Comparing a Neural Network and a Graph Neural Network for Predicting PFAS Bioactivity

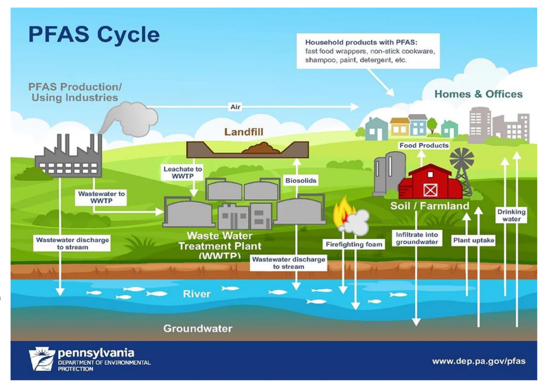
Wilson Schwegler

Abstract

Per- and polyfluoroalkyl substances (PFAS) are synthetic chemicals widely used in industrial and consumer products due to their unique properties, such as resistance to heat, water, and oil. However, these compounds do not break down easily in the environment and can accumulate in human body, posing significant health risks. Traditional methods for identifying bioactivity are often costly and time consuming for researchers and manufactures. Therefore, we present a cost-effective machine learning approach to predict the bioactivity of PFAS molecules, streamlining the identification process and enabling more efficient testing and safer product design.

Introduction

- PFAS are a class of synthetic chemicals known for their widespread usage in consumer products. Unfortunately, their persistence in the environment has led to long-term contamination of soil and drinking water. Additionally, PFAS can accumulate in the human body over time, posing health risks. Addressing these challenges requires innovative approaches to identify harmful bioactive PFAS molecules and replace them with less harmful alternatives.
- Bioactivity prediction is a critical step in mitigating PFAS-related risks. Traditional neural networks (NNs) utilize molecular fingerprints to represent the molecule's structure, while graph neural networks (GNNs) utilize a graph representation, incorporating the molecular structure with important attributes like spatial relationships, bond angle, and bond length. By comparing these two machine learning models, this study seeks to establish a baseline of knowledge to aid researchers in selecting the most effective approach for bioactivity prediction in their own research.

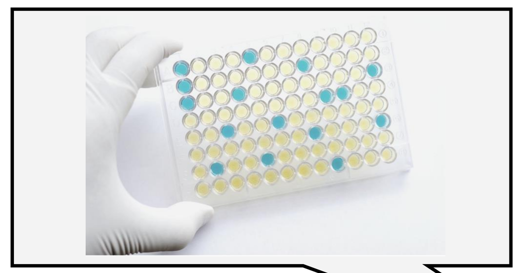


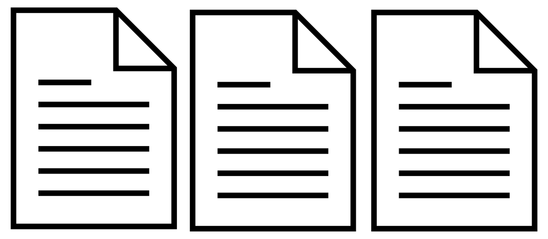
Conclusion

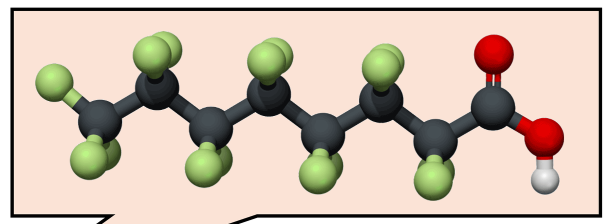
- A neural network based approach is better suited to reduce false positives, as it excels at recognizing the more subtle trends that contribute to bioactivity.
- A graph neural network approach is more effective at reducing false negatives, as it excels at identifying the structural trends that predict bioactivity.
- This research is important for materials development and product manufacturing as it can lead to the creation of safer products while reducing environmental and health impacts.

