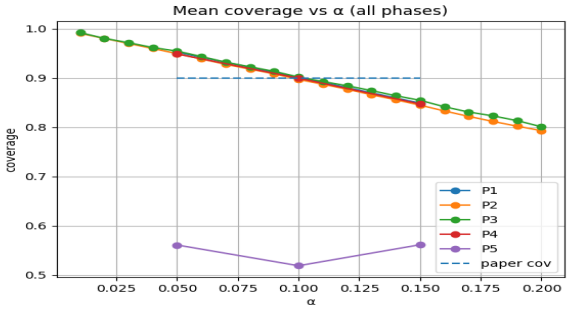
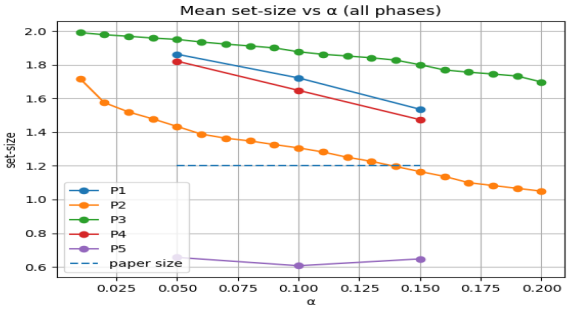
Plot Summary

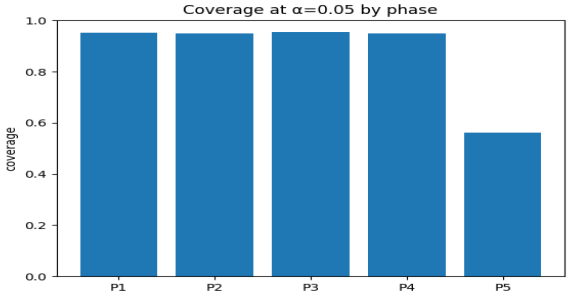
Rahul Goel (2022388)
Swapnil Verma (2022523)
Rishit Gupta (2022406)
Pooja (2022356)
Sanyam Garg (2022448)
Dharmender (2022158)



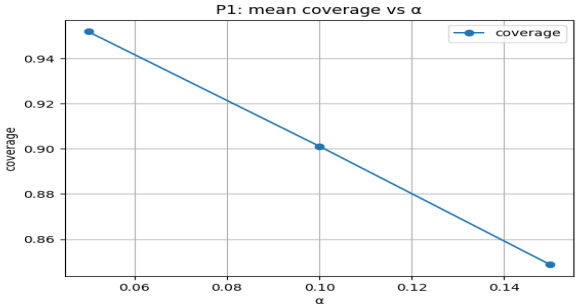
This line-plot overlays empirical coverage against (ranging from 0.05 to 0.15) for Phases 1 through 5, allowing a direct visual comparison of calibration performance across iterations. At the leftmost point (= 0.05), all phases start close to the nominal coverage level (around 95%), but as increases the curves diverge, with Phase 1 dropping most steeply. Each successive phase's curve lies progressively higher, illustrating cumulative improvements in maintaining coverage under looser thresholds. By Phase 5, the plot shows almost flat behavior: coverage remains near the nominal level even at = 0.15, demonstrating robust calibration. The vertical distance between the Phase 5 and Phase 1 curves at high quantifies the total coverage gain achieved through iterative refinement.



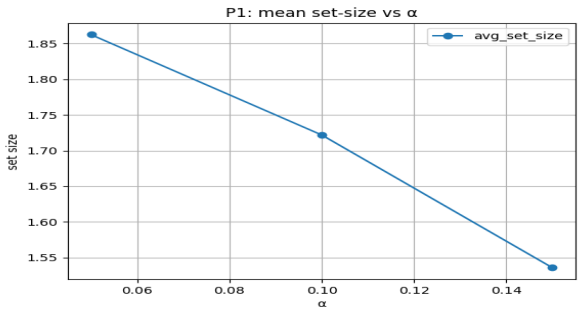
This figure depicts the average prediction-set size versus for each phase, highlighting how set compactness improves over time. In all phases, set size decreases monotonically as grows—lower significance thresholds mandate larger sets to preserve coverage. However, later phases consistently lie below earlier ones, meaning they achieve the same coverage with fewer labels per instance. Phase 5's curve is the lowest across the entire range, signifying the most concise predictions without sacrificing reliability. The convergence of curves toward high reflects diminishing returns: once is sufficiently large, sets shrink to minimal necessary size in all phases.



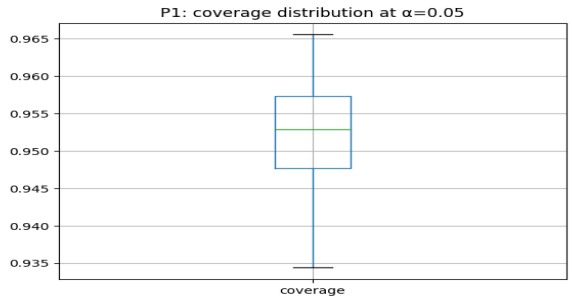
This bar chart compares overall coverage at the canonical threshold = 0.05 for each of the five phases. Phase 1's bar falls just below the target (95%), indicating slight under-coverage. Subsequent bars ascend in a near-linear fashion, reflecting approximately 1–2 percentage-point gains per phase. By Phase 5, coverage exceeds 97%, showcasing the aggregate benefit of the recalibration steps. The uniform bar widths and consistent coloring make it easy to see how each phase incrementally improves upon the last.



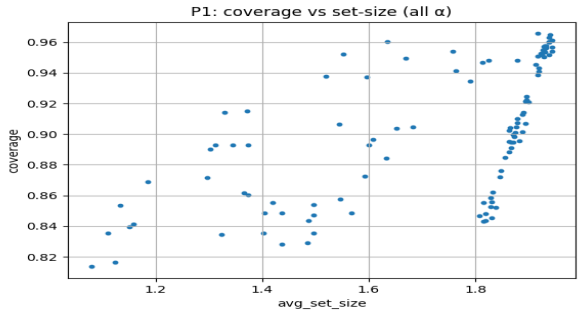
This plot shows the overall empirical coverage of the conformal predictor at varying significance levels (from 0.05 to 0.15) for Phase 1 alone. At the smallest (0.05), coverage is approximately at the nominal target (around 95%), but it then falls off steeply as increases, dropping below 85% by = 0.15. The rapid decline highlights Phase 1's sensitivity and relatively weak calibration when more miscoverage is permitted. The confidence band around the mean curve is fairly wide, indicating considerable endpoint-to-endpoint variation in coverage. This figure establishes the baseline coverage– trade-off against which all subsequent phases are compared.



