Position Papers of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 93–97
DOI: 10.15439/2025F1171
ISSN 2300-5963 ACSIS, Vol. 44

Multitask Learning for Six-Pack Toxicity Prediction

Chun-Wei Tung*, Chia-Chi Wang†, Run-Hsin Lin*, Shan-Shan Wang*
0000-0003-3011-8440, 0000-0002-1272-1513, 0000-0002-1205-2874, 0000-0001-6163-120X
*Institute of Biotechnology and Pharmaceutical Research, National Health Research Institutes,
35 Keyan Road, Zhunan, Miaoli, Taiwan
Email: {cwtung, rhlin, shanwang}@nhri.edu.tw
†Department and Graduate Institute of Veterinary Medicine, School of Veterinary Medicine,
National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, Taiwan
Email: ccwang@ntu.edu.tw

***Abstract*—The assessment of the six-pack toxicity, the crucial six systems and organ toxicities, is vital for ensuring the safe use of chemicals. Computational models capable of providing reliable predictions are acceptable for regulatory use to replace animal testing. However, data scarcity issues hindered the development of prediction models. This study proposed the first application of multitask learning to the six-pack toxicity for addressing data scarcity issues. Five algorithms were implemented and compared. Results showed that the distinct chemical space of tasks impedes the learning of shared representation of conventional algorithms, with performance worse than baseline models. In contrast, the MTForestNet algorithm built on a biological readacross concept performed best, with 3.1% and 3.3% improvement on AUC and accuracy, respectively. These findings demonstrate that biologically informed multitask learning can effectively overcome data scarcity and enhance toxicity prediction.**

***Index Terms*—multitask learning, biological readacross, six-pack toxicity, distinct chemical space, MTForestNet.**

I. Introduction

TOXICITY prediction plays a pivotal role in the early stages of drug discovery and chemical safety assessment. Among the large number of toxicity endpoints for testing, there is a suite of six key toxicity endpoints, commonly known as the 'six-pack': acute oral toxicity, acute dermal toxicity, acute inhalation toxicity, skin irritation, eye irritation, and skin sensitization. These endpoints provide important information about the system and organ toxicity of testing chemicals and are crucial for regulatory decision-making and risk assessment of industrial chemicals, pharmaceuticals, and consumer products.

The assessment of the six-pack toxicity is traditionally based on animal testing. However, the traditional experimental approaches to assess these toxicities are time-consuming and costly, and ethical concerns are raised due to extensive animal testing. In recent years, computational methods, particularly machine learning, have emerged as powerful alternatives for toxicity prediction. Several studies have developed machine learning models for predicting the six-pack toxicity [1], [2], [3].

Despite the efforts made by the scientific community, dataset size poses a major limitation on advancing the prediction performance of six-pack toxicity. It is unlikely to have a huge increase in the testing data due to the high cost and labor-intensive experiments. Compared to the conventional single-task models developed by previous studies, multitask learning algorithms capable of leveraging the shared knowledge among relevant learning tasks can be promising solutions to the prediction of six-pack toxicity.

Several multitask learning algorithms have been proposed and implemented with success for toxicity prediction. For example, three deep learning-based multitask learning algorithms, including conventional, bypass, and progressive multitask learning algorithms, were shown to outperform single-task models for several drug development-relevant datasets [4]. The three algorithms were implemented as an open-sourced library, DeepChem [4]. In addition, AutoGluon-Tabular [5], a powerful automated machine learning algorithm, implemented a multilabel learning algorithm that can be potentially useful for multitask learning. By leveraging shared knowledge, multitask learning can improve prediction accuracy, especially when training data for individual tasks is limited or imbalanced.

While the abovementioned algorithms performed well on the benchmark datasets, each dataset contains a large portion of shared training samples among tasks in the dataset [6], [7], [8], [9], [10], and therefore ensures the successful transfer of knowledge among tasks. However, the majority of learning tasks of toxicity datasets are with distinct chemical spaces containing little or no shared samples, which hinders the application of the DeepChem-based methods. To solve the issue of distinct chemical space, MTForestNet was proposed with a progressive multitask learning strategy concatenating chemical features and outputs of individual classifiers of tasks from

Topical area: Advanced Artificial Intelligence in Applications

the previous layer for accuracy improvement [11]. The algorithm showed superior performance compared to other algorithms on the zebrafish toxicity dataset, consisting of 48 tasks, and is expected to be useful for other toxicity datasets with distinct chemical space.

