Supporting Information: Graph Neural Networks for Identifying Protein-Reactive Compounds

Victor Hugo Cano Gil and Christopher N. Rowley*

Department of Chemistry, Carleton University, Ottawa, ON

E-mail: christopherrowley@cunet.carleton.ca

Phone: +(613) 520-2600 x 1647

Hyperparameter Optimization of Morgan Fingerprint Models

The hyperparameters of all Morgan-fingerprint-based classifiers were optimized through an exhaustive search. The search covered the radii (r=2,3,4,5) and the bitsize (b=512, 1024, 2048, 4098) used to generate the Morgan fingerprint features as well as the parameters of the classifier. The model with the highest AUC on the internal test set is the model presented in the main text of the paper. The classifiers are the unmodified implementations in scikit-learn library version 1.2.2.

- *Logistic Regression (LR).* The optimal logistic regression classifier used a regularization constant of 0.1 and used a fingerprint with a radius of 3 and a size of 4098 bits. This optimal bitsize was anomalously large, consistent with the LR model having limited flexibility.

- *Support Vector Classifier (SVC).* Classifier hyperparameters included in the exhaustive search included the using radial basis function, linear, 3-order polynomial, and sigmoid kernels and regularization parameters (C) of 0.1,1, 10, 100. A 'scaled' kernel coefficient was used for the radial basis functions, polynomial, and sigmoid kernels. The SVC model with the highest AUC used a Morgan fingerprint size of radii of 3 and a bitsize of 2048 as features and used a radial basis function kernel and C of 1. This model used Morgan fingerprint features with a radii of 3 and 2048 bits.

- *Random Forest (RF).* By default, the Sci-Kit learn will increase the tree depth until all the instances in the training set are predicted to be part of the same same class; however, we found that a maximum tree depth of 21 was close to the optimal. This model used Morgan fingerprint features with a radii of 3 and 2048 bits.

- *Histogram Gradient Boost (HGB).* The optimal model had a maximum number of leaf nodes of 21 with no maximum depth. This model used Morgan fingerprint features with a radii of 3 and 2048 bits.

- *Multi-layer Perceptron (MLP).* A exhaustive hyperparameter search for the MLP classifier model was performed with 2,3,4,5 hidden layers with 100, 200, 300, 500 nodes per layer. *tanh* and *ReLU*[S1] activation functions were both tested. Strength of the $L2$ regularization term (a) covered the values 0.0001, 0.001, 0.005. The Adam[S2] algorithm was used to optimize the neural network weights. The optimal model had an a term of 0.0001 with ReLU activation functions. The network had three hidden layers with 100 nodes per layer. This model used Morgan fingerprint features with a radii of 3 and 2048 bits.

Hyperparameter Optimization of Graph Neural Networks

The hyperparameter search was performed using randomized search with 10-fold cross validation. Table S1 summarizes the best hyperparameters for each graph architecture.

Table S1: Optimal hyperparameters for GNNs

	GCNII	GT	GraphSage	GAT	GatedGCN	GATv2	GCN	GIN	GMM
Upsample	True	True	True	True	True	True	True	True	True
Balance Class Weights	True	True	False	False	True	True	True	True	False
Batch Size	64	32	32	16	16	16	16	128	16
Number of Layers	6	6	6	6	4	4	2	2	6
Readout Layer	Mean	Mean	Mean	Max	Mean	Mean	Mean	Mean	Mean
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
Learning Rate	5E-05	5E-05	5E-05	5E-05	5E-05	5E-05	5E-05	5E-05	5E-05
Units	128	32	64	32	32	32	64	64	64
Use Edge Features	True	True	False	False	False	False	False	True	False
Activation Function	SeLU [S3]	SeLU	ReLU	ReLU	SeLU	SeLU	ReLU	ReLU	SeLU
Dropout Rate	0.15	0.1	0.25	0.25	0.1	0.15	0.25	0.15	0.1
Number of Parameters	63391	177759	24511	14431	22335	14303	22815	31585	374211

Graph Neural Networks Features

Table S2 summarizes the features used in graph models and the way each feature was encoded.

