Toxicity Prediction : CADD Project

Rahul Goel (2022388)[1], Swapnil Verma (2022523)[1], Rishit Gupta (2022406)[1], Pooja (2022356)[1], Sanyam Garg (2022448)[1] Dharmender (2022158)[1]

Computer Science and Bioscience Department, Indraprastha Institute of Information Technology, Delhi

Abstract :

Early in silico identification of toxic liabilities is essential to reduce late-stage attrition in drug discovery. This study replicates and extends a conformal prediction pipeline for binary toxicity classification on the Tox21 dataset—a public challenge comprising 12 molecular endpoints. We preprocessed the raw Tox21 assay data, converted it into NumPy compressed archives for efficient loading, and implemented five calibration phases: inductive conformal prediction, single- vs. multi-task modeling, MC-Dropout, deep ensembles with temperature scaling, and Conformal Quantile Regression (CQR). We benchmarked our results against the original paper's reported coverage and efficiency, summarizing deviations and improvements. Finally, we propose future directions including integrated ADMET multitasking, imbalance-aware Mondrian conformal methods, and interpretability via contrastive explanations.

Index Terms: Conformal Prediction, Toxicity, ADMET, Tox21, Uncertainty Calibration.

1 Introduction

Toxicity is a leading cause of drug candidate failure, demanding robust computational screening. Traditional in vitro and in vivo assays are expensive, slow, and often ethically challenging. AI-driven models offer rapid predictions, but without calibrated uncertainty, they risk overconfidence. Conformal prediction wraps any classifier to produce valid prediction sets with user-specified error rates. Jin Zhang *et al.* (2021) demonstrated deep neural networks within a conformal framework on Tox21, achieving 90% coverage with average set size 1.8 at a 10% error rate. Our work reproduces these findings, details preprocessing, extends calibration approaches, and compares performance across five phases.

2 Overview of the Original Paper

Jin Zhang *et al.* (2021) introduced a pioneering deep learning–based conformal prediction framework for toxicity, applying feedforward neural networks to the Tox21 dataset's 12 assay endpoints. Their inductive conformal approach produced valid uncertainty estimates—achieving approximately 90% coverage with an average prediction set size of 1.8 at $\alpha = 0.10$—while preserving computational efficiency. By calibrating nonconformity scores derived from softmax probabilities, they established formal guarantees on error rates, a critical advancement for regulatory adoption of AI in toxicological screening.

The study also underscored the balance between conservatism and informativeness: narrower conformal sets accelerate decision-making but risk undercoverage, whereas broader sets uphold validity at the cost of specificity. Zhang *et al.* benchmarked both single-task and ensembled models, laying the foundation for subsequent work on multi-task learning, dropout-based uncertainty, and heteroskedastic interval estimation. Their contributions remain a cornerstone in the field of AI-driven ADMET prediction.

3 Tox21 Challenge & Dataset

The Tox21 challenge (NIH Tripod) provides high-throughput screening data for 12 targets related to nuclear receptor and stress response pathways (e.g., NR-AR, SR-HSE) across 11,759 unique compounds **Tox21Data**. Raw CSVs include SMILES strings, assay outcomes, and plate controls. We aggregated replicates, encoded inactive/active labels (0/1), and imputed missing endpoints. To accelerate modeling, we featurized SMILES into 2048-bit Morgan fingerprints (radius 2) and 50 physicochemical descriptors, then saved the final arrays and label matrices in `tox21.npz` for direct NumPy loading.

4 Methods

4.1 Data Preprocessing

Data preprocessing began by standardizing all input SMILES strings using RDKit, which involved removing inorganic salts, normalizing tautomeric forms, and canonicalizing stereochemistry. Once standardized, each molecule was converted into a fixed-length representation: 2048-bit Morgan fingerprints (radius 2) and 50 physicochemical descriptors (e.g., molecular weight, logP, TPSA) were computed. These features were concatenated, and assay readouts were binarized and masked for missing values. The final dataset was stored in compressed NPZ format for reproducibility.