This study explores the application of multitask learning models to predict all six toxicity endpoints concurrently. A total of five algorithms were implemented and compared for their application to the prediction of six-pack toxicity. Results showed that the model based on MTForestNet performed best on predicting the independent test dataset with the highest average area under the receiver operating characteristic curve (AUC) value of 0.825, showing a 3.1% improvement over single-task models. The other models showed no improvement or much worse performance. The low percentage of shared samples among the six tasks further supports the usefulness of MTForestNet on predicting chemical toxicity.

II. Materials and Methods

A. Dataset

The six-pack toxicity dataset was obtained from a previous study [3] collecting the largest dataset of toxicity data from the U.S. National Toxicology Program and OECD eCHem-Portal. The dataset was randomly divided into 70% training, 10% validation, and 20% test sets for model training, tuning, and independent test, respectively. A summary of the dataset is shown in Table I. In this study, the widely used extended connectivity fingerprint (ECFP) with a diameter of 6 was utilized to encode the chemical feature vector. Specifically, a 1024-dimensional vector representing the binary occurrence of specific substructures was utilized for machine learning.

B. Single-task learning algorithm

In this study, random forest [12] was utilized as the baseline algorithm for evaluating the performance improvement based on multitask learning algorithms. Random forest was extensively used and proved to have robust and high performance in a large number of cheminformatics tasks [13], [14], [15], [16], [17]. The parameters utilized to implement random forest classifiers were set as follows: mtry=log2(total feature number) and n_estimators=500. With the parameters, a single-task random forest classifier with 500 trees and log2(total feature number) features sampled from all features was developed for each task.

C. Multitask learning algorithms

Five algorithms were implemented and compared in this study. Accuracy was utilized as the objective function to tune or select models based on the validation sets for all algorithms. DeepChem package [4] was utilized to implement three multitask learning algorithms of multitask network (DC_MTN), progressive network (DC_Progressive), and bypass network (DC_Bypass). DC_MTN incorporates shared layers for learning a joint representation of all tasks with six separate output layers, each corresponding to a specific task. DC_Progressive prevents catastrophic forgetting by adding a new column for each task and using lateral connections to transfer knowledge from previously learned tasks. DC_Bypass combines the learnable shared representation and a column of weights that bypass the shared representation for each task. The hyperparameters of the three networks were set as follows: learning_rate=0.001; dropouts=[0.20, 0.10, 0.05]; layer_sizes=[400, 200, 100]; penalty=0.001; weight_decay_penalty_type='l2'.

The multilabel learning algorithm of AutoGluon-Tabular trained an individual model for each label, with the inclusion of previous labels as features. In this way, the dependence of labels can be modeled. The default setting of AutoGluon-Tabular was applied in this study with eight classifiers, including two neural networks based on Torch and FastAI, LightGBM boosted trees, CatBoost boosted trees, XGBoost, random forest, extremely randomized trees, and k-nearest neighbors were automatically trained and stacked to achieve the highest performance on the validation set. The parameter of auto_stack was set to true for automatic model stacking in the model development. Medium (AG_Medium) and best (AG_Best) quality models were built for performance comparison using the quality parameter.

MTForestNet was proposed to deal with the distinct chemical space of tasks with little or no shared samples. The idea is based on the biological data-based read-across, where the label (target endpoint) of chemicals tends to be similar if the bioactivity profile of chemicals is similar [18], [19], [20]. MTForestNet utilized random forest as a base learner for building models, each for a task. The predicted outputs of single-task models were then fed into the next layer, where the feature vector was refined to concatenate both the chemical

TABLE I.
OVERVIEW OF DATASET SAMPLE SIZES

Task	Toxic/Nontoxic	Training	Validation	Test
Acute Dermal Toxicity	870/939	1266	181	362
Acute Inhalation Toxicity	436/428	604	87	173
Acute Oral Toxicity	6391/4723	7779	1112	2223
Eye Irritation	1824/1841	2565	367	733
Skin Irritation	1315/1311	1837	263	526
Skin Sensitization	1510/1256	1935	277	554