Table S2: Encodings of the GNN node and edge features

Node Features		Edge Features	
Description	Encoding	Encoding	Description
Atomic Symbol	One-hot	Bond Type	One-hot
Total Number of Hydrogens	One-hot	Conjugated	Binary
Chiral Centre	Binary	Rotatable	Binary
Aromatic	Binary	Part of Ring	Binary
Part of Ring	Binary		
Hetero	Binary		
Hydrogen Donor	Binary		
Hydrogen Acceptor	Binary		
Part of Ring of Size N	One-hot		
Gasteiger Charge	Float		

We have also examined at how removing each feature from the GCNII model affected external test set metrics. The GCNII model was retrained with each edge and node feature removed. The change in performance of the new models are summarized in Table S3. Recall was the metric most affected by feature imputation. In particular, the removal of the Atomic Symbol feature had a significant detrimental effect. Precision and AUROC were much less affected by feature imputation.

Table S3: Feature Impact Analysis, showing how metrics of the GCNII model change when a particular feature is removed.

	External	External	External
	Test	Test	Test
Feature	AUROC	Precision	Recall
Removed	Impact	Impact	Impact
Atomic Symbol	-0.02	0.00	-0.12
Total Number Hydrogens	-0.03	-0.01	-0.06
Chiral Centre	-0.01	+0.01	-0.02
Aromatic	0.00	-0.01	-0.08
Part of Ring	0.00	+0.01	-0.08
Hetero	-0.01	-0.04	0.00
Hydrogen Donor	0.00	-0.02	-0.04
Hydrogen Acceptor	0.00	0.00	+0.01
Part of Ring of Size N	+0.01	0.00	-0.09
Gasteiger Charge	0.00	0.00	-0.05
Bond Type	0.00	0.00	-0.03
Conjugated	-0.01	-0.02	0.00
Rotatable	0.00	-0.03	+0.01
Part of Ring	0.00	0.00	-0.06

Filters Removed from Other Screening Tools

PAINS

```
[#6]=,:[#6]:[#7]([#6])~[#6]:[#6]=,:[#6][#6]~[#6]:[#7] dyes5A(27)

[#6]-,:[#6]:[#7+]=,:[#6][#6]=[#6][#7][#6;X4] dyes3A(19)

c:1:c(:c:c:c:c:1)-[#7](-[#6](-[#1])-[#1])-[#6](-[#1])=[#6](-[#1])-[#6]=!@[#6](-[#1])-[#6
```

Eli Lilly

no_interesting_atoms

too_few_rings

too_many_rings

biotin

too_many_aromatic_rings_in_ring_system

ring_system_too_large

ring_system_too_large_with_aromatic

fmoc

positive

negative

too_long_carbon_chain

diphosphate

quaternary_amine

sulfonic_acid

quaternary_amine

crown_2_2_cyclic

crown_3_3_cyclic

crown_2_3_cyclic

crown_2_2

crown_3_3

crown_2_3

phenylenediamine

LongCChain

Comparison of External Test to the Training Set

To determine how similar training data was to external test data, we performed analysis to search through the external test set to find the compound with the largest Tanimoto similarity coefficient and smallest pairwise distance of each compound in the training set. A histogram of the frequency of each coefficient is presented in Figure S1. The most frequent Tanimoto coefficients were in the interval between 0.2 and 0.4, while the most frequent pairwise distance were in the interval between 0.1 and 0.35. This indicates that while there are some structures that are similar (at worst differing by a functional group or two) they are few in number, and the two datasets are largely dissimilar.

