



The
University
Of
Sheffield.



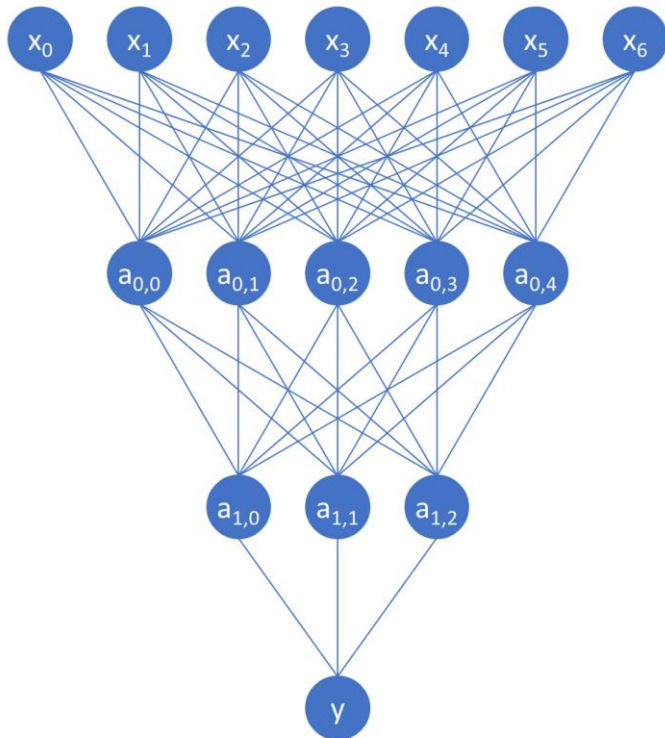
Interpreting neural network models for toxicity prediction by extracting learned chemical features