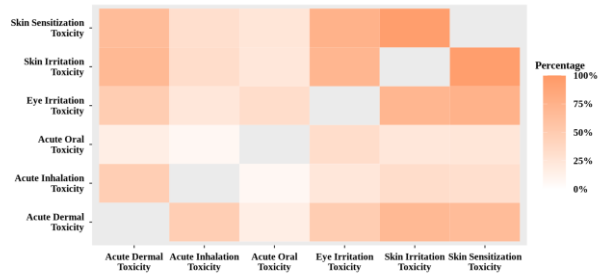

Fig. 1 The percentage of shared chemicals for each pair of tasks

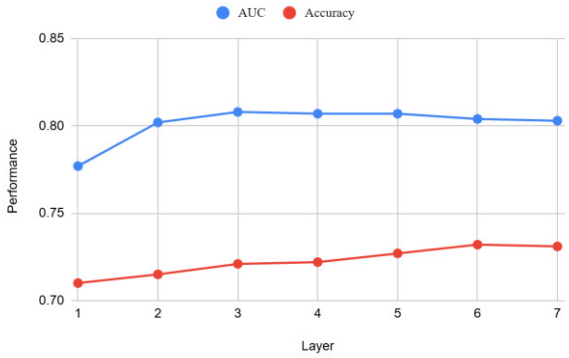

Fig. 2 The validation performance of MTForestNet

fingerprint and the six outputs from the models of the previous layer. The validation set was utilized to determine the size of the model giving the highest validation performance.

D. Hardware

The experiments were conducted in a computer equipped with two Intel® Xeon® Gold 6330, one NVIDIA RTX A6000, and 2 TiB RAM. The operating system is Ubuntu 22.04.

III. Results

A. Tasks with low percentages of shared chemicals

The percentages of shared chemicals among tasks were first analyzed to give an overview of the similarity of the six tasks. As shown in Fig.1, overall medium to low percentages of shared chemicals among tasks indicated that the six datasets lack sufficient information for learning a shared representation. The two skin-relevant tasks of skin sensitization and skin irritation shared the highest percentages of samples, where 94.94% of chemicals have both labels. The task of acute oral is associated with the lowest percentage of shared chemicals of 7.77% and 16.28% for acute inhalation and acute dermal, respectively. Among the 15 pairs of tasks, 5 pairs of tasks are associated with a percentage of shared samples less than or equal to 30%. Only 3 pairs of tasks are associated with a percentage of shared chemicals greater than or equal to 70%. The average percentage of samples shared in all pairs of tasks is 44.45%. In summary, the low percentages of shared chemicals may hinder the learning of shared representation for conventional multitask algorithms.

B. Validation performance

The application of multitask learning algorithms for predicting six-pack toxicity includes three steps of model training based on the training sets, model tuning/validation based on the validation set, and model testing using the test sets. This section provides the validation results of the implemented models. The detailed performance comparison is shown in Table II. The baseline models based on random forest provide reasonably good performance for all tasks, with an average AUC and accuracy of 0.777 and 0.711, respectively.

The validation performance of the three DeepChem-based models is much worse than that of the baseline models, with at least a 10% decrease in the average AUC. The average AUC and accuracy values are 0.673 and 0.545 for DC_MTN, 0.659 and 0.556 for DC_Bypass, and 0.600 and 0.395 for DC_Progressive, respectively. As the chemical spaces are distinct for each task, the low performance of DeepChem-based models is expected.

The AutoML models based on AutoGluon-Tabular provide slightly worse performance compared to the random forest.

TABLE II.
VALIDATION PERFORMANCE

Model	Acute Dermal Toxicity	Acute Inhalation Toxicity	Acute Oral Toxicity	Eye Irritation	Skin Irritation	Skin Sensitization
Random forest	0.773/0.680	0.794/**0.770**	0.840/0.761	0.729/0.665	0.803/0.730	0.724/0.657
MTForestNet	**0.813**/0.713	**0.833**/**0.770**	0.829/**0.772**	**0.758**/**0.689**	**0.847**/**0.768**	**0.746**/**0.679**
AG_Medium	0.773/0.707	0.723/0.690	0.841/0.763	0.708/0.649	0.804/0.722	0.724/0.671
AG_Best	0.791/**0.718**	0.777/0.701	**0.842**/0.761	0.738/0.678	0.728/0.668	0.728/0.668
DC_MTN	0.676/0.595	0.713/0.464	0.742/0.621	0.609/0.515	0.714/0.617	0.586/0.455
DC_Bypass	0.687/0.565	0.689/0.582	0.734/0.592	0.603/0.531	0.684/0.581	0.556/0.487
DC_Progressive	0.702/0.585	0.500/0.019	0.500/0.279	0.622/0.521	0.698/0.560	0.577/0.407

Performance is expressed as AUC/Accuracy. Bold numbers show the best performance in the specific task.