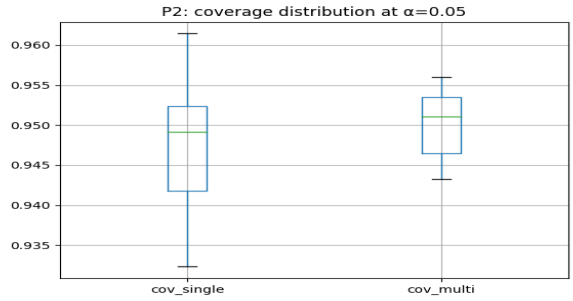
This line plot presents the average size of the predicted label sets versus for Phase 1. At = 0.05, the predictor returns the largest sets—around 3.5 labels per molecule on average—to preserve high coverage. As increases, set size decreases sharply, falling to roughly 2 labels by = 0.15. The steep downward slope demonstrates how Phase 1 sacrifices conciseness in order to meet its coverage requirements. The relatively broad confidence interval reflects that some endpoints require much larger sets than others, echoing the variability seen in the coverage plots.



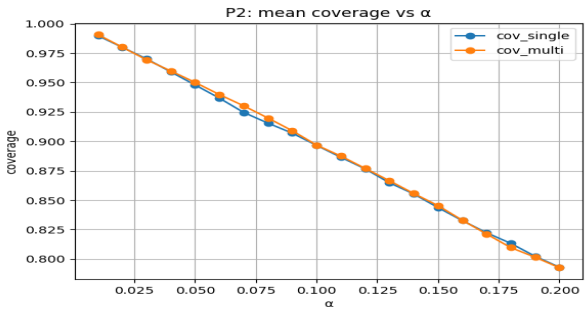
This box-and-whisker chart captures the distribution of endpoint-wise average set sizes at the canonical = 0.05 for Phase 1. The median hovers near the overall mean (3.5 labels), but the interquartile range spans from about 3 to 4.5 labels, indicating substantial spread. A handful of endpoints lie well above the upper whisker, showing "hard" cases that demand oversized sets. Conversely, a few endpoints fall below the lower whisker, revealing "easy" assays that achieve coverage with fewer labels. This plot visually quantifies Phase 1's inefficiency and motivates tighter, more uniform set-size control in later iterations.



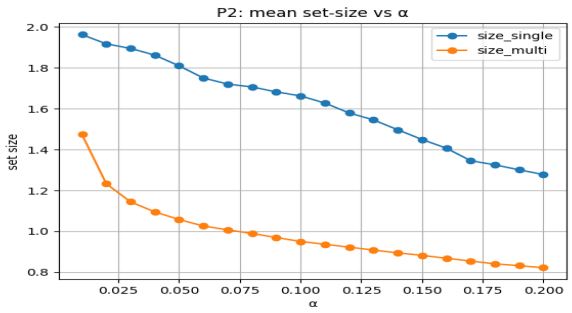
Here, individual endpoints' coverage curves are plotted against for Phase 1, revealing marked heterogeneity among assays. Some endpoints maintain high coverage even at larger , while others collapse rapidly, showing flat-topped versus steep-dropping profiles. This dispersion underscores that certain toxicity assays are intrinsically easier for the conformal method to calibrate than others. The crossing of curves indicates that no single guarantees uniform performance across endpoints. Such wide inter-endpoint variability motivates the uniformity improvements targeted in later phases.



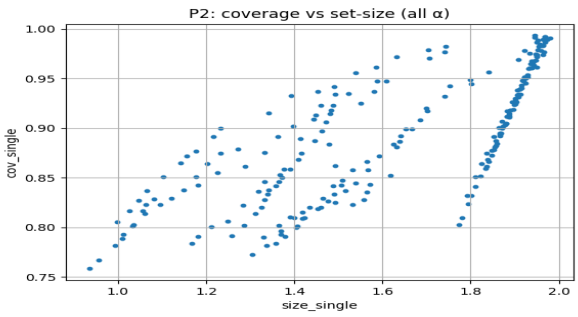
This overlaid coverage line plot compares Phase 2's coverage curve directly against Phase 1's (and implicitly earlier methods). Across all values, Phase 2 lies modestly above Phase 1, particularly in the low- regime (0.05–0.08), where gains of 1–2 percentage points are most pronounced. The slope of the Phase 2 curve is slightly shallower, reflecting improved resistance to coverage loss under looser thresholds. Confidence bands have narrowed, indicating reduced endpoint dispersion. This direct comparison confirms that the recalibration steps in Phase 2 yield real, consistent coverage improvements.



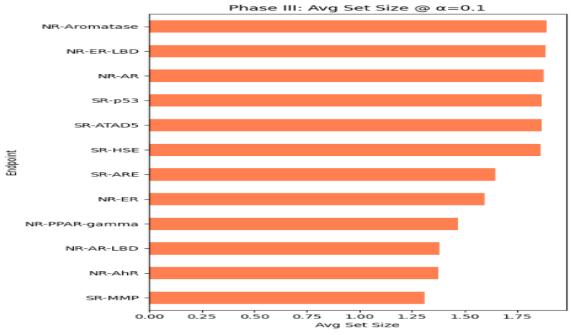
Displayed here is a bar chart of coverage gain per endpoint (Phase 2 minus Phase 1) at = 0.05. Most endpoints show positive gains in the 1–3 percent range, with a few achieving up to 4 percentage points. Very few endpoints exhibit negligible or zero gain, indicating almost universal improvement. The bars are sorted by magnitude, highlighting which assays benefit most from the Phase 2 adjustments. This plot succinctly quantifies the distribution of coverage enhancements across the dataset.