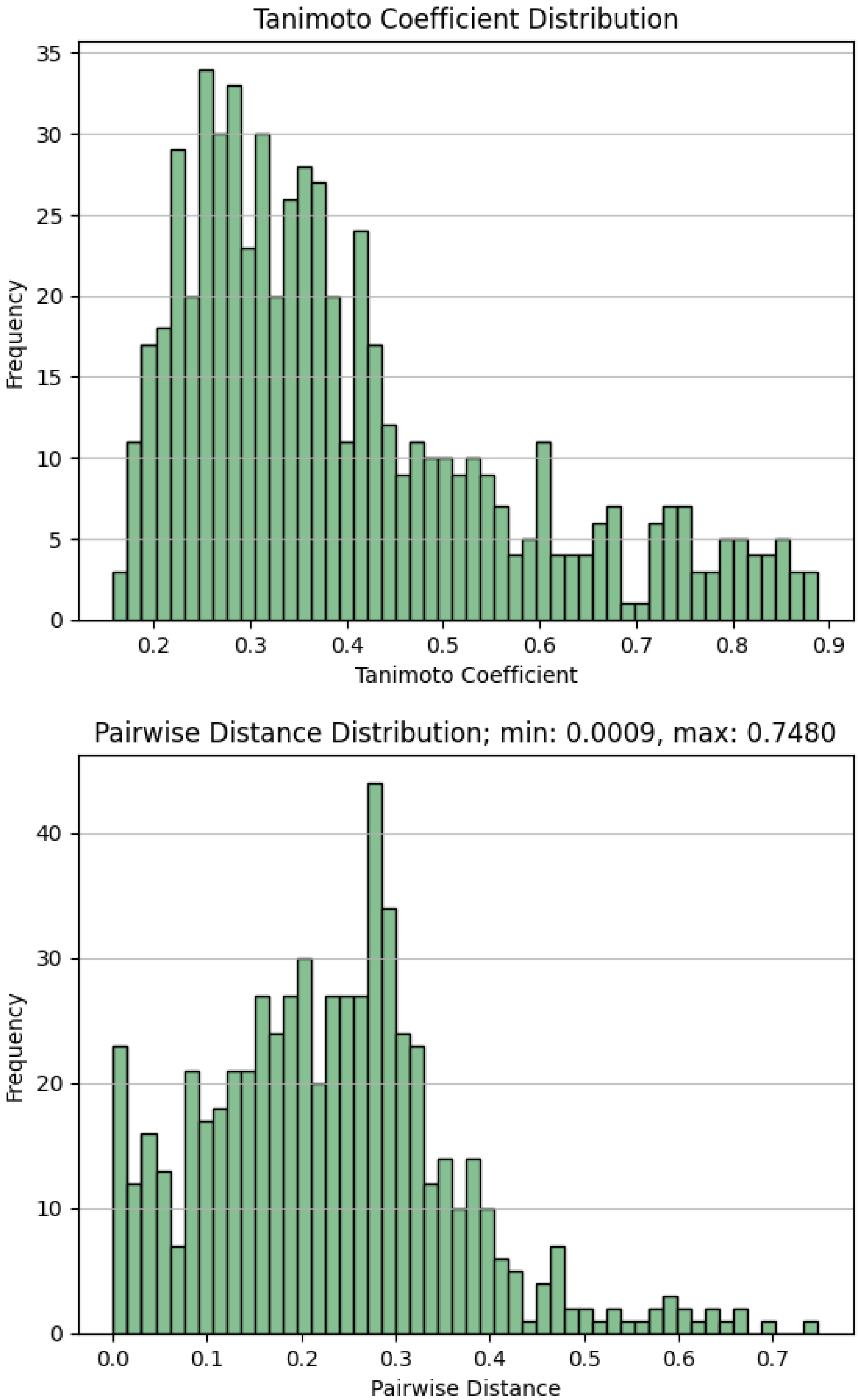


Figure S1: Distributions of Tanimoto coefficients (top figure) and pairwise distances (bottom figure) constructed by finding for each structure in the external test set most similar (or dissimilar) structure from the training set. Both figures were calculated independently from each other. The distributions show that the compounds in the two sets are largely dissimilar.

