INTEGRATING GRAPH NEURAL NETWORKS AND SPECTROPHOTOMETRY FOR pIC50 PREDICTION OF PESTICIDES

Ang Yu Xi Sophie[1], Yap Xiu Huan[2], Shen Bingquan[2]
[1]Raffles Institution, 1 Raffles Institution Lane, Singapore 575954
[2]DSO National Laboratories, 12 Science Park Drive, Singapore 118225

1. 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 to determine empirical pIC50 values for two pesticides of differing potency. A strong correlation between predicted and experimental values highlights the potential of combining graph-based deep learning for toxicity prediction.

2. Introduction

Molecular property prediction is a key area of computational chemistry, aimed at developing models that map molecular structures to their properties.[1] 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.[2] The pIC50 value, which quantifies the half maximal inhibitory concentration, is used to determine the potency of a drug against a variety of enzymes or biological targets associated with the pathogenesis of multiple disorders.[3] Experimental identification of the bioactivity of each potential peptide is an expensive and time-consuming task.[4] Deep representation learning has been reported as a promising approach for molecular property prediction, outperforming fixed molecular representations.[1]

However, accurately predicting molecular properties poses a significant challenge due to the complex relationships between molecular structures and their properties.[5] Molecules are typically represented in three ways: fixed representations, including fingerprints and structural keys, that signify the presence of specific structural patterns; linear notations, such as Simplified Molecular Input Line Entry System (SMILES) strings; and molecular graphs.[6] With the advent of deep learning, various neural networks have been proposed for molecular representation learning, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) and graph neural networks (GNNs).[1]

One major task for AI in drug discovery is molecular property prediction.[1] Graph Isomorphism Networks (GINs) address this 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.[5] By leveraging knowledge gained from related tasks, transfer learning enables models to generalise better to target tasks with limited training data, enhancing prediction accuracy. Thus, to enhance model accuracy, pretraining on large datasets followed by task-specific fine-tuning is crucial.

The EU has established a new chemical assessment paradigm (REACH, enacted in 2007) to evaluate chemicals globally.[5] 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.[7] 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.[8] The coverage and comprehensiveness of metabolic data presents a critical challenge with respect to QSAR modelling[8], requiring manual recalibration for new datasets or molecular classes.

In this study, we developed a GIN model pretrained on a PubChem toxicity dataset of over 100,000 molecules and fine-tuned it on over 14,000 molecules with known pIC50 values to predict pesticide inhibition. Unlike the static descriptors of QSAR, the GIN directly learns from molecular graph representations and captures topological and relational features more comprehensively. Additionally, the GIN can incorporate new data seamlessly through retraining. Validation through acetylcholinesterase (AChE) spectrophotometric assays revealed strong agreement between predicted and experimental pIC50 values. B

3. Materials and Methods

3.1 Dataset and Data Preprocessing

SMILES Representation: 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. In this representation, nodes represent the atoms in the molecule and edges represent the bonds between atoms. Bond types (single, double, etc.) were encoded as edge attributes. This transformation enabled the model to process molecular structures in a graph format, capturing both the local connectivity of atoms and the overall topology of the molecule.

Feature extraction: 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, hybridization state and presence of formal charges, to allow the GIN to learn how individual atoms contribute to the molecule's overall activity. In addition, graph-level descriptors such as molecular weight, lipophilicity and topological polar surface area (TPSA) were computed, summarising global molecular properties particularly relevant for pIC50 prediction. Such preprocessing pipeline ensured that the GIN could learn from both the detailed local information at the atom level and the broader structural and physicochemical properties of each molecule.

