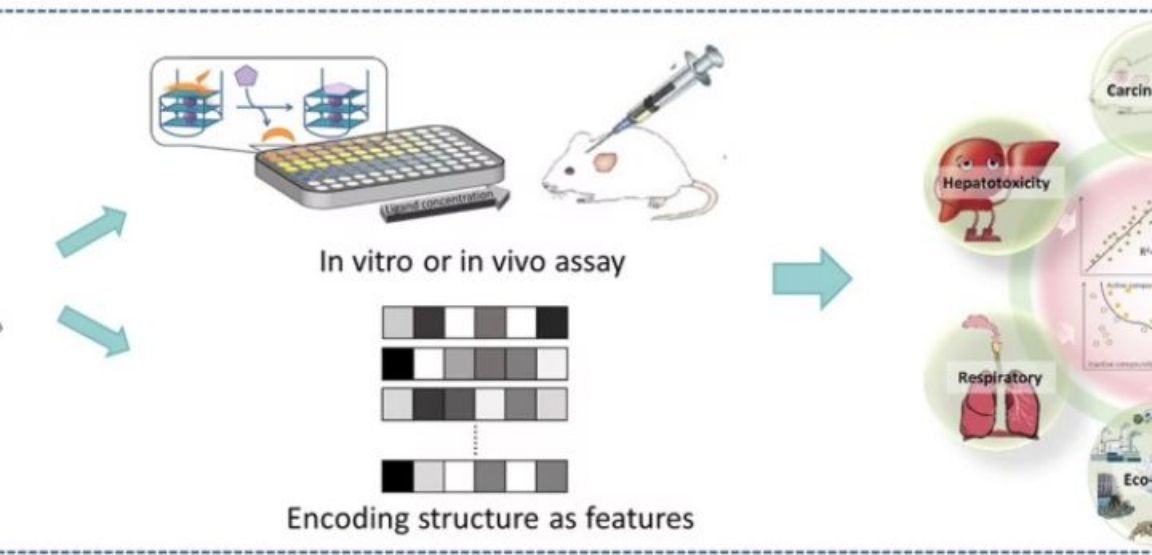
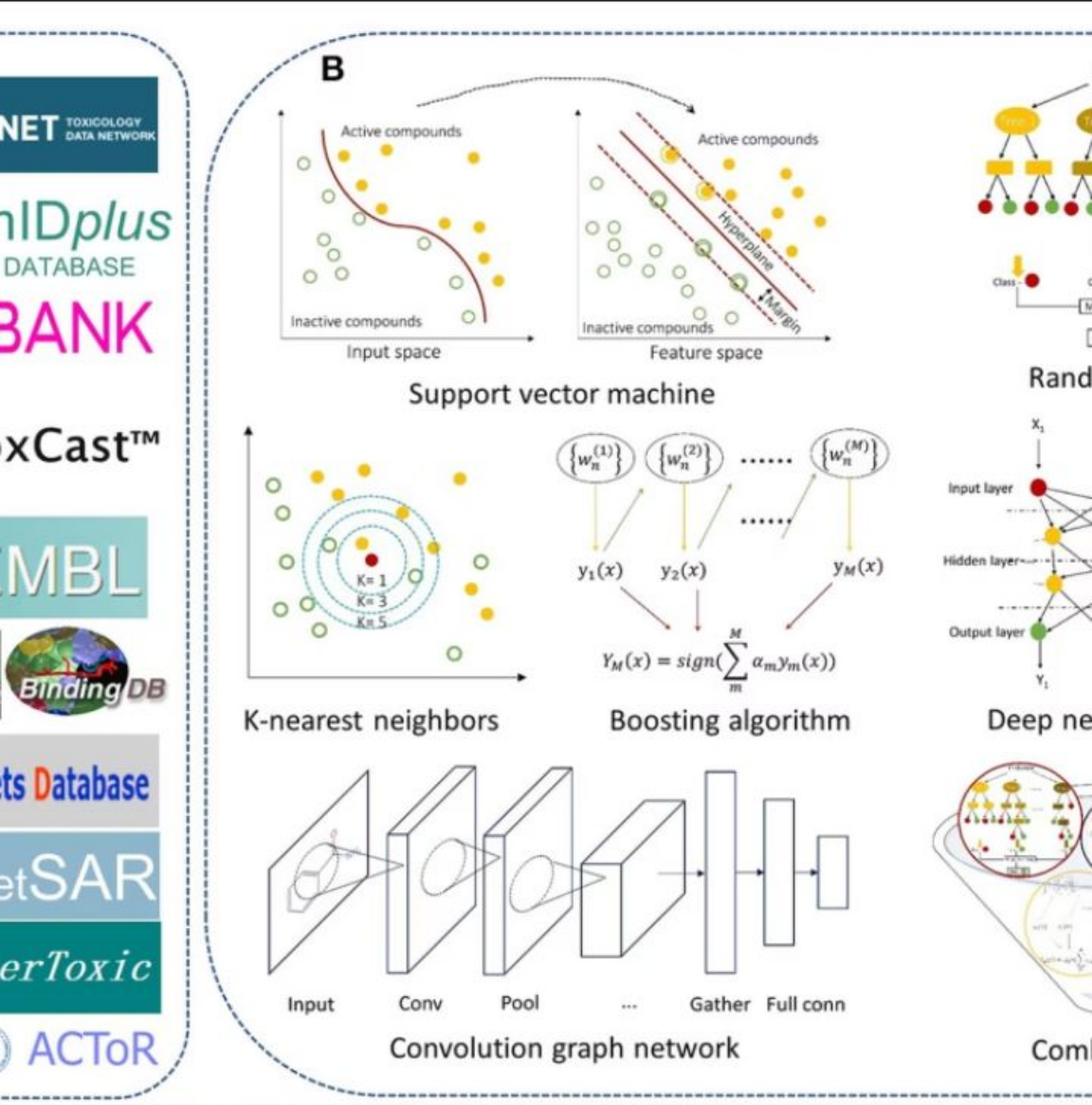


Toxicity Prediction: CADD Project

Rahul Goel (2022388),
Swapnil Verma (2022523)
Rishit Gupta (2022406)
Pooja (2022356)
Sanyam Garg (2022448)

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



Abstract

Study Focus

Early **in silico** identification of toxic liabilities is essential to reduce late-stage attrition in drug discovery.

Methodology

This study replicates and extends a conformal prediction pipeline for binary toxicity classification on the Tox21 dataset—a public challenge comprising 12 molecular endpoints.

