AmesFormer: A Graph Transformer Neural Network for Muatgenicity Prediction

Luke Thompson
Supervisor: Slade Matthews

THE UNIVERSITY OF
SYDNEY

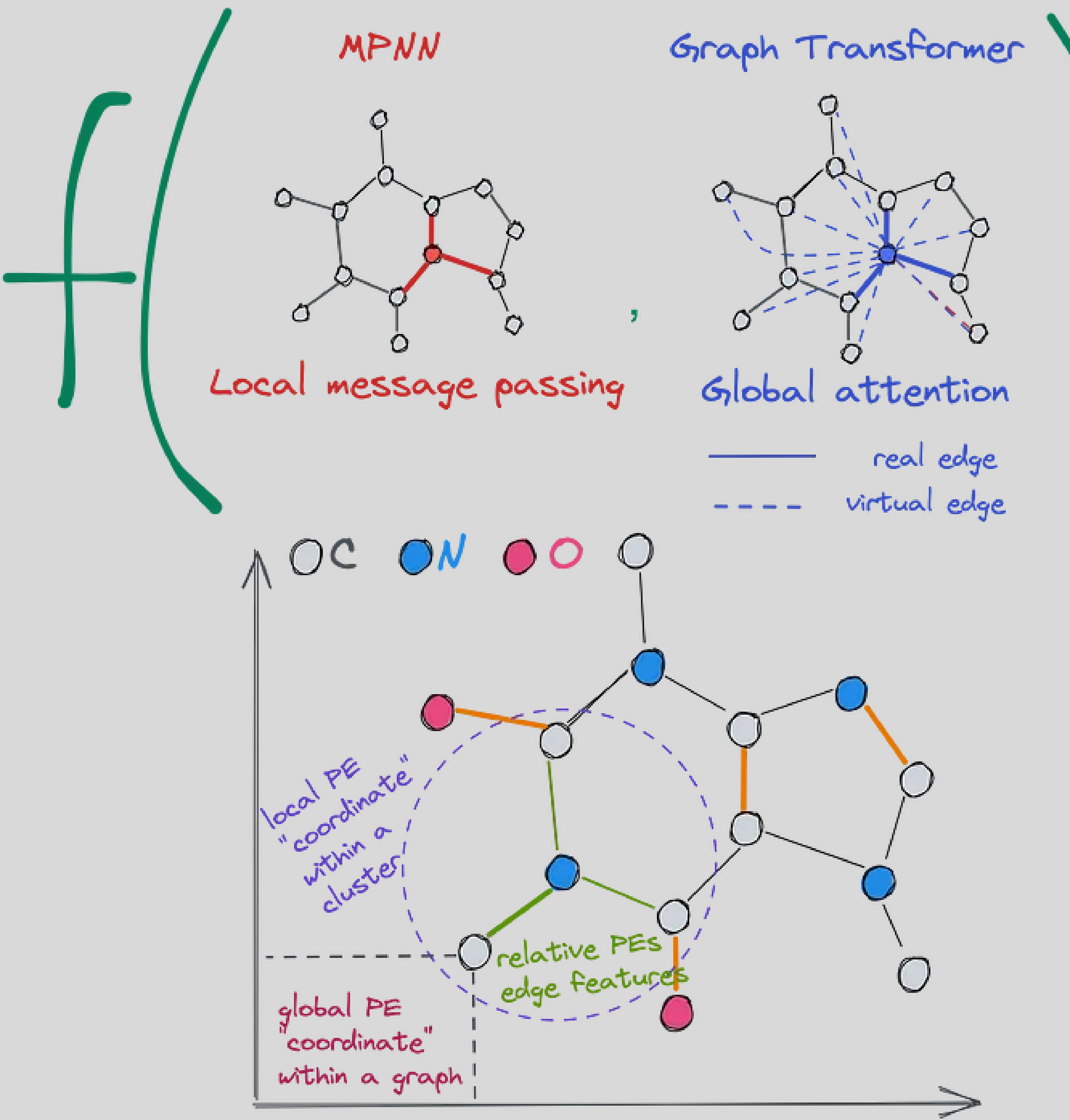
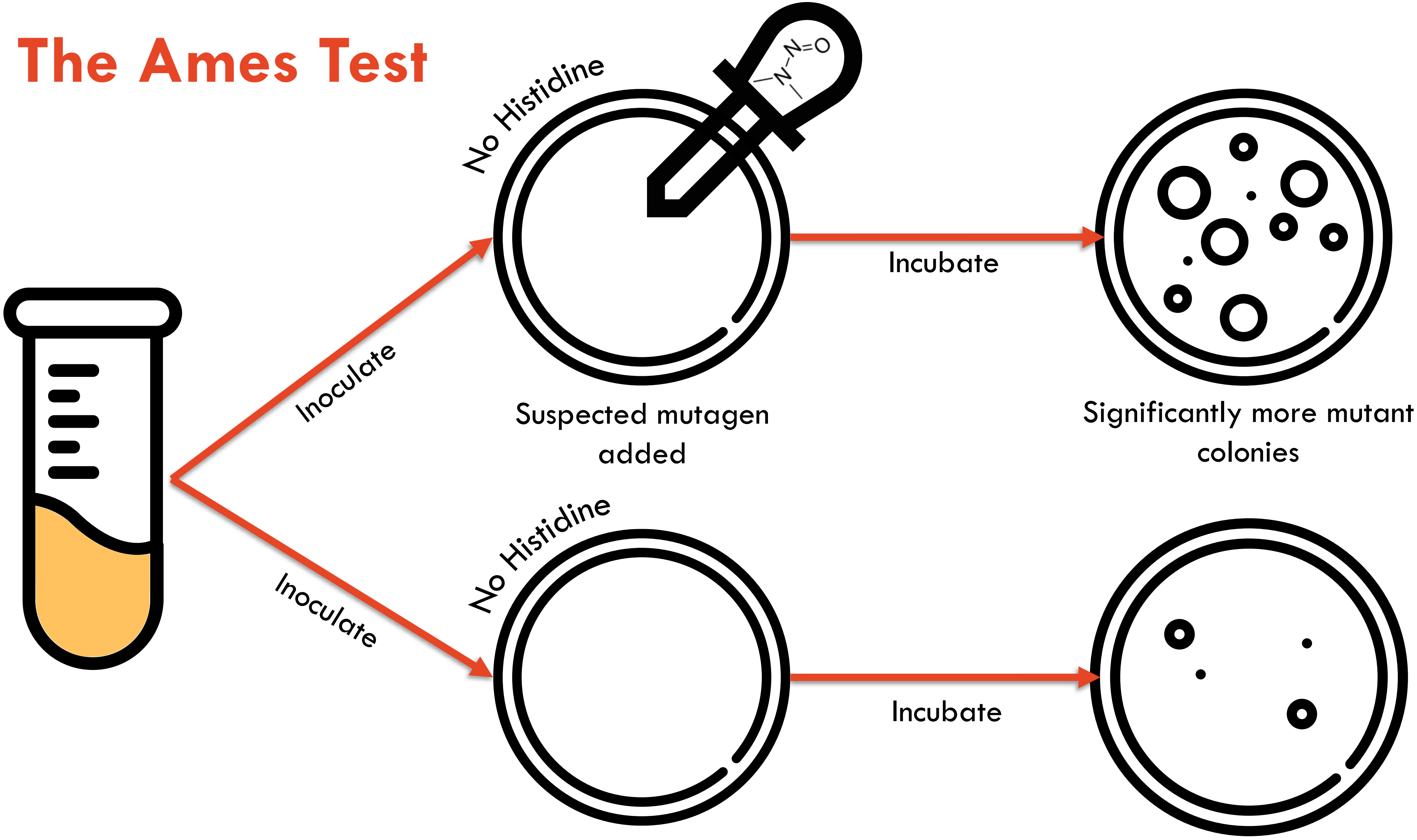


Image: https://towardsdatascience.com/graphgps-navigating-graph-transformers-c2cc223a051c

The Ames Test



Inoculate

No Histidine

Suspected mutagen added

Incubate

Significantly more mutant colonies

Inoculate

No Histidine

Control added

Incubate

Few natural mutant colonies

Mutagenicity Detection is a Contemporary Issue

ACCC recalls more jeans containing hazardous dye linked to cancer

By consumer affairs reporter Amy Bainbridge

Posted Thu 15 May 2014 at 3:53pm, updated Thu 15 May 2014 at 6:34pm

Potential contamination of Australian metformin medicines

Low levels of contamination with N-nitrosodimethylamine (NDMA)

Published: 18 November 2020

Textiles recalled after tests for azo dyes

Date	15 May 2014
Topics	Protecting yourself

Product Safety Recall

Do you own this product?

The Laundress Fabric Conditioners

The Laundress brand fabric conditioner products

Sold in various sizes and container styles

Dates sold: 1 January 2021 – 31 March 2023

Why the product is recalled: The recalled products can contain a chemical impurity, ethylene oxide.

Hazard: Ethylene oxide is a carcinogen. Consumers may be exposed to this chemical through inhalation during use of the product, or through skin exposure to residue on clothing. Direct long-term exposure may cause adverse health effects, including cancer.

Five popular sunscreens recalled after a cancer-causing ingredient was added to the batches

Five popular Australian sun safety products have been urgently recalled after a cancer-causing ingredient was detected in the batches.

Georgina Noack

Computational Ames Models

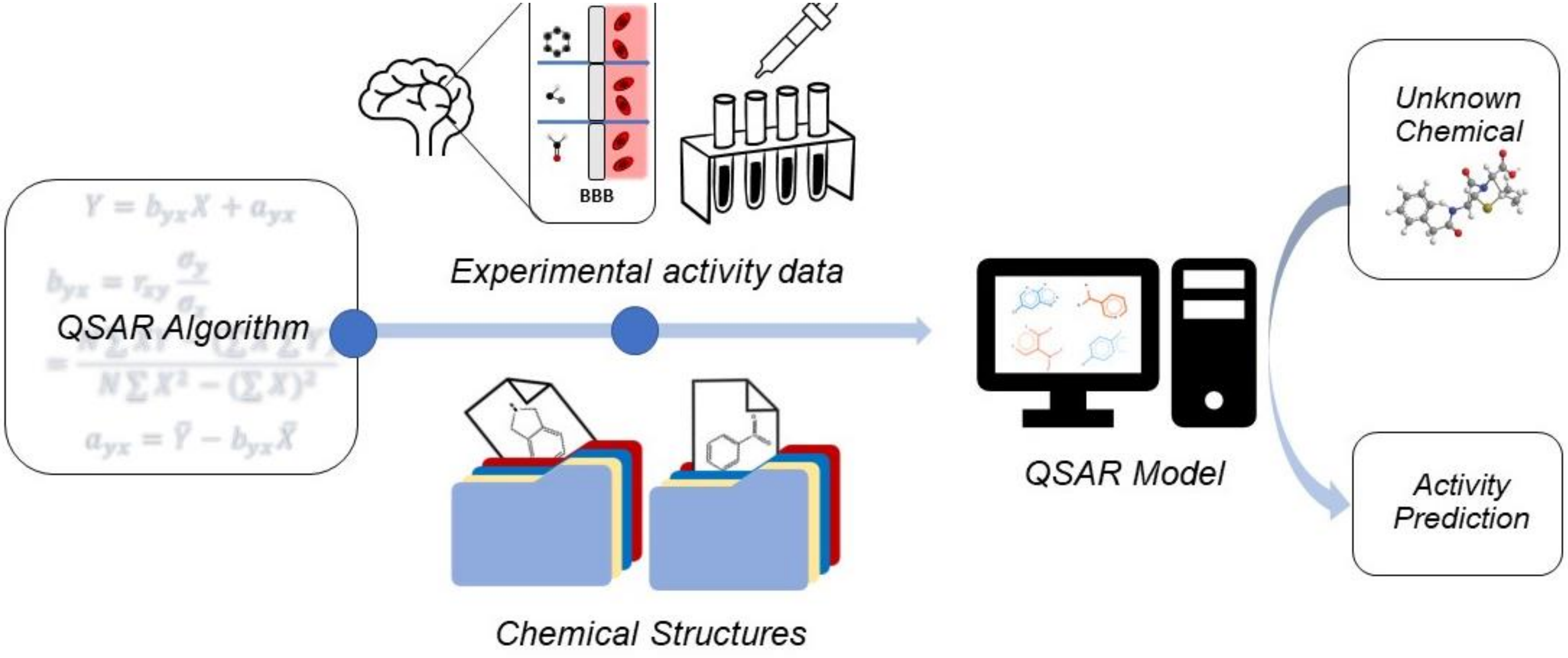
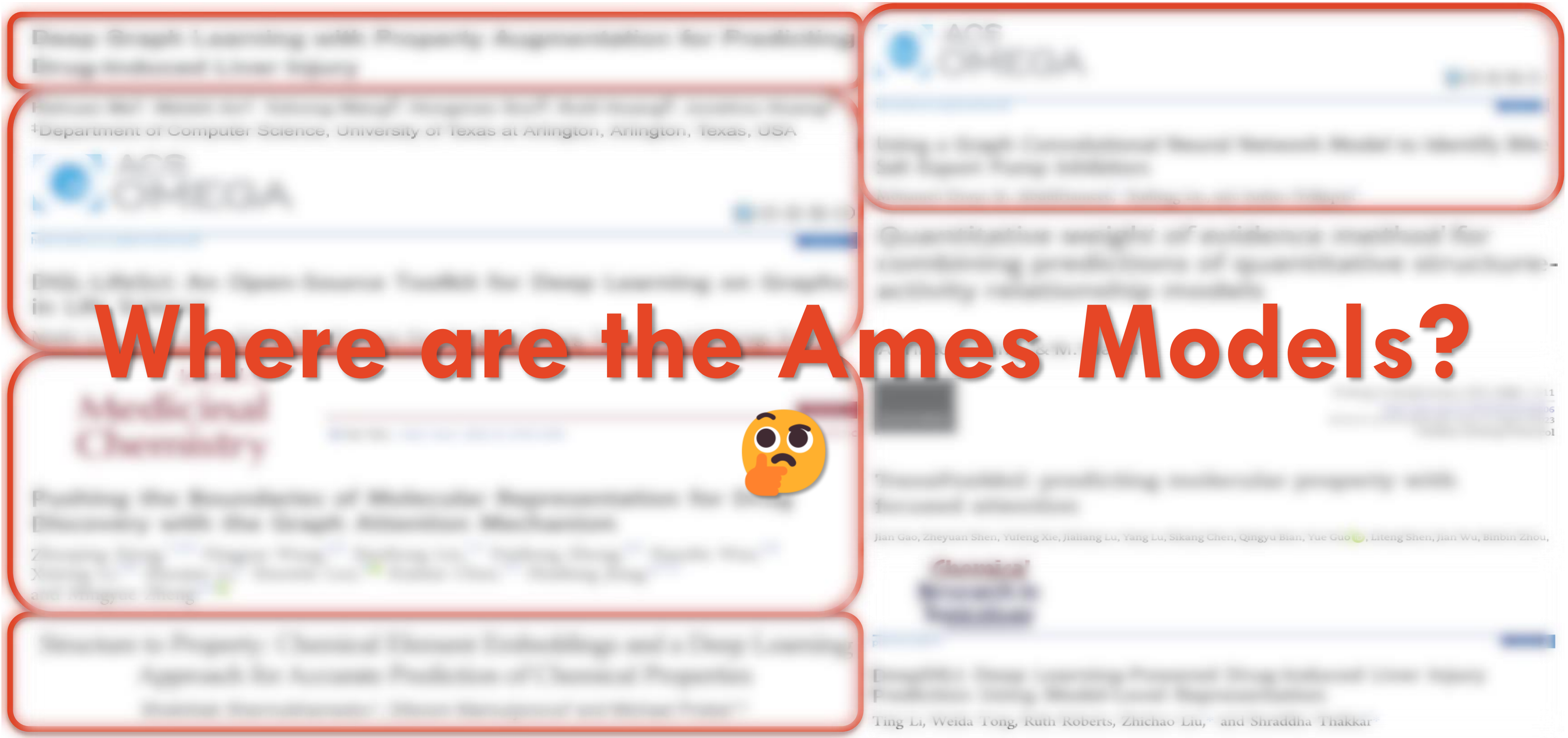


