Comparing GNN Approaches for Molecule Identification

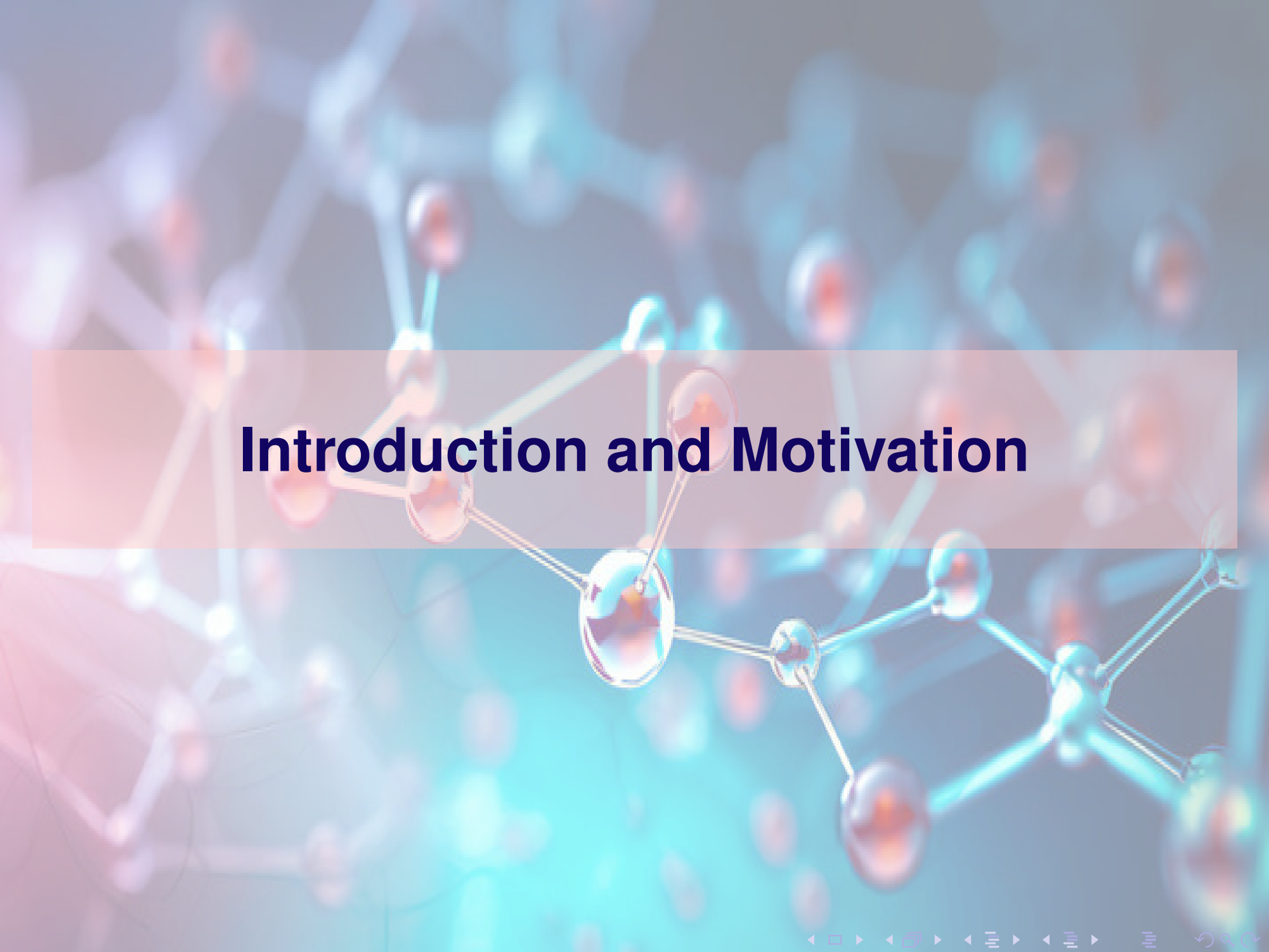
Ivana Milutinović
Teaching Associate

University of Novi Sad,
Faculty of Sciences

Data Science Conference 2024, November 18

About me

- MSc in Artificial Intelligence
- First year PhD student @ University of Novi Sad, Faculty of Sciences
- Teaching Associate (soon Teaching Assistant) @ University of Novi Sad, Faculty of Sciences
- Fields of interest:
  - Machine Learning and Deep Learning
  - Graph Neural Networks
  - Bioinformatics and Computational Biology
  - Medical Data Analysis
- Contact:
  - LinkedIn profile: https://www.linkedin.com/in/ivana-milutinovic1109/
  - Email: ivana.milutinovic@dmi.uns.ac.rs