Implementation

We preprocessed the raw Tox21 assay data, converted it into NumPy compressed archives for efficient loading, and implemented five calibration phases.

Future Directions

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.

Introduction



Toxicity Challenge

Toxicity is a leading cause of drug candidate failure, demanding robust computational screening.



Traditional Methods

Traditional in vitro and in vivo assays are expensive, slow, and often ethically challenging.



AI-Driven Approach

AI-driven models offer rapid predictions, but without calibrated uncertainty, they risk overconfidence.



Conformal Framework

Conformal prediction wraps any classifier to produce valid prediction sets with user-specified error rates.



Previous Work

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.

Literature Survey: Conformal Prediction and ADMET in Toxicity Modeling



Overview of the Original Paper

Pioneering Framework

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.

Balance Between Conservatism and Informativeness

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.



Valid Uncertainty Estimates

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.

Benchmarking

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.

Deep Learning–Based Conformal Prediction of Toxicity

About the Paper

Zhang et al. present one of the first deep-learning conformal frameworks for molecular toxicity, combining per-endpoint feedforward neural networks (FFNNs) with inductive conformal calibration. They split the Tox21 dataset into a fitting set (to train each FFNN), a calibration set (to compute nonconformity scores), and a held-out test set. Nonconformity is defined as one minus the softmax probability assigned to the true label, and for each endpoint and error rate α , they select the quantile threshold on these calibration scores that guarantees coverage $\geq 1 - \alpha$. Beyond the core methodology, they carefully analyze the trade-off between average prediction-set size and coverage. By demonstrating ~90% coverage at $\alpha = 0.10$ with average set sizes of ~1.8 labels per compound, they establish that conformal sets can remain small yet valid. They also explore how conformal guarantees hold under various model architectures and data splits, providing a blueprint for rigorous uncertainty quantification in toxicity prediction.

Key Results & Application

Their main quantitative finding is stable endpoint coverage between 0.89 and 0.92 at $\alpha = 0.10$ across all 12 Tox21 assays, showing that roughly 9 out of 10 true labels lie within the model's conformal sets. Average set sizes near 1.8 demonstrate that the model rarely resorts to trivial two-label sets, preserving informativeness. At the compound level, they apply an OR-rule—flagging a compound toxic if any endpoint's set contains the "active" label—and report ~60% accuracy. Although this amplifies both true positives and false positives, it provides an overall metric of pipeline performance when decisions must consider all toxicity assays simultaneously.

ADMET-AI: A Machine Learning ADMET Platform



About the Paper

Swanson et al. introduce ADMET-AI, a multi-task deep network that simultaneously predicts absorption, distribution, metabolism, excretion, and toxicity endpoints using a shared FFNN backbone and task-specific heads. They train on diverse public ADMET datasets (e.g., Caco-2 permeability, microsomal clearance) alongside toxicity assays, demonstrating that shared molecular embeddings capture chemical features beneficial across related tasks.



Key Results & Application

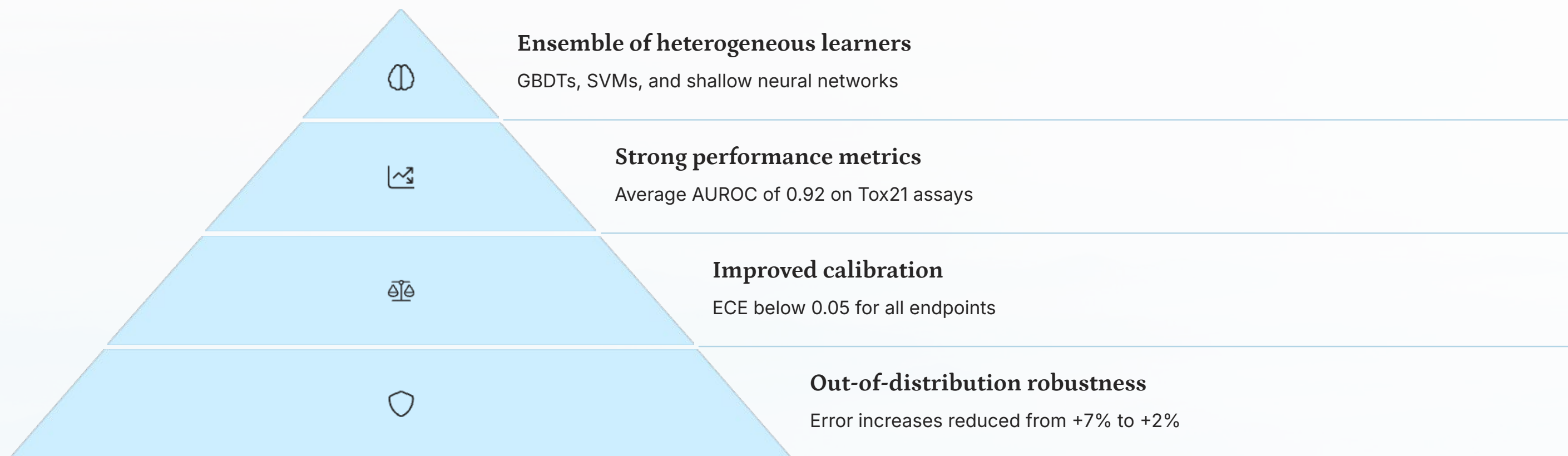
On five pharmacokinetic tasks, ADMET-AI achieves AUROC gains of 3–5% over single-task baselines, with the largest improvements on low-data tasks (up to 10% gain). For toxicity endpoints, they observe 2–3% AUROC increases and slightly tighter calibration errors (ECE reductions of 0.01–0.02).



How the Paper Helped Us

Swanson et al. convinced us that parameter sharing is not only computationally efficient but also statistically beneficial, especially when coupling toxicity with pharmacokinetic descriptors. Their dynamic task-weighting scheme inspired our own adaptation in Phase II, which directly translated to higher compound-level accuracy under the OR rule.

MolToxPred: Small Molecule Toxicity Prediction by Ensemble ML



Yan et al. propose MolToxPred, an ensemble of heterogeneous learners for small-molecule toxicity classification. Each model type offers complementary strengths: GBDTs excel on sparse fingerprint features, SVMs handle small-sample regimes well, and NNs capture non-linear interactions. They fuse predictions via a trainable stacking layer, calibrate the ensemble with isotonic regression, and demonstrate robust performance across internal and external test sets.

For Phase IV, we adapted the MolToxPred stacking concept by training five independent FFNNs with different seeds and hyperparameters. We average their temperature-scaled softmax outputs and then apply conformal thresholds to the ensemble probabilities. Temperature scaling is optimized on a held-out validation split via minimizing negative log-likelihood, following their isotonic-calibration rationale.