Image: https://www.fda.gov/drugs/regulatory-science-action/new-developments-regulatory-qsar-modeling-new-qsar-model-predicting-blood-brain-barrier-permeability

Explosion in AI Research for Pharmacology / Tox



‡Department of Computer Science, University of Texas at Arlington, Arlington, Texas, USA

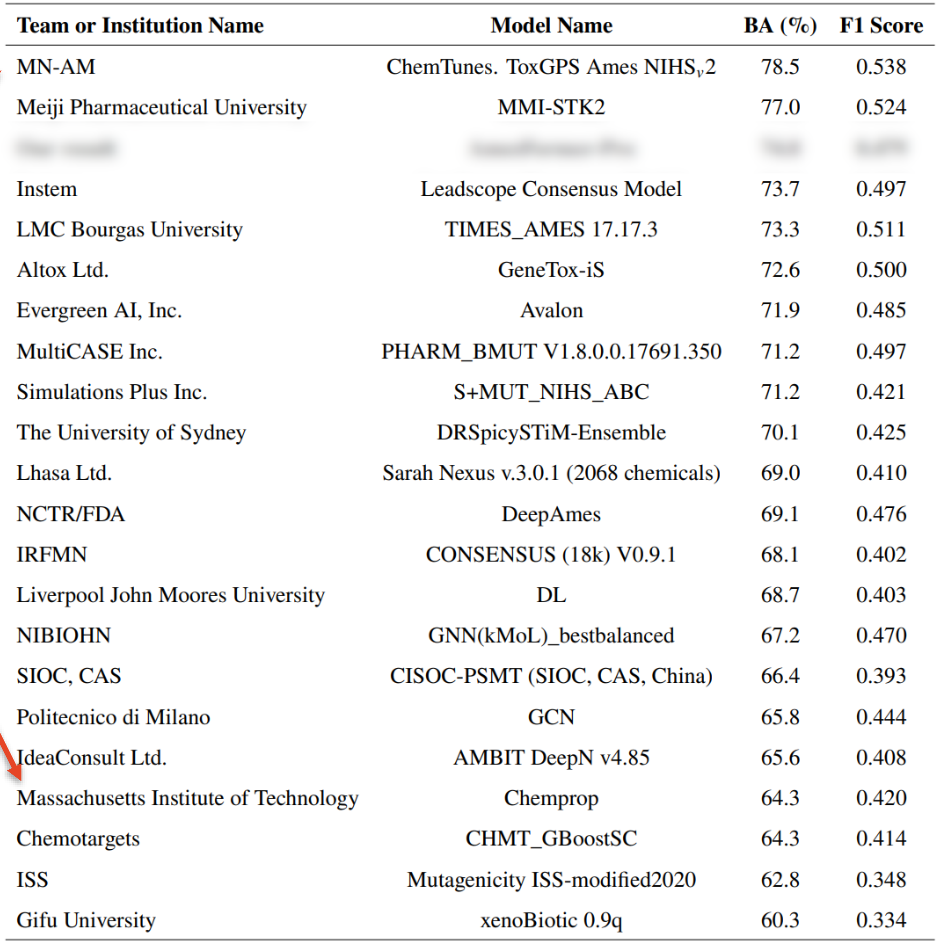
Jian Gao, Zheyuan Shen, Yufeng Xie, Jialiang Lu, Yang Lu, Sikang Chen, Qingyu Bian, Yue Guo, Liteng Shen, Jian Wu, Binbin Zhou,

Ting Li, Weida Tong, Ruth Roberts, Zhichao Liu, and Shraddha Thakkar

Where are the Ames Models?

What Do Existing Models Look Like?

- Big Players
  - MN-AM – US FDA-affiliated
  - MIT – World #1 University
- Old Architectures
  - "Classical machine learning"
- Australia uses TIMES_AMES
  - Costs >$50k / year
- Still not good enough to replace *in vitro* testing



Team or Institution Name	Model Name	BA (%)	F1 Score
MN-AM	ChemTunes. ToxGPS Ames NIHS$_{v}$2	78.5	0.538
Meiji Pharmaceutical University	MMI-STK2	77.0	0.524
[redacted]	[redacted]	[redacted]	[redacted]
Instem	Leadscope Consensus Model	73.7	0.497
LMC Bourgas University	TIMES_AMES 17.17.3	73.3	0.511
Altox Ltd.	GeneTox-iS	72.6	0.500
Evergreen AI, Inc.	Avalon	71.9	0.485
MultiCASE Inc.	PHARM_BMUT V1.8.0.0.17691.350	71.2	0.497
Simulations Plus Inc.	S+MUT_NIHS_ABC	71.2	0.421
The University of Sydney	DRSpicySTiM-Ensemble	70.1	0.425
Lhasa Ltd.	Sarah Nexus v.3.0.1 (2068 chemicals)	69.0	0.410
NCTR/FDA	DeepAmes	69.1	0.476
IRFMN	CONSENSUS (18k) V0.9.1	68.1	0.402
Liverpool John Moores University	DL	68.7	0.403
NIBIOHN	GNN(kMoL)_bestbalanced	67.2	0.470
SIOC, CAS	CISOC-PSMT (SIOC, CAS, China)	66.4	0.393
Politecnico di Milano	GCN	65.8	0.444
IdeaConsult Ltd.	AMBIT DeepN v4.85	65.6	0.408
Massachusetts Institute of Technology	Chemprop	64.3	0.420
Chemotargets	CHMT_GBoostSC	64.3	0.414
ISS	Mutagenicity ISS-modified2020	62.8	0.348
Gifu University	xenoBiotic 0.9q	60.3	0.334

How can we Make the Best Ames Model?

- What models performed best on other biology tasks?
    - Benchmark molecular prediction
    - Multi-endpoint toxicity prediction
- Use state-of-the-art techniques from AI literature
    - Transformers – ChatGPT
    - Graph neural networks – Facebook friend recommendation
    - Special encodings – Extra chemical information
    - Harder math 🤓
- A graph transformer?

Hypotheses

We hypothesise a graph transformer for Ames mutagenicity will:

1. Be the most effective when trained on the largest existing Ames datasets

2. Achieve state-of-the-art predictive performance

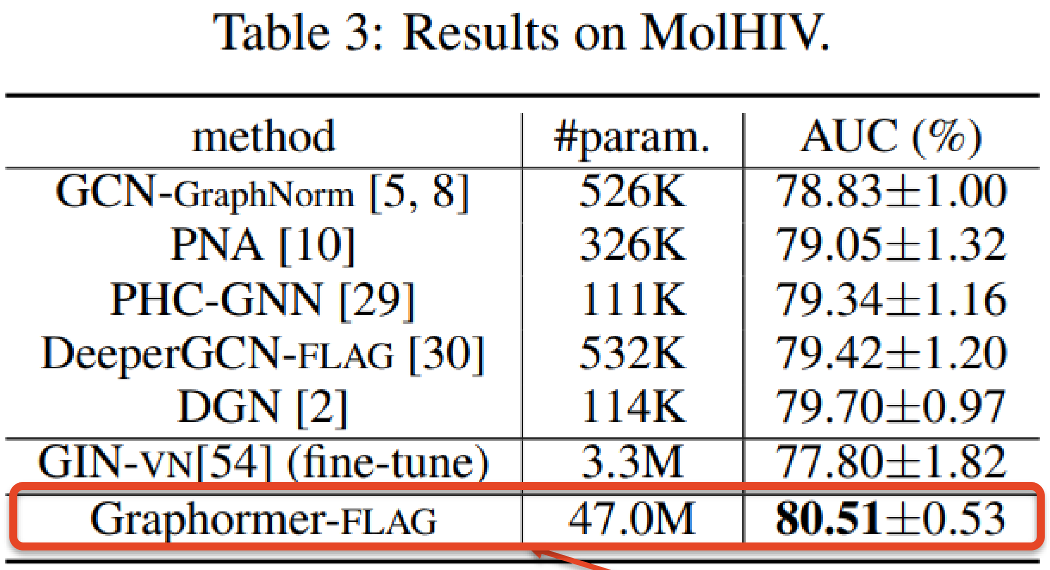
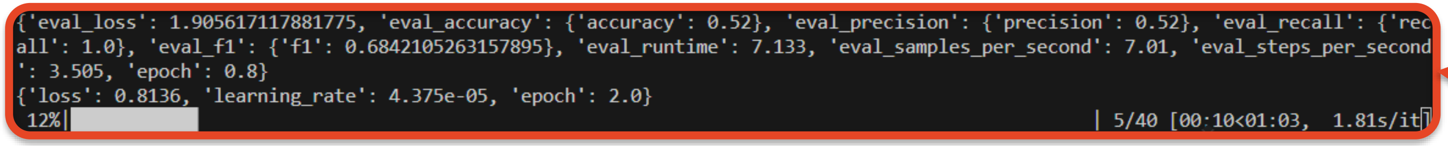


Table 3: Results on MolHIV.

method	#param.	AUC (%)
GCN-GraphNorm [5, 8]	526K	78.83±1.00
PNA [10]	326K	79.05±1.32
PHC-GNN [29]	111K	79.34±1.16
DeeperGCN-FLAG [30]	532K	79.42±1.20
DGN [2]	114K	79.70±0.97
GIN-VN[54] (fine-tune)	3.3M	77.80±1.82
Graphormer-FLAG	47.0M	**80.51**±0.53

Image: http://arxiv.org/pdf/2106.05234.pdf

The basis of our architecture!



```
{'eval_loss': 1.905617117881775, 'eval_accuracy': {'accuracy': 0.52}, 'eval_precision': {'precision': 0.52}, 'eval_recall': {'rec
all': 1.0}, 'eval_f1': {'f1': 0.6842105263157895}, 'eval_runtime': 7.133, 'eval_samples_per_second': 7.01, 'eval_steps_per_second
': 3.505, 'epoch': 0.8}
{'loss': 0.8136, 'learning_rate': 4.375e-05, 'epoch': 2.0}
 12%|                                                                                    | 5/40 [00:10<01:03,  1.81s/it
```

Training
data? 🤔

Bruteing through
pages. 🤓

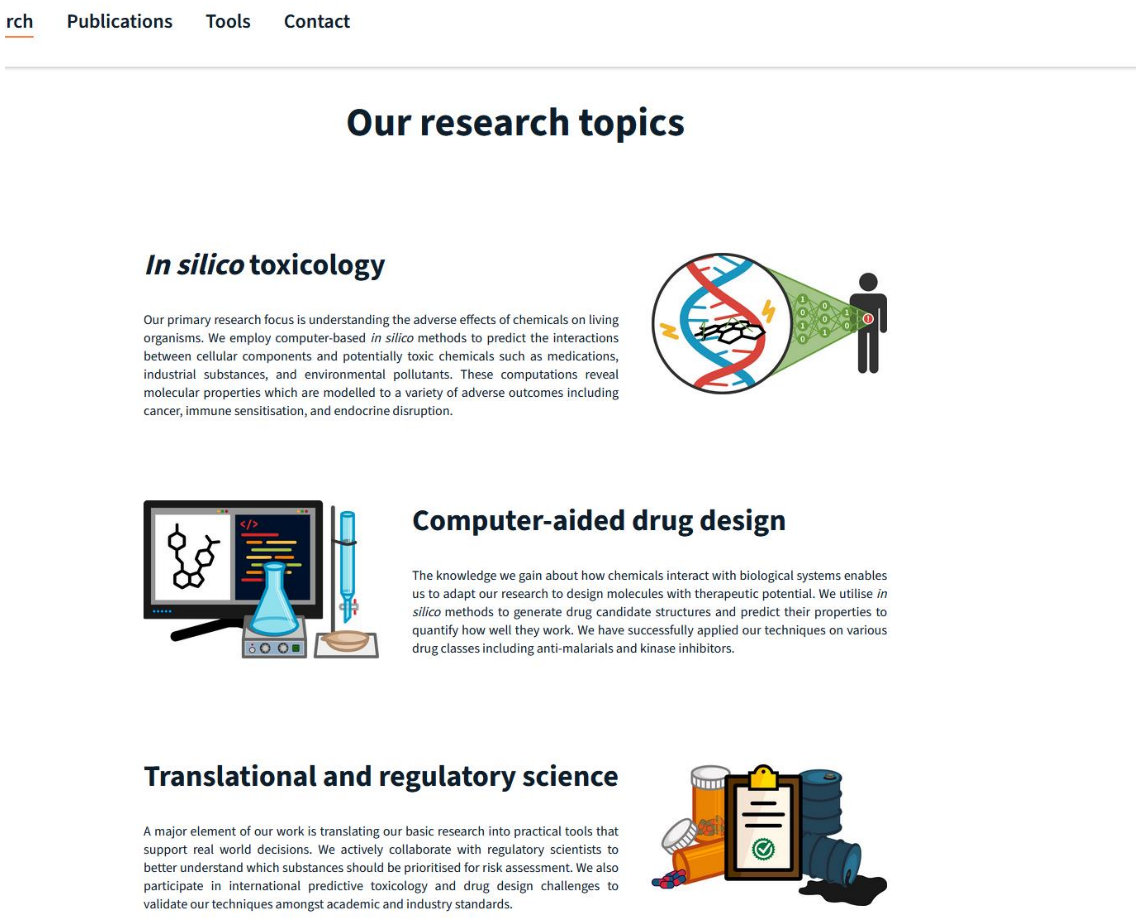
Aims

Hence, we aim to:

- To construct a graph transformer *incorporating our lab's unique domain knowledge*

- To compare the performance of our model with others from the literature

- To deploy this model on our lab website
  - Enabling regulatory, industrial use



rch Publications Tools Contact

Our research topics

***In silico* toxicology**

Our primary research focus is understanding the adverse effects of chemicals on living organisms. We employ computer-based *in silico* methods to predict the interactions between cellular components and potentially toxic chemicals such as medications, industrial substances, and environmental pollutants. These computations reveal molecular properties which are modelled to a variety of adverse outcomes including cancer, immune sensitisation, and endocrine disruption.