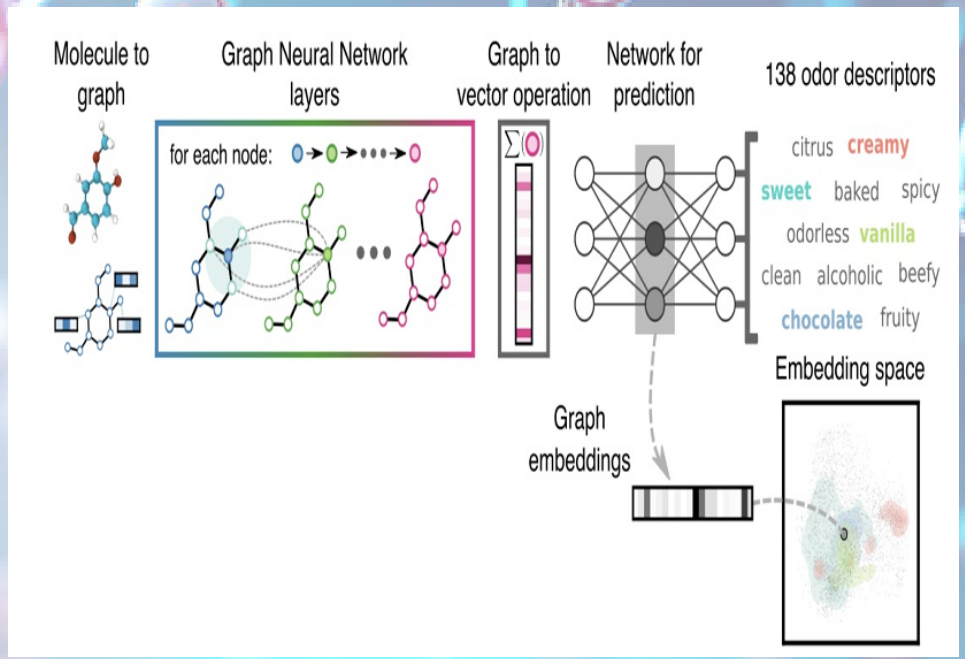
Introduction and Motivation

Introduction and Motivation

- The analysis of molecules in biological processes often requires precise methods to identify their activity
- Traditional approaches mainly rely on heuristic methods and manual analysis of molecular structures, which involve simple calculations and limited chemical parameters
- These approaches in molecule analysis are often not efficient enough to address the complexity of chemical structures and activities

