Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 3515

Advancing mutagenicity predictions in drug discovery with an explainable few-shot deep learning framework

Luis H. M. Torres, *[a] Sofia M. da Silva,[b] Joel P. Arrais,[a] Catarina Pimentel[b] and Bernardete Ribeiro[a]

The Ames mutagenicity test serves as a cornerstone for evaluating the mutagenic potential of chemical compounds, which is critical in drug discovery and safety assessments. However, existing computational methods struggle to utilize the contribution of individual bacterial strains used in the Ames test, limiting the accuracy of overall mutagenicity predictions. To address this, we introduce Meta-GTMP, a few-shot learning framework that combines graph neural networks (GNNs) and Transformers to integrate the local molecular graph structure with the global information in graph embedding representations for mutagenicity prediction using limited labeled data. A multi-task meta-learning strategy further optimizes the model parameters across individual strain-specific few-shot tasks, leveraging their complementarity to predict the overall Ames result. Computational experiments conducted on the ISSSTY v1-a dataset demonstrate that Meta-GTMP outperforms standard graph-based models, achieving notable improvements in sensitivity (+6.82%) and ROC-AUC score (+2.50%). Laboratory validation tests using six chemically diverse compounds with unknown mutagenicity labels confirmed the model's effectiveness, achieving high accuracy in distinguishing mutagenic and non-mutagenic samples. Importantly, Meta-GTMP makes explainable predictions through a node-edge attribute masking strategy, identifying significant molecular substructures responsible for mutagenicity. These insights are essential in drug discovery, positioning Meta-GTMP as a robust and explainable tool for using mutagenicity predictions to enhance the identification, selection and rational design of safer and more effective potential drug candidates.

Received 23rd June 2025
Accepted 30th October 2025

DOI: 10.1039/d5dd00276a

rsc.li/digitaldiscovery

1 Introduction

The integrity of the genetic material in cells can be compromised by a variety of chemical agents. This phenomenon, known as genotoxicity is a crucial aspect to consider when evaluating the safety of new chemical entities and potential drug candidates. In drug discovery, strict regulations regarding genotoxicity ensure the safe usage of new and already existing substances.[1] One effective approach to assess the risk of genotoxicity relies on the Ames mutagenicity test, which measures the ability of chemical compounds to induce genetic mutations in DNA. The Ames test is an experimental *in vitro* assay designed to detect genetic mutations induced by a given compound across various strains of bacteria. It serves as a preliminary screening tool in drug discovery to estimate the mutagenicity of drug candidates and can be valuable in the regulatory process prior to the compound registration and approval.[2,3]

The primary concept of the Ames test is to identify chemical compounds capable of causing DNA mutations that revert the inability of certain amino acid dependent bacteria to survive and grow without supplementation of these amino acids, by regaining the ability to synthesize them.[2] Typically, a minimum of five different histidine-dependent *Salmonella typhimurium* strains are used for the Ames test. For four of these strains (TA98, TA100, TA1535, and TA1537, TA97 or TA97a), growth in the absence of histidine is restored upon base-pair substitution or frameshift mutations in the histidine genetic marker. However, these strains have limited sensitivity in detecting certain types of mutagens, such as oxidants and cross-linking compounds.[4] For this reason, another strain (the histidine-dependent *S. typhimurium* strain TA102 or the tryptophan-dependent *Escherichia coli* WP2) is also included in the test.[5] Mutations in the histidine or tryptophan markers of the strains TA102 or *E. coli* WP2, respectively, can be reverted by transitions or transversions.[6] According to the OECD guidelines, the Ames test is considered positive and a compound mutagenic if it causes a significant increase in revertant colonies in at least one

[a]*Univ Coimbra, Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, 3030-790 Coimbra, Portugal. E-mail: luistorres@dei.uc.pt*

[b]*Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, 2780-157, Portugal*

of the bacterial strains. A negative result, indicative of a non-mutagenic compound, requires no significant increase across tested doses for all the bacterial strains.[4] The result may also be inconclusive if no clear positive or negative response is obtained for individual strains.[7]

Recent research in the field of computational toxicology has been focused on the development of quantitative structure–activity relationship (QSAR) models to predict the mutagenic properties of chemical compounds[8–10] including several machine learning (ML) approaches.[11–14] Nonetheless, the lack of structure–activity relationships and the limited access to high quality labeled data are holding back computational genotoxicity from reaching higher predictive levels.[15] In addition, the individual impact of different bacterial strains used in the Ames test has not been extensively studied. Hence, most *in silico* methods only consider the overall Ames result for small drug repositories, overlooking the outcomes of individual experiments conducted for each strain.[16–18] To bridge this gap, molecular property predictors have introduced multi-task learning strategies such as few-shot learning or meta-learning to model multi-target properties with limited data.[19–21]

In the last decade, deep learning (DL) methods have gained prominence as valuable methods for QSAR modeling and molecular property prediction in drug discovery.[22,23] This trend can be attributed to the ability of DL to model complex and high-dimensional non-linear functions, and advancements that have made DL methods more robust to low-data environments and able to generalize for challenging predictive tasks.[24] Few-shot learning has shown a great potential as a data-efficient strategy used to adapt DL methods in scenarios where data is limited and with high class imbalance, without having a negative impact on performance. This meta-learning approach[25] has proved to be pivotal in QSAR studies to predict multiple chemical properties across few-shot tasks with few labeled compounds.[26–28] However, the development of data-driven DL models to predict mutagenic properties and replicate the Ames test experiment still remains a key area for improvement in drug discovery.

Compounds can be represented by molecular graphs, with atoms described as a set of nodes and chemical bonds as edges.[29] Graph-based DL methods such as graph neural networks (GNNs) use molecular graphs to learn node-edge representations using neighborhood aggregation and to generate graph-level embeddings for molecular property prediction.[30–33] Although GNNs can discriminate local information, they encounter difficulties in capturing the long-range dependencies important for compound classification. In computational drug discovery, Transformer networks account for the global-semantic structure within molecular embeddings and preserve the long-range structural information.[34–36] Recent advancements in molecular property discovery have led to the development of hybrid DL approaches including graph-based networks using Transformers.[37–39] These methods hold great potential for molecular representation learning to predict the chemical properties of drug candidates.

The research question we propose to answer is whether it is possible to adapt these DL methods using a few-shot meta-learning framework to model the Ames test and predict the mutagenic properties of chemical compounds with limited data. To address this challenge, a few-shot GNN Transformer model, Meta-GTMP is proposed to explore the local and the global information within molecular graph embeddings for Ames mutagenicity prediction, as depicted in Fig. 1A. Meta-GTMP consists in a few-shot meta-learning framework that jointly learns across different few-shot tasks specific for each bacterial strain involved in the Ames test. The proposed approach leverages the complementarity among the five different Ames test strains to iteratively update model parameters and predict the overall Ames mutagenicity result with just a few labeled compounds, as shown in Fig. 1B. In this work, we conduct a performance comparison with the standard graph-based methods in Ames mutagenicity prediction by removing the Transformer component of the Meta-GTMP framework. The results obtained show that Meta-GTMP achieves the best overall performance compared to the graph-based baselines, while providing explainable insights into the mutagenicity predictions using a node-edge attribute masking strategy. In this study, we sequentially masked individual atoms (nodes) and chemical bonds (edges) in each molecule to observe the changes in Meta-GTMP predictions and highlight the key molecular substructures that are likely to influence their mutagenic and non-mutagenic properties. Finally, to validate the computational results, laboratory experiments were conducted using six candidate compounds encompassing a diverse range of chemical structures, each with unknown mutagenic labels. This experimental validation step was crucial in reaffirming the model's ability to accurately distinguish between mutagenic and non-mutagenic samples. The results of these experiments unequivocally demonstrated the effectiveness of Meta-GTMP in identifying compounds with mutagenic potential.

1.1 Problem formulation

In Ames mutagenicity prediction, the main objective is to determine whether the individual results obtained across each one of the five Ames bacterial strains produce a mutagenic or non-mutagenic overall result. In this specific problem, chemical compounds are classified considering six different mutagenicity labels, one for each bacterial strain and one for the overall Ames result. Typically, DL-based approaches describe input molecules using Simplified Molecular Input Line Entry System (SMILES) representations, which consist in 1D sequences describing individual atom and chemical bond information within a drug compound. Nonetheless, since the 1D sequence is not a natural representation to describe the spatial relation among atoms *via* chemical bonds, some important structural information of drugs can be lost, degrading the predictive performance. Therefore, the Meta-GTMP model converts input SMILES into a more complex representation, a 2D molecular graph. To define molecular graphs, we use the notation $G = (V, E)$, where V refers to a set of nodes and E describes the set of edges. Edges are denoted by $e = (v, u)$, where v and u are adjacent nodes connected in a neighborhood $N(v)$ with $u \in N(v)$ (see

