | Prototype representations                      |
| $h$                               | Unitless | Graph embedding                                |
| $\sigma$                          | Unitless | Sigmoid activation function                    |
| $\mathcal{L}$                     | Unitless | Loss function                                  |
| $L_{CE}$                          | Unitless | Cross-entropy loss                             |
| $L_{Clst}$                        | Unitless | Cluster loss                                   |
| $L_{Sep}$                         | Unitless | Separation loss                                |
| $L_{Div}$                         | Unitless | Diversity loss                                 |
| $\eta$                            | Unitless | Learning rate                                  |
| $\lambda_1, \lambda_2, \lambda_3$ | Unitless | Regularization parameters                      |
| $h_v^{k+1}$                       | Unitless | Node representation at layer $k + 1$           |
| $h_u^k$                           | Unitless | Representation vector of node $u$ at layer $k$ |
| $\tilde{A}$                       | Unitless | Normalized adjacency matrix                    |

|                  |          |   |
|------------------|----------|---|
| $\hat{A}$        | Unitless | Adjacency matrix with self-connections                            |
| $\hat{D}$        | Unitless | Diagonal degree matrix where $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$ |
| $W^k$            | Unitless | Trainable weight matrix at layer $k$                              |
| $\mathcal{N}(v)$ | Unitless | Neighborhood of node $v$ , i.e., set of adjacent nodes            |
| $s_{\max}$       | Unitless | Cosine similarity threshold for prototype diversity loss          |
| $C$              | Unitless | Number of classes in the classification task                      |

---

### 1.3 Abbreviations and Acronyms

| symbol     | description                            |
|------------|--|
| A          | Assumption                             |
| DD         | Data Definition                        |
| GD         | General Definition                     |
| GS         | Goal Statement                         |
| IM         | Instance Model                         |
| LC         | Likely Change                          |
| PS         | Physical System Description            |
| R          | Requirement                            |
| SRS        | Software Requirements Specification    |
| TM         | Theoretical Model                      |
| ProtGNN    | Prototype-based Graph Neural Network   |
| Re-ProtGNN | Re-implementation of the ProtGNN model |
| GNN        | Graph Neural Network                   |
| GCN        | Graph Convolutional Network            |
| GIN        | Graph Isomorphism Network              |

---

### 1.4 Mathematical Notation

| Notation           | Description                                   | Source |
|--------------------|---|--------|
| $num\_rows(\cdot)$ | Number of rows in a matrix                    | -      |
| $num\_cols(\cdot)$ | Number of columns in a matrix                 | -      |
| $dim(\cdot)$       | Dimension of a matrix                         | -      |
| $M[i, j]$          | Element at row $i$ , column $j$ in matrix $M$ | -      |

|                      |  |  |
|----------------------|--|--|
| $f(\cdot)$           | A function representing the graph encoder, as defined in TM4.3.3 | -  |
| $\cos(\cdot, \cdot)$ | Cosine similarity function                                       | <a href="#">ScienceDirect (2024)</a>           |
| $\text{ReLU}(\cdot)$ | ReLU activation function   | <a href="#">Wikipedia contributors (2024f)</a> |
| $\nabla$             | Gradient operator  | <a href="#">Wikipedia contributors (2024c)</a> |

---

## 2 Introduction

Graph Neural Networks (GNNs) have shown strong performance in node classification, graph classification, and link prediction tasks. However, their black-box nature makes it difficult to understand their decision-making process, limiting their usage in critical areas such as medical diagnosis.

To address this, the paper *ProtGNN: Towards Self-Explaining Graph Neural Networks* [Zhang et al. \(2022\)](#) proposes ProtGNN (Prototype-based Graph Neural Network), an interpretable GNN model designed to improve explainability by learning representative prototypes for each class. The goal of this project is to re-implement ProtGNN and validate its results. Further details on the problem and motivation behind this project can be found in the Problem Statement document [Xue \(2025\)](#).

The introduction section outlines the Software Requirement Specifications (SRS) for Re-ProtGNN (Re-implementation of the ProtGNN model), detailing its purpose, scope, target audience, and organization.

### 2.1 Purpose of Document

The purpose of this SRS document is to define and communicate the requirements, constraints, and models for the Re-ProtGNN project. It serves as a formal reference and links to the other documents of this project.

### 2.2 Scope of Requirements

Re-ProtGNN is a re-implementation of the existing ProtGNN model, aimed at reproducing its results and validating its performance on graph datasets. The model operates in a supervised setting, performing graph classification while incorporating built-in interpretability.

The input graphs for training must be valid labeled graphs, ensuring they contain both structural and node feature information necessary for classification. Further details on data constraints and assumptions can be found in (Section [4.3.2](#)).

### 2.3 Characteristics of Intended Reader

This document is intended for machine learning researchers, AI practitioners, and students who are interested in interpretable GNNs and prototype-based learning. To effectively un-

derstand and utilize this document, the reader should have:

- High-school knowledge of calculus, particularly partial derivatives and gradient-based optimization.
- Completion of an undergrad-level course or equivalent experience in supervised learning (e.g., logistic regression) and neural networks (including backpropagation, loss functions, and optimization).
- Familiarity with graph-based learning (e.g., message-passing mechanisms in GNNs) is recommended but not required.
- Experience with Python and PyTorch, including implementing and training neural networks.
- Basic knowledge of model explainability in AI, including the concept of prototypes in classification models.

This document is designed to be accessible to readers with the above qualifications. However, researchers with limited exposure to graph neural networks can still benefit from the explanations provided, as long as they have a strong foundation in deep learning and optimization methods.