29/09/2022

Moritz Walter

3rd *In Silico* Toxicology Conference 2022

Chemical feature visualisation

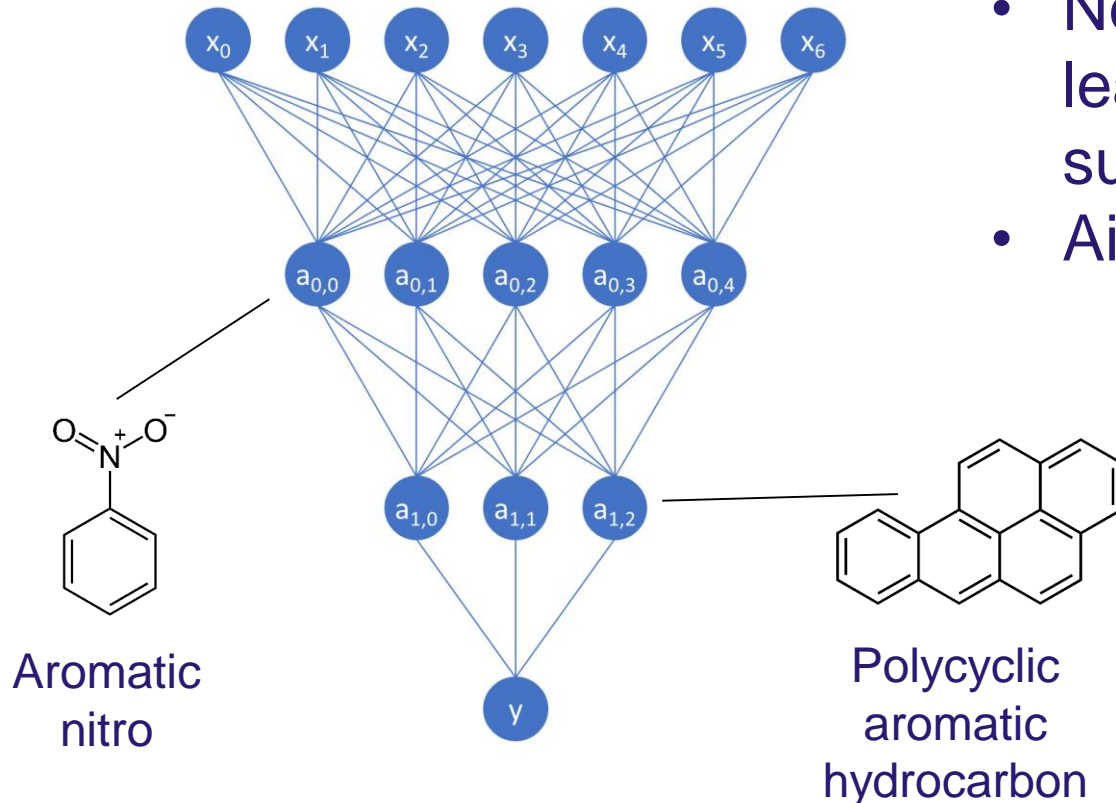


- Neural network model: hidden layer neurons learn representation of data suitable to solve supervised task (classification/regression)
- Aim: find chemical features detected in neurons

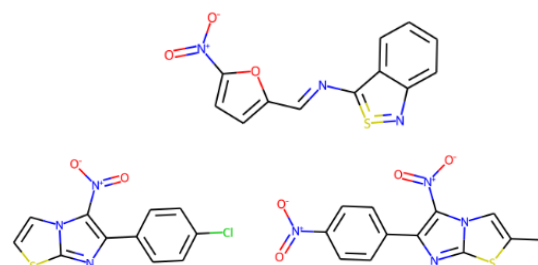
Chemical feature visualisation

Modelled endpoint: mutagenicity

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



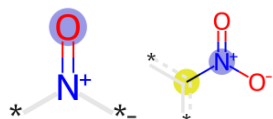
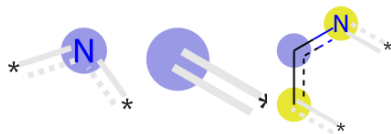
Automatic substructure extraction



Training compounds with high activation

+

Input fingerprint bits with high weight



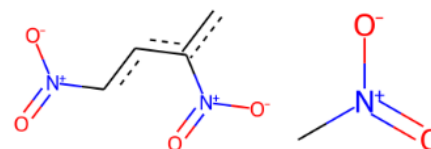
FCA



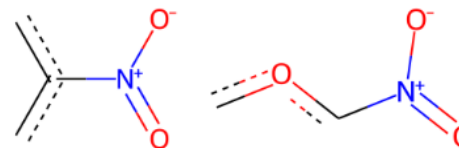
formal concept X

Extent: compound 3465, compound 482, ...
Intent: Bit 24, Bit 346, Bit 1098, Bit 1532, ...

-
-
-



Extracting chemical substructures

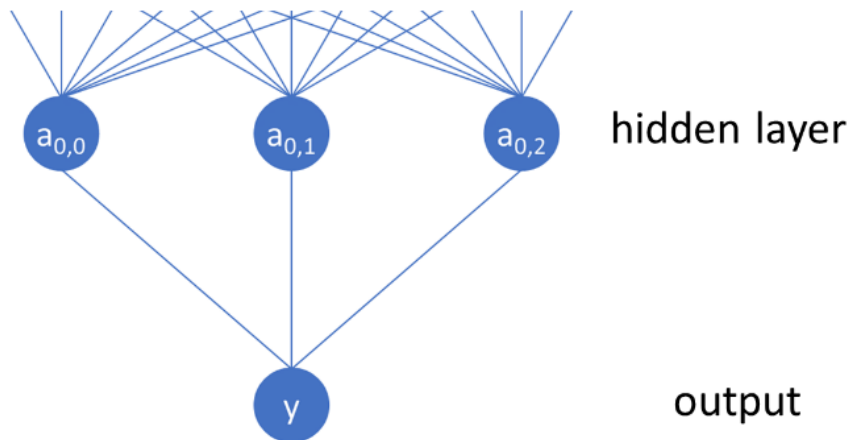


- FCA (Formal Concept Analysis) identifies combinations of compounds and FP bits (formal concepts)
- From those chemical substructures are extracted if associated with neuron activation

From substructures to atom attributions

1 Determine attribution of neurons for individual prediction

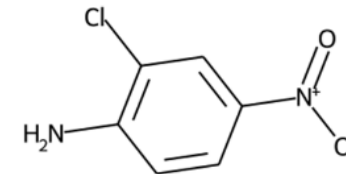
- Integrated gradients (IG) on hidden neurons



Neuron attributions:

0.24 , 0.14 , -0.05

positive negative
attributions attribution



Prediction: 0.83

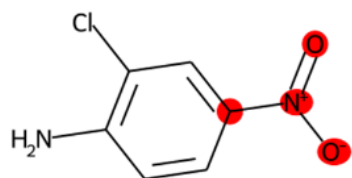
2 Map neuron attributions onto structure

- Find most specific matching substructure(s) in subnetworks
- Share attribution between atoms of substructure (here: unweighted)

Neuron 0

Neuron attribution: 0.24

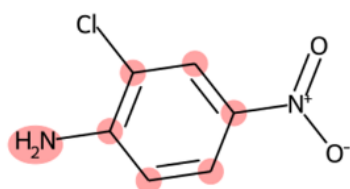
Atom attribution: 0.06



Neuron 1

Neuron attribution: 0.14

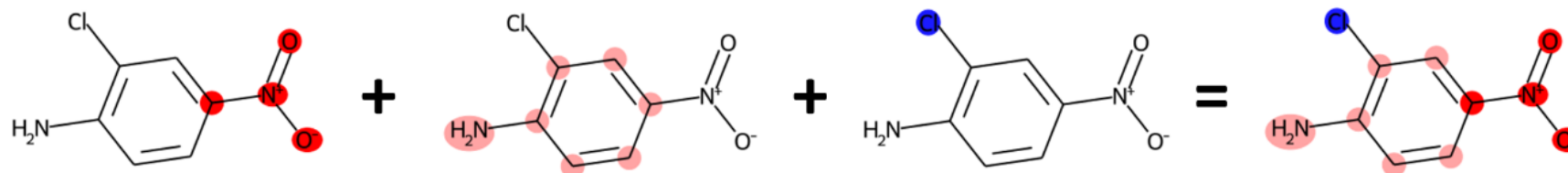
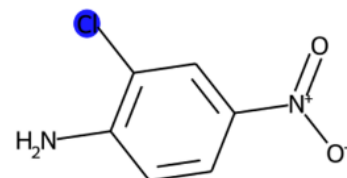
Atom attribution: 0.02