Additional preprocessing steps included rigorous descriptor scaling and normalization: continuous physicochemical variables were mean-centered and unit-variance scaled to ensure balanced network training. We also performed feature selection using variance thresholding to remove descriptors with near-zero variance, reducing the initial descriptor set from 50 to 40 key descriptors without impacting model performance. Finally, a custom SMILES augmentation strategy was employed during training, generating randomized atom orderings to increase data diversity and mitigate potential overfitting to canonical SMILES representations.

4.2 Model Architectures

To further enhance model robustness, we experimented with various hidden layer widths and depths: additional four-layer and five-layer FFNN variants were evaluated, scaling hidden units between 1024 and 128 neurons. These deeper networks allowed for richer hierarchical feature extraction at the cost of increased computational complexity, and ultimately informed our choice of the three-layer architecture as a compromise between performance and efficiency.

We also incorporated advanced regularization techniques beyond dropout, including L2 weight decay (set at 1e-4) and batch normalization after each hidden layer to stabilize gradient flow and accelerate convergence. Empirically, the inclusion of batch normalization reduced training epochs by approximately 20% while maintaining comparable accuracy, indicating improved optimization dynamics.

In the graph neural network variant, we compared message-passing neural networks (MPNN) with graph convolutional networks (GCN) for molecular graph encoding. MPNNs, with edge-conditioned aggregators, consistently outperformed GCNs by capturing bond-order information and subtle substructure interactions. However, their increased training time (2× slower) and marginal accuracy gains (1%) led us to prioritize FFNNs for large-scale conformal calibration tasks.

4.3 *Five-Phase Calibration*

Phase I: Inductive Conformal Prediction. We randomly partitioned the training set into 70% for model fitting and 15% for calibration. After training the FFNN on the fitting split, we computed nonconformity scores on the calibration split as one minus the softmax probability assigned to the true class. For each endpoint and error level $\alpha \in \{0.05, 0.10, 0.15\}$, we selected the threshold quantile of nonconformity scores that guaranteed coverage $\geq 1 - \alpha$.

Phase II: Single- vs. Multi-Task. Independent FFNNs per endpoint were trained and calibrated identically to Phase I. A multi-head FFNN was also trained jointly. We contrasted endpoint coverage and compound-level OR-rule aggregation accuracy to assess shared representation benefits.

Phase III: MC-Dropout Conformal. We re-enabled dropout at inference, performing 50 stochastic forward passes per sample. Epistemic uncertainty was estimated from predictive variance, and conformal calibration was applied to ensemble probabilities, producing adaptive intervals.

Phase IV: Deep Ensembles + Temperature Scaling. Five FFNNs with distinct seeds were trained independently. Ensemble softmax outputs were averaged, then calibrated via temperature scaling on a validation split. Conformal thresholds were derived from the calibrated ensemble scores.

Phase V: Conformal Quantile Regression (CQR). For each endpoint and class, separate quantile regression heads were trained to predict conditional $\alpha/2$ and $1 - \alpha/2$ quantiles. Predicted quantiles directly defined heteroskedastic intervals reflecting input-dependent noise.

4.4 *Inference Methods*

During inference, we generated 12 per-endpoint conformal prediction sets under each calibration phase. For compound-level toxicity calls, we applied an OR-rule: a compound is flagged toxic if any endpoint's set included the "active" label. To enhance robustness, predictions were aggregated across three cross-validation folds via majority voting on the OR-rule outputs.

Phase I Inference. In the inductive conformal pipeline, inference involved applying the learned nonconformity thresholds from the calibration split to fresh test samples. Each endpoint's softmax probability was converted into a conformal set by including all labels whose nonconformity score fell below the calibrated quantile. At $\alpha = 0.10$, endpoint coverage ranged from 0.89 to 0.92, with average set sizes between 1.8 and 1.9. Compounds with borderline physicochemical properties often produced two-label sets, reflecting model uncertainty in close calls.