ChEMBL Predictions

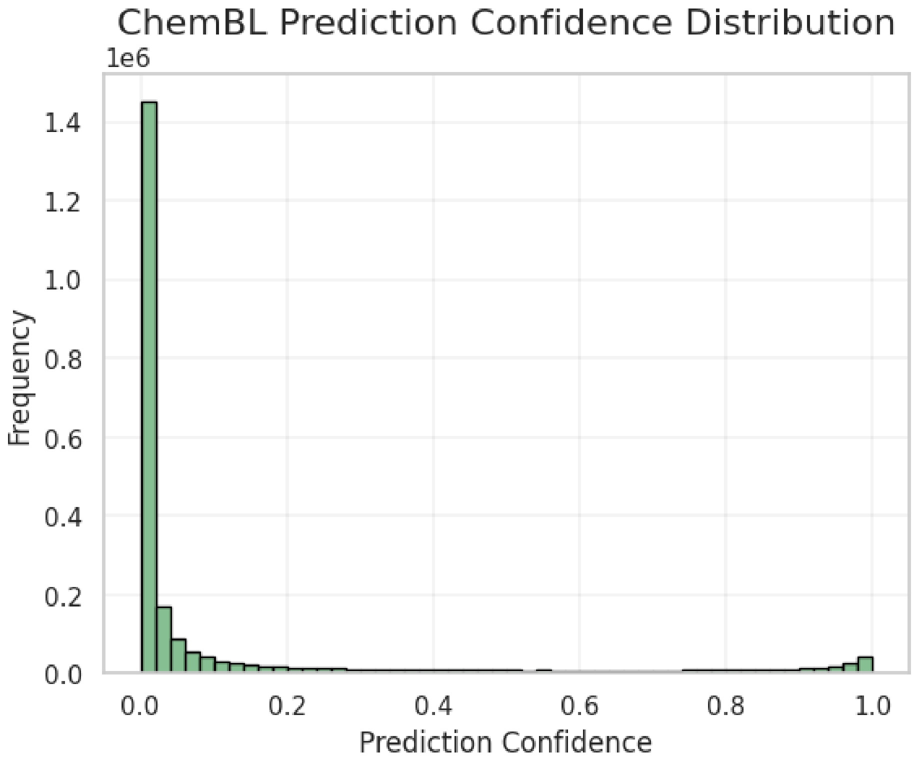


Figure S2: Prediction confidence distribution of the GCNII model when applied to ChemBL[S4] dataset.

Comparison of Eli Lilly and GNN predictions

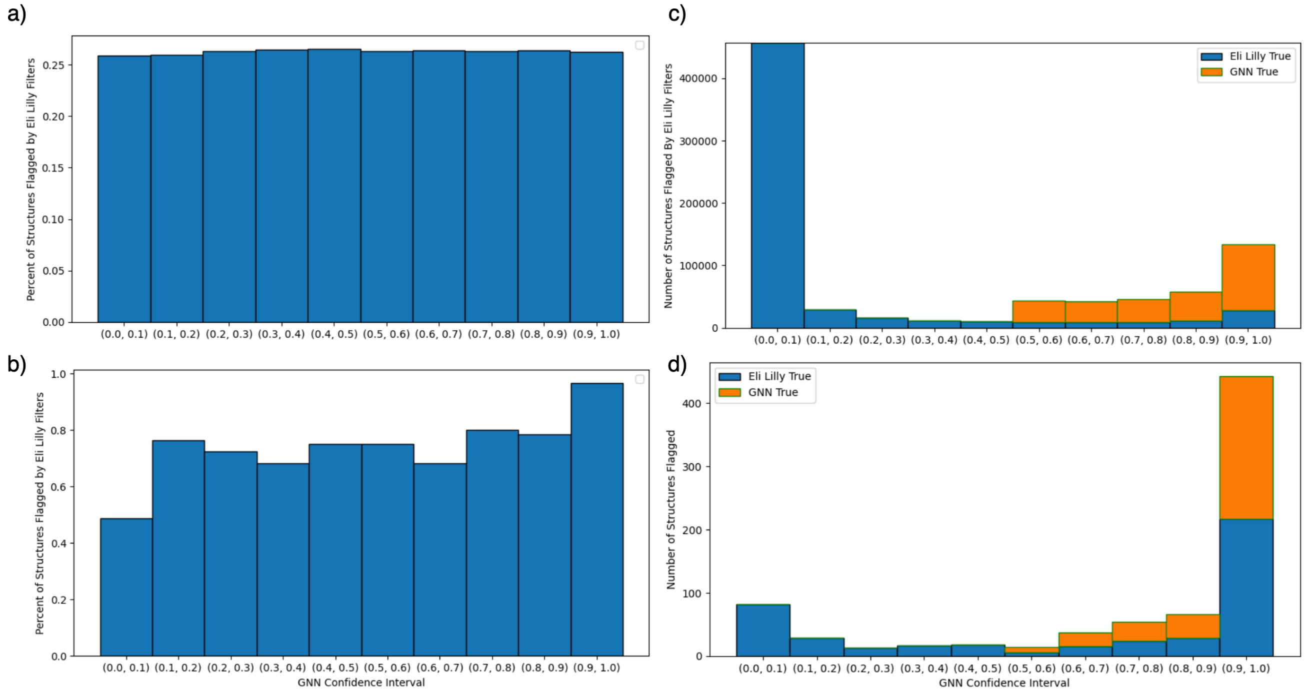


Figure S3: a) and b) - Fraction of structures tagged by Eli Lilly at different GNN confidence intervals in a) ChEMBL dataset and external and b) external test set. c) and d) - Number of structures tagged by Eli Lilly at different GNN confidence intervals in c) ChEMBL dataset and external and d) external test set.

