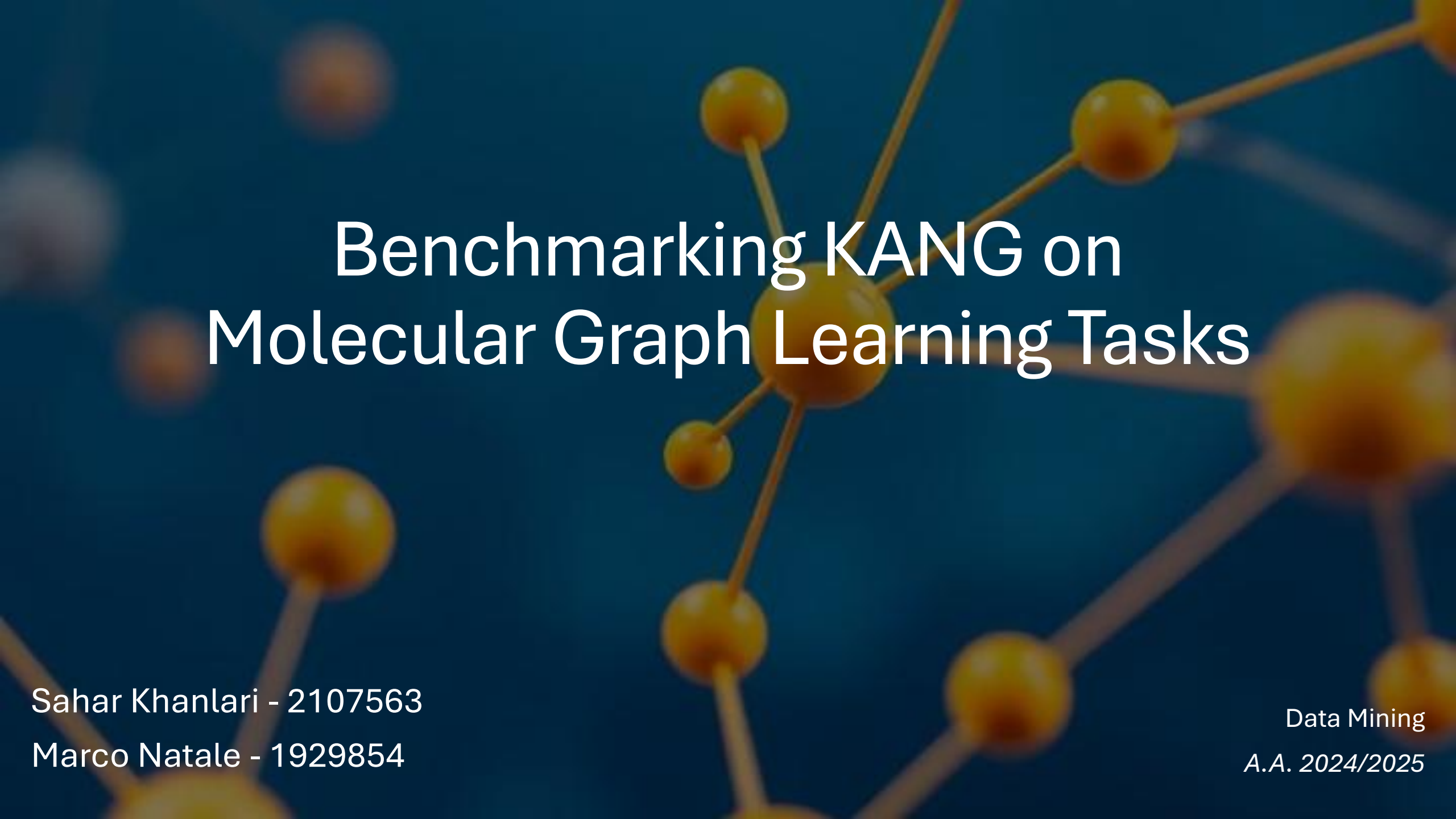


Benchmarking KANG on Molecular Graph Learning Tasks

Sahar Khanlari - 2107563

Marco Natale - 1929854

Data Mining

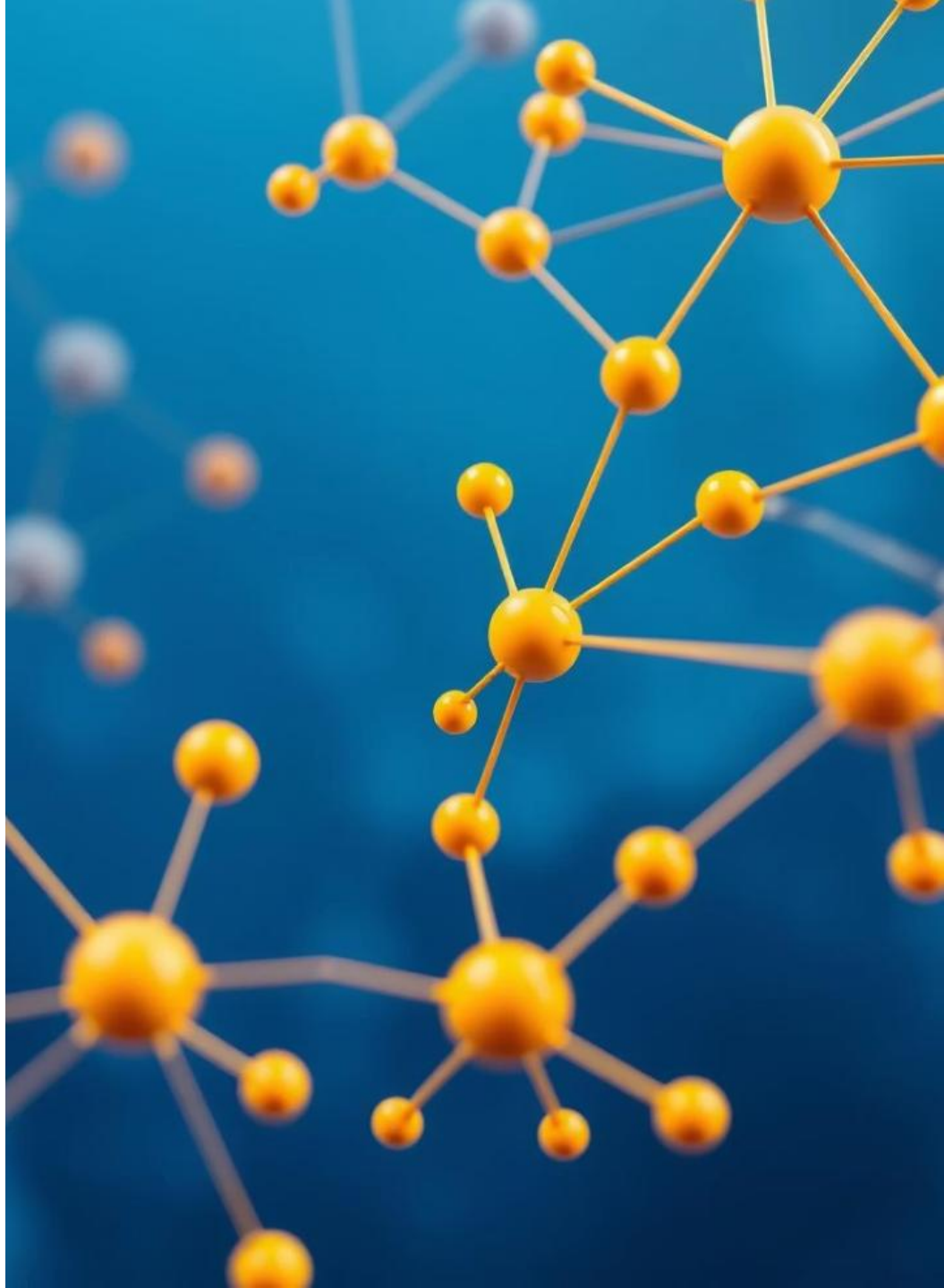
A.A. 2024/2025

Introduction

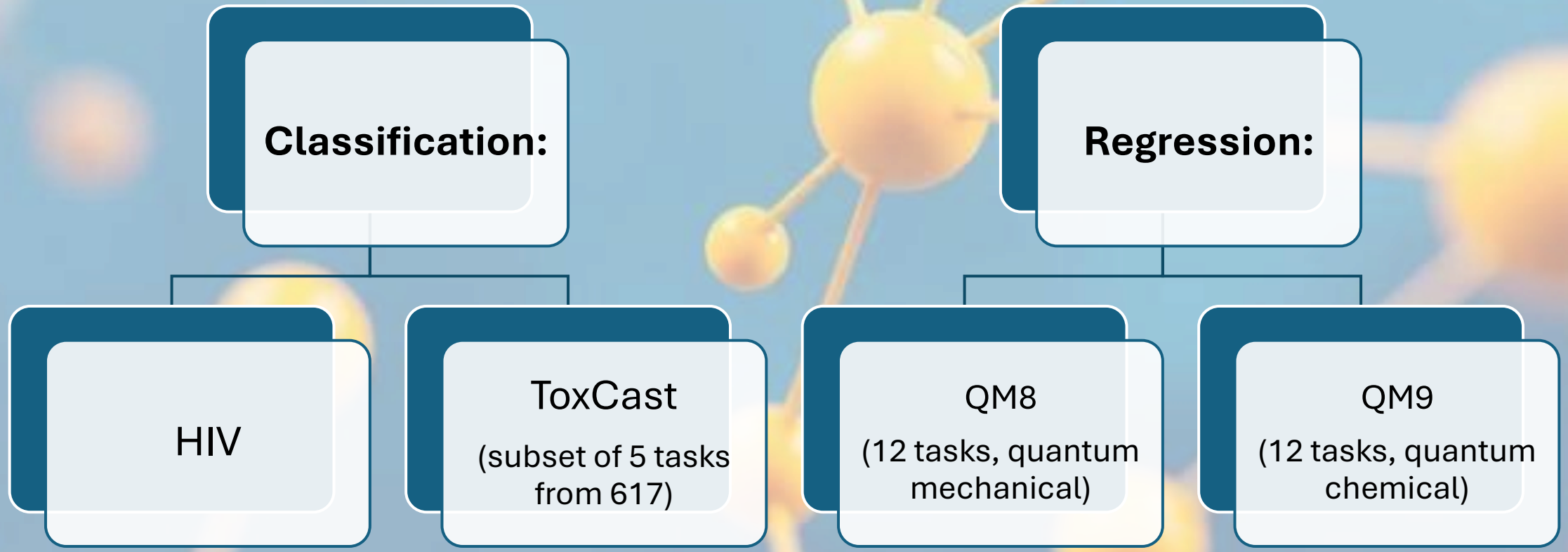
Objective: Evaluate the performance of the KANG model across classification and regression tasks using MoleculeNet datasets.

Molecular property prediction is crucial for drug discovery.

Only 2D molecular structures



Overview of Datasets



Experimental Setup

- SMILES to graph conversion with **RDKit**
- Graph generation: atoms → nodes, bonds → edges
- Hyperparameter tuning with **Optuna**
- Best val score saved and evaluated on test set

Data Representation

Node features:

- Atomic number
- Degree
- Formal charge
- Aromaticity
- Total hydrogens
- Chirality
- Hybridization
- Atomic mass

Edge features:

- Bond type
- Conjugation
- Ring membership
- Stereochemistry

Improvements

- Improve performance of the KANG model on regression and classification molecular datasets.
- Address limitations in feature representation, training configuration, and loss function design.

Methodology Updates

- **Feature Engineering**
  - Added three new atom features.
  - Normalized two numeric atom features (others were one-hot already).
- **Training Configuration**
  - Expanded hyperparameters search grid.
  - Compressed grid values to [-1.1, 1.1].
  - Extended epochs and early stopping patience.
- **New Classification Loss**
  - Implemented Focal Loss for graph classification tasks.
  - Helps tackle class imbalance by focusing learning on hard samples.
  - we introduced **FocalLoss** to improve robustness on imbalanced data.