Neuron 2

Neuron attribution: -0.05

Atom attribution: -0.05



Integrated gradients on input features as comparison¹

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



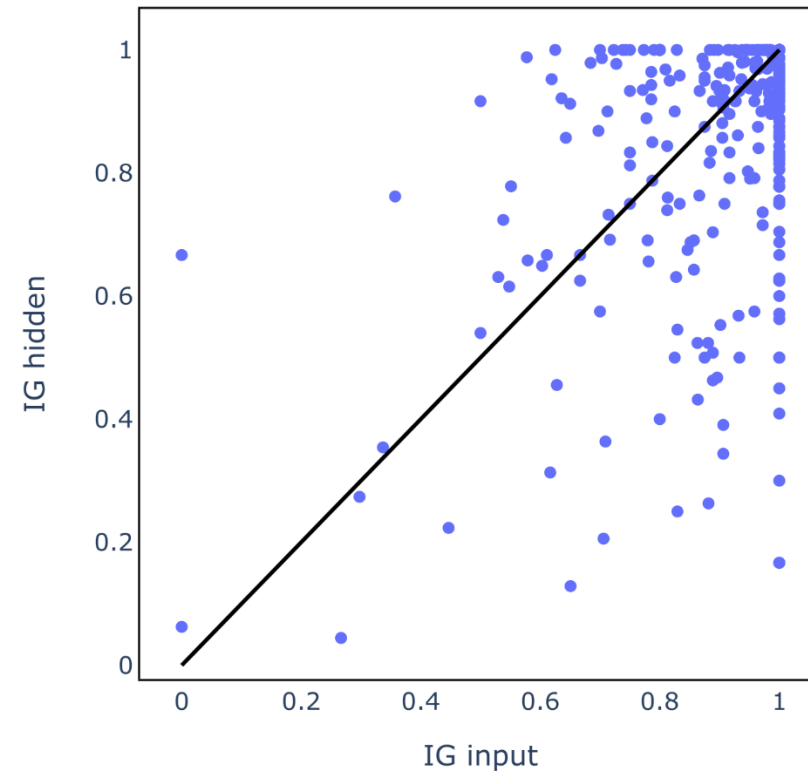
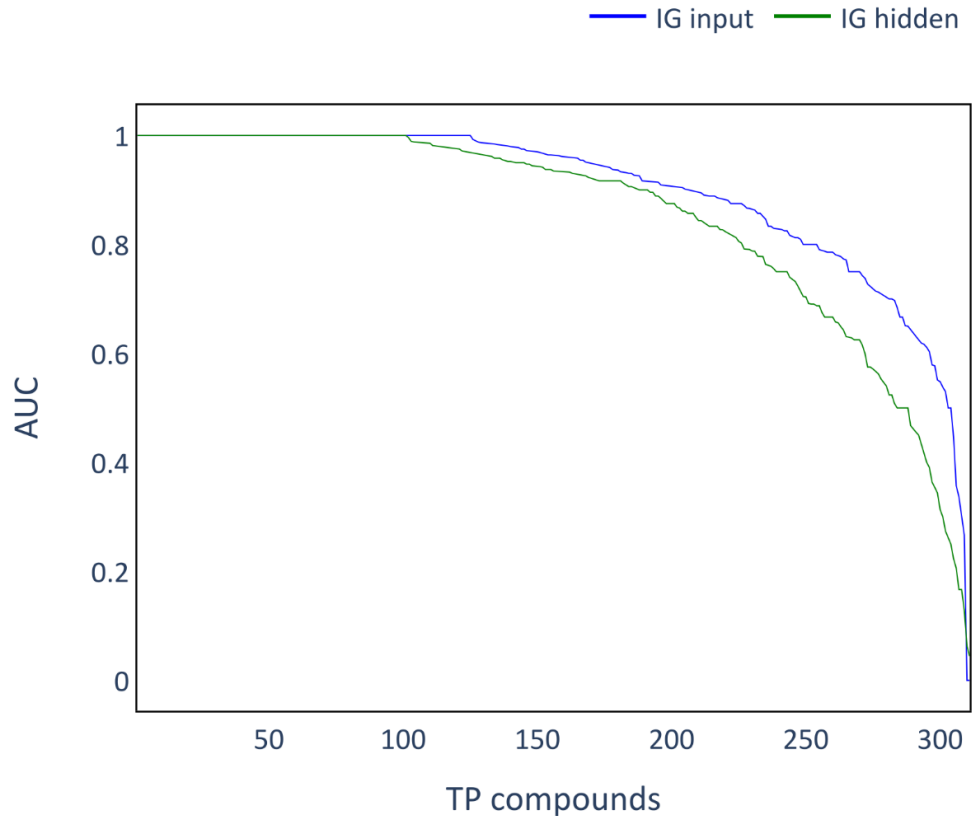
Evaluation

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



Explanatory performance

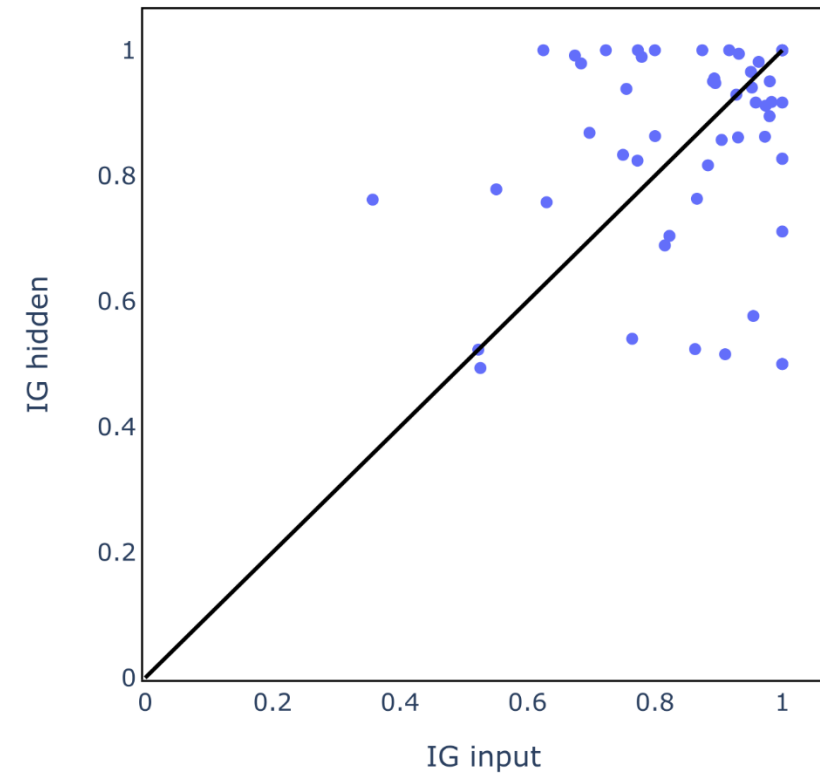
	Median AUC	AUC ≥ 0.8
IG input	0.964	255/306
IG hidden neurons	0.935	227/306