Computer-aided drug design

The knowledge we gain about how chemicals interact with biological systems enables us to adapt our research to design molecules with therapeutic potential. We utilise *in silico* methods to generate drug candidate structures and predict their properties to quantify how well they work. We have successfully applied our techniques on various drug classes including anti-malarials and kinase inhibitors.

Translational and regulatory science

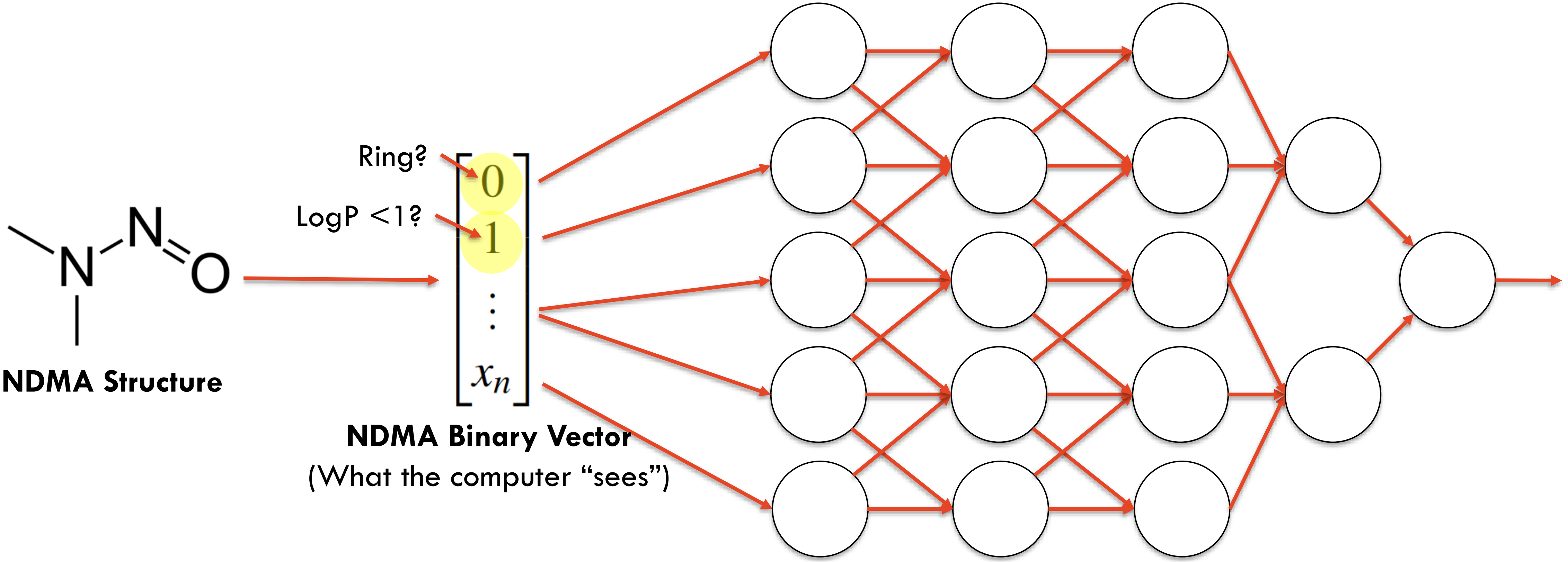
A major element of our work is translating our basic research into practical tools that support real world decisions. We actively collaborate with regulatory scientists to better understand which substances should be prioritised for risk assessment. We also participate in international predictive toxicology and drug design challenges to validate our techniques amongst academic and industry standards.

Our lab website 🧑‍💻

Methods

Understanding Neural Networks

Conventional Neural Networks for Ames Mutagenicity



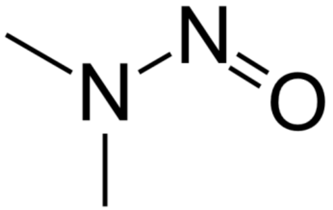
NDMA Structure

NDMA Binary Vector
(What the computer “sees”)

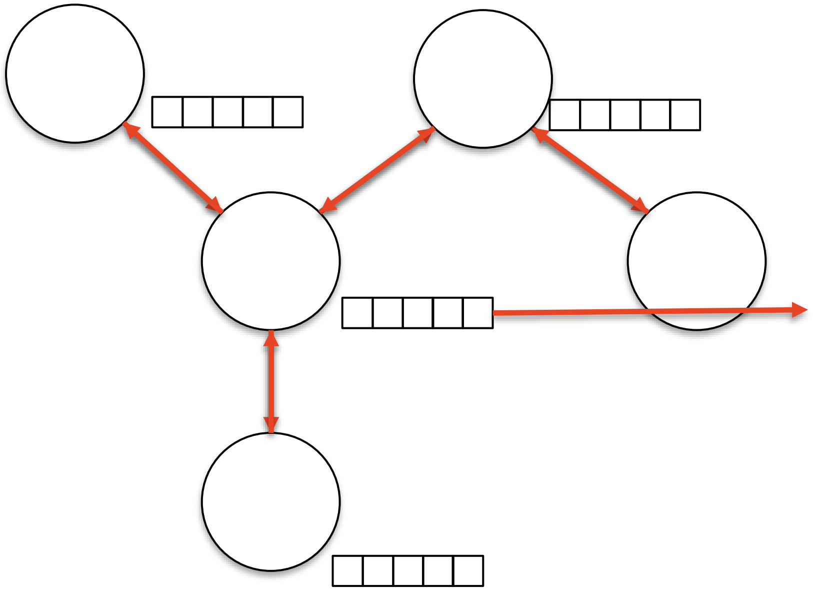
Ames Positivity +/-

Conventional Neural Network
Neuron values update to learn Ames positivity

The **Graph** in Graph Transformers



NDMA Structure

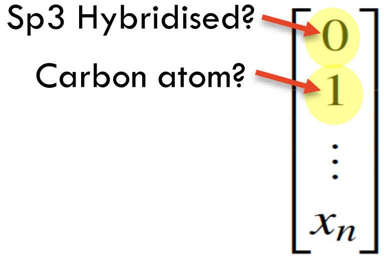


Graph Neural Network
Molecular Structure imbued within
the network structure

Neural Network
Processes
aggregated vectors

Ames Positivity
+/-

Example atom vector



$$\begin{bmatrix} 0 \\ 1 \\ \vdots \\ x_n \end{bmatrix}$$

Methods

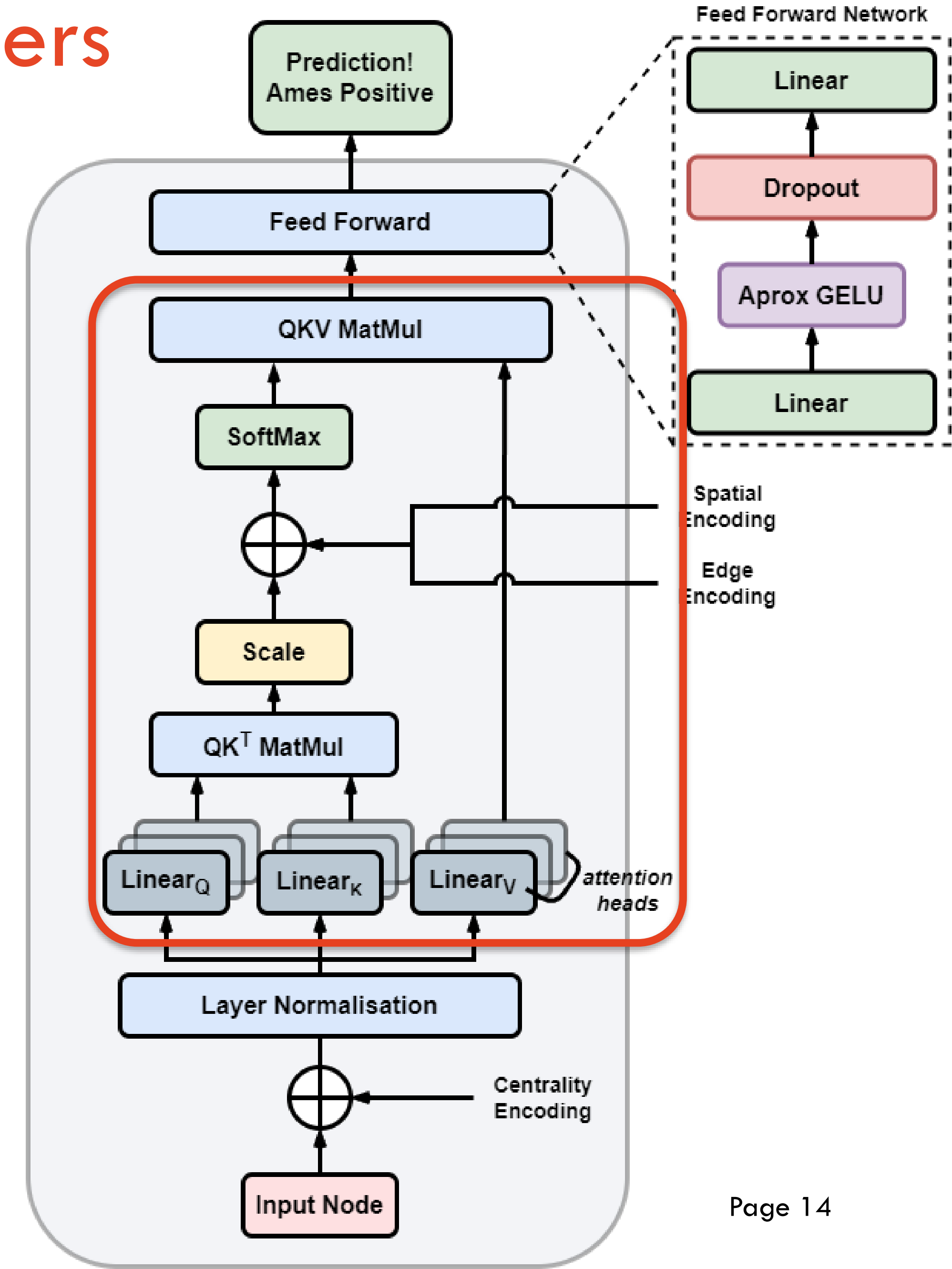
Understanding AmesFormer

The **Transformer** in Graph Transformers

Attention

- Prioritise the most important atomic features
    - Is chirality probably more important than conjugation?
- Allow the network to always see its local environment

- Results in much better *learned* molecular representations



The **Transformer** in Graph Transformers

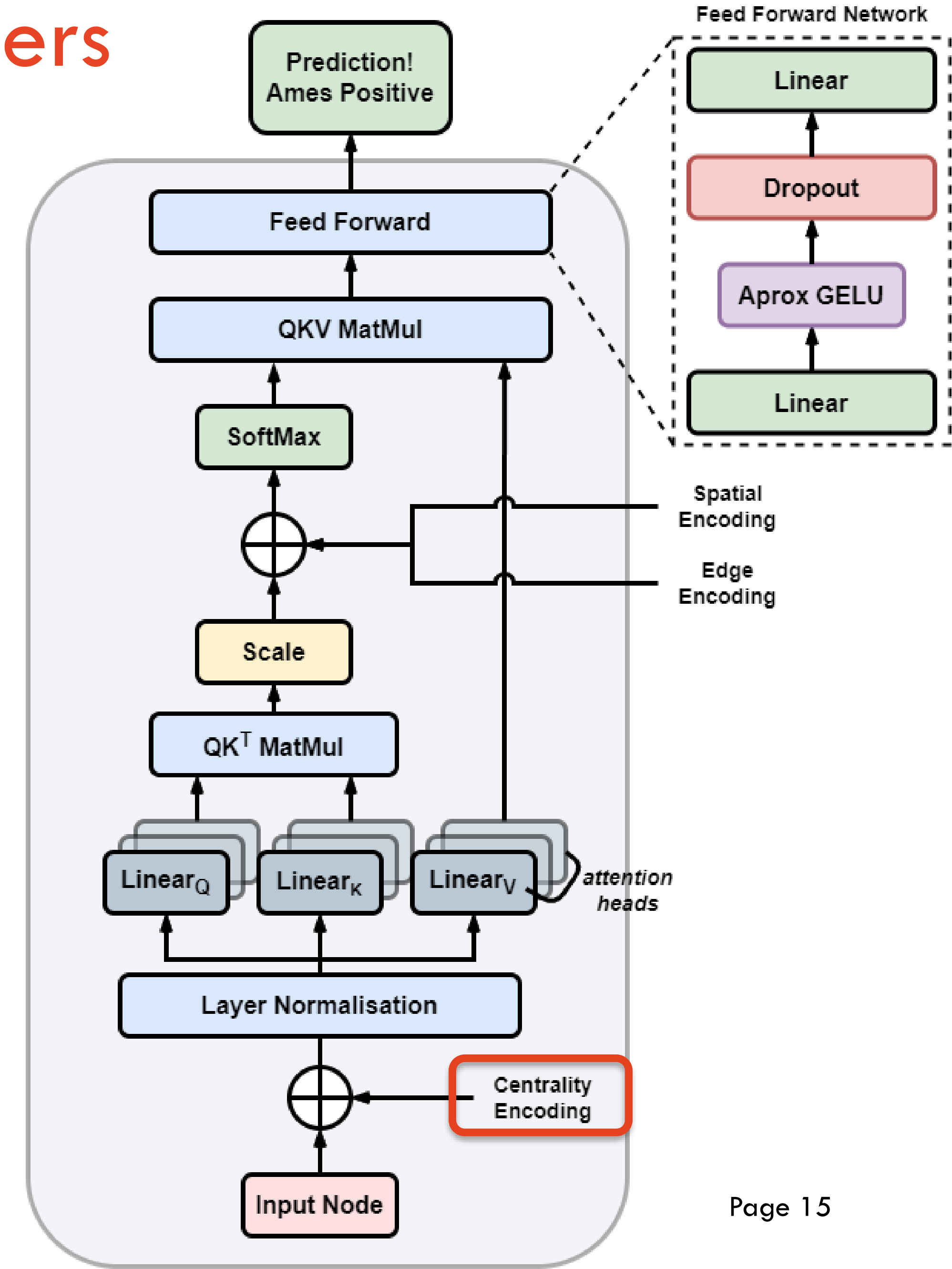
Centrality encoding

- Introduced at the beginning
- Summed with the atom feature vector
- “How many bonds does this atom make?