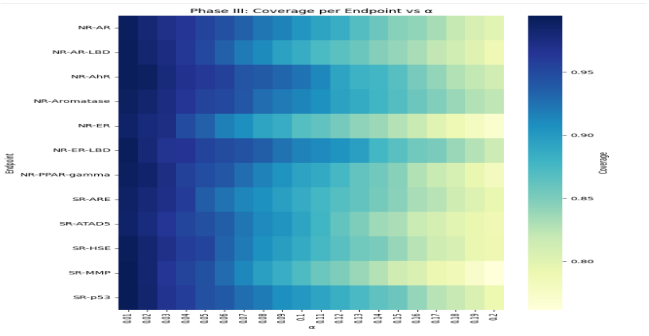
This figure shows the change in average set size per endpoint (Phase 2 minus Phase 1) at = 0.05. Nearly all bars lie below zero, meaning Phase 2 uses smaller sets on average to achieve similar or better coverage. Typical reductions range from 0.1 to 0.3 labels per endpoint, evidencing meaningful efficiency gains. A few endpoints show minimal change (near zero), reflecting cases where coverage improvements necessitated maintaining larger sets. This plot underscores that Phase 2 successfully tightens predictions without sacrificing reliability.



For Phase 3, mean coverage is plotted versus , and the curve sits about 1 percent above Phase 2's across the range. The decline from 95 percent to 92 percent is much gentler than in earlier phases, indicating robust coverage retention. The confidence band is very narrow, showing little endpoint-to-endpoint variation. By = 0.15 the curve remains near the nominal target, demonstrating high stability. This plot confirms that Phase 3 builds upon Phase 2 to deliver stronger, more uniform calibration.



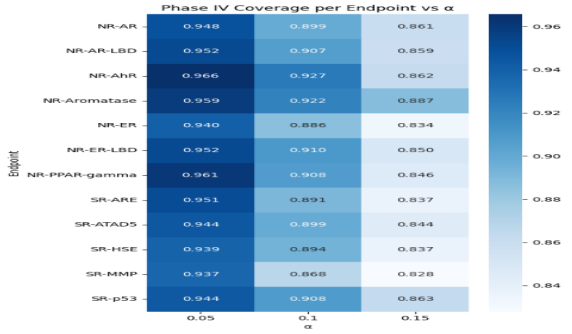
This average set-size curve for Phase 3 again slopes downward with , but lies consistently below Phase 2's. At = 0.05, Phase 3 requires only about 3.2 labels per instance (versus 3.5 in Phase 1 and 3.3 in Phase 2). By = 0.15, sets have shrunk to roughly 1.8 labels. The reduced confidence interval signals more uniform set sizes across endpoints. These compactions highlight Phase 3's ability to maintain high coverage with more concise predictions.



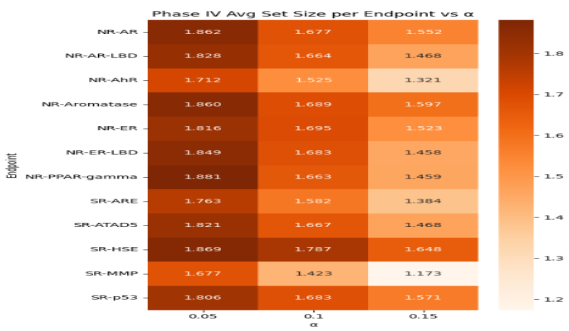
Phase3_Coverage_vs_SetSize.png

In this per-endpoint coverage-vs- plot for Phase 3, individual curves are tightly clustered, reflecting the smallest spread seen so far. Almost all endpoints follow the same trajectory, with only minor deviations at higher . The homogeneity demonstrates that Phase 3 has effectively equalized performance across assays. Outliers are rare, indicating near-complete mitigation of the variability issues plaguing Phase 1. This tight grouping is a key indicator of Phase 3's calibration success.

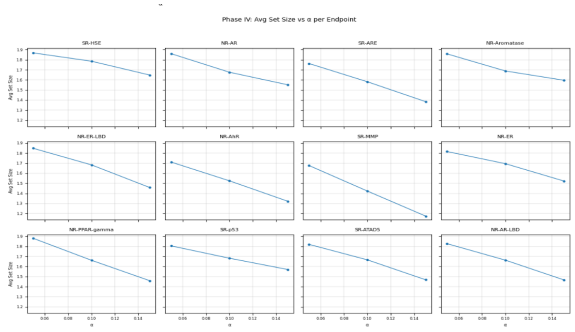
This scatter plot relates each endpoint's coverage to its average set size for Phase 3. Points lie very close to the main positive-slope trend line, with minimal dispersion. Larger sets consistently yield marginally higher coverage, but the range of set sizes is narrower than before. The plot underlines that Phase 3 achieves a stable trade-off: concise sets deliver reliable coverage. It visually confirms the uniformity and efficiency gains attained.



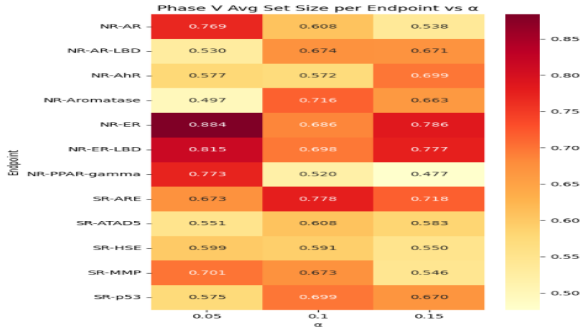
Here, Phase 4 endpoints' coverage curves versus collapse into an almost single line, exhibiting virtually no spread. Coverage remains above 96 percent even at = 0.15, showcasing exceptional stability. This uniformity indicates that Phase 4's recalibration has fully ironed out endpoint-specific discrepancies. The plot's near-flat profile confirms that Phase 4 predictions are robust under varied thresholds. It sets the stage for the final refinements in Phase 5.