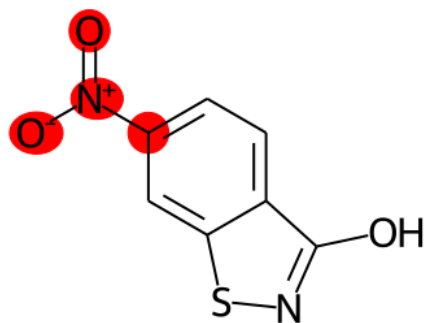
Alert performances

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



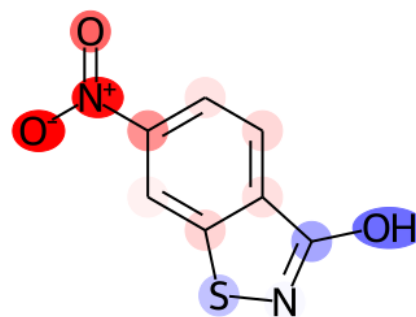
Individual compounds

Derek Alert



Arom. nitro

IG input

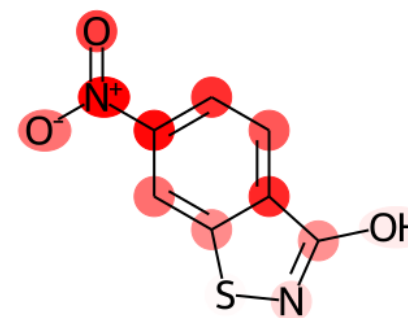


AUC = 1

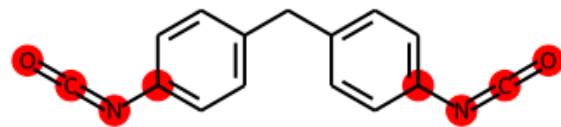
Contributing to toxic prediction

Contributing to non-toxic prediction

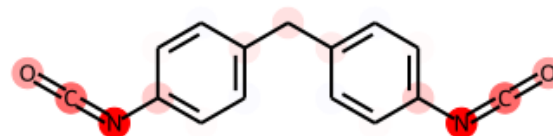
IG hidden



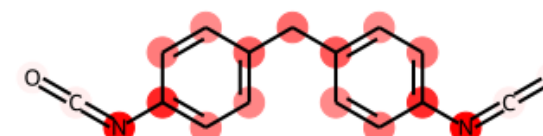
AUC = 0.83



Isocyanate



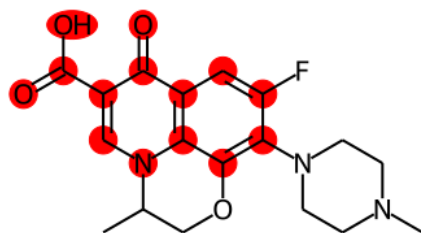
AUC = 1



AUC = 0.5

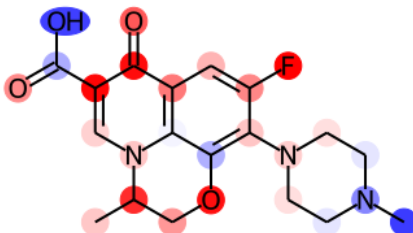