## 2.4 Organization of Document

This SRS document is based on the template by [Smith and Lai \(2005\)](#); [Smith et al. \(2007\)](#). It systematically presents the system’s objectives, theoretical foundations, assumptions, and models. For a structured reading approach, readers may start with the goal statements (Section 4.1.3), followed by the theoretical models (Section 4.3.3) and general definitions (Section 4.3.4), which build upon these goals. The document concludes with instance models (Section 4.3.7), providing a concrete understanding of the system’s actual implementation.

# 3 General System Description

This section provides general information about the system. It identifies the interfaces between the system and its environment, describes the user characteristics and lists the system constraints.

## 3.1 System Context

- User Responsibilities:
  - Download the MUTAG dataset and load it to Re-ProtGNN.
  - Run Re-ProtGNN on a system with appropriate computational resources.





Figure 1: System Context

- Interpret the model’s output, including classifications and prototype-based explanations.
- Re-ProtGNN Responsibilities:
  - Train a GNN model using gradient-based optimization.
  - Provide classification predictions.
  - Learn prototype representations.
  - Offer visualization tools for understanding learned prototypes.

## 3.2 User Characteristics

The end user of Re-ProtGNN includes anyone working with GNNs or those interested in interpretable deep learning models. While prior experience with interpretable machine learning is beneficial, it is not required. The basic skill set needed to use Re-ProtGNN includes basic computer literacy and a high school-level understanding of graphs.

## 3.3 System Constraints

The Re-ProtGNN system has the constraints that the input must be a valid graph dataset (i.e., adjacency matrices, node features).

# 4 Specific System Description

This section first presents the problem description, which gives a high-level view of the problem to be solved. This is followed by the solution characteristics specification, which presents the assumptions, theories, definitions and finally the instance models.

## 4.1 Problem Description

Traditional GNNs excel in graph classification tasks but function as black-box models, making it difficult for users to understand their classification decisions. ProtGNN addresses this

lack of interpretability by incorporating prototype-based explanations into graph classification. This project, Re-ProtGNN, is a re-implementation of ProtGNN, aimed at validating its reproducibility and performance on benchmark datasets.

#### 4.1.1 Terminology and Definitions

This subsection provides a list of terms that are used in the subsequent sections and their meaning, with the purpose of reducing ambiguity and making it easier to correctly understand the requirements:

- **Graph:** A structure consisting of nodes and edges, typically represented by a node feature matrix and an adjacency matrix.
- **Graph Classification:** The task of assigning a label to an entire graph based on its structure and node attributes.
- **Graph Neural Networks:** A type of deep learning model designed to process graph-structured data by aggregating and transforming node information.
- **Prototype:** A representative subgraph that highlights the key structural patterns influencing classification decisions.
- **MUTAG Dataset:** A benchmark dataset for molecular graph classification, distinguishing between mutagenic and non-mutagenic compounds [Debnath et al. \(1991\)](#).
- **Supervised Learning:** A training approach where the model learns from labeled data, mapping input graphs to their corresponding class labels.

#### 4.1.2 Physical System Description

The physical system of Re-ProtGNN, as shown in Figure 2, includes the following elements:

- PS1: **Graph Inputs:** The dataset consists of labeled graphs used for graph classification (e.g., MUTAG).
- PS2: **GNN Encoder:** A GNN-based encoder (e.g., GCN, GIN) extracts node and graph-level embeddings through message passing. Graph embeddings are used for both classification and prototype learning.
- PS3: **Prototype Layer:** A prototype layer learns representative subgraph embeddings that characterize structural patterns of different classes. Prototypes are updated during training to match key graph structures associated with class labels.
- PS4: **Fully Connected Layer:** The model classifies graphs based on their learned representations. The most similar prototype is retrieved to justify the classification result.

- PS5: Training Process: The model is trained using supervised learning, optimizing a combined cross-entropy loss and prototype-based losses.
- PS6: Inference Process: The trained model classifies new graphs and provides prototype-based explanations. Prototype similarity scores are used to explain classification decisions.

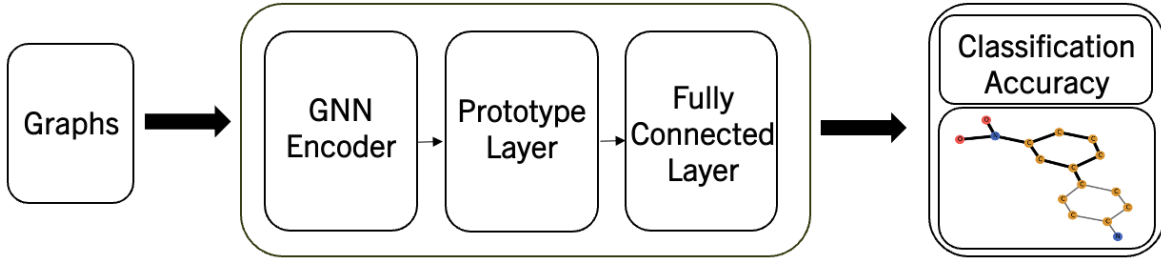


Figure 2: The ProtGNN Framework

### 4.1.3 Goal Statements

Given the input graphs, the goal statement is:

- GS1: Output a trained model that classifies the input graphs and identifies a few prototypes (i.e., key graph structures that contribute to classification) for each class.

## 4.2 Solution Characteristics Specification

The instance models that govern Re-ProtGNN are presented in Subsection 4.3.7. The information to understand the meaning of the instance models and their derivation is also presented, so that the instance models can be verified.