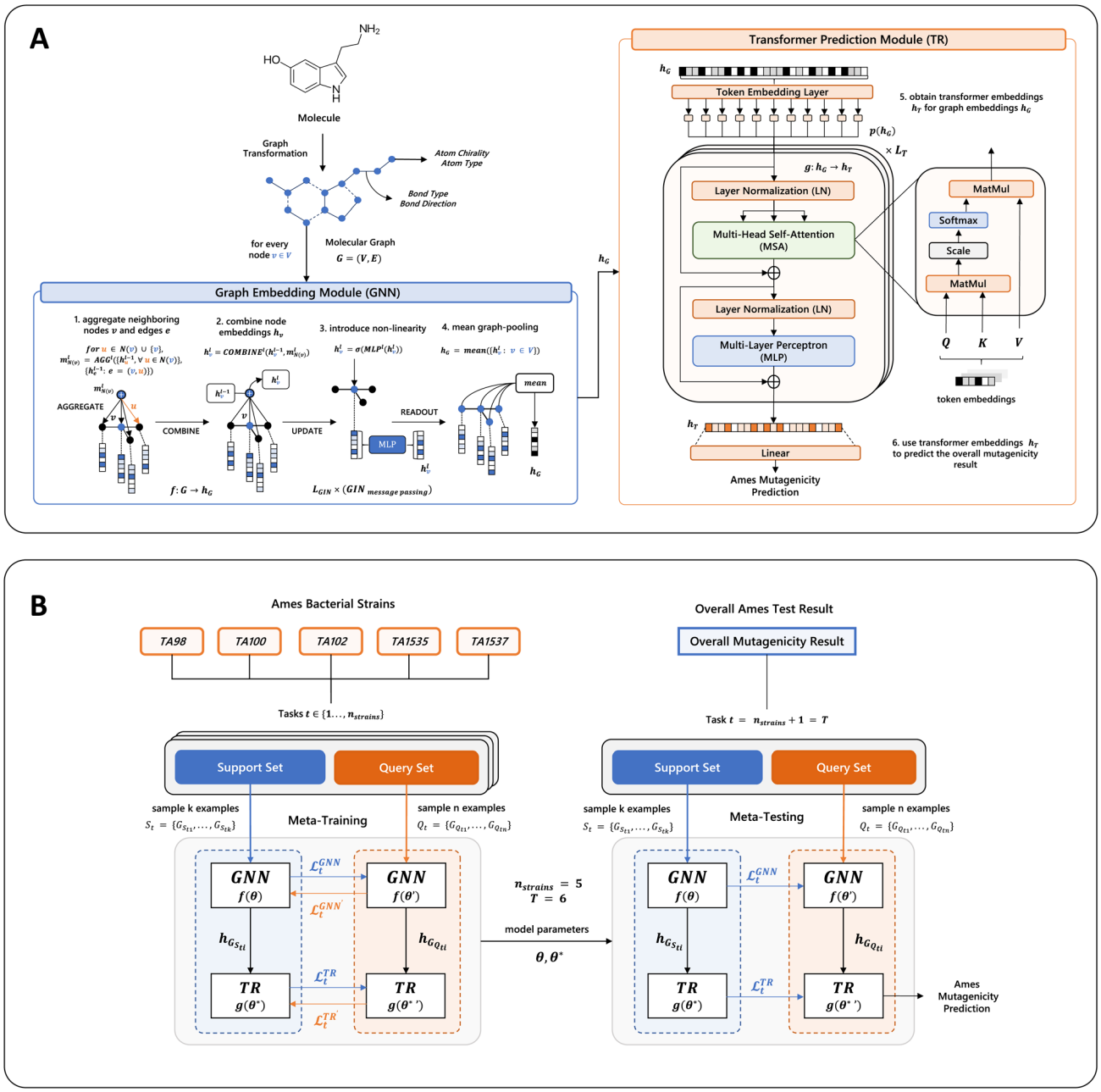

Fig. 1 Overview of Meta-GTMP. (A) Depiction of the GNN-Transformer for Ames mutagenicity prediction. Molecules are described by molecular graphs $G = (V, E)$, where V is the set of nodes and E the set of edges. Edges are represented by $e = (v, u)$, where v and u are adjacent nodes in a neighborhood $N(v)$. Input node and edge features include specific molecular attributes: atom number, atom chirality and bond type, bond direction. Molecular graphs G serve as input to a graph isomorphism network (GIN) with $L_{GIN} = 5$ layers to generate graph embeddings h_G. The GNN iteratively computes the AGGREGATE, COMBINE, and UPDATE steps, performed simultaneously for all nodes $v \in V$. At the final GNN layer L_{GIN}, graph embeddings h_G are obtained with a mean-pooling READOUT operation. The Transformer takes graph embeddings h_G as 1D feature vectors with an embedding size of 300, which are converted into a sequence of patch tokens x_p. Transformer embeddings h_T obtained by the linear projection of individual patches are propagated across multi-head self-attention (MSA) layers. Transformer blocks include MSA layers followed by MLP, preceded by layer normalization (LN) and residual connections. A linear layer followed by sigmoid activation uses the output Transformer embedding h_T to obtain the prediction for each strain and for the overall Ames mutagenicity (condensed in a value $\in \{0, 1\}$). (B) Graphical schematic of the few-shot meta-learning framework for Ames mutagenicity prediction. The two-module framework is composed by two distinct parts: a GNN module f and a Transformer (TR) module g with parameters θ and θ', respectively. For each bacterial strain involved in the Ames test, few-shot tasks t consist of a random task-specific support set S_t with a set of positive samples k_+ and negative samples k_- provided for training, and the remaining n data points for each task t are used as a disjoint query set Q_t provided for evaluation. The Meta-GTMP model considers five different meta-training tasks for each bacterial strain used in the Ames test and leverages their complementarity in a joint learning procedure to predict the overall mutagenicity in a meta-testing task. The updated parameters from meta-training are used to initialize Meta-GTMP and predict the overall mutagenicity result in a final meta-testing task using a random support set of size (k_+, k_-) for k-shot experiments with limited data.

Fig. 1A). The input features for nodes (atoms) and edges (bonds) (h_v^0, h_e^0) in molecular graphs G are: atom number, atom chirality and bond type, bond direction. In this work, we focus on predicting the Ames mutagenicity of a limited amount of chemical compounds, so that $\{f_\theta(G), g_{\theta^*}(h_G)\} : S \Rightarrow \{0,1\} \in Y$, where S is the space of molecular graphs G, h_G are the output graph embeddings from a GNN f_θ, g_{θ^*} is a Transformer (TR), and Y are the mutagenicity labels for each bacterial strain and for the overall Ames mutagenicity result.

1.2 Meta-GTMP framework

A few-shot meta-learning framework was developed to optimize a GNN-Transformer architecture and leverage the contribution of multiple Ames bacterial strains for Ames mutagenicity prediction. This meta-learning framework is used to optimize two neural network models: a graph isomorphism network (GNN) f_θ with parameters θ and a Transformer (TR) g_{θ^*} with parameters θ*. Both models are trained using a meta-learning framework across few-shot tasks $t \in \{1\ldots, n_{\text{strains}}\}$ for each bacterial strain used in the Ames test, with $n_{\text{strains}} = 5$. In meta-training, models f_θ and g_{θ^*} are trained across few-shot tasks using a random support set S_t of molecular graphs $G_{S_{t_i}}$ and evaluated on a disjoint query set Q_t of molecular graphs $G_{Q_{t_j}}$. In meta-testing, the updated parameters are used to initialize the GNN-Transformer model and generalize to new chemical compounds and predict the overall Ames mutagenicity result. This strategy leverages the complementarity among five different bacterial strains involved in the Ames test by the means of integrating information of individual predictive tasks with a joint learning procedure to infer the overall Ames mutagenicity result with a limited number of chemical compounds (see Fig. 1A and B). More details of the Meta-GTMP framework are provided in the Methods and in the Supplementary material sections.

1.3 Mutagenicity data and evaluation strategy

Data was collected using a public compound repository known as ISSSTY v1-a, which includes publicly available information on the mutagenic effects of chemical compounds on different bacterial strains in Ames test experiments. This dataset, which includes information on the mutagenicity of 7367 compounds, was collected and organized by the Istituto Superiore di Sanità (ISS).[40,41] Compounds are classified according to their mutagenic or non-mutagenic activity for five different target properties corresponding to each one of the bacterial strains involved in the Ames mutagenicity test and for a final target property corresponding to the overall Ames mutagenicity result. In this work, we evaluate the ability to perform the binary classification of compounds using a k-shot meta-learning framework taking into account five meta-training tasks for each one of the bacterial strains used in the Ames test and one final meta-testing task for the overall Ames result. In meta-training, a support set of size k is randomly sampled to serve as an input to the Meta-GTMP model and update model parameters for each bacterial strain: {TA98, TA100, TA102 (WP2), TA1535, and TA1537} across few-shot tasks $t \in \{1\ldots, n_{\text{strains}}\}$, with $n_{\text{strains}} = 5$. Then, both models are used to evaluate the mutagenicity of a disjoint query set of compounds using the remaining n samples for that task with a few gradient descent steps. In meta-testing, a support set of k examples is randomly sampled for the overall Ames test task $t = n_{\text{strains}} + 1 = T = 6$ and models are initialized using the updated parameters from meta-training. Next, both models are evaluated using a disjoint query set of new compounds with the remaining n samples for this final test task, to predict the overall Ames mutagenicity result (see Fig. 1B). ROC-AUC is used as the major metric to evaluate model performance translating the ability of Meta-GTMP to predict the Ames mutagenicity and correctly classify compounds in imbalanced scenarios. In addition, we also report the Sensitivity (Sn), Specificity (Sp), Precision (Pr), Accuracy (Acc) and F1 score (F1s). Sn and Sp evaluate the ability to identify mutagenic and non-mutagenic compounds, respectively. Pr is the proportion of correctly predicted mutagenic compounds and Acc is the percentage of correct predictions. F1s is the harmonic mean of Sn and Pr. We conduct k-shot experiments with random support sets of size (5 +,5−) and (10 +,10 −) for $k = 5$ (5-shot) and $k = 10$ (10-shot), respectively. These computational experiments are repeated 30 times, using random support sets each time, to obtain a robust estimate of Meta-GTMP performance. To address class imbalance, we implement a weighted binary cross-entropy loss to avoid majority-class bias during training and improves performance based on a imbalanced set of strain-specific tasks (see Methods section). More details about the data and performance metrics are provided in the SI Material section.

2 Results

2.1 Computational results of Meta-GTMP and graph-based baselines

To provide a comparative baseline, we report the computational results of Meta-GTMP alongside GNN baseline methods, including a graph isomorphism network (GIN),[20,42] a graph convolutional network (GCN)[43] and GraphSAGE.[44] The GNN baselines are derived by removing the Transformer component of Meta-GTMP, retaining the core GNN structure. All models are pre-trained using GNN models of Hu *et al.* (2020)[30] for improved initialization. In addition, the GNN baselines also use a few-shot meta-learning framework and standard binary cross-entropy loss. All models take into account the positive and negative samples for each strain and for the overall Ames result. Here, we consider $n_{\text{strains}} = 5$ few-shot tasks for meta-training for each Ames bacterial strain: {TA98, TA100, TA102, TA1535, and TA1537} and 1 final task in meta-testing to predict the overall mutagenicity result. Tables 1 and 2 show the mean and standard deviations of performance results obtained by Meta-GTMP and GNN baselines across 30 experiments with (5 +,5 −) (5-shot) and (10 +,10 −) (10-shot) random support sets on the overall test task. The Δ (metric) column shows the difference in results of Meta-GTMP and the best GNN baseline for each performance metric. The scatter plots overlaid with boxplots in Fig. 2 show the average metric scores and standard deviations in

Table 1 Average metric scores obtained across 30 computational experiments in the 5-shot setting considering all five strains (All-strains) for the binary classification on the overall Ames test task. All models are trained and tested using a few-shot meta-learning framework in the 5-shot setting

5-Shot (5 +,5 −)					
Metric	GIN	GCN	GraphSAGE	Meta-GTMP (GIN + TR)	Δ (metric)
Specificity (Sp)	0.9997 ± 0.0011	0.9234 ± 0.0615	0.9703 ± 0.0291	0.9815 ± 0.0098	−0.0182
Sensitivity (Sn)	0.8444 ± 0.0152	0.7912 ± 0.0481	0.7377 ± 0.0369	0.9126 ± 0.0106	+0.0682
Precision (Pr)	0.9999 ± 0.0001	0.9933 ± 0.0049	0.9972 ± 0.0025	0.9985 ± 0.0007	−0.0014
Accuracy (Acc)	0.8550 ± 0.0141	0.8002 ± 0.0412	0.7535 ± 0.0331	0.9173 ± 0.0092	+0.0623
F1-score (F1s)	0.9156 ± 0.0089	0.8799 ± 0.0279	0.8475 ± 0.0235	0.9536 ± 0.0054	+0.0380
ROC-AUC	0.9221 ± 0.0073	0.8573 ± 0.0150	0.8540 ± 0.0142	**0.9471 ± 0.0021**	+0.0250

Table 2 Average metric scores obtained across 30 computational experiments in the 10-shot setting considering all five strains (All-strains) for the binary classification on the overall Ames test task. All models are trained and tested using a few-shot meta-learning framework in the 10-shot setting

10-Shot (10 +,10 −)					
Metric	GIN	GCN	GraphSAGE	Meta-GTMP (GIN + TR)	Δ (metric)
Specificity (Sp)	0.9923 ± 0.0079	0.8931 ± 0.0621	0.9475 ± 0.0486	0.9854 ± 0.0133	−0.0069
Sensitivity (Sn)	0.8513 ± 0.0190	0.8197 ± 0.0449	0.7659 ± 0.0395	0.9038 ± 0.0109	+0.0525
Precision (Pr)	0.9994 ± 0.0007	0.9910 ± 0.0048	0.9953 ± 0.0040	0.9989 ± 0.0010	−0.0005
Accuracy (Acc)	0.8607 ± 0.0176	0.8246 ± 0.0387	0.7781 ± 0.0342	0.9093 ± 0.0095	+0.0486
F1-score (F1s)	0.9193 ± 0.0110	0.8965 ± 0.0259	0.8651 ± 0.0237	0.9489 ± 0.0056	+0.0296
ROC-AUC	0.9218 ± 0.0096	0.8564 ± 0.0184	0.8567 ± 0.0136	**0.9446 ± 0.0035**	+0.0228

5-shot (Fig. 2A) and 10-shot (Fig. 2B) settings on the final overall test task.

Meta-GTMP demonstrates a significant and consistent improvement over all GNN baselines, in particular over the GIN model, across multiple performance metrics in the 5-shot and 10-shot settings. These results underscore the accuracy and robustness of Meta-GTMP, especially in the context of few-shot learning with the limited data available and high class imbalance of the Ames dataset. Meta-GTMP achieves substantial improvements for different metrics. In terms of Sensitivity (Sn), it outperforms GIN by +6.82% in the 5-shot setting and +5.25% in the 10-shot setting. Sn is a crucial metric for mutagenicity prediction, as it minimizes the false negatives, ensuring the accurate identification of mutagenic compounds. In addition, Meta-GTMP shows improvements of +6.23% and +4.86% in Accuracy (Acc) across 5-shot and 10-shot settings, outperforming all GNN baselines in making correct predictions. The model also exceeds GIN in F1-score (F1s), with improvements of +3.80% in the 5-shot and +2.96% in the 10-shot settings, highlighting its balanced performance between precision (Pr) and recall. Overall, the key improvements in ROC-AUC scores, by +2.50% in 5-shot and +2.28% in 10-shot experiments, demonstrate the superior performance of Meta-GTMP in discriminating between the mutagenic and non-mutagenic compounds with limited and highly imbalanced data. The robustness of Meta-GTMP predictions is further demonstrated by significantly lower variances across performance metrics, including Sn and ROC-AUC, compared to the GNN baselines. In the 5-shot setting, the Meta-GTMP standard deviation in Sn is 0.0106 compared to 0.0152 for GIN, and for ROC-AUC is 0.0021 compared to 0.0073 for GIN. Similar trends are observed in the 10-shot setting, with Meta-GTMP exhibiting smaller standard deviations across metrics, such as Sn (0.0109 *vs.* 0.0190 for GIN) and ROC-AUC (0.0035 *vs.* 0.0096 for GIN).