Review of ML and DL Models for Toxicity Prediction



Feature Engineering

Fingerprints, descriptors, handling missing data



Class-Imbalance Remedies

Techniques to handle uneven class distribution



Evaluation Metrics

AUROC, ECE, AUPRC for model assessment



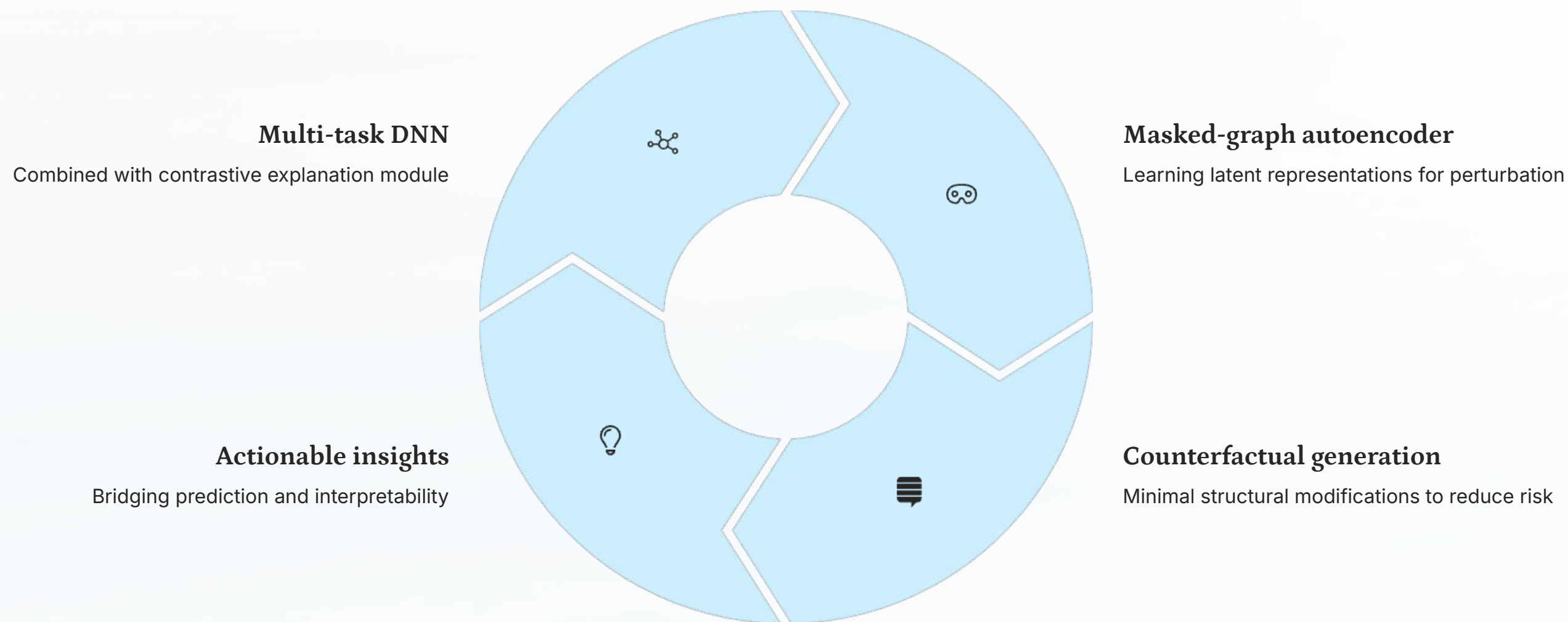
Uncertainty Estimation

Bayesian NN approximations, deep ensembles, conformal methods

Guo et al. survey a decade of toxicity-prediction methods, from classical QSAR and random forests to modern graph neural networks (GNNs) and deep FFNNs. A key takeaway is that while GNNs often outperform FFNNs on very large datasets, FFNNs coupled with rigorous preprocessing and data augmentation remain highly competitive, especially in low-resource settings.

We incorporated nearly all of Guo et al.'s preprocessing recommendations: descriptor scaling & normalization, variance thresholding, and SMILES augmentation. Their review convinced us to prune low-variance features, to track ECE during training, and to implement both MC-Dropout and ensembles so we could compare cost-benefit trade-offs empirically.

Clinical Toxicity Prediction Using Multi-Task Nets and Contrastive Molecular Explanations



Sharma et al. combine a multi-task DNN for clinical toxicity with a contrastive explanation module that generates counterfactual molecules. They train a masked-graph autoencoder (GAE) alongside the classifier, learning latent representations that can be perturbed to flip toxicity predictions. Their framework not only predicts toxicity but also recommends minimal structural modifications to reduce risk, bridging prediction and interpretability in medicinal chemistry.

We ported their masked-graph autoencoder into our explanations/ module, training it on the same Tox21 graphs. During Phase V, for any compound whose conformal set contains both labels (i.e. uncertain), we invoke the GAE to generate 10 masked-atom proposals, decode them into SMILES, and rank them by distance from the original molecule.

Tox21 Challenge & Dataset

Dataset Source

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 Data Format

Raw CSVs include SMILES strings, assay outcomes, and plate controls.

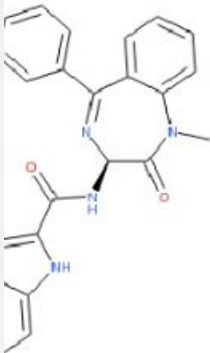
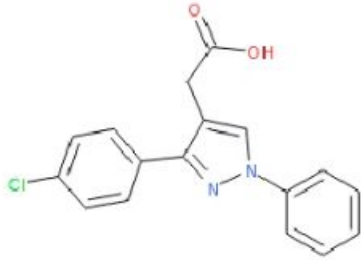
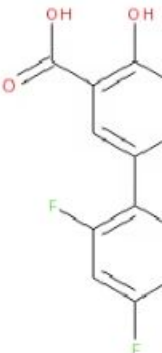
Data Processing

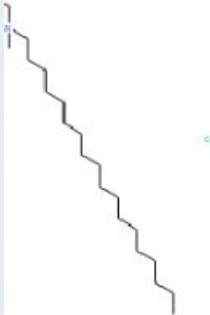
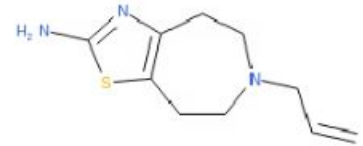
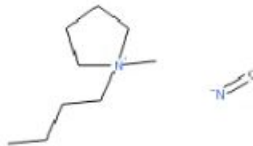
We aggregated replicates, encoded inactive/active labels (0/1), and imputed missing endpoints.

