Chemical Research in Toxicology

pubs.acs.org/crt

Article

Mechanistic Task Groupings Enhance Multitask Deep Learning of Strain-Specific Ames Mutagenicity

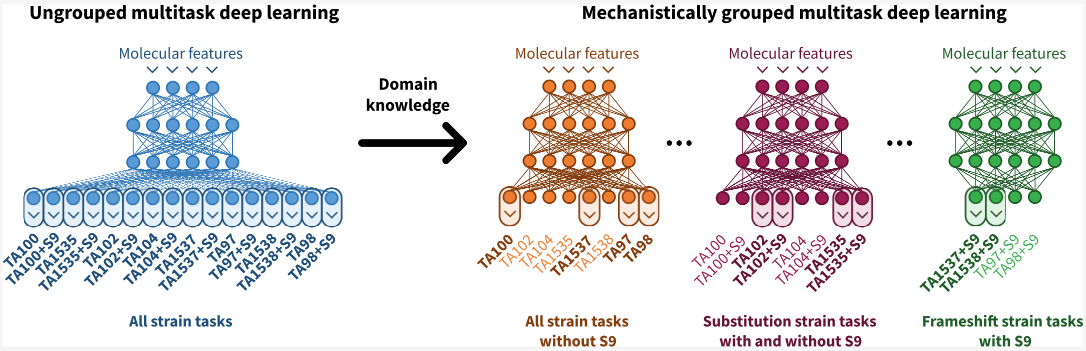
Raymond Lui,* Davy Guan, and Slade Matthews

Cite This: *Chem. Res. Toxicol.* 2023, 36, 1248−1254

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

Downloaded via UNIV OF SYDNEY on September 18, 2023 at 04:29:18 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.


Ungrouped multitask deep learning
Mechanistically grouped multitask deep learning
Molecular features
Domain knowledge
TA100 TA100+S9 TA1535 TA1535+S9 TA102 TA102+S9 TA104 TA104+S9 TA1537 TA1537+S9 TA97 TA97+S9 TA1538 TA1538+S9 TA98 TA98+S9
All strain tasks
TA100 TA102 TA104 TA1535 TA1537 TA1538 TA97 TA98
All strain tasks without S9
TA100 TA100+S9 TA102 TA102+S9 TA104 TA104+S9 TA1535 TA1535+S9
Substitution strain tasks with and without S9
TA1537+S9 TA1538+S9 TA97+S9 TA98+S9
Frameshift strain tasks with S9

ABSTRACT: The Ames test is a gold standard mutagenicity assay that utilizes various *Salmonella typhimurium* strains with and without S9 fraction to provide insights into the mechanisms by which a chemical can mutate DNA. Multitask deep learning is an ideal framework for developing QSAR models with multiple end points, such as the Ames test, as the joint training of multiple predictive tasks may synergistically improve the prediction accuracy of each task. This work investigated how toxicology domain knowledge can be used to handcraft task groupings that better guide the training of multitask neural networks compared to a naïve ungrouped multitask neural network developed on a complete set of tasks. Sixteen *S. typhimurium* ± S9 strain tasks were used to generate groupings based on mutagenic and metabolic mechanisms that were reflected in correlation data analyses. Both grouped and ungrouped multitask neural networks predicted the 16 strain tasks with a higher balanced accuracy compared with single task controls, with grouped multitask neural networks consistently featuring incremental increases in predictivity over the ungrouped approach. We conclude that the main variable driving these performance improvements is the general multitask effect with mechanistic task groupings acting as an enhancement step to further concentrate synergistic training signals united by a common biological mechanism. This approach enables incorporation of toxicology domain knowledge into multitask QSAR model development allowing for more transparent and accurate Ames mutagenicity prediction.

■ INTRODUCTION

The Ames test is a gold standard mutagenicity assay and constitutes genotoxicity test batteries worldwide.[1−3] The *in vitro* assay detects mutagenic activity by measuring colony growth of histidine auxotrophic *Salmonella typhimurium* reverse mutated by test chemicals.[4,5] There are two primary variables in the Ames protocol that enable the elucidation of the mechanisms by which a chemical can induce mutations.

First, various *S. typhimurium* mutant strains have been developed, each with a distinct mutation along their histidine operons for the detection of substitution and frameshift reverse mutations. This work focuses on eight strains (Figure 1.A) that have been comprehensively outlined by Klapacz and Gollapudi[6] and have readily available data for computational modeling. Substitution strains featuring mutations of the first histidine-synthesizing enzyme, HisG, include TA100 and TA1535 which detect transversion and transition reversions at the *hisG46* mutation and TA102 and TA104 which detect transversion and transition reversions at the *hisG428* mutation (Figure 1.A.i).[7−9] Frameshift strains featuring mutations of the downstream histidine-synthesizing enzymes, HisC and HisD, include TA98 and TA1538 which detect indel reversions at the *hisD3052* mutation and TA97 and TA1537 which detect indel reversions at the *hisD6610* and *hisC3076* mutations, respectively (Figure 1.A.ii).[7−9]

Second, a metabolic activation system is included in the Ames test protocol to simulate *in vivo* biotransformation of promutagenic chemicals. An S9 homogenate fraction is prepared from the liver tissue of various mammalian species