Individual compounds

Derek Alert



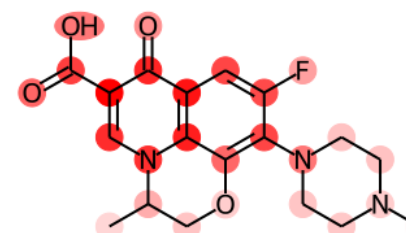
Quinolone-3-carboxylic acid

IG input

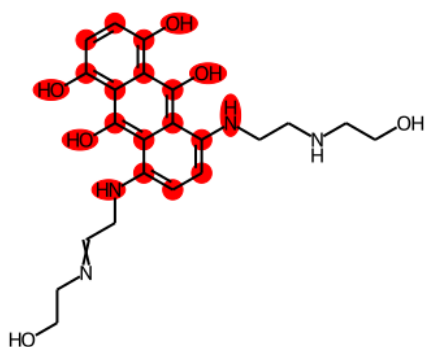


AUC = 0.577

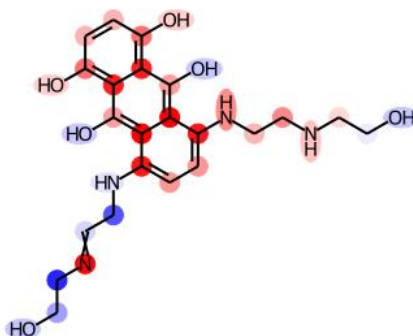
IG hidden



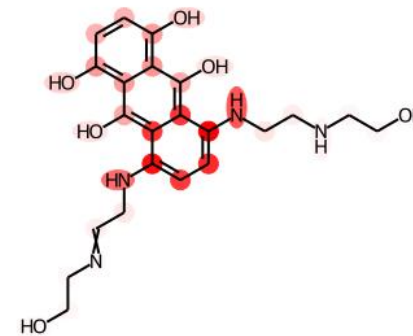
AUC = 0.988



Hydroxylated anthraquinone



AUC = 0.8

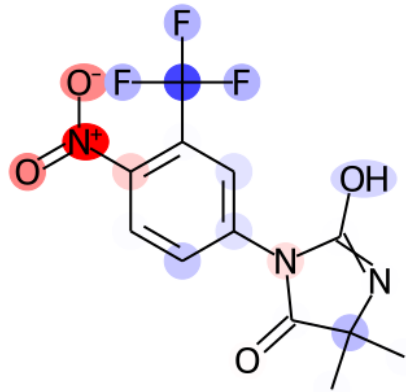


AUC = 1

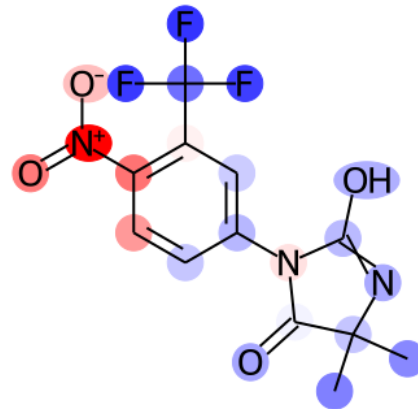


Negative prediction

IG input



IG hidden



Taken from model trained on experimental Ames labels



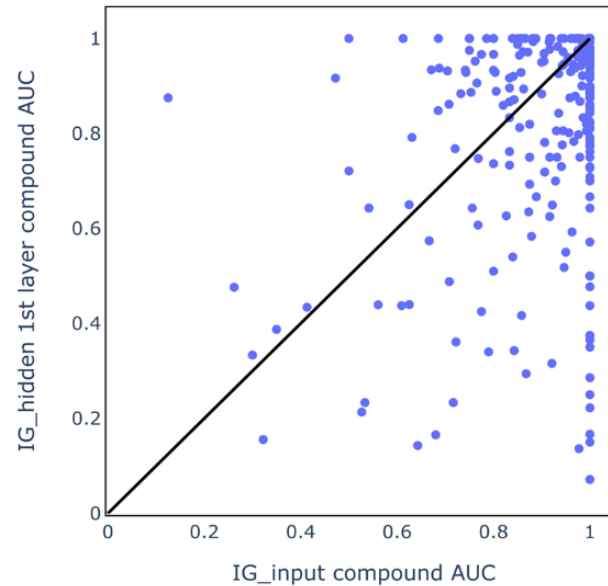
2-layer network (deep neural network)