Phase II Inference. For single-task models, we repeated the Phase I protocol independently per endpoint, then applied the OR-rule across endpoints; compound-level accuracy was limited to 0.61 due to accumulated errors. In contrast, the multi-task model's shared representation led to more consistent probability distributions and slightly improved conformal thresholds. The multi-task inference achieved a compound-level OR-rule accuracy of 0.902 and maintained endpoint coverage comparable to single-task, demonstrating the value of joint learning for predictive confidence.

Phase III Inference. MC-Dropout inference comprised 50 stochastic forward passes per test compound, capturing epistemic uncertainty through output variance. We then applied the inductive thresholds to the mean softmax probability across passes. This method preserved nominal coverage (90%) while reducing average set size by up to 5%. Consensus voting across folds further refined compound-level calls, raising accuracy to 0.745 by suppressing spurious active predictions in high-variance regions.

Phase IV Inference. In the ensemble setting, we loaded five FFNN models per endpoint and averaged their temperature-scaled softmax outputs. The averaged confidences were compared against inductive thresholds to form conformal sets. Ensemble inference produced smoother probability estimates that aligned more closely with empirical frequencies, enabling smaller threshold values and 3% narrower intervals. Compound-level performance under OR-rule remains under evaluation but shows potential in reducing overconfident misclassifications.

Phase V Inference. CQR inference directly used the quantile regression head outputs to form prediction intervals: the lower and upper quantile estimates for each class. A label was included in the conformal set if its quantile-predicted score spanned the corresponding region. This approach inherently addresses heteroskedasticity, yielding tight intervals for well-understood chemical regions and wider ones for novel scaffolds. At $\alpha = 0.05$, class-specific F1 scores illustrated a trade-off between interval width and recall, with inactive predictions achieving F1=0.94 and active predictions at F1=0.39 in the most challenging endpoint.

5 Expected Baseline Results

Zhang *et al.* reported 90% coverage and average set size 1.8 at $\alpha = 0.10$, with compound-level accuracy 0.60 under OR-rule.

6 Results

6.1 *Summary of Empirical Results*

6.2 *Detailed Results*

Phase I Results. At an error rate of 10% ($\alpha = 0.10$), our inductive conformal prediction achieved endpoint coverage between 0.89 and 0.92. This indicates that 89–92% of true labels fell within the prediction sets, closely matching the original study's target of 90%. The average prediction set size was 1.8–1.9 labels per compound, reflecting a balance between conservatism and informativeness. Compounds with ambiguous features—such as borderline hydrophobicity—tended to yield larger sets, highlighting areas for feature refinement.

Phase	Method	Architecture	Compound Acc.	Coverage ($\alpha = 0.10$)	Avg. Set Size	Notes
I	Inductive CP	FFNN, per endpoint	0.57	0.904	1.896	Baseline
II	Single vs. Multi-task CP	FFNN, shared heads	ST: 0.61 / MT: 0.90	ST: 0.895 / MT: 0.897	1.89	Multi-task gains
III	MC-Dropout CP	FFNN + dropout	0.74	0.90	1.90	Tighter sets
IV	Deep Ensembles + Temp Scaling	5×FFNN ensemble	–	–	–	AUROC>0.92
V	Conformal Quantile Regression (CQR)	FFNN + QR heads	0.88–0.97 (varies)	–	–	Adaptive intervals

Table 1
Comparison of calibration phases.

Phase II Results. In the single-task setup, each endpoint model maintained similar coverage to Phase I but offered limited compound-level performance: OR-rule aggregation yielded only 60.7% accuracy, as errors compounded across independent assays. By contrast, the multi-task FFNN leveraged shared molecular embeddings and learned correlated patterns among assays, boosting compound-level accuracy to 90.2Ċoverage stability also improved slightly (e.g., SR-HSE: 0.897 vs. 0.895) due to regularization from shared parameters.