TABLE III.
INDEPENDENT TEST

Model	Acute Dermal Toxicity	Acute Inhalation Toxicity	Acute Oral Toxicity	Eye Irritation	Skin Irritation	Skin Sensitization
Random forest	0.836/0.732	0.758/0.676	0.832/0.745	0.767/0.703	0.822/0.751	0.751/0.679
MTForestNet	**0.865**/0.765	**0.842**/**0.740**	0.819/0.752	**0.795**/**0.719**	**0.851**/**0.795**	**0.779**/**0.708**
AG_Medium	0.826/0.729	0.765/0.711	**0.838**/0.757	0.746/0.689	0.804/0.743	0.749/0.671
AG_Best	0.729/**0.826**	0.760/0.728	**0.838**/**0.760**	0.771/0.700	0.815/0.743	0.762/0.702
DC_MTN	0.747/0.622	0.654/0.483	0.730/0.628	0.596/0.521	0.683/0.615	0.589/0.473
DC_Bypass	0.748/0.569	0.642/0.591	0.745/0.624	0.600/0.546	0.672/0.592	0.593/0.487
DC_Progressive	0.729/0.602	0.500/0.019	0.500/0.280	0.617/0.530	0.687/0.558	0.584/0.429

Performance is expressed as AUC/Accuracy. Bold numbers show the best performance in the specific task.

The average AUC and accuracy values of AG_Medium are 0.762 and 0.700, respectively. AG_Best delivers a slightly better AUC of 0.767 and slightly worse accuracy of 0.699.

The MTForestNet, designed for dealing with the distinct chemical space of tasks, performed best. Fig. 2 shows the training process with accuracy and AUC performance for each layer. The optimal number of layers of MTForestNet was determined to be six according to the accuracy of the validation set. Its average AUC is 0.804, which is 3.3% better than the baseline models. With an average accuracy of 0.732, MTForestNet provides 2.1% performance improvement over the baseline models.

Table II showed that MTForestNet performed best in 5 out of the 6 tasks in terms of AUC and accuracy. AG_Best is the best model for acute dermal toxicity and acute oral toxicity in terms of accuracy and AUC, respectively. However, AG_Best is worse than the baseline models for the other tasks, resulting in a worse average AUC and accuracy compared to the baseline model.

C. Independent test

The independent test showed similar results that MTForestNet is the only algorithm providing a superior performance over the baseline model, with an average AUC and accuracy of 0.825 and 0.747, respectively. A 3.1% and 3.3% improvement on the average AUC and accuracy was achieved compared to the random forest models. The average AUC and accuracy of random forest models are 0.794 and 0.714, respectively.

Table III showed that MTForestNet performed best in 5 and 4 tasks in terms of AUC and accuracy, respectively. While with a slightly worse mean AUC of 0.779, AG_Best models provide good accuracy of 0.743, which is close to MTForestNet models and better than the baseline models. AG_Best is the best model in 1 and 2 tasks in terms of AUC and accuracy, respectively, as shown in Table III. As for the DeepChem-based models, their performance is the worst among the evaluated algorithms and is much worse than the baseline models. The average AUC and accuracy are 0.667 and 0.557 for DC_MTN, 0.667 and 0.568 for DC_Bypass, and 0.603 and 0.403 for DC_Progressive, respectively.

D. Comparison to existing methods

There are three recently published methods aiming to predict six-pack toxicity [1], [2], [3]. However, a careful evaluation found that the three studies divide the whole dataset into training and validation sets without an independent test. All three studies applied multiple machine learning algorithms and picked the best results from validation results. In this case, the prediction performance may be overestimated. Nevertheless, a comparison to existing methods can still provide some information on the current status of prediction models for six-pack toxicity.