### 4.2.1 Types

| symbol       | description |
|--------------|-------------|
| $\mathbb{Z}$ | Integer     |
| $\mathbb{R}$ | Real number |

## 4.3 Scope Decisions

Re-ProtGNN focuses on validating the classification accuracy and prototype quality of the ProtGNN model using the MUTAG dataset [Debnath et al. \(1991\)](#).

### 4.3.1 Modelling Decisions

Re-ProtGNN follows the same model design and has the same Loss Function (see IM4.3.3) as ProtGNN.

### 4.3.2 Assumptions

This section simplifies the original problem and helps in developing the theoretical model by filling in the missing information for the physical system. The numbers given in the square brackets refer to the theoretical model [TM], general definition [GD], data definition [DD], instance model [IM], or likely change [LC], in which the respective assumption is used.

- A1: The node features of input graphs are relevant to classification, as required by TM4.3.3 to produce meaningful graph embeddings.
- A2: The structural patterns of input graphs are relevant to classification, as required by TM4.3.3 to produce meaningful graph embeddings.
- A3: For the training phase IM1 to produce an accurate model and the inference phase IM2 to generate reasonable predictions and prototypes, each input graph must belong to one of the two predefined classes (i.e., mutagenic or non-mutagenic) as defined in the MUTAG dataset [Debnath et al. \(1991\)](#).

### 4.3.3 Theoretical Models

This section focuses on the general equations and laws that Re-ProtGNN is based on.

---

**RefName:** TM1:GNNFP

**Label:** Graph Neural Network (GNN) Forward Pass

---

**Equation:**  $h_v^{k+1} = \text{ReLU} \left( \sum_{u \in \mathcal{N}(v)} W^k h_u^k \tilde{A}_{uv} \right),$

**Description:** The above equation describes the message-passing paradigm of Graph Convolutional Networks (GCNs), where the representation of each node  $v$  is iteratively updated by aggregating representations from its neighboring nodes  $\mathcal{N}(v)$ , using a trainable weight matrix  $W^k$  and an activation function  $\text{ReLU}(\cdot)$ .

- $h_u^k$ : Representation vector of node  $u$  at the  $k$ -th layer. (Type: a vector of  $\mathbb{R}$ )
- $\tilde{A} = \hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}}$ : Normalized adjacency matrix. (Type: a matrix of  $\mathbb{R}$ )
- $\hat{A} = A + I$ : Adjacency matrix of graph  $G$  with self-connections. (Type: a matrix of  $\mathbb{R}$ )
- $\hat{D}$ : Diagonal degree matrix where  $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$ . (Type: a matrix of  $\mathbb{R}$ )
- $\text{ReLU}(\cdot)$ : ReLU activation function.
- $W^k$ : Trainable weight matrix at the  $k$ -th layer. (Type: a matrix of  $\mathbb{R}$ )

**Notes:** None

**Source:** [Kipf and Welling \(2016\)](#)

**Ref. By:** A1, A2, TM4.3.3, GD2, GD5, GD6, DD2, IM1

**Preconditions for TM1:GNNFP:** None

**Derivation for TM1:GNNFP:** Not Applicable

---

---

**RefName:** TM2:LF

**Label:** Loss Function for Re-ProtGNN

---

**Equation:**  $\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{\text{Clst}} + \lambda_2 \mathcal{L}_{\text{Sep}} + \lambda_3 \mathcal{L}_{\text{Div}}$

**Description:** The overall loss function consists of multiple terms:

- $\mathcal{L}_{CE} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$ : Cross-entropy loss for classification. (Type:  $\mathbb{R}$ )
- $\mathcal{L}_{\text{Clst}} = \frac{1}{n} \sum_{i=1}^n \min_{j: p_j \in P_{y_i}} \|f(x_i) - p_j\|_2^2$ : Cluster loss to ensure embeddings are close to prototypes of the same class. (Type:  $\mathbb{R}$ )
- $\mathcal{L}_{\text{Sep}} = -\frac{1}{n} \sum_{i=1}^n \min_{j: p_j \notin P_{y_i}} \|f(x_i) - p_j\|_2^2$ : Separation loss to push embeddings away from prototypes of other classes. (Type:  $\mathbb{R}$ )
- $\mathcal{L}_{\text{Div}} = \sum_{k=1}^C \sum_{\substack{i \neq j \\ p_i, p_j \in P_k}} \max(0, \cos(p_i, p_j) - s_{\max})$ : Diversity loss to prevent prototypes from collapsing. (Type:  $\mathbb{R}$ )
- $y$ : Ground-truth label of an input graph. (Type:  $\mathbb{Z}$ )
- $\hat{y}$ : Predicted label of an input graph. (Type:  $\mathbb{R}$ )
- $p_j$ : A prototype embedding representing a characteristic subgraph associated with a specific class. (Type: a vector of  $\mathbb{R}$ )
- $x_i$ : The node feature representation of the  $i$ -th input graph in the dataset. (Type: a matrix of  $\mathbb{R}$ )
- $f(\cdot)$ : A function representing the graph encoder, as defined in TM4.3.3.
- $s_{\max}$ : Threshold of the cosine similarity measured by  $\cos(\cdot, \cdot)$  in the diversity loss. (Type:  $\mathbb{R}$ )
- $\lambda_1, \lambda_2, \lambda_3$ : Hyper-parameters (see DD2) balancing between prediction accuracy (imposed by cross-entropy loss), and prototype finding (imposed by cluster loss, separation loss, and divergence loss). (Type:  $\mathbb{R}$ )

**Notes:** The cross-entropy loss is a common loss used in many ML models, while the other losses (i.e., cluster loss, separation loss, and diversity loss) are proposed in the context of Re-ProtGNN.