$$\vec{h}_i = \vec{h}_i + z_{\deg(v_i)}$$

Atom feature vector

Bond count



The **Transformer** in Graph Transformers

Spatial encoding

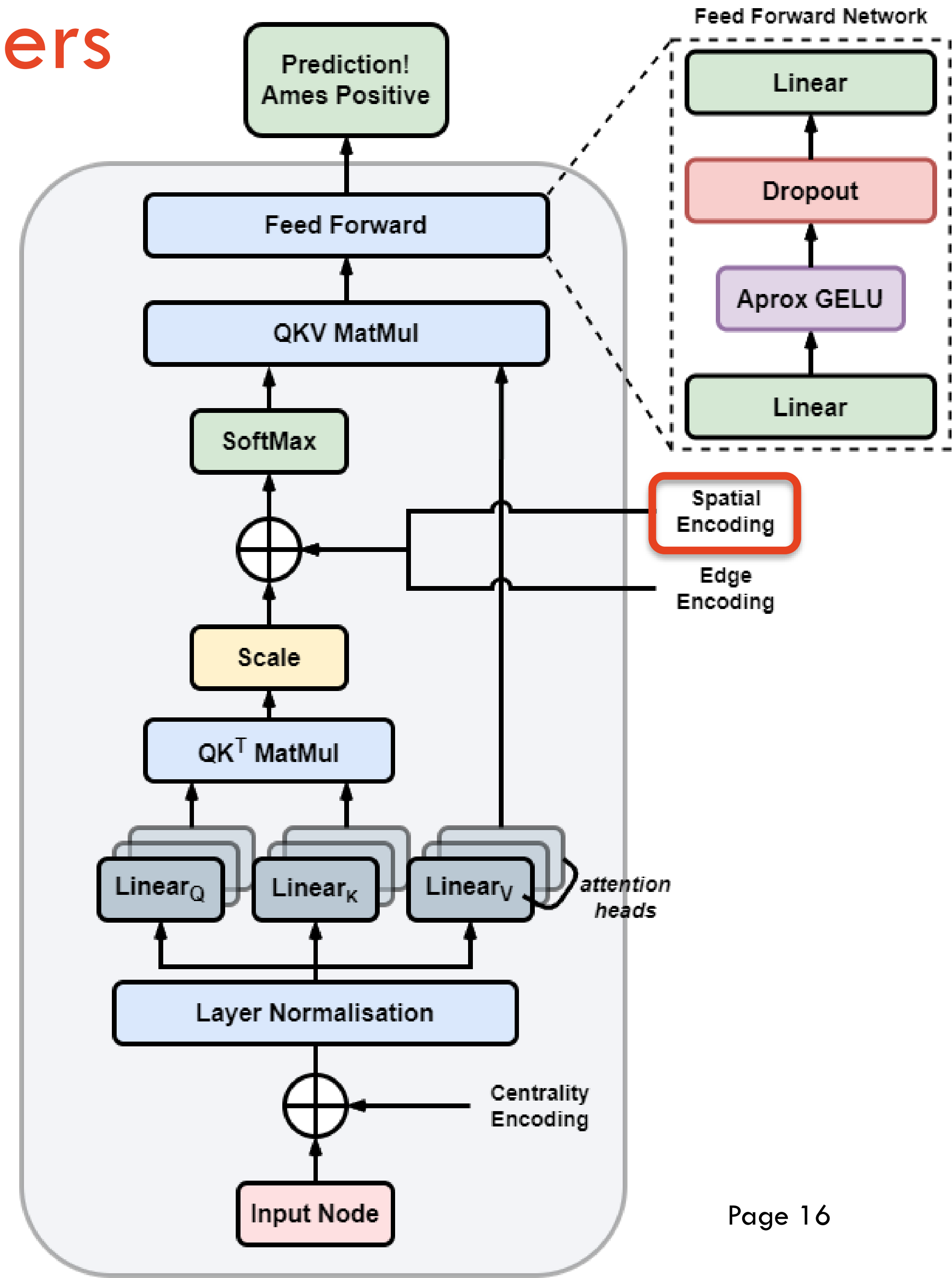
- Biases the attention – The amount each atom feature attends the others
- “How much does every other atom affect me?
- Upshot: Pay less attention to distant atoms, as they likely exert less electrostatic forces

Each atom pair

$$b_\phi(v_i, v_j)$$

Biasing scalar

Shortest path distance



The **Transformer** in Graph Transformers

Edge encoding

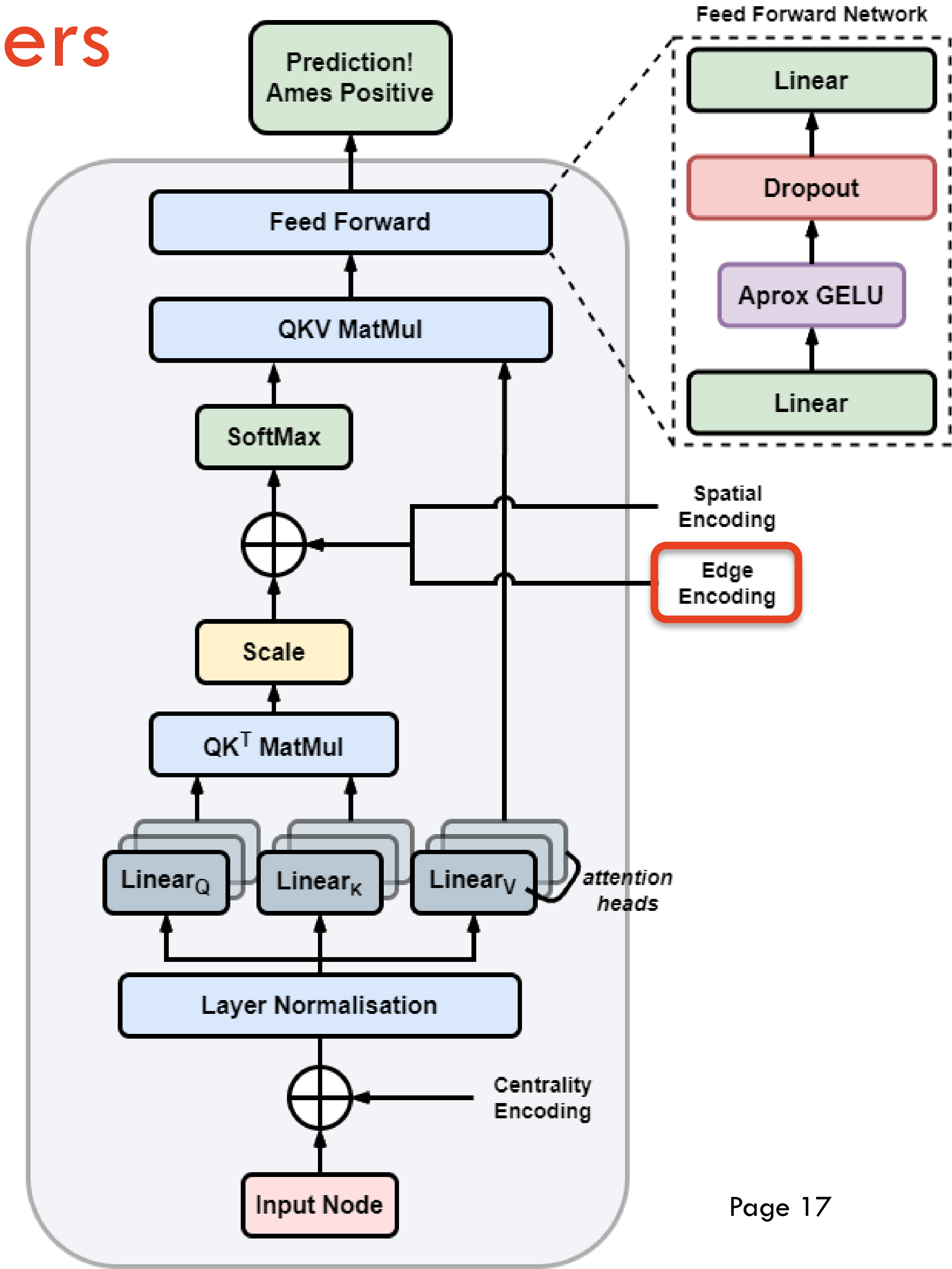
- Biases the attention – "How important are the bonds my neighbours form?"
- Basically, the mean of the dot products of all bond features on each shortest path times a bias

$$\frac{1}{N}\sum_{n=1}^{N}\vec{e}_n \cdot (w_n^E)^T$$

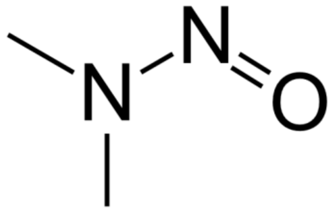
Average them!