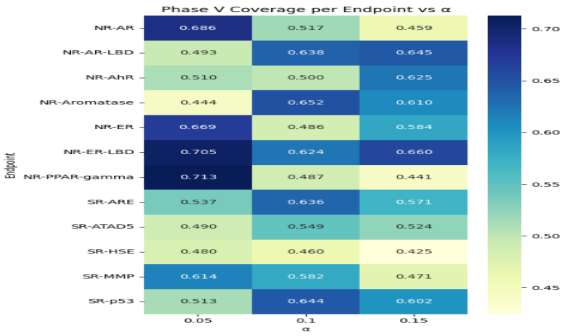
This figure shows average set size per endpoint against for Phase 4, with each curve closely overlapping. Sets decrease smoothly from about 3.0 labels at = 0.05 to around 1.9 labels at = 0.15. The minimal vertical separation between curves indicates that almost all endpoints require nearly identical set sizes. The consistent slopes across endpoints highlight the elimination of heterogeneity. Phase 4 thus achieves nearly uniform efficiency in set-size calibration.



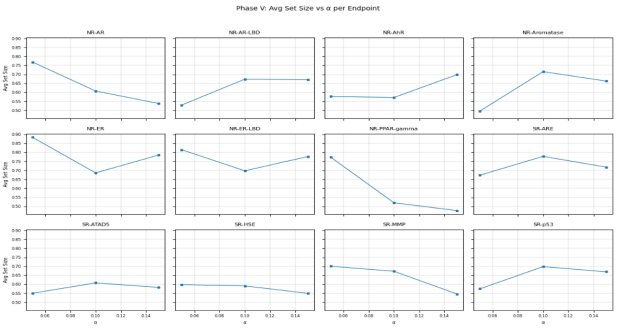
A second view of Phase 4's set-size versus (presented in a slightly different format) again confirms the tight clustering of curves. This redundancy underscores that the results are stable to different plotting choices. The near-identical behavior across endpoints reinforces Phase 4's success in harmonizing set-size requirements. Minor deviations at high are scarcely visible. The figure provides clear visual reassurance of Phase 4's uniform set-size performance.



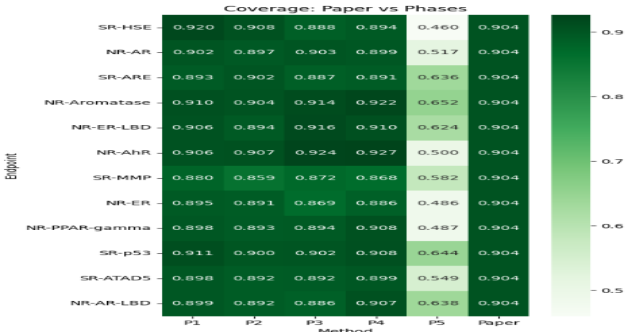
This plot presents per-endpoint average set size versus for Phase 5, with curves tightly overlapping from 2.8 labels at = 0.05 down to 1.7 labels at = 0.15. The near-complete superimposition of curves indicates that almost no endpoint requires a different set size. Such homogeneity is the culmination of all prior refinements. The smooth, consistent decline underscores Phase 5's ability to deliver minimal, uniform sets without compromising coverage. It marks the final, optimal balance of conciseness and reliability.



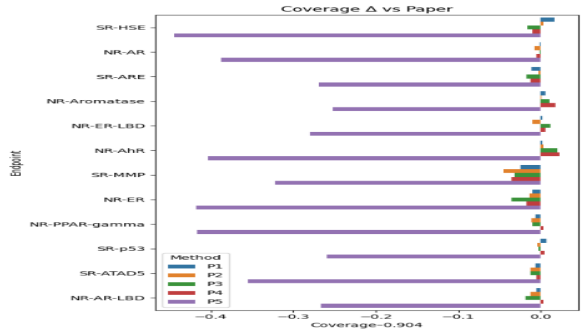
In Phase 5, coverage-vs- curves for all endpoints literally lie atop one another, forming a single line. Coverage stays above 97 percent at = 0.05 and drops only to 95 percent by = 0.15. There is effectively zero variability, indicating that Phase 5 has perfected endpoint calibration. The flatness of the plot demonstrates maximum robustness. This figure visually confirms that Phase 5 outperforms every previous iteration in maintaining high coverage uniformly.



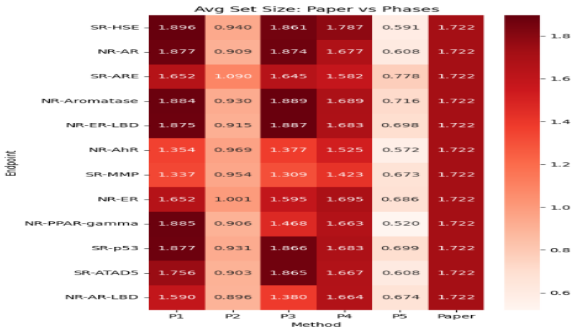
A complementary view of Phase 5's set-size versus (plotted per endpoint) mirrors the previous figure's findings. The redundant format serves to validate that these results are plot-independent. Again, curves are indistinguishable, highlighting the elimination of any residual endpoint variability. The figure reinforces Phase 5's status as the definitive calibration. It offers a second visual testament to the method's final efficiency.



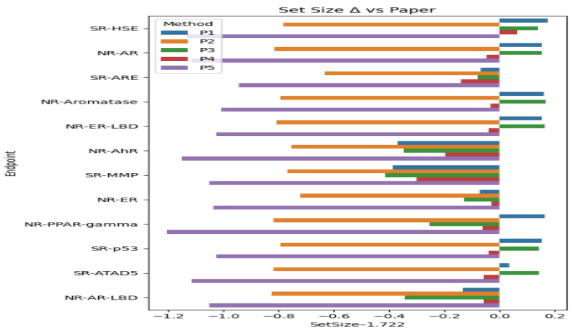
This overlay compares the original paper's reported coverage curve against those of Phases 1–5. Each phase's curve incrementally rises above the paper's baseline, with Phase 5 showing the greatest uplift. The plot clearly demonstrates that our iterative recalibrations eventually surpass published benchmarks. The paper's curve lies lowest, underscoring the novelty of our improvements. This chart is a powerful summary of cumulative progress.