**Source:** [Wikipedia contributors \(2024b\)](#), [Zhang et al. \(2022\)](#)

**Ref. By:** [GD3](#), [GD4](#), [GD5](#), [GD6](#), [IM1](#)

**Preconditions for [TM2:LF](#):** None

**Derivation for [TM2:LF](#):** Not Applicable

---

#### 4.3.4 General Definitions

This section collects the laws and equations that will be used in building the instance models.

|             |   |
|-------------|---|
| Number      | GD1   |
| Label       | <b>Chain Rule</b>   |
| SI Units    | Unitless  |
| Equation    | $\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$   |
| Description | The chain rule states that the derivative of a function with respect to an independent variable can be computed as the product of the derivative of the function with respect to an intermediate variable and the derivative of the intermediate variable with respect to the independent variable. |
| Source      | <a href="#">Wikipedia contributors (2024a)</a>  |
| Ref. By     | <a href="#">GD3</a> , <a href="#">GD4</a> , <a href="#">GD5</a>   |

|             |   |
|-------------|---|
| Number      | GD2   |
| Label       | <b>Sigmoid Function for Fully Connected Layer</b>   |
| SI Units    | Unitless  |
| Equation    | $\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}}, \quad \text{where } z = W^\top h + b$  |
| Description | <p><math>h</math>: Graph embedding obtained after processing input node features <math>X</math> through the GNN encoder (see TM4.3.3). (Type: a vector of <math>\mathbb{R}</math>)</p> <p><math>W</math>: Trainable weight matrix of Fully Connected Layer (Section 4.1.2) that maps graph embeddings to classification results. (Type: a matrix of <math>\mathbb{R}</math>)</p> <p><math>b</math>: Bias term introduced to adjust predictions when feature values are near zero. (Type: a vector of <math>\mathbb{R}</math>)</p> |
| Source      | <a href="#">Wikipedia contributors (2024a)</a>  |
| Ref. By     | GD3, GD4, GD5   |

|             |  |
|-------------|--|
| Number      | GD3  |
| Label       | <b>Gradient of Loss Function with respect to the bias b of Fully Connected Layer</b>   |
| SI Units    | Unitless   |
| Equation    | $\nabla_b \mathcal{L} = (\hat{y} - y)$   |
| Description | With $\hat{y}$ defined by the Sigmoid Function (see GD2), we derive the gradient of Loss Function (see TM4.3.3) with respect to $b$ using the chain rule (see GD1). This gradient allows us to minimize Loss Function during model training (see IM1). |
| Source      | <a href="#">Turin (2020)</a>   |
| Ref. By     | IM1  |



|             |  |
|-------------|--|
| Number      | GD4  |
| Label       | <b>Gradient of Loss Function with respect to the weight <math>W</math> of Fully Connected Layer</b>  |
| SI Units    | Unitless   |
| Equation    | $\nabla_W \mathcal{L} = (\hat{y} - y)h$  |
| Description | With $\hat{y}$ defined by the Sigmoid Function (see GD2), we derive the gradient of Loss Function (see TM4.3.3) with respect to $W$ using the chain rule (see GD1). This gradient allows us to minimize Loss Function during model training (see IM1). |
| Source      | <a href="#">Turin (2020)</a>   |
| Ref. By     | IM1  |

|             |   |
|-------------|---|
| Number      | GD5   |
| Label       | <b>Gradient of Loss Function with respect to the weight <math>W_g</math> of GNN Encoder</b>   |
| SI Units    | Unitless  |
| Equation    | $\nabla_{W_g} \mathcal{L} = (\hat{y} - y)W \sum_{u \in \mathcal{N}(v)} \tilde{A}_{vu} h_u$  |
| Description | <p><math>W_g</math>: Trainable weight matrix of GNN Encoder (see TM4.3.3) that maps input graphs to graph embeddings. (Type: a matrix of <math>\mathbb{R}</math>)</p> <p>With <math>\hat{y}</math> defined by the Sigmoid Function (see GD2), we derive the gradient of Loss Function (see TM4.3.3) with respect to <math>W_g</math> using the chain rule (see GD1). This gradient allows us to minimize Loss Function during model training (see IM1).</p> |
| Source      | <a href="#">Turin (2020)</a>  |
| Ref. By     | IM1   |

|             |  |
|-------------|--|
| Number      | GD6  |
| Label       | <b>Gradient of Loss Function with respect to P</b>   |
| SI Units    | Unitless   |
| Equation    | $\nabla_P \mathcal{L} = 2(h_i - p_j) - 2(h_i - p_k) + \sum_{j \neq i} \mathbf{1}[\cos(p_i, p_j) > s_{\max}] \left( \frac{p_j}{\ p_i\  \ p_j\ } - \frac{(p_i \cdot p_j) p_i}{\ p_i\ ^3 \ p_j\ } \right)$  |
| Description | <p><math>h_i</math>: Embedding representation of input graph <math>i</math>, computed by the graph encoder <math>f(\cdot)</math> (see TM4.3.3). (Type: a vector of <math>\mathbb{R}</math>)</p> <p>We derive the gradient of the Loss Function (see TM4.3.3) with respect to P. This allows us to minimize the Loss Function for training the model (see IM1).</p> |
| Source      | <a href="#">Turin (2020)</a>   |
| Ref. By     | IM1  |

### Detailed derivation of gradients for Re-ProtGNN

To optimize Re-ProtGNN, we minimize the objective function:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{\text{Clst}} + \lambda_2 \mathcal{L}_{\text{Sep}} + \lambda_3 \mathcal{L}_{\text{Div}}$$

where

- $\mathcal{L}_{CE}$ : Cross-entropy loss for classification.
- $\mathcal{L}_{\text{Clst}}$ : Cluster loss ensuring embeddings stay close to correct prototypes.
- $\mathcal{L}_{\text{Sep}}$ : Separation loss ensuring embeddings are far from incorrect prototypes.
- $\mathcal{L}_{\text{Div}}$ : Diversity loss preventing prototype collapse.