3.2 Model Architecture and Pretraining

GIN Components: The GIN was chosen for this study due to its ability to distinguish non-isomorphic graphs effectively, making it ideal for molecular graph representation. The key components of the GIN architecture are as follows: (i) the input layer consisting of molecular graphs converted from SMILES strings; (ii) graph convolution layers aggregating information

from neighbouring nodes, iteratively updating each node's representation to encode local structural and chemical information; (iii) hidden layers triggering nonlinear transformations using activation functions (e.g. ReLU) and batch normalisation to stabilise training; (iv) the readout layer, a global pooling layer combining node features into a fixed-size graph-level representation; and (v) the task-specific output layer – during pretraining, the output is node-level predictions to learn molecular substructures, while during fine-tuning, the output is a single graph-level target (i.e. pIC50).

Pretraining on the PubChem Dataset: 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 given the node representations of their local sub-graphs[9]. The model parameters were then optimised using gradient descent and a loss function that measured the accuracy of node-level predictions.

Fine-tuning for pIC50 Prediction: After pretraining, the GIN was fine-tuned on a smaller dataset with corresponding pIC50 values. This phase involved task-specific adjustments, where (i) the output layer was modified to predict graph-level instead of node-level properties; (ii) both node-level features (e.g., atomic types) and graph-level descriptors (e.g., molecular weight, lipophilicity, TPSA) were utilised to enhance prediction accuracy; and (iii) the pre-trained weights were used to initialise the GIN, allowing the model to retain its learned molecular features. A regression loss function (e.g., Mean Squared Error) was employed to measure the difference between predicted and actual pIC50 values. The model was fine-tuned for several epochs to reduce training and validation loss. This two-step approach of pretraining and fine-tuning allowed the GIN to combine general molecular feature extraction with task-specific optimisation, significantly improving its accuracy and robustness in predicting pIC50 values.

3.3 Experimental Setup

Colorimetric Determination of Acetylcholinesterase (AChE) Activity: To validate the predictions of the GIN, the toxicities of two pesticides, Profenofos and Dichlorvos, were experimentally determined using an AChE inhibition assay. A preliminary experiment to determine AChE activity was conducted before testing pesticide inhibition to optimise the dilution factor of the AChE enzyme – the goal was to achieve a baseline V_{max} (i.e., maximum reaction rate) of approximately 100 without the addition of pesticides. After attaining the optimal dilution factor of enzyme, the chemical procedure was as follows: (i) preparation: stock solutions of the two pesticides were prepared, serially diluted to create eight concentrations for testing covering a wide range to accurately capture dose-response behaviour; (ii) enzyme incubation: in 96-well plates, triplicates were prepared for each pesticide concentration to ensure reproducibility. Control wells (without pesticide) were included to establish baseline AChE activity. AChE enzyme solution was added to each well containing one of the pesticide concentrations, the mixtures were incubated at 25°C for 15 minutes to allow the pesticides to interact with and inhibit the enzyme; (iii) addition of Ellman's Assay: 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[10]; (iv) spectrophotometry: a spectrophotometer was used to measure the colour intensity of the solutions and the reaction rates (i.e. V_{max}) at a wavelength of 412 nm, reflecting the enzyme's activity at

each pesticide concentration; and (v) repetition for consistency: to enhance the reliability of results, the entire experiment was repeated four times for each pesticide, and the mean values for each measurement were calculated. This comprehensive setup provided a robust dataset for statistical analysis and reduced the influence of experimental variability.

Data Collection: To quantify enzyme inhibition, the following steps were performed: (i) V_{max} Measurement: the spectrophotometer recorded the V_{max} values for each pesticide concentration in triplicate; (ii) averaging V_{max}: the V_{max} values for each concentration were averaged across the triplicates; (iii) enzyme activity calculation: the enzyme activity at each concentration was expressed as a percentage relative to a control (enzyme activity without pesticide) using the formula: Enzyme activity (%) = V_{max} (sample) / V_{max} (control) x 100;
and (iv) percentage inhibition: The percentage inhibition of AChE was calculated as: inhibition (%) = 100 - Enzyme Activity (%). These calculations were repeated for all eight pesticide concentrations for both pesticides.