Received: December 1, 2022
Published: July 21, 2023

© 2023 American Chemical Society

https://doi.org/10.1021/acs.chemrestox.2c00385

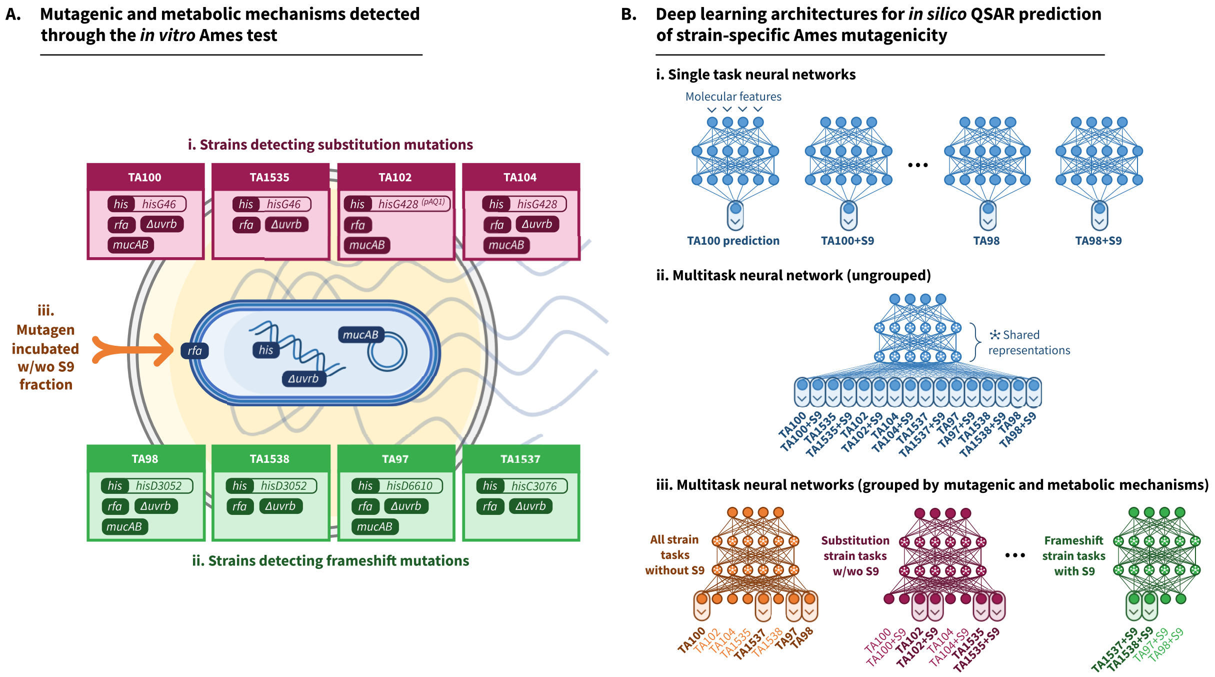


Figure 1. Overview of study background. (A) Mutagenic and metabolic mechanisms detected through the *in vitro* Ames test: (i) substitution mutations, (ii) frameshift mutations, and (iii) mutagens requiring metabolic activation. (B) Deep learning architectures for *in silico* QSAR prediction of strain-specific Ames mutagenicity: (i) single task neural networks, (ii) ungrouped multitask neural network, (iii) multitask neural networks grouped by mutagenic and metabolic mechanisms.

and then combined with cofactors to form a final S9 mix to be applied with the test chemical and *S. typhimurium* onto the agar medium.[5] Chemicals are tested with and without the S9 mix to determine if the chemical is a pro- and/or direct mutagen, respectively (Figure 1.A.iii).[8] Elucidation of both mutagenic mechanisms, through different *S. typhimurium* strains, and metabolic mechanisms, through inclusion/exclusion of the S9 mix, offers a mechanistic fingerprint of the mutagenicity of a test chemical.

Quantitative structure−activity relationships (QSARs) are mathematical models that describe the biological activity of chemicals as a function of their molecular properties. Once trained with a set of chemicals, QSAR models are used to infer the activity of new query chemicals based on their molecular properties.[10,11] Deep learning is a machine learning method which specifically employs the deep neural network algorithm to transform input molecular features into an activity prediction output through interconnected layers of neuron units comprised of numerical weights and biases.[12] Neural networks are trained to learn QSAR patterns using backpropagation where these weight and bias values are iteratively optimized to minimize a loss function quantifying the difference between the predicted and actual activities, until an ideal set of neurons is obtained which accurately transform molecular feature inputs into the correct activity prediction.[12] Neural networks can be developed to predict each *S. typhimurium* ± S9 activity label in a process defined as "single task" learning; i.e., each neural network learns and outputs a single predictive task (Figure 1.B.i).