Methods

Open access data





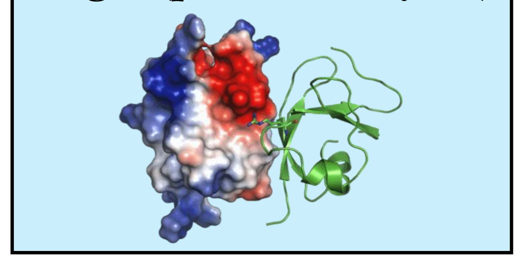


Bioassay results

Structure

PCBA-624288	PCBA-624296	PCBA-624417	PCBA-651635	...	PCBA-686970	smiles
0	0	1	0		0	CCC1NC(=O)c2cccnc2-n2c1nc1ccc(F)cc1c2=O
1	0	0	0		1	O=C(CSc1nc2ccccc2s1)N1c2ccccc2Sc2ccc(C(F)(F)F)cc21
0	0	1	1		1	CCOc1cc(/C=N/NC(=O)c2ccccn2)ccc1OC(=O)c1ccc(F)cc1
0	1	0	0		0	CCn1cc(CN2CCC(C(=O)Nc3cccc(-c4cccc(F)c4)c3)CC2)cn1

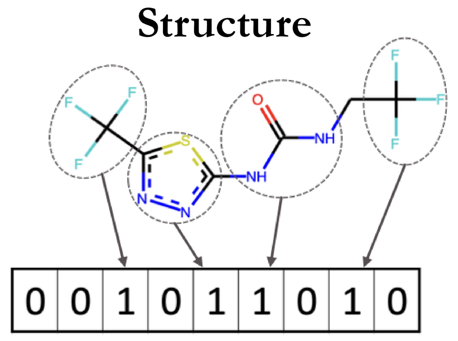
Target (protein/enzyme)



Neural Network

Training Data Set

Structure



Molecular Fingerprint

+

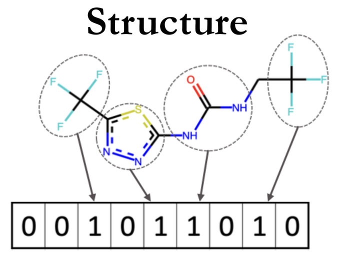
Bioassay result

Target 2
Active

Train the Model

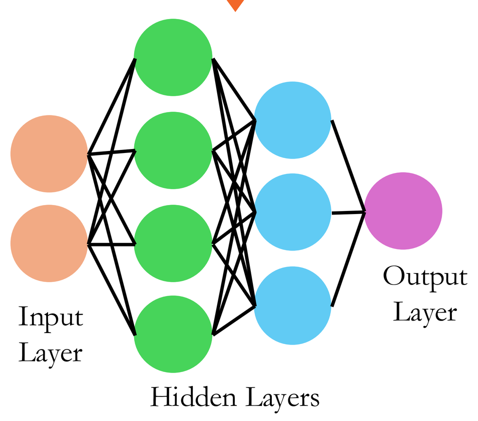
Testing Data Set

Structure



Molecular Fingerprint

Test the Model



Predicted result

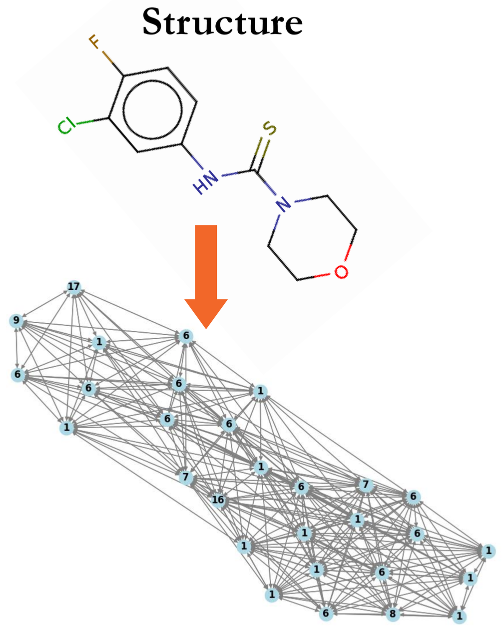
Target 2
Active

- Convert SMILES strings into molecular fingerprints.
- Molecular fingerprints are binary strings (0s and 1s) representing chemical structures.
- They encode key structural features for efficient representation.