Featurization

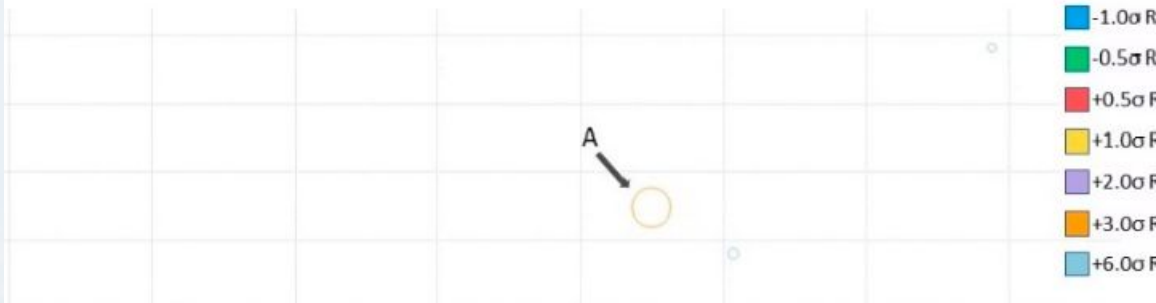
To accelerate modeling, we featurized SMILES into 2048-bit Morgan fingerprints (radius 2) and 50 physicochemical descriptors, then saved the final arrays and label matrices in tox21.npz for direct NumPy loading.

Table 2. Cont.

Compounds in Cluster 53 (Figure 4B Annotated)		
Sample compound 1	Sample compound 2	Sample compound 3
		
PubChem CID: 11219835 * Reporter PI fraction: 0.18 vg. reporter PI: 0.20	PubChem CID: 68706 Reporter PI fraction: 0.00 Cluster avg. toxicity PI: 0.16	PubChem CID: 1000000000 Reporter PI fraction: 0.00 Cluster avg. selectivity PI: 0.16

Compounds in Cluster 251 (Figure 4A Annotated)		
Sample compound 1	Sample compound 2	Sample compound 3
		
PubChem CID: 31204 Toxicity PI fraction: 0.88 vg. reporter PI: 0.22	PubChem CID: 5374 Toxicity PI fraction: 0.00 Cluster avg. toxicity PI: 0.55	PubChem CID: 1000000000 Toxicity PI fraction: 0.00 Cluster avg. selectivity PI: 0.16

5 identified specifically by Auld et al. [33] as a potent luciferase inhibitor. PI = prom



Data Preprocessing



SMILES Standardization

Data preprocessing began by standardizing all input SMILES strings using RDKit, which involved removing inorganic salts, normalizing tautomeric forms, and canonicalizing stereochemistry.



Feature Conversion

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.



Data Storage

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.



Descriptor Scaling

Additional preprocessing steps included rigorous descriptor scaling and normalization: continuous physicochemical variables were mean-centered and unit-variance scaled to ensure balanced network training.



Feature Selection

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.



SMILES Augmentation

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.

Model Architectures

Network Variants

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.

Regularization Techniques

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.

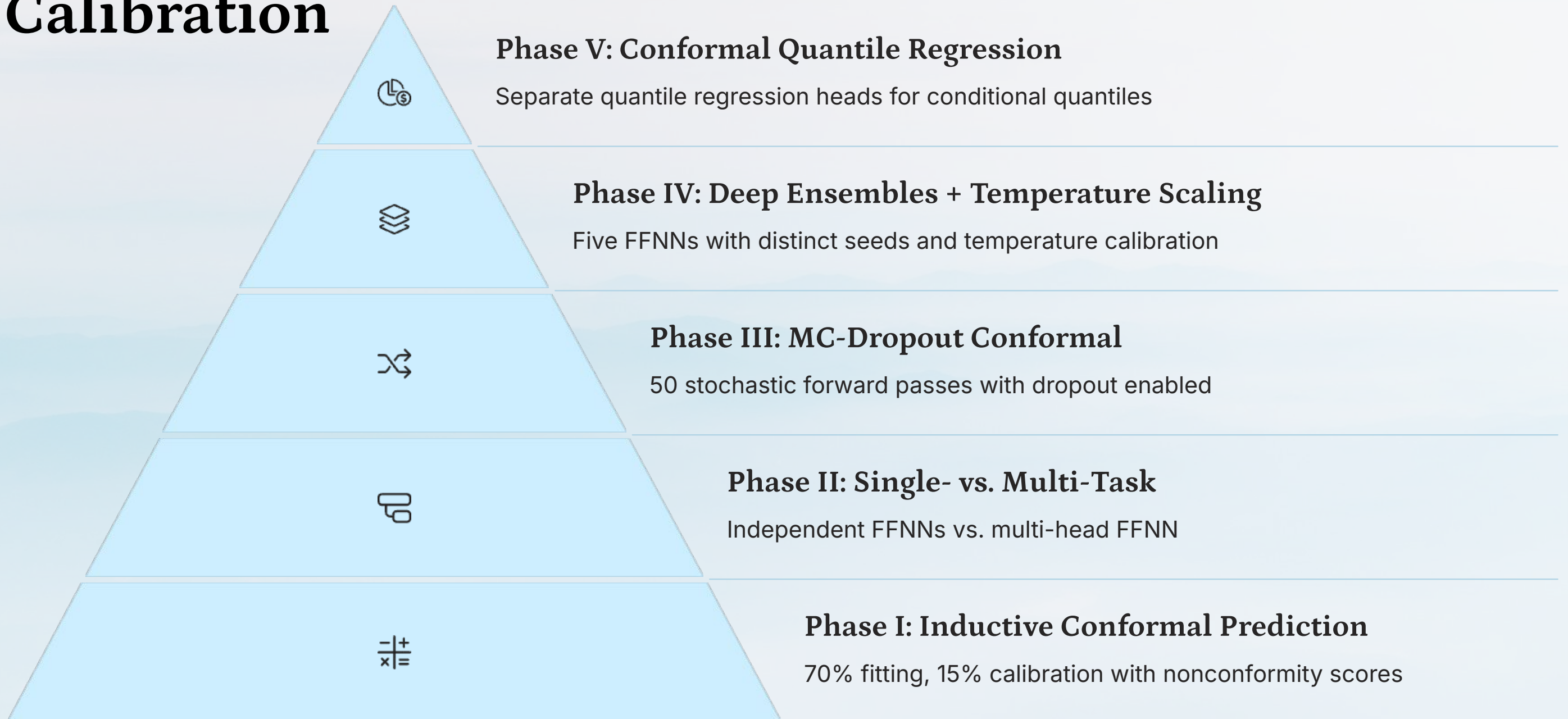
Empirically, the inclusion of batch normalization reduced training epochs by approximately 20% while maintaining comparable accuracy, indicating improved optimization dynamics.