The Ames dataset presents notable challenges due to this highly imbalanced positive-negative class distribution across the five bacterial strains and for the final overall Ames result (see Table 7). The five bacterial strains used for meta-training exhibit a predominance of negative (non-mutagenic) samples over the positive (mutagenic) samples. For instance, strain TA98 contains 2782 negatives compared to 1676 positives, strain TA1535 has 2103 negatives *versus* 436 positives, and strain TA1537 has 1779 negatives compared to the 365 positives. In the overall Ames meta-testing task, while the class imbalance persists, there are 3103 positives and only 231 negative samples. As a result, the GIN model as well as other GNN baselines performances reflect the inability to address the issue of class imbalance effectively during meta-training and meta-testing stages. The binary cross-entropy loss does not penalize errors on the minority class (positive), while the errors on the majority class (negative) are overemphasized across meta-training tasks. This imbalance causes GIN to overfit to the negative class, leading to an increased Specificity (Sp) and Pr at the expense of the Sn, Acc and ROC-AUC score in the final meta-testing task. Consequently, GIN misclassifies the mutagenic compounds as non-mutagenic and fails to minimize the false negatives, reducing its practical utility in real-world drug discovery

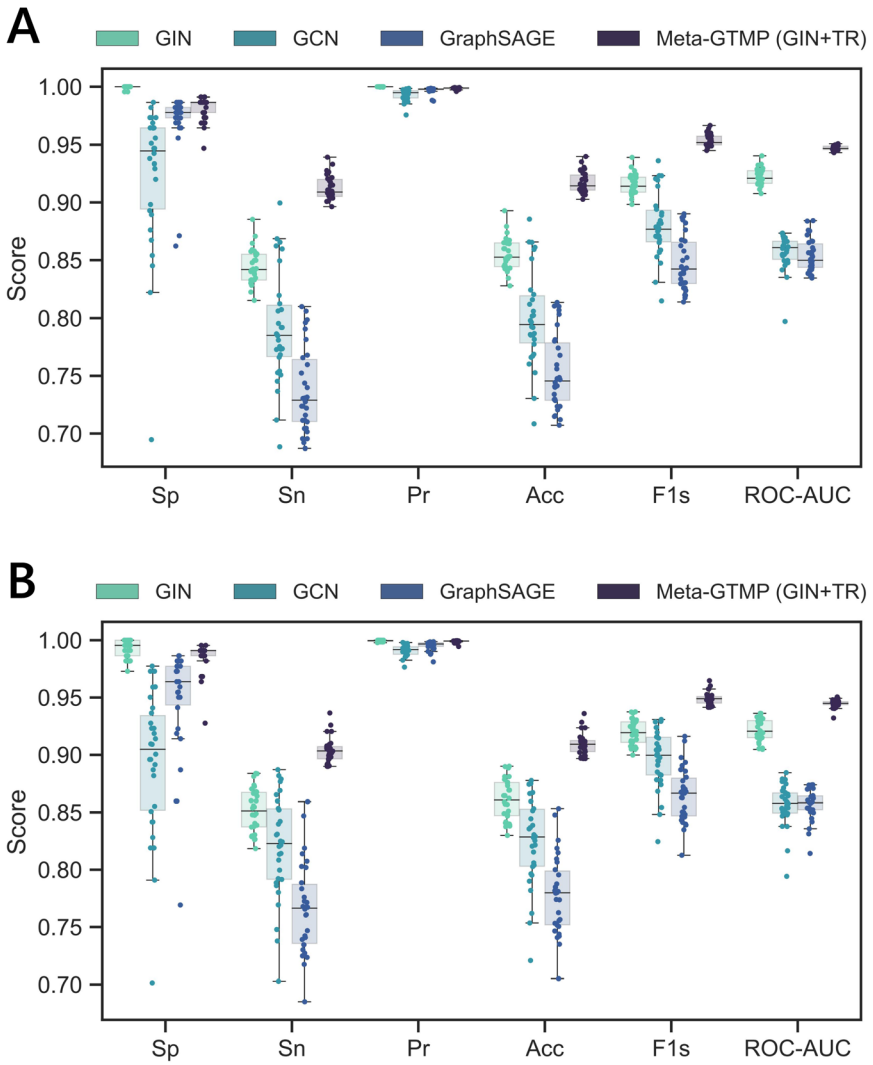

Fig. 2 Analysis of Meta-GTMP performance across 30 few-shot experiments with random support sets for Ames mutagenicity prediction. (A) Scatter plots overlaid with boxplots showing the comparison of metric scores obtained by the Meta-GTMP model and graph-based baselines in 5-shot experiments with 30 random support sets of size (5 +,5 −). (B) Scatter plots overlaid with boxplots showing the comparison of metric scores obtained by the Meta-GTMP model and graph-based baselines in 10-shot experiments with 30 random support sets of size (10 +,10 −).

scenarios where the accurate prediction of mutagenic compounds is critical.

Meta-GTMP addresses these limitations by considering a weighted binary cross-entropy loss (see the Methods section) that assigns higher penalties to the errors on the minority class. Hence, this approach ensures a balanced learning across meta-training tasks to effectively generalize for the overall Ames meta-testing task. In real-world mutagenicity prediction, the sensitivity is often critical because failing to identify a mutagenic compound (false negatives) poses significant risks that can lead to more severe downstream consequences in drug development and in drug safety evaluations. Therefore, the Meta-GTMP higher Sn (+6.82%, 5-shot), Acc (+6.23%, 5-shot) and ROC-AUC (+2.50%, 5-shot) ensures that mutagenic compounds are less likely to be overlooked, addressing a key limitation of GIN and other GNN baselines. By prioritizing certain compounds for experimental testing, while minimizing false negatives, Meta-GTMP can significantly reduce the costs and time spent in drug discovery, preserving high standards of safety and efficacy. These results highlight the potential of Meta-GTMP as a valuable tool to accelerate drug discovery and development, setting a novel approach to infer the mutagenic and non-mutagenic properties of drug candidates.

2.2 Computational results of single-task Meta-GTMP models

In our computational experiments, we test the ability of Meta-GTMP to predict the overall Ames mutagenicity considering a single strain as the meta-training task. The main goal is to evaluate the performance of single task (ST) models for each bacterial strain used in the Ames test to determine which strain has the most significant contribution to model the Ames mutagenicity. In this sense, we conduct computational experiments to compare the performances of Meta-GTMP models which use a single meta-training task for each strain involved in

Table 3 Average metric scores for binary classification on the overall Ames test task across 30 computational experiments in the 5-shot setting by single-task Meta-GTMP models (ST)

5-Shot (5 +,5 −)						
Metric	ST1-TA98	ST2-TA100	ST3-TA102	ST4-TA1535	ST5-TA1537	Δ (All − strains)
Specificity (Sp)	0.9205 ± 0.0155	0.8223 ± 0.0603	0.9686 ± 0.0296	0.8918 ± 0.0461	0.8499 ± 0.2192	−0.0129
Sensitivity (Sn)	0.8106 ± 0.0149	0.9050 ± 0.0207	0.6598 ± 0.0684	0.6479 ± 0.0933	0.6622 ± 0.1105	−0.0076
Precision (Pr)	0.9929 ± 0.0012	0.9859 ± 0.0043	0.9967 ± 0.0028	0.9883 ± 0.0031	0.9863 ± 0.0151	−0.0018
Accuracy (Acc)	0.8180 ± 0.0130	0.8994 ± 0.0153	0.6808 ± 0.0623	0.6644 ± 0.0842	0.6751 ± 0.0895	−0.0179
F1-score (F1s)	0.8924 ± 0.0086	0.9436 ± 0.0092	0.7919 ± 0.0503	0.7788 ± 0.0667	0.7868 ± 0.0683	−0.0100
ROC-AUC	**0.8655 ± 0.0039**	0.8636 ± 0.0201	0.8142 ± 0.0252	0.7698 ± 0.0280	0.7561 ± 0.0631	−0.0816

Table 4 Average metric scores for binary classification on the overall Ames test task across 30 computational experiments in the 10-shot setting by single-task Meta-GTMP models (ST)

10-Shot (10 +,10 −)						
Metric	ST1-TA98	ST2-TA100	ST3-TA102	ST4-TA1535	ST5-TA1537	Δ (All − strains)
Specificity (Sp)	0.8507 ± 0.0760	0.8629 ± 0.0255	0.9594 ± 0.0381	0.8696 ± 0.0521	0.8233 ± 0.1208	−0.0260
Sensitivity (Sn)	0.8546 ± 0.0294	0.8847 ± 0.0197	0.7406 ± 0.0481	0.6639 ± 0.0624	0.7173 ± 0.0532	−0.0191
Precision (Pr)	0.9879 ± 0.0057	0.9891 ± 0.0018	0.9963 ± 0.0032	0.9865 ± 0.0046	0.9834 ± 0.0095	−0.0026
Accuracy (Acc)	0.8544 ± 0.0230	0.8833 ± 0.0168	0.7552 ± 0.0427	0.6776 ± 0.0555	0.7243 ± 0.0422	−0.0260
F1-score (F1s)	0.9161 ± 0.0148	0.9339 ± 0.0102	0.8487 ± 0.0305	0.7919 ± 0.0452	0.8282 ± 0.0316	−0.0150
ROC-AUC	0.8527 ± 0.0256	**0.8738 ± 0.0040**	0.8500 ± 0.0113	0.7667 ± 0.0194	0.7703 ± 0.0368	−0.0708

the Ames test. In Tables 3 and 4, we show the mean and standard deviations of performance results for single-task Meta-GTMP models using information of each individual strain across 30 experiments with (5 +,5 −) (5-shot) and (10 +,10 −) (10-shot) random support sets. The Δ (All − strains) column quantifies the differential between the top-performing single-task Meta-GTMP models and Meta-GTMP considering all five strains for meta-training (see Tables 1 and 2). In Fig. 3, scatter plots overlaid with boxplots show the difference between the ROC-AUC scores obtained across 30 experiments by Meta-GTMP considering all five bacterial strains (All-strains) as meta-training tasks and by single-task Meta-GTMP models using a single meta-training task for one single bacterial strain in 5-shot and 10-shot experiments.

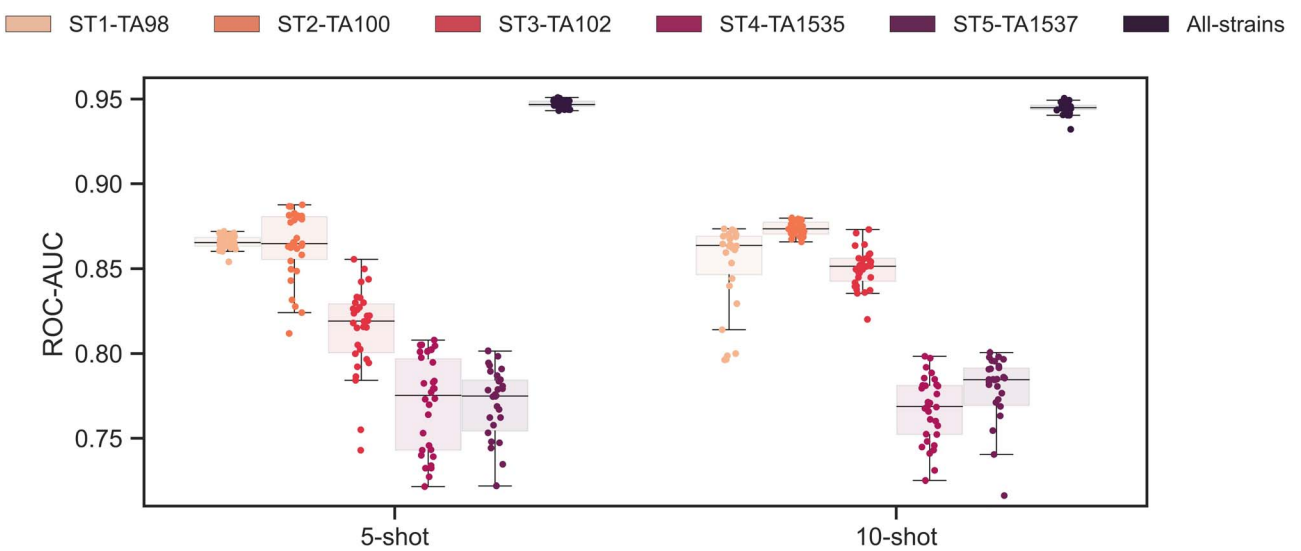

Fig. 3 Scatter plots overlaid with boxplots showing the comparison between ROC-AUC scores obtained by the single-task (ST) Meta-GTMP models for each individual strain and the Meta-GTMP model considering all Ames test strains in meta-training (All-strains) across 30 computational experiments in the 5-shot and 10-shot settings.

