AMESFORMER: STATE-OF-THE-ART MUTAGENICITY PREDICTION WITH GRAPH TRANSFORMERS

A PREPRINT

Luke A. Thompson
Sydney Pharmacy School
The University of Sydney
Sydney, Australia
ltho4642@uni.sydney.edu.au

Josiah G. Evans
Independent Researcher
Seoul, Republic of Korea
josiah.g.e@proton.me

Slade T. Matthews
Sydney Pharmacy School
The University of Sydney
Sydney, Australia
slade.matthews@sydney.edu.au

October 11, 2024

ABSTRACT

The Ames mutagenicity test is a gold standard assay for the safety assessment of new chemicals. However, many *in silico* models rely on challenging-to-interpret ensemble strategies and molecular fingerprint data which neglects gestalt molecular structure. To improve upon these models, we propose AmesFormer, a graph transformer neural network which shows state-of-the-art performance when paired with our new Ames dataset. We briefly review the current state of Ames modelling with a focus on graph neural networks. We then benchmark AmesFormer on a standardised test dataset against 22 other Ames models, achieving state of the art (SOTA) performance. We then uniquely report the calibration performance of our model and attempts to improve it using temperature scaling. We support our findings with reference to other models from the literature and with developments in machine learning (ML) and graph theory. Overall, we present a high-performance, accessible, and open-source computational model for Ames mutagenicity, with significant potential for regulatory and drug development applications.

Keywords Ames · Mutagenicity · QSAR · GNNs

1 Introduction

1.1 The Ames Assay

The Ames test is a widely-used *in vitro* mutagenicity assay essential to drug development. Ames data is explicitly required by many regulatory guidelines, such as International Council for Harmonisation (ICH) guideline S2 (R1) (ICH 2013). It thus represents a high bar to market access for pharmaceuticals, with many compounds rejected early in the case of an Ames-positive outcome (Honma et al. 2019).

In the Ames assay, a histidine-deficient substrate is inoculated with strain of auxotrophic mutant histidine-dependent Salmonella typhimirium (Ames et al. 1973). This dependence is introduced via a mutation at the histidine operon (Ames et al. 1973). A suspected mutagenic compound is then introduced. If the mutagen-containing substrate shows significantly greater colony growth than a control, the test-molecule has reversed the histidine-dependence mutation and is mutagenic. S9 rodent liver homogenate is included to enable the detection of pro-mutagens (Maron et al. 1983). This simplicity has resulted in the Ames test having an excellent inter-laboratory replicability of 85% (Kamber et al. 2009).

Different *S. typhimurium* strains detect various mutagenicity mechanisms, which can be categorized into substitution mutations (SNPs) or frameshift mutations (Lui et al. 2023). Other strains can detect skin sensitisation and oxidative mutagenisis (Patlewicz et al. 2010; Levin et al. 1982).

1.2 QSAR Modelling

With the going rate of an Ames test at $2000 AUD (USD 1 364) per chemical, and the CAS registry growing by over 4000 molecules daily, the cost of *in vitro* screening for every new chemical is prohibitively high (Honma et al. 2019).

To address this problem, *in silico* quantitative structure activity relationship (QSAR) models have been developed to provide cheaper and higher-throughput methods of Ames screening (Furuhama et al. 2023). The use of such models is recommended within ICH guideline M7 (R1) for the control of mutagenic impurities in pharmaceuticals and the European Union's Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) agreement (European Communities 2006; ICH 2017; Honma et al. 2019). The Australian Industrial Chemicals Introduction Scheme (AICIS) and the United States Food and Drug Administration (US FDA) have also provided guidelines for implementing *in silico* toxicity evaluation methods (AICIS 2022; Han 2023).

1.3 Current Models

The input to many Ames QSAR models are molecular fingerprints (MFs). MF, as seen in Figure 1, are vectorised binary representations of a molecules chemical features (Capecchi et al. 2020).

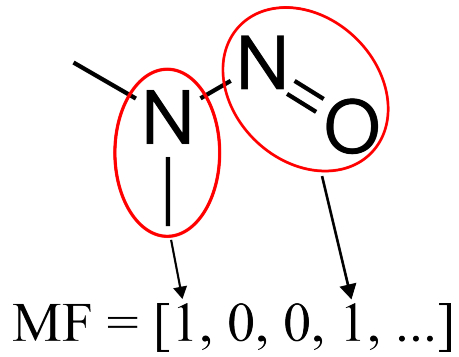


Figure 1: An illustrative example of the hashing of different N-Nitrosodimethylamine (NDMA) substructures into a MF bit vector.

Two fingerprinting methodologies are commonly applied to molecules within the Lipinski Limits (Lipinski et al. 2001). Structural key MFs, such as Molecular ACCess Systems (MACCS) keys seen in Figure 2a, encode the presence of a set of predefined chemical substructures into a binary vector (Durant et al. 2002; Seo et al. 2020). Hash MF *de novo* numerically encode non-predefined substructures. The archetypal Extended-connectivity Fingerprint 4 (ECFP-4) (Morgan FP) MF shown in Figure 2b hashes the local environment (of arbitrary size) around each atom and encodes this into the bit vector representation (Rogers et al. 2010).

$$\text{MACCS} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{166} \end{bmatrix} \rightarrow \{0,1\}^{166}$$

(a) MACCS fingerprint bit vectors encode the presence (1) or absence (0) of a set of 166 predefined molecular substructures. The index of each substructure is unique.

$$\text{ECFP} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \rightarrow \{0,1\}^{n}$$

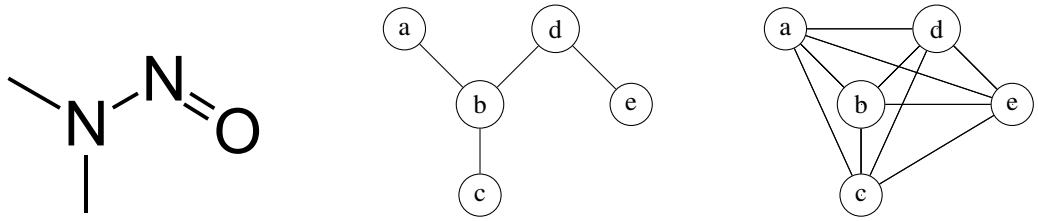
(b) Morgan fingerprints encode an arbitrarily sized set of non-predefined molecular substructures into a bit vector of arbitrary length.

Figure 2: Different molecular fingerprinting techniques and their dimensionalities.

Of the 21 Ames models presented in the Second Ames International Challenge by Furuhama et al. (2023), 19 utilise MF-type input data.

1.4 Graphs and Graph Neural Networks (GNNs)

A graph $G = (V, E)$ models entities as a V set of v nodes, and a E set of pairwise entity relationships e, known as edges, in non-Euclidean space (Scarselli et al. 2009). Node and edge features are typically contained in a real-valued d-dimensional vectors $\vec{v}_i \in \mathbb{R}^d$, $\vec{e}_i \in \mathbb{R}^d$.



(a) Skeleton structure of NDMA. (b) NDMA as a bipartite graph. (c) NDMA as a complete graph.

Figure 3: Representations of NDMA as a skeleton diagram and as two common molecular graphs.

As shown in Figure 3, molecules are analagous to graphs when we consider atoms as nodes, bonds as edges, and the qualities of atoms and bonds as the vectors associated with each node and edge.

A graph neural network (GNN) is a neural network (NN) operating on graph-structured data. It may learn to predict the qualities of nodes (e.g., hybridisation state), edges (e.g., bond length) or the whole graph (e.g., Ames positivity) (Jin et al. 2022).

GNNs first AGGREGATE the information contained in the existing feature vectors $\vec{h}_j^{(l-1)}$ of each node's n-hop neighbours $\mathcal{N}(i)$ using a permutationally invariant aggregation function such as mean or sum eq. (1). An UPDATE function then combines the aggregated information, $\vec{a}_i^{(l)}$, with the feature vector of the node $\vec{h}_i^{(l)}$ to form the new node representation $\vec{h}_i^{(l)}$ eq. (2).

$$\vec{a}_i^{(l)} = \text{AGGREGATE}^{(l)}\left(\{\vec{h}_j^{(l-1)} : j \in \mathcal{N}(i)\}\right), \tag{1}$$

$$\vec{h}_i^{(l)} = \text{UPDATE}^{(l)}\left(\vec{h}_i^{(l-1)}, \vec{a}_i^{(l)}\right) \tag{2}$$

$$\vdots$$

$$\vec{h}_G = \text{READOUT}\left(\{\vec{h}_i^{(L)}\}_{i \in V}\right) \tag{3}$$

Repeat for L layers

After L iterations or *layers* a READOUT graph-pooling function then combines the representation of every node $\{\vec{h}_i^{(L)}\}_{i \in V}$ into a unified whole-graph representation eq. (3).

1.5 The Transformer

The transformer introduced by Vaswani et al. (2017) represents a major leap in the capability of ML models across natural language processing (NLP), image processing and biomedical tasks (Devlin et al. 2018; Brown et al. 2020; Parmar et al. 2018; Radford et al. 2021; Elnaggar et al. 2020).

Given a node feature vector $\vec{h}$, we compute query Q, key K and value V matrices using three randomly initialised learnable matrices W_Q, W_K, and W_V:

$$Q = \vec{h}W_Q \qquad \text{shape: } (\vec{h}, d_k) \tag{4}$$

$$K = \vec{h}W_K \qquad \text{shape: } (\vec{h}, d_k) \tag{5}$$

$$V = \vec{h}W_V \qquad \text{shape: } (\vec{h}, d_v) \tag{6}$$

In the above, d_k denotes the dimensionalities of K, and d_v is the dimensionality of V.

The attention matrix A captures the similarity between the Q and K matrices. A is then passed through a softargmax, ensuring all attention "weights" in A sum to one. This enables the final output of the attention mechanism to be a weighted sum of the value vectors V, where the weights are given by the computed attention scores:

$$A = \frac{QK^T}{\sqrt{d_k}} \qquad \text{shape: } (\vec{h}, \vec{h}) \tag{7}$$

$$\text{Attention}(Q, K, V) = \text{softmax}(A)VO \qquad \text{shape: } (\vec{h}, d_v) \tag{8}$$

Multiple attention calculations or "heads" are performed in parallel on the same input vector using different Q, K and V matrices, then passed through a linear output layer O (Vaswani et al. 2017). Their outputs are later concatenated to form the final representation. This concatenated output is fed through a conventional NN.