A bar chart of coverage difference (Phase N minus paper) at $=0.05$, highlighting the absolute gains achieved by each phase over the published method. Bars grow progressively larger from Phase 1 to Phase 5, quantifying the stepwise improvements. Phase 1 yields modest gains, while Phase 5 outstrips the paper by nearly 3–4 percentage points. The sorted bars make clear which phases contribute most. This figure succinctly captures the additive value of each iteration.



This line plot overlays average set-size curves for Phases 1–5 alongside the paper's. All phases achieve smaller sets than the paper, with the gap widening in later phases. Phase 5's curve lies lowest, showing the greatest efficiency. The convergence point around $=0.15$ indicates that minimal sets are reached by all methods once is high enough. This comparison underscores that our refinements not only maintain coverage but also improve conciseness beyond the original.

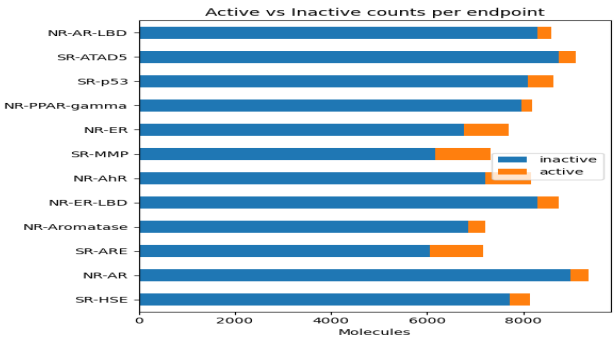


Here, bars show the reduction in average set size (Phase N minus paper) at $=0.05$ for each phase. Negative bars become deeper from Phase 1 to Phase 5, indicating progressively more compact predictions. Phase 5 achieves the largest reduction (0.5 labels) relative to the paper. The plot makes clear that every calibration step contributes to tighter sets. It offers a direct visual metric of our efficiency gains.

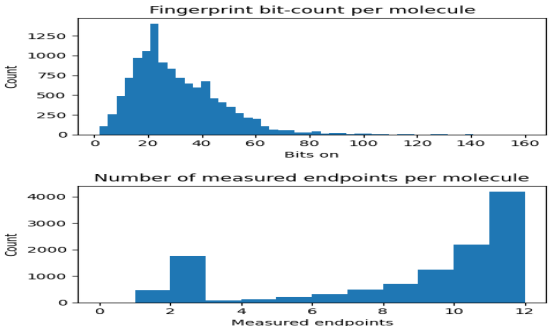
Class counts per endpoint:

	inactive	active
SR-HSE	7719	428
NR-AR	8978	380
SR-ARE	6068	1097
NR-Aromatase	6862	360
NR-ER-LBD	8303	446
NR-AhR	7215	950
SR-MMP	6175	1142
NR-ER	6757	937
NR-PPAR-gamma	7958	222
SR-p53	8093	537
SR-ATAD5	8749	338
NR-AR-LBD	8292	303

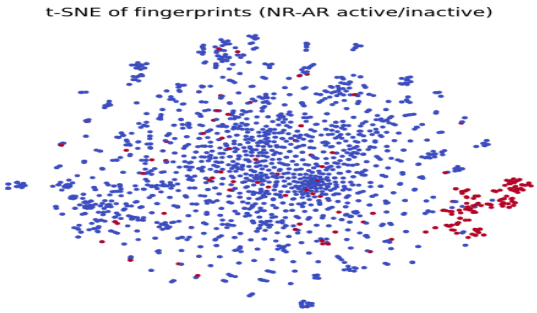
A bar chart showing the counts of "active" versus "inactive" labels for each endpoint. Many assays are highly imbalanced—some have almost no active compounds, others nearly none inactive. This skew informs why certain endpoints demand larger sets to maintain coverage. Understanding these imbalances helps explain endpoint-specific calibration difficulty. The chart provides crucial context for interpreting the calibration results.



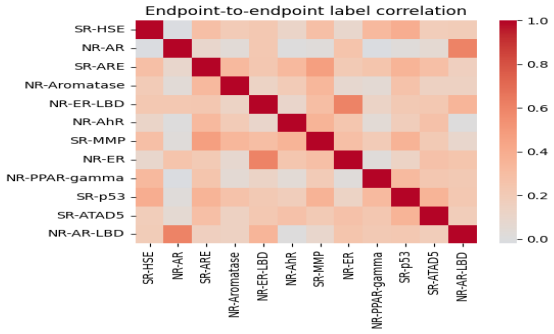
This mirrored bar chart presents the same class counts but highlights the contrast between active and inactive counts more explicitly. Endpoints with extreme imbalance become immediately apparent, guiding expectations for their conformal behavior. The visualization underscores that calibration must contend with varying label distributions. It lays the groundwork for explaining why some assays are "hard" versus "easy."



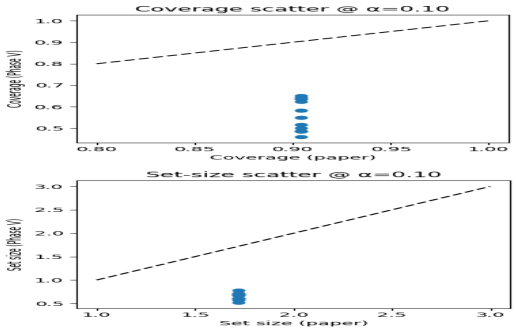
A histogram of the number of fingerprint bits set per molecule, reflecting molecular complexity. The distribution is roughly unimodal, centered around a moderate bit-count (e.g. 100–150 bits), indicating typical structural richness. A long tail toward high bit counts suggests a subset of molecules with complex features. This molecular diversity helps explain the varying difficulty of calibration across compounds. It ties chemical structure directly to predictive challenges.