Table S4: GradCAM heat maps of compounds in the atypical subset of the external test set that were positively classified

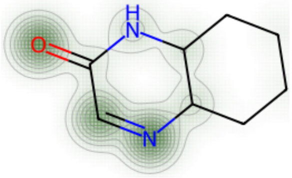
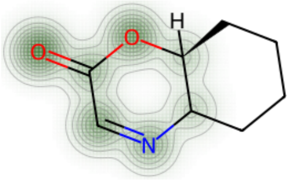
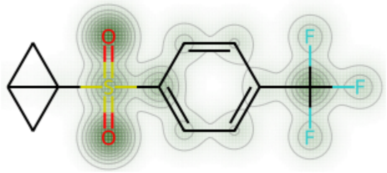
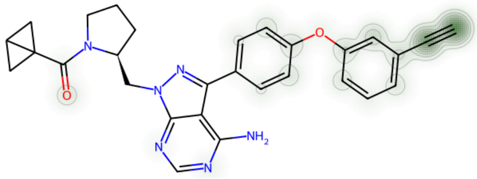
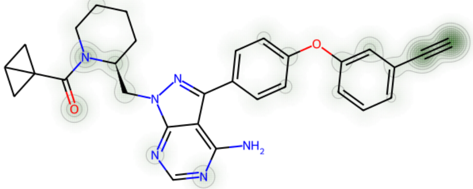
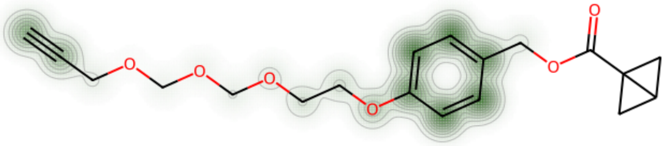
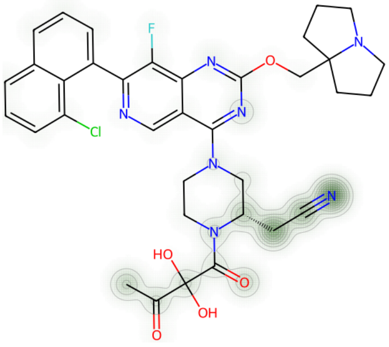
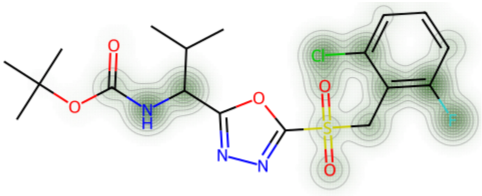
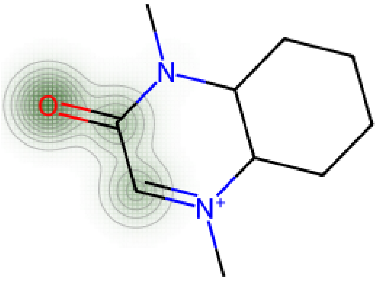
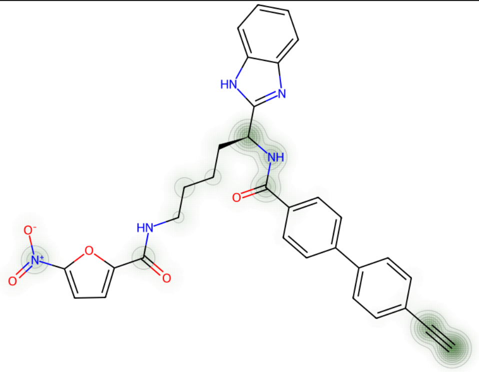
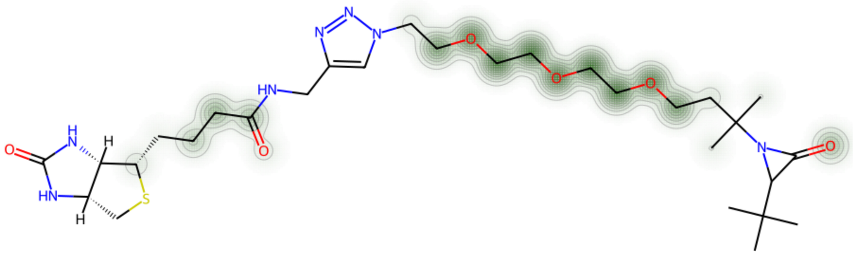
	