Multitask deep learning is an approach where multiple predictive tasks are jointly learned in parallel within the same multitask neural network. By backpropagating a combined loss function for multiple tasks, the weights and biases are optimized with respect to multiple training signals which may synergistically improve the prediction accuracy of each task in a process known as inductive transfer.[13] The resulting shared neuron representation will be trained to recognize relatedness between QSAR patterns for multiple biological activities, enabling more generalizable predictions compared to a series of equivalent single task neural networks which each learn QSARs independently.[14−19] A multitask neural network developed to predict all *S. typhimurium* ± S9 activities will learn to use similarities between the various mutagenic and metabolic mechanisms to predict strain-specific mutagenicity with greater accuracy (Figure 1.B.ii).

However, with the same notion that combining training signals from related tasks improves prediction performance, the inclusion of tasks with insufficient relatedness can also impart antagonistic signals, which hinder the neural network from properly learning other tasks. These effects may be most noticeable when a given set of tasks is naïvely combined into one multitask neural network in which the loss values of antagonistic tasks are jointly backpropagated with other tasks. Task weighting is a strategy that addresses this by assigning weights to the task loss values as they are summed, effectively suppressing antagonistic noise and allowing more synergistic training signals to be backpropagated.[20,21] Task grouping is another "informed" multitask learning strategy which involves clustering a given set of tasks into explicit subsets of synergistic tasks that are used to train separate multitask neural networks in a constrained manner free from antagonistic signals.[22−25]

This work aims to investigate how toxicology domain knowledge can be used to handcraft task groupings and introduce "rule hints" that better guide the training of multitask neural networks.[26] By grouping together *S. typhimurium* ± S9 strain tasks based on their mutagenic and/or metabolic mechanisms, a series of multitask neural networks developed on these groupings can learn more specialized shared representations (Figure 1.B.iii). Accordingly, multiple predictions may be generated for each strain task if

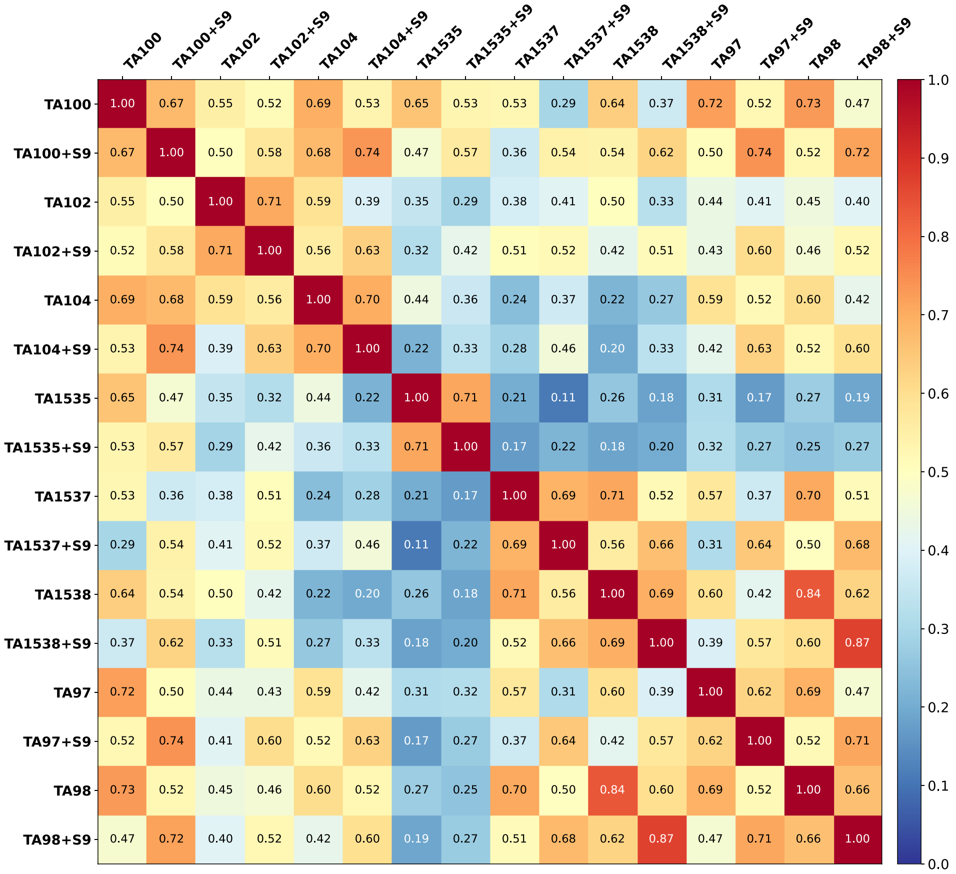


Figure 2. Correlation heatmap matrix for 16 *S. typhimurium* ± S9 strain task data sets curated for this study.

they appear in multiple groupings, since they will be used to provide hints for the canonical prediction of another strain task. Uncovering mechanistic task groupings for multitask learning is an interesting avenue as it allows the computational toxicologist to integrate their domain expertise directly into the QSAR model development process. This can bridge the gap between mechanistic-driven and data-driven modeling allowing for more transparent model development for regulatory toxicology use, and, more broadly, more insightful models in pharmacology and toxicology.