Dot product bond features along shortest paths

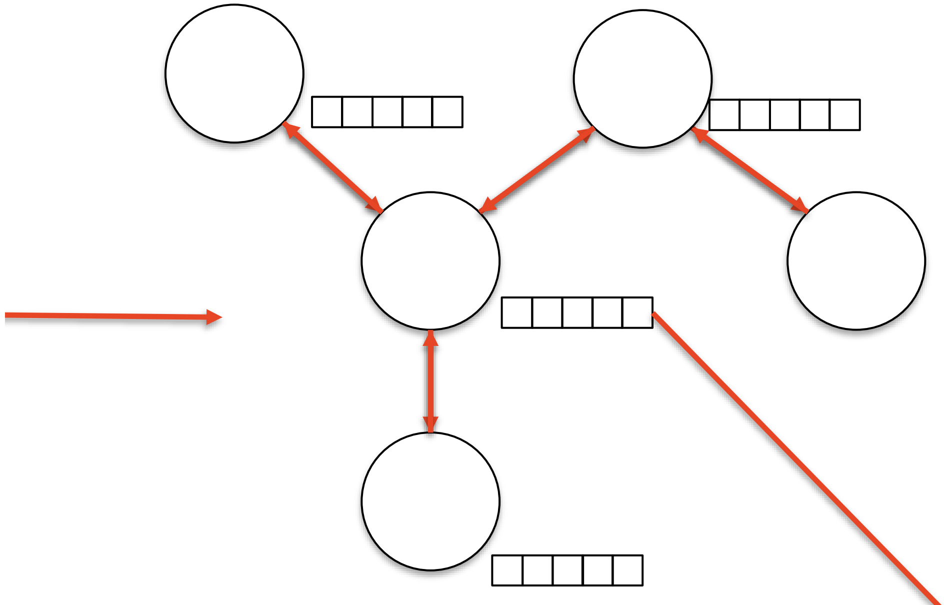
Biasing matrix



The Final Architecture of AmesFormer

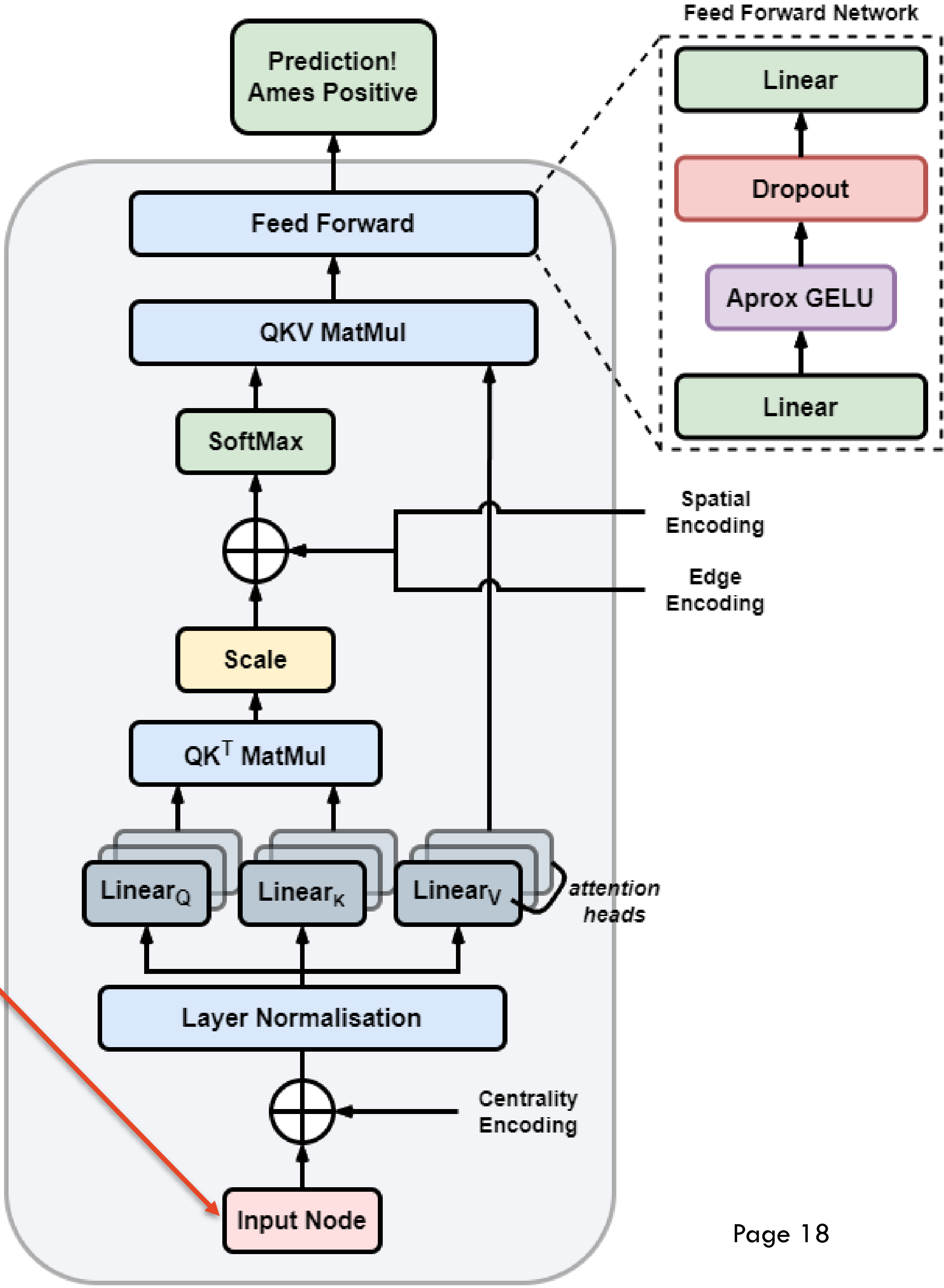


NDMA Structure



Graph Neural Network

Molecular Structure imbued within the network structure

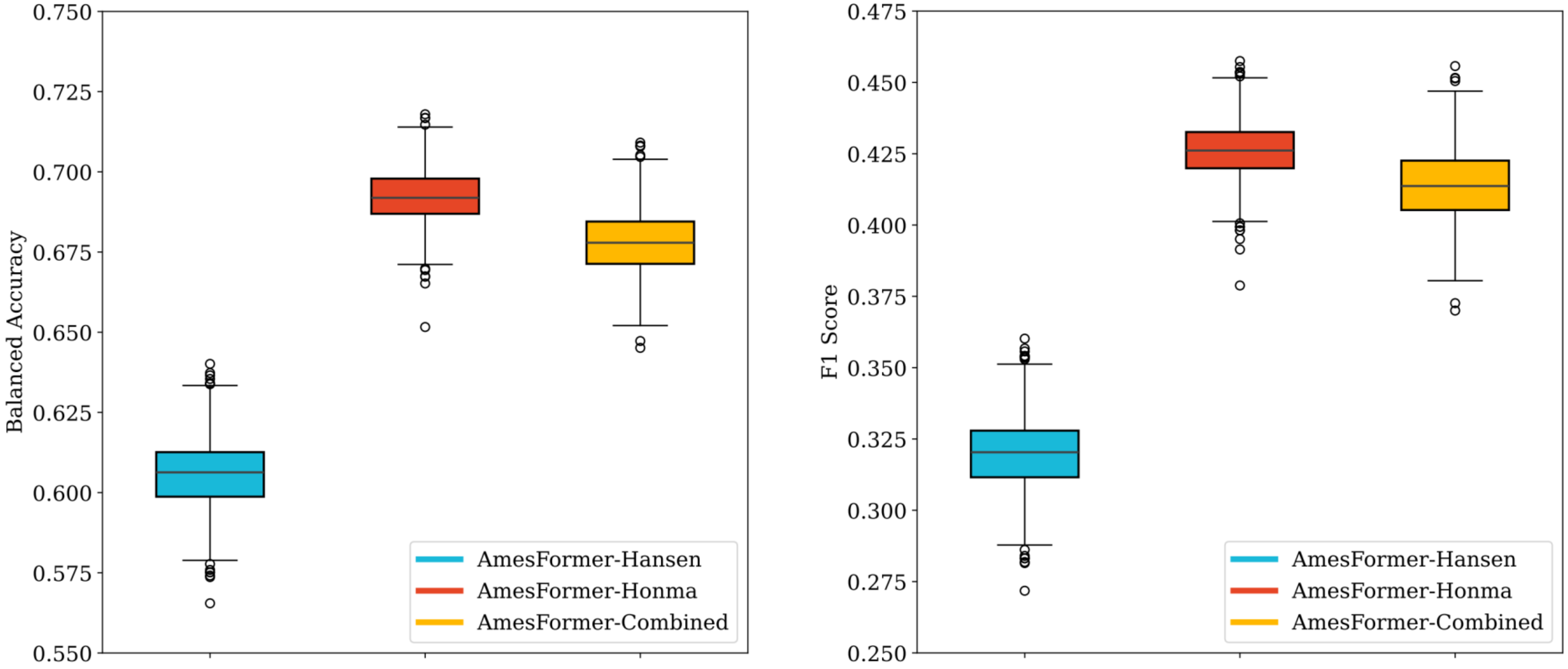


Results

Hypothesis 1 – Is more data always better?

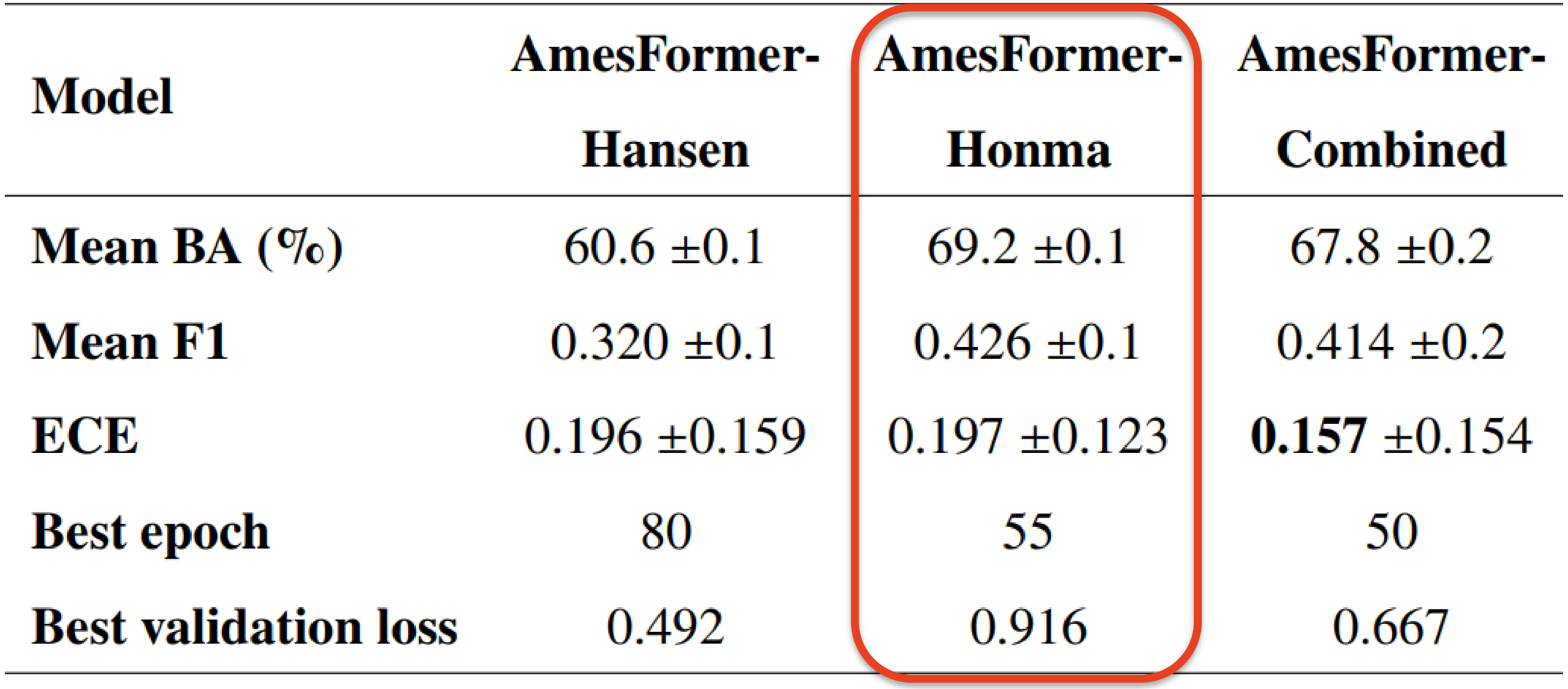
Testing Our Hypotheses – Is More Data Better?

- We trained three models – One on each Ames dataset
  - Surprisingly, the 2nd largest dataset produced the best performing model



Testing Our Hypotheses – Is More Data Better?

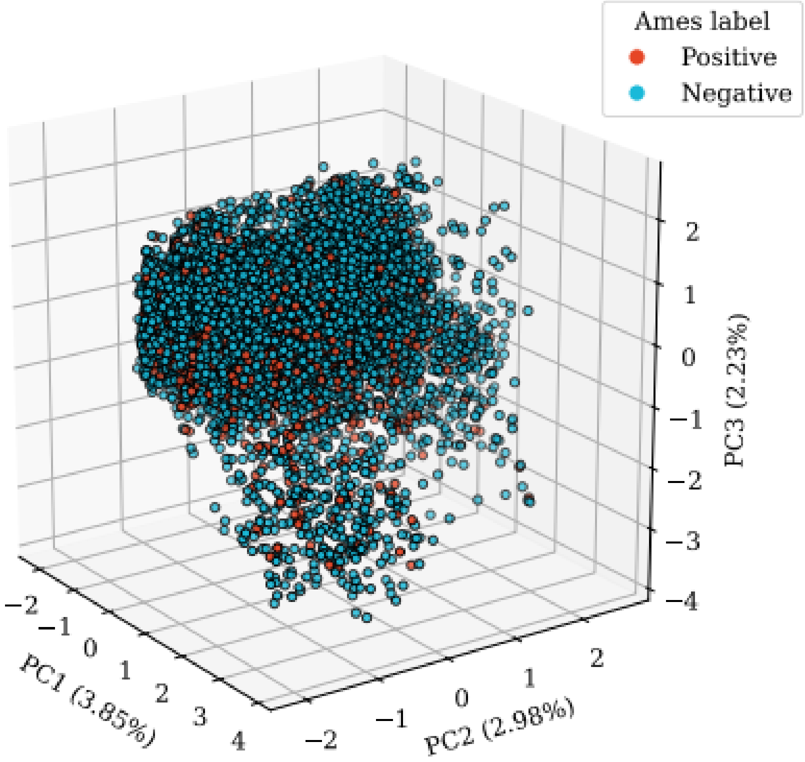
- We trained three models – One on each Ames dataset
  - Surprisingly, the 2nd largest dataset produced the best performing model



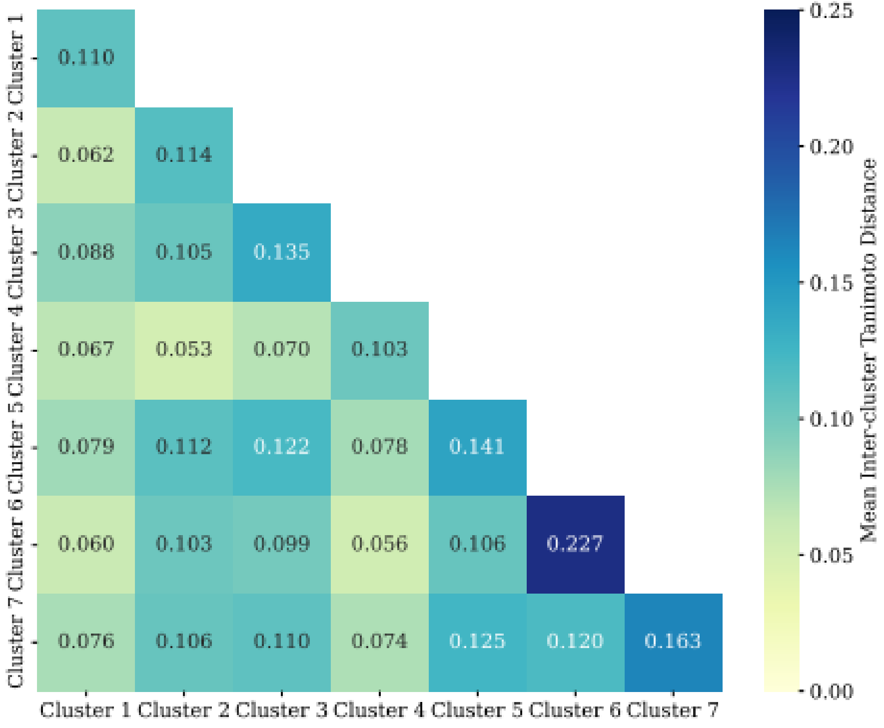
Model	**AmesFormer-Hansen**	**AmesFormer-Honma**	**AmesFormer-Combined**
Mean BA (%)	60.6 ±0.1	69.2 ±0.1	67.8 ±0.2
Mean F1	0.320 ±0.1	0.426 ±0.1	0.414 ±0.2
ECE	0.196 ±0.159	0.197 ±0.123	**0.157** ±0.154
Best epoch	80	55	50
Best validation loss	0.492	0.916	0.667

Understanding Our Results – Why isn't More Data Better?

- The best dataset showed the most chemical diversity – Silhouette Score of 0.488
  - Others had silhouettes of 0.378 and 0.384
  - I.e. It covered the broadest range of molecular structures



(c) Honma dataset PCA.