Table 3 shows that single-task Meta-GTMP models for TA98 or TA100 strains outperform the other single-task models exhibiting higher ROC-AUC results in 5-shot experiments. It is interesting to note that individual strains with higher class imbalance or limited data such as TA1535 and TA1537 are expected to negatively affect predictive performance. ROC-AUC results for 10-shot experiments (see Table 4), show that individual strains with a higher number of labeled compounds and a lower class imbalance, TA98 and TA100 achieve more robust results in Ames mutagenicity prediction (see Fig. 3). However, by comparing the performance of single-task models with our multi-task Meta-GTMP approach (All-strains), we observe that single-task models for individual strains are missing crucial information to discriminate mutagenic and non-mutagenic compounds. In 5-shot and 10-shot experiments, the Meta-GTMP model considering all the five Ames bacterial strains (All-strains) as meta-training tasks outperforms all the single-task models, suggesting that taking into account the unique

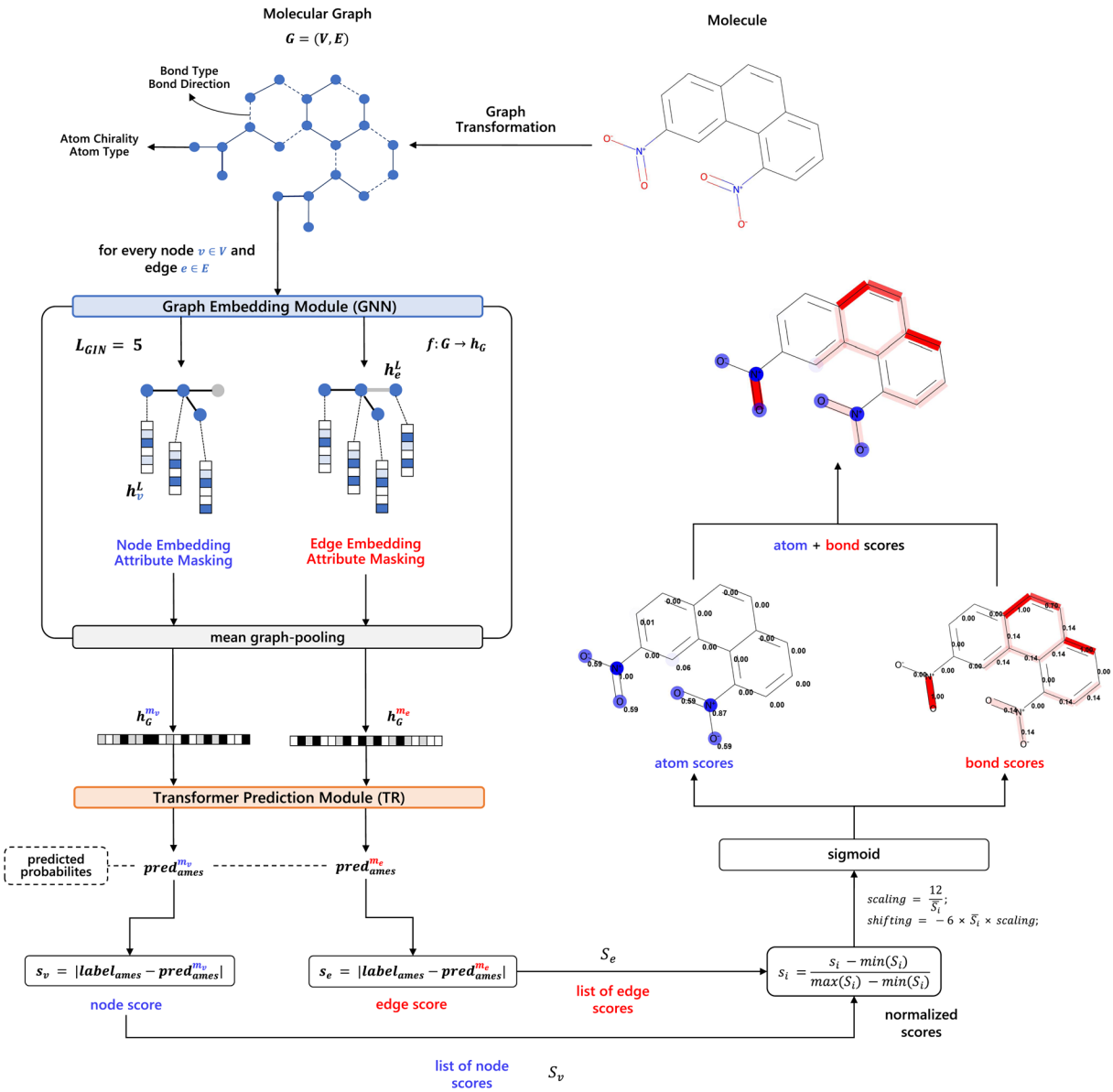

Fig. 4 Workflow of the explainability study using the Meta-GTMP predictions and a node-edge attribute masking strategy. Molecular graphs G representing compounds serve as input to the GNN embedding module of Meta-GTMP to obtain node and edge embeddings (h_v, h_e) and graph-level embeddings h_G at the last GIN layer $L_{GIN} = 5$. Here, for each node v, node embeddings h_v are masked by setting the node embedding features to zero and the modified graph is used to obtain a modified graph embedding $h_G^{m_v}$. In the same way, for each edge e, edge embeddings h_e are masked by setting the edge embedding features to zero and the modified graph is used to obtain a modified graph embedding $h_G^{m_e}$. The modified graph embeddings $h_G^{m_v}, h_G^{m_e}$ serve as input to the Meta-GTMP Transformer prediction module to obtain the predicted probabilities $\text{pred}_{\text{ames}}^{m_v}, \text{pred}_{\text{ames}}^{m_e}$. The node and edge scores s_v, s_e, for each node v and edge e masked, are obtained by the absolute difference of $\text{pred}_{\text{ames}}^{m_v}, \text{pred}_{\text{ames}}^{m_e}$. with the ground-truth labels $\text{label}_{\text{ames}}$ for the overall Ames test task. These scores are normalized using a min–max scaling followed by sigmoid activation and visualized using a color gradient (blue for nodes, red for edges) with more intense colors denoting higher node (atom) or edge (bond) scores.

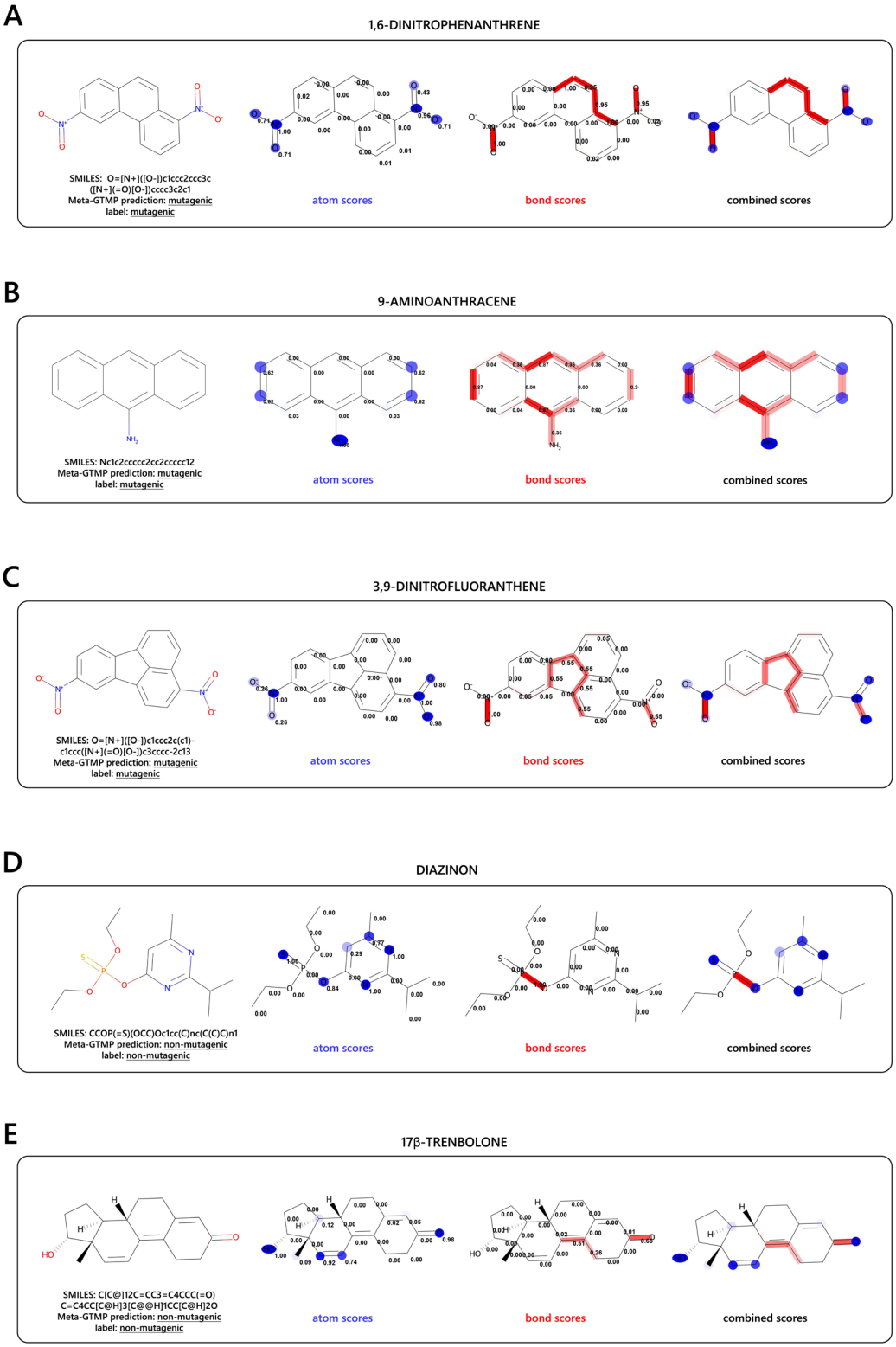

Fig. 5 Visual representation of the atom and chemical bond normalized scores in mutagenic and non-mutagenic predictions for five selected compounds (3 mutagenic: 1,6-dinitrophenanthrene (A), 9-aminoanthracene (B), 3,9-dinitrofluoranthene (C); 2 non-mutagenic: diazinon (D), 17β-trenbolone (E)). The node-edge masking explainability results are obtained using the predictions of Meta-GTMP for the overall Ames test task and the mutagenicity labels in the 5-shot setting. The highlighted areas for each compound indicate the regions of higher scores or importance to determine the mutagenic or non-mutagenic activity, with the blue color gradient used for atoms and the red color gradient used for chemical bonds.

contribution of each individual strain and their complementarity has a positive impact in the predictive performance, leading to more accurate and robust models of mutagenic toxicity.

The multi-task Meta-GTMP model effectively integrates diverse strain-specific data, overcoming the challenges posed by the positive-negative class imbalance for individual bacterial strains. By incorporating data from strains such as TA98 and TA100, which have a larger number of samples and lower class imbalance, and from strains like TA1535 and TA1537, which are more limited in data and highly imbalanced, Meta-GTMP learns complementary dependencies that the single-task models fail to capture. The integration of this strain-specific information provides more robust and accurate predictions for the overall meta-testing task. The ability of Meta-GTMP to leverage the individual contributions of each strain helps to mitigate the negative effects of class imbalance and limited data, leading to performance improvements when distinguishing mutagenic from non-mutagenic compounds.

2.3 Chemistry-oriented explanations of mutagenic and non-mutagenic predictions using a node-edge attribute masking strategy

In this work, we introduce an explainability study of Ames mutagenicity using the predictions made by the Transformer module of Meta-GTMP and the output node and edge embeddings of the Meta-GTMP graph embedding module. A node-edge attribute masking strategy is used to obtain chemical explanations for the mutagenic and non-mutagenic predictions made by Meta-GTMP. As described in the previous sections, each node $v \in V$ (atoms) and edge $e = (v, u)$ with $u \in N(v)$ (chemical bonds) in molecular graphs G is embedded into a high-dimensional space of node and edge embeddings (h_v, h_e) to capture the molecular and structural dependencies relevant to determine the Ames mutagenicity result. Meta-GTMP predicts the mutagenicity of compounds in the overall Ames test task, with node and edge embeddings serving as deep representations obtained by the GNN embedding module to build the graph-level embeddings h_G (mean graph-pooling) used as the input to the Transformer prediction module, which computes the final prediction. To identify key molecular features associated with the mutagenicity and non-mutagenicity, we apply node and edge attribute masking to systematically inactivate the node and edge embeddings in molecular graphs and observing the impact on Meta-GTMP predictions. Hence, by comparing the node-edge masked predictions with the original mutagenicity labels, we are able to quantify the contributions of atoms and chemical bonds to determine the Ames mutagenicity result. In Fig. 4, we show the workflow of this explainability study, outlining the steps from the initial molecular graph to the visualization of the nodes (atoms) and the edges (bonds) that determine the mutagenicity or non-mutagenicity. In this study, we calculate the atom (node) and bond (edge) scores derived from the Meta-GTMP predictions for a total of 3324 molecules, the query set of the overall Ames test task in the 5-shot setting (using a random support set of 10 compounds with $n_+ = 5$ positives and $n_- = 5$ negatives of a total of 3103 positives + 231 negatives = 3334 samples). The node scores s_v provide insights into the importance of individual atoms, highlighting specific atomic positions critical to determine the mutagenicity or non-mutagenicity. The edge scores s_e quantify the importance of chemical bonds in influencing the mutagenic or non-mutagenic properties. Therefore, the combined analysis of atom (node) and bond (edge) scores is crucial for identifying key molecular substructures and functional groups that play an important role to determine the Ames mutagenicity result. In Fig. 5, we analyse five compounds showing the atoms and bonds with higher scores, suggesting the existence of substructures that contribute more prominently to the mutagenic and non-mutagenic properties.