pIC50 Determination: The pIC50 values of the pesticides were determined using the following approach: (i) data plotting: the percentage inhibition values at each pesticide concentration were entered into PRISM GRAPHPAD software; (ii) sigmoidal curve fitting: a dose-response curve of percentage inhibition against log-transformed pesticide concentration was plotted *(Fig. 1)*; (iii) pIC50 determination: the Sigmoidal, 4PL function was used to determine the IC50 value, which was converted to pIC50 using the formula: $pIC50 = -\log_{10}(IC50)$. The resulting pIC50 values provided a quantitative measure of each pesticide's inhibitory potency, which was then compared to the GIN's predicted values for model validation.

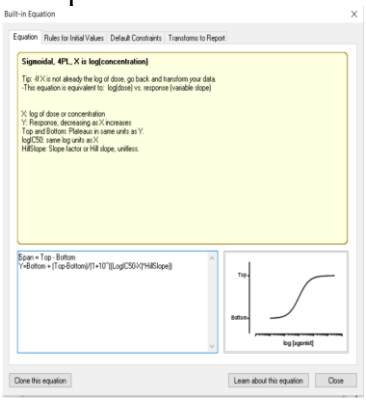

Fig. 1. Built-in equation for Sigmoidal, 4Pl function of the PRISM GRAPHPAD software

4. Results

4.1 GIN Prediction Results

Prediction accuracy:

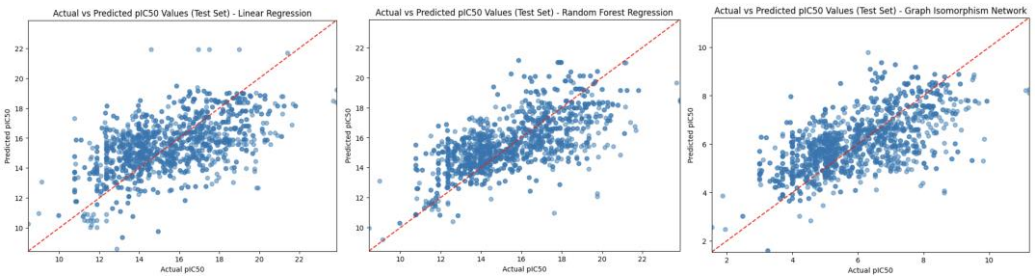

Fig. 2. Scatter plots of predicted vs actual pIC50 values using simpler models: (from left to right) simple linear regression, random forest regression and GIN

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

Fig. 3. Table of values of mean squared error for each model employed

```
Epoch 1, Training Loss: 0.7303854137496218
Epoch 1, Validation Loss: 0.6841405176598093
Model saved at epoch 1 with validation loss: 0.6841405176598093
Epoch 2, Training Loss: 0.580572730042244
Epoch 2, Validation Loss: 0.7463098336821017
Epoch 3, Training Loss: 0.523014206052478
Epoch 3, Validation Loss: 0.6647760103578153
Model saved at epoch 3 with validation loss: 0.6647760103578153
Epoch 4, Training Loss: 0.4844845746710001
Epoch 4, Validation Loss: 0.6399535050858622
Model saved at epoch 4 with validation loss: 0.6399535050858622
Epoch 5, Training Loss: 0.4589613482274644
Epoch 5, Validation Loss: 0.7705744919569596
Epoch 6, Training Loss: 0.4385773071337267
Epoch 6, Validation Loss: 0.7156230910964634
Epoch 7, Training Loss: 0.42326210919625123
Epoch 7, Validation Loss: 0.6127975868142169
Model saved at epoch 7 with validation loss: 0.6127975868142169
Epoch 8, Training Loss: 0.4083780018521137
Epoch 8, Validation Loss: 0.6275735769582831
Epoch 9, Training Loss: 0.3965139038901511
Epoch 9, Validation Loss: 0.6954140481741532
Epoch 10, Training Loss: 0.38277937610292695
Epoch 10, Validation Loss: 0.5683993489845939
Model saved at epoch 10 with validation loss: 0.5683993489845939
```