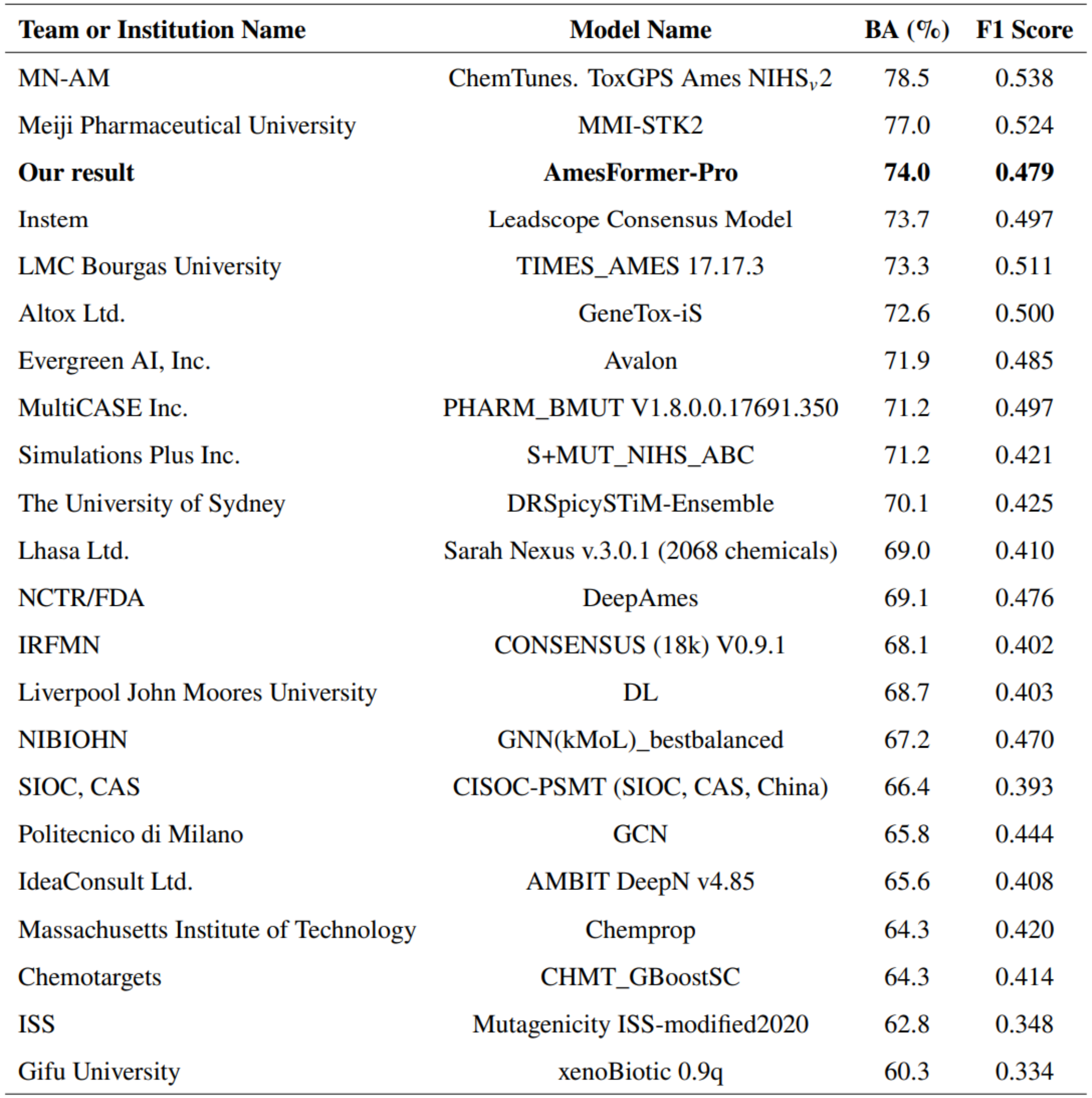
(d) UMAP clusters of the Honma dataset.

Results

Hypothesis 2 – Is Our Model State-of-the-Art?

Testing Our Hypotheses – Is Our Model State-of-the-Art?

- Our model is the third best predictor of Ames mutagenicity
- We beat several established teams & companies
- Significant improvement (3.9%) over previous lab result



Team or Institution Name	Model Name	BA (%)	F1 Score
MN-AM	ChemTunes. ToxGPS Ames NIHS$_v$2	78.5	0.538
Meiji Pharmaceutical University	MMI-STK2	77.0	0.524
Our result	**AmesFormer-Pro**	**74.0**	**0.479**
Instem	Leadscope Consensus Model	73.7	0.497
LMC Bourgas University	TIMES_AMES 17.17.3	73.3	0.511
Altox Ltd.	GeneTox-iS	72.6	0.500
Evergreen AI, Inc.	Avalon	71.9	0.485
MultiCASE Inc.	PHARM_BMUT V1.8.0.0.17691.350	71.2	0.497
Simulations Plus Inc.	S+MUT_NIHS_ABC	71.2	0.421
The University of Sydney	DRSpicySTiM-Ensemble	70.1	0.425
Lhasa Ltd.	Sarah Nexus v.3.0.1 (2068 chemicals)	69.0	0.410
NCTR/FDA	DeepAmes	69.1	0.476
IRFMN	CONSENSUS (18k) V0.9.1	68.1	0.402
Liverpool John Moores University	DL	68.7	0.403
NIBIOHN	GNN(kMoL)_bestbalanced	67.2	0.470
SIOC, CAS	CISOC-PSMT (SIOC, CAS, China)	66.4	0.393
Politecnico di Milano	GCN	65.8	0.444
IdeaConsult Ltd.	AMBIT DeepN v4.85	65.6	0.408
Massachusetts Institute of Technology	Chemprop	64.3	0.420
Chemotargets	CHMT_GBoostSC	64.3	0.414
ISS	Mutagenicity ISS-modified2020	62.8	0.348
Gifu University	xenoBiotic 0.9q	60.3	0.334

Understanding Our Results – Why is AmesFormer so Good?

- Representational Power
  - **We can always tell different molecules apart**
  - Earlier models use those “bit vectors”, these are **condensed** representations of the molecule
  - Hence, similar, but pharmacologically distinct molecules can produce the same vector, and thus same prediction, despite differing toxicity
  - This is known as *bit clashing – Causes activity cliffs*

Why doesn’t AmesFormer suffer the same problem?

Understanding Our Results – Why is AmesFormer so Good?

1. Representational Power via the W-L Test
   - **We avoid this problem using our spatial encoding**
   - The spatial encoding is equivalent to the shortest-path-enhanced Weisfeiler-Lehmen graph isomorphism test
   - An inductive proof is available in Chengxuan, et al. 2021

A.1 SPD can Be Used to Improve WL-Test

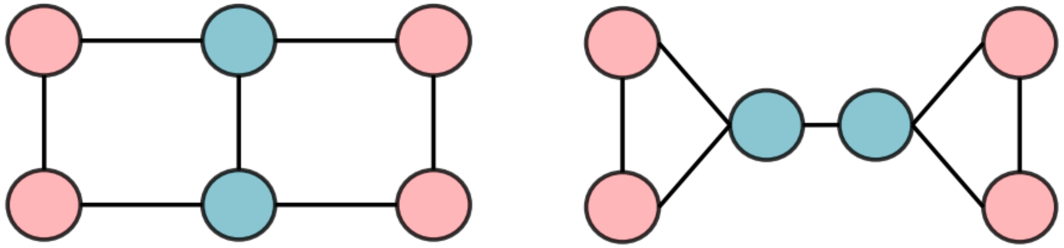


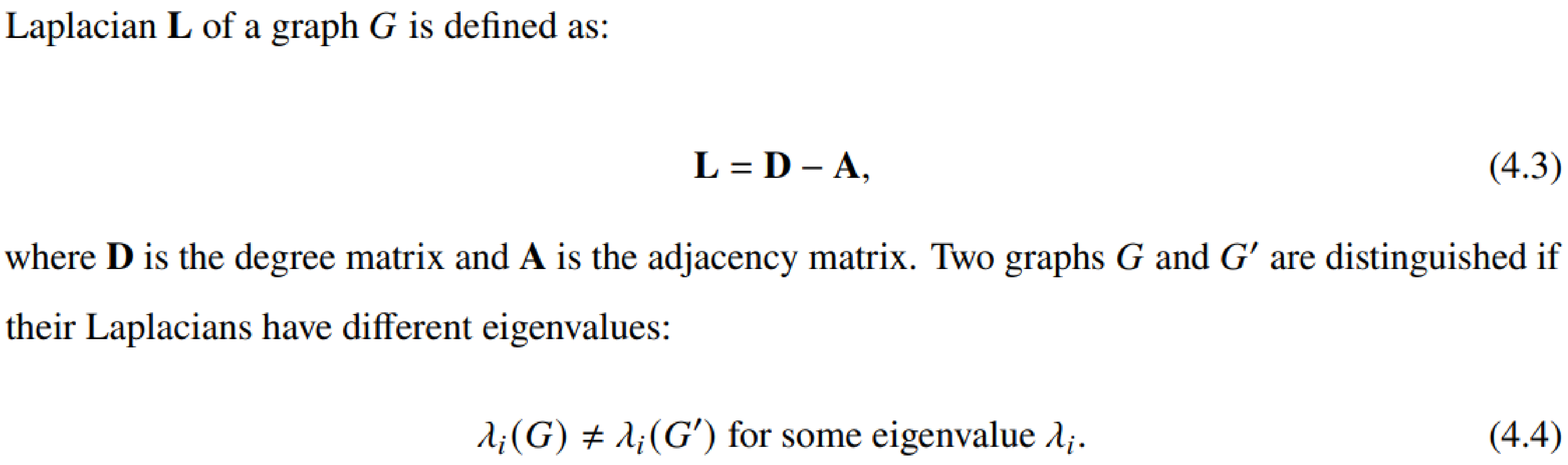
Figure 2: These two graphs cannot be distinguished by 1-WL-test. But the SPD sets, i.e., the SPD from each node to others, are different: The two types of nodes in the left graph have SPD sets $\{0,1,1,2,2,3\}$, $\{0,1,1,1,2,2\}$ while the nodes in the right graph have SPD sets $\{0,1,1,2,3,3\}$, $\{0,1,1,1,2,2\}$.

Image: https://arxiv.org/pdf/2106.05234

Understanding Our Results – Why is AmesFormer Good?

2. Representational Power via the Graph Laplacian

- **Our GNN can differentiate any two graphs which differ in the spectral properties of their graph Laplacian**
- A constructive proof is shown in Kanatsoulis & Ribeiro, 2023



Laplacian $\mathbf{L}$ of a graph G is defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{4.3}$$

where $\mathbf{D}$ is the degree matrix and $\mathbf{A}$ is the adjacency matrix. Two graphs G and G' are distinguished if their Laplacians have different eigenvalues:

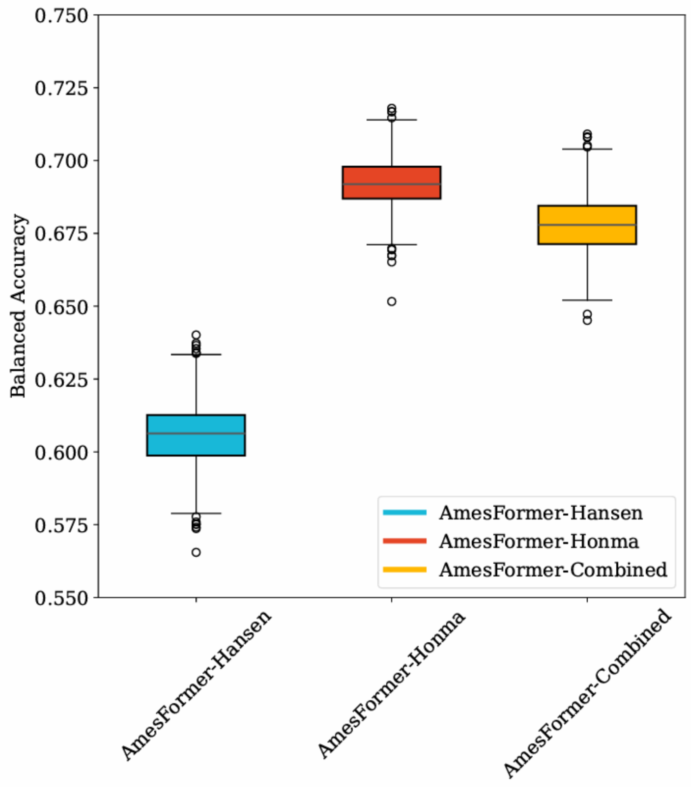
$$\lambda_i(G) \neq \lambda_i(G') \text{ for some eigenvalue } \lambda_i. \tag{4.4}$$

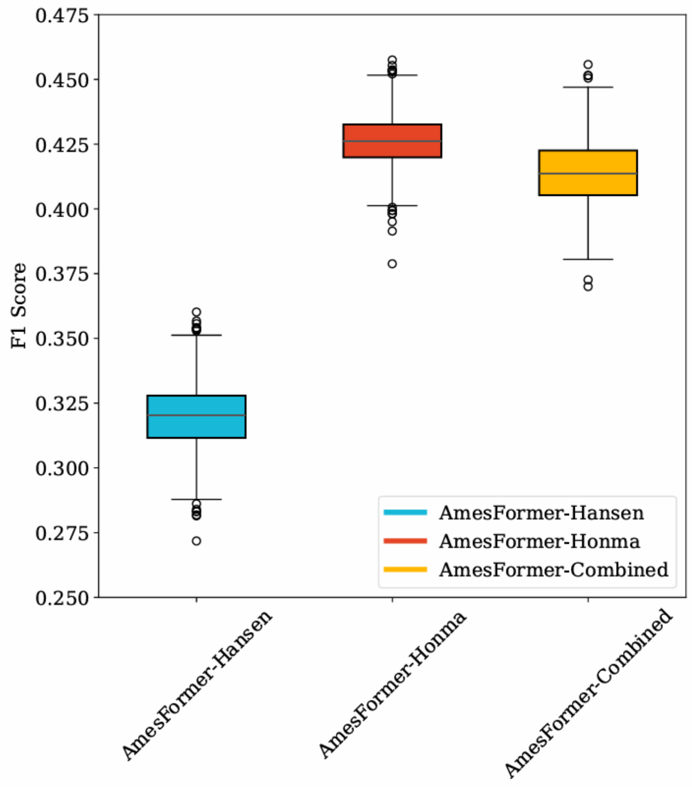
Certainty

How do we Know These Results are Accurate?