■ METHODS

Data Sources and Curation. This study centers around eight *S. typhimurium* strains (in alphanumeric order): TA100, TA102, TA104, TA1535, TA1537, TA1538, TA97, TA98—each strain with and without S9 fraction for a total of 16 predictive strain tasks. These strains were selected as they have been well-documented by Klapacz and Gollapudi[6] and were readily accessible in the OECD QSAR Toolbox.[27,28] Moreover, these strains offer equal consideration of mutagenic (substitutions vs frameshifts) and metabolic mechanisms (with and without S9) while covering a large range of data set sizes and class balance. The ISSSTY, OASIS, EFSAP, and MHLW data sets were extracted from the QSAR Toolbox 4.4.1 and then concatenated to form a master data set for this study.

Ames test end point curation involved removing chemicals that contained inconclusive, equivocal, or blank/uncategorised classifications. Chemical structure curation closely followed the protocol of Fourches *et al.*:[29] structure filtering, including the removal of inorganics, organometallics, and salt and solvent counterions, as well as the consolidation of mixtures into their active constituent; structure cleaning, including neutralization and final recalculation and standardization of all 2D coordinates; consolidation of duplicate entries from the four data sets using the standardized structures and the removal of chemicals possessing the same structures but conflicting Ames end points; final manual inspection and clean up, including random shuffling and assignment of chemical identifiers. End point and chemical structure curation were conducted using KNIME 4.3.2 (with KNIME RDKit Extensions 4.2.0.v202103031419) and ChemAxon Standardizer 21.7.0.

Molecular Featurization and Data Partitioning. Binary Morgan fingerprints, a reimplementation of extended connectivity fingerprints,[30] with length 1,024 bits and radius of 2 bond lengths were computed for all chemical structures.

Partitioning of the data into 70% training, 20% validation, and 10% test sets was conducted by using two approaches to test the performance of the neural network QSAR models within and outside their applicability domains. In-domain partitioning involved dissimilarity-based compound selection to select test and validation instances evenly across the chemical space of the data set; this assesses the ability for a QSAR model to reliably predict within the documented chemical space in which it was trained. Out-of-domain partitioning involved scaffold splitting to select test instances containing unique Bemis-Murcko scaffolds that would not be present in the remaining training set; this assesses the ability for a QSAR model to generalize on unseen chemical species on which it was not trained. Molecular featurization and data partitioning were performed using KNIME 4.3.2 (with KNIME RDKit Extensions 4.2.0.v202103031419).

The number of validation and test molecules within the training set applicability domain for each of the approaches was computed with a distance from the centroid measure. Distances of all training molecules to a centroid confidence area around the grand mean were calculated and the 95th percentile set as the domain threshold; validation and test instances below the threshold are considered in domain. This was performed with the Applicability Domain Toolbox 1.0[31] in MATLAB R2022b.

Table 1. Averaged In-Domain Test Set Performance Across 16 *S. typhimurium* ± S9 Strain Tasks, Predicted by Neural Networks Developed with Single Task (STL), Ungrouped Multitask (uMTL), Grouped Multitask (gMTL), and Overall Learning Architectures[a]

Deep learning architecture	Balanced accuracy	Sensitivity	Specificity	ROC AUC
STL	0.610 (0.536–0.689)	0.273 (0.128–0.423)	0.948 (0.911–0.979)	0.745 (0.635–0.842)
uMTL	0.754 (0.662–0.842)*	0.571 (0.391–0.738)*	0.937 (0.895–0.972)	0.894 (0.831–0.944)*
gMTL	0.771 (0.676–0.857)*	0.603 (0.425–0.767)**	0.940 (0.898–0.975)	0.875 (0.800–0.934)*
Overall	0.783 (0.690–0.866)**	0.631 (0.451–0.783)**	0.935 (0.892–0.970)	0.888 (0.818–0.942)*

[a] 95% (shown) and 83% confidence intervals were averaged across the 16 strain tasks (**nonoverlap with STL 95% CIs, p ≈ 0.01; *non-overlap with STL 83% CIs, p ≈ 0.05).

Neural Network Learning Architectures. The primary aim of this study was to compare the ability for three neural network decoder architectures, single task learning (STL), ungrouped multitask learning (uMTL), and grouped multitask learning (gMTL), to learn strain-specific Ames mutagenicity end points (Figure 1.B). A total of 16 single task (STL) neural networks that model each strain task independently were developed as controls. One ungrouped multitask (uMTL) neural network that simultaneously models all 16 strain tasks was developed to act as a baseline for the current state of the art multitask QSAR models. A total of eight grouped multitask (gMTL) neural networks were developed using task groupings designed around the mechanistic details of the Ames protocol: all strains with S9 (eight strain tasks), all strains without S9 (eight strain tasks), substitution strains ± S9 (eight strain tasks), frameshift strains ± S9 (eight strain tasks), substitution strains with S9 (four strain tasks), frameshift strains with S9 (four strain tasks), substitution strains without S9 (four strain tasks), and frameshift strains without S9 (four strain tasks). All neural networks were developed with PyTorch 1.12.1 and Python 3.7.13.