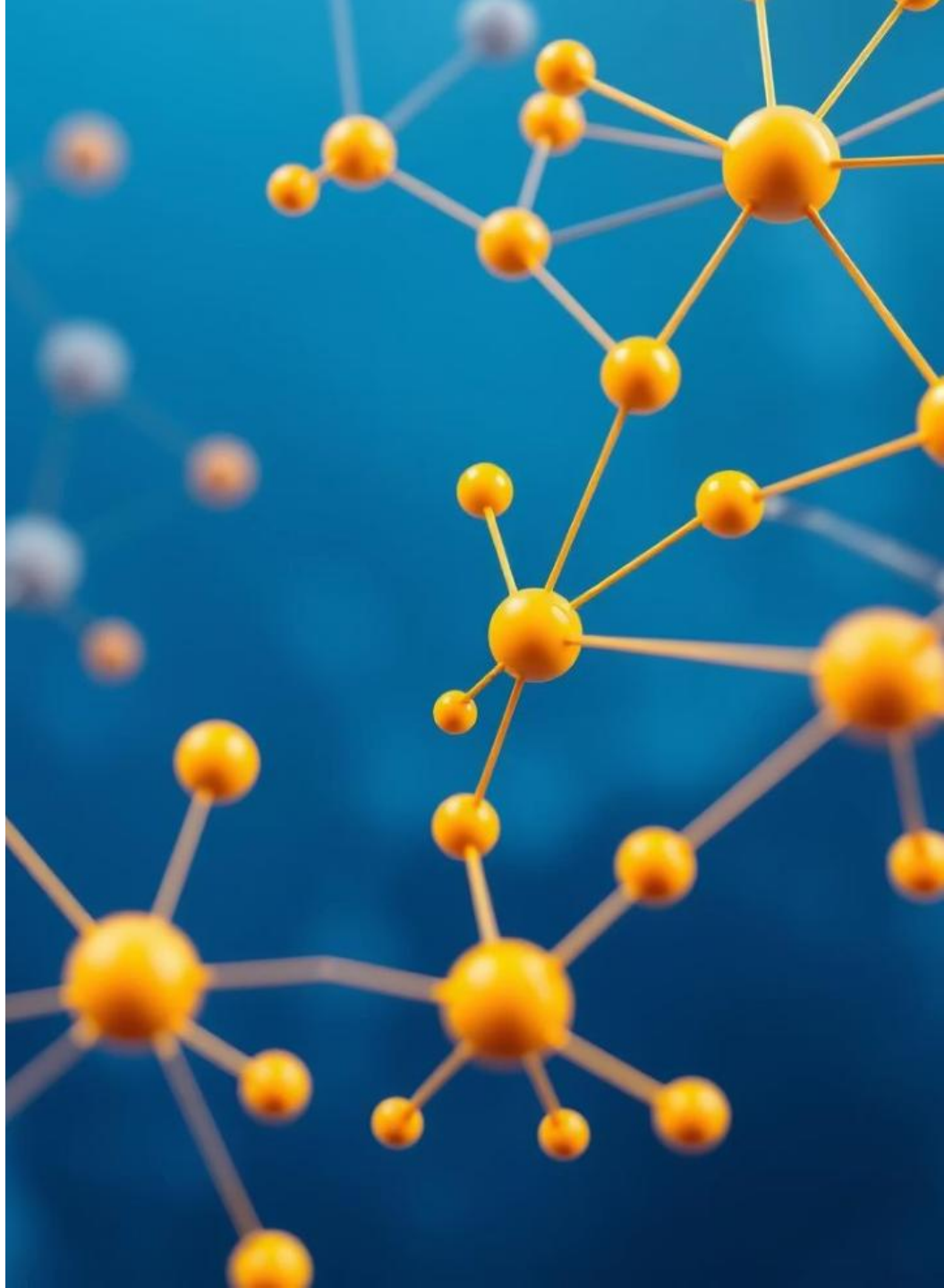
Breakdown of HIV Dataset

- HIV dataset is extremely imbalanced, with only ~3.5% positives.
- That makes the use of Focal Loss and ROC-AUC evaluation justified.

Class	Count	Proportion
Negative (0)	39,684	96.49%
Positive (1)	1,443	3.51%
Total	41,127	100%