Fig. 4. Training and validation losses over 10 training epochs, with the GIN model saved at the best validation loss

```
new_molecule1_smiles = "COP(=O)(OC)OC=C(Cl)Cl"
predicted_pIC50 = predict_pIC50(new_molecule1_smiles, model, data_mean, data_std)
print(f"Predicted pIC50 for {new_molecule1_smiles}: {predicted_pIC50}")

new_molecule2_smiles = "CCCSP(=O)(OCC)OC1=C(C=C(C=C1)Br)Cl"
predicted_pIC50 = predict_pIC50(new_molecule2_smiles, model, data_mean, data_std)
print(f"Predicted pIC50 for {new_molecule2_smiles}: {predicted_pIC50}")
```

```
Predicted pIC50 for COP(=O)(OC)OC=C(Cl)Cl: 3.00315197293448
Predicted pIC50 for CCCSP(=O)(OCC)OC1=C(C=C(C=C1)Br)Cl: 4.224558182399754
```

Fig. 5. GIN's predicted pIC50 values, where molecule 1 represents Dichlorvos and molecule 2 represents Profenofos

4.2 Experimental Results

Inhibition data:

Sample	Conc (nM)	Log conc	Well	Vmax	Mean Vmax	Mean activity (a.u.)	Inhibition (%)
1	0	-3	A1	71.82	73.853	100	0
			A2	75.24			
			A3	74.5			
2	100	2	B1	70.45	72.113	97.64396842	2.356031576
			B2	72.08			
			B3	73.81			
3	250	2.4	C1	65.11	66.607	90.18861793	9.81138207
			C2	68.12			
			C3	66.59			
4	500	2.7	D1	55.51	57.213	77.4687555	22.5312445
			D2	58.32			
			D3	57.81			
5	1000	3	E1	47.42	48.767	66.03252407	33.96747593
			E2	49.55			
			E3	49.33			
6	6691	3.83	F1	9.21	6.843	9.265703492	90.73429651
			F2	6.7			
			F3	4.62			
7	13382	4.13	G1	1.097	1.196	1.619433198	98.3805668
			G2	1.09			
			G3	1.4			
8	26764	4.43	H1	-0.09	0.573	0.775865571	99.22413443
			H2	0.22			
			H3	1.59			

Sample	Conc (nM)	Log conc	Well	Vmax	Mean Vmax	Mean activity (a.u.)	Inhibition (%)
1	0	-3	A1	73.99	70.67	100	0
			A2	68.16			
			A3	69.86			
2	100	2	B1	72.77	69.38	98.1746144	1.825385595
			B2	67.37			
			B3	68			
3	250	2.4	C1	61.86	59.173	83.73142776	16.26857224
			C2	58.57			
			C3	57.09			
4	500	2.7	D1	48.45	46.81	66.23744163	33.76255837
			D2	45.95			
			D3	46.03			
5	1000	3	E1	36.67	35.777	50.6254422	49.3745578
			E2	34.38			
			E3	36.28			
6	6691	3.83	F1	1.64	3.293	4.659685864	95.34031414
			F2	4.45			
			F3	3.79			
7	13382	4.13	G1	3.38	1.86	2.631951323	97.36804868
			G2	1.04			
			G3	1.16			
8	26764	4.43	H1	3.03	0.51	0.721664072	99.27833593
			H2	-0.6			
			H3	-0.9			