Introduction and Motivation

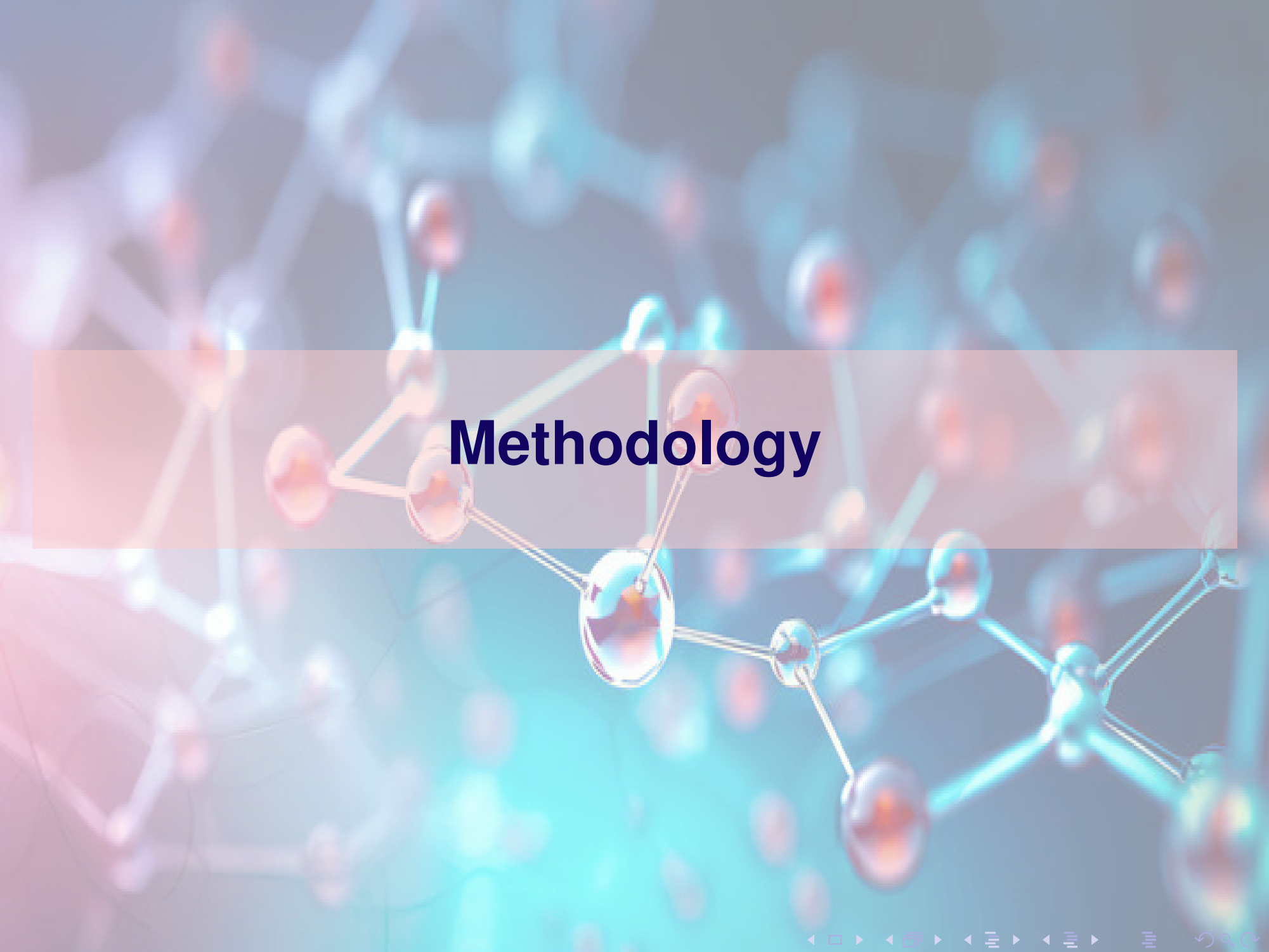
- Advanced methods use machine learning techniques and graph neural networks
- They enable a deeper analysis of molecular structures in the form of graphs, including atomic interactions, molecular topology, and their chemical properties
- They facilitate and improve the accuracy and efficiency in predicting the biological activity of molecules



Introduction and Motivation

- The idea of this work:
  1. Modeling molecules into a graph structure
  2. Applying graph neural network models to predict molecular activity
  3. Analyzing and comparing the obtained results

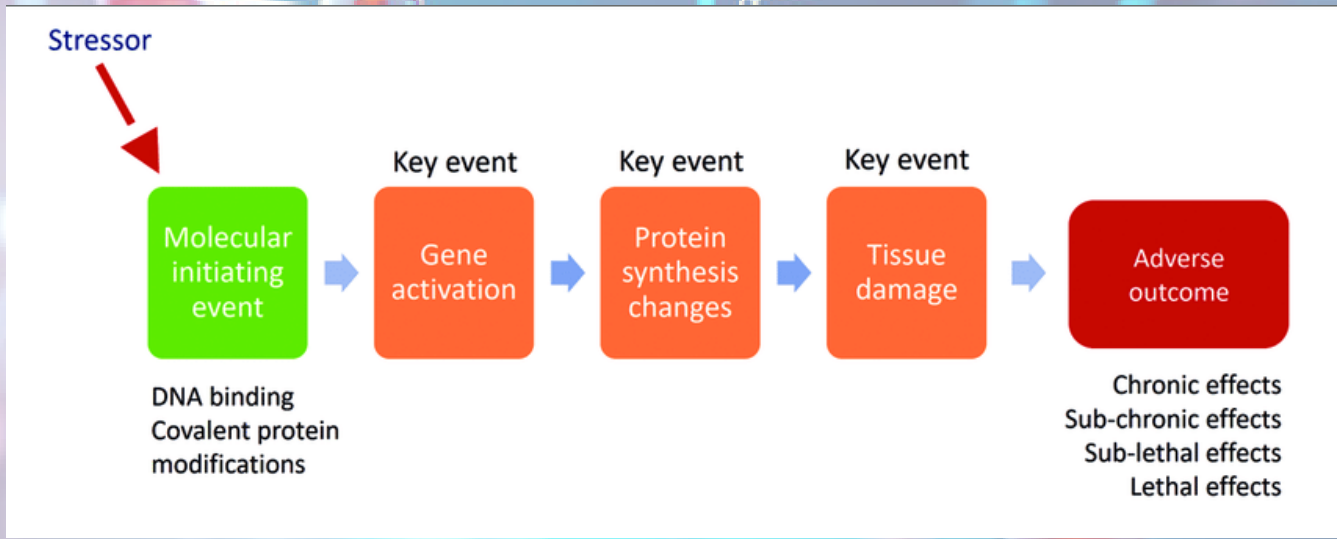
- This work applies supervised learning, where molecules are labeled according to their activity in biological processes, and the model's goal is to predict this activity based on the molecular graph structure.