Bayesian Uncertainty Estimation via MC Dropout

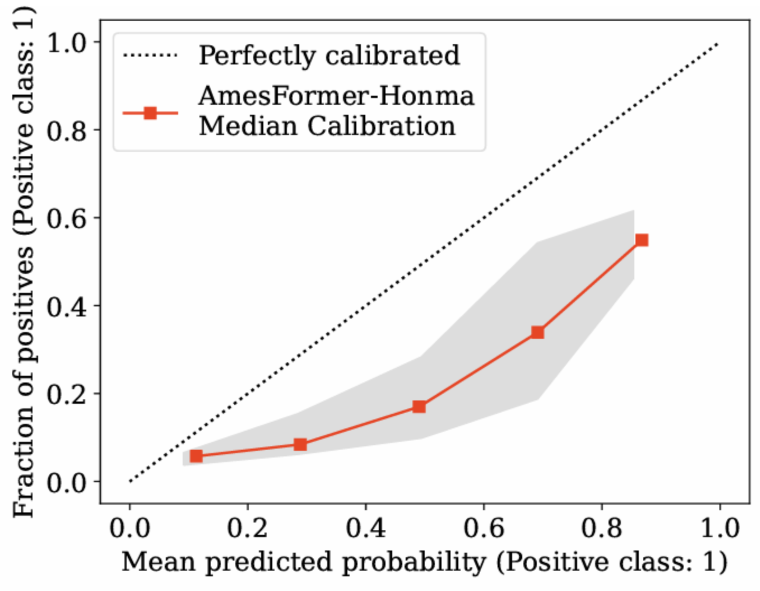
- We use Monte Carlo (MC) dropout to generate CIs for our results – BAC





Bayesian Uncertainty Estimation via MC Dropout

- We use Monte Carlo (MC) dropout to generate CIs for our results – F1

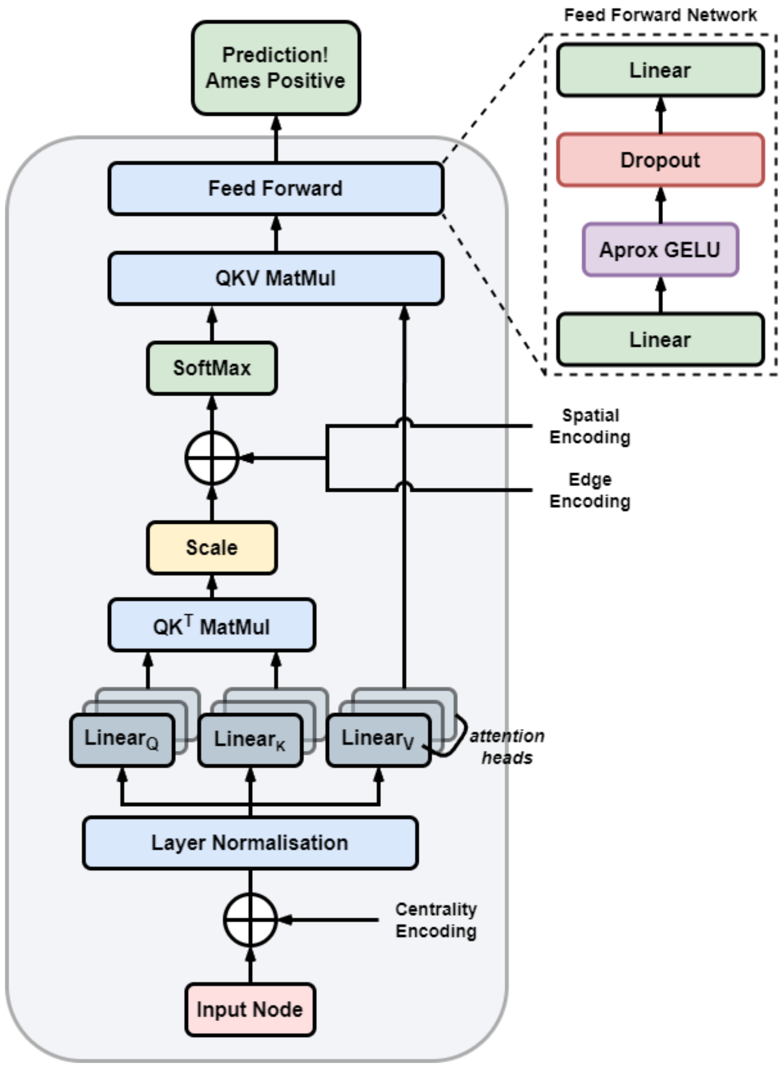


(b) The median calibration curve of AmesFormer-Honma over 1000 Monte Carlo dropout samples with an associated ECE of 0.197 (95% CI: 0.087, 0.333).

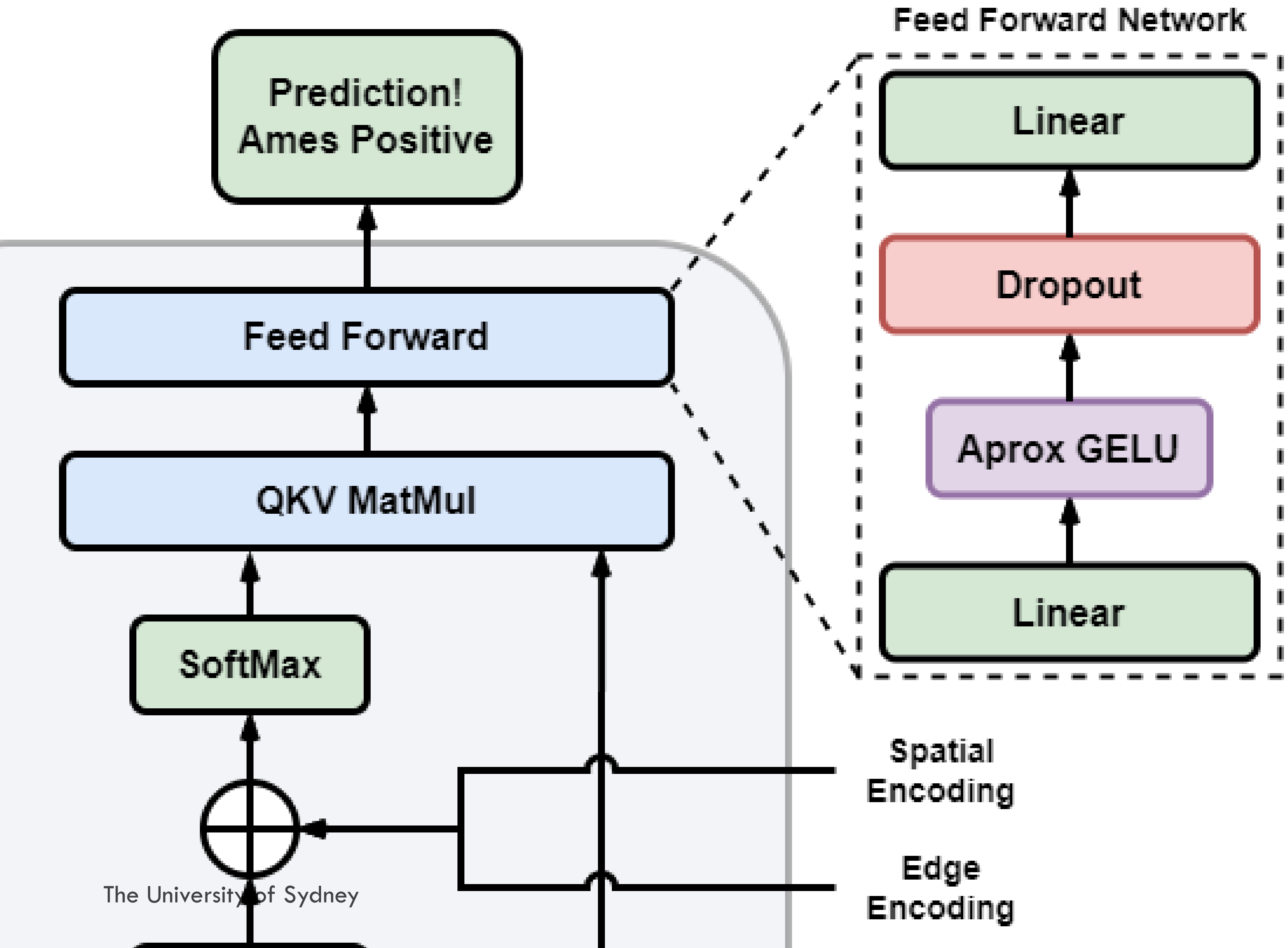
Bayesian Uncertainty Estimation via MC Dropout

- But…
  - We can extend this methodology to the regulatory context by sampling the uncertainty of our inference (I.e., when we are using the model live)
  - Over 1000 passes we are integrating under the distribution of predictions to gauge our uncertainty
  - **We can therefore sample our uncertainty for the prediction of that *particular* chemical**
  - Recommended by the OECD QSAR Reporting Guideline

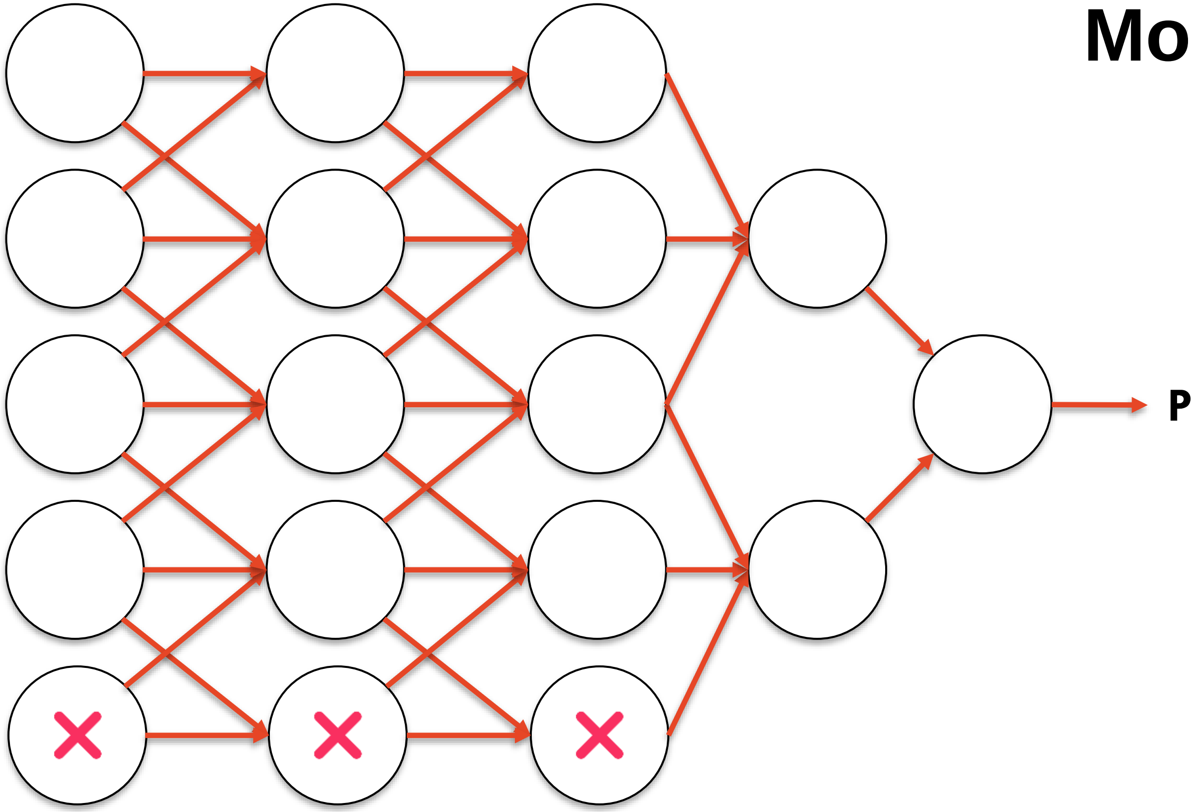
Bayesian Uncertainty Estimation via MC Dropout