Sample	Conc (nM)	Log conc	Well	Vmax	Mean Vmax	Mean activity (a.u.)	Inhibition (%)
1	0	-3	A1	108.21	102.15	100	0
			A2	102.1			
			A3	96.14			
2	100	2	B1	99.05	100.023	97.91776799	2.082232012
			B2	101.73			
			B3	99.29			
3	250	2.4	C1	86.63	87.463	85.62212433	14.37787567
			C2	88.67			
			C3	87.09			
4	500	2.7	D1	76.64	78.45	76.79882526	23.20117474
			D2	80.88			
			D3	77.83			
5	1000	3	E1	68.85	64.137	62.78707783	37.21292217
			E2	63.52			
			E3	60.04			
6	6691	3.83	F1	17.71	14.347	14.04503182	85.95496818
			F2	15.14			
			F3	10.19			
7	13382	4.13	G1	8.214	6.666	6.525697504	93.4743025
			G2	7.07			
			G3	4.714			
8	26764	4.43	H1	3.614	4.502	4.407244249	95.59275575
			H2	4.491			
			H3	5.4			

Sample	Conc (nM)	Log conc	Well	Vmax	Mean Vmax	Mean activity (a.u.)	Inhibition (%)
1	0	-3	A1	78.53	76.283	100	0
			A2	76.01			
			A3	74.31			
2	100	2	B1	69.68	70.907	92.95255824	7.047441763
			B2	72.58			
			B3	70.46			
3	250	2.4	C1	60.41	63.92	83.79324358	16.20675642
			C2	68.74			
			C3	62.61			
4	500	2.7	D1	49.88	54.26	71.12987166	28.87012834
			D2	58.97			
			D3	53.93			
5	1000	3	E1	34.18	40.537	53.14028027	46.85971973
			E2	42.96			
			E3	44.47			
6	26764	4.43	F1	3.42	4.514	5.91743901	94.08256099
			F2	5.843			
			F3	4.279			
7	33456	4.52	G1	-0.7	2.647	3.469973651	96.53002635
			G2	5.66			
			G3	2.98			
8	40147	4.6	H1	-0.96	2.987	3.915682393	96.08431761
			H2	5.79			
			H3	4.13			

*Fig. 6. Experimental data for **Profenofos** across four experiments, read from left to right, top to bottom, with log concentration (x-axis) and percentage inhibition (y-axis) boxed*

Sample	Conc (nM)	Log conc	Well	Vmax	Mean Vmax	Mean activity (a.u.)	Inhibition (%)
1	0	-3	A1	72.07	68.817	100	0
			A2	68.46			
			A3	65.92			
2	199	2	B1	68.06	65.003	94.4577648	5.542235204
			B2	64.15			
			B3	62.8			
3	497.5	2.4	C1	62.22	60.343	87.68618219	12.31381781
			C2	61.19			
			C3	57.62			
4	995	2.7	D1	50.95	51.05	74.18225148	25.81774852
			D2	53.59			
			D3	48.61			
5	1990	3	E1	39.38	37.927	55.11283549	44.88716451
			E2	37.69			
			E3	36.71			
6	112256.68	3.75	F1	-0.6	0.18	0.261563277	99.73843672
			F2	NoFit			
			F3	0.96			
7	16885.03	3.93	G1	1.131	1.02	1.4821919	98.5178081
			G2	1.02			
			G3	0.91			
8	22513.37	4.05	H1	NoFit	NoFit		
			H2	NoFit			
			H3	NoFit			

Sample	Conc (nM)	Log conc	Well	Vmax	Mean Vmax	Mean activity (a.u.)	Inhibition (%)
1	0	-3	A1	66.41	65.153	100	0
			A2	66.27			
			A3	62.78			
2	100	2	B1	59.33	56.957	87.42037972	12.57962028
			B2	55.24			
			B3	56.3			
3	250	2.4	C1	46.22	44.647	68.52639172	31.47360828
			C2	43.96			
			C3	43.76			
4	500	2.7	D1	32.23	30.757	47.20734272	52.79265728
			D2	30.81			
			D3	29.23			
5	1000	3	E1	16.7	14.32	21.97903397	78.02096603
			E2	13.19			
			E3	13.07			
6	2000	3.3	F1	1.2	0.79	1.212530505	98.78746949
			F2	1.471			
			F3	-0.3			
7	2500	3.4	G1	0	0	0	100
			G2	0			
			G3	0			
8	4000	3.6	H1	0	1.8	2.762727733	97.23727227
			H2	1.8			
			H3	0			