For further details, interested readers may consult an excellent review on the transformer and its variants by Lin et al. (2021).

1.6 Related Work

A diverse array of ML architectures have been applied to Ames mutagenicity QSAR studies. As early as 2004, nearest-neighbour and decision-tree models were employed by Votano et al. (2004). In the same lineage, newer descriptor-based models often leverage support vector machines, random forest models, and gradient boosted trees, among others (Xu et al. 2012; Tintó-Moliner et al. 2020; Chu et al. 2021). A number of novel approaches have also been applied, including weighted ensembles and multitask learning models which exploit strain-specific Ames data (Li, Liu, et al. 2023; Feeney et al. 2023). The majority of contemporary models belong to this family of models using classical ML techniques (Furuhama et al. 2023).

When considering the family of deep-learning Ames models, of which AmesFormer is one, a number of convolutional (Shi, Yang, et al. 2019; Tran et al. 2024) and graph-convolutional (Li, Zhang, et al. 2021) approaches have been recently applied. Along similar lines, Hung et al. (2021) presented a Bayesian GNN and attentional graph convolutional network (GCN) for Ames mutagenicity, achieving near-SOTA results. The closest cousin of our model in the literature is the recent Ligandformer by Guo, Liu, et al. (2022) which shares our basic transformer architecture, but lacks the structural encodings we employ.

Direct comparison of performance between these models is hampered by their use of varying test datasets and reporting methodologies. To remedy this in our case, we report our model's test performance on a standardised dataset provided as part of the Furuhama et al. (2023) Second Global Ames prediction challenge, the sequel to Honma et al. (2019).

2 Methods

2.1 Datasets

We train two AmesFormer models. The first dubbed "AmesFormer-Honma" was trained on the Honma training dataset which was provided as part of the Furuhama et al. (2023) Second Ames International QSAR Challenge. The second model, "AmesFormer-Pro" was trained on our new partially open-source Ames dataset constructed as part of this study. The use of custom datasets was within the rules of the Furuhama et al. (2023) contest. The non-proprietary compounds of our dataset are available in our repository.

Both AmesFormer models were evaluated on the Honma test set — These are the final results we report. We were kindly granted access to the Honma training and test datasets after the contest had concluded by the authors and the National Institute of Health Sciences, Japan (NIHS-J).

2.2 Dataset Construction

To construct both datasets, Simplified Molecular-input Line-entry System (SMILES) were converted to graphs using Pytorch-Geometric with hydrogen atoms omitted to save computational resources in line with many previous GNN architectures (Xiong et al. 2020; Jin et al. 2022).

2.2.1 Honma Dataset

We re-coded the Honma dataset for binary classification by considering Ames class 1 and class 2 compounds as Ames-positive, and Ames class 3 compounds as Ames-negative. We featurised the molecular graphs using the RDKit functions described in Table 2.

2.2.2 Our Dataset

Our new dataset, dubbed the "combined" dataset, was composed from four sources: The "Honma" dataset, the European Union Reference Laboratory (EURL) dataset, the ISSTOX dataset, the "Hansen" dataset (Hansen et al. 2009; Benigni et al. 2013; Madia et al. 2020; Furuhama et al. 2023). These datasets were identified via a literature search.

We first cleaned and canonicalised all SMILES in each dataset using RDKit by transforming them to RDKit molecule objects and back to SMILES strings. We then combine these four cleaned datasets, eliminating duplicates. If SMILES clashed, the Ames results presented in newer datsaets was prioritised, as shown in the Table 1. SMILES were also dropped if they could not be disambiguated, were broken, or if RDKit was unable to generate Mol objects based on their SMILES.

The Honma- and ISSTOX-derived Ames results were recoded as shown in Table 1. Neither the Hansen nor EURL datasets required recoding.

Table 1: Overview of the combined dataset. Duplicates were removed, with higher priority (lower number) sources preferred. The number of datapoints we incorporate into combined from the initial amount present in each dataset is dented # Incorporated and #Initially. The recoding of Ames test results unsuitable for binary classification is denoted as: original $\rightarrow$ new code.

Name	**Source**	**Priority**	**# Incorporated**	**# Originally**	**Recoding**
Honma	(Furuhama et al. 2023)	1	12134	13729	$1 \rightarrow 1$ $2 \rightarrow 1$ $3 \rightarrow 0$
EURL	(Madia et al. 2020)	2	197	211	No recoding necessary
ISSTOX	(Benigni et al. 2013)	3	6460	7367	$3 \rightarrow 1$ $2 \rightarrow 0$ $1 \rightarrow 0$ Inconclusive $\rightarrow 0$
Hansen	(Hansen et al. 2009)	4	3728	6514	No recoding necessary
Total	**This paper**		**22519**	**27821**	

Molecular graphs were generated using the RDKit functions presented in Table 2 using Pytorch-Geometric. All dataset construction and analysis code is available in the supplementary materials.

Table 2: RDKit atom and bond encodings used to featurise molecular graph nodes and edges.

Atom Encodings	
Atomic Number	GetAtomicNum()
Chirality	GetChiralTag()
Degree	GetTotalDegree()
Formal Charge	GetFormalCharge()
Hydrogen count	GetTotalNumHs()
Radical electron count	GetNumRadicalElectrons()
Hybridisation State	GetHybridization()
Aromatic bond (bool)	GetIsAromatic()
Part of a ring? (bool)	IsInRing()
Bond Encodings	
Bond type	GetBondType()
Stereoisomerism	GetStereo()
Conjugation	GetIsConjugated()

2.3 Training Process

We first split each dataset into training, validation and held-out test sets as described in Table 3 and Table 4. We then performed manual and Bayesian hyperparameter optimisation via Optuna (3.6.1) on the Honma training and validation data to discover optimal hyperparameters for both our models (Akiba et al. 2019). We ran 100 trials with hyperband pruning (Li, Jamieson, et al. 2016).

We then trained AmesFormer-Honma and AmesFormer-Pro on a combined train and validation dataset using these discovered hyperparameters. Each AmesFormer model was trained using binary cross entropy loss (BCE Loss) loss for 100 epochs with a batch size of 32 on an NVIDIA Titan V graphics processing unit (GPU) supplied in an NVIDIA Accelerated Data Science Grant.

Our batch size is substantially smaller compared to those used in Graphormer (256), Hung et al. (2021) (128), LigandFormer (256) and other molecular GNNs present in the literature (Chengxuan et al. 2021; Guo, Liu, et al. 2022; Yuan et al. 2020). This choice was a deliberate attempt to empirically approximate an optimal "temperature", the ratio of batch size to learning rate, as defined in McCandlish et al. (2018). Despite similar naming, this temperature is unrelated to the sigmoid temperature we discuss later.

We use the AdamW optimiser and set ϵ to 1×10^{-8}, and $(\beta 1, \beta 2)$ to $(0.9, 0.999)$. We employ a simple fixed learning rate (LR) scheduler set to 1.5×10^{-4}. We discontinued training at 70 epochs as it was approximately the end of the minimum validation loss window before the onset of overfitting for AmesFormer-Honma. We stopped the second checkpoint at 100 epochs. A complete overview of AmesFormer hyperperameters is available in the appendix Table 7.

Both AmesFormer-Honma and AmesFormer-pro were evaluated on the same test set (n = 1584), the makeup of which was determined by the NIHS-J. Under contest conditions, the test set was not given to participants until models had already been submitted to the authors. Whilst we had access to this data from the beginning, we self-blinded and did not use it for training/validation. In the 2nd Ames International challenge, participants were permitted to submit an unlimited number of models (Furuhama et al. 2023). The model from each team showing the best test set performance was then used for the final comparison. In line with this, we tested two checkpoints for each AmesFormer: One at 70 epochs and one at 100 epochs as describe above.

2.4 The Architecture of AmesFormer

AmesFormer is fundamentally a full reimplementation of Graphormer using PyTorch (2.2.2) and PyTorch-Geometric (2.5.2) (Paszke et al. 2019; Fey et al. 2019).

AmesFormer contains three transformer blocks, each with four attention heads of dimension 128. The feed forward neural network (FFN) uses Gaussian error linear unit (GELU), is of dimension 80, and has dropout set to 5%. We set max shortest path distance (SPD) to five as higher values significantly increased training time. Final readout is accomplished using a virtual node which aggregates all node representations but does not contain SPD information. Both AmesFormer models contain a 264 696 total parameters.

2.4.1 Structural Encodings

We retain the three structural encodings implemented in the original Graphormer (Chengxuan et al. 2021).

To construct the *centrality encoding* we append the degree of each node to their feature vectors $\vec{h}_i$, and parameterise this by a learnable scalar z (9) (Chengxuan et al. 2021).

$$\vec{h}_i = \vec{h}_i \| z_{\deg(v_i)} \tag{9}$$

The *spatial encoding* is a function $\phi(v_i, v_j) \to \mathbb{R}$ representing the SPD between any two nodes, v_i, v_j in a graph G. The SPD is multiplied by a learnable scalar b and added to the self-attention calculation as an attention prior or bias (Chengxuan et al. 2021). The same biasing scalar b is shared between all layers.

The *edge encoding* aggregates edge information along the shortest path $SP_{ij} = (e_1, e_2, \ldots, e_N)$ between node pairs (v_i, v_j) (Chengxuan et al. 2021). It is computed as the mean of the dot products between edge feature vectors $\vec{e}_n$ and a corresponding weight embedding w_n^E. This approach integrates structural features directly into the attention mechanism.

This yields the Graphormer attention formula representing the attention coefficients between node i and j:

$$A_{ij} = \underbrace{\frac{(h_i W_Q)(h_j W_K)^T}{\sqrt{d_k}}}_{Self-attention} + \underbrace{b_\phi(v_i, v_j)}_{Spatial\ encoding} + \underbrace{\frac{1}{N}\sum_{n=1}^{N} \vec{e}_n \cdot (w_n^E)^T}_{Edge\ encoding} \tag{10}$$

Contrasting Chengxuan et al. (2021), we implement a breadth first search (BFS) algorithm in Rust to calculate SPDs for the spatial and centrality encodings (algorithm 1).

