The University Of Sheffield.

Lhasa Limited

Sheffield Chemoinformatics Research Group

Interpreting neural network models for toxicity prediction by extracting learned chemical features

30/06/2022

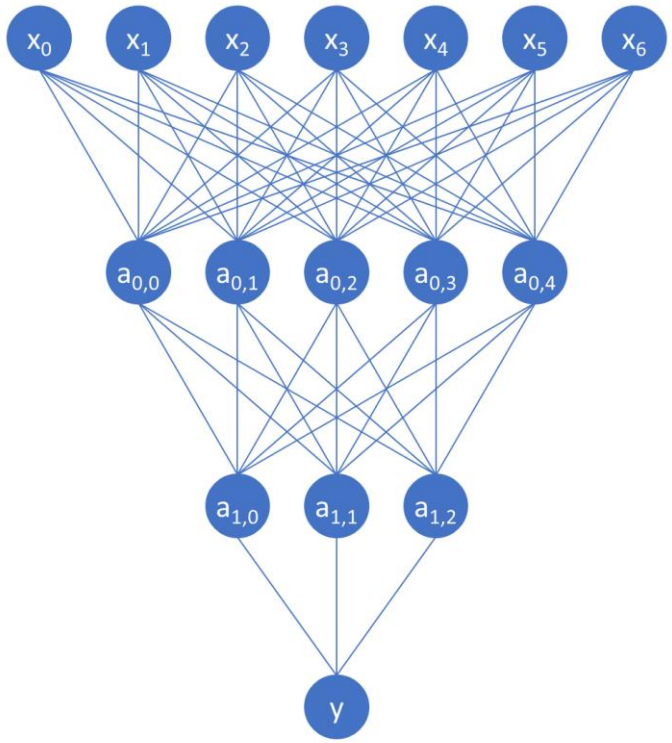
Moritz Walter

Strasbourg Summer School in Chemoinformatics 2022

A WORLD TOP 100 UNIVERSITY

The University Of Sheffield.

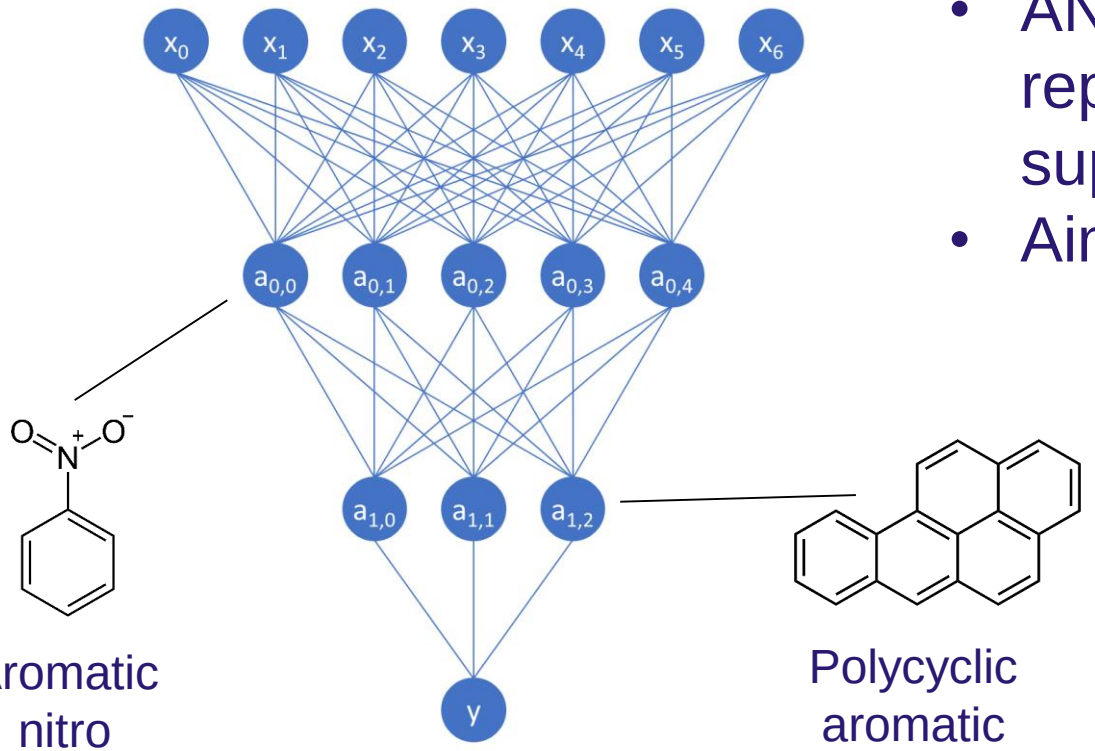
Chemical feature visualisation



- ANN/DNN model: hidden layer neurons learn representation of data suitable to solve supervised task (classification/regression)
- Aim: find chemical features detected in neurons