Sample	Conc (nM)	Log conc	Well	Vmax	Mean Vmax	Mean activity (a.u.)	Inhibition (%)
1	0	-3	A1	55.59	54.77	100	0
			A2	54.81			
			A3	53.91			
2	100	2	B1	50.67	49.437	90.26291766	9.737082344
			B2	49.24			
			B3	48.4			
3	250	2.4	C1	38.77	38.397	70.10589739	29.89410261
			C2	38.44			
			C3	37.98			
4	500	2.7	D1	28.64	27	49.29706043	50.70293957
			D2	27.33			
			D3	25.03			
5	1000	3	E1	13.15	13.29	24.26510864	75.73489136
			E2	14.54			
			E3	12.18			
6	2000	3.3	F1	4.92	5.62	10.26109184	89.73890816
			F2	6.21			
			F3	5.73			
7	2500	3.4	G1	-0.03	3.438	6.277159029	93.72284097
			G2	6.08			
			G3	4.264			
8	4000	3.6	H1	NoFit	1.48	2.702209239	97.29779076
			H2	2.89			
			H3	0.07			

Sample	Conc (nM)	Log conc	Well	Vmax	Mean Vmax	Mean activity (a.u.)	Inhibition (%)
1	0	-3	A1	102.564	103.862	100	0
			A2	103.264			
			A3	105.757			
2	100	2	B1	97.764	98.119	94.47054746	5.529452543
			B2	95.914			
			B3	100.679			
3	250	2.4	C1	83.086	86.974	83.73996264	16.26003736
			C2	79.85			
			C3	97.986			
4	500	2.7	D1	71.771	70.838	68.20396295	31.79603705
			D2	68.764			
			D3	71.979			
5	1000	3	E1	51.307	49.979	48.12058308	51.87941692
			E2	47.729			
			E3	50.9			
6	5657	3.75	F1	8.871	8.119	7.817103464	92.18289654
			F2	7.586			
			F3	7.9			
7	8485	3.93	G1	8.464	4.698	4.523309776	95.47669022
			G2	-0.786			
			G3	6.414			
8	11313	4.05	H1	3.021	3.052	2.938514567	97.06148543
			H2	0.457			
			H3	5.679			

*Fig. 7. Experimental data for **Dichlorvos** across four experiments, read from left to right, top to bottom, with log concentration (x-axis) and percentage inhibition (y-axis) boxed*

Sigmoidal fit:

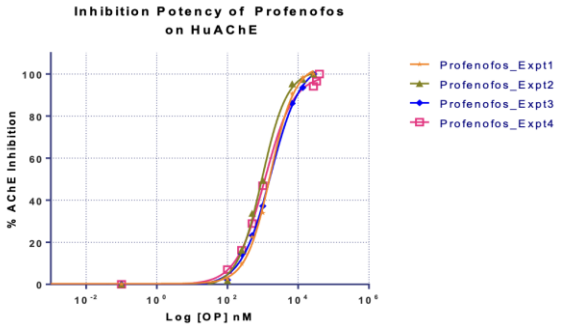

*Fig. 8. Sigmoidal curves of percentage AChE inhibition against log-transformed concentration of **Profenofos** across four experiments, as plotted on PRISM*

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	

*Fig. 9. Statistical analysis of critical values, particularly pIC50 (in red) and R^2 (in blue), for **Profenofos***