Algorithm 1 Breadth-First Search for Shortest Paths

Require: Graph $G = (V, E)$, source node $s \in V$
Require: N = Maximum shortest path distance (SPD)
Ensure: Shortest paths from s to all other nodes in G

Initialize Q as an empty queue
Initialize $visited$ as an empty set
Initialize $nodepaths$, $edgepaths$ as matrices of size $|V| \times N$ with all values set to -1
Initialize $nodepaths[s, 0] \leftarrow s$
Enqueue s into Q
Add s to $visited$
while Q is not empty **do**
  Dequeue a node v from Q
  for each neighbor w of v in G **do**
    if $w \notin visited$ **then**
      Add w to $visited$
      Copy $node\_paths[v]$ to $node\_paths[w]$
      Find the first empty slot in $node\_paths[w]$ and set it to w
      $E_i \leftarrow$ Get the index of the edge (v, w) in E
      if no entry with a value of -1 in $edgepaths[w]$ **then**
        Skip to the next iteration
      end if
      $free\_idx \leftarrow$ The index of the first entry with a value of -1
      Set $edge\_paths[w, free\_idx]$ to E_i
      Enqueue w into Q
    end if
  end for
end while
return $paths$

2.5 Temperature Scaling

In an attempt to improve model calibration we perform temperature scaling on the sigmoid temperature. We optimise expected calibration error (ECE) on the validation dataset using 50 iterations of Brent's method bounded bounded between 0.1 and 5.0 (Guo, Pleiss, et al. 2017).

We perform the following transformation on the final sigmoid layer of AmesFormer:

$$p(y = 1|x) = \frac{1}{1 + \exp(-z/T)}$$

Where:

- $p(y = 1|x)$ is the probability of the positive class given input x
- z is the logit (the input to the sigmoid function)
- T is the temperature parameter

2.6 Statistical Methodologies

2.6.1 Performance Measurement

We assess the performance of AmesFormer using balanced accuracy (BA) and F1 score:

$$\text{BA} = \frac{\text{specificity} + \text{recall}}{2} \tag{11}$$

$$F1 = 2\left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right) \tag{12}$$

Where:

- TP: True positives; TN: True negatives; FP: False positives; FN: False negatives.
- Specificity: The ratio of correctly predicted negatives to the total actual negatives, defined as $\frac{TN}{TN+FP}$
- Precision: The ratio of correctly predicted positives to the total predicted positives, defined as $\frac{TP}{TP+FP}$.
- Recall: The ratio of correctly predicted positives to all observations in the actual class, defined as $\frac{TP}{TP+FN}$.

BA, shown in eq. (11), is the percentage of molecules that the model correctly predicts, adjusted to ensure equal weighting of model performance on both the majority and minority classes (Brodersen et al. 2010). As both datasets show considerable class imbalance, accuracy alone would be a misleading performance metric (Provost et al. 2001).

The F1 score eq. (12) is the harmonic mean of recall and precision (Hicks et al. 2022). Like BA, F1 is robust to class imbalance (Hicks et al. 2022). It is also suited to regulatory toxicological tasks such as Ames mutagenicity where both false positives and false negatives have serious financial or safety implications (Liu et al. 2017; Hicks et al. 2022).

2.6.2 Calibration Analysis

Calibration reflects how accurately a model's predicted probabilities reflect the true likelihood of an outcome. For example, if a well calibrated model outputs an 80% probability of an event occurring, the event should actually be observed 80% of the time (Guo, Pleiss, et al. 2017).

For our calibration analysis, we consider the sigmoid of the logits as representing the probability the model ascribes to a given molecule being Ames-positive.

We assess the calibration before and after temperature scaling using a calibration curve and expected calibration error (ECE) eq. (13) (Pakdaman Naeini et al. 2015). A lower ECE indicates better calibration and quantifies the difference between predicted probabilities and actual outcomes by averaging the absolute differences across bins of predictions.

$$\text{ECE} = \sum_{i=1}^{N} \frac{|B_i|}{n} \left|\text{acc}(B_i) - \text{conf}(B_i)\right| \tag{13}$$

Where:

- N: The total number of bins used to group predictions based on their predicted confidence.
- B_i: The set of indices of samples that fall into the i-th confidence bin.
- $|B_i|$: The number of samples in the i-th bin, i.e., the size of bin B_i.
- n: The total number of samples in the dataset, i.e., $n = \sum_{i=1}^{N} |B_i|$.
- acc(B_i): The accuracy of the model for the samples in bin B_i, defined as the proportion of correctly classified samples in B_i.
- conf(B_i): The average confidence of the model for the samples in bin B_i, defined as the mean of the predicted probabilities assigned to the predicted class for the samples in B_i.

3 Results

3.1 Exploratory Data Analysis

3.1.1 Ames Outcome Distributions

The Honma training set and Honma test set contained roughly the same proportion of Ames-positive and Ames-negative species. By virtue of including other datasets with less class imbalance, the combined training set included many more Ames-positive compounds (28.4%).

Table 3: Overview of the Honma dataset split.

Split	Train	Validation	Test
Size	9707	2426	1589
% positive	14.4	14.4	14.9

Table 4: Overview of the combined dataset split.

Split	Train	Validation	Test
Size	18015	4503	1589
% positive	28.4	29.6	14.9

3.1.2 Physicochemical Property Distributions

The distributions of physicochemical properties, including C-LogP, TPSA, molecular weight, H-bond donor and acceptor counts, and the number of rotatable bonds, were compared across the three datasets using the Kolmogorov-Smirnov test with Holm-Bonferroni corrections ($\alpha = 0.008$) (Holm 1979). The resulting p-values are summarised in Figure 9.

For C-LogP, TPSA, and H-bond acceptors, significant differences ($p < 0.008$) were observed for all pairwise dataset comparisons. Conversely, molecular weight did not exhibit significant differences between datasets (all $p = 1.000$). H-bond donor counts showed significant differences between the Combined Train dataset and both Honma Train and Honma Test datasets ($p < 0.008$). These results indicate that, while visual inspection of property distributions suggests general similarity, statistically significant differences exist for most properties across datasets.

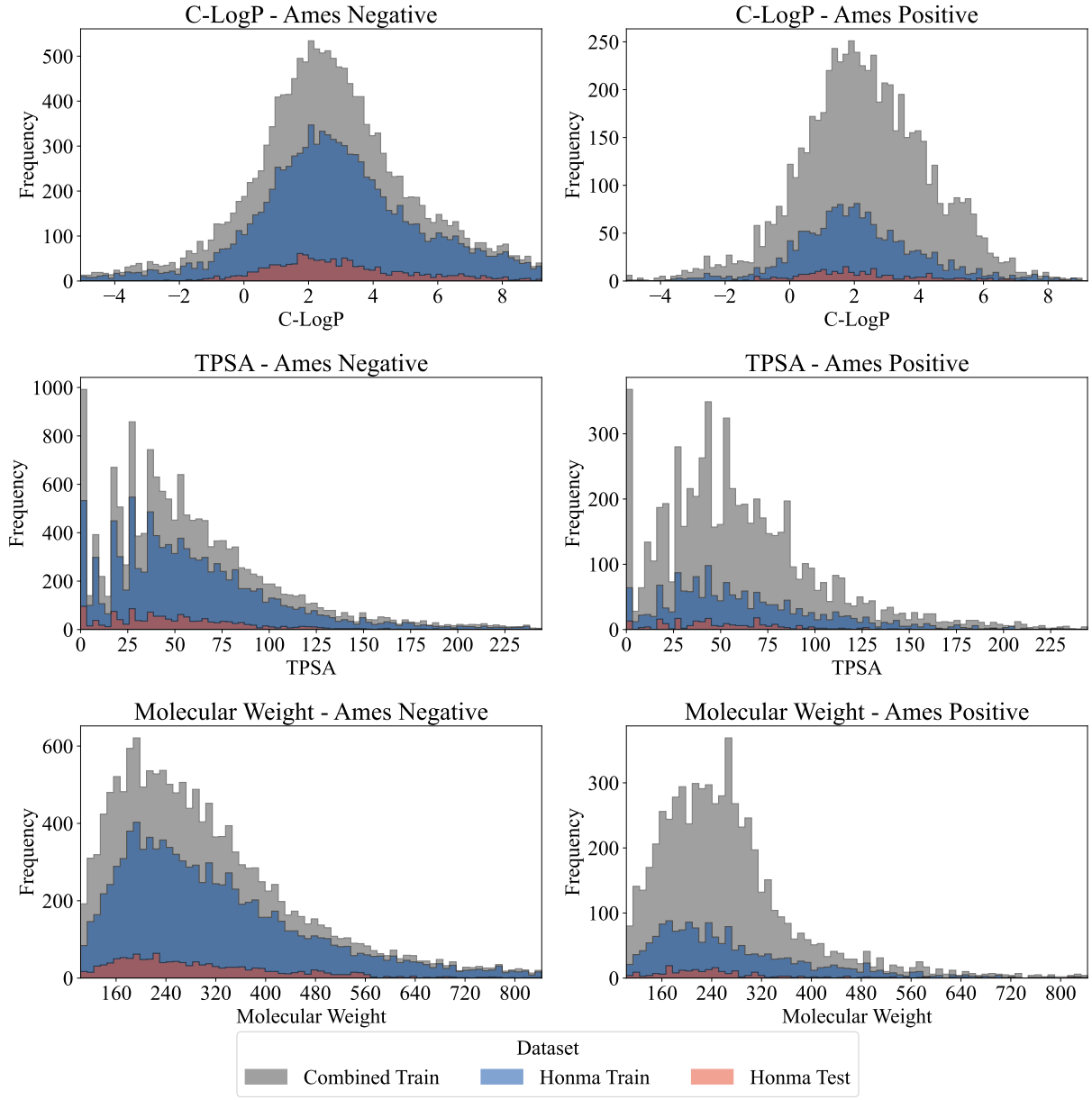


Figure 4: Continuous physicochemical property descriptors of the molecules contained within the three examined datasets. C-LogP, the Crippen LogP value; topological polar surface area (TPSA) in square Angstroms ($Å^2$), molecular weight in grams per mol (g/mol). Histogram bins were determined using the Freedman-Diaconis rule and descriptor values were generated using RDkit (Freedman et al. 1981; Landrum 2012).

