# INTEGRATING GRAPH NEURAL NETWORKS AND SPECTROPHOTOMETRY FOR pIC50 PREDICTION OF PESTICIDES

**Member:**
Ang Yu Xi Sophie (Raffles Institution)

**Mentors:**
Shen Bingquan, Yap Xiu Huan
(DSO National Laboratories)

## Abstract

Accurate pIC50 prediction is vital for assessing the toxicity and potency of chemical compounds, including pesticides. This study leverages computational and experimental approaches to achieve this goal. A Graph Isomorphism Network (GIN) was pretrained on a large PubChem dataset (117,520 molecules) to learn node-level molecular features and fine-tuned on a smaller dataset (14,611 molecules) containing experimentally determined pIC50 values. The model predicts pIC50 values directly from molecular SMILES strings. Validation of the GIN's predictions was conducted using spectrophotometric assays for two pesticides, of differing potency Profenofos and Dichlorvos, to determine empirical pIC50 values.

## Introduction

Molecular property prediction is a key area of computational chemistry, aimed at developing models that map molecular structures to their properties. The accurate prediction of a compound's inhibitory concentration (pIC50) is a cornerstone of toxicological studies, allowing chemical structures, simulation, and physical data to be integrated when predicting health risks and other toxicological information. Graph Isomorphism Networks (GINs) rise to the challenge by representing molecules as graphs, capturing both local and global features to link chemical topology to biological activity. Furthermore, transfer learning has emerged as an essential technique to address the scarcity of labelled data and high dimensionality of feature spaces. In the USA, quantitative structure–activity relationships (QSARs) predictions are used to evaluate two to three thousand chemicals each year and to assess a significant portion of the toxicity information. However, traditional QSAR models often rely on predefined molecular descriptors (e.g., atom counts, bond types, molecular weight) and linear or nonlinear regression models to establish relationships between molecular features and biological activity. Unlike the static descriptors of QSAR, the GIN directly learns from molecular graph representations and captures topological and relational features more comprehensively. This project presents a modern, deep-learning-driven alternative to traditional QSAR approaches, offering greater flexibility, and adaptability in predicting pIC50 values.

## Methods

### Data Preprocessing Pipeline

Molecular structures in both datasets were represented using the SMILES notation. To prepare the data for the GIN model, each SMILES string was converted into a graph structure.

Feature extraction was conducted to provide meaningful inputs to the GIN model at both the node and graph levels. To extract node level features, each atom in the graph was assigned a feature vector encoding its chemical properties, such as atom type, hybridisation state and presence of formal charges. In addition, graph-level descriptors such as molecular weight, lipophilicity and topological polar surface area were computed, summarising global molecular properties particularly relevant for pIC50 prediction.

### Model Architecture and Pretraining

To build a robust feature extractor, the GIN was pretrained on a large toxicity dataset from PubChem via a self-supervised learning task, where the goal was to predict node-level features, such as atomic environments or chemical properties. Using the generative reconstruction technique, we masked the features of a random batch of nodes, forwarded the masked graph through the GIN encoder and reconstructed the masked node features. After pretraining, the GIN was fine-tuned on a smaller dataset with corresponding pIC50 values over several epochs.


*Fig. 1. Scatter plots of predicted vs actual pIC50 values using simpler models: (from left to right) simple linear regression, random forest regression and GIN*

| Model | Mean squared error |
|---|---|
| Simple linear regression | 1.4881 |
| Random forest regression | 0.5789 |
| Graph isomorphism network | 0.5698 |

*Fig. 2. Table of values of mean squared error for each model employed*

## Experimental Setup

To validate the predictions of the GIN, the toxicities of two pesticides, Profenofos and Dichlorvos, were experimentally determined using an AChE inhibition assay. Stock solutions were serially diluted to create eight concentrations for testing. AChE enzyme solution was added to each well containing one of the pesticide concentrations, the mixtures were incubated at 25°C for 15 minutes. Ellman's reagent (5,5′-dithiobis-(2-nitrobenzoic acid)) and the substrate acetylcholine were added to the wells. Ellman's assay produces a yellow colour as AChE catalyses the breakdown of acetylcholine, with the intensity proportional to the enzyme's activity. A spectrophotometer was used to measure the colour intensity of the solutions and thence the reaction rates. After the experiment, the percentage inhibition of AChE was calculated and pIC50 values of the pesticides were determined using the PRISM GRAPHPAD software, which plotted a dose-response curve of percentage inhibition against log-transformed pesticide concentration.

## Results

**GIN Prediction Results:** The GIN model was saved at the best validation loss of **0.568** after going through 10 training epochs. GIN's predicted pIC50 values for **Profenofos** and **Dichlorvos** were, to 3 significant figures, **4.22** and **3.00** respectively.

**Experimental Results**: The experimentally-obtained pIC50 values for Profenofos and Dichlorvos were **3.12** and **3.01** respectively. The comparison between both sets of results is demonstrated in Fig. 3.