### Step 1: Gradient of Cross-Entropy Loss on W and b

For a given prediction  $\hat{y}$  with label  $y$ , the cross-entropy loss is

$$\mathcal{L}_{CE} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}).$$

Taking the derivative with respect to  $\hat{y}$  gives

$$\frac{d}{d\hat{y}} \mathcal{L}_{CE} = -\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}}.$$

Using the property of the sigmoid function

$$\hat{y} = \sigma(Z) = \frac{1}{1 + e^{-z}},$$

We differentiate  $\hat{y}$  with respect to  $z$

$$\frac{d}{dZ}\sigma(Z) = \sigma(Z)(1 - \sigma(Z)) = \hat{y}(1 - \hat{y})$$

and apply the chain rule

$$\frac{d}{dZ}\mathcal{L}_{CE} = (\hat{y} - y).$$

Since  $Z = Wh + b$ , we apply the chain rule again to get the gradients with respect to  $W$  and  $b$ :

$$\begin{aligned}\frac{d\mathcal{L}_{CE}}{dW} &= \frac{d}{dZ} \frac{dZ}{dW} \mathcal{L}_{CE} = (\hat{y} - y)h, \\ \frac{d\mathcal{L}_{CE}}{db} &= \frac{d}{dZ} \frac{dZ}{db} \mathcal{L}_{CE} = (\hat{y} - y).\end{aligned}$$

## Step 2: Gradient of Cross-Entropy Loss on $W_g$

The GNN encoder generates the embeddings  $h$  using message passing:

$$h_v = \sigma \left( \sum_{u \in \mathcal{N}(v)} \tilde{A}_{vu} W_g h_u \right),$$

where  $\tilde{A}$  is the normalized adjacency matrix,  $W_g$  is the trainable weight matrix of the GNN, and  $\sigma(\cdot)$  is a non-linear activation function.

Taking the derivative of the loss function with respect to  $W_g$ , we apply the chain rule:

$$\nabla_{W_g} \mathcal{L} = \frac{d\mathcal{L}}{d\hat{y}} \cdot \frac{d\hat{y}}{dh} \cdot \frac{dh}{dW_g}.$$

From Step 1, we already derived:

$$\frac{d\mathcal{L}}{d\hat{y}} = (\hat{y} - y),$$

and

$$\frac{d\hat{y}}{dh} = W.$$

Since the GNN encoder propagates node features through message passing, differentiating  $h$  with respect to  $W_g$  gives:

$$\nabla_{W_g} h_v = \sum_{u \in \mathcal{N}(v)} \tilde{A}_{vu} h_u.$$

Thus, combining these terms, the final gradient is:

$$\nabla_{W_g} \mathcal{L} = (\hat{y} - y)W \sum_{u \in \mathcal{N}(v)} \tilde{A}_{vu} h_u.$$

This expression shows that computing  $\nabla_{W_g} \mathcal{L}$  requires propagating gradients through the entire graph, making the computation iterative rather than explicit.

### Step 3: Gradient of Prototype Cluster Loss on P

The cluster loss is defined as

$$\mathcal{L}_{\text{Clst}} = \sum_{i=1}^n \min_{j: p_j \in P_{y_i}} \|h_i - p_j\|^2$$

Differentiating with respect to  $p_j$ ,

$$\frac{d}{dp_j} \mathcal{L}_{\text{Clst}} = -2(h_i - p_j)$$

### Step 4: Gradient of Prototype Separation Loss on P

The separation loss is

$$\mathcal{L}_{\text{Sep}} = - \sum_{i=1}^n \min_{j: p_j \notin P_{y_i}} \|h_i - p_j\|^2$$

Differentiating with respect to  $p_j$

$$\frac{d}{dp_j} \mathcal{L}_{\text{Sep}} = 2(h_i - p_j)$$

### Step 5: Gradient of the Diversity Loss on P

To prevent prototype collapse, we enforce diversity via cosine similarity

$$\mathcal{L}_{\text{Div}} = \sum_{k=1}^C \sum_{i \neq j, p_i, p_j \in P_k} \max(0, \cos(p_i, p_j) - s_{\max})$$

where

$$\cos(p_i, p_j) = \frac{p_i \cdot p_j}{\|p_i\| \|p_j\|}$$

Using the gradient of cosine similarity

$$\frac{d}{dp_i} \cos(p_i, p_j) = \frac{p_j}{\|p_i\| \|p_j\|} - \frac{(p_i \cdot p_j) p_i}{\|p_i\|^3 \|p_j\|}$$

The final update step for prototype diversity

$$\frac{d}{dp_i} \mathcal{L}_{\text{Div}} = \sum_{j \neq i} \mathbf{1}[\cos(p_i, p_j) > s_{\max}] \left( \frac{p_j}{\|p_i\| \|p_j\|} - \frac{(p_i \cdot p_j) p_i}{\|p_i\|^3 \|p_j\|} \right)$$

## Gradients Summary

To sum up, Cross-entropy loss  $\mathcal{L}_{CE}$  updates the weight  $W$  and the bias  $b$  of the Fully Connected Layer, as well as updates the weight  $W_g$  of GNN Encoder. Cluster Loss, separation Loss, and diversity Loss update the prototype  $P$ :

$$\begin{aligned}\frac{d\mathcal{L}}{db} &= \frac{d\mathcal{L}_{CE}}{db}, \\ \frac{d\mathcal{L}}{dW} &= \frac{d\mathcal{L}_{CE}}{dW}, \\ \frac{d\mathcal{L}}{dW_g} &= \frac{d\mathcal{L}_{CE}}{dW_g}, \\ \frac{d\mathcal{L}}{dP} &= \frac{d\mathcal{L}_{Clst}}{dP} + \frac{d\mathcal{L}_{Sep}}{dP} + \frac{d\mathcal{L}_{Div}}{dP}.\end{aligned}$$

This leads to the final formula seen in GD3, GD4, GD5, GD6.