Hyperparameter Optimization. A grid search was performed to optimize the following hyperparameters of each neural network: number of layers (2, 3, 4); decoder "shape" governed by the number of neurons per layer (linear, top skewed, bottom skewed, outer skewed, inner skewed), learning rate (0.001, 0.0001), and number of training epochs (100, 150, 200). Binary cross entropy was used as the loss function, with the Adam algorithm used as the optimizer. The 20% validation set was used to assess the balanced accuracy of each grid configuration, and the most optimally configured neural network (i.e., highest balanced accuracy) for each strain task was selected for final testing. Hyperparameter optimization was conducted with Optuna 3.0.0 and Python 3.7.13.

Test Performance. The 25 optimized neural networks were each retrained with a combined set of their 70% train and 20% validation sets then tested on the 10% test set. Balanced accuracy was used as the primary performance metric, as its consideration of sensitivity and specificity of final predictions reflects the practical use of these models for toxicological assessment. Area under the receiver operator characteristic curve (ROC AUC) was used as a secondary performance metric. To gauge the variability of each neural network on the study chemical space, 1,000 bootstrap sets were sampled from the test set, and the two metrics were calculated for each bootstrap set. 95% and 83% confidence intervals were computed with the resulting bootstrapped distributions.[32] Test performance analysis was conducted using Python 3.7.13. The hyperparameter optimization and final testing protocols were repeated twice for each of the in-domain and out-of-domain approaches.

■ RESULTS AND DISCUSSION

Mutagenic and Metabolic Mechanisms Are Reflected in Strain-Specific Ames Test Data. Figure 2 reports the pairwise correlation of positive and negative Ames mutagenicity classes for each strain task data set. Higher correlation between substitution strain tasks was noted toward the top left corner, while higher correlation between frameshift strain tasks was noted toward the bottom right corner. Strain tasks with/without S9 are generally more correlated with other S9/non-S9 strain tasks, respectively, manifesting as a partial checkerboard pattern in the heatmap. Notably, high correlation is seen between TA98±S9 and TA1538±S9 strains most likely due to underlying similarities in their genetic makeup, and thus sensitivity to mutagens (Figure 1.A.ii). Conversely, low correlation is seen between TA1535 and TA1537+S9, most likely due to distinct differences in both mutagenic and metabolic mechanisms (Figure 1.A). These trends demonstrate that mutagenic and metabolic mechanisms are reflected in the total data structure and support the investigation of domain knowledge-driven mechanistic task groupings for the development of multitask neural networks.

Strain-Specific Ames Mutagenicity Is Better Predicted with Multitask Learning. Ungrouped multitask (uMTL) neural networks were generally more predictive than single task (STL) neural networks. In-domain test results revealed uMTL predicted with an averaged balanced accuracy of 0.754 across all 16 strain tasks with nonoverlapping 83% confidence intervals compared to 0.610 for STL (Table 1). Specifically, 15 of the 16 strain tasks were better predicted with uMTL compared to STL controls, of which 11 were significantly better (five at the 95% confidence level and six at the 83% confidence level) (Table S3). Out-of-domain test results delivered similar trends, with uMTL predicting with a task-averaged balanced accuracy of 0.716 compared to 0.700 for STL (Table 2). Specifically, 12 of the 16 strain tasks were better predicted with uMTL compared with STL controls (Table S3).

Grouped multitask (gMTL) neural networks demonstrated higher predictivity over single task (STL) neural networks. In-domain test results show gMTL predicted with an averaged balanced accuracy of 0.771 across all 16 strain tasks with nonoverlapping 83% confidence intervals compared with 0.610 for STL (Table 1). Specifically, all 16 strains were better predicted with gMTL compared to STL controls, of which 15 were significantly better (seven at the 95% confidence level and eight at the 83% confidence level) (Table S3). Out-of-domain test results displayed similar trends, with gMTL predicting with a task-averaged balanced accuracy of 0.745 compared to 0.700 for STL (Table 2). Specifically, 13 of the 16 strain tasks were better predicted with gMTL compared with STL controls, of which one was significantly better at the 83% confidence level (Table S3).

These results reproduce the established conclusion that multitask learning, whether ungrouped or grouped, delivers higher predictivity over its single task counterparts. In particular, this study demonstrates that multitask learning is an ideal learning framework for strain-specific Ames mutagenicity QSAR. The interrelatedness between the mutagenic and metabolic mechanisms of the strain tasks allows the

Table 2. Averaged Out-of-Domain Test Set Performance Across 16 *S. typhimurium* ± S9 Strain Tasks, Predicted by Neural Networks Developed with Single Task (STL), Ungrouped Multitask (uMTL), Grouped Multitask (gMTL), and Overall Learning Architectures[a]

Deep learning architecture	Balanced accuracy	Sensitivity	Specificity	ROC AUC
STL	0.700 (0.592−0.813)	0.448 (0.238−0.675)	0.951 (0.913−0.979)	0.799 (0.664−0.915)
uMTL	0.716 (0.602−0.828)	0.542 (0.317−0.749)	0.890 (0.834−0.943)	0.823 (0.713−0.913)
gMTL	0.745 (0.628−0.858)	0.575 (0.347−0.788)	0.915 (0.863−0.957)	0.832 (0.721−0.923)
Overall	0.756 (0.638−0.865)	0.599 (0.367−0.801)	0.913 (0.861−0.958)	0.832 (0.720−0.923)

[a] 95% (shown) and 83% confidence intervals were averaged across the 16 strain tasks.

synergistic incorporation of training hints in multitask neural networks that boost predictive performance compared to independent single task neural networks for each strain task. Furthermore, these results confirm that the multitask effect is still active and potent despite using smaller four and eight strain task subsets of the original set of 16.