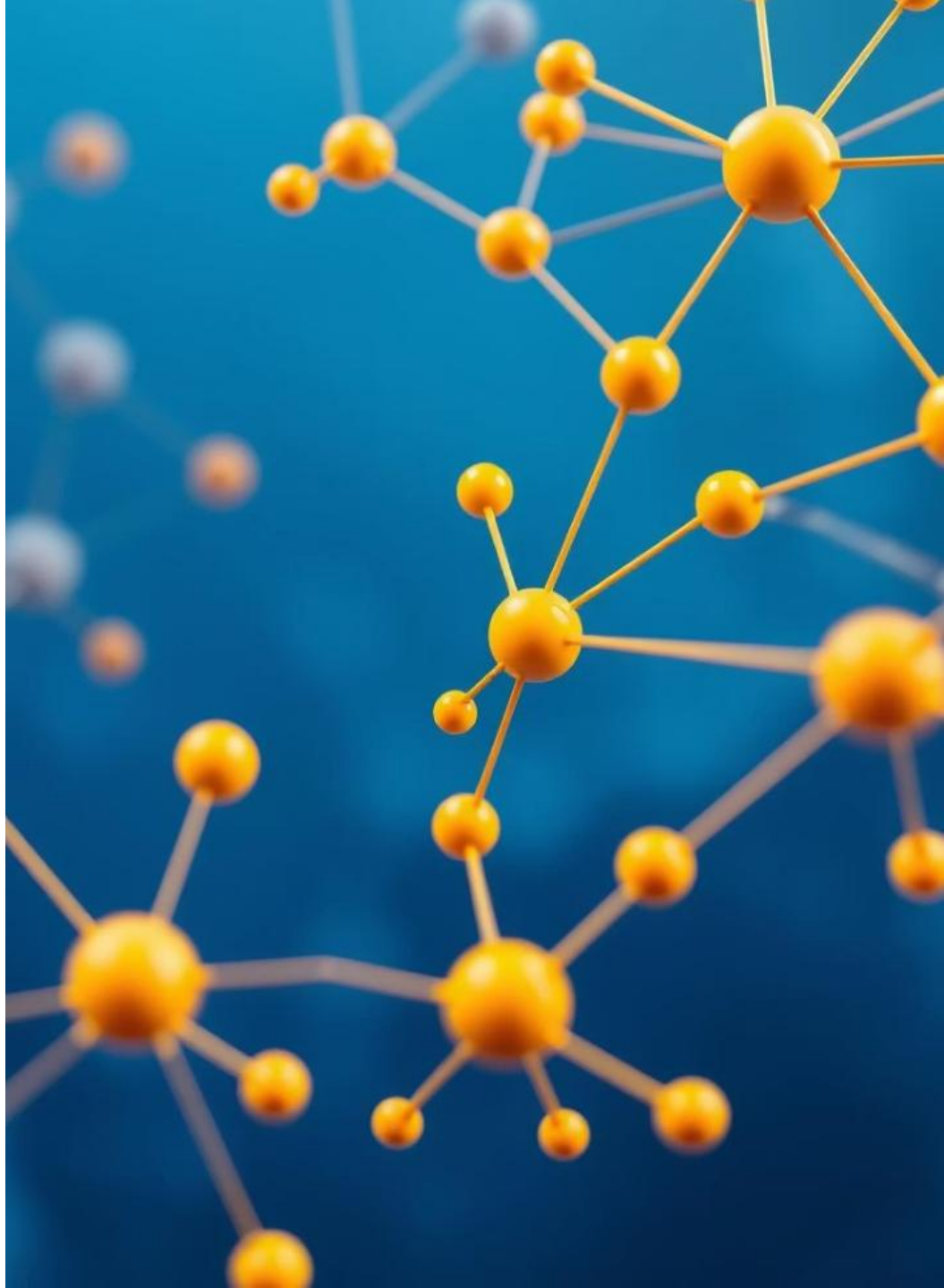
Graph Classification - HIV

- Metric: ROC-AUC
- State-of-the-Art model: CIN++ (0.8063)
- **KANG result:** 0.8104