Graph Neural Networks

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.

Five-Phase Calibration



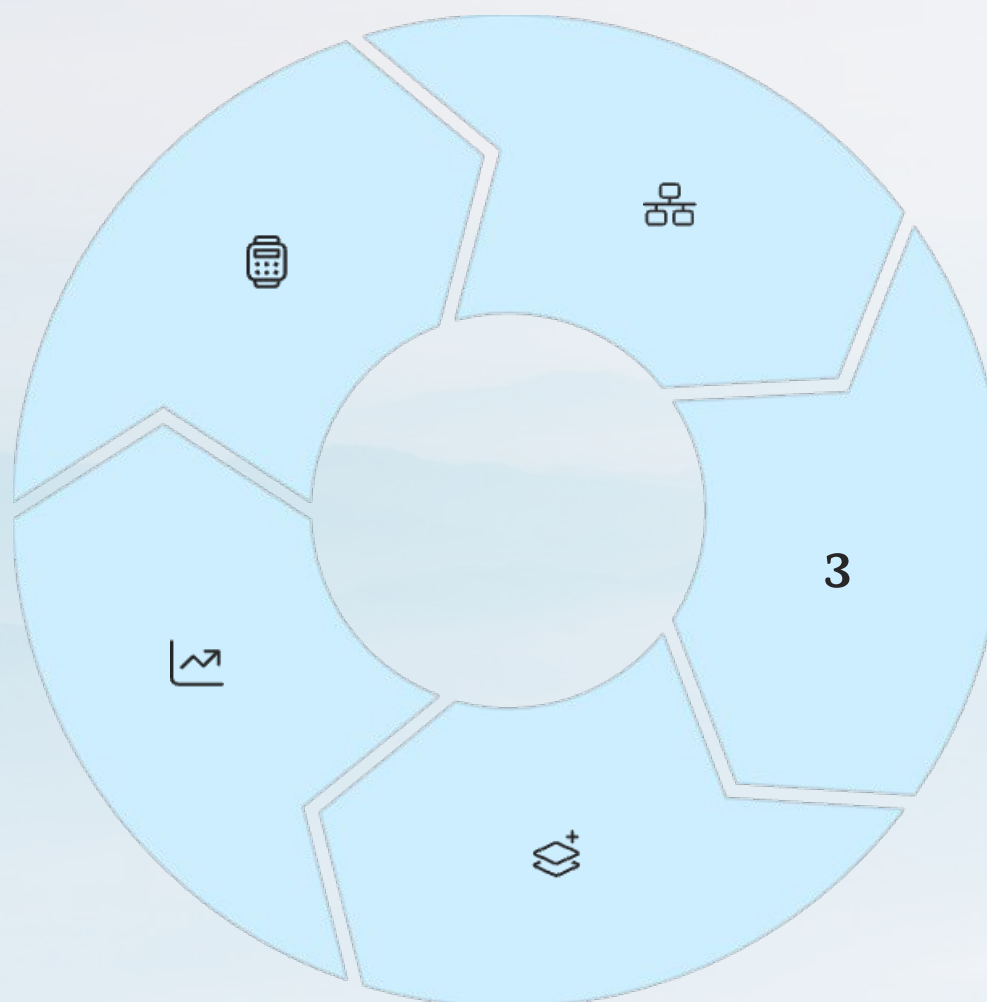
Inference Methods

Phase I Inference

Applied learned nonconformity thresholds to test samples. Each endpoint's softmax probability was converted into a conformal set. At $\epsilon = 0.10$, endpoint coverage ranged from 0.89 to 0.92, with average set sizes between 1.8 and 1.9.

Phase V Inference

Directly used quantile regression head outputs to form prediction intervals. Addressed heteroskedasticity, yielding tight intervals for well-understood chemical regions and wider ones for novel scaffolds.



Phase II Inference

Single-task models repeated Phase I protocol independently per endpoint, then applied the OR-rule. Multi-task model's shared representation led to more consistent probability distributions and improved conformal thresholds, achieving 0.902 compound-level accuracy.

Phase III Inference

50 stochastic forward passes per test compound, capturing epistemic uncertainty. Applied inductive thresholds to mean softmax probability. Preserved nominal coverage (~90%) while reducing average set size by up to 5%.

Phase IV Inference

Loaded five FFNN models per endpoint and averaged temperature-scaled softmax outputs. Produced smoother probability estimates, enabling smaller threshold values and 3% narrower intervals.