Methodology

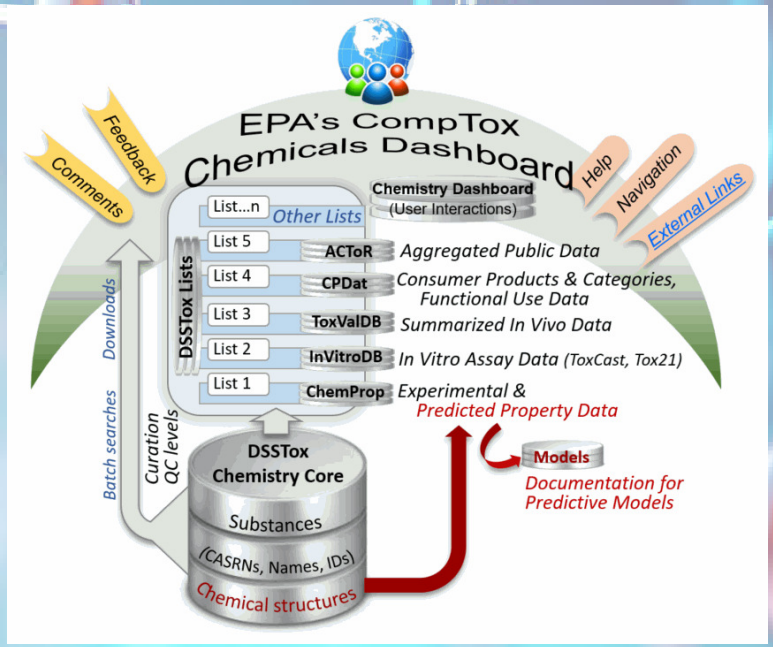
Methodology - Biological Background

- The Adverse Outcome Pathway (AOP) framework connects molecular initiating events (MIEs) through a series of intermediate steps, known as key events (KEs), leading to an adverse outcome (AO)
- Molecular targets identified as MIEs and KEs can help predict the impact of chemicals on the female reproductive system



Methodology - Database

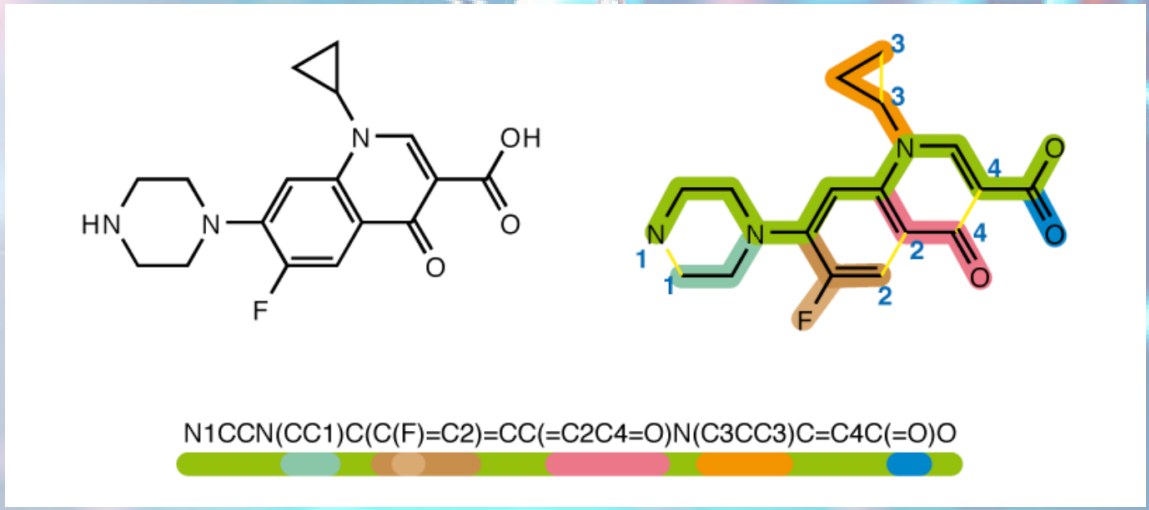
- Data on molecules related to tests relevant to the AOP titled “Reduction of aromatase (CYP19a1) impairing female fertility” (AOP7) was obtained from the CompTox Dashboard
- Data on the activity of molecules tested in these studies was also collected
- The dataset includes information on approximately 11,630 molecules and their activity accross 17 target tests, linked to three biological targets.