	
	
	
	
	

Table S5: negative classification

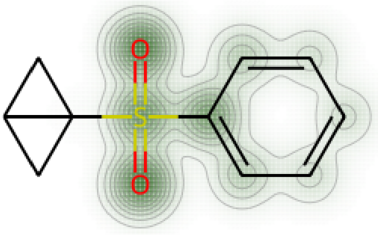
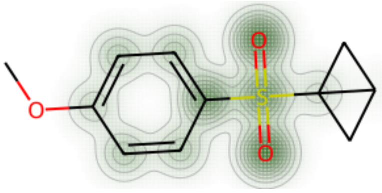
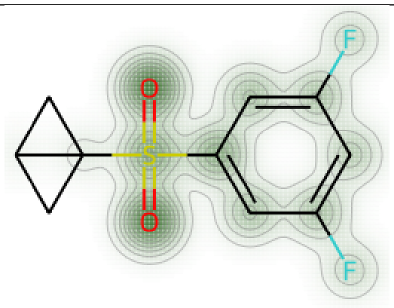
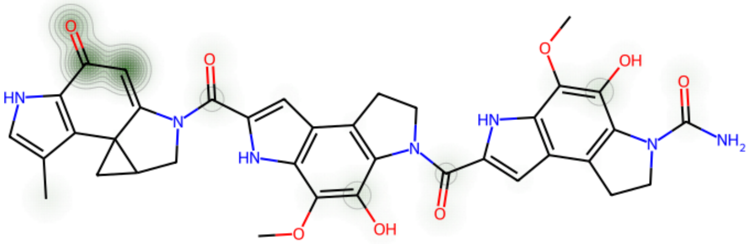
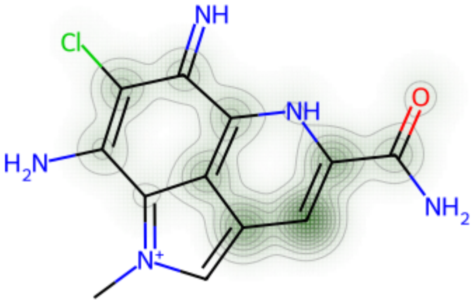
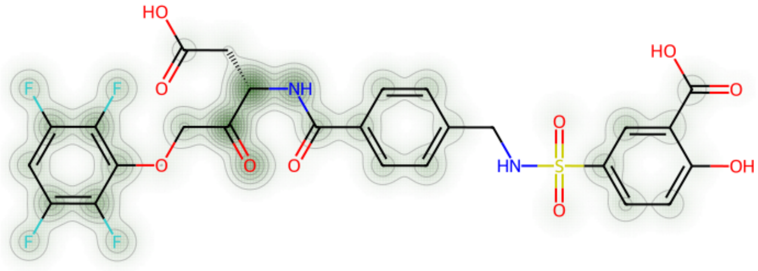
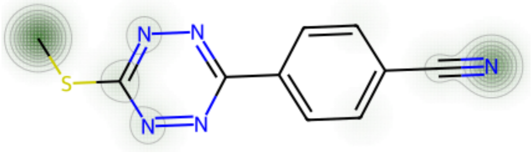
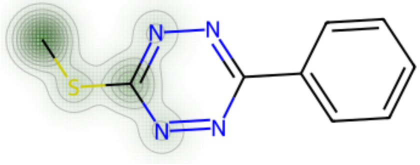
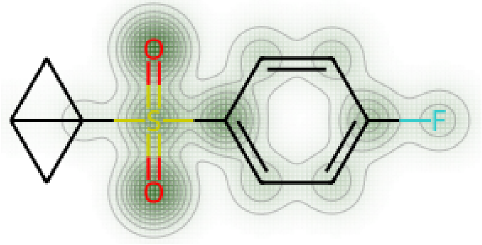
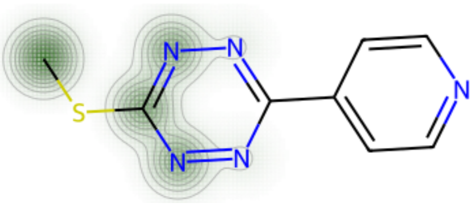
	