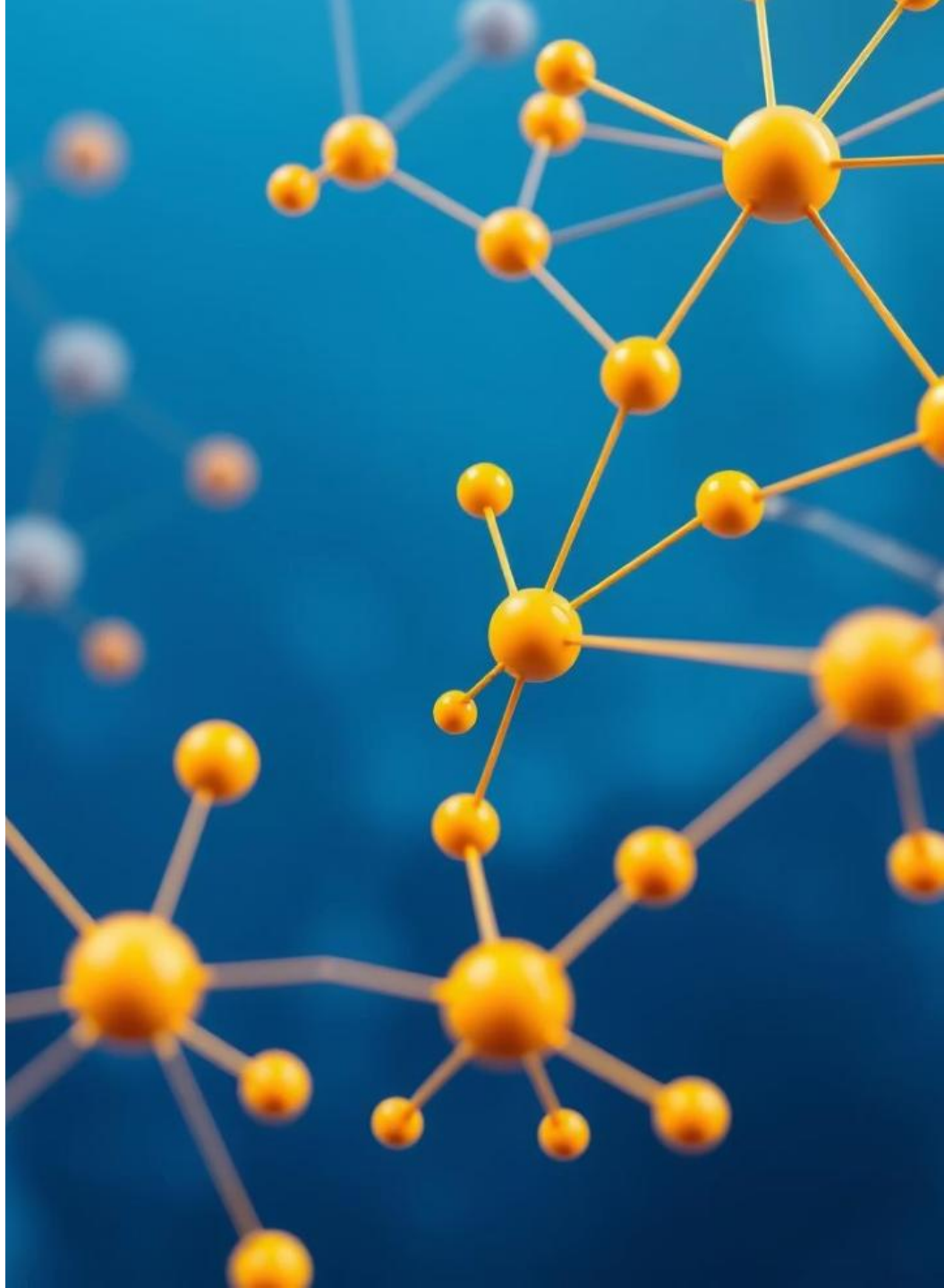
Graph Classification - ToxCast

- 5 selected binary tasks from 617
- Tasks include: AhR, Aromatase, AutoFluor, p53 (2 types)
- State-of-the-Art model: DumplingGNN (0.782)
- **KANG average result:** 0.7922