Methodology - Preprocessing

- Three main types of information in the dataset are crucial:
  1. Molecule names
  2. Their SMILES representation
  3. Activity in biological assays: inactive, partially active, and active

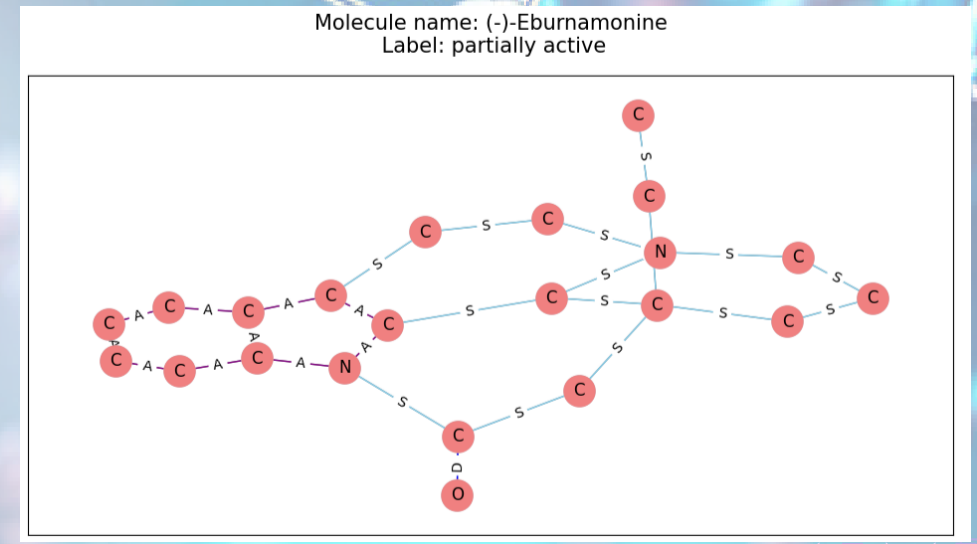
- SMILES (*Simplified Molecular Input Line Entry System*) - notation that encodes the structure of molecules into a line of text
- Allows efficient storage and processing of molecular information



Methodology - Preprocessing

- The ChemIDPlus API Beta database was used to supplement and correct missing molecular information that was not initially available
- The collected molecules are composed of 53 different chemical elements (atoms) and 4 different types of chemical bonds

- Based on the SMILES representation, each molecule was modeled as a graph:
  1. Atoms are represented as nodes, which correspond to letters in the SMILES
  2. Bonds are represented as edges, corresponding to specific symbols in the SMILES



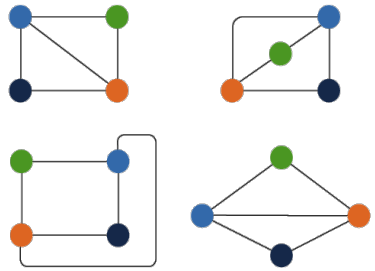
Methodology - Nodes features

- Since rich features for the nodes (atoms) were missing, *DeepWalk* and *Word2Vec* were used to generate their initial vector representations
- For each molecular graph, 150 random walks of length 35 were created
- The walks are used as input sequences for the Word2Vec model, which maps the graph nodes to 64-dimensional embeddings

- The embedding for each node in every graph is combined with the *one-hot* encoded atomic labels (chemical elements).
- Finally, each node embedding has a size of 117:
  - 64-dimensional embedding obtained from the Word2Vec model
  - 53-dimensional encoded atomic label