Chemical feature visualisation

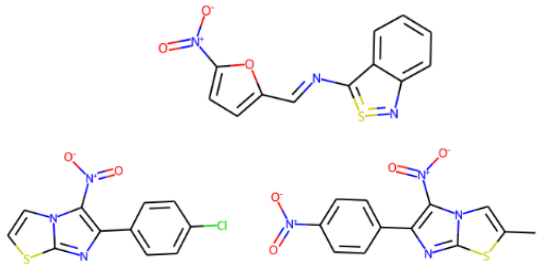
Modelled endpoint: mutagenicity



- ANN/DNN model: hidden layer neurons learn representation of data suitable to solve supervised task (classification/regression)
- Aim: find chemical features detected in neurons

The University Of Sheffield.

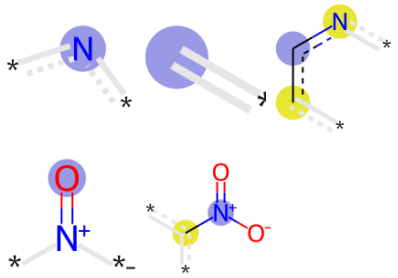
Automatic substructure extraction



Training compounds with high activation

+

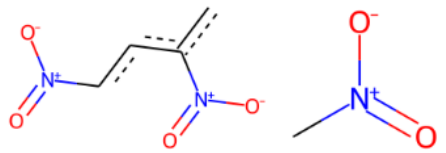
Input fingerprint bits with high weight



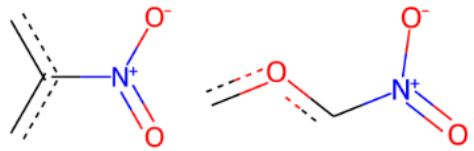
→ **FCA** →

formal concept X
Extent: compound 3465, compound 482,…
Intent: Bit 24, Bit 346, Bit 1098, Bit 1532, …

▪
▪
▪



Extracting chemical substructures



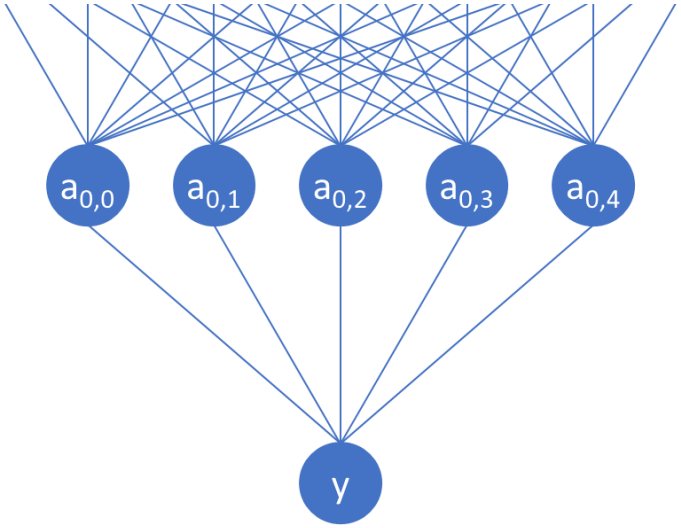
- FCA (Formal Concept Analysis) identifies combinations of compounds and FP bits (formal concepts)
- From those chemical substructures are extracted if associated with neuron activation

The University Of Sheffield.