	
	
	
	

Table S5: (continued)

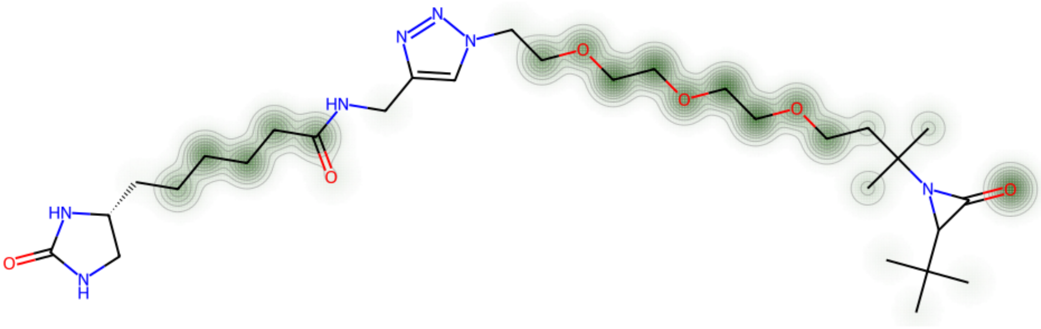
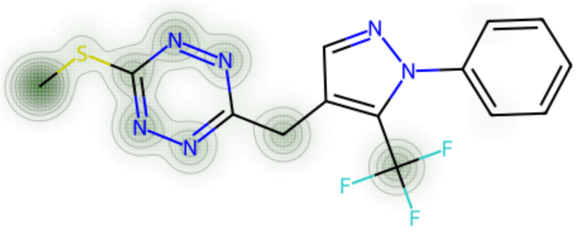
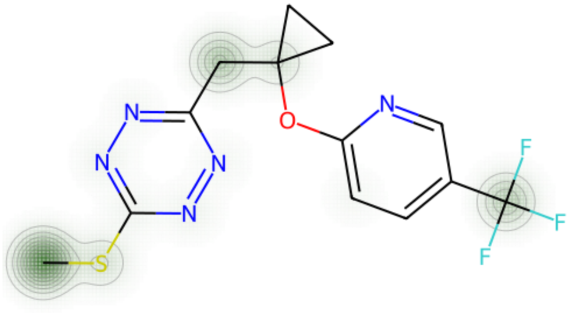
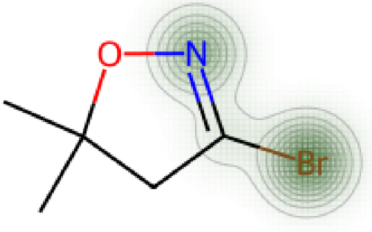
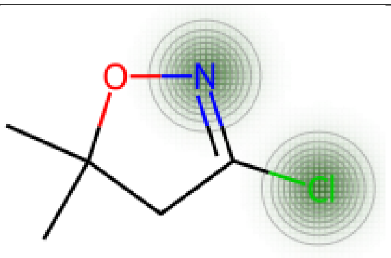
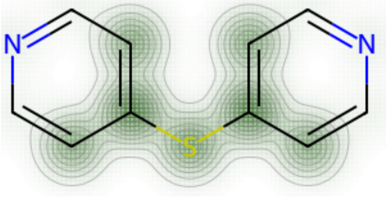
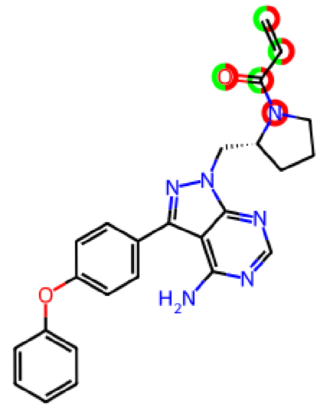
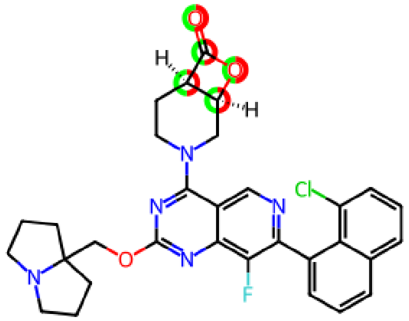
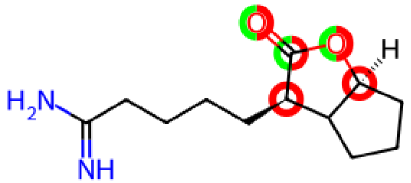
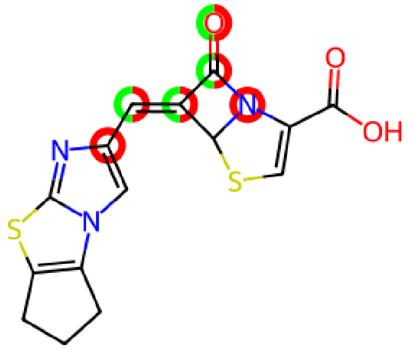
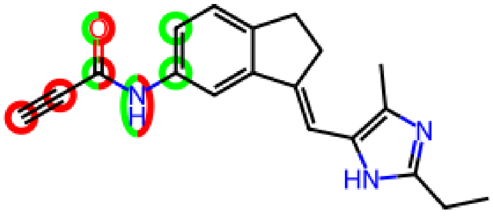
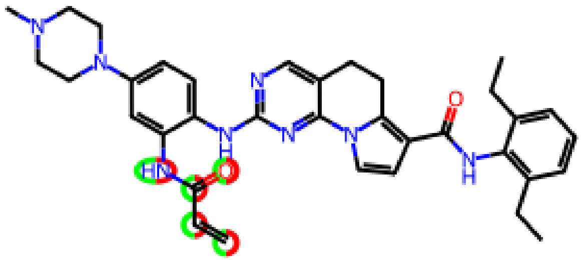
	