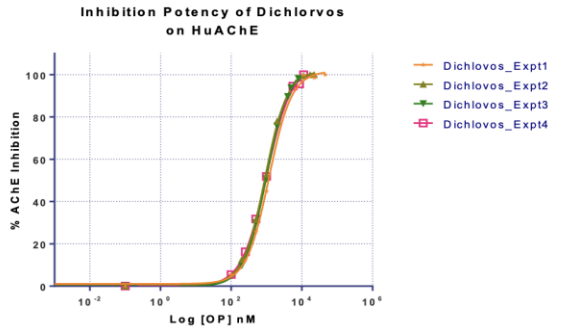

*Fig. 10. Sigmoidal curves of percentage AChE inhibition against log-transformed concentration of **Dichlorvos** across four experiments, as plotted on PRISM*

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. 11. Statistical analysis of critical values, particularly pIC50 (in red) and R^2 (in blue), for **Dichlorvos***

4.3 Comparison of GIN Predictions and Experimental pIC50 Values

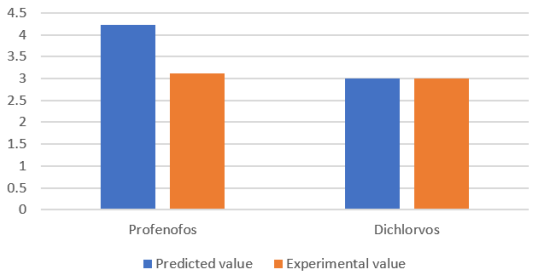

Fig. 12. Bar graph of predicted vs experimental pIC50 values for Profenofos and Dichlorvos

5. Discussion

5.1 Interpretation of Model Performance

Strengths: 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. The pretraining on a large toxicity dataset ensured the model learned generalisable molecular features, while fine-tuning on the smaller dataset of labelled pIC50 values enabled it to specialise in this specific task.

Challenges: Despite the model's success, certain challenges were identified: (i) potential overfitting: despite cross validation, the smaller fine-tuning dataset may limit generalisation and increase the risk of overfitting, especially if the model becomes too specialised to the training data; (ii) prediction discrepancies: variability in enzyme activity or spectrophotometric readings, compounded by the training dataset's limited chemical diversity, may contribute to differences between predicted and experimental pIC50 values; (iii) rare or unique molecular structures: underrepresented or absent structural motifs in the training data may reduce prediction accuracy for complex molecules; (iv) simplified assumptions in computational modelling: while GINs excel at capturing molecular topology and physicochemical properties, certain biochemical factors influencing pIC50 values, such as solvent effects, metabolic pathways or protein-ligand dynamics, are not explicitly modelled, contributing to prediction errors; and (v) substructure representation: GINs may struggle with complex substructures, such as large aromatic systems with multiple halogen and functional group substitutions, which present intricate patterns of interactions that can be difficult to fully capture.

5.2 Experimental verification

Agreement between predicted and experimental values: The comparison between the GIN model's predicted pIC50 values and those experimentally-obtained through enzyme inhibition assays provides insights into the model's performance and its ability to generalize from the training data to unseen test compounds. In this section, we analyse the correlation between the predicted and experimental pIC50 values for Dichlorvos and Profenofos — which were used to validate the model. For Dichlorvos, the GIN model demonstrated a high degree of accuracy in predicting the pIC50 value, with a difference of 0.00330 *(Fig. 12)*. The empirically determined pIC50 was found to be closely aligned with the model's predicted value, indicating that the model was able to effectively capture the relationship between the SMILES representation of Dichlorvos and its toxicity. In contrast, for Profenofos, the discrepancy between the predicted and experimental

values for Profenofos was more pronounced, with a larger difference of 1.10581 *(Fig. 12)*, which suggests that the model struggled with the complex molecular features of this compound.

Evaluation: 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.