Conversely, the atoms and chemical bonds of lower scores indicate specific regions where structural changes have minimal impact to determine the Ames mutagenicity result. These five compounds were chosen to represent a diverse set of chemical structures, including different types of molecular substructures and functional groups responsible for their mutagenic and non-mutagenic properties. In Fig. 5A, the atom scores in 1,6-dinitrophenanthrene highlight the nitroso (N═O) groups, a well-known mutagenic moiety, which leads to the formation of reactive intermediates, implicated in the formation of DNA adducts, segments of DNA associated with mutagenic effects.[45,46] The chemical bond scores also highlight the N═O bonds of the nitroso groups, which have the potential of interacting with DNA segments and promote genotoxicity.[47,48] In addition, the C–C bonds of the phenanthrene ring are also highlighted, which provide a conjugated system that can stabilize the derived reactive intermediates. In Fig. 5B, the atom scores in 9-aminoanthracene emphasize the nitrogen atom integrated in the amino (NH2) group, which can lead to the formation of reactive electrophiles called nitrenium ions.[49] These electrophilic species form strong covalent bonds with important nucleophilic centers such as the DNA, leading to significant genotoxic effects.[50,51] In addition, the atom and chemical bond scores in the anthracene ring highlight the planarity and aromatic nature of the 9-aminoanthracene, which enables the intercalation with DNA chains, a key mechanism underlying their mutagenic potential.[52,53] In Fig. 5B, the bond scores denote the C–N bond connecting the amine group (NH2) with the anthracene ring system, highlighting its role in stabilizing the derived reactive intermediates. In Fig. 5C, the atom scores obtained for 3,9-dinitrofluoranthene focus on the nitrogen atoms of the nitroso groups, which play a central role in the formation of nitrenium ions, capable of directly interacting with DNA. Similarly, the bond scores also emphasize the N═O bonds, which lead to the formation of these reactive intermediates with the potential of interacting with DNA and cause mutagenic effects.[45,46] In non-mutagenic compounds such as diazinon in Fig. 5D, the atom scores place emphasis in the sulfur atom (S) in the thiophosphate group, which is typically not reactive towards DNA, reducing the mutagenic potential.[54] The oxygen atom is also highlighted as part of ester group (C–O), which maintains the structural integrity of the compound rather than promoting the formation of reactive

species.[55] In addition, atom scores emphasize the C–N atoms in the pyrimidine ring, which are crucial for maintaining the structure of the ring, affecting how diazinon interacts with biological targets.[56] The bond scores highlight the C–O ester bond in diazinon, which improves chemical stability, solubility and resistance to bioactivation.[55] In Fig. 5E, the atom scores in 17β-trenbolone highlight the hydroxyl (OH) group, which influences the binding to the androgen receptors and ensures a biological functionality without inducing mutagenicity.[57] The carbonyl (C═O) group is also highlighted, which increases the binding affinity of 17β-trenbolone to androgen receptors without non-specific mutagenic interactions.[58] The bond scores emphasize the C–C bonds of the rigid steroid structure of 17β-trenbolone, increasing the overall stability, preventing the formation of electrophilic species, and thereby reducing the risk of genotoxicity.[59] From this explainability study, we conclude that this node-edge attribute masking strategy allows the detection of molecular substructures and functional groups that collectively determine the mutagenic or non-mutagenic properties. The atom and bond scores are calculated for a query set with a total of 3324 compounds in the 5-shot setting accessible *via* the Github link provided in the Code Availability section.

2.4 Experimental validation of Meta-GTMP mutagenicity predictions

The molecular substructures identified using the node-edge attribute masking strategy described in the previous section were used as queries in the PubChem[60] substructure search tool to select six new compounds (three predicted mutagenic and three predicted non-mutagenic) with a diverse set of chemical structures and unknown mutagenicity labels for laboratory validation experiments. These substructures correspond to highly informative atomic and bond-level features identified by the model, often overlapping with known mutagenic motifs (nitroso, nitroaromatic, aromatic amine structures) and non-mutagenic features. They were encoded as SMARTS patterns and submitted to PubChem's substructure search API. We include a detailed list of selected compounds in the SI Material, along with their corresponding SMILES string representations and Meta-GTMP mutagenicity predictions. The substructure search was based on the following criteria: (1) unknown mutagenicity labels of the selected compounds to validate the applicability of Meta-GTMP in a potential drug discovery scenario; (2) compliance with the Lipinski rule of five, which evaluates the drug-likeness of chemical compounds to ensure a more favorable pharmacological profile and (3) availability and feasibility of the selected compounds for laboratory testing, ensuring that they could be sourced and tested within a reasonable timeframe and budget. In Fig. 6, we describe the six selected compounds and their chemical structures, half of them were predicted to be mutagenic, and the remaining three compounds were predicted to be non-mutagenic, according to the 5-shot and 10-shot Meta-GTMP models. More details regarding the most important molecular substructures identified for each one of these compounds using a node-edge attribute masking strategy are provided in the SI Material.

Laboratory experiments were conducted to validate the computational Meta-GTMP predictions for the six selected compounds with unknown Ames test results. These chemical compounds were tested using a miniaturized version of the Ames bacterial reversion test, in 6-well cell culture plates,[61,62]

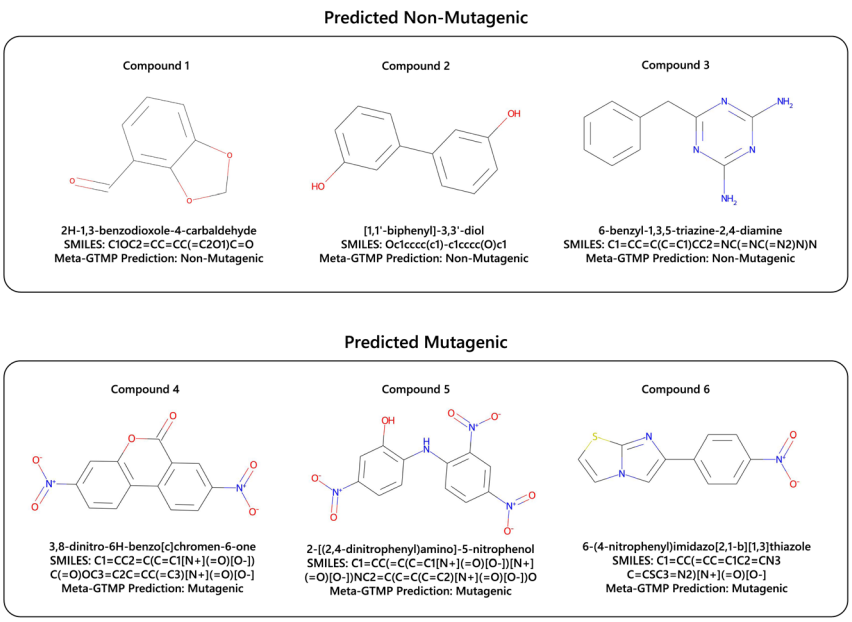

Fig. 6 List of the six compounds selected for laboratory validation experiments. A total of 3 compounds predicted mutagenic (3,8-dinitro-6*H*-benzo[*c*]chromen-6-one; 2-[(2,4-dinitrophenyl)amino]-5-nitrophenol and 6-(4-nitrophenyl)imidazo[2,1-b][1,3]thiazole) and 3 predicted non-mutagenic (2*H*-1,3-benzodioxole-4-carbaldehyde; [1,1′-biphenyl]-3,3′-diol and 6-benzyl-1,3,5-triazine-2,4-diamine) were selected for laboratory validation experiments and the Ames mutagenicity predictions for each selected compound were computed by the Meta-GTMP model.

following the OECD testing guideline 471.[4,6] As *S. typhimurium* TA98 and TA100 strains together are capable of detecting 93% of the mutagens identified by all the other strains,[17,63] we established a tiered approach to the Ames test workflow. Each compound was first tested with strain TA98, then with strain TA100, followed by *S. typhimurium* strains TA1535, TA1537, and *E. coli* WP2, which according to the OECD guidelines is equivalent to *S. typhimurium* TA102.[4] Only those compounds that did not test positive with TA98 were tested with the rest of the strains. Table 6 summarizes the results for all tested compounds and further details can be found in the SI Material.

2.5 Comparison of predicted probabilities: Meta-GTMP *vs.* GNN baselines

To further validate our findings, we extended the experimental validation of the computational results to the GNN baseline models, including: GIN, GCN, and GraphSAGE. By directly computing the predicted probabilities across the six selected compounds for experimental validation, we show that Meta-GTMP consistently outperforms the GNN baselines when predicting the probability of a compound to be mutagenic or non-mutagenic. Predicted probabilities for non-mutagenic compounds (Compounds **1–3**) and mutagenic compounds (Compounds **4–6**) under the 5-shot and 10-shot settings are displayed in Table 5.

For non-mutagenic compounds (Compounds **1–3**), Meta-GTMP achieves predicted probabilities close to 0.999 in 5-shot and 10-shot settings, outperforming the GNN baselines across all compounds. Similarly, for mutagenic compounds (Compounds **4–6**), Meta-GTMP achieves near-perfect predicted probabilities (*e.g.*, 0.999982 for Compounds **5** and **6**, 5-shot setting), aligning closely with the experimental labels. In contrast, GNN baselines exhibit higher variability and struggle to match the experimental outcomes, with probabilities often falling below the threshold required for confident and reliable predictions.

Table 5 Predicted probabilities obtained by Meta-GTMP and GNN baselines for the six compounds selected for experimental validation. Predicted probabilities $\in \{0, 1\}$ are shown for non-mutagenic compounds (Compounds **1–3**) and mutagenic compounds (Compounds **4–6**) under 5-shot and 10-shot settings. The highest predicted probability for each selected compound is highlighted in bold

Compound		GIN	GCN	GraphSAGE	Meta-GTMP
Experimental label: Non-mutagenic					
1	5-shot	0.940237	0.533031	0.495889	**0.999966**
	10-shot	0.943482	0.564391	0.686700	**0.999966**
2	5-shot	0.891616	0.851538	0.803546	**0.999979**
	10-shot	0.904768	0.702467	0.792602	**0.999962**
3	5-shot	0.981856	0.898893	0.839920	**0.999978**
	10-shot	0.977554	0.910749	0.882489	**0.999965**
Experimental Label: Mutagenic					
4	5-shot	0.981704	0.880847	0.819704	**0.999980**
	10-shot	0.993135	0.935832	0.913474	**0.999972**
5	5-shot	0.857370	0.897551	0.739320	**0.999982**
	10-shot	0.847780	0.983059	0.710503	**0.999953**
6	5-shot	0.941193	0.954725	0.892516	**0.999982**
	10-shot	0.921292	0.996580	0.914871	**0.999900**

Meta-GTMP achieves the highest predicted probabilities across all the six selected compounds, consistently aligning with the experimental labels and outperforming the GNN baselines across the 5-shot and 10-shot settings. This predictive performance reflects the robustness and reliability of the Meta-GTMP model, particularly in addressing the challenges posed by high class imbalance and limited data in mutagenicity prediction. These results highlight the practical advancements of Meta-GTMP for applications such as drug discovery and toxicology, where the prediction of mutagenic properties is critical.

2.6 Conclusion

The Ames mutagenicity test is a widely used experimental method for evaluating the mutagenic properties of chemical compounds. In this research, we consider the individual contributions of each bacterial strain involved in the Ames test to predict mutagenic toxicity with high sensitivity and specificity in highly imbalanced scenarios with limited data, which is crucial in drug discovery. To address these challenges, we introduced a few-shot GNN-Transformer, Meta-GTMP, to capture the local structure of molecular graphs and the global information of molecular graph embeddings for mutagenicity prediction. Additionally, a multi-task few-shot learning framework is proposed to leverage the complementarity among individual predictive tasks for each strain of bacteria in a joint learning procedure to model the results of the Ames mutagenicity test with just a few labeled compounds.