From substructures to atom attributions

1 Determine importance of neuron for individual prediction

- Integrated gradients (IG) on hidden neurons

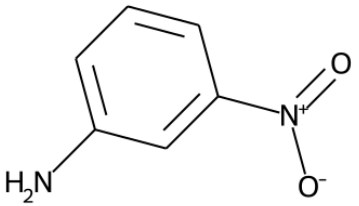


Hidden layer

Neuron IG attributions:

0.32 – 0.22 – -0.12 – 0.42 – 0.07

Output



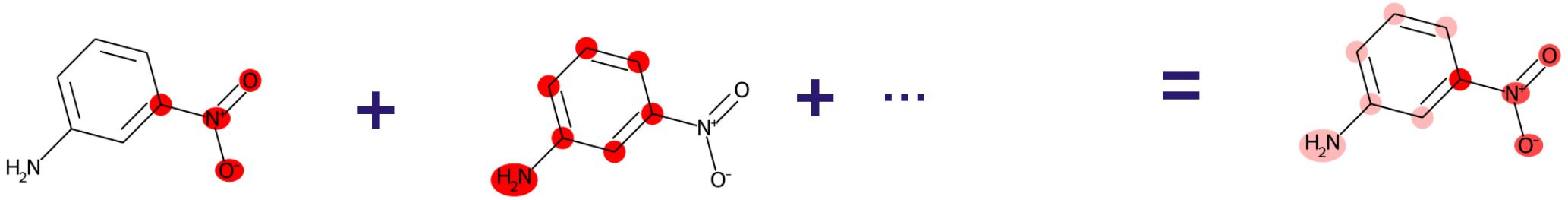
Prediction: 0.91

From substructures to atom attributions

2 Map neuron importances onto structure:

- Find most specific matching substructure(s) in trees
- Share attribution between atoms of substructure

Neuron 0 (0.32) Neuron 1 (0.22)



Integrated gradients on input features as comparison[1]

1 Preuer et al. 2019: Interpretable deep learning in drug discovery

Neural network model

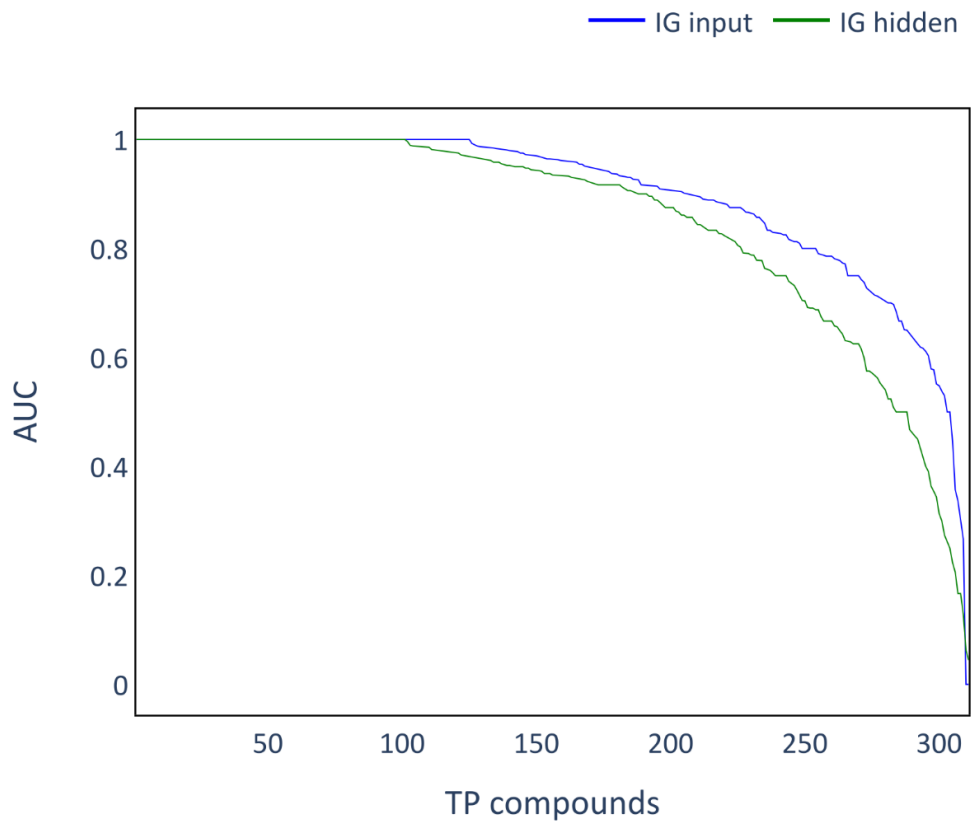
- Dataset on Ames mutagenicity (~8k)
  - Hansen (curated), ISSSTY, ECVAM, CGX, Snyder
- Derek expert system used to label compounds (structural alerts for mutagenicity)
- Model architecture: 1 hidden layer (512 neurons)
- Input: Morgan FP (radius=1, 2048 bits)
- High performance on test set: ACC: 0.91, ROC-AUC: 0.97, Recall: 0.91, Precision: 0.92