Graph Neural Network

Training Data Set

Structure



Graph Representation

+

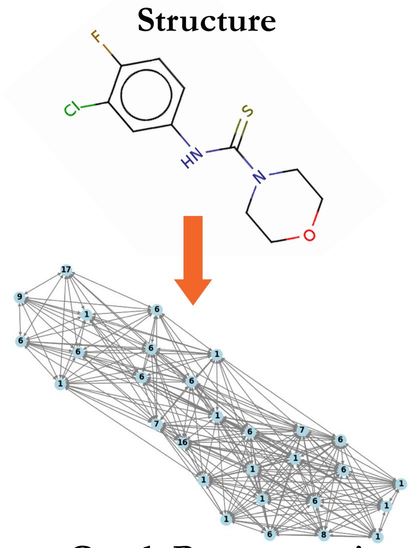
Bioassay result

Target 2
Active

Train the Model

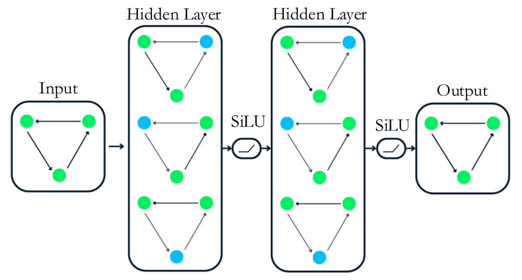
Testing Data Set

Structure



Graph Representation

Test the Model



Predicted result

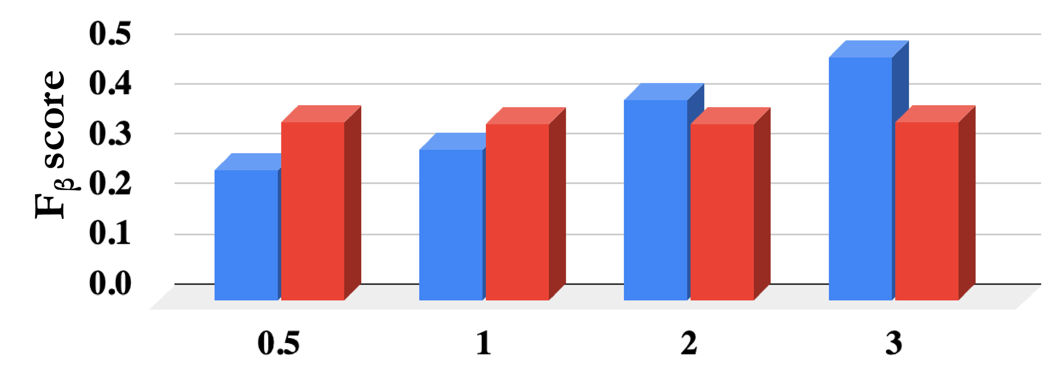
Target 2
Active

- Convert SMILES strings into 3D molecular structures using UFF (Universal Force Field).
- Represent molecules as graphs using atom types, bond lengths, and angles.
- Apply random rotations for equivariant data augmentation.