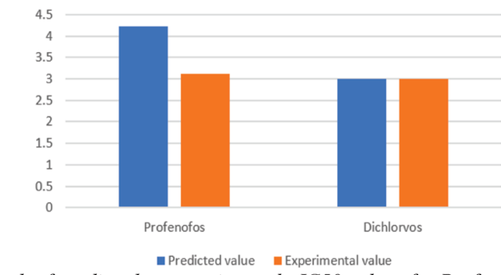


*Fig. 3. Bar graph of predicted vs experimental pIC50 values for Profenofos and Dichlorvos*

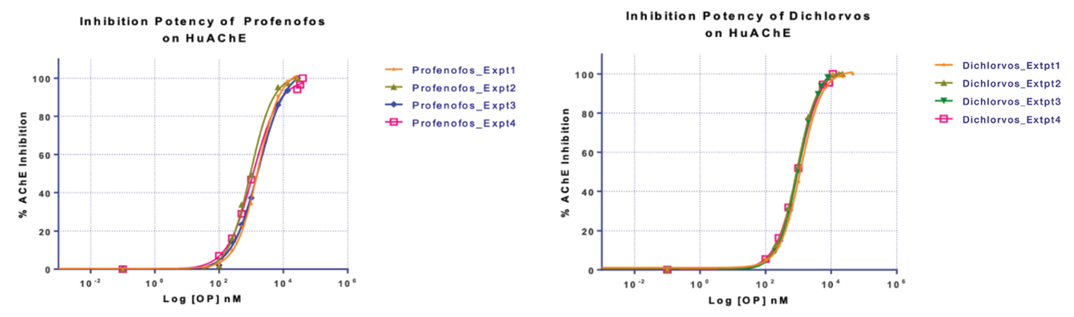


*Fig. 4. Sigmoidal curves of percentage AChE inhibition against log-transformed concentration of Profenofos (left) and Dichlorvos (right) across four experiments, as plotted on PRISM*

| Best-fit values | Expt 1 | Expt 2 | Expt 3 | Expt 4 | Mean |
|---|---|---|---|---|---|
| Top | 104.5 | 102.1 | 105.4 | 99.58 | 102.895 |
| Bottom | 0.3186 | -1.396 | -0.7566 | -0.05916 | -0.47329 |
| LogIC50 | 3.215 | 2.978 | 3.227 | 3.055 | 3.11875 |
| HillSlope | 1.254 | 1.224 | 1.036 | 1.072 | 1.1465 |
| IC50 | 1642 | 951.2 | 1686 | 1134 | 1353.3 |
| Span | 104.2 | 103.5 | 106.2 | 99.64 | 103.385 |
| **Goodness of Fit** | | | | | |
| Degrees of Freedom | 4 | 4 | 4 | 4 | |
| R square | 0.9984 | 0.9977 | 0.9988 | 0.999 | 0.998475 |
| Absolute Sum of Squares | 22.05 | 29.52 | 14.98 | 12.34 | |
| Sy.x | 2.348 | 2.716 | 1.935 | 1.757 | |

| Best-fit values | Expt 1 | Expt 2 | Expt 3 | Expt 4 | Mean |
|---|---|---|---|---|---|
| Top | 101.7 | 101.9 | 103.9 | 104.2 | 102.925 |
| Bottom | 1.092 | 0.686 | -0.1098 | -0.06993 | 0.3995675 |
| LogIC50 | 3.07 | 2.96 | 3 | 2.996 | 3.0065 |
| HillSlope | 1.356 | 1.397 | 1.37 | 1.233 | 1.339 |
| IC50 | 1176 | 911.6 | 1001 | 989.7 | 1019.575 |
| Span | 100.6 | 101.2 | 104 | 104.3 | 102.525 |
| **Goodness of Fit** | | | | | |
| Degrees of Freedom | 4 | 4 | 4 | 4 | |
| R square | 0.999 | 0.9993 | 0.9997 | 0.9995 | 0.9994 |
| Absolute Sum of Squares | 13.28 | 7.926 | 3.662 | 6.321 | |
| Sy.x | 1.822 | 1.408 | 0.9568 | 1.257 | |

*Fig. 5. Statistical analysis of critical values, particularly pIC50 (in red) and $R^2$ (in blue), for Profenofos (left) and Dichlorvos (right)*

## Discussion

The GIN demonstrated relatively strong performance in predicting pIC50 values from molecular SMILES strings, highlighting the model's ability to capture the relationship between molecular structure and inhibitory potency. Comparing the predicted and experimental values, the GIN achieved a better correlation for Dichlorvos. In contrast, for Profenofos, the discrepancy between the predicted and experimental values for Profenofos was more pronounced, suggesting that the model struggled more with capturing the complex molecular features of this compound. Dichlorvos achieved an $R^2$ value of 0.9994, indicating an almost perfect fit between the experimental data and the sigmoidal curve. This high $R^2$ suggests that the enzyme inhibition data followed a consistent pattern with minimal variability, making it easier for the GIN model to predict the pIC50 value accurately. Profenofos, in contrast, achieved an $R^2$ of 0.9985, still high but slightly lower than Dichlorvos. This minor drop in $R^2$ reflects a slight increase in variability in the experimental data.

The observed discrepancy may stem from (i) chemical complexity: Dichlorvos is a simpler molecule with a more straightforward structure; Profenofos' bulkier, more intricate structure, with a large aromatic ring system adds more variability to the molecular structure; (ii) training data bias: the pretraining dataset may have overrepresented simpler molecules and underrepresented complex aromatic systems; (iii) interaction dynamics: while Dichlorvos might have more predictable interactions, Profenofos' steric hindrance and complex binding dynamics may pose challenges for the model.

## Conclusion

This study highlights the GIN's ability to predict pIC50 values from SMILES strings, bridging computational modelling and experimental validation in toxicity assessment. AChE inhibition assays confirmed strong prediction accuracy for Dichlorvos, with more variability for Profenofos. While reliable for simpler molecules, performance declined with complex structures, emphasising the need for diverse data and better feature representation. Future improvements could include expanding the dataset with more complex molecules, enhancing feature extraction for subtle structural details, refining the model for nuanced interactions and optimising training methods to improve accuracy and generalisability.

The GIN model has broad real-world potential in drug discovery, toxicology, and environmental science. It can predict the efficacy and toxicity of drug candidates, assess pesticide toxicity to protect non-target organisms and aid in hazard classification and risk assessment for industrial chemicals, pharmaceuticals, and pollutants.