Methodology - Data Augmentation

- The initial dataset is highly imbalanced, with molecules that are inactive in biological processes predominating
- A data augmentation technique was applied for the graphs of minority classes, based on the creation of isomorphic graphs through node permutations
- The permutation is performed by randomly rearranging the node indices and updating the adjacency matrix to reflect this change



- This method resulted in 20,884 graph-based molecules, of which:
  - 7,291 molecules belong to the inactive class (numerically 0)
  - 6,945 molecules belong to the partially active class (numerically 1)
  - 6,648 molecules belong to the active class (numerically 2)

Methodology - Models

Graph Convolutional Neural Network (GCN)

- Each graph convolution involves two key operations: aggregation and combination of information from neighboring nodes.

GraphSAGE

- The process of learning node representations is based on three key steps: neighbor sampling, aggregation, and combination.

Graph Attention Network (GAT)

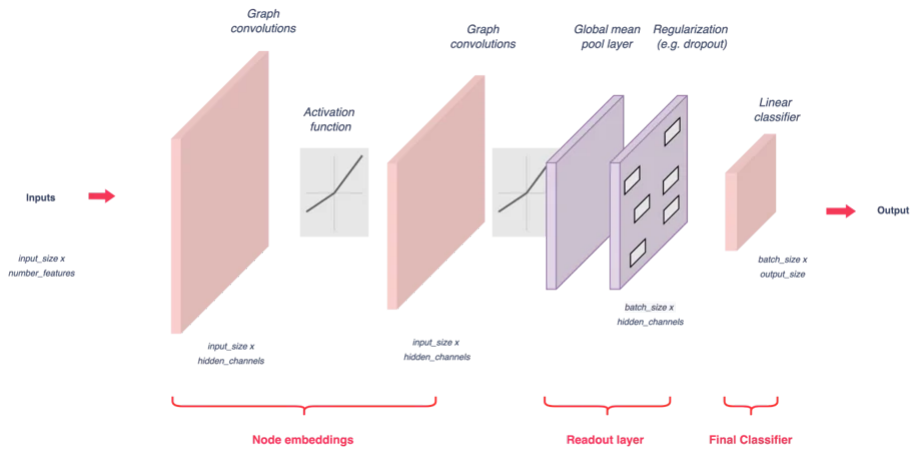
- The attention mechanism enables automatically assignment different weights to neighbors based on their importance during the aggregation of their features.

Graph Isomorphism Network (GIN)

- Includes aggregation methods that allow the distinction between graphs with different structures, even if they are equivalent in terms of their relationships.

Methodology - Hyperparameters

- The *Optuna* library was used for hyperparameter optimization of all the mentioned models through automatic tuning.