Mechanistic Task Groupings Enhance the Multitask Effect. Grouped multitask (gMTL) neural networks were generally more predictive than ungrouped multitask (uMTL) neural networks. In-domain test results revealed gMTL predicted with an averaged balanced accuracy of 0.771 across all 16 strain tasks compared to 0.754 for uMTL (Table 1). Specifically, 10 of the 16 strain tasks were better predicted with gMTL compared with uMTL (Table S3). Out-of-domain test results displayed similar trends, with gMTL predicting with a task-averaged balanced accuracy of 0.745 compared with 0.716 for uMTL (Table 2). Specifically, 10 of the 16 strain tasks were better predicted with gMTL compared to uMTL (Table S3).

Increases in predictivity from uMTL to gMTL were not as significant as the transition from STL to uMTL. This suggests that MTL itself, purely defined as the grouping and prediction of more than one task in a neural network, is the significant factor which improves predictive performance over STL. However, the extra step of explicitly searching for optimal task groupings, as demonstrated with the gMTL variable in this study, serves more as a subsequent optimization or enhancement step to maximize any unfound predictive performance from existing ungrouped multitask models.

These enhancements may stem from the task groupings having a concentrating effect on multitask learning. By selecting a specific set of tasks to develop grouped multitask neural networks, antagonistic training signals from other tasks that would otherwise be present in an ungrouped multitask neural network are filtered out, enabling incremental increases in predictivity. While this was generally observed for 10 of the 16 strain tasks as aforementioned, the synergistic effects of task grouping were most apparent for five strain tasks where uMTL could not outperform STL controls (TA1535 in in-domain testing and TA100+S9, TA1535, TA1535+S9, and TA1537 in out-of-domain testing) (Table S3). Of these five strain tasks, gMTL reversed the decline in balanced accuracy for all but one (TA1537), demonstrating that using groupings may better aid QSAR learning than naïve uMTL for certain tasks.

In our study, the best mechanistic task grouping for each grouped multitask neural network for each strain task was not generally definitive between in-domain and out-of-domain tests (Table S4). This may indicate that the minimum requisite for mechanistic task grouping to be effective is a common mechanism that unites tasks as a grouping; the exact combination of tasks may depend on the inherent structure of the input training and test data. Nonetheless, there were several strain tasks in our study that were best predicted by the same grouped multitask neural network regardless of testing strategy that may shed some potential mechanistic interpretability. The "substitution strains ± S9" grouping resulted in a grouped multitask neural network that best predicted the substitution strains TA102, TA102+S9, and TA1535+S9, the "all strains without S9" grouping best predicted the non-S9 frameshift strains TA1537 and TA97, and "the frameshift strains ± S9" grouping best predicted the S9 frameshift strain TA97+S9.

An Overall Approach to Task Groupings for Multitask Learning. The notion of an overall/holistic approach to task groupings is explored where instead of strictly defining STL as modeling one task, uMTL as modeling *n* tasks (16, in our study), and gMTL as modeling subsets of between 1 and *n* tasks, a "grouping" can be any size from 1 to *n*. In this manner, there is no hard delineation between STL, uMTL, or gMTL as different learning architectures, but rather all neural networks are free to use any number of tasks that best complement the canonical prediction of a particular task. The best learning architecture for each strain task, whether STL, uMTL, or gMTL (i.e., bolded models in Table S3) were selected as the "overall" learning architecture for that strain task.

The overall learning architecture was the highest performing approach in this study. By combining the best tested architecture for each of the strain tasks, balanced accuracy was further improved across all 16 strain tasks. In-domain test results show the overall approach predicted with an averaged balanced accuracy of 0.783 across all 16 strain tasks with nonoverlapping 95% confidence intervals compared to 0.610 for STL, as well as outperforming gMTL (0.771) and uMTL (0.754) (Table 1). Similarly, out-of-domain test results maintained this rank order, with the overall approach achieving 0.756 task-averaged balanced accuracy, over gMTL (0.745), uMTL (0.716), and STL (0.700) (Table 2). These results highlight the importance of considering how predictive tasks are related and that strategically using STL, uMTL, and gMTL together can maximize predictivity of multitask QSAR models.

Future Directions for Strain-Specific Ames Mutagenicity QSAR. The work presented here involves applying toxicology domain knowledge to handcraft task groupings based on mutagenic and metabolic mechanisms. There exist alternative task grouping strategies, most recently task affinity grouping devised by Fifty *et al.*[25] which employs a data-driven methodology of computing pairwise intertask affinity values from the training dynamics of an ungrouped multitask neural network and then searching for task groupings which maximize total intertask affinity. An interesting avenue of research would be to compare task affinity groupings with mechanistic task groupings to observe whether a data-driven grouping would recreate similar groupings defined by natural biological mechanisms. Nonetheless, should a deep learning practitioner be limited in the time-cost associated with exhaustive combinatorial searches that usually accompany data-driven methods, our work presents mechanistic task groupings as a viable, more transparent alternative as well as a practical use of the domain knowledge that a toxicologist will have developed over the course of their education and career.