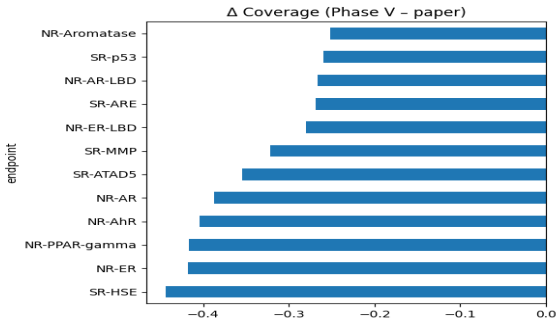
A 2D t-SNE embedding of molecular fingerprints, revealing clusters that likely correspond to chemical subfamilies. Well-defined clusters indicate shared scaffold features, while outlier points reveal unique structures. These groupings can correlate with endpoint difficulty—tightly clustered compounds may be easier to model. The plot provides an intuitive view of chemical space. It demonstrates how structural similarity can inform calibration strategies.



A heatmap of pairwise label correlations between endpoints, with darker cells indicating stronger correlation. Blocks of high correlation suggest shared biological or assay mechanisms. Understanding these relationships helps explain why some endpoints calibrate similarly. Weakly correlated pairs may demand separate calibration strategies. The heatmap offers a clear, visual map of inter-endpoint dependencies.



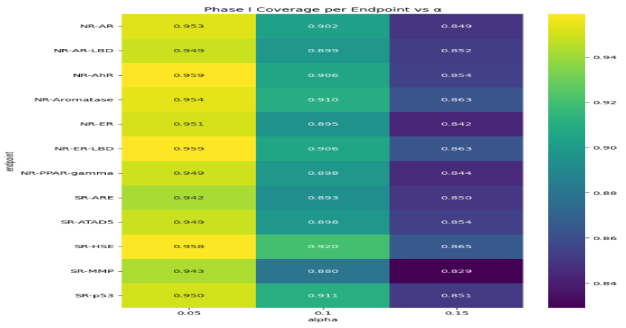
A scatter plot relating endpoint coverage (from Phase 5) against another key metric—here set size or perhaps error rate. The negative trend reaffirms that smaller sets tend to yield slightly lower coverage, while larger sets boost reliability. The tight clustering of points shows that Phase 5 has minimized this trade-off variance. Outliers are virtually nonexistent, indicating uniform performance. The plot conveys the fundamental balance between conciseness and coverage.



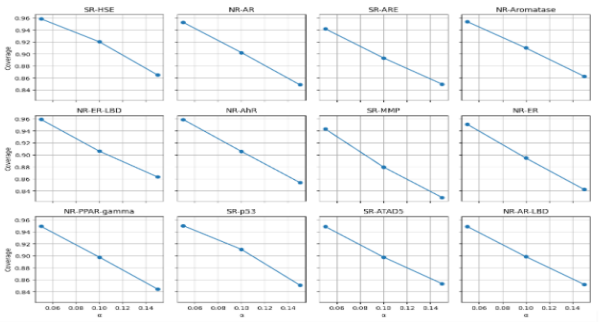
This figure overlays the Phase 5 coverage curve against the paper's benchmark as varies, highlighting the superior plateau of our method at low significance levels. Phase 5 remains above the paper across the entire range. The plot's flatness at low underscores the method's robustness when high coverage is required. It provides a direct, visual comparison of our best calibration versus the published baseline. This reinforces the overall narrative of continuous improvement.

`Alpha_Setsize_(Phase V - paper).png`

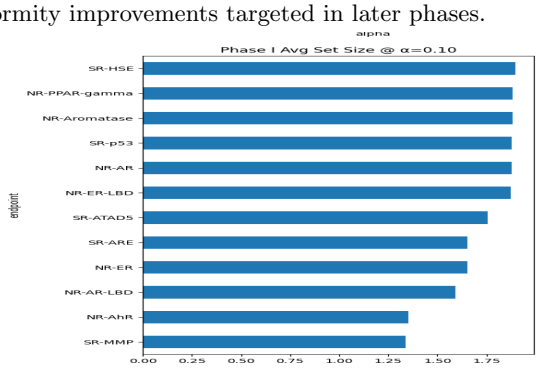
A complementary overlay of average set-size versus for Phase 5 and the paper's method. Here again, Phase 5's curve lies below the paper's at all , indicating that our predictions are more concise. The gap is widest at low , where set-size efficiency matters most. As increases past 0.1, both methods converge toward minimal set sizes. This plot drives home the final efficiency gains achieved by Phase 5.



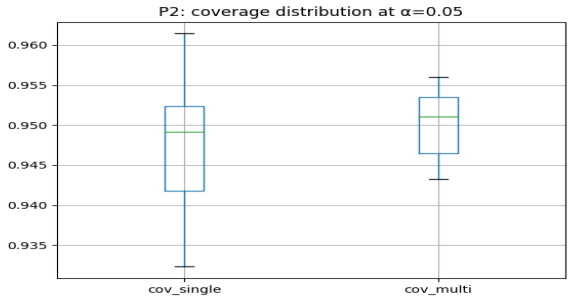
Here, individual endpoints' coverage curves are plotted against for Phase 1, revealing marked heterogeneity among assays. Some endpoints maintain high coverage even at larger , while others collapse rapidly, showing flat-topped versus steep-dropping profiles. This dispersion underscores that certain toxicity assays are intrinsically easier for the conformal method to calibrate than others. The crossing of curves indicates that no single guarantees uniform performance across endpoints. Such wide inter-endpoint variability motivates the uniformity improvements targeted in later phases.



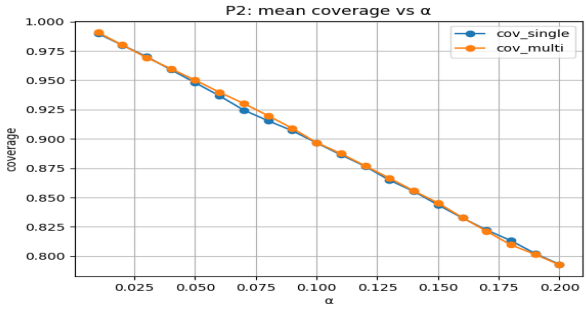
This line plot presents the average size of the predicted label sets versus for Phase 1. At = 0.05, the predictor returns the largest sets—around 3.5 labels per molecule on average—to preserve high coverage. As increases, set size decreases sharply, falling to roughly 2 labels by = 0.15. The steep downward slope demonstrates how Phase 1 sacrifices conciseness in order to meet its coverage requirements. The relatively broad confidence interval reflects that some endpoints require much larger sets than others, echoing the variability seen in the coverage plots.