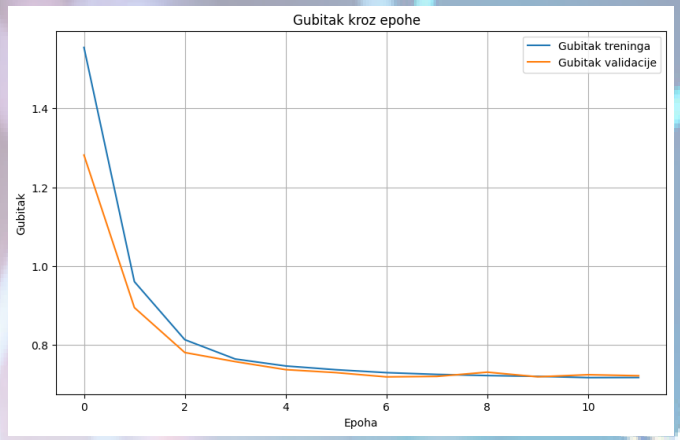
Methodology - Training

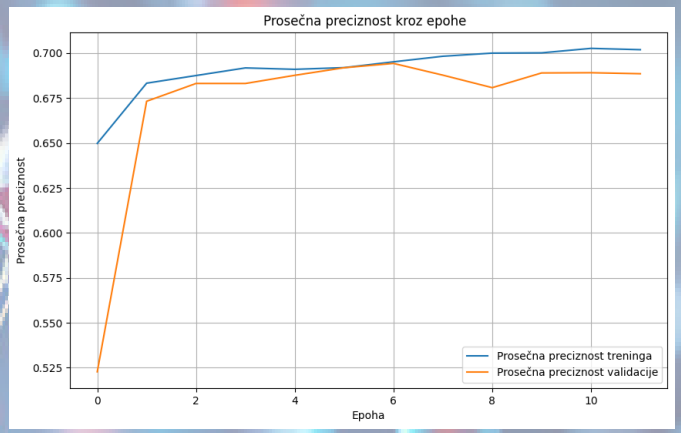
- The data was stratified into training and test sets (80:20), and the training set was further split into training and validation sets (80:20).
- Class labels were converted to *one-hot* format.
- Batches and other hyperparameters were defined based on the results of hyperparameter optimization.
- Various *callbacks* were added:
  - *LearningRateScheduler,*
  - *EarlyStopping,*
  - *F1ScoreCallback*
- Training was run for up to 50 epochs on a GPU-supported cluster.

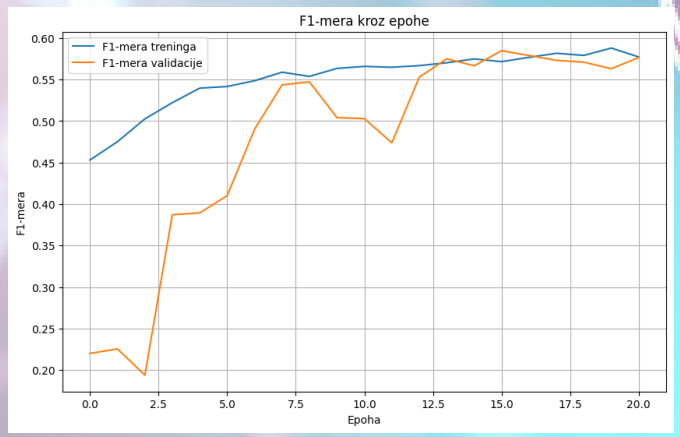
Results

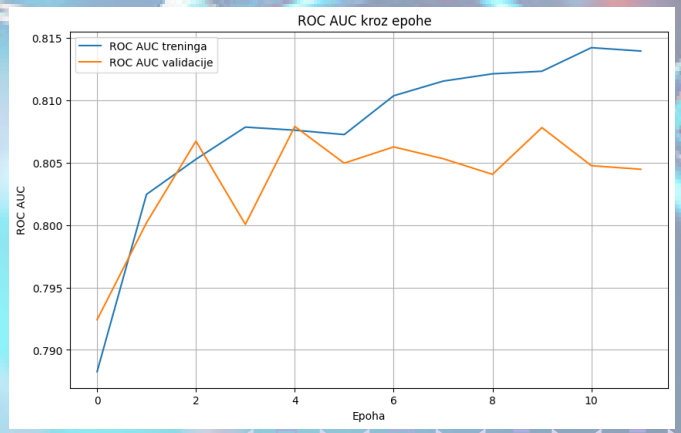
Results - Training metrics values

The training of the GraphSAGE model was the quickest and required the fewest epochs.









Results - Evaluation

- Metrics used for model evaluation:
  - Loss function - Categorical Cross Entropy (Softmax)
  - Accuracy
  - F1-score
  - Area under the ROC curve (*ROC-AUC*)
  - Average precision

Average values for the mentioned metrics with standard deviation obtained from evaluating models on the test set

Model	Loss	Accuracy (%)	F1-score (%)	ROC-AUC (%)	Average Precision (%)
GCN	0.7578±0.002	62.58±0.552	54.44±1.7	79.42±0.19	66.6±0.24
GAT	0.7454±0.009	62.94±1.65	55.63±5.04	78.93±0.52	65.29±0.96
GIN	0.9681±0.016	63.33±0.37	57.48±0.405	80.03±0.31	67.73±0.57
GraphSAGE	**0.7178±0.006**	**64.27±0.43**	**60.64±0.54**	**81±0.06**	**69.51±0.08**

Ideas for future work

Ideas for future work

Idea 1

Increasing the number of features for each node in the graph, i.e., incorporating natural, physicochemical properties of atoms and molecules.