A limitation of this study was the data sets that were retrieved from the OECD QSAR Toolbox, which do not contain explicit molecular structures for the metabolites produced by the S9 incubation. We hypothesize that the test chemicals were simply added to the S9, incubated, and then transferred directly to the Ames assay with no characterization of produced metabolites. This results in data set entries containing the same parent molecular structure across S9 and non-S9 end points, potentially making it difficult for neural networks, or any machine learning algorithm for that matter, to learn QSARs between one parent molecular structure and the S9 and non-S9 end points concurrently assigned to it. The only true solution is to check whether the authors of these *in vitro* Ames test studies identified any metabolite structures, and if so, which metabolites were mutagenic positive.

Finally, it is important to note that the learning algorithm is only half the story of QSARs; the other half being the molecular features that inherently feed into the algorithm. This study has demonstrated how optimization of a highly predictive deep learning method can maximize the utilization of data. As predictive performance plateaus and insignificant gains begin to occur, it is important to consider the bigger picture of QSAR modeling and ensure that quality data is being fed into machine learning models. For example, the calculation of mechanistic electronic descriptors that are highly representative of the mutagen-DNA interaction[33] can ensure that there are realistic QSARs for the neural network to learn. Alternatively, extending the deep learning paradigm to include data-driven encodings, such as graph convolutional embeddings, may allow the neural networks to extract more detailed signals from the underlying data structure.[34] Conclusively, these improvements, in conjunction with the presented work demonstrating the potential for mechanistic task groupings to enhance the multitask effect, serve as complementary optimization steps for the design of new deep learning toxicological QSAR models.

■ ASSOCIATED CONTENT

Data Availability Statement

All data, source code, final models, and full results are available in the following GitHub repository: https://github.com/luiraym/gmtames.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemrestox.2c00385.

> Tables reporting descriptive statistics and percentage overlap for each strain task data set, neural network test set performance for each strain task, and best mechanistic task groupings for each strain task (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Raymond Lui − *Computational Pharmacology and Toxicology Laboratory, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW 2006, Australia;* orcid.org/0000-0003-4673-9030; Email: rlui9522@uni.sydney.edu.au

Authors

Davy Guan − *Computational Pharmacology and Toxicology Laboratory, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW 2006, Australia;* orcid.org/0000-0001-6290-3166

Slade Matthews − *Computational Pharmacology and Toxicology Laboratory, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW 2006, Australia;* orcid.org/0000-0002-1652-543X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrestox.2c00385

Funding

We thank The University of Sydney Faculty of Medicine and Health, in partnership with the Australian Industrial Chemicals Introduction Scheme (AICIS), for their support and guidance through the Ignition Seed Fund Grant. We also thank NVIDIA for their Titan V sponsorship through the Accelerated Data Science GPU Grant.

Notes

The authors declare no competing financial interest.

■ REFERENCES

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

(2) EMA. *ICH guideline M7(R1) on assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk*; 2017.

(3) OECD. *Test No. 471: Bacterial Reverse Mutation Test*; 2020.

(4) Ames, B. N.; Lee, F. D.; Durston, W. E. An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proc. Natl. Acad. Sci. U. S. A.* **1973**, *70* (3), 782−6.

(5) Ames, B. N.; Durston, W. E.; Yamasaki, E.; Lee, F. D. Carcinogens are Mutagens: A Simple Test System Combining Liver Homogenates for Activation and Bacteria for Detection. *Proc. Natl. Acad. Sci. U. S. A.* **1973**, *70* (8), 2281−2285.

(6) Klapacz, J.; Gollapudi, B. B. Genetic Toxicology. In *Casarett & Doull's Toxicology: The Basic Science of Poisons*, 9th ed.; Klaassen, C. D., Ed.; McGraw-Hill Education: New York, NY, 2019.

(7) Klaassen, C. D. *Casarett & Doull's Toxicology: The Basic Science of Poisons*, 9th ed.; McGraw-Hill Education: 2018.

(8) Maron, D. M.; Ames, B. N. Revised methods for the Salmonella mutagenicity test. *Mutation Research/Environmental Mutagenesis and Related Subjects* **1983**, *113* (3), 173−215.

(9) Kulis-Horn, R. K.; Persicke, M.; Kalinowski, J. Histidine biosynthesis, its regulation and biotechnological application in Corynebacterium glutamicum. *Microbial Biotechnology* **2014**, *7* (1), 5−25.

(10) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977−5010.

(11) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49* (11), 3525−3564.

(12) Xu, Y., Deep Neural Networks for QSAR. In *Artificial Intelligence in Drug Design*; Heifetz, A., Ed.; Springer US: New York, NY, 2022; pp 233−260.

(13) Caruana, R. Multitask Learning. *Machine Learning* **1997**, *28* (1), 41−75.