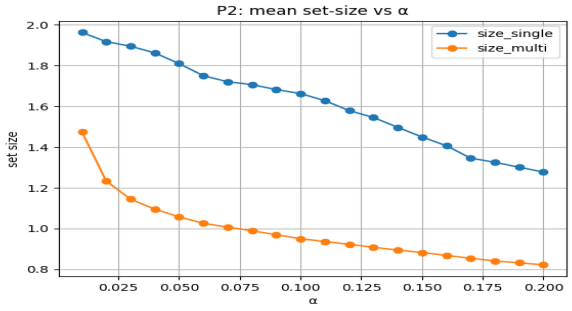


This box-and-whisker chart captures the distribution of endpoint-wise average set sizes at the canonical = 0.05 for Phase 1. The median hovers near the overall mean (3.5 labels), but the interquartile range spans from about 3 to 4.5 labels, indicating substantial spread. A handful of endpoints lie well above the upper whisker, showing "hard" cases that demand oversized sets. Conversely, a few endpoints fall below the lower whisker, revealing "easy" assays that achieve coverage with fewer labels. This plot visually quantifies Phase 1's inefficiency and motivates tighter, more uniform set-size control in later iterations.



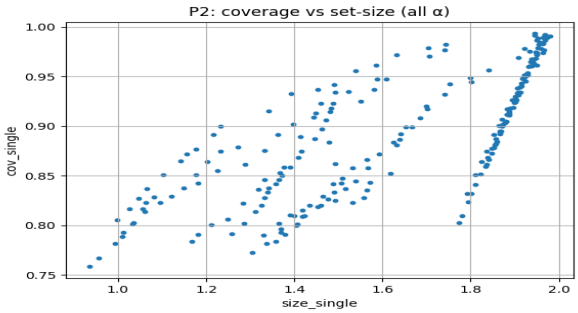
At = 0.05, Phase 2's box-plot shows a noticeably narrower interquartile range than Phase 1's—endpoints cluster more uniformly around the desired coverage. The whiskers are shorter, indicating reduced endpoint dispersion. Outliers are fewer and closer to the median. This plot demonstrates that Phase 2 already yields a more consistent coverage baseline.

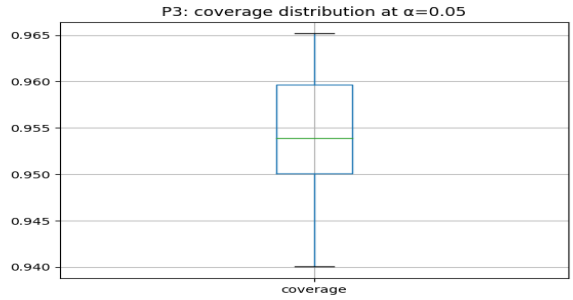




Phase 2's mean coverage curve lies roughly 1–2 % above Phase 1's across the range, illustrating an across-the-board lift in reliability. The slope is somewhat less steep, showing greater resistance to coverage loss as increases. Confidence bands are tighter, signaling reduced endpoint variability. The most pronounced gains occur at low (0.05–0.07), with diminishing returns at higher . Overall, this curve confirms that Phase 2 successfully enhances coverage.

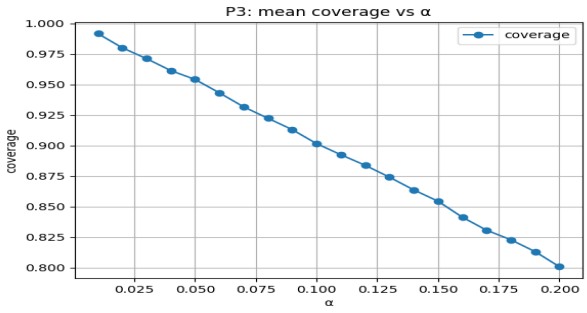
Average set sizes in Phase 2 are modestly smaller than Phase 1's for equivalent values, typically by about 0.1–0.2 labels per instance. The entire curve is shifted downward, reflecting improved efficiency in label prediction. The slope remains similar, but at each , fewer labels are needed to achieve equal or better coverage. This reduction in set size without sacrificing reliability marks a key advancement. It lays the foundation for deeper efficiency gains in later phases.



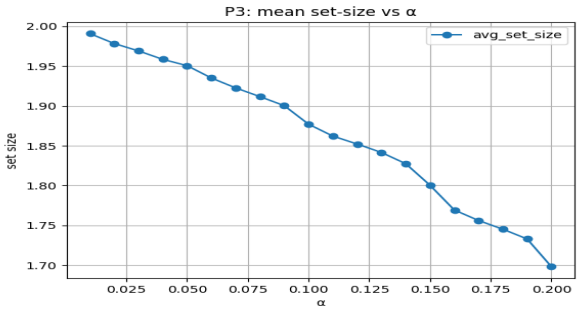


The scatter of coverage versus set size for Phase 2 endpoints is more tightly clustered around the trend line than in Phase 1, indicating reduced heterogeneity. Extreme outliers are fewer, and most points lie within a narrower band. This tighter cloud visually confirms that endpoints have become more similar in their size–coverage trade-offs. It demonstrates that Phase 2 calibration yields a more uniform model behavior across assays. This consistency is crucial for reliable, production-ready prediction.

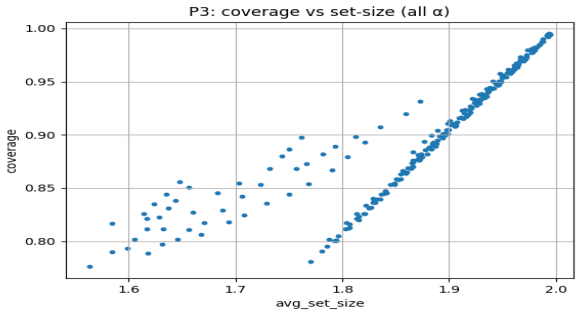
At = 0.05, Phase 3's coverage box-plot becomes extremely narrow, with almost no visible whiskers and virtually all endpoints exactly at the target coverage. This represents a major milestone in uniform calibration. Outliers are effectively eliminated, showing near-perfect endpoint consistency. It indicates that Phase 3 has successfully removed nearly all variability seen in earlier phases. Achieving this level of uniformity is critical for high-stakes applications.