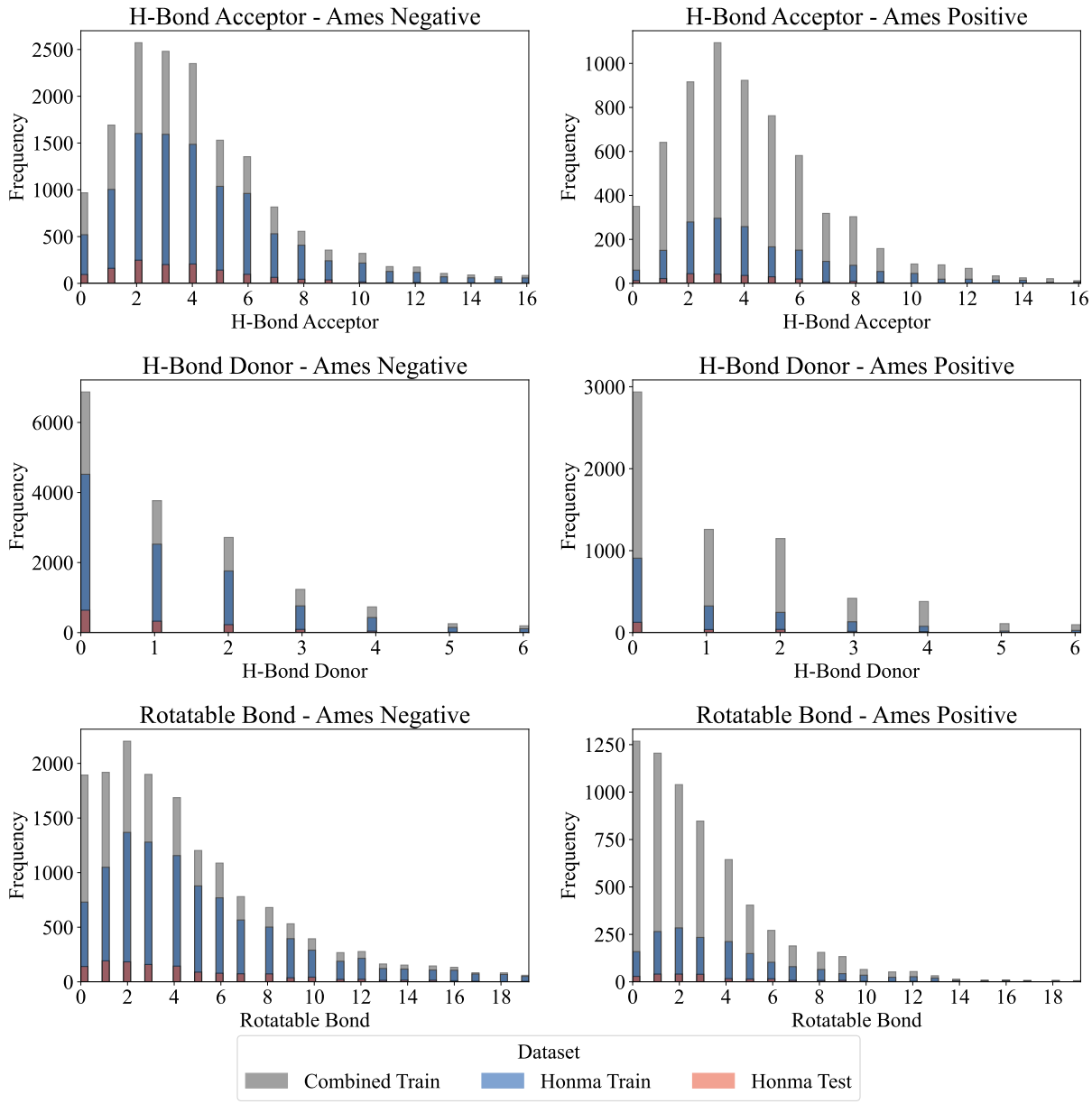
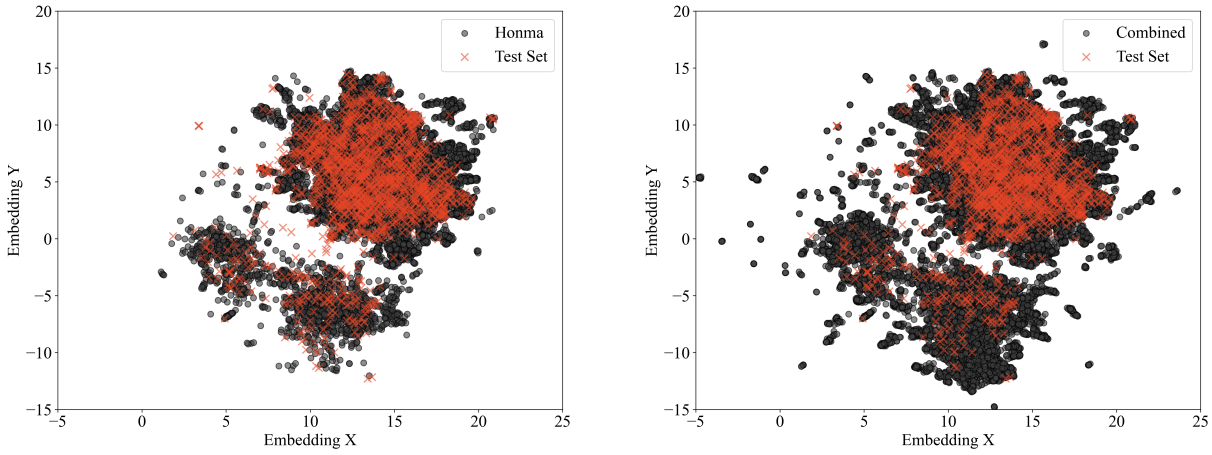


Figure 5: Discrete physicochemical property descriptors of the molecules contained within the three examined datasets. The number of hydrogen bond acceptors, the number of hydrogen bond donors, the number of rotatable bonds. Histogram bins were determined using the Freedman-Diaconis rule and descriptor values were generated using RDkit (Freedman et al. 1981; Landrum 2012).

3.1.3 Manifold Projection

The supervised Uniform Manifold Approximation and Projection (UMAP) Figure 6 displays the greater chemical diversity of the combined dataset when compared to the Honma dataset, as it covers a wider region of UMAP space. When projected with the combined dataset, the Honma test set shows far fewer "orphan" datapoints, those without any close neighbours in the training set. This is particularly noticable in the gulf between the two large clusters, and around the sparse north-west segment of the plot. Thus, it appears that the combined dataset covers a broader chemical space than the Honma dataset training alone, and contains more compounds similar to those of the test set.



(a) UMAP of the Honma dataset and Honma test set.

(b) UMAP of the combined dataset and Honma test set.

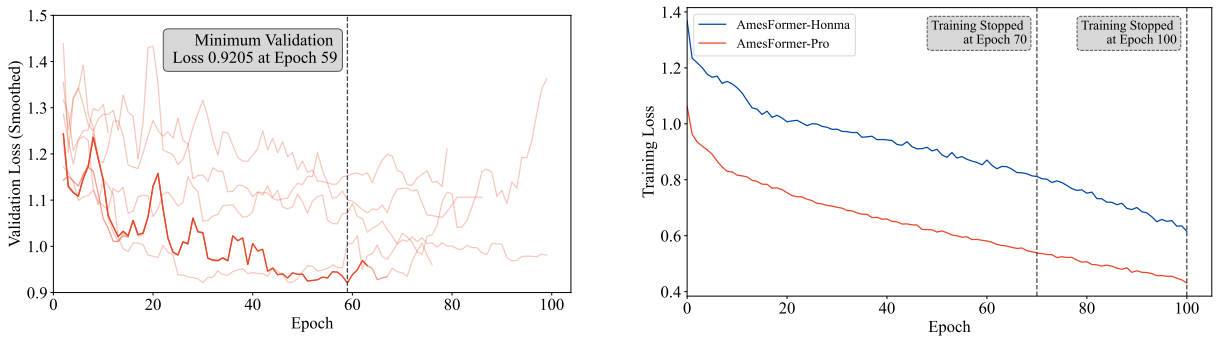
Figure 6: UMAP embeddings on a shared manifold comparing the chemical space of the Honma and Combined datasets to the Honma test dataset. Molecules were represented as 2048-bit, radius three Morgan fingerprints, with Jaccard distance as the ambient space metric.

The greater similarity of the combined training set with the Honma test set shown by the physicochemical property descriptors and UMAP embeddings suggests the combined dataset contains a more representative chemical diversity. As prior work has shown models struggle to extrapolate beyond the chemical space of their training data, this greater overlap of chemical properties likely enabled learning of more robust structure-property relationships (Tropsha 2010). This may partly explain the enhanced performance observed in downstream modelling.

3.2 Insights from Training AmesFormer

In Figure 7a we present the validation loss produced by a series of hyperparameter optimisation runs on the Honma training and validation dataset. We highlight the run with the hyperparameters shown in Table 7 which produced the lowest validation loss of 0.921 at epoch 59.

The training loss of both AmesFormer-Honma and AmesFormer-Pro decline monotonically to epoch 100 as seen in Figure 8d. We may attribute the significantly better convergence of AmesFormer-Pro to the greater size of the combined dataset improving generalisation performance (Hestness et al. 2017).



(a) A series of hyperparameter optimisation runs on the Honma training and validation dataset. The best-loss run is highlighted, with the best epoch and loss denoted. We apply window three rolling average smoothing to aid trend analysis.

(b) The training loss of AmesFormer-Honma and AmesFormer-Pro on the combined training/validation sets of their respective datasets. Checkpoints and associated training losses are denoted with a vertical dashsed lines.

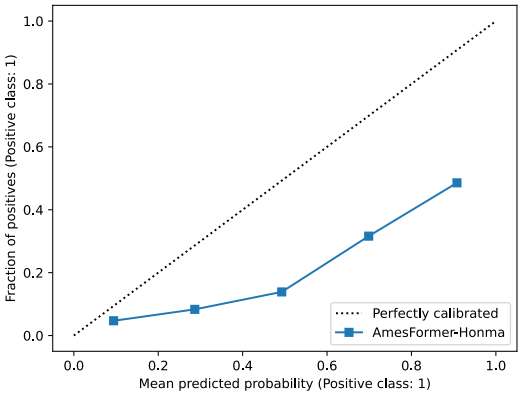
Figure 7: BCE Loss losses associated with hyperparameter optimisation, and with the training of AmesFormer-Honma and AmesFormer-Pro.