	
	
	

Table S6: Selected examples of how GradCAM heatmaps correspond to SMARTS-based filters. GradCAM tags are indicated in green, filter-based tags are indicated in red.

GNN Training Progress

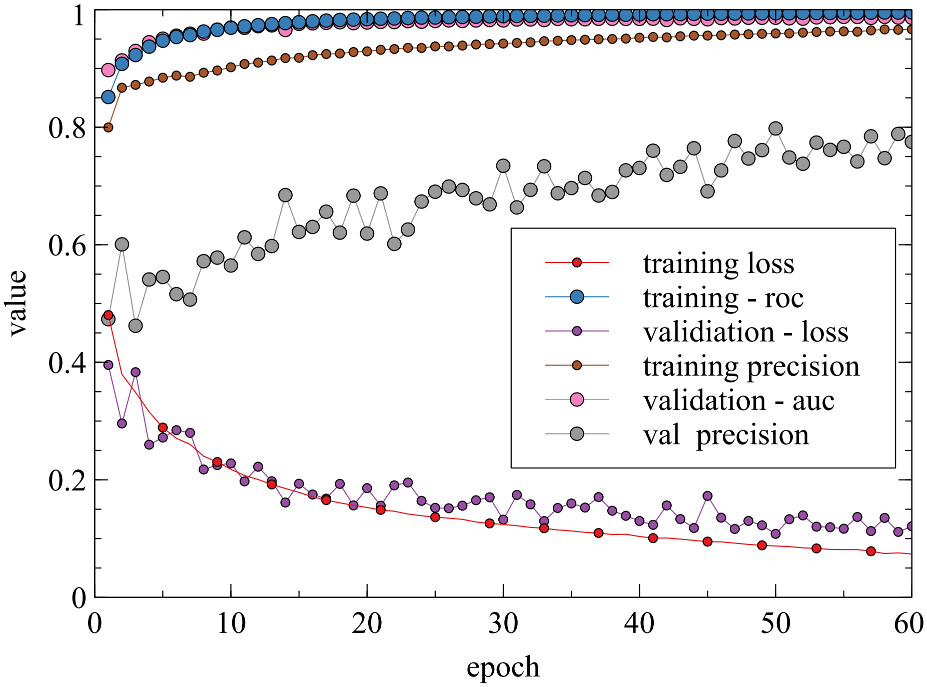


Figure S4: Training process of the GIN network. The loss function of the validation set decreases only incrementally after 25 epochs.

References

(S1) Fukushima, K. Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics* **1975**, *20*, 121–136.

(S2) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2017; http://arxiv.org/abs/1412.6980.

(S3) Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-Normalizing Neural Networks. 2017; http://arxiv.org/abs/1706.02515.

(S4) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2011**, *40*, D1100–D1107.