Moreover, we implemented a node-edge attribute masking strategy within the Meta-GTMP graph embedding module. This strategy computes a set of atom and bond scores for each molecule, offering valuable insights into the key molecular substructures and functional groups influencing mutagenicity. These explainable insights guided the selection of diverse compounds with unknown mutagenicity labels for experimental validation, which confirmed the Meta-GTMP's computational predictions. These insights are essential in drug discovery and valuable for the identification of chemical properties associated with mutagenicity to inform the selection and design of promising drug candidates. By highlighting important molecular features and functional groups, Meta-GTMP offers a practical method for lead optimization, facilitating the generation of safer and more effective *drug-like* compounds.

The traditional Ames test is labor-intensive, expensive, and time-consuming. The Meta-GTMP model accelerates this process with high reliability and high sensitivity, making the identification of mutagenic and non-mutagenic compounds faster and more cost-effective. Beyond its computational efficiency, the explainability of Meta-GTMP ensures its practical utility across drug discovery pipelines, where understanding the relationship between the chemical structure and mutagenic potential is critical. The results demonstrate that Meta-GTMP achieves substantial improvements over existing methods,

providing a scalable and explainable framework for modeling mutagenicity in low-data and highly imbalanced scenarios, which are common in drug discovery. This makes Meta-GTMP a powerful tool for improving the identification, selection, and rational design of potential drug candidates, offering significant advancements in lead optimization and mutagenicity assessments.

All the compounds predicted by the Meta-GTMP model to be non-mutagenic (compounds **1**, **2**, and **3**) showed no effect in any of the strains tested. For *S. typhimurium* strains TA98 and TA100, and *E. coli* WP2 (equivalent to TA102),[4] a positive result implies that the number of colonies (revertants) induced by the tested compound should be at least two-fold higher than the spontaneous reversion rate (negative control),[61,64,65] which was not observed for these compounds. For the *S. typhimurium* strains TA1535 and TA1537, a positive response is considered whenever the compound causes a three-fold increase compared to the negative control.[61,64,65] For these strains, we found that the spontaneous reversion rate was low, yet within the published ranges.[61,64,65] Moreover, the number of colonies obtained at all concentrations with compounds **1**, **2** and **3** was within the range observed for the negative control (DMSO). Therefore, we concluded that none of these compounds elicited a positive response, indicating that they were non-mutagenic for all strains. The compounds predicted to be mutagenic (compounds **4**, **5**, and **6**) were also confirmed by the Ames *in vitro* assay, showing a clear positive response in the TA98 strain (Table 6). Indeed, after the incubation of TA98 with each of the compounds, an increase in the number of revertants ranging from 4.5 to > 105 fold above the negative control was observed. In summary, the *in vitro* results of the Ames mutagenicity test showed 100% agreement with the Meta-GTMP predictions for both mutagenic and non-mutagenic compounds, further corroborating the robustness of the Meta-GTMP framework.

3 Methods

3.1 Meta-GTMP architecture

3.1.1 Graph embedding module. A graph isomorphism network (GIN)[42] with $L_{\mathrm{GIN}} = 5$ layers is used as a GNN embedding module to generate graph embeddings h_{G} for molecular graphs G (see Fig. 1A). The GNN computes AGGREGATE and COMBINE operations by summing up node and edge features. During each message-passing iteration l, node embeddings h_v^l are updated by aggregating prior node representations h_v^{l-1} with representations of neighboring nodes and edges (h_u^{l-1},h_e^{l-1}). Node embeddings h_v on the $l-th$ layer are given by the COMBINE and UPDATE steps after neighborhood aggregation,

$$m_{N(v)}^l = \mathrm{AGGREGATE}^l\left(\left\{h_u^{l-1}, \forall u \in N(v)\right\}, \left\{h_e^{l-1} : e = (v,u)\right\}\right) \quad (1)$$

$$h_v^l = \sigma(\mathrm{MLP}^l(\mathrm{COMBINE}^l(h_v^{l-1}, m_{\mathrm{N}(v)}^l))) \quad (2)$$

with m the "neural message" transmitted across GNN layers, h_u^l the embedding of neighboring nodes, and h_e^l the embedding of the connection between nodes u and v, with $u \in N(v)$. An UPDATE step uses a multi-layer perceptron (MLP) followed by an activation function σ = ReLU. These operations can be expressed as follows

$$h_v^l = \mathrm{ReLU}\left(\mathrm{MLP}^l\left(\sum_{u \in N(v) \cup v} h_u^{l-1} + \sum_{e=(v,u):u \in N(v) \cup v} h_e^{l-1}\right)\right). \quad (3)$$

In the final iteration of message-passing, graph embeddings h_{G} are obtained using a READOUT operation, which involves aggregating node embeddings h_v at the last GNN layer L_{GIN} to generate a graph-level representation h_{G} with a mean-pooling operation

$$h_G = \mathrm{mean}\left(\left\{h_v^{L_{GIN}} : v \in V\right\}\right). \quad (4)$$

This GNN module is pre-trained with GNN models proposed by Hu *et al.* (2020)[30] to obtain a better initialization. In this module, we consider 5 GNN layers and graph embeddings of size 300.

3.1.2 Transformer prediction module. A Transformer encoder with $L_{\mathrm{T}} = 5$ blocks is used as the Ames mutagenicity prediction module (see Fig. 1A). Unlike the standard Transformer,[66] it acts as a vision Transformer (ViT)[67,68] that takes graph embeddings h_{G} as 1D feature vectors with embedding size of 300. Input graph embeddings h_{G} are converted into a sequence of patches p in a D-dimensional space with N patches of size P. In this process, the Transformer accepts the

Table 6 Summary of the results obtained in Ames laboratory test experiments[a]

Compounds		Meta-GTMP prediction	Ames test result/strain					Ames test overall result
			TA98	TA100	TA1535	TA1537	WP2	
1	2H-1,3-benzodioxole-4-carbaldehyde	N	N	N	N	N	N	N
2	[1,1′-biphenyl]-3,3′-diol	N	N	N	N	N	N	N
3	6-Benzyl-1,3,5-triazine-2,4-diamine	N	N	N	N	N	N	N
4	3,8-Dinitro-6*H*-benzo(*c*)chromen-6-one	P	P	NT	NT	NT	NT	P
5	2-((2,4-dinitrophenyl)(amino)-5-nitrophenol	P	P	NT	NT	NT	NT	P
6	6-(4-nitrophenyl)imidazo[2,1-b][1,3]thiazole	P	P	NT	NT	NT	NT	P

[a] NT – not tested, P – positive (mutagenic), N – negative (non-mutagenic).

input embeddings h_G and converts them into a sequence of patch tokens x_p,

$$p(h_G) = [x_p^1, x_p^2, \ldots, x_p^N] \quad (5)$$

where x_p^i represent individual patch tokens. The Transformer converts graph embeddings h_G into $N = \frac{300}{P}$ patch tokens of size P, which are linearly projected to build Transformer embeddings h_T. Transformer blocks propagate embeddings h_T across multi-head self-attention (MSA) layers. MSA takes 3 inputs: queries, keys, and values (q, k, v) stacked into matrices (Q, K, V) to optimize the dot-product MSA operation. MSA calculates the dot-product for each query in Q and all keys in K and applies a softmax function to obtain the attention weights for each value in V

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \quad (6)$$

In the MSA operation, we consider multiple linear projection heads H and the final attention score is given by

$$\text{MSA}(Q,K,V) = \text{CONCAT}(\text{head}_1, \ldots, \text{head}_H)W \quad (7)$$

$$\text{head}_j = \text{Attention}(QW_j^Q, KW_j^K, VW_j^V) \quad (8)$$

with (W_j^Q, W_j^K, W_j^V) the linear projection matrices obtained by the projection of (Q, K, V) for each head j. Individual Transformer blocks include a MSA layer followed by MLP. MSA and MLP are preceded by layer normalization (LN) and residual connections. For 1D patch token sequences x_p^i, Transformer embeddings h_T obtained across Transformer blocks l can be formulated as

$$h_T^0 = [x_p^1 K, x_p^2 K, x_p^3 K, \ldots, x_p^N K] \quad (9)$$

$$h_T^{l^*} = \text{MSA}\left(\text{LN}\left(h_T^{l-1}\right)\right) + h_T^{l-1} \quad (10)$$

$$h_T^l = \text{MLP}\left(\text{LN}\left(h_T^{l^*}\right)\right) + h_T^{l^*} \quad (11)$$

$$y = \text{LN}\left(h_T^{L_T}\right) \quad (12)$$

where $l = \{1, \ldots, L_T\}$, h_T^l are Transformer embeddings at layer l, K are the linear projections of patch embeddings with and y is the output vector. A linear layer followed by sigmoid activation uses the output of the last Transformer block to predict the mutagenicity result (condensed in a value $\in \{0, 1\}$).

3.1.3 Few-shot meta-learning framework. A meta-learning framework based on model-agnostic meta-learning (MAML)[20,26] was developed to learn complementary information across few-shot tasks and model the contribution of five different strains of bacteria for Ames mutagenicity prediction (see Fig. 1B). This strategy leverages the complementarity among different strains involved in the Ames test by the means of integrating information of these individual predictive tasks with a joint learning procedure. The meta-learning framework optimizes two neural network models: a GNN and a Transformer (TR). Both models update model parameters across few-shot tasks (meta-training) for each strain using a random support set for training and a disjoint query set for evaluation. The updated parameters are used to initialize both models and generalize to new compounds and predict the overall Ames mutagenicity result in the test data (meta-testing). In meta-training, a support set S_t of molecular graphs $G_{S_{t_i}}$ of size k is randomly sampled to serve as an input to the GNN-Transformer and compute the support losses $\mathscr{L}_t^{\text{GNN}}, \mathscr{L}_t^{\text{TR}}$ for each strain: {TA98, TA100, TA102, TA1535, and TA1537} across tasks $t \in \{1\ldots, n_{\text{strains}}\}$. Support losses are then used to iteratively update model parameters $\theta \to \theta'$, $\theta^* \to \theta^{*\prime}$. Both models compute the query losses $\mathscr{L}_t^{\text{GNN}'}, \mathscr{L}_t^{\text{TR}'}$ using a query set Q_t of molecular graphs $G_{Q_{t_j}}$ with the remaining n samples for that task. In meta-training, model parameters are updated by applying just a few gradient steps

$$\theta_t = \theta - \alpha \nabla_\theta \mathscr{L}_t^{\text{GNN}}(\theta) \quad (13)$$

$$\theta_t^* = \theta^* - \alpha^* \nabla_{\theta^*} \mathscr{L}_t^{\text{TR}}(\theta^*) \quad (14)$$

where α and α^* are the size of the steps used for the gradient descent updates. In meta-testing, a support set of k examples is randomly sampled for the overall Ames test task $t = n_{\text{strains}} + 1 = T$ and model parameters are initialized using the updated parameters from meta-training, θ', $\theta^{*\prime}$. Then, the GNN and the Transformer modules are used to predict the overall Ames mutagenicity result of a query set of new compounds with the remaining n samples for this overall test task.