We first compare our results with the study [3] using the same dataset. Only accuracies rounded to two decimal places were fully disclosed in their paper, with an average value of 0.75 based on the validation set. Their average accuracy value is the same as that of the developed MTForestNet model based on the test set, indicating that MTForestNet performed very well without the need to exhaustively train and select models.

The other two studies used a smaller dataset [1], [2] for model development. There is no accuracy information reported by StopTox [1]. Instead, a balanced accuracy representing a mean of sensitivity and specificity was given based on their validation set with an average value of 0.735. Please note that the results were based on a selection of chemicals suitable for the StopTox models. There are 5.4% deemed to be not suitable for the StopTox models. Without a selection of chemicals, the MTForestNet model with an average value of balanced accuracy of 0.7445 based on the test set provides better performance. The latest study [2] exhaustively trained all models by using the combination of three algorithms and four representations of chemicals. The selection of the best models based on their validation set yields average AUC values of 0.832 and 0.802 for models based on fingerprint and descriptor, and physicochemical properties, respectively. MTForestNet with an average AUC of 0.825 based on the test set is better than the models based on physicochemical properties and comparable to the models based on fingerprint and descriptor. While they proposed to combine the best-performing models to vote for the final prediction with a higher AUC

of 0.838 based on their validation set, the iterative use of samples from the validation set is prone to overfit the validation set without generalization ability to unseen samples.

Overall, MTForestNet provides an easy-to-use and robust method for predicting six-pack toxicity. The models developed in this study were rigorously validated and independently tested, and performed better than existing methods.

E. Comparison of training times

While good performance was achieved by the MTForestNet, it would be interesting to know the efficiency of the algorithms. We therefore compare the training time of the models. The baseline model requires 58 seconds for training six models. The DeepChem algorithms with early stop enabled are efficient, although with the worst performance. The training times are 40 seconds, 1 minute and 46 seconds, and 7 minutes and 19 seconds for DC_MTN, DC_Bypass, and DC_Progressive. The AG_Medium and AG_Best took the longest training time of 8 minutes and 28 seconds and 5 hours, 48 minutes and 34 seconds, respectively. MTForestNet maintains a well-balanced training time of 7 minutes and 26 seconds and the best prediction performance. Please note that only DeepChem-based models were trained using a GPU. CPU-based training was conducted for the other algorithms, and the model training may be further accelerated by using a GPU.

IV. Conclusion

Distinct chemical space is a unique attribute of biochemical datasets with little or no common chemicals shared among the tasks. Conventional multitask learning algorithms relying on learning a shared representation obtained from the common chemicals may not provide beneficial effects on the prediction performance. This study implemented and compared three types of multitask learning algorithms. Based on the validation and independent test results, we found that the biological readacross-based MTForestNet performed best. Overall, this work represents a significant step toward a biologically grounded and performance-enhancing solution suitable for computational toxicology tasks.

Acknowledgment

This work was supported by the National Science and Technology Council of Taiwan (NSTC-113-2628-E-400-001-MY3).

References

[1] J. V. B. Borba *et al.*, "STopTox: An in Silico Alternative to Animal Testing for Acute Systemic and Topical Toxicity," *Environ. Health Perspect.*, vol. 130, no. 2, p. 27012, Feb. 2022, doi: 10.1289/EHP9341.

[2] Y. N. Fuadah, M. A. Pramudito, L. Firdaus, F. J. Vanheusden, and K. M. Lim, "QSAR Classification Modeling Using Machine Learning with a Consensus-Based Approach for Multivariate Chemical Hazard End Points," *ACS Omega*, vol. 9, no. 51, pp. 50796–50808, Dec. 2024, doi: 10.1021/acsomega.4c09356.

[3] Y. Chushak, J. M. Gearhart, and R. A. Clewell, "Structural alerts and Machine learning modeling of 'Six-pack' toxicity as alternative to animal testing," *Comput. Toxicol.*, vol. 27, p. 100280, Aug. 2023, doi: 10.1016/j.comtox.2023.100280.

[4] B. Ramsundar *et al.*, "Is Multitask Deep Learning Practical for Pharma?," *J. Chem. Inf. Model.*, vol. 57, no. 8, pp. 2068–2076, Aug. 2017, doi: 10.1021/acs.jcim.7b00146.

[5] N. Erickson *et al.*, "AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data," Mar. 13, 2020, *arXiv*: arXiv:2003.06505. doi: 10.48550/arXiv.2003.06505.