3.3 Accuracy and Calibration Performance

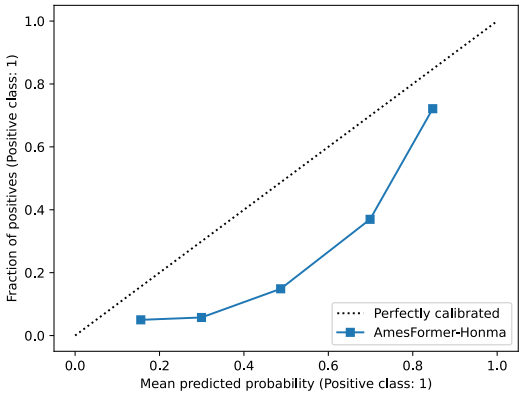
AmesFormer-Pro produced a BA of 82.1% and an F1 score of 0.674 on the Honma test set. Sigmoid temperature scaling failed to improve the ECE of 0.075, producing a ECE of 0.155 at a sigmoid temperature of $T = 3.14$. The good ECE results of untuned AmesFormer-Pro suggest that its outputs are sufficiently calibrated to serve as a reliable indicator of model confidence.

AmesFormer-Honma showed weaker performance, with a BA of 74.0 and F1 score of 0.479. AmesFormer-Honma was initially miscalibrated with an ECE of 0.200, which worsened to 0.259 with $T = 2.08$.

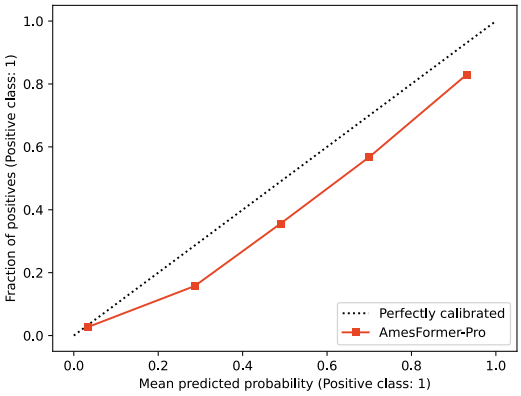
Previous studies have demonstrated that imbalanced datasets can lead to systematic biases in model predictions, particularly for minority classes (Johnson et al. 2019). AmesFormer-Honma exhibited a systematic underestimation of Ames-positive compounds, as seen in Figure 8a. This bias can be attributed to the low proportion of Ames-positive molecules (14.4%) in the Honma training set, which likely impaired the model's ability to effectively capture the key features of mutagenicity. In contrast, the significantly greater number of positive examples (28.4%) in the combined training set likely mitigated the tendency towards underestimation by providing more balanced representation of the feature space as shown in Figure 8c.



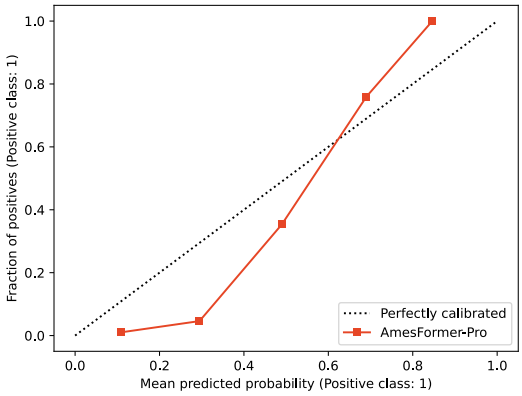
(a) AmesFormer-Honma calibration prior to sigmoid temperature scaling. $T = 1.00$, ECE = 0.200.



(b) AmesFormer-Honma calibration with "optimal" sigmoid temperature scaling. $T = 2.08$, ECE = 0.259.



(c) AmesFormer-Pro calibration prior to sigmoid temperature scaling. $T = 1.00$, ECE = 0.075.



(d) AmesFormer-Pro calibration with "optimal" sigmoid temperature scaling. $T = 3.14$, ECE = 0.155.

Figure 8: Calibration curves with five bins for AmesFormer-Honma and AmesFormer-Pro before and after sigmoid temperature scaling.

3.4 Performance and Comparison with the Literature

AmesFormer-Pro ranked first whilst AmesFormer-Honma ranked third in BA among the 21 models that reported results in the 2nd Global Ames Challenge, as shown in Table 5 (Furuhama et al. 2023). This represents a 12% and 3.9% improvement, respectively, over the previous results from our lab, achieved by the DRSpicySTiM-Ensemble model (Furuhama et al. 2023). AmesFormer-Pro showed the best overall F1 score (0.674), whilst AmesFormer-Honma was showed poorer a F1 of 0.479.

Table 5: A comparison of the performance of AmesFormer with models presented in Furuhama et al. (2023). Our results for AmesFormer are shown in bold. See (Furuhama et al. 2023) for model version information.

Team or Institution Name	Model Name	Datapoints	BA (%)	F1
Our result	**AmesFormer-Pro**	**22518**	**82.1**	**0.674**
MN-AM	ChemTunes. ToxGPS Ames	13730* + unk. prop	78.5	0.538
Meiji Pharmaceutical University	MMI-STK2	13730* + 8103†	77.0	0.524
Our result	**AmesFormer-Honma**	13730*	74.0	0.479
Instem	Leadscope Consensus	13730* + 9248 prop	73.7	0.497
LMC Bourgas University	TIMES_AMES	4127 + 1734 prop	73.3	0.511
Altox Ltd.	GeneTox-iS	unk. prop	72.6	0.500
Evergreen AI, Inc.	Avalon	13730* + unk. prop	71.9	0.485
MultiCASE Inc.	PHARM_BMUT	13730* + unk. prop	71.2	0.497
Simulations Plus Inc.	S+MUT_NIHS_ABC	13730*	71.2	0.421
The University of Sydney	DRSpicySTiM-Ensemble	13730*	70.1	0.425
Lhasa Ltd.	Sarah Nexus (2068)	13730* + 11774	69.0	0.410
NCTR/FDA	DeepAmes	13730*	69.1	0.476
IRFMN	CONSENSUS (18k)	13730* + Unk. prop	68.1	0.402
Liverpool John Moores University	DL	13730*	68.7	0.403
NIBIOHN	GNN(kMoL)_bestbalanced	13730*	67.2	0.470
SIOC, CAS	CISOC-PSMT	Unk. prop	66.4	0.393
Politecnico di Milano	GCN	13730*	65.8	0.444
IdeaConsult Ltd.	AMBIT DeepN	13730*	65.6	0.408
Massachusetts Institute of Technology	Chemprop	13730*	64.3	0.420
Chemotargets	CHMT_GBoostSC	13730* + unk.	64.3	0.414
Istituto Superiore di Sanità	ISS-modified2020	13730* + 7367‡	62.8	0.348
Gifu University	xenoBiotic	13730* + unk. prop	60.3	0.334

*Honma dataset, †Hansen Dataset, ‡ISSSTOX

4 Discussion

4.1 Understanding the Performance of AmesFormer

4.1.1 The Impact of Data

As seen in Table 5, the MN-AM and Meiji Pharmaceutical University models incorporate significantly more training data points than AmesFormer-Honma . In this light, it is unsurprising that they report BAs and F1s greater than this model as dataset size strongly correlates with model ability (Halevy et al. 2009). This phenomenon has been specifically demonstrated for classical ML- and NN-based QSAR classifiers (Rácz et al. 2021; Frey et al. 2023).

4.1.2 AmesFormer is a Particularly Strong Learner

AmesFormer-Honma performs better than many other models incorporating data beyond the Honma training set. AmesFormer-Pro also shows superior performance to MMI-STK2 despite being trained on a similarly sized dataset. Thus, even when considering performance relative to dataset size, AmesFormer appears to be a particularly strong learner.

GNNs have consistently posted excellent results in cheminformatics tasks. The basis of AmesFormer, Graphormer, achieved SOTA performance on the ZINC-500K molecular prediction task at the time of its publishing (Chengxuan et al.

2021). Since then, Menegaux et al. (2023) and (Ma et al. 2023) have achieved SOTA mean absolute error (MAE) results on ZINC-500K, supporting the notion that GNNs are sufficiently strong learners to post best-in-class performance on cheminformatics tasks. The excellent few-shot learning capabilities of GNNs also specifically benefits performance on the Ames task as the small size of available datasets necessitates data-efficient learning schemes (Guo, Zhang, et al. 2021). Even the most modest benchmark graph property prediction datasets, such as Open Graph Benchmark MolHIV, contain four-fold as much data as preexisting Ames datasets (Hu et al. 2020).

Furthermore, Kriege et al. (2019) found that the Weisfeiler-Lehman (WL) graph isomorphism kernel is the best-performing kernel for Gaussian Processes (GP)-based classification of mutagenicity. Interestingly, Chengxuan et al. (2021) showed that Graphormer (and hence AmesFormer) is equivalent to a shortest-path-distance enhanced variant of the WL test. As the WL-equivalent edge embeddings of AmesFormer share a common theoretical foundation with the WL kernel, it is reasonable to hypothesise that AmesFormer may inherit the excellent performance of the WL kernel for mutagenicity (Kriege et al. 2019). However, as the learning paradigm of GPs differs fundamentally from that of our graph transformer, such an extrapolation should be taken with some caution.

4.1.3 Versus Two Other Excellent Models

The best-performing models in the Furuhama et al. (2023) Ames prediction challenge, ChemTunes ToxGPS and MMI-STK2, are both ensemble classifiers.

In the cheminformatics context, ensembles have a unique advantage as they access richer chemical information compared to GNNs operating on graph-structured data. To illustrate this point for the Ames prediction task, consider an ensemble of a NN and a GNN. The NN operating on fingerprint data may access the imprecise atom and bond information encoded by the MF, and the global features encoded by some descriptors or MF (i.e., Mordred descriptors which encode LogP solubility information). As in AmesFormer, the constituent GNN of the ensemble may still access the detailed atom-level and bond-level information. However, this combination allows the ensemble model to access both the precise atom-level information yielded by the GNN, and imprecise whole-molecule information seen by the NN which processes MF.

As gestalt molecular properties, notably solubility and TPSA, strongly correlate with mutagenicity, we hypothesise this extra MF-encoded global molecular accessed by ensembles contributes to their excellent performance. The benefits of such whole-molecule data also holds for GNNs. Rampáek et al. (2022) showed incorporation of graph-level features via concatenation with node feature vectors can produce modest performance improvements over Graphormer.

4.2 The Usefulness of Calibration and Confidence

Both calibration and confidence estimates are intrinsically linked in their use in pre-clinical and regulatory toxicology models (Marshall et al. 2016).