Algorithm 1 Meta-GTMP: Few-shot meta-learning framework for Ames mutagenicity prediction

Require: Support data: (S_t, Y_t), Query data: (Q_t, Y_t'), $\alpha, \beta, \alpha^*, \beta^*$: update step sizes
  $\theta \leftarrow$ Pre-trained GNN
  while ***not done*** **do**
    Sample a batch of tasks $t \sim \rho(T)$ across all *S. T.* strains and overall Ames result
    for all t **do**
      Sample k examples $\{G_{S_{t_1}}, G_{S_{t_2}}, \ldots G_{S_{t_k}}\} \in S_t$
      for $i = 1$ to k **do**
        $y_{t_i}, h_{G_{S_{t_i}}} = \text{GNN}(G_{S_{t_i}}, \theta)$
      end for
      $\mathscr{L}_t^{GNN} \leftarrow \{y_{t_1}, y_{t_2}, \ldots, y_{t_k}\}$
      for $i = 1$ to k **do**
        $y_{t_i}, h_{T_{S_{t_i}}} = \text{TR}(h_{G_{S_{t_i}}}, \theta^*)$
      end for
      $\mathscr{L}_t^{TR} \leftarrow \{y_{t_1}, y_{t_2}, \ldots, y_{t_k}\}$
      $\theta' = \theta - \alpha \nabla \mathscr{L}_t^{GNN}$
      $\theta^{*\prime} = \theta^* - \alpha^* \nabla \mathscr{L}_t^{TR}$
      Sample n examples $\{G_{Q_{t_1}}, G_{Q_{t_2}}, \ldots G_{Q_{t_n}}\} \in Q_t$
      for $j = 1$ to n **do**
        $y'_{t_j}, h_{G_{Q_{t_j}}} = \text{GNN}(G_{Q_{t_j}}, \theta')$
      end for
      $\mathscr{L}_t^{GNN'} \leftarrow \{y'_{t_1}, y'_{t_2}, \ldots, y'_{t_n}\}$
      for $j = 1$ to n **do**
        $y'_{t_j}, h_{T_{Q_{t_j}}} = \text{TR}(h_{G_{Q_{t_j}}}, \theta^{*\prime})$
      end for
      $\mathscr{L}_t^{TR'} \leftarrow \{y'_{t_1}, y'_{t_2}, \ldots, y'_{t_n}\}$
    end for
    $\theta \leftarrow \theta\beta \nabla_\theta \sum_{t \sim \rho(T)} \mathscr{L}_i^{GNN'}$
    $\theta^* \leftarrow \theta^* \beta^* \nabla_{\theta^*} \sum_{t \sim \rho(T)} \mathscr{L}_i^{TR'}$
  end while

3.1.4 Weighted loss for mutagenicity prediction. In Meta-GTMP, the loss function for the GNN and Transformer modules, $\mathscr{L}^{\text{GNN}}$ and $\mathscr{L}^{\text{TR}}$ is a binary cross-entropy loss. To address the problem of class imbalance, a weighted binary cross-entropy loss is introduced to greatly penalize failed predictions on rare-class samples. Hence, the loss defines a weight c for the minority class,

$$\mathscr{L} = -\frac{1}{k}\sum_{i=1}^{k} c\ y_i\ \log\left(y_i^{'}\right) + (1-y_i)\log\left(1-y_i^{'}\right) \qquad (15)$$

where y' are the predictions and y the mutagenicity labels with k the number of samples. Since we have different positive-negative ratios for each strain used in the Ames test, the value c is determined by exploring values in a reasonable range $c \in \{0.1, \ldots, 10\}$. In our computational experiments, we select a value of $c = 5$ given the task-specific variability across the Ames test bacterial strains. This strategy is particularly important for strains with a severe imbalance between positive and negative samples, such as TA1535 and TA1537, which have very few mutagenic compounds in the ISSSTY dataset. By up-weighting the contribution of rare positive examples during loss computation we improve sensitivity without overfitting to the dominant negative class. Moreover, this approach aligns with the few-shot setting in which balanced support sets are sampled during training, helping the model generalize across highly imbalanced tasks. Together, the use of class weighting and balanced episodic training allows Meta-GTMP to learn generalizable representations across all strains, mitigating the adverse impact of limited imbalanced data on predictive performance.

3.2 Computational experimental settings

3.2.1 Dataset. Data is collected using a public compound repository known as ISSSTY v1-a[40,41] (accessed on March 14, 2024), which includes publicly available information on the mutagenic effects of compounds on five *S. typhimurium* strains across Ames mutagenicity test experiments. This repository includes data of 7367 compounds and contains information on the Ames test performed with five bacterial strains {TA98, TA100, TA102, TA1535, and TA1537} to obtain the overall Ames mutagenicity result.[18] We selected this dataset due to the high-quality experimental annotations and the inclusion of strain-specific mutagenicity results across five bacterial strains. This is essential for our few-shot meta-learning framework, which depends on a multi-task structure to define task-specific support and query sets. The ISSSTY dataset therefore provides both the biological granularity and structured diversity necessary to model cross-strain patterns under low-data scenarios. Compounds are represented using molecular graphs obtained from SMILES (Simplified Molecular-Input Line-Entry System) using the RDKit.Chem library,[69] which are pre-processed, so that SMILES are canonicalized and duplicates are removed.[70] While RDKit performs basic molecule sanitization automatically (*e.g.*, valence checks and aromaticity detection), we did not apply tautomer normalization or rare functional group filtering. This decision preserves the structural fidelity of reported compounds and avoids potential distortion of their bioactivity profiles, while we recognize that such normalization may improve consistency and robustness in future studies. After the pre-processing stage, we obtain a dataset of 6445 compounds with six different labels per compound: five different labels for each different bacterial strain {TA98, TA100, TA102, TA1535 and TA1537}, and one for the overall Ames mutagenicity label. In this process, the labels are computed to follow the standard Organisation for Economic Cooperation and Development Guidelines for the Testing of Chemicals (OECD)-5 classification: TA98, TA100, TA1535, TA1537 (or TA97) and TA102 (or *E. coli*)[4] after aggregating the labels including derivations of the same strain. These labels can be either: positive (mutagenic), negative (non-mutagenic), inconclusive or equivocal. Equivocal labels are produced if compounds do not yield positive results in any strains and there is at least one equivocal result across the five strains. Inconclusive labels are returned if there is not enough data to make a determination. Equivocal and inconclusive designations are merged into a new label: undetermined. At this stage, most compounds with an undetermined label in one strain can have a distinct label for at least one other strain. If a compound has an undetermined label for at least one strain, but had either a positive or negative label in the remaining strains, it is still included in the final data. However, if a compound has an undetermined label for all tested strains, they are removed, to obtain a final dataset with 5536 compounds. In Table 7, we report the distribution of compounds for the bacterial strains used in the Ames test and for the overall Ames mutagenicity labels.

External validation using other known mutagenicity datasets was considered. However, most publicly available datasets report only the overall (aggregated) Ames result, a binary label indicating whether a compound is mutagenic in any strain, and do not include the strain-specific annotations necessary to generate a set of few-shot tasks. As a result, such datasets are not directly compatible with our task-based meta-learning setup. Given these constraints, ISSSTY v1-a was selected as the most appropriate benchmark to develop and evaluate Meta-GTMP. Future work may explore extensions of Meta-GTMP that adapt the model for compatibility with aggregated-label datasets or hybrid evaluation strategies which may incorporate scaffold-based task generation.

3.2.2 Implementation. Meta-GTMP is implemented in Python 3.9.16 and PyTorch 1.13.0 with CUDA 11.6, along with functions in Scikit-learn 1.2.2, NumPy 1.22.3, Pandas 1.5.3 and RDKit 2022.03.5. The Meta-GTMP model was trained across ($n_{\text{strains}} \times epochs$) iterations with $n_{\text{strains}} = 5$ as the number of meta-training tasks and *epochs* as the number of epochs. The best model is selected at the epoch giving the best ROC-AUC

Table 7 Distribution of positive, negative and undetermined samples for each one of the five bacterial strains and for the overall Ames mutagenicity label

Ames mutagenicity dataset			
Strain	# Positive (1)	# Negative (0)	# Undetermined (−1)
TA98	1676	2782	1078
TA100	2096	2721	719
TA102	226	587	4723
TA1535	436	2103	2997
TA1537	365	1779	3392
Overall	3103	231	2202

score in the query set of the overall Ames test task and we allow it to run for at most 2000 epochs. Additionally, we consider learnings rates of 1×10^{-3} for the GNN, 1×10^{-4} for the Transformer, update steps of 5 for meta-training and 10 for meta-testing. In Meta-GTMP, the GNN embedding module includes 5 GIN message-passing layers and embedding dimension of 300. For Meta-GTMP and graph-based baselines, the GNN models are pre-trained with GNN models of Hu *et al.* (2020)[30] for improved initialization. The GNN and Transformer modules of Meta-GTMP have the main hyperparameters displayed on the SI Material section. In this work, we do not mainly focus on hyperparameter optimization, especially for GNN baselines. Thus, we did not put an extensive effort into optimizing model hyperparameters, leaving it for future work. In model convergence, we consider a random seed of 1. The computational results are fully reproducible by using 30 different random seeds $\in \{2, \ldots, 31\}$ for each one of the 30 computational experiments conducted in 5-shot and 10-shot settings. The explainability study using a node-edge attribute masking strategy with the output node-edge embeddings of the Meta-GTMP GNN embedding module uses the RDKit.Chem library[69] to visualize node (atom) and edge (bond) importance scores for each compound using blue and red color gradients, respectively. The model is trained to generalize across related tasks defined by bacterial strains. This setup reflects the intended application of few-shot learning frameworks, which aim to perform well under task-level variability with limited data. Nonetheless, we acknowledge that scaffold-aware task construction and external validation using complementary datasets could further strengthen the evaluation of generalization. However, many external datasets do not provide the necessary strain-level annotations required for task definition in our proposed framework. Future extensions of this work may explore strategies to incorporate scaffold-aware sampling protocols or adapt the model for compatibility with aggregated-label datasets to support broader benchmarking.

3.2.3 Comparison with machine learning (ML) baselines. For comparative purposes, we implemented a set of classical machine learning (ML) models using standard cheminformatics descriptors. The results are available in the SI Material section and compared against 5-shot and 10-shot Meta-GTMP models. Specifically, we used RDKit fingerprints and Morgan fingerprints generated *via* RDKit, alongside four widely used classifiers, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gaussian Process, implemented with scikit-learn. These models serve as representative baselines for molecular property prediction. While we acknowledge that further tuning, descriptor selection, and bias mitigation strategies could improve their performance, our focus in this work is on evaluating the proposed Meta-GTMP model, a graph-based few-shot learning approach designed for multi-strain-based mutagenicity prediction under low-data settings. Therefore, the classical ML models were included as reference points under standard conditions rather than fully optimized competitors, consistent with our treatment of the proposed Meta-GTMP model, where we likewise did not perform extensive hyperparameter tuning. This approach ensures a fair and balanced comparison across methods. Hence, the focus of this work is on evaluating the proposed few-shot graph-based meta-learning framework Meta-GTMP under low-data settings. Consequently, we chose standard descriptors and machine learning algorithms to serve as baseline methods for comparison.

3.3 Ames laboratory test experimental settings

3.3.1 Chemical compounds. The compound versions used in laboratory experiments were commercially sourced and can be uniquely identified using their CAS (Chemical Abstract Service) numbers. Hence, the compounds used in experiments were commercially sourced from vendor catalogs *via* the CAS numbers linked in the original PubChem records (accessed May 20, 2024). The tested compounds in laboratory validation experiments included: 2*H*-1,3-benzodioxole-4-carbaldehyde (CAS No 7797-83-3, BLD PharmaTech GmbH), 3,8-dinitro-6*H*-benzo(*c*)chromen-6-one (CAS No 63636-7, Vitas M Chemical limited), 6-benzyl-1,3,5-triazine-2,4-diamine (CAS No 1853-88-9, Specs), 6-(4-nitrophenyl)imidazo[2,1-b][1,3]thiazole (CAS No 7120-14-1, Vitas M Chemical limited), 2-((2,4-dinitrophenyl)(amino)-5-nitrophenol (CAS No 304479-24-1, Specs) and [1,1′-biphenyl]-3,3′-diol (CAS No 612-76-0, Fluorochem Limited). All compounds were purchased in batch through the global chemical marketplace MolPort. The chemicals 2-aminoanthracene (CAS No 613-13-8, Sigma), 2-nitrofluorene (CAS No 607-57-8CAS, Sigma), sodium azide (CAS No 26628-22-8, Sigma), 9-aminoacridine (CAS No 90-45-9, Merck), and methyl methaneosulfonate (CAS No 66-27-3, Sigma) were used as the positive controls for mutagenicity.

3.3.2 Bacterial strains. The histidine-dependent *S. typhimurium* strains TA98, TA100, TA1535 and TA1537, and the tryptophan-dependent *E. coli* WP2 (*uvra* pKM101), which is equivalent to the *S. typhimurium* TA102,[4] were purchased from Trinova Biochem (Moltox). These bacterial strains used in the Ames test were grown in liquid nutrient broth (Oxoid No. 2) at 37° with agitation.

3.3.3 Ames test. The Ames bacterial reverse mutation assay was performed according to the OECD Test Guideline No. 471,[4,6] albeit in a modified, miniaturized version.[61,62,64,65,71] The assay was conducted in 6-well cell culture plates; each well contained 5 mL minimal agar medium consisting of Vogel–Bonner medium E and glucose,[6] and 400 μL of top agar containing 0.6% NaCl, 0.05 mM histidine and biotin (for *S. typhimurium* strains) or 0.05 mM tryptophan (for the *E. coli* WP2 strain). Bacterial cultures grown to late log growth phase (3 to 5×108 cells per mL)[61] were collected by centrifugation, washed, and resuspended in phosphate buffer (PBS). Each chemical compound was tested at 1, 0.3, 0.1, 0.03, 0.01, and 0.003 mg per well. Briefly, compounds were first dissolved in DMSO and serially diluted from the highest concentration (100 mg mL^{-1}) at half-log intervals to prepare five stock solutions of 100, 30, 10, 3, and 1 mg mL^{-1}. Each stock solution (10 μL) was mixed with 20 μL of each strain culture and 100 μL of PBS, and incubated at 37 °C for 30 min. To test mutagenicity induced by metabolic activation, a 10% S9 mix (10% Mutazyme, Trinova Biochem,

Moltox) was used instead of PBS. After incubation, 400 μL of melted top agar were added to each condition/strain, and cultures were poured into the wells. The plates were incubated at 35 °C for 48 hours for colony counting. The Ames test was performed in triplicate for each chemical compound.