Bayesian Uncertainty Estimation via MC Dropout



Bayesian Uncertainty Estimation via MC Dropout



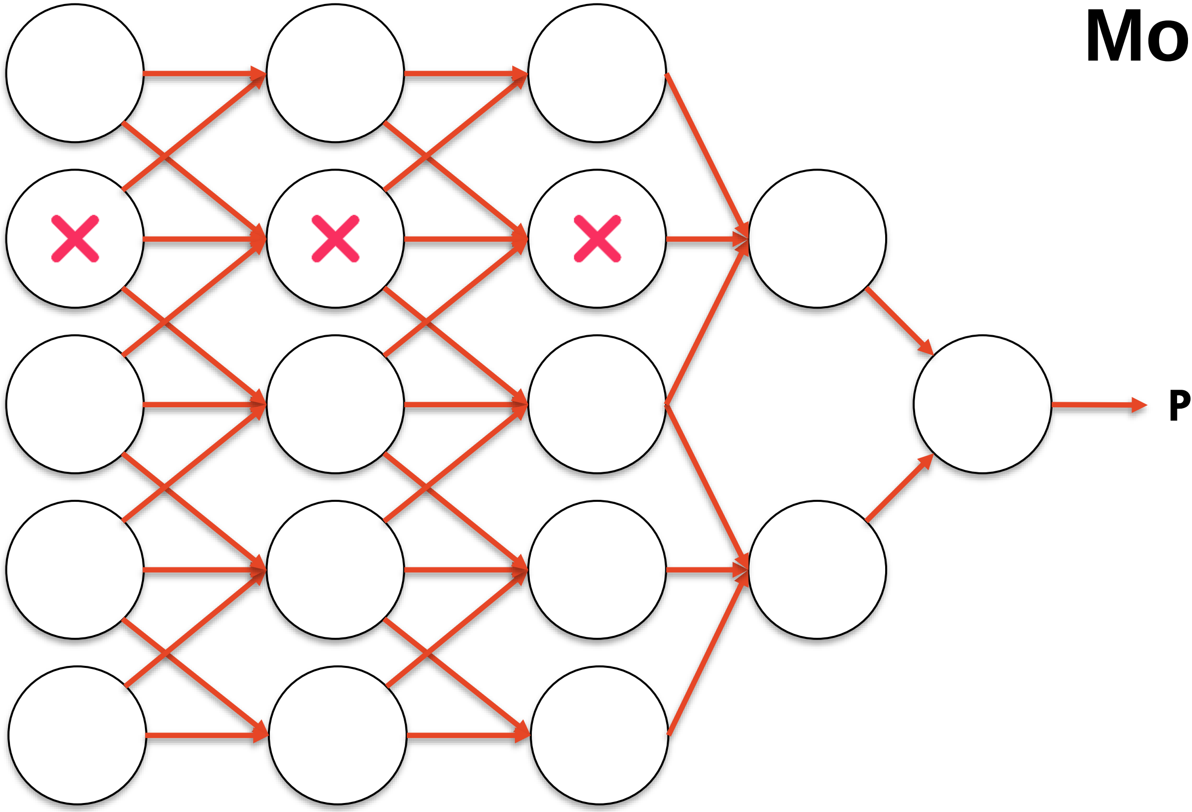
Molecule Pass 1

Ames Positivity +/-

Feed Forward Network

The FFN in the Transformer Diagram

Bayesian Uncertainty Estimation via MC Dropout

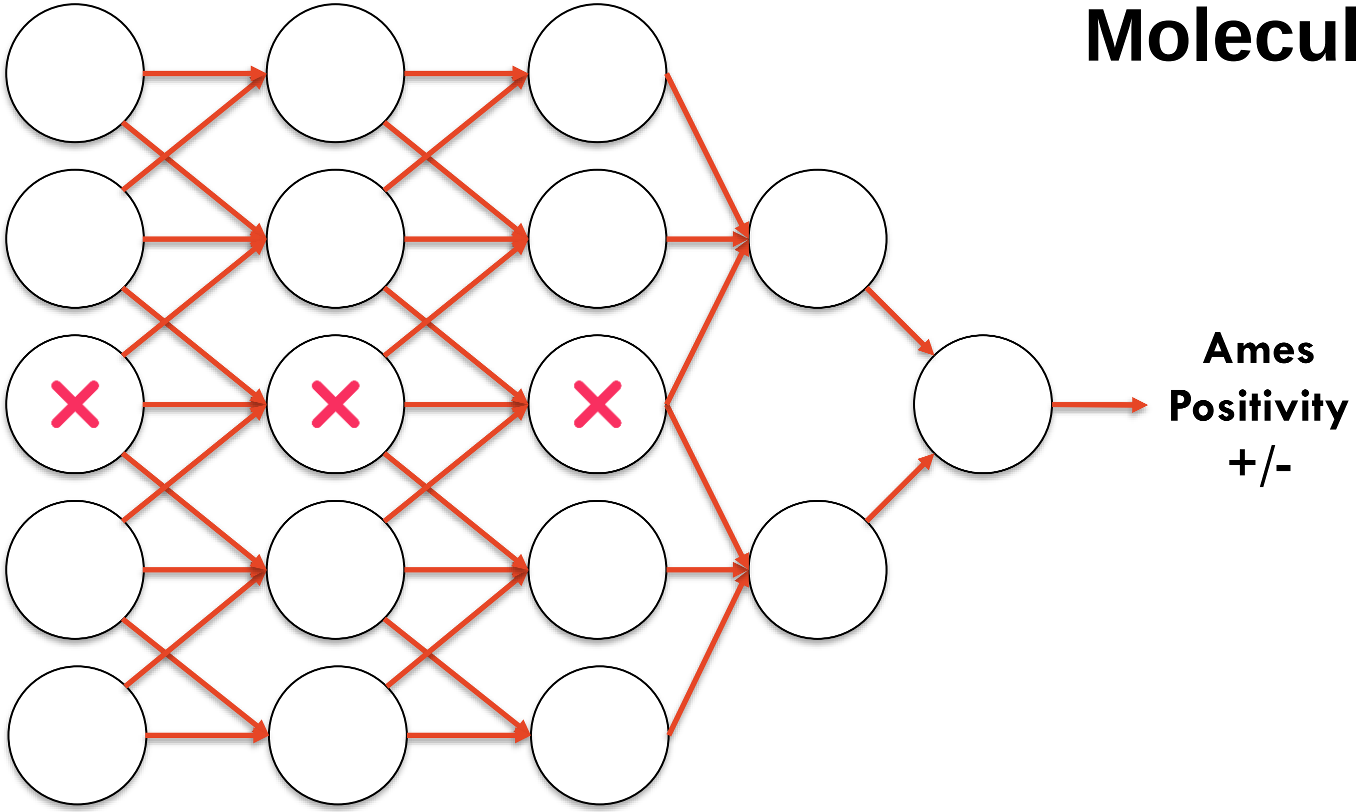


Molecule Pass 2

Feed Forward Network

The FFN in the Transformer Diagram

Bayesian Uncertainty Estimation via MC Dropout



Molecule Pass 3

Feed Forward Network
The FFN in the Transformer Diagram

Future Directions

One Hard Thing That Sounds Easy

Future Directions – Taking the Number 1 Spot

- Our performance is very good, but two models are better – Why?
- Both better models are “ensembles”
  - Combinations of multiple different models – Logistic regression, simpler graphs, etc
- These models **can see** *whole graph* properties – Solubility, etc
- AmesFormer **cannot see** these properties, it only sees the more detailed atom and bond infor

How can we incorporate these *whole molecule* properties into AmesFormer?

Future Directions – Taking the Number 1 Spot

It's tough…

Node-wise Approach

- Add whole-graph data to each atom
- Pros
  - Done in literature (GraphGPS)
  - Trivial to implement
- Cons
  - Massive data duplication – There's only one set of graph properties, but we add them to every node
  - Computationally inefficient

Attentional Approach

- Add whole-graph data to the graph attention calculation
- Pros
  - No duplication – Improved efficiency
- Cons
  - Unproven
  - Hard to implement
  - Network can't "see" whole-graph data before attention, less opportunities to incorporate it into the molecular representation

Future Directions

One Easy(ish) Thing That Sounds Hard

Future Directions – Improving Accessibility

- Our models are relatively efficient, but still required days to train on a $US 2000 graphics card
  - For more complex tasks, like general mutagenicity, this would be longer
- This is **out of reach** for many small academic labs & startups

How can we make our model more computationally efficient and accessible to compute-poor users?

Future Directions – Improving Accessibility

Improve **attention**

The most computationally expensive part of AmesFormer

- Currently, we do multiple attention calculations in parallel
    - Each attention *head* learns different things to “attend” – Great performance!
    - But do all heads actually learn to attend something valuable?
    - **No** – So, can we:
        - Remove useless heads, retain the good ones?
        - Maintain the same performance whilst improving computational efficiency?

We can use GFiSH-Former by Tan, et al. 2022 to accompish this

Future Directions – Improving Accessibility

1. Eigenvalue decomposition – Attention covariance matrices are low-rank 🤓
    - I.e., Most of the information in them is useless, we only need the most important 10%
2. Calculate ~3 heads – This should be enough to capture ~90% of variance
    - Way less than the 32 currently calculated for AmesFormer
3. Calculate the remaining 29 as a *finite admixture* of those 3

Future Directions – Improving Accessibility

The head we're calling

E.g., head 4

$$\boxed{\mathbf{A}_j} = \sum_{k=1}^{M} \phi(p_{kj}(\mathbf{Q}_k {\mathbf{K}_k}^\top + \sigma_k \odot \epsilon_j)), \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}),$$

Image: https://openreview.net/pdf?id=0VFQhPGF1M3

Future Directions – Improving Accessibility

Is a mixture of our 3
main heads M

$$\mathbf{A}_j = \boxed{\sum_{k=1}^{M}} \phi(p_{kj}(\mathbf{Q}_k \mathbf{K}_k{}^\top + \sigma_k \odot \epsilon_j)), \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}),$$

Image: https://openreview.net/pdf?id=0VFQhPGF1M3

Future Directions – Improving Accessibility

With a non-linear transformation

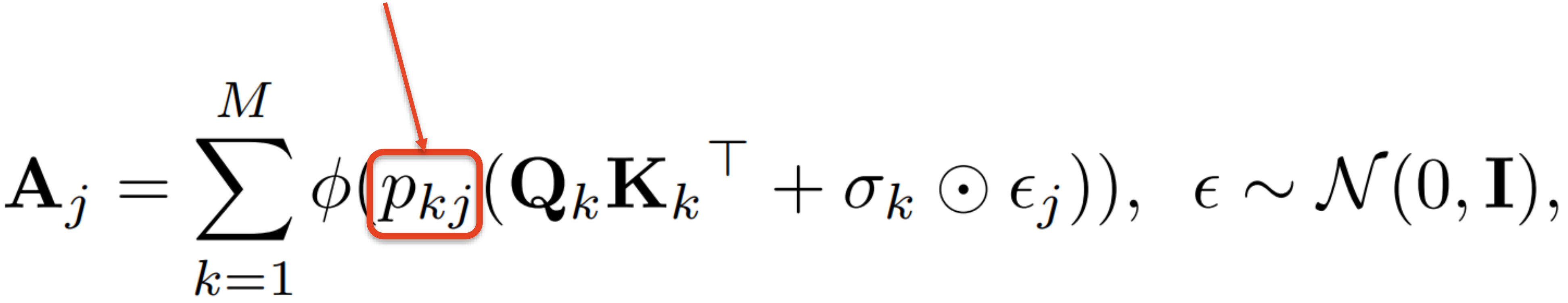
E.g., Gaussian

$$\mathbf{A}_j = \sum_{k=1}^{M} \boxed{\phi}(p_{kj}(\mathbf{Q}_k \mathbf{K}_k{}^{\top} + \sigma_k \odot \epsilon_j)), \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}),$$

Image: https://openreview.net/pdf?id=0VFQhPGF1M3

Future Directions – Improving Accessibility

Weighted by a parameter determing much each of the 3 main heads should contribute



$$\mathbf{A}_j = \sum_{k=1}^{M} \phi(\boxed{p_{kj}}(\mathbf{Q}_k \mathbf{K}_k{}^\top + \sigma_k \odot \epsilon_j)), \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}),$$

Image: https://openreview.net/pdf?id=0VFQhPGF1M3

Future Directions – Improving Accessibility

Where this is the actual content of the main head (e.g., head 2)

$$\mathbf{A}_j = \sum_{k=1}^{M} \phi(p_{kj}(\boxed{\mathbf{Q}_k \mathbf{K}_k{}^\top} + \sigma_k \odot \epsilon_j)), \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}),$$

Image: https://openreview.net/pdf?id=0VFQhPGF1M3

Future Directions – Improving Accessibility

Perturbed by some isotropic Gaussian noise sampled from a distribution with mean 0 and covariance of the identity matrix

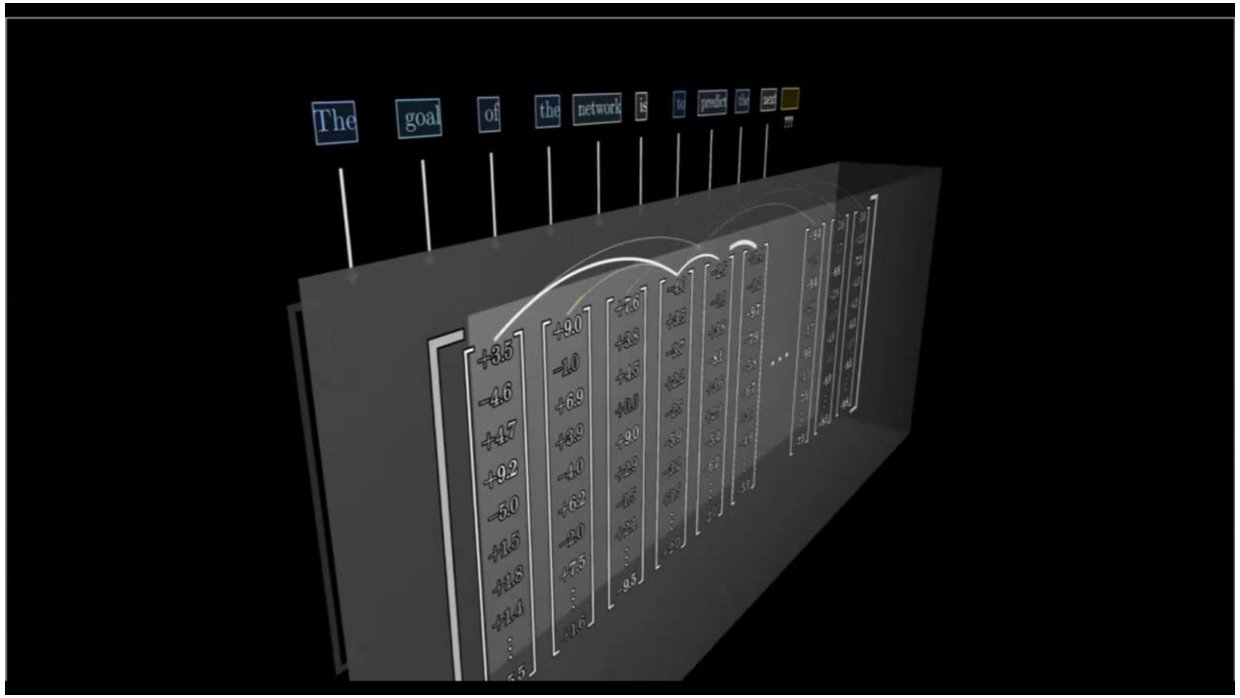
$$\mathbf{A}_j = \sum_{k=1}^{M} \phi(p_{kj}(\mathbf{Q}_k \mathbf{K}_k{}^{\top} + \boxed{\sigma_k \odot \epsilon_j)), \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}),}$$

Image: https://openreview.net/pdf?id=0VFQhPGF1M3

Future Directions

With these improvements we can:

- Improve performance
- Whilst making our model cheaper and easier to run



Video: https://youtu.be/eMlx5fFNoYc?si=NKgOvfLV9cDTMcWi

Summary

- Ames is important for public safety
- We take advantage of the recent explosion in AI research & apply it to Ames
- Our graph transformer is state-of-the-art
- Serious potential for regulatory application

- This work received a grant from ACTRA and will be presented at the ASM in August

THE UNIVERSITY OF SYDNEY