Consider a scenario in which a new chemical entity is due to be introduced to the market. The choice of whether to undertake further *in vitro* testing should be influenced by the *confidence* of the model's prediction. For this reason, the Organisation for Economic Co-operation and Development (OECD) recommends uncertainty metrics for all regulatory QSAR models (OECD 2014).

To be useful, this confidence must be *calibrated* to reflect the true likelihood that the predicted label aligns with the ground truth (Guo, Pleiss, et al. 2017). A miscalibrated model that outputs predictions misaligned with actual likelihoods can lead to poor decision-making, such as unnecessary testing or the oversight of harmful compounds. In a regulatory context, where resources and time are limited, well-calibrated confidence estimates enable efficient prioritisation of chemicals for further evaluation.

Our attempts at temperature scaling to improve the ECE of AmesFormer were unsuccessful. This outcome may stem from overfitting the temperature parameter to the validation set, a common issue when the validation data is not fully representative of the test distribution. Overfitting in temperature scaling has been noted in previous studies, particularly when the optimization process is limited to a fixed set of data, resulting in suboptimal generalization to unseen examples (Guo, Pleiss, et al. 2017).

Moreover, our current root-finding approach to temperature scaling may be too simplistic. Quasi-Newton algorithms or stochastic gradient descent (SGD) could potentially explore a broader range of temperature values and avoid local minima, potentially improving calibration. Alternative techniques, such as isotonic regression (Zadrozny et al. 2002) and Platt scaling (Platt 2000; Niculescu-Mizil et al. 2005), may also help improve calibration performance

4.3 Future Directions

Australian chemical regulators have expressed interest in AmesFormer for *in silico* screening of new chemical entities. In line with this, we intend to construct an OECD QSAR model reporting format (QMRF) document for AmesFormer in the near-future. In this regulatory environment, additional training data gradually emerges as novel chemicals are encountered and screened.

Implementing an incremental learning framework would allow AmesFormer to assimilate new data on-the-fly, potentially accelerating regulatory responses to emerging chemical classes. This approach could be particularly valuable in scenarios where rapid assessment of novel compounds is crucial for public safety and environmental protection. However, the implementation of such a system necessitates careful consideration of potential pitfalls, particularly the phenomenon of catastrophic forgetting (French 1999). This occurs when neural networks abruptly lose previously learned knowledge when trained on new tasks without proper mechanisms to retain earlier information (French 1999).

To mitigate this issue, techniques such as elastic weight consolidation (EWC) or Learning without Forgetting (LwF) would need to be incorporated to balance the stability and plasticity of the model (Kirkpatrick et al. 2017; Li and Hoiem 2016).

The successful implementation of an incrementally learning AmesFormer could serve as a paradigm for adaptive *in silico* toxicity prediction in regulatory frameworks globally, potentially revolutionizing the efficiency and responsiveness of chemical safety assessments.

5 Conclusion

We present AmesFormer, a state of the art (SOTA) molecular graph transformer for Ames mutagenicity prediction. Our open-source model delivers superior performance backed by comprehensive calibration metrics to validate its robustness. Finally, we make available the open-source components of the dataset we constructed to achieve this benchmark performance.

A Data Availability

All model, data cleaning and results figure construction code is available at https://github.com/luke-a-thompson/AmesFormer. The Combined dataset excluding proprietary data from the Honma dataset is available in the "datacleaning/" directory.

B AmesFormer Hyperparameters

Below we show a complete overview of the hyperparameters used to train both AmesFormer-Honma and AmesFormer-Pro.

Table 6: The hyperparameters of AmesFormer, discovered via manual and Optuna hyperparameter optimisation. Dropout was not optimised.

Parameter	Value	Note
Batch size	32	
LR scheduler	Fixed	
LR	1.5×10^{-4}	
AdamW $\beta 1$	0.9	
AdamW $\beta 2$	0.999	
AdamW ϵ	1×10^{-8}	
Gradient clipping	5	
Dropout	0.05	Preset, not discovered
Attention Dropout	0.10	
Weight Decay	0.00	
Max SPD	5	For edge and spatial encodings

C AmesFormer Model Configuration

The following table presents a detailed overview of architectural specifics of AmesFormer. Our model is closely based on Graphormer by Chengxuan et al. (2021). Please see their work for further details.

As part initial model design, we experimented with a number of alternative attention and normalisation options. Manual testing did not find these options preferable to a canonical transformer implementing multi-head attention (MHA), pre-LayerNorm.

It is worth noting that Shi, Zheng, et al. (2022) showed improved performance for Graphormer when shifting the layernorm after the MHA module, in line with the original transformer (Vaswani et al. 2017). We did not experiment with this modification ourselves, however it is an open direction for a potential marginal improvement.

We encourage interested readers to train their own AmesFormer models using these alternative archtectures as all the required code is available on our GitHub repository.

Table 7: The architectural parameters used to consturct AmesFormer. d refers to dimensionality.

Module	Value	Others briefly tested
Optimiser	AdamW	SGD
LR scheduler	Fixed	Polynomial, Plateau, Greedy (Subramanian et al. 2023), One-cycle (Smith et al. 2017)
Transformer Blocks	3	
Attention heads	4	Heterogenous increasing, decreasing
Attention head d	128	
FFN d	80	
FFN activation function d	GELU	Approximate GELU
Attention type	MHA	Linear (Katharopoulos et al. 2020), FiSH (Nguyen et al. 2022)
Norm type	LayerNorm	None, RMSNorm, CRMSNorm, MAXNorm
Residual type	PreNorm	ReZero (Bachlechner et al. 2020)
Edge embedding d	128	
Max SPD	5	
Max centrality encoding	5	

D Kolmogorov-Smirnov Significance Matrix

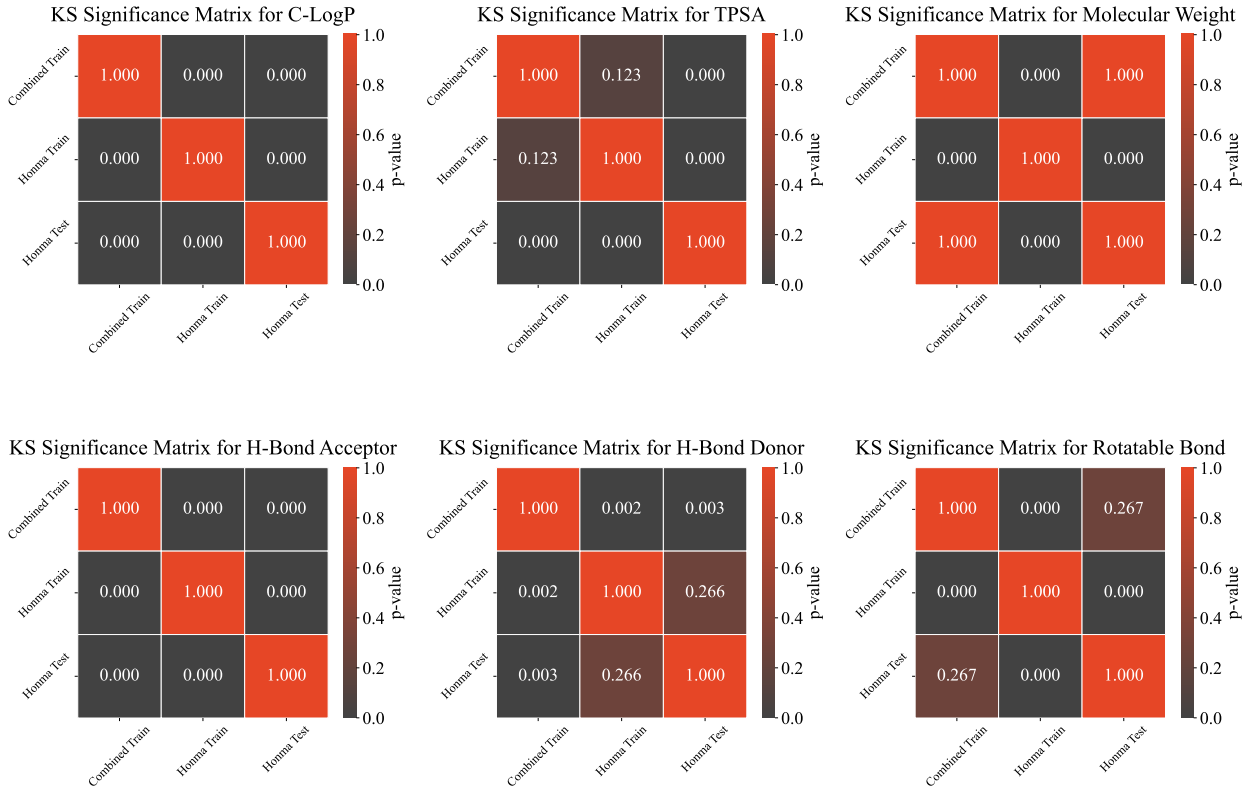


Figure 9: Matrix of Kolmogorov-Smirnov p-values with Holm-Bonferroni adjustments ($a = 0.008$). The matrices compare the distributions of phyicochemical properties between the Honma train, Combined train and Honma test datasets.

References

AICIS (2022). *Guide to categorising your chemical importation and manufacture - Step 4.4 Work out your human health hazard characteristics*. English. URL: https://www.industrialchemicals.gov.au/guide-categorising-your-chemical-importation-and-manufacture/step-4-work-out-your-introductions-risk-human-health/step-44-work-out-your-human-health-hazard-characteristics.

Akiba, Takuya et al. (July 2019). "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *arXiv e-prints*. _eprint: 1907.10902, arXiv:1907.10902. DOI: 10.48550/arXiv.1907.10902.

Ames, B. N., F. D. Lee, and W. E. Durston (Mar. 1973). "An improved bacterial test system for the detection and classification of mutagens and carcinogens." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 70.3. Place: United States, pp. 782–786. ISSN: 0027-8424 1091-6490. DOI: 10.1073/pnas.70.3.782.

Bachlechner, Thomas et al. (Mar. 2020). "ReZero is All You Need: Fast Convergence at Large Depth". In: *arXiv e-prints*. _eprint: 2003.04887, arXiv:2003.04887. DOI: 10.48550/arXiv.2003.04887.