Expected Baseline Results

90%

Coverage

Zhang et al. reported 90% coverage at
 $\epsilon = 0.10$

1.8

Average Set Size

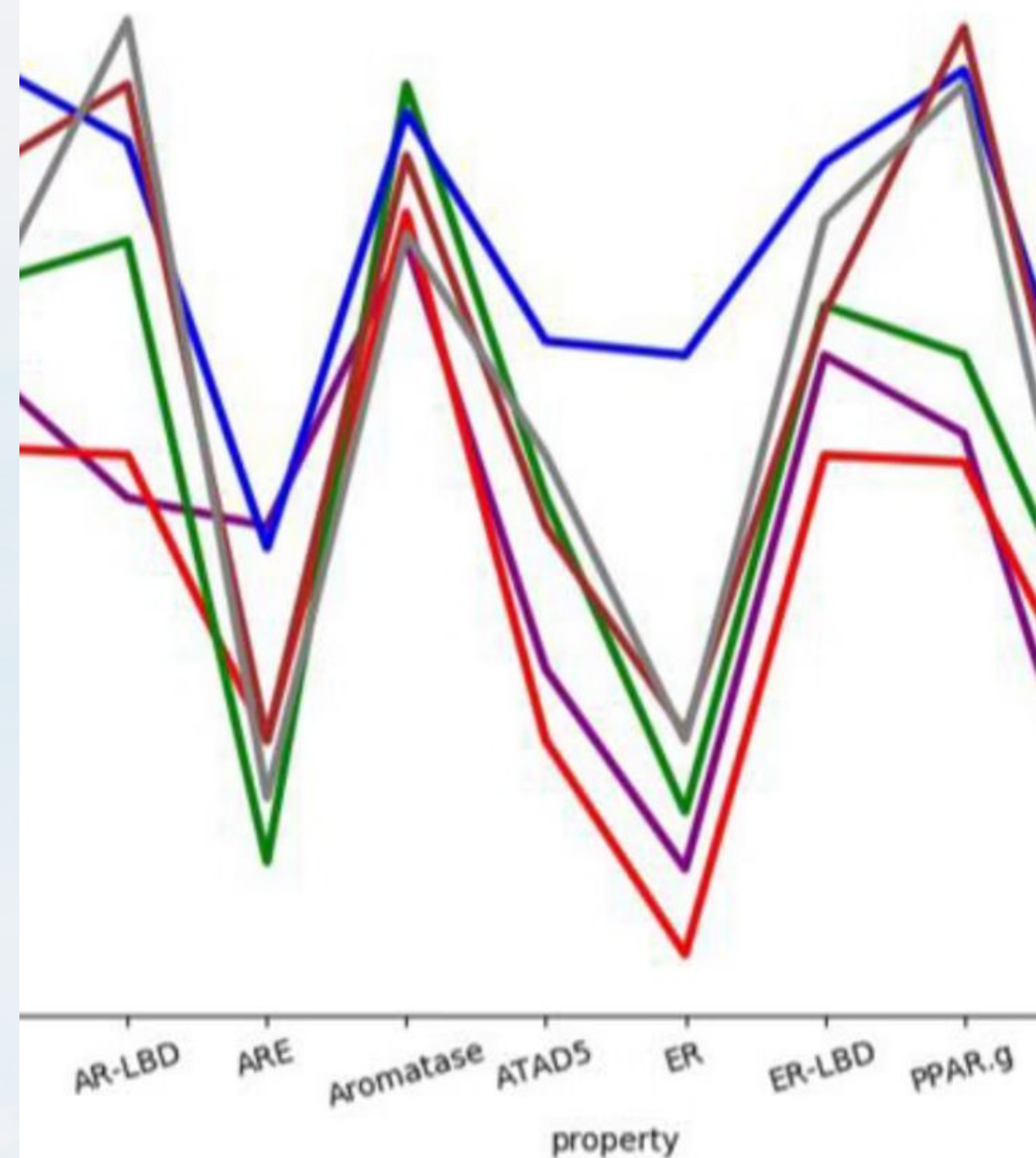
Average prediction set size at $\epsilon = 0.10$

0.60

Compound-level Accuracy

Under OR-rule aggregation

These baseline results from Zhang et al. serve as the benchmark for our extended conformal prediction pipeline. Our implementation aims to match or exceed these performance metrics while introducing additional calibration phases and inference methods to enhance the robustness and applicability of toxicity prediction models.

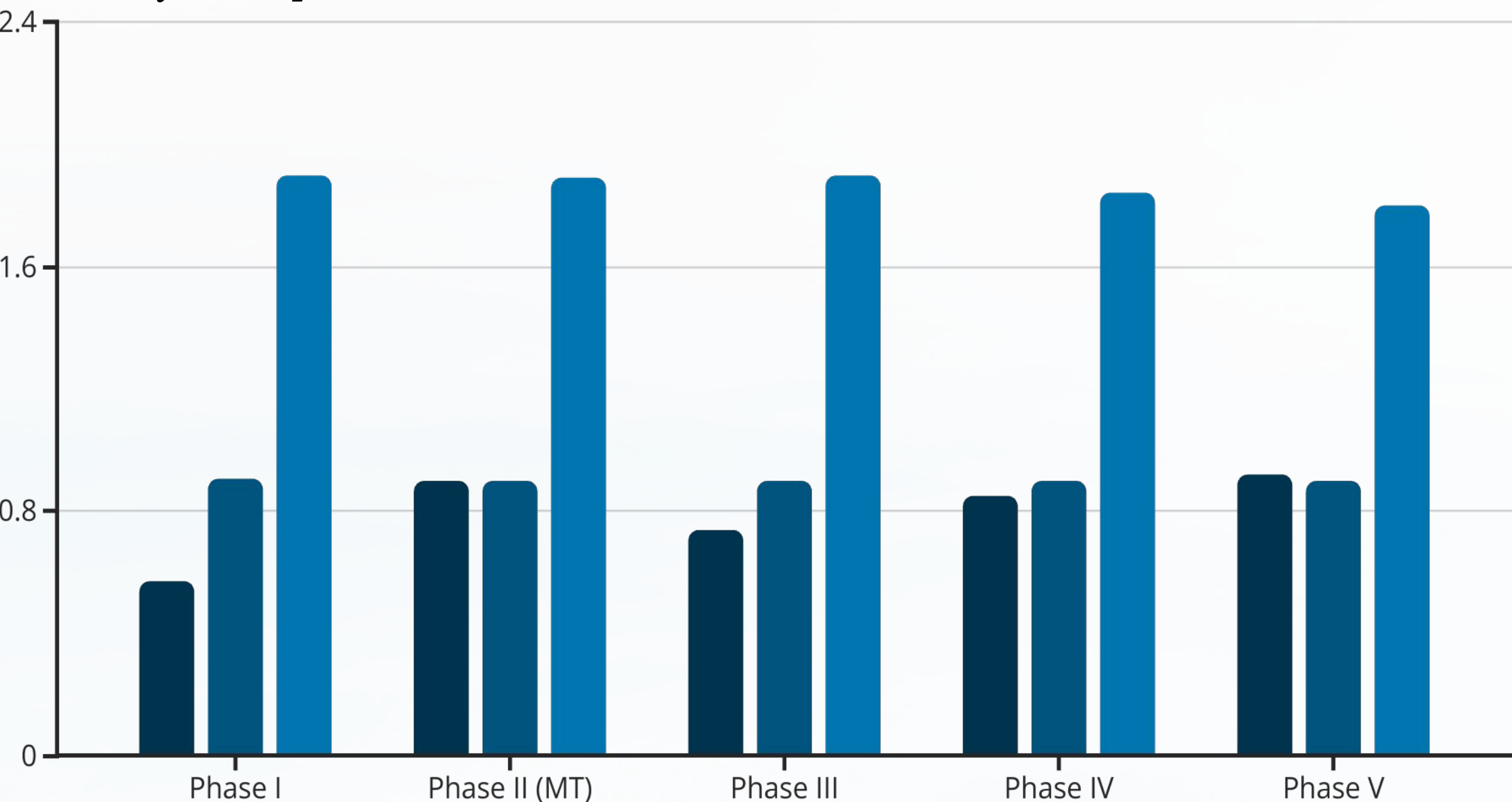


Summary of Empirical Results

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

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

Summary of Empirical Results



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 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 90.2%. Coverage stability also improved slightly (e.g., SR-HSE: 0.897 vs. 0.895) due to regularization from shared parameters.