### 4.3.5 Data Definitions

This section collects and defines all the data needed to build the instance models. The dimension of each quantity is also given.

|             |  |
|-------------|--|
| Number      | DD1  |
| Label       | <b>Learning Rate</b>   |
| Symbol      | $\eta$   |
| SI Units    | Unitless   |
| Equation    | Not Applicable   |
| Description | The learning rate is a hyperparameter that controls the step size at each iteration while moving toward a minimum of the loss function. It plays an important role in balancing convergence speed and stability. |
| Sources     | <a href="#">Wikipedia contributors (2024d)</a>   |
| Ref. By     | IM1  |

|             |  |
|-------------|--|
| Number      | DD2  |
| Label       | <b>Regularization Parameters</b>   |
| Symbol      | $\lambda_1, \lambda_2, \lambda_3$  |
| SI Units    | Unitless   |
| Equation    | Not Applicable   |
| Description | The regularization parameters control the trade-off between the main classification loss and additional constraints (see TM4.3.3) to enhance interpretability and prevent overfitting. |
| Sources     | <a href="#">Wikipedia contributors (2024e)</a>   |
| Ref. By     | TM4.3.3  |

#### 4.3.6 Data Types

Not applicable, as the inputs and outputs consist of straightforward types, such as sequences of integers or matrices of real numbers.

#### 4.3.7 Instance Models

This section transforms the problem defined in Section 4.1 into one which is expressed in mathematical terms. It uses concrete symbols defined in Section 4.3.5 to replace the abstract symbols in the models identified in Sections 4.3.3 and 4.3.4.

The goal GS1 is achieved through IM1 and IM2. IM1 trains the model to ensure that IM2 delivers accurate classifications and reliable interpretations.

|             |   |
|-------------|---|
| Number      | IM1   |
| Label       | <b>Model Training via Gradient Descent</b>  |
| Input       | Node features $X$ , adjacency matrix $A$ , prototypes $P$ , learning rate $\eta$ .  |
| Output      | Optimized parameters $W, b, P$  |
| Description | <p>The model is trained by iteratively updating the parameters via gradient descent:</p> <ol style="list-style-type: none"> <li>1. Compute embeddings using GNN layers (see TM4.3.3).</li> <li>2. Compute the loss function <math>\mathcal{L}</math> (see TM4.3.3).</li> <li>3. Update parameters using <math>W \leftarrow W - \eta \frac{\partial \mathcal{L}}{\partial W}</math>, <math>b \leftarrow b - \eta \frac{\partial \mathcal{L}}{\partial b}</math>, <math>P \leftarrow P - \eta \frac{\partial \mathcal{L}}{\partial P}</math> (see DD1, GD3, GD4, GD5, GD6, TM4.3.3).</li> </ol> |
| Sources     | <a href="#">Zhang et al. (2022)</a>   |
| Ref. by     | A3, R2, GD3, GD4, GD5, GD6, IM2   |

|             |   |
|-------------|---|
| Number      | IM2   |
| Label       | <b>Model Inference via Forward Propagation</b>  |
| Input       | Node features $X$ , adjacency matrix $A$ , trained parameters $W, b, P$ .   |
| Output      | Predicted class labels $\hat{y}$ .  |
| Description | During inference, the model forwards input features through the trained model (see IM1) and directly gets the classification results. |
| Sources     | <a href="#">Run:AI (2024)</a>   |
| Ref. by     | A3, R3  |

#### 4.3.8 Input Data Constraints

Table 3 shows the data constraints on the input output variables. The column for physical constraints gives the physical limitations on the range of values that can be taken by the variable. The column for software constraints restricts the range of inputs to reasonable values. The software constraints will be helpful in the design stage for picking suitable algorithms. The constraints are conservative, to give the user of the model the flexibility to experiment with unusual situations. The column of typical values is intended to provide a feel for a common scenario. The uncertainty column provides an estimate of the confidence with which the physical quantities can be measured. This information would be part of the

input if one were performing an uncertainty quantification exercise.

The specification parameters in Table 3 are listed in Table 5.

Table 3: Input Variables

| Var | Physical Constraints               | Software Constraints               | Typical Value          | Uncertainty |
|-----|------------------------------------|------------------------------------|------------------------|-------------|
| $X$ | $X[i, j] \in \{-\infty, +\infty\}$ | $X[i, j] \in \{-\infty, +\infty\}$ | depends on the dataset | -           |
| $A$ | $A[i, j] \in \{0, 1\}$             | $A[i, j] \in \{0, 1\}$             | depends on the dataset | -           |
| $y$ | $y \in \{-\infty, +\infty\}$       | $y \in \{1, \dots, C\}$            | depends on the dataset | -           |

Table 5: Specification Parameter Values

| Var | Value                  |
|-----|------------------------|
| $C$ | depends on the dataset |

#### 4.3.9 Properties of a Correct Solution

A correct solution must ensure that the predicted class is a valid label and that the prototypes are valid graphs, following the constraints in Table 7. In addition, each prototype should be a subgraph of at least one input graph.