	Median AUC	Median alert AUC
IG input	0.984	0.907
IG hidden	0.938	0.906

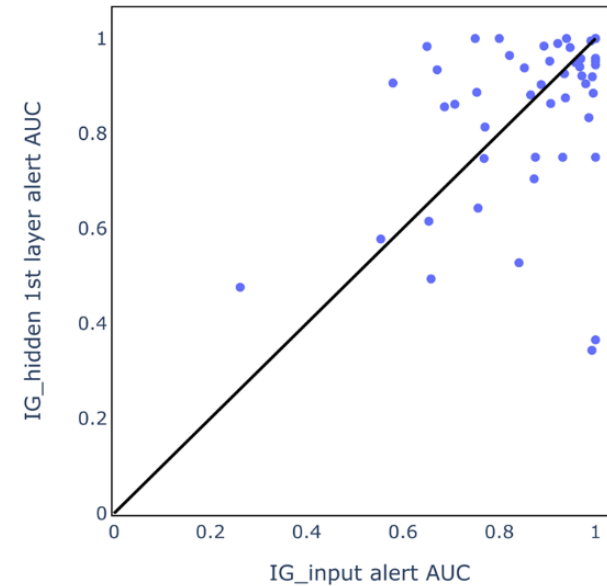
A: individual compound AUCs

B: average alert AUCs

A

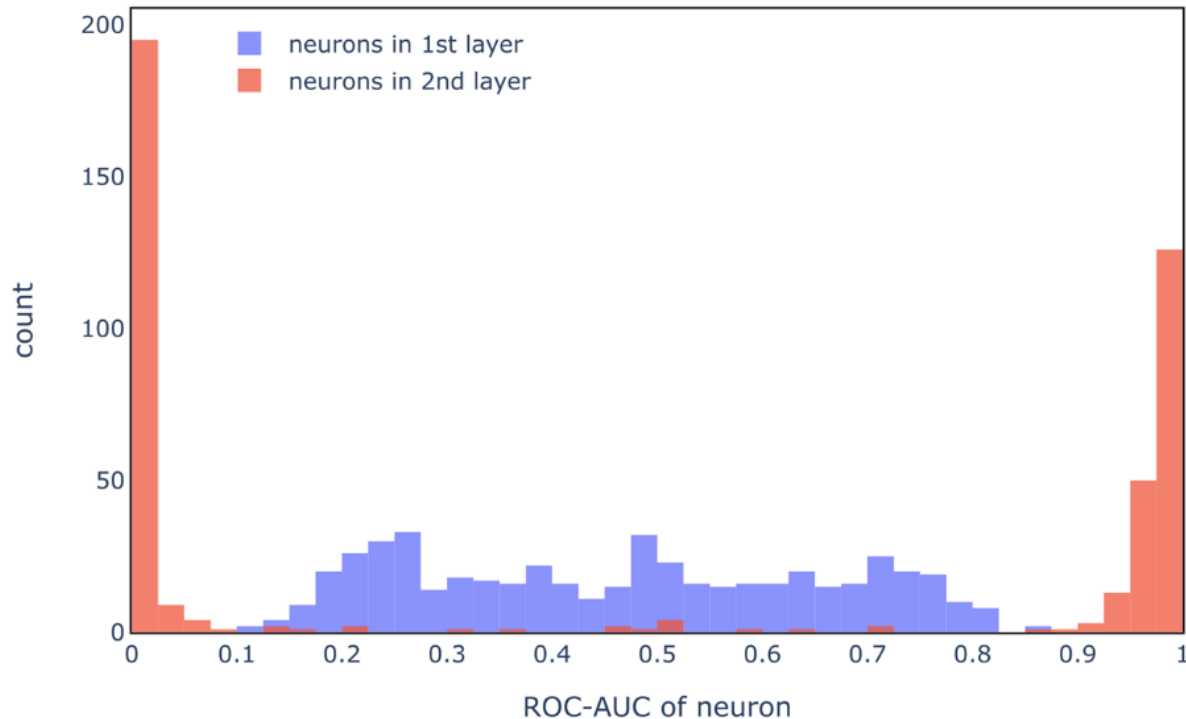


B





Neurons in 2nd layer don't detect specific chemical features



It is tested how well each individual neuron 'ranks' toxic compounds higher than non-toxic ones (based on neuron activation)



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



Prof. Dr. Val Gillet



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

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

Integral approximated using a sum:

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

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 $x' \rightarrow x$
 m : number of steps
 k : current step



Neuron correlations DNN