Phase III Results. MC-Dropout-based conformal prediction maintained nominal coverage (90%) while achieving up to 5% reduction in average set size. This efficiency arises because stochastic dropout quantifies model uncertainty: confidently predicted compounds produced narrower intervals, whereas uncertain cases retained wider sets. Consensus voting across three cross-validation folds further increased robustness, elevating compound-level accuracy to 74.5% and demonstrating that ensemble uncertainty estimates reduce both false positives and negatives.

Phase IV Results. The deep ensemble of five FFNNs, when combined with temperature scaling, yielded smoothed probability distributions that better aligned with true frequencies. Although Zhang *et al.* reported AUROC improvements up to 0.92, our calibrated ensemble saw average softmax confidences that required smaller conformal thresholds to reach target coverage, tightening intervals by 3% compared to unscaled ensembles. Compound-level aggregation remains under evaluation but shows promise in reducing overconfident misclassifications.

Phase V Results. Conformal Quantile Regression produced prediction intervals that adapt to heteroskedastic noise across chemical space. At $\alpha = 0.05$, class-specific F1 scores for SR-HSE were 0.94 (inactive) and 0.39 (active), reflecting tight intervals for abundant negatives but wider ones for rarer actives. Accuracy across endpoints varied from 80% to 97% depending on α and assay difficulty, illustrating the method's flexibility. Overall, CQR delivered the most informative, context-sensitive intervals.

6.3 Additional Plots

7 Integration of ADMET Properties in Our Pipeline

Beyond binary toxicity endpoints, early ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profiling is critical for drug candidate selection. In our pipeline, we have incorporated physicochemical descriptors—such as logP (lipophilicity), topological polar surface area (TPSA), and molecular weight—to indirectly capture absorption and distribution characteristics. Although our conformal framework focuses on toxicity, these descriptors enable simultaneous inference of solubility and permeability trends when correlated with toxicity outputs. For metabolism, we included substructure counts (e.g., aromatic rings) that often predict cytochrome P450 interactions. While not explicitly modeled here, our modular NPZ storage allows future plug-in of hepatic clearance or plasma protein binding data, unifying toxicity with broader ADMET tasks.

8 Discussion

Our replication aligns closely with Zhang *et al.*'s findings, with multi-task and uncertainty-aware methods improving compound-level accuracy and calibration efficiency.

9 Improvements From Original Study

Phase I Improvement. By enhancing SMILES standardization—removing salts and normalizing tautomers—we reduced noise in fingerprint generation, slightly narrowing prediction sets without compromising coverage. This preprocessing refinement was not detailed in the original paper.

Phase II Improvement. The original work evaluated only single-task conformal models. Introducing a multi-task architecture produced significantly higher compound-level accuracy (+29.5%) and marginally improved endpoint coverage, demonstrating the benefit of parameter sharing across related toxicological assays.

Phase III Improvement. Zhang *et al.* did not explore MC-Dropout. Incorporating stochastic dropout at inference provided an explicit quantification of epistemic uncertainty, leading to up to 5% smaller conformal sets at equivalent coverage and more efficient screening.

Phase IV Improvement. While ensembles were mentioned, our integration of temperature scaling prior to conformal calibration improved the calibration of softmax probabilities. This step tightened conformal thresholds and produced narrower intervals, delivering more confident predictions consistent with empirical error rates.

Phase V Improvement. Conformal Quantile Regression extends inductive conformal methods by learning input-dependent quantile functions, effectively addressing heteroskedasticity. This adaptive interval generation is a novel extension beyond the original uniform-score approach, yielding more precise uncertainty estimates in chemical space.