Table 7: Output Variables

| Var       | Physical Constraints                                 |
|-----------|--|
| $X_p$     | $num\_rows(X_p) < num\_rows(X)$                      |
| $A_p$     | $num\_rows(A_p) = num\_cols(A_p), dim(A_p) < dim(A)$ |
| $\hat{y}$ | $\hat{y} \in \{1, \dots, C\}$                        |

## 5 Requirements

This section provides the functional requirements, the business tasks that the software is expected to complete, and the nonfunctional requirements, the qualities that the software is



expected to exhibit.

## 5.1 Functional Requirements

- R1: The system should be able to load the MUTAG dataset [Debnath et al. \(1991\)](#) .
- R2: During the training phase, the system should update model parameters and learn prototype representations by optimizing the Loss Function (see IM1).
- R3: During the inference phase, the system should compute classification accuracy. (see IM2).

## 5.2 Nonfunctional Requirements

- NFR1: **Accuracy** The system should achieve a satisfactory classification accuracy ( $\geq 80\%$ ) on the MUTAG dataset [Debnath et al. \(1991\)](#).
- NFR2: **Usability** The system should visualize learned prototypes as graphs represented by images rather than node features and adjacency matrices to enhance interpretability.
- NFR3: **Portability** Re-ProtGNN should be compatible with Windows and macOS and run with PyTorch 1.8.0 and Torch-Geometric 2.0.2.

## 5.3 Rationale

The rationale for the scope decisions (Section 4.3) is that Re-ProtGNN uses the MUTAG dataset [Debnath et al. \(1991\)](#) because the original ProtGNN model was evaluated on it. Additionally, MUTAG is a well-established benchmark for graph classification, ensuring a fair validation of ProtGNN’s performance.

In addition, the model is designed to follow the structure of the original ProtGNN framework to ensure consistency.

## 6 Likely Changes

- LC1: A3 The system may be extended to support training on a broader range of datasets, such as BA-shape, to further validate the reproducibility and robustness of ProtGNN’s results.
- LC2: The model may use alternative GNN encoders, such as Graph Isomorphism Networks (GINs), to evaluate the impact of different architectures on classification accuracy and interpretability.

## 7 Unlikely Changes

LC3: It is unlikely that the system will transition to a fully unsupervised setting, as ProtGNN relies on labeled data to learn prototype representations.

LC4: The system is unlikely to remove the prototype-based interpretability mechanism, as this is a fundamental aspect of ProtGNN.

## 8 Traceability Matrices and Graphs

The purpose of the traceability matrices is to provide easy references on what has to be additionally modified if a certain component is changed. Every time a component is changed, the items in the column of that component that are marked with an “X” may have to be modified as well. Table 9 shows the dependencies of theoretical models, general definitions, data definitions, and instance models with each other. Table 10 shows the dependencies of instance models, requirements, and data constraints on each other. Table 11 shows the dependencies of theoretical models, general definitions, data definitions, instance models, and likely changes on the assumptions.

|         | TM4.3.3 | TM4.3.3 | GD1 | GD2 | GD3 | GD4 | GD5 | GD6 | DD1 | DD2 | IM1 | IM2 |
|---------|---------|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| TM4.3.3 |         |         |     |     |     |     |     |     |     |     |     |     |
| TM4.3.3 | X       |         |     |     |     |     |     |     |     |     |     |     |
| GD1     |         |         |     |     |     |     |     |     |     |     |     |     |
| GD2     | X       |         |     |     |     |     |     |     |     |     |     |     |
| GD3     | X       | X       | X   | X   | X   | X   | X   | X   | X   | X   | X   |     |
| GD4     | X       | X       | X   | X   | X   | X   | X   | X   | X   | X   | X   |     |
| GD5     | X       | X       | X   | X   | X   | X   | X   | X   | X   | X   | X   |     |
| GD6     | X       | X       | X   | X   | X   | X   | X   | X   | X   | X   | X   |     |
| DD1     |         |         |     |     |     |     |     |     |     |     |     |     |
| DD2     | X       |         |     |     |     |     |     |     |     |     |     |     |
| IM1     | X       | X       | X   | X   | X   | X   | X   | X   | X   | X   |     |     |
| IM2     | X       | X       | X   | X   | X   | X   | X   | X   | X   | X   | X   |     |

Table 9: Traceability Matrix Showing the Connections Between Items of Different Sections

## 9 Development Plan

Not applicable.

|       | IM1 | IM2 | R1 | R2 | R3 | NFR1 | NFR2 | NFR?? | NFR3 |
|-------|-----|-----|----|----|----|------|------|-------|------|
| IM1   |     |     |    |    |    |      |      |       |      |
| IM2   | X   |     |    |    |    |      |      |       |      |
| R1    |     |     |    |    |    |      |      |       |      |
| R2    | X   |     |    |    |    |      |      |       |      |
| R3    | X   | X   |    |    |    |      |      |       |      |
| NFR1  |     |     |    |    |    |      |      |       |      |
| NFR2  |     |     |    |    |    |      |      |       |      |
| NFR?? |     |     |    |    |    |      |      |       |      |
| NFR3  |     |     |    |    |    |      |      |       |      |

Table 10: Traceability Matrix Showing the Connections Between Requirements and Instance Models

## 10 Values of Auxiliary Constants

Not applicable, as no auxiliary constants are defined in this document.