A comparison between the predicted and experimental pIC50 values revealed a stronger correlation for Dichlorvos (COP(=O)(OC)OC=C(Cl)Cl) than Profenofos (CCCSP(=O)(OCC)OC1=C(C=C(C=C1)Br)Cl). The observed discrepancy in model accuracy may stem from several factors: (i) chemical complexity: Dichlorvos is a simpler molecule with a more straightforward structure, consisting of a phosphate group, two ester linkages and a conjugated alkene. Meanwhile, Profenofos' bulkier, more intricate structure, with a large aromatic ring system (i.e. benzene with bromine and chlorine substituents) attached to a phosphorothioate group adds more variability to the molecular structure, including additional aromaticity and halogenation, which may challenge the model's ability to capture binding dynamics; (ii) training data bias: the pretraining dataset may have overrepresented simpler molecules like Dichlorvos and underrepresented complex aromatic systems, reducing accuracy for molecules like Profenofos; (iii) interaction dynamics: while Dichlorvos, being a relatively smaller and more linear molecule, might have more predictable interactions, Profenofos' steric hindrance and complex binding dynamics may exceed the model's current capacity to account for factors like solubility and metabolic transformations that exert a greater influence on the biological activity of Profenofos.

5.3 Implications and Future Work

Room for improvement of GIN model: The lower accuracy for Profenofos is likely due to its complex structure, underrepresented in the training data and the GIN's difficulty handling aromatic and halogenated features. 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.

Potential real-world applications: 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.

6. Conclusion

Main findings: This study highlights the GIN's ability to predict pIC50 values from SMILES, bridging computational modeling 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 increasing complexity of chemical structures, emphasising the need for diverse data and better feature representation. The research demonstrates GINs' potential in drug discovery, toxicology and environmental science.

Acknowledgements

I would like to express my deepest gratitude to Bingquan Shen from the Defence Science Organisation (DSO) for his invaluable guidance, insightful feedback and for making this project possible. I would also like to thank Yap Xiu Huan from DSO for her kind supervision and support through this research. Special thanks to Shen Xiaoting at the Chemical, Toxins and Radionuclear (CTRN) programme for her assistance with the spectrophotometric assays, experimentation and data collection. Lastly, I am grateful to my family and friends for their unwavering support throughout this journey. This research was supported by DSO.

References

1. Deng, J., Yang, Z., Wang, H. et al. A systematic study of key elements underlying molecular property prediction. Nat Commun 14, 6395 (2023). https://doi.org/10.1038/s41467-023-41948-6
2. Luechtefeld T, Hartung T. Computational approaches to chemical hazard assessment. Altex. 2017;34:459–478. doi: 10.14573/altex.1710141
3. Thakur, A. (2022). PIC50: an open source tool for interconversion of PIC50 values and IC50 for efficient data representation and analysis. ResearchGate. https://doi.org/10.1101/2022.10.15.512366
4. Singh, D., Mahadik, A., Surana, S., & Arora, P. (2022). Proteochemometric method for PIC50 prediction of flaviviridae. BioMed Research International, 2022, 1–7. https://doi.org/10.1155/2022/7901791
5. Jie Shen and Christos A. Nicolaou. Molecular property prediction: recent trends in the era of artificial intelligence. Drug Discovery Today: Technologies, 32-33:29–36, 2019. ArtificialIntelligence
6. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in ai-driven drug discovery: a review and practical guide. J. Cheminformatics 12, 1–22 (2020)
7. Ministry of Environment, Republic of Korea (2001) Study on improvement of new chemical substance hazard assessment system
8. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014 Jun 26;57(12):4977-5010. doi: 10.1021/jm4004285. Epub 2014 Jan 6. PMID: 24351051; PMCID: PMC4074254
9. Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In Procs. of SIGKDD, 2022
10. Aynacı Koyuncu, Elif & Yaşar, Ahmet & Arslan, Fatma & Sari, Nursen. (2019). Synthesis of novel Schiff base derivatives of tacrine and investigation of their acetylcholinesterase inhibition potency. Macedonian Journal of Chemistry and Chemical Engineering. 38. 75. 10.20450/mjcce.2019.1561