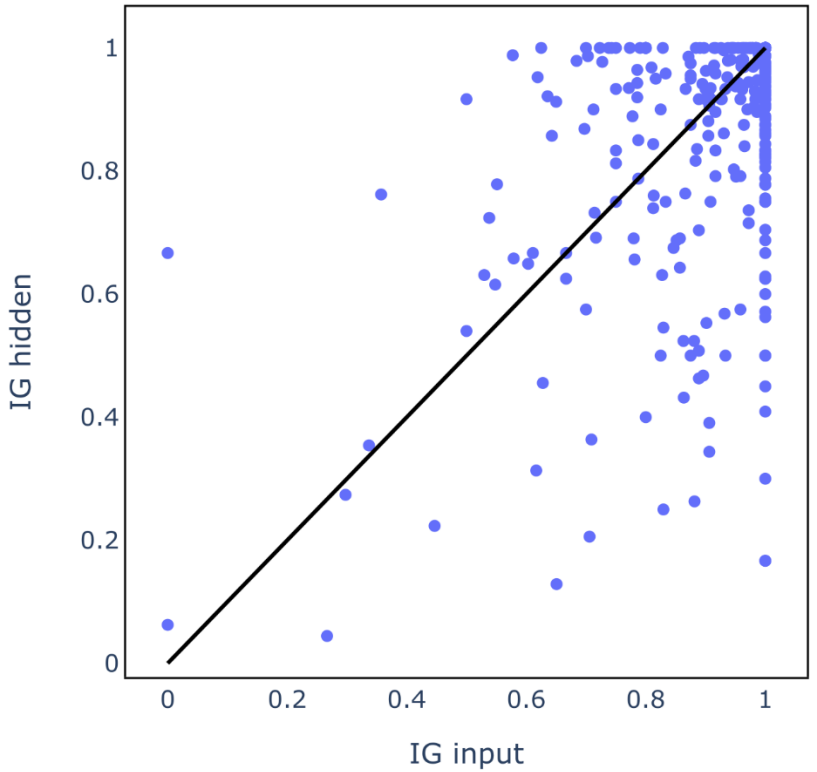
Evaluation

- Individual compounds: attribution AUCs for TP compounds
- Alerts: compute average AUCs for compounds matching a given alert

Explanatory performance

	Median AUC	AUC ≥0.8
IG input	0.964	255/306
IG hidden neurons	0.935	227/306

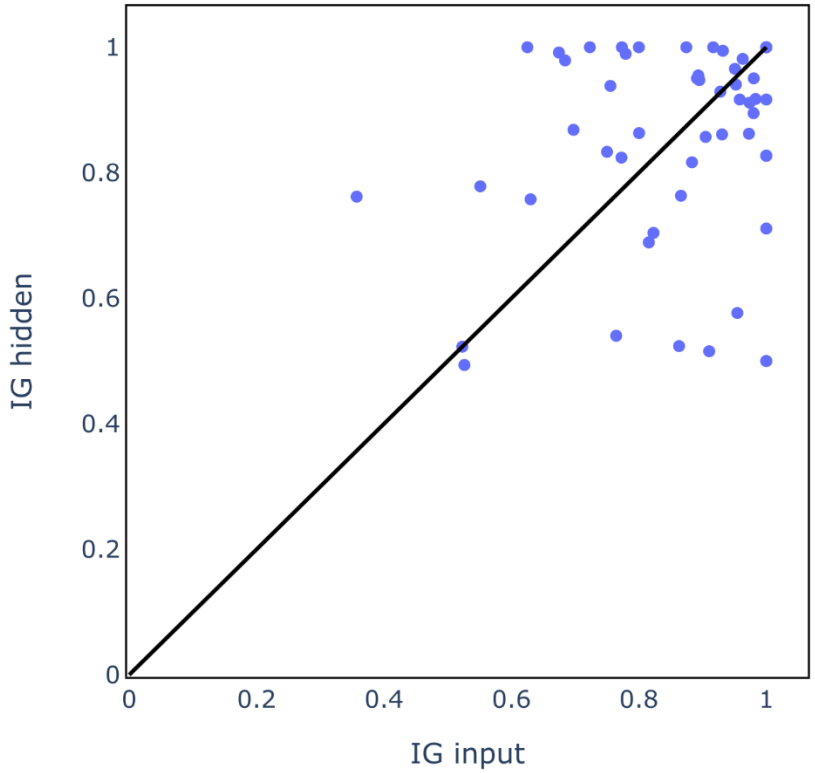




The University Of Sheffield.