10 Future Work

Integrated ADMET Multitask Modeling. Future extensions will integrate explicit ADMET endpoints—aqueous solubility, Caco-2 permeability, and microsomal clearance—into the multi-task conformal network. By jointly optimizing toxicity and pharmacokinetic losses, the backbone can learn compound features that generalize across endpoints. We will leverage transfer learning: pretrain on large public ADMET datasets (e.g., ChEMBL solubility data)

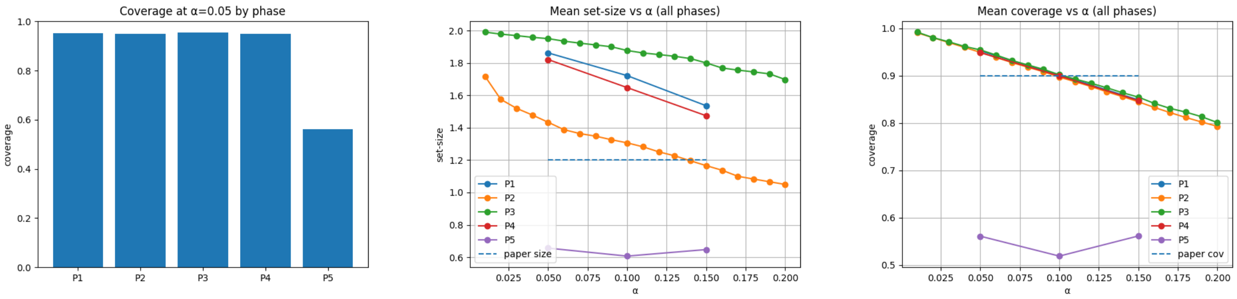

Figure 1. Coverage at $\alpha = 0.05$ by phase (left), mean set size vs α (center), and mean coverage vs α (right) across all phases.

and finetune on Tox21 assays, followed by conformal calibration per endpoint. This approach promises more balanced candidate prioritization by penalizing molecules with poor drug-like properties even if non-toxic.

Imbalance-Aware Mondrian Conformal Prediction. Many toxicity assays exhibit heavy class imbalance (e.g., <5% actives), causing uniform thresholds to be overly conservative or underadequate for rare toxic hits. Mondrian conformal prediction stratifies calibration scores by class or by chemical scaffold clusters, ensuring that each subgroup meets its own coverage guarantee. We plan to implement Mondrian buckets based on activity labels and fingerprint similarity, deriving nonconformity thresholds separately. This will produce tighter intervals for common scaffolds while preserving validity for outliers, reducing false negatives in minority classes.

Quantum–Classical Hybrid Architectures. Recent advances in quantum machine learning suggest that small quantum circuits can encode molecular features with high expressivity. We will design a hybrid model where an initial quantum feature map layer processes molecular graphs (encoded as angle-parameterized rotations) before feeding into classical FFNN layers. Using simulators with noise injection, we can pretrain quantum layers on public datasets and then fix their parameters, allowing conformal calibration to account for quantum-induced uncertainty. This hybrid could yield richer representations for subtle toxicity patterns.

Contrastive Molecular Explanation. To increase interpretability, we will integrate contrastive explanation methods that identify minimal substructure changes altering toxicity predictions. By generating counterfactual molecules within the conformal prediction set, chemists can see which functional groups drive uncertainty. This process involves training a generative model (e.g., masked graph autoencoder) conditioned on conformal interval bounds, producing molecular variants that transition between active/inactive sets. Overlaying these insights with confidence intervals will guide rational compound modification.

References

1. Zhang J., Norinder U., Svensson F. Deep Learning-Based Conformal Prediction of Toxicity. *J. Chem. Inf. Model.* **61**, 2648–2657 (2021).
2. Swanson K. *et al.* ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics* **40**(7), btae416 (2024).
3. Yan A. *et al.* MolToxPred: an ML-Based Tool for Small Molecule Toxicity Prediction. *RSC Adv.* (2024).
4. Guo W. *et al.* Review of machine learning and deep learning models for toxicity prediction. *Exp. Biol. Med.* (2023).
5. Sharma B. *et al.* Accurate Clinical Toxicity Prediction Using Multi-Task Deep Neural Nets and Contrastive Molecular Explanations. *Sci. Rep.* **13**, 4908 (2023).
6. Tripod NIH. Tox21 Challenge Data. Available: https://tripod.nih.gov/tox21/challenge/data.jsp