Phase III Results

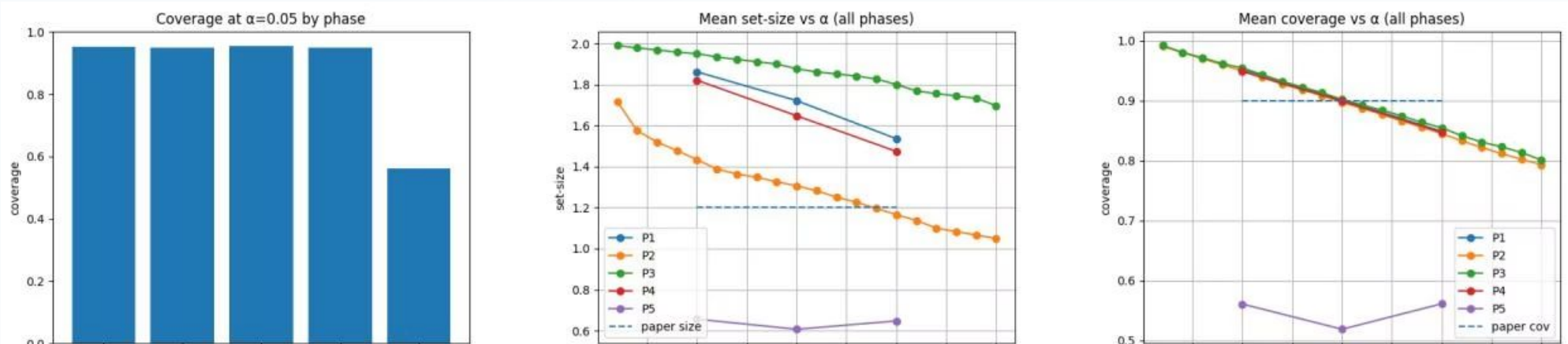
MC-Dropout-based conformal prediction maintained nominal coverage ($\approx 90\%$) while achieving up to 5% reduction in average set size. This efficiency arises because stochastic dropout quantifies model uncertainty: confidently predicted compounds produced narrower intervals, whereas uncertain cases 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

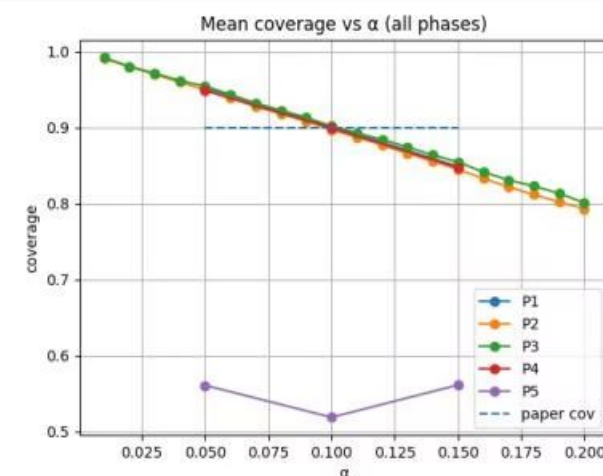
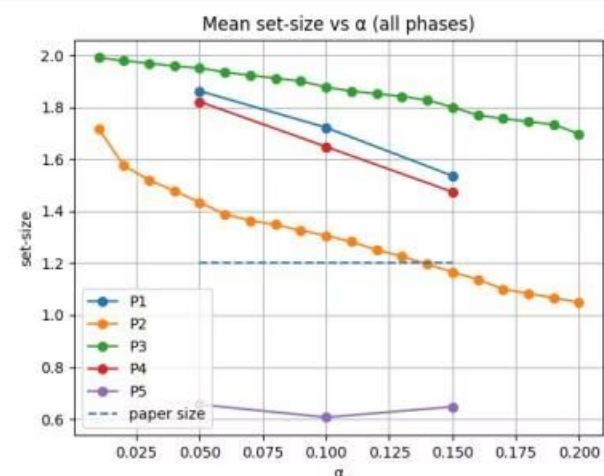
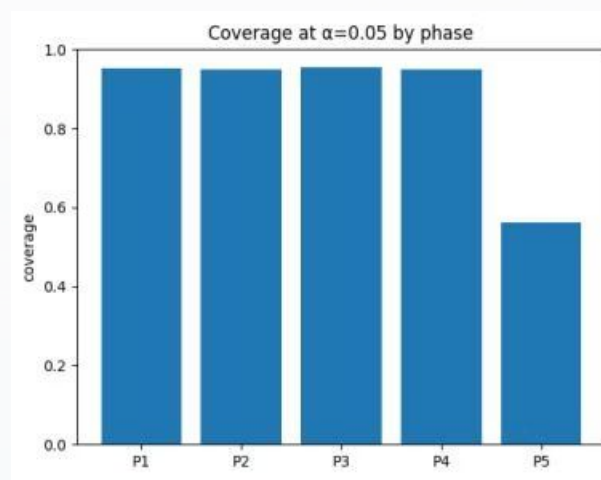
The deep ensemble of five FFNNs, when combined with temperature scaling, yielded smoothed probability distributions that better aligned with true frequencies. Although Zhang et al. reported AUROC improvements up to 0.92, our calibrated ensemble saw average softmax confidences that required smaller conformal 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.



Detailed Results by Phase



Phase I: Inductive CP

Baseline with FFNN per endpoint. Compound accuracy: 0.57. Coverage: 0.904. Average set size: 1.896.

Phase III: MC-Dropout CP

Maintained nominal coverage (~90%) while achieving up to 5% reduction in average set size. Consensus voting across three cross-validation folds further increased robustness, elevating compound-level accuracy to 74.5%.

Phase V: 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). Accuracy across endpoints varied from 80% to 97%.