Benigni, Romualdo et al. (Mar. 2013). "New perspectives in toxicological information management, and the role of ISSTOX databases in assessing chemical mutagenicity and carcinogenicity". In: *Mutagenesis* 28.4. _eprint: https://academic.oup.com/mutage/article-pdf/28/4/401/7428062/get016.pdf, pp. 401–409. ISSN: 0267-8357. DOI: 10.1093/mutage/get016. URL: https://doi.org/10.1093/mutage/get016.

Brodersen, Kay H et al. (2010). "The Balanced Accuracy and Its Posterior Distribution". eng. In: *2010 20th International Conference on Pattern Recognition*. ISSN: 1051-4651. IEEE, pp. 3121–3124. ISBN: 1-4244-7542-2.

Brown, Tom B. et al. (May 2020). "Language Models are Few-Shot Learners". In: *arXiv e-prints*. _eprint: 2005.14165, arXiv:2005.14165. DOI: 10.48550/arXiv.2005.14165.

Capecchi, Alice, Daniel Probst, and Jean-Louis Reymond (June 2020). "One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome". In: *Journal of Cheminformatics* 12.1, p. 43. ISSN: 1758-2946. DOI: 10.1186/s13321-020-00445-4.

Chengxuan, Ying et al. (2021). "Do Transformers Really Perform Bad for Graph Representation?" In: *arXiv.org*. ISSN: 2331-8422. DOI: 10.48550/arxiv.2106.05234.

Chu, Charmaine S.M. et al. (Dec. 2021). "Machine learning Predicting Ames mutagenicity of small molecules". In: *Journal of Molecular Graphics and Modelling* 109, p. 108011. ISSN: 1093-3263. DOI: 10.1016/j.jmgm.2021.108011. URL: https://www.sciencedirect.com/science/article/pii/S1093326321001820.

Devlin, Jacob et al. (Oct. 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv e-prints*. _eprint: 1810.04805, arXiv:1810.04805. DOI: 10.48550/arXiv.1810.04805.

Durant, Joseph L. et al. (2002). "Reoptimization of MDL Keys for Use in Drug Discovery". In: *Journal of Chemical Information and Computer Sciences* 42.6, pp. 1273–1280. ISSN: 0095-2338. DOI: 10.1021/ci010132r.

Elnaggar, Ahmed et al. (July 2020). "ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing". In: *arXiv e-prints*. _eprint: 2007.06225, arXiv:2007.06225. DOI: 10.48550/arXiv.2007.06225.

European Communities (2006). *Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC*. Tech. rep., pp. 1–849.

Feeney, Samuel V. et al. (2023). "Multiple Instance Learning Improves Ames Mutagenicity Prediction for Problematic Molecular Species". In: *Chemical research in toxicology* 36.8, pp. 1227–1237. ISSN: 0893-228X. DOI: 10.1021/acs.chemrestox.2c00372.

Fey, Matthias and Jan Eric Lenssen (Mar. 2019). "Fast Graph Representation Learning with PyTorch Geometric". In: *arXiv e-prints*, arXiv:1903.02428. DOI: 10.48550/arXiv.1903.02428. URL: https://ui.adsabs.harvard.edu/abs/2019arXiv190302428F.

Freedman, David and Persi Diaconis (Dec. 1981). "On the histogram as a density estimator:L2 theory". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57.4, pp. 453–476. ISSN: 1432-2064. DOI: 10.1007/BF01025868. URL: https://doi.org/10.1007/BF01025868.

French, Robert M. (1999). "Catastrophic forgetting in connectionist networks". In: *Trends in Cognitive Sciences* 3.4, pp. 128–135. ISSN: 1364-6613. DOI: https://doi.org/10.1016/S1364-6613(99)01294-2. URL: https://www.sciencedirect.com/science/article/pii/S1364661399012942.

Frey, Nathan C. et al. (Nov. 2023). "Neural scaling of deep chemical models". In: *Nature Machine Intelligence* 5.11, pp. 1297–1305. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00740-3. URL: https://doi.org/10.1038/s42256-023-00740-3.

Furuhama, A. et al. (2023). "Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR international challenge project". In: *SAR and QSAR in environmental research* 34.12, pp. 983–1001. ISSN: 1062-936X. DOI: 10.1080/1062936X.2023.2284902.

Guo, Chuan, Geoff Pleiss, et al. (June 2017). "On Calibration of Modern Neural Networks". In: *arXiv e-prints*. _eprint: 1706.04599, arXiv:1706.04599. DOI: 10.48550/arXiv.1706.04599.

Guo, Jinjiang, Qi Liu, et al. (Feb. 2022). "Ligandformer: A Graph Neural Network for Predicting Compound Property with Robust Interpretation". In: *arXiv e-prints*. _eprint: 2202.10873, arXiv:2202.10873. DOI: 10.48550/arXiv.2202.10873.

Guo, Zhichun, Chuxu Zhang, et al. (Feb. 2021). "Few-Shot Graph Learning for Molecular Property Prediction". In: *arXiv e-prints*. _eprint: 2102.07916, arXiv:2102.07916. DOI: 10.48550/arXiv.2102.07916.

Halevy, Alon, Peter Norvig, and Fernando Pereira (2009). "The Unreasonable Effectiveness of Data". In: *IEEE Intelligent Systems* 24.2, pp. 8–12. DOI: 10.1109/MIS.2009.36.

Han, Jason J. (Mar. 2023). "FDA Modernization Act 2.0 allows for alternatives to animal testing." eng. In: *Artificial organs* 47.3. Place: United States, pp. 449–450. ISSN: 1525-1594 0160-564X. DOI: 10.1111/aor.14503.

Hansen, Katja et al. (2009). "Benchmark data set for in silico prediction of Ames mutagenicity". In: *Journal of chemical information and modeling* 49.9, p. 2077. ISSN: 1549-960X. DOI: 10.1021/ci900161g.

Hestness, Joel et al. (Dec. 2017). "Deep Learning Scaling is Predictable, Empirically". In: *arXiv e-prints*. _eprint: 1712.00409, arXiv:1712.00409. DOI: 10.48550/arXiv.1712.00409.

Hicks, Steven A. et al. (Apr. 2022). "On evaluation metrics for medical applications of artificial intelligence." eng. In: *Scientific reports* 12.1. Place: England, p. 5979. ISSN: 2045-2322. DOI: 10.1038/s41598-022-09954-8.

Holm, Sture (1979). "A simple sequentially rejective multiple test procedure". In: *Scandinavian journal of statistics*. Publisher: JSTOR, pp. 65–70.

Honma, Masamitsu et al. (2019). "Improvement of quantitative structureactivity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project". In: *Mutagenesis* 34.1. Place: UK Publisher: Oxford University Press, pp. 3–16. ISSN: 0267-8357. DOI: 10.1093/mutage/gey031.

Hu, Weihua et al. (May 2020). "Open Graph Benchmark: Datasets for Machine Learning on Graphs". In: *arXiv e-prints*. _eprint: 2005.00687, arXiv:2005.00687. DOI: 10.48550/arXiv.2005.00687.

Hung, Chiakang and Giuseppina Gini (2021). "QSAR modeling without descriptors using graph convolutional neural networks: the case of mutagenicity prediction". In: *Molecular diversity* 25.3, pp. 1283–1299. ISSN: 1381-1991. DOI: 10.1007/s11030-021-10250-2.

ICH (2013). *ICH guideline S2 (R1) on genotoxicity testing and data interpretation for pharmaceuticals intended for human use*. Tech. rep. ICH.

– (2017). *ASSESSMENT AND CONTROL OF DNA REACTIVE (MUTAGENIC) IMPURITIES IN PHARMACEUTICALS TO LIMIT POTENTIAL CARCINOGENIC RISK*. Tech. rep. ICH.

Jin, Jiarui et al. (2022). "Refined Edge Usage of Graph Neural Networks for Edge Prediction". In: *arXiv.org*. ISSN: 2331-8422. DOI: 10.48550/arxiv.2212.12970.

Johnson, Justin M. and Taghi M. Khoshgoftaar (Mar. 2019). "Survey on deep learning with class imbalance". In: *Journal of Big Data* 6.1, p. 27. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0192-5. URL: https://doi.org/10.1186/s40537-019-0192-5.

Kamber, Markus et al. (May 2009). "Comparison of the Ames II and traditional Ames test responses with respect to mutagenicity, strain specificities, need for metabolism and correlation with rodent carcinogenicity". In: *Mutagenesis* 24.4. _eprint: https://academic.oup.com/mutage/article-pdf/24/4/359/3787533/gep017.pdf, pp. 359–366. ISSN: 0267-8357. DOI: 10.1093/mutage/gep017. URL: https://doi.org/10.1093/mutage/gep017.

Katharopoulos, Angelos et al. (June 2020). "Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention". In: *arXiv e-prints*. _eprint: 2006.16236, arXiv:2006.16236. DOI: 10.48550/arXiv.2006.16236.

Kirkpatrick, James et al. (Mar. 2017). "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the National Academy of Science* 114.13. _eprint: 1612.00796, pp. 3521–3526. DOI: 10.1073/pnas.1611835114.

Kriege, Nils M., Fredrik D. Johansson, and Christopher Morris (Mar. 2019). "A Survey on Graph Kernels". In: *arXiv e-prints*. _eprint: 1903.11835, arXiv:1903.11835. DOI: 10.48550/arXiv.1903.11835.

Landrum, Greg (2012). *RDKit - MACCS Keys Implementation*. Repository of Python3 Code.

Levin, D. E. et al. (Dec. 1982). "A new Salmonella tester strain (TA102) with A X T base pairs at the site of mutation detects oxidative mutagens." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 79.23. Place: United States, pp. 7445–7449. ISSN: 0027-8424 1091-6490. DOI: 10.1073/pnas.79.23.7445.

Li, Lisha, Kevin Jamieson, et al. (Mar. 2016). "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization". In: *arXiv e-prints*. _eprint: 1603.06560, arXiv:1603.06560. DOI: 10.48550/arXiv.1603.06560.

Li, Shimeng, Li Zhang, et al. (Mar. 2021). "MutagenPred-GCNNs: A Graph Convolutional Neural Network-Based Classification Model for Mutagenicity Prediction with Data-Driven Molecular Fingerprints". In: *Interdisciplinary sciences, computational life sciences* 13.1, pp. 25–33. ISSN: 1913-2751. DOI: 10.1007/s12539-020-00407-2. URL: https://doi.org/10.1007/s12539-020-00407-2.