|         | A1 | A2 | A3 |
|---------|----|----|----|
| TM4.3.3 | X  | X  |    |
| TM4.3.3 | X  | X  |    |
| GD1     |    |    |    |
| GD2     | X  |    |    |
| GD3     | X  | X  | X  |
| GD4     | X  | X  | X  |
| GD5     | X  | X  | X  |
| GD6     | X  | X  | X  |
| DD1     |    |    |    |
| DD2     | X  | X  |    |
| IM1     | X  | X  | X  |
| IM2     |    |    | X  |
| LC1     |    |    | X  |
| LC2     |    |    |    |
| LC3     |    |    |    |
| LC4     |    |    |    |

Table 11: Traceability Matrix Showing the Connections Between Assumptions and Other Items

## References

- Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds: Correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34(2):786–797, February 1991. doi: 10.1021/jm00106a046.
- Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, September 2016. URL <https://arxiv.org/abs/1609.02907>.
- Run:AI. Understanding machine learning inference, 2024. URL <https://www.run.ai/guides/machine-learning-inference/understanding-machine-learning-inference>. Accessed: February 6, 2025.
- ScienceDirect. Cosine similarity, 2024. URL <https://www.sciencedirect.com/topics/computer-science/cosine-similarity>. Accessed: February 6, 2025.
- W. Spencer Smith and Lei Lai. A new requirements template for scientific computing. In J. Ralyté, P. Ågerfalk, and N. Kraiem, editors, *Proceedings of the First International Workshop on Situational Requirements Engineering Processes – Methods, Techniques and Tools to Support Situation-Specific Requirements Engineering Processes, SREP’05*, pages 107–121, Paris, France, 2005. In conjunction with 13th IEEE International Requirements Engineering Conference.
- W. Spencer Smith, Lei Lai, and Ridha Khedri. Requirements analysis for engineering computation: A systematic approach for improving software reliability. *Reliable Computing, Special Issue on Reliable Engineering Computation*, 13(1):83–107, February 2007.
- Arseny Turin. Logistic regression from scratch. <https://towardsdatascience.com/logistic-regression-from-scratch-69db4f587e17>, 2020.
- Wikipedia contributors. Chain rule — Wikipedia, the free encyclopedia, 2024a. URL [https://en.wikipedia.org/wiki/Chain\\_rule](https://en.wikipedia.org/wiki/Chain_rule). Accessed: February 6, 2025.
- Wikipedia contributors. Cross-entropy — Wikipedia, the free encyclopedia, 2024b. URL <https://en.wikipedia.org/wiki/Cross-entropy>. Accessed: February 6, 2025.
- Wikipedia contributors. Gradient — Wikipedia, the free encyclopedia, 2024c. URL <https://en.wikipedia.org/wiki/Gradient>. Accessed: February 6, 2025.
- Wikipedia contributors. Learning rate — Wikipedia, the free encyclopedia, 2024d. URL [https://en.wikipedia.org/wiki/Learning\\_rate](https://en.wikipedia.org/wiki/Learning_rate). Accessed: February 6, 2025.

Wikipedia contributors. Regularization (mathematics) — Wikipedia, the free encyclopedia, 2024e. URL [https://en.wikipedia.org/wiki/Regularization\\_\(mathematics\)](https://en.wikipedia.org/wiki/Regularization_(mathematics)). Accessed: February 6, 2025.

Wikipedia contributors. Rectifier (neural networks) — Wikipedia, the free encyclopedia, 2024f. URL [https://en.wikipedia.org/wiki/Rectifier\\_\(neural\\_networks\)](https://en.wikipedia.org/wiki/Rectifier_(neural_networks)). Accessed: February 6, 2025.

Yuanqi Xue. Problem statement and goals for re-protgnn, 2025. URL <https://github.com/Yuanqi-X/Re-ProtGNN/blob/main/docs/ProblemStatementAndGoals/ProblemStatement.pdf>. Accessed: February 6, 2025.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. Protgnn: Towards self-explaining graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9127–9135, June 2022.

## Appendix — Reflection

[Not required for CAS 741 —SS]

The information in this section will be used to evaluate the team members on the graduate attribute of Lifelong Learning.

The purpose of reflection questions is to give you a chance to assess your own learning and that of your group as a whole, and to find ways to improve in the future. Reflection is an important part of the learning process. Reflection is also an essential component of a successful software development process.

Reflections are most interesting and useful when they're honest, even if the stories they tell are imperfect. You will be marked based on your depth of thought and analysis, and not based on the content of the reflections themselves. Thus, for full marks we encourage you to answer openly and honestly and to avoid simply writing "what you think the evaluator wants to hear."

Please answer the following questions. Some questions can be answered on the team level, but where appropriate, each team member should write their own response:

1. What went well while writing this deliverable?
2. What pain points did you experience during this deliverable, and how did you resolve them?
3. How many of your requirements were inspired by speaking to your client(s) or their proxies (e.g. your peers, stakeholders, potential users)?
4. Which of the courses you have taken, or are currently taking, will help your team to be successful with your capstone project.
5. What knowledge and skills will the team collectively need to acquire to successfully complete this capstone project? Examples of possible knowledge to acquire include domain specific knowledge from the domain of your application, or software engineering knowledge, mechatronics knowledge or computer science knowledge. Skills may be related to technology, or writing, or presentation, or team management, etc. You should look to identify at least one item for each team member.
6. For each of the knowledge areas and skills identified in the previous question, what are at least two approaches to acquiring the knowledge or mastering the skill? Of the identified approaches, which will each team member pursue, and why did they make this choice?