(14) Varnek, A.; Gaudin, C.; Marcou, G.; Baskin, I.; Pandey, A. K.; Tetko, I. V. Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net Approaches to Model Tissue-Air Partition Coefficients. *J. Chem. Inf. Model.* **2009**, *49* (1), 133−144.

(15) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55* (2), 263−274.

(16) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science* **2016**, DOI: 10.3389/fenvs.2015.00080.

(17) Hughes, T. B.; Dang, N. L.; Miller, G. P.; Swamidass, S. J. Modeling Reactivity to Biological Macromolecules with a Deep Multitask Network. *ACS Central Science* **2016**, *2* (8), 529−537.

(18) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57* (8), 2068−2076.

(19) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57* (10), 2490−2504.

(20) Gong, T.; Lee, T.; Stephenson, C.; Renduchintala, V.; Padhy, S.; Ndirango, A.; Keskin, G.; Elibol, O. H. A Comparison of Loss Weighting Strategies for Multi task Learning in Deep Neural Networks. *IEEE Access* **2019**, *7*, 141627−141632.

(21) Humbeck, L.; Morawietz, T.; Sturm, N.; Zalewski, A.; Harnqvist, S.; Heyndrickx, W.; Holmes, M.; Beck, B. Don't Overweight Weights: Evaluation of Weighting Strategies for Multi-Task Bioactivity Classification Models. *Molecules* **2021**, *26* (22), 6959.

(22) Kang, Z.; Grauman, K.; Sha, F. Learning with whom to share in multi-task feature learning. *Proceedings of the 28th International Conference on International Conference on Machine Learning*; Omnipress: Bellevue, Washington, USA, 2011; pp 521−528.

(23) Kumar, A.; Daumé, H. Learning task grouping and overlap in multi-task learning. *Proceedings of the 29th International Coference on International Conference on Machine Learning*; Omnipress: Edinburgh, Scotland, 2012; pp 1723−1730.

(24) Standley, T.; Zamir, A.; Chen, D.; Guibas, L.; Malik, J.; Savarese, S. Which tasks should be learned together in multi-task learning? *Proceedings of the 37th International Conference on Machine Learning*; JMLR.org: 2020; Article 846.

(25) Fifty, C.; Amid, E.; Zhao, Z.; Yu, T.; Anil, R.; Finn, C. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems* **2021**, *34*, 27503−27516.

(26) Caruana, R., A Dozen Tricks with Multitask Learning. In *Neural Networks: Tricks of the Trade*, Second ed.; Montavon, G., Orr, G. B., Müller, K.-R., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; pp 163−189.

(27) Dimitrov, S. D.; Diderich, R.; Sobanski, T.; Pavlov, T. S.; Chankov, G. V.; Chapkanov, A. S.; Karakolev, Y. H.; Temelkov, S. G.; Vasilev, R. A.; Gerova, K. D.; Kuseva, C. D.; Todorova, N. D.; Mehmed, A. M.; Rasenberg, M.; Mekenyan, O. G. QSAR Toolbox - workflow and major functionalities. *SAR and QSAR in Environmental Research* **2016**, *27* (3), 203−219.

(28) Schultz, T. W.; Diderich, R.; Kuseva, C. D.; Mekenyan, O. G. The OECD QSAR Toolbox Starts Its Second Decade. In *Computational Toxicology: Methods and Protocols*; Nicolotti, O., Ed.; Springer New York: New York, NY, 2018; pp 55−77.

(29) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189−1204.

(30) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(31) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17* (5), 4791−4810.

(32) Austin, P. C.; Hux, J. E. A brief note on overlapping confidence intervals. *Journal of Vascular Surgery* **2002**, *36* (1), 194−195.

(33) Townsend, P. A.; Grayson, M. N. Density Functional Theory Transition-State Modeling for the Prediction of Ames Mutagenicity in 1,4 Michael Acceptors. *J. Chem. Inf. Model.* **2019**, *59* (12), 5099−5103.

(34) Montanari, F.; Kuhnke, L.; Ter Laak, A.; Clevert, D.-A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules* **2020**, *25* (1), 44.

Recommended by ACS

Identifying Differences in the Performance of Machine Learning Models for Off-Targets Trained on Publicly Available and Proprietary Data Sets

Aljoša Smajić, Gerhard F. Ecker, *et al.*

JULY 13, 2023
CHEMICAL RESEARCH IN TOXICOLOGY

READ

Multiple Instance Learning Improves Ames Mutagenicity Prediction for Problematic Molecular Species

Samuel V. Feeney, Slade Matthews, *et al.*

JULY 21, 2023
CHEMICAL RESEARCH IN TOXICOLOGY

READ

High Throughput Read-Across for Screening a Large Inventory of Related Structures by Balancing Artificial Intelligence/Machine Learning and Human Knowledge

Chihae Yang, Tara Barton-Maclaren, *et al.*

JULY 03, 2023
CHEMICAL RESEARCH IN TOXICOLOGY

READ

MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information

Wouter Heyndrickx, Hugo Ceulemans, *et al.*

AUGUST 29, 2023
JOURNAL OF CHEMICAL INFORMATION AND MODELING

READ

Get More Suggestions >