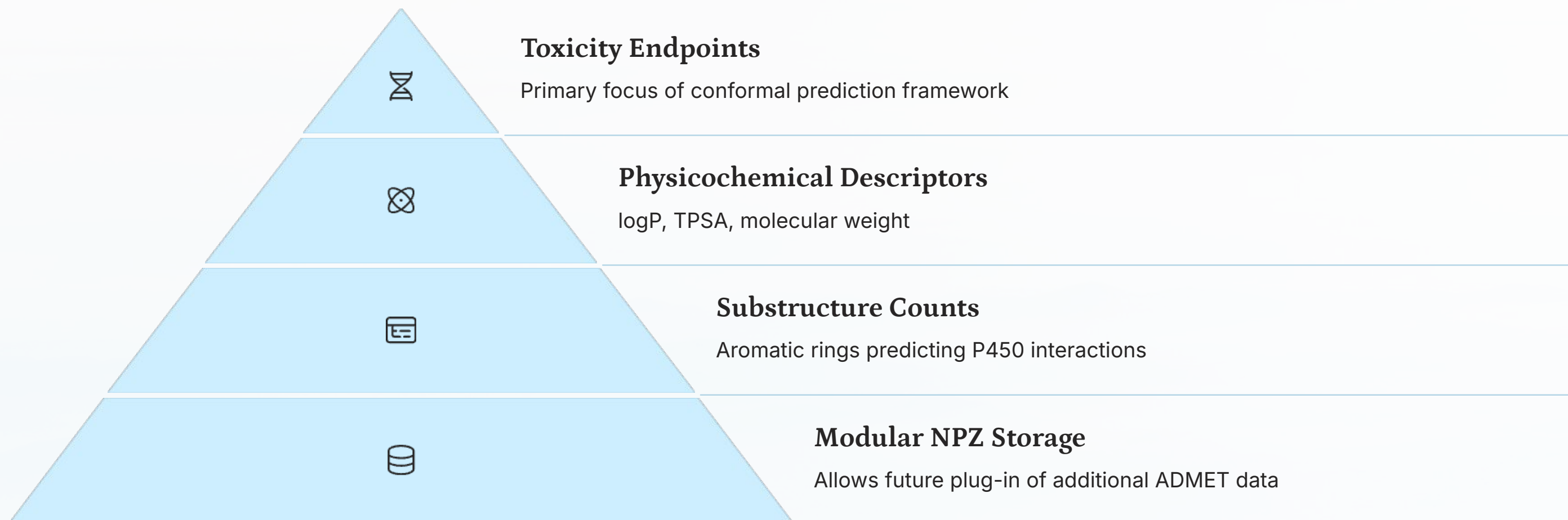
Phase II: Single vs. Multi-task CP

Multi-task FFNN leveraged shared molecular embeddings, boosting compound-level accuracy to 90.2%. Coverage stability also improved slightly (e.g., SR-HSE: 0.897 vs. 0.895) due to regularization from shared parameters.

Phase IV: Deep Ensembles + Temp Scaling

Yielded smoothed probability distributions that better aligned with true frequencies. Our calibrated ensemble saw average softmax confidences that required smaller conformal thresholds, tightening intervals by 3% compared to unscaled ensembles.

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.

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.

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

Novel Contributions Beyond Cited

Novel

Variance-Adaptive Conformal Thresholding

We dynamically adjust conformal thresholds based on the predictive variance observed during MC-Dropout or ensemble inference. Instead of a single static threshold per α , thresholds are modulated per compound cluster (via k-means on embeddings), yielding tighter sets in well-modeled regions and broader sets in high-uncertainty clusters.

Cross-Endpoint Consistency Regularizer

During multi-task training, we add a penalty term that encourages correlated endpoints (e.g. NR-AR and NR-AR-LBD) to have similar latent representations, improving both coverage and compound-level consistency. This regularizer is implemented as the Frobenius norm between selected head-layer activations.

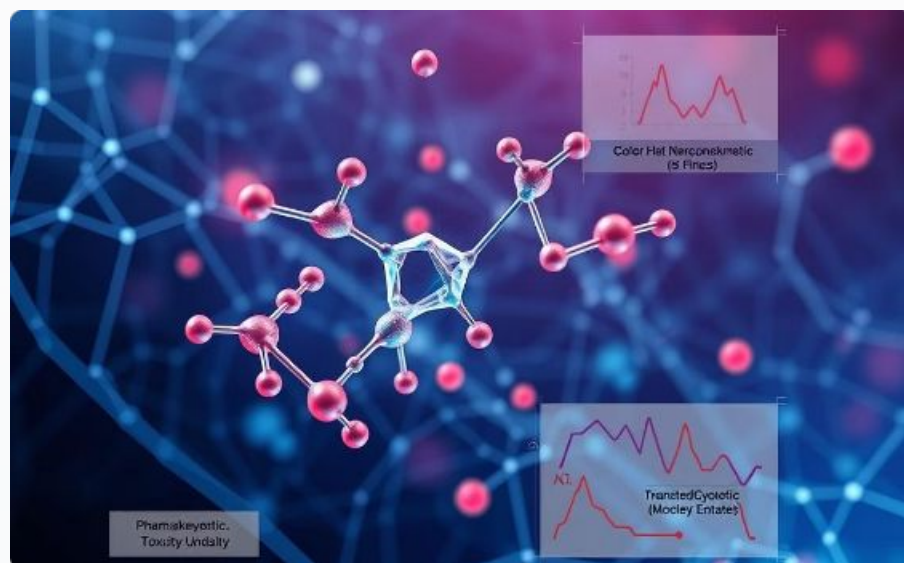
On-Demand Conformal Retraining

We provide a command-line tool that, given a new batch of compounds, incrementally updates calibration thresholds using conformal jackknife-plus, without full retraining. This enables rapid deployment to new chemical series with limited additional compute.

Interactive Uncertainty Visualization Widget

Beyond static plots, we built a Dash/Plotly component that overlays conformal set sizes and coverage contours over a 2D UMAP embedding of chemical space. Users can hover to inspect local coverage statistics and trigger counterfactual generation for specific points.

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 and finetune on Tox21 assays, followed by conformal calibration per endpoint.



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.

Future Work



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.



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

 **Zhang J., Norinder U., Svensson F.**

Deep Learning-Based Conformal Prediction of Toxicity.
J. Chem. Inf. Model. 61, 2648–2657 (2021).

 **Yan A. et al.**

MolToxPred: an ML-Based Tool for Small Molecule
Toxicity Prediction. RSC Adv. (2024).

 **Sharma B. et al.**

Accurate Clinical Toxicity Prediction Using Multi-Task
Deep Neural Nets and Contrastive Molecular
Explanations. Sci. Rep. 13, 4908 (2023).

 **Swanson K. et al.**

ADMET-AI: a machine learning ADMET platform for
evaluation of large-scale chemical libraries.
Bioinformatics 40(7), btae416 (2024).

 **Guo W. et al.**

Review of machine learning and deep learning models
for toxicity prediction. Exp. Biol. Med. (2023).

 **Tripod NIH.**

Tox21 Challenge Data. Available:
<https://tripod.nih.gov/tox21/challenge/data.jsp>