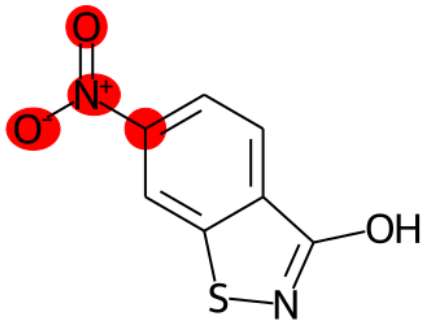
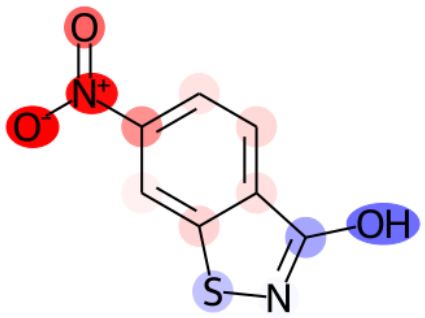
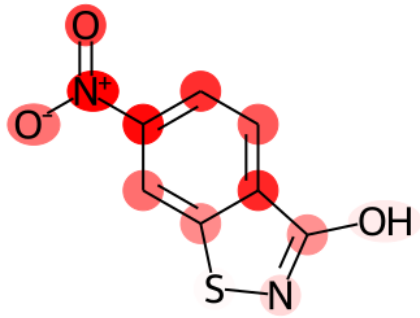
Alert performances

	Median AUC	AUC ≥0.8
IG input	0.894	36/52
IG hidden	0.903	37/52

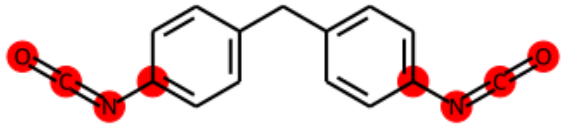
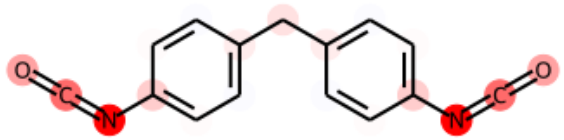
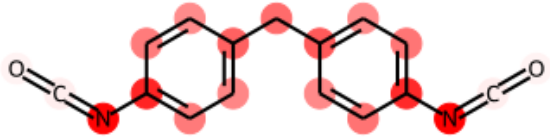


The University Of Sheffield.

Individual compounds

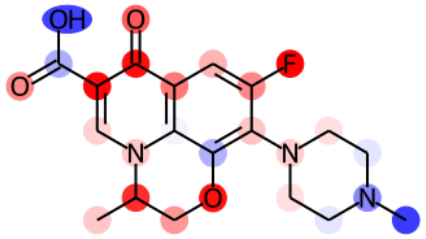
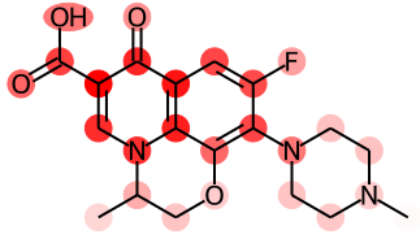
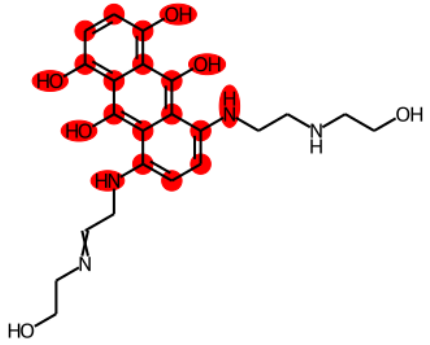
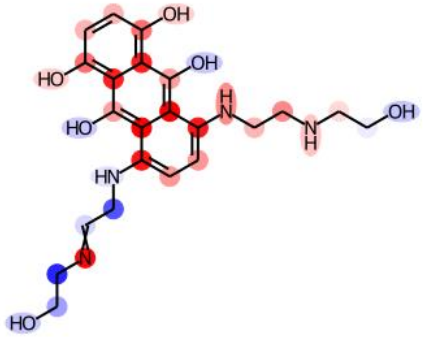
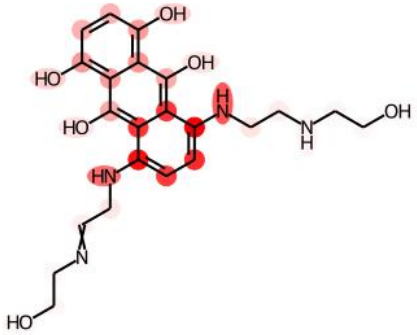
Derek Alert	IG input	IG hidden
		