Author contributions

Luis H. M. Torres: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, writing - review & editing, visualization. Sofia M. da Silva: Validation, formal analysis, investigation, data curation, writing – original draft, writing - review & editing, visualization. Joel P. Arrais: Writing – review & editing, supervision. Catarina Pimentel: Writing – review & editing, supervision. Bernardete Ribeiro: Writing – review & editing, supervision.

Conflicts of interest

The authors declare no competing interests.

Data availability

All data used in this study are available from publicly available repositories. The dataset used in the computational experiments, ISSSTY v1-a, can be accessed *via* the Istituto Superiore di Sanità Toxicology Portal at: **https://www.iss.it/isstox**. Data used in the experimental validation stage for candidate compound selection are available from the PubChem database at: **https://pubchem.ncbi.nlm.nih.gov**. The data used, source code and implementation details of the Meta-GTMP model are openly available in the DOI link: **https://doi.org/10.5281/zenodo.17373535** and GitHub repository: **https://github.com/ltorres97/Meta-GTMP**. This includes all relevant scripts, documentation, and instructions for reproducing the computational experiments presented in this study.

Supplementary information is available. See DOI: **https://doi.org/10.1039/d5dd00276a**.

Acknowledgements

This research was funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020, MOSTMICRO-ITQB R&D Unit (UIDB/04612/2020, UIDP/04612/2020) and LS4FUTURE Associated Laboratory (LA/P/0087/2020) and by national funds through the Portuguese Recovery and Resilience Plan (PRR) through the project C645008882-00000055, Center for Responsible AI (**https://www.centerforresponsible.ai/**).

References

1 G. Brambilla and A. Martelli, *Pharmacol. Res.*, 2009, **60**, 1–17.
2 B. N. Ames, F. D. Lee and W. E. Durston, *Proc. Natl. Acad. Sci. U. S. A.*, 1973, **70**, 782–786.
3 D. M. Maron and B. N. Ames, *Mutation Research/Environmental Mutagenesis and Related Subjects*, 1983, vol. 113, pp. 173–215.
4 OECD, OECD Publishing, 2020, p. 20745788, DOI: **10.1787/9789264071247–en**.
5 D. E. Levin, M. Hollstein, M. F. Christman, E. A. Schwiers and B. N. Ames, *Proc. Natl. Acad. Sci. U. S. A.*, 1982, **79**, 7445–7449.
6 K. Mortelmans and E. Zeiger, *Mutat. Res., Fundam. Mol. Mech. Mutagen.*, 2000, **455**, 29–60.
7 D. D. Levy, E. Zeiger, P. A. Escobar, A. Hakura, B. jan M. van der Leede, M. Kato, M. M. Moore and K. ichi Sugiyama, *Mutat. Res., Genet. Toxicol. Environ. Mutagen.*, 2019, **848**, 403074.
8 M. Honma, A. Kitazawa, A. Cayley, R. V. Williams, C. Barber, T. Hanser, R. Saiakhov, S. Chakravarti, G. J. Myatt, K. P. Cross, E. Benfenati, G. Raitano, O. Mekenyan, P. Petkov, C. Bossa, R. Benigni, C. L. Battistelli, A. Giuliani, O. Tcheremenskaia, C. DeMeo, U. Norinder, H. Koga, C. Jose, N. Jeliazkova, N. Kochev, V. Paskaleva, C. Yang, P. R. Daga, R. D. Clark and J. Rathman, *Mutagenesis*, 2019, **34**, 41–48.
9 R. Benigni, A. Bassan and M. Pavan, *Expert Opin. Drug Metab. Toxicol.*, 2020, **16**, 651–662.
10 A. Cassano, G. Raitano, E. Mombelli, A. Fernández, J. Cester, A. Roncaglioni and E. Benfenati, *J. Environ. Sci. Health*, 2014, **32**, 273–298.
11 M. K. Leong, S. W. Lin, H. B. Chen and F. Y. Tsai, *Toxicol. Sci.*, 2010, **116**(2), 498–513.
12 N. K. Shinada, N. Koyama, M. Ikemori, T. Nishioka, S. Hitaoka, A. Hakura, S. Asakura, Y. Matsuoka and S. K. Palaniappan, *Mutagenesis*, 2022, **37**(3–4), 191–202.
13 S. J. Webb, T. Hanser, B. Howlin, P. Krause and J. D. Vessey, *J. Cheminf.*, 2014, **6**, 1758–2946.
14 S. K. Chakravarti and S. R. M. Alla, *Front. Artif. Intell.*, 2019, **2**, DOI: **10.3389/frai.2019.00017**.
15 R. T. Naven, S. Louise-May and N. Greene, *Expert Opin. Drug Metab. Toxicol.*, 2010, **6**, 797–807.
16 D. Thorne, J. Kilford, M. Hollings, A. Dalrymple, M. Ballantyne, C. Meredith and D. Dillon, *Mutat. Res., Genet. Toxicol. Environ. Mutagen.*, 2015, **782**, 9–17.
17 R. V. Williams, D. M. DeMarini, L. F. Stankowski, P. A. Escobar, E. Zeiger, J. Howe, R. Elespuru and K. P. Cross, *Mutat. Res., Genet. Toxicol. Environ. Mutagen.*, 2019, **848**, 41–48.
18 M. J. Martínez, M. V. Sabando, A. J. Soto, C. Roca, C. Requena-Triguero, N. E. Campillo, J. A. Páez and I. Ponzoni, *J. Chem. Inf. Model.*, 2022, **62**(24), 6342–6351.
19 I. Olier, N. Sadawi, G. R. Bickerton, J. Vanschoren, C. Grosan, L. Soldatova and R. D. King, *Mach. Learn.*, 2018, **107**, 285–311.
20 Z. Guo, C. Zhang, W. Yu, J. Herr, O. Wiest, M. Jiang and N. V. Chawla, *Proceedings of the World Wide Web Conference, WWW 2021*, 2021.
21 H. Altae-Tran, B. Ramsundar, A. S. Pappu and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 283–293.

22 R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta and P. Kumar, *Mol. Diversity*, 2021, **25**, 1315–1360.
23 A. Lavecchia, *Drug Discovery Today*, 2019, **24**, 2017–2032.
24 A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay and T. Doğan, *Briefings Bioinf.*, 2019, **20**, 1878–1912.
25 S. Silva-Mendonça, A. R. de Sousa Vitória, T. W. de Lima, A. R. Galvão-Filho and C. H. Andrade, *Artif. Intell. Life Sci.*, 2023, **4**, 100086.
26 C. Finn, P. Abbeel and S. Levine, *34th International Conference on Machine Learning, ICML 2017*, 2017, vol. 3, pp. 1856–1868.
27 I. Olier, N. Sadawi, G. R. Bickerton, J. Vanschoren, C. Grosan, L. Soldatova and R. D. King, *Mach. Learn.*, 2018, **107**, 285–311.
28 L. H. Torres, B. Ribeiro and J. P. Arrais, *Expert Syst. Appl.*, 2023, **225**, 120005.
29 J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li and M. Sun, *AO*, 2020, **1**, 57–81.
30 W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande and J. Leskovec, *8th International Conference on Learning Representations, ICLR 2020*, 2020.
31 O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel and T. Langer, *Drug Discovery Today: Technol.*, 2020, **37**, 1–12.
32 Y. Wang, J. Wang, Z. Cao and A. B. Farimani, *Nat. Mach. Intell.*, 2022, **4**, 279–287.
33 X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu and H. Wang, *Nat. Mach. Intell.*, 2022, **4**, 127–134.
34 Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang and J. Huang, *arXiv*, 2020, preprint, arXiv:2007.02835, DOI: 10.48550/arXiv.2007.02835.
35 S. Wang, Y. Guo, Y. Wang, H. Sun and J. Huang, *ACM-BCB 2019 - Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019, pp. 429–436.
36 J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh and P. Das, *Nat. Mach. Intell.*, 2022, **4**, 1256–1264.
37 D. Chen, K. Gao, D. D. Nguyen, X. Chen, Y. Jiang, G. W. Wei and F. Pan, *Nat. Commun.*, 2021, **12**, 3521.
38 S. Zheng, Z. Lei, H. Ai, H. Chen, D. Deng and Y. Yang, *J. Cheminf.*, 2021, **13**, 87.
39 H. Li, D. Zhao and J. Zeng, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2022, pp. 857–867.
40 R. Benigni, C. L. Battistelli, C. Bossa, O. Tcheremenskaia and P. Crettaz, *Mutagenesis*, 2013, **28**, 401–409.
41 R. Benigni, C. L. Battistelli, C. Bossa, O. Tcheremenskaia and P. Crettaz, *ISSTOX*, 2019.
42 K. Xu, S. Jegelka, W. Hu and J. Leskovec, *7th International Conference on Learning representations, ICLR 2019*, 2019.
43 M. Defferrard, X. Bresson and P. Vandergheynst, *Adv Neural Inf Process Syst*, 2016, 3844–3852.
44 W. L. Hamilton, R. Ying and J. Leskovec, *Adv Neural Inf Process Syst*, 2017, 1025–1035.
45 C. You, J. Wang, X. Dai and Y. Wang, *Nucleic Acids Res.*, 2015, **43**, 1012–1018.
46 H. L. Wong, S. E. Murphy, M. Wang and S. S. Hecht, *Carcinogenesis*, 2003, **24**, 291–300.
47 J. Kobayashi, *Nitric Oxide - Biol. Chem*, 2018, **73**, 66–73.
48 T. Haack, L. Erdinger and G. Boche, *Mutat. Res., Genet. Toxicol. Environ. Mutagen.*, 2001, **491**, 183–193.
49 A. Furukawa, S. Ono, K. Yamada, N. Torimoto, M. Asayama and S. Muto, *Genes Environ.*, 2022, **44**, 10.
50 R. J. Turesky, *Drug Metab. Rev.*, 2002, **34**, 625–650.
51 M. Bellamri, S. J. Walmsley and R. J. Turesky, *Genes Environ.*, 2021, **43**(1), 29.
52 H. C. Becker and B. Nordén, *J. Am. Chem. Soc.*, 1999, **121**, 11947–11952.
53 E. A. Kataev, T. A. Shumilova, B. Fiedler, T. Anacker and J. Friedrich, *J. Org. Chem.*, 2016, **81**, 6505–6514.
54 S. A. Rather, M. Y. Bhat, F. Hussain and Q. N. Ahmed, *J. Org. Chem.*, 2021, **86**, 13644–13663.
55 Y. A. Yusof, Z. A. A. Hasan and Z. A. Maurad, *Int. J. Toxicol.*, 2024, **43**, 157–164.
56 A. A. Kirilchuk, A. B. Rozhenko and J. Leszczynski, *Comput. Theor. Chem.*, 2017, 1103.
57 A. K. Roy, R. K. Tyagi, C. S. Song, Y. Lavrovsky, S. C. Ahn, T. S. Oh and B. Chatterjee, *Ann. N. Y. Acad. Sci.*, 2001, **949**, 44–57.
58 L. Varticovski, D. A. Stavreva, A. McGowan, R. Raziuddin and G. L. Hager, *Mol. Cell. Endocrinol.*, 2022, **539**, 111415.
59 M. Richold, *Arch. Toxicol.*, 1988, **61**, 249–258.
60 S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang and S. H. Bryant, *Nucleic Acids Res.*, 2016, **44**, D1202–D1213.
61 K. Pant, S. Bruce, J. Sly, M. K. Laforce, S. Springer, M. Cecil, E. Andrus, E. Dakoulas, V. O. Wagner, N. J. Hewitt and R. Kulkarni, *Environ. Mol. Mutagen.*, 2016, **57**.
62 M. S. Diehl, S. L. Willaby and R. D. Snyder, *Environ. Mol. Mutagen.*, 2000, **36**.
63 K. P. Cross and D. M. DeMarini, *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 2023, p. 827.
64 N. Flamand, J. R. Meunier, P. A. Meunier and C. Agapakis-Caussé, *Toxicol. in Vitro*, 2001, **15**.
65 J. Nicolette, E. Dakoulas, K. Pant, M. Crosby, A. Kondratiuk, J. Murray, P. Sonders, R. Kulkarni, G. Jayakumar, M. Mathur, A. Patel, R. Vicente, K. Datta and K. Kolaja, *Regul. Toxicol. Pharmacol.*, 2018, **100**.
66 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Adv Neural Inf Process Syst*, 2017.
67 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, *9th International Conference on Learning representations, ICLR 2021*, 2020.
68 K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang and D. Tao, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.
69 G. Landrum, *RDKit*, 2021.
70 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
71 O. V. Egorova, N. A. Ilyushina and V. N. Rakitskii, *Toxicol. in Vitro*, 2020, **69**.