[6] Z. Tan, Y. Li, W. Shi, and S. Yang, "A Multitask Approach to Learn Molecular Properties," *J. Chem. Inf. Model.*, vol. 61, no. 8, pp. 3824–3834, Aug. 2021, doi: 10.1021/acs.jcim.1c00646.

[7] X. Qian *et al.*, "An Interpretable Multitask Framework BiLAT Enables Accurate Prediction of Cyclin-Dependent Protein Kinase Inhibitors," *J. Chem. Inf. Model.*, vol. 63, no. 11, pp. 3350–3368, Jun. 2023, doi: 10.1021/acs.jcim.3c00473.

[8] Y. Yuan Li *et al.*, "Co-model for chemical toxicity prediction based on multi-task deep learning," *Mol. Inform.*, vol. 42, no. 5, p. e2200257, May 2023, doi: 10.1002/minf.202200257.

[9] X. Lin, Z. Quan, Z.-J. Wang, H. Huang, and X. Zeng, "A novel molecular representation with BiGRU neural networks for learning atom," *Brief. Bioinform.*, vol. 21, no. 6, pp. 2099–2111, Dec. 2020, doi: 10.1093/bib/bbz125.

[10] Y. Wang *et al.*, "Multitask CapsNet: An Imbalanced Data Deep Learning Method for Predicting Toxicants," *ACS Omega*, vol. 6, no. 40, pp. 26545–26555, Oct. 2021, doi: 10.1021/acsomega.1c03842.

[11] R.-H. Lin, P. Lin, C.-C. Wang, and C.-W. Tung, "A novel multitask learning algorithm for tasks with distinct chemical space: zebrafish toxicity prediction as an example," *J. Cheminformatics*, vol. 16, no. 1, p. 91, Aug. 2024, doi: 10.1186/s13321-024-00891-4.

[12] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[13] C.-C. Wang *et al.*, "Using random forest to predict antimicrobial minimum inhibitory concentrations of nontyphoidal Salmonella in Taiwan," *Vet. Res.*, vol. 54, no. 1, p. 11, Feb. 2023, doi: 10.1186/s13567-023-01141-5.

[14] C.-Y. Chou, P. Lin, J. Kim, S.-S. Wang, C.-C. Wang, and C.-W. Tung, "Ensemble learning for predicting ex vivo human placental barrier permeability," *BMC Bioinformatics*, vol. 22, no. Suppl 10, p. 629, Sep. 2022, doi: 10.1186/s12859-022-04937-y.

[15] C.-C. Wang, Y.-C. Liang, S.-S. Wang, P. Lin, and C.-W. Tung, "A machine learning-driven approach for prioritizing food contact chemicals of carcinogenic concern based on complementary in silico methods," *Food Chem. Toxicol. Int. J. Publ. Br. Ind. Biol. Res. Assoc.*, vol. 160, p. 112802, Feb. 2022, doi: 10.1016/j.fct.2021.112802.

[16] H.-L. Lin, Y.-W. Chiu, C.-C. Wang, and C.-W. Tung, "Computational prediction of Calu-3-based in vitro pulmonary permeability of chemicals," *Regul. Toxicol. Pharmacol. RTP*, vol. 135, p. 105265, Nov. 2022, doi: 10.1016/j.yrtph.2022.105265.

[17] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, 2003, doi: 10.1021/ci034160g.

[18] C.-C. Wang, Y.-C. Lin, Y.-C. Lin, S.-R. Jhang, and C.-W. Tung, "Identification of informative features for predicting proinflammatory potentials of engine exhausts," *Biomed. Eng. Online*, vol. 16, no. Suppl 1, p. 66, Aug. 2017, doi: 10.1186/s12938-017-0355-6.

[19] Y. Low *et al.*, "Integrative chemical-biological read-across approach for chemical hazard classification," *Chem. Res. Toxicol.*, vol. 26, no. 8, pp. 1199–1208, Aug. 2013, doi: 10.1021/tx400110f.

[20] Y. Guo, L. Zhao, X. Zhang, and H. Zhu, "Using a hybrid read-across method to evaluate chemical toxicity based on chemical structure and biological data," *Ecotoxicol. Environ. Saf.*, vol. 178, pp. 178–187, Aug. 2019, doi: 10.1016/j.ecoenv.2019.04.019.