Arom. nitro	AUC = 1	AUC = 0.83

Contributing to toxic prediction
Contributing to non-toxic prediction

		
Isocyanate	AUC = 1	AUC = 0.5

The University Of Sheffield.

Individual compounds

Derek Alert	IG input	IG hidden
		
Quinolone-3-carboxylic acid	AUC = 0.577	AUC = 0.988
		
Hydroxylated anthraquinone	AUC = 0.8	AUC = 1

Conclusion

- Method to visualize chemical features learned in hidden layers
- Extracted fragments can be used to interpret neural network model
- Method limited by quality of extracted fragments
- Different explanation methods have strengths and weaknesses

→ Benchmarking required

The University Of Sheffield.

Acknowledgement

Sheffield Chemoinformatics Research Group

Prof. Dr. Val Gillet

Lhasa Limited

Dr. Sam Webb

Integrated gradients (IG)

- Determines importance of each input feature for given prediction
- Integration of gradients (of model output wrt feature) along straight path between baseline (bit vector of 0s) and instance

$$a_i(x) = (x_i - x_i') \int_{a=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

a_i: attribution for feature i
x_i: feature i
x'_i: feature i in baseline (0)
F: NN model
x_i: feature i
α: path x' -> x
m: number of steps
k: current step

Integral approximated using a sum:

$$a_i(x) \approx (x_i - x_i') \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}$$

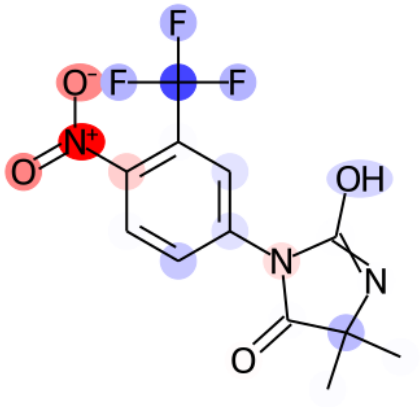
The University Of Sheffield.

Alert performances

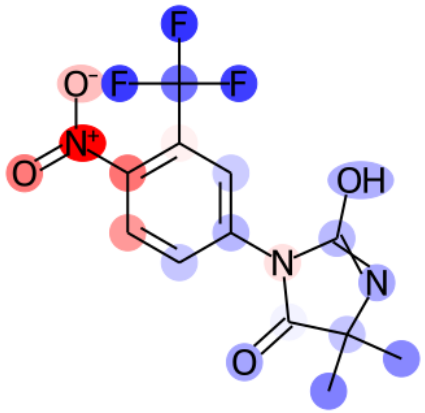
Alert	Proportion train set	IG input	IG hidden neurons
Aromatic nitro	0.130	**0.983**	0.908
Alkylating agent	0.058	0.900	**0.918**
PAHs	0.043	**0.764**	0.540
Epoxide	0.033	**0.974**	0.912
N-Nitroso	0.029	**0.980**	0.950
Isocyanate	0.002	**1**	0.5
Aromatic nitroso	0.007	**1**	0.711
Hydrox. anthraquinone	0.007	0.63	**0.758**
Quinolone-3-carboxylic acid	0.003	0.674	**0.992**

Negative prediction

IG input



IG hidden



Taken from model trained on experimental Ames labels
Prediction: 0.41
Label: negative