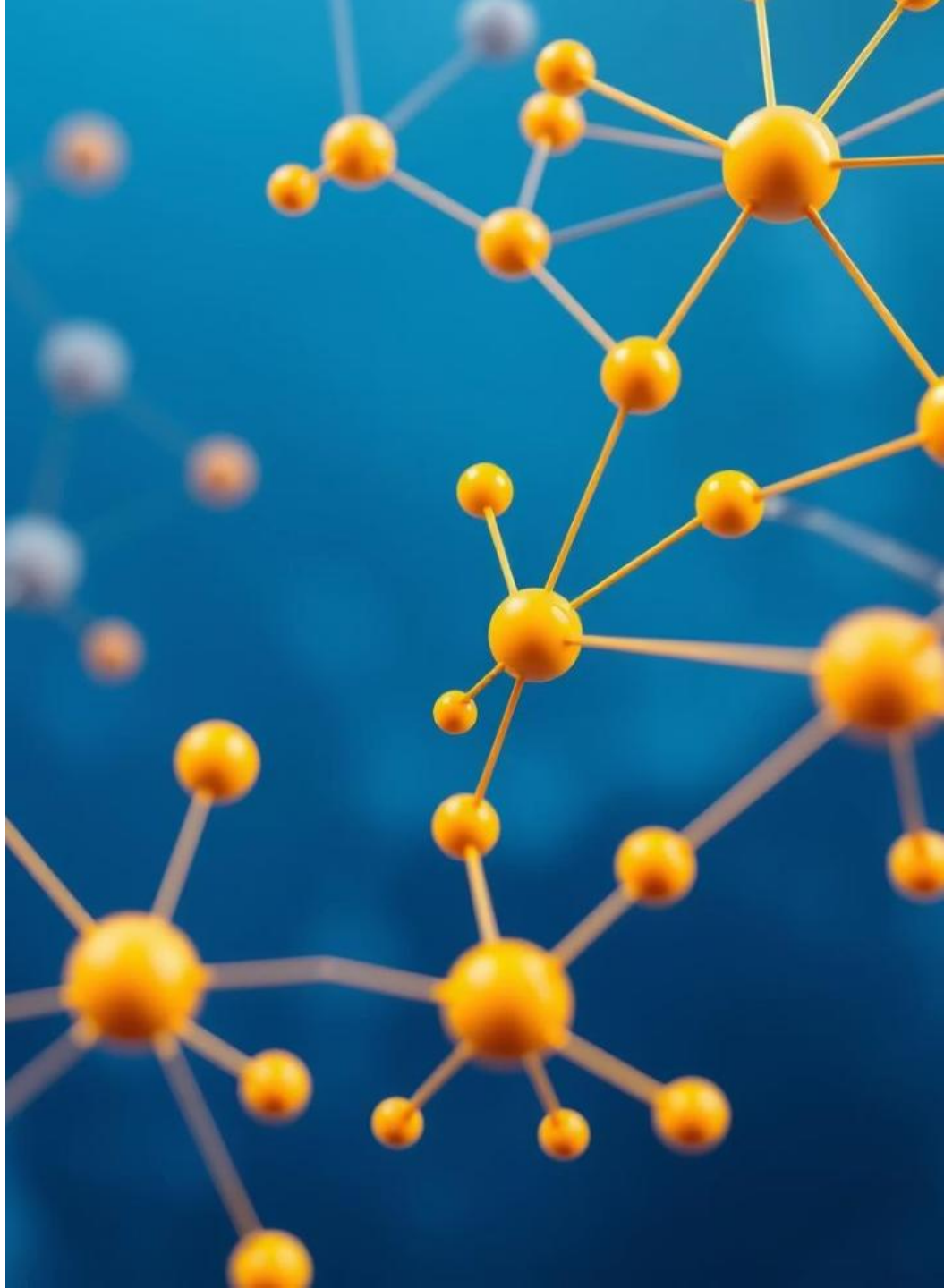
Graph Regression - QM8

- 12 quantum mechanical tasks
- Metric: MAE
- SoTA: D-MPNN (0.0190)
- **KANG average result:** 0.0164



Graph Regression - QM9

- 12 quantum chemical tasks
- Metric: MAE
- SoTA: D-MPNN (0.00814)
- **KANG average result:** 8.5740



Performance Summary Table

Dataset	Metric	SoTA Model	SoTA Value	Previous KANG Value	Updated KANG Value
QM8	MAE	D-MPNN	**0.0163**	0.0221	0.0164
QM9	MAE	D-MPNN	**2.694**	7.1407	8.5740
HIV	ROC-AUC	CIN++	0.8063	0.6997	**0.8104**
ToxCast	ROC-AUC	DumplingGNN	0.782	0.7896	**0.7922**

- **Improvements** on QM8, HIV, ToxCast
- **QM9** still **underperforms** despite retraining with best hyperparameters.

Conclusions

- Significant underperformance in QM9
- Comparable QM8 performance
- ToxCast better performance, but only 5 out of the total tasks were evaluated
- Better performance for HIV
- Further Improvements:
  - 3D Structure
  - Multi-Task Learning

Thank You

Sahar Khanlari - 2107563

Marco Natale - 1929854

A.A. 2024/2025