Idea 2

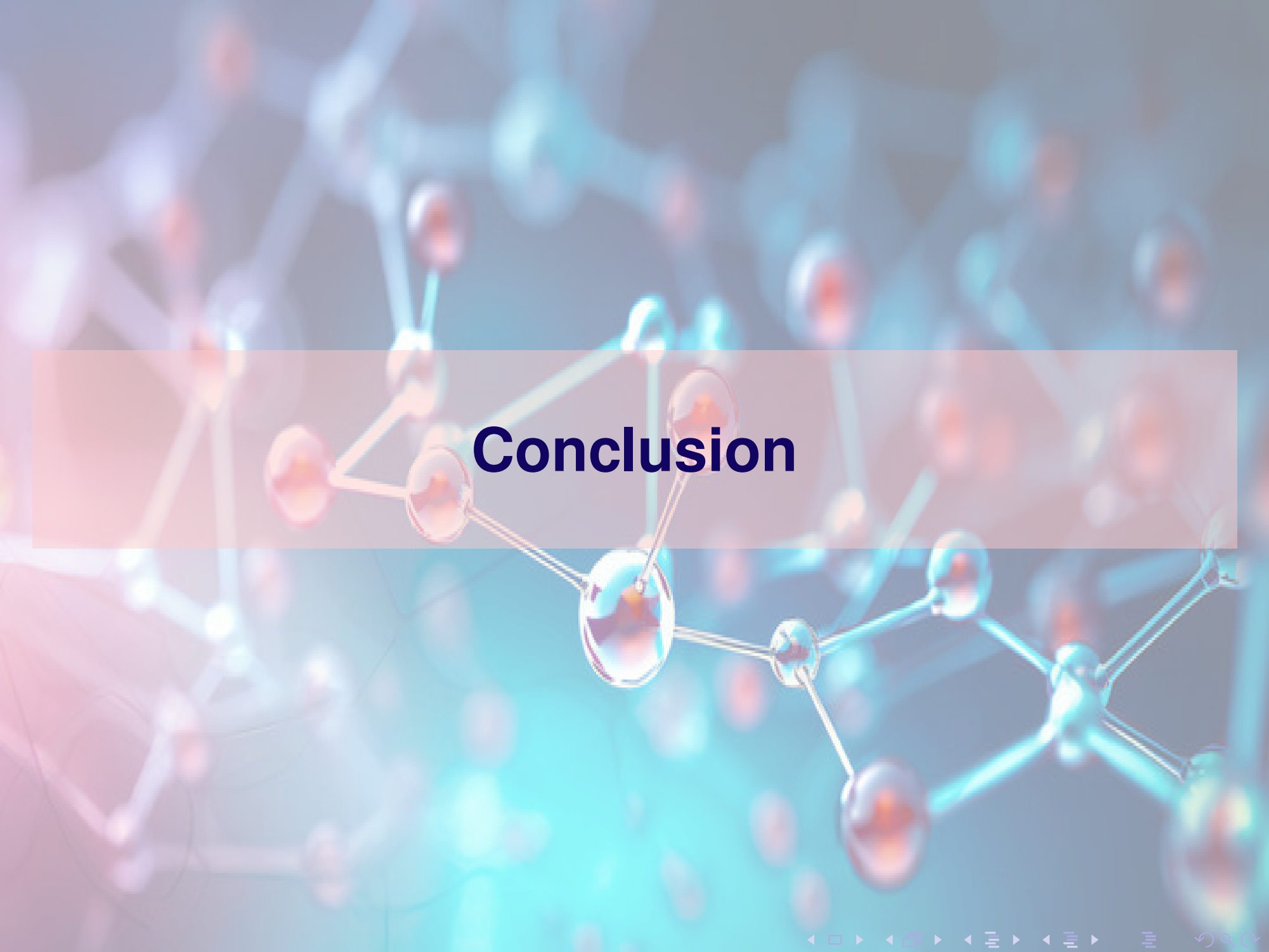
Exploring different data augmentation techniques, ensuring that these methods are biologically and chemically valid to preserve molecular characteristics.

Idea 3

Investigating deeper or more complex architectures, such as combining different graph neural network models (e.g., GCN, GAT, GraphSAGE) to enhance performance and capture diverse graph features.

Idea 4

Directly obtaining molecular data in graph format, without the need for conversion from one representation to another.



Conclusion

Conclusion

- We covered the demonstration of applying Graph Neural Networks to a problem in the fields of biology and chemistry
- The procedures for data acquisition, preprocessing and augmentation were explained, as well as the applied models and the training process
- The results of training and evaluation were presented
- Ideas for improving the obtained results in future work were outlined


Thank you
for your attention!
Q&A
Feel free to ask any question!