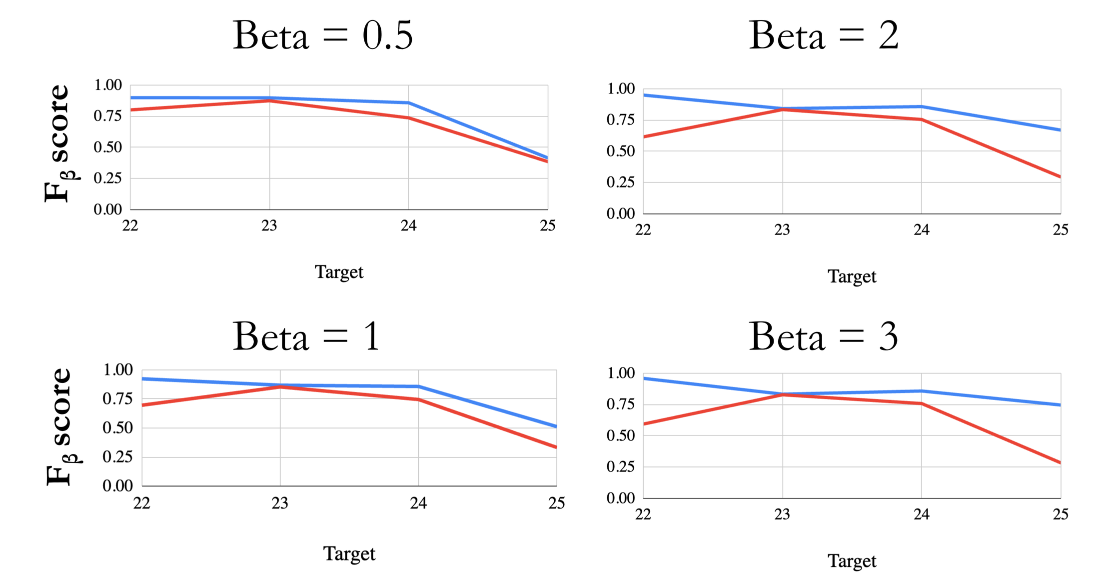
Results

F-Beta Score

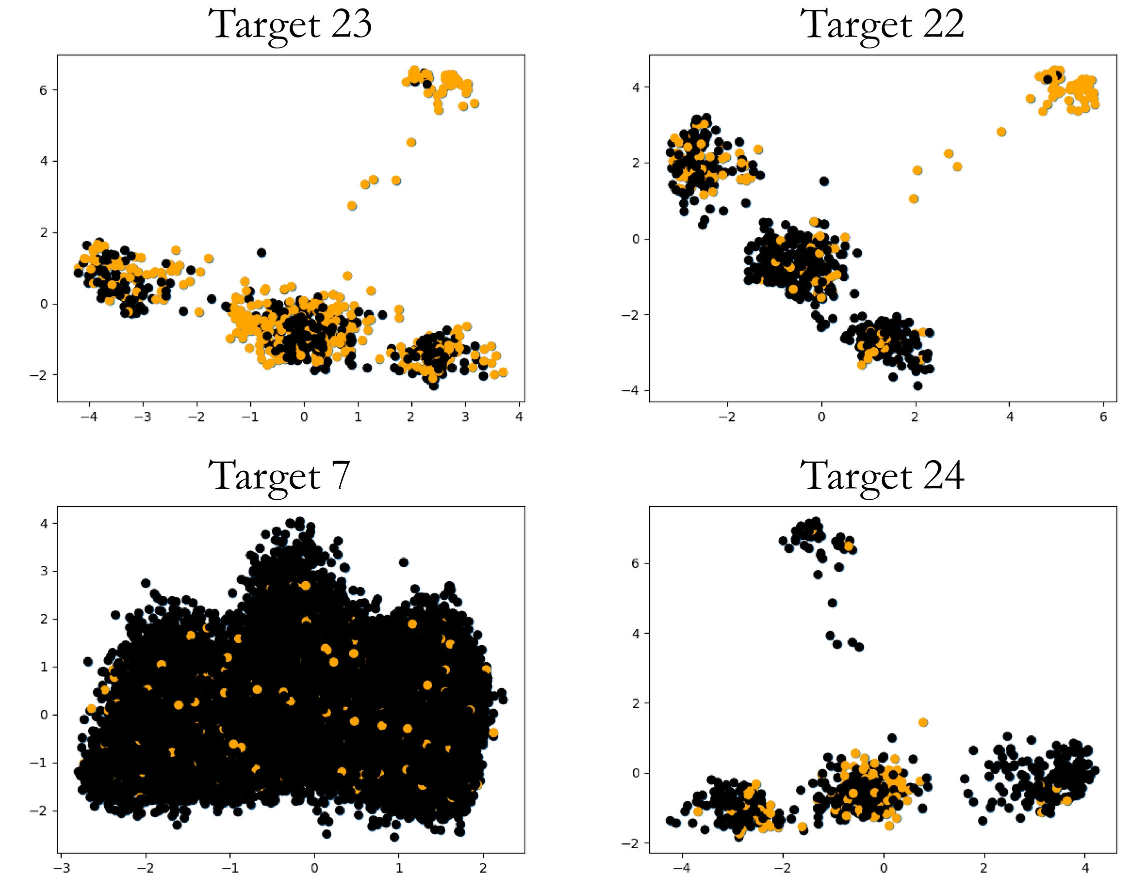
$$F_\beta = \frac{(1+\beta^2)\cdot \text{true positive}}{(1+\beta^2)\cdot \text{true positive} + \beta^2 \cdot \text{false negative} + \text{false positive}}$$



In Depth Analysis



PCA Dimension Reduction



Compared neural and graph neural networks for PFAS bioactivity prediction: Graph neural networks perform better in structure-dependent cases like cytochrome C enzymes.

Lawrence Livermore National Laboratory PyG scikit learn TensorFlow RDKit