Li, Ting, Zhichao Liu, et al. (Oct. 2023). "DeepAmes: A deep learning-powered Ames test predictive model with potential for regulatory application." eng. In: *Regulatory toxicology and pharmacology : RTP* 144. Place: Netherlands, p. 105486. ISSN: 1096-0295 0273-2300. DOI: 10.1016/j.yrtph.2023.105486.

Li, Zhizhong and Derek Hoiem (June 2016). "Learning without Forgetting". In: *arXiv e-prints*. _eprint: 1606.09282, arXiv:1606.09282. DOI: 10.48550/arXiv.1606.09282.

Lin, Tianyang et al. (June 2021). "A Survey of Transformers". In: *arXiv e-prints*. _eprint: 2106.04554, arXiv:2106.04554. DOI: 10.48550/arXiv.2106.04554.

Lipinski, C. A. et al. (Mar. 2001). "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings." eng. In: *Advanced drug delivery reviews* 46.1-3. Place: Netherlands, pp. 3–26. ISSN: 0169-409X. DOI: 10.1016/s0169-409x(00)00129-0.

Liu, Jie et al. (Nov. 2017). "Predicting Organ Toxicity Using in Vitro Bioactivity Data and Chemical Structure." eng. In: *Chemical research in toxicology* 30.11. Place: United States, pp. 2046–2059. ISSN: 1520-5010 0893-228X. DOI: 10.1021/acs.chemrestox.7b00084.

Lui, Raymond, Davy Guan, and Slade Matthews (2023). "Mechanistic Task Groupings Enhance Multitask Deep Learning of Strain-Specific Ames Mutagenicity". In: *Chemical research in toxicology* 36.8, pp. 1248–1254. ISSN: 0893-228X. DOI: 10.1021/acs.chemrestox.2c00385.

Ma, Liheng et al. (May 2023). "Graph Inductive Biases in Transformers without Message Passing". In: *arXiv e-prints*. _eprint: 2305.17589, arXiv:2305.17589. DOI: 10.48550/arXiv.2305.17589.

Madia, Federica et al. (2020). "EURL ECVAM Genotoxicity and Carcinogenicity Database of Substances Eliciting Negative Results in the Ames Test: Construction of the Database". In: *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* 854-855, p. 503199. ISSN: 1383-5718. DOI: https://doi.org/10.1016/j.mrgentox.2020.503199. URL: https://www.sciencedirect.com/science/article/pii/S1383571820300693.

Maron, Dorothy M. and Bruce N. Ames (1983). "Revised methods for the Salmonella mutagenicity test". In: *Mutation Research/Environmental Mutagenesis and Related Subjects* 113.3, pp. 173–215. ISSN: 0165-1161. DOI: 10.1016/0165-1161(83)90010-9.

Marshall, S. F. et al. (Mar. 2016). "Good Practices in Model-Informed Drug Discovery and Development: Practice, Application, and Documentation." eng. In: *CPT: pharmacometrics & systems pharmacology* 5.3. Place: United States, pp. 93–122. ISSN: 2163-8306. DOI: 10.1002/psp4.12049.

McCandlish, Sam et al. (2018). *An Empirical Model of Large-Batch Training*. _eprint: 1812.06162. URL: https://arxiv.org/abs/1812.06162.

Menegaux, Romain et al. (Apr. 2023). "Self-Attention in Colors: Another Take on Encoding Graph Structure in Transformers". In: *arXiv e-prints*. _eprint: 2304.10933, arXiv:2304.10933. DOI: 10.48550/arXiv.2304.10933.

Nguyen, Tan et al. (2022). "Improving Transformer with an Admixture of Attention Heads". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 27937–27952. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b2e4edd53059e24002a0c916d75cc9a3-Paper-Conference.pdf.

Niculescu-Mizil, Alexandru and Rich Caruana (2005). "Predicting good probabilities with supervised learning". In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML '05. event-place: Bonn, Germany. New York, NY, USA: Association for Computing Machinery, pp. 625–632. ISBN: 1-59593-180-5. DOI: 10.1145/1102351.1102430. URL: https://doi.org/10.1145/1102351.1102430.

OECD (2014). *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. Type: doi:https://doi.org/10.1787/9789264085442-en. URL: https://www.oecd-ilibrary.org/content/publication/9789264085442-en.

Pakdaman Naeini, Mahdi, Gregory Cooper, and Milos Hauskrecht (Feb. 2015). "Obtaining Well Calibrated Probabilities Using Bayesian Binning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 29.1. DOI: 10.1609/aaai.v29i1.9602. URL: https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Parmar, Niki et al. (Feb. 2018). "Image Transformer". In: *arXiv e-prints*. _eprint: 1802.05751, arXiv:1802.05751. DOI: 10.48550/arXiv.1802.05751.

Paszke, Adam et al. (Dec. 2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *arXiv e-prints*, arXiv:1912.01703. DOI: 10.48550/arXiv.1912.01703. URL: https://ui.adsabs.harvard.edu/abs/2019arXiv191201703P.

Patlewicz, G. et al. (2010). "Can mutagenicity information be useful in an Integrated Testing Strategy (ITS) for skin sensitization?" In: *SAR and QSAR in environmental research* 21.7-8, pp. 619–656. ISSN: 1062-936X. DOI: 10.1080/1062936X.2010.528447.

Platt, John (June 2000). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Adv. Large Margin Classif.* 10.

Provost, Foster, Tom Fawcett, and Ron Kohavi (Apr. 2001). "The Case Against Accuracy Estimation for Comparing Induction Algorithms". In: *Proceedings of the Fifteenth International Conference on Machine Learning*.

Rácz, Anita, Dávid Bajusz, and Károly Héberger (Feb. 2021). "Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification." eng. In: *Molecules (Basel, Switzerland)* 26.4. Place: Switzerland. ISSN: 1420-3049. DOI: 10.3390/molecules26041111.

Radford, Alec et al. (Feb. 2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *arXiv e-prints*. _eprint: 2103.00020, arXiv:2103.00020. DOI: 10.48550/arXiv.2103.00020.

Rampáek, Ladislav et al. (May 2022). "Recipe for a General, Powerful, Scalable Graph Transformer". In: *arXiv e-prints*. _eprint: 2205.12454, arXiv:2205.12454. DOI: 10.48550/arXiv.2205.12454.

Rogers, David and Mathew Hahn (May 2010). "Extended-Connectivity Fingerprints". In: *Journal of chemical information and modeling* 50.5, pp. 742–754. ISSN: 1549-9596. DOI: 10.1021/ci100050t.

Scarselli, F. et al. (2009). "The Graph Neural Network Model". In: *IEEE transactions on neural networks* 20.1, pp. 61–80. ISSN: 1045-9227. DOI: 10.1109/TNN.2008.2005605.

Seo, Myungwon et al. (Jan. 2020). "Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP) for natural product-based drug development". In: *Journal of Cheminformatics* 12.1, p. 6. ISSN: 1758-2946. DOI: 10.1186/s13321-020-0410-3.

Shi, Tingting, Yingwu Yang, et al. (2019). "Molecular image-based convolutional neural network for the prediction of ADMET properties". In: *Chemometrics and intelligent laboratory systems* 194, p. 103853. ISSN: 0169-7439. DOI: 10.1016/j.chemolab.2019.103853.

Shi, Yu, Shuxin Zheng, et al. (Mar. 2022). "Benchmarking Graphormer on Large-Scale Molecular Modeling Datasets". In: arXiv:2203.04810. DOI: 10.48550/arXiv.2203.04810.

Smith, Leslie N. and Nicholay Topin (Aug. 2017). "Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates". In: *arXiv e-prints*. _eprint: 1708.07120, arXiv:1708.07120. DOI: 10.48550/arXiv.1708.07120.

Subramanian, Shreyas and Vignesh Ganapathiraman (2023). "Zeroth order GreedyLR: An adaptive learning rate scheduler for deep neural network training". In: *PRML 2023*. URL: https://www.amazon.science/publications/zeroth-order-greedylr-an-adaptive-learning-rate-scheduler-for-deep-neural-network-training.

Tintó-Moliner, A. and M. Martin (2020). "Quantitative weight of evidence method for combining predictions of quantitative structure-activity relationship models". In: *SAR and QSAR in environmental research* 31.4, pp. 261–279. ISSN: 1062-936X. DOI: 10.1080/1062936X.2020.1725116.

Tran, Thi Tuyet Van, Hilal Tayara, and Kil To Chong (2024). "AMPred-CNN: Ames mutagenicity prediction model based on convolutional neural networks". In: *Computers in Biology and Medicine* 176, p. 108560. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2024.108560. URL: https://www.sciencedirect.com/science/article/pii/S0010482524006449.

Tropsha, Alexander (July 2010). "Best Practices for QSAR Model Development, Validation, and Exploitation." eng. In: *Molecular informatics* 29.6-7. Place: Germany, pp. 476–488. ISSN: 1868-1743. DOI: 10.1002/minf.201000061.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *arXiv.org*. ISSN: 2331-8422. DOI: 10.48550/arxiv.1706.03762.

Votano, Joseph R. et al. (2004). "Three new consensus QSAR models for the prediction of Ames genotoxicity". eng. In: *Mutagenesis* 19.5. Place: Oxford Publisher: Oxford University Press, pp. 365–377. ISSN: 0267-8357.

Xiong, Ruibin et al. (Feb. 2020). "On Layer Normalization in the Transformer Architecture". In: *arXiv e-prints*. _eprint: 2002.04745, arXiv:2002.04745. DOI: 10.48550/arXiv.2002.04745.

Xu, Congying et al. (2012). "In silico Prediction of Chemical Ames Mutagenicity". eng. In: *Journal of chemical information and modeling* 52.11. Place: Washington, DC Publisher: American Chemical Society, pp. 2840–2847. ISSN: 1549-9596.

Yuan, Zhuoning et al. (Dec. 2020). "Large-scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification". In: *arXiv e-prints*. _eprint: 2012.03173, arXiv:2012.03173. DOI: 10.48550/arXiv.2012.03173.

Zadrozny, Bianca and Charles Elkan (2002). "Transforming classifier scores into accurate multiclass probability estimates". In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '02. event-place: Edmonton, Alberta, Canada. New York, NY, USA: Association for Computing Machinery, pp. 694–699. ISBN: 1-58113-567-X. DOI: 10.1145/775047.775151. URL: https://doi.org/10.1145/775047.775151.