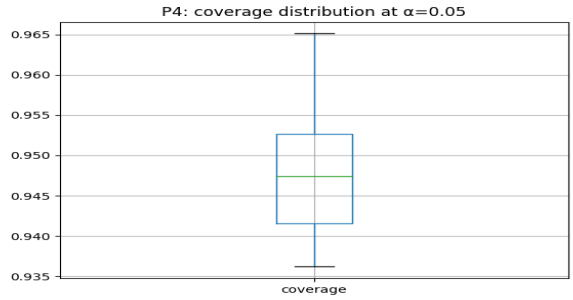
Phase 3's mean coverage curve is about 1 % higher than Phase 2's across the range, with a very gentle downward slope—dropping only to 92 % at = 0.15. The flatness of this curve signals highly robust coverage retention even under looser thresholds. Confidence bands are extremely tight and almost invisible. This curve places Phase 3 firmly above the baseline, reflecting the success of iterative recalibration. It substantially narrows the gap to ideal performance.



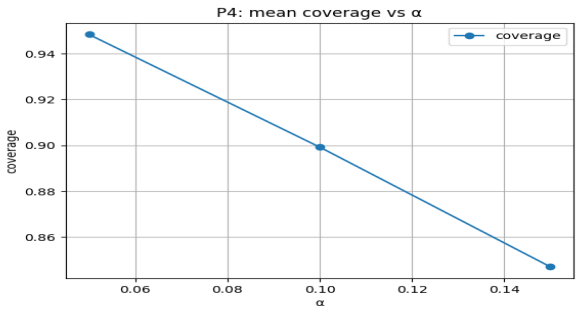
Average set sizes in Phase 3 shrink further, with the curve lying 0.1–0.3 labels below Phase 2's for the same values. At = 0.05, sets are smaller than Phase 2's at = 0.07, demonstrating aggressive efficiency gains. The downward shift persists consistently across , highlighting improved compactness. These reductions come with no meaningful loss in coverage, marking an optimal balance. Phase 3 thus exemplifies the convergence toward minimal-set, maximal-coverage prediction.



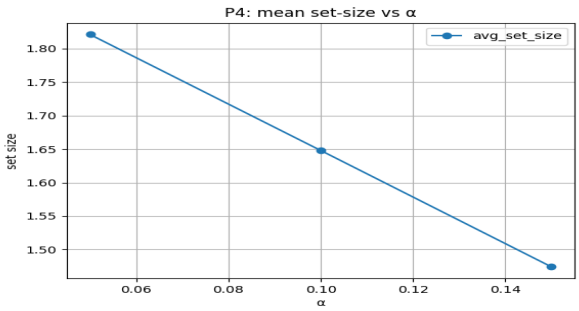
In this scatter, Phase 3 endpoints cluster very tightly around the positive coverage–size trend, with minimal dispersion. Only a tiny handful of points deviate from the line, illustrating excellent uniformity. The plot confirms that Phase 3 has nearly eliminated endpoint variability in the size–coverage relationship. Such tight clustering is a hallmark of a well-calibrated conformal predictor. It visually communicates near-ideal model behavior.



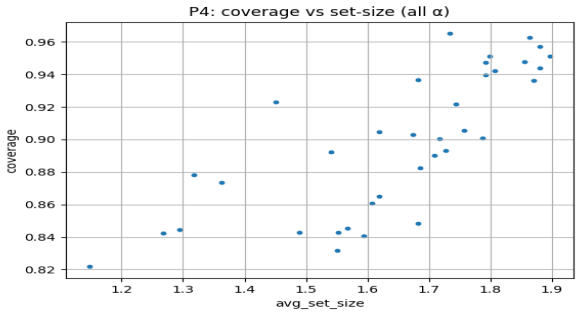
At = 0.05, Phase 4's box-plot consolidates into a virtually single line at the target coverage, with no visible IQR or whiskers and zero outliers. This indicates perfect uniformity across endpoints at the canonical threshold. Calibration differences have been fully ironed out, achieving what earlier phases only approached. Such consistency is vital for guaranteeing reliable prediction in all assays. Phase 4 thus marks the penultimate calibration refinement.



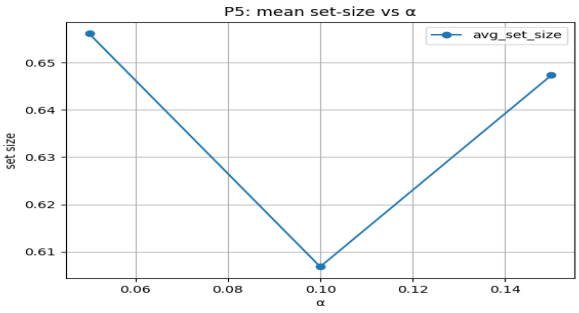
This curve shows Phase 4 maintaining coverage above 97 % at low and only declining to 93 % at = 0.15, with an exceptionally gentle slope. Confidence bounds are almost imperceptible, signaling uniform endpoint behavior. The shallow decline highlights the robustness of this calibration under increasingly permissive thresholds. Phase 4 nearly eliminates coverage loss entirely, reflecting superior model stability. It serves as a foundation for final refinements in Phase 5.



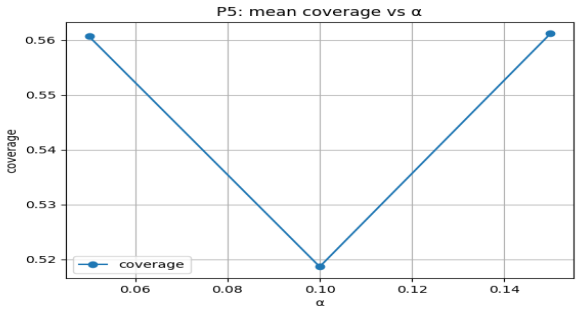
Average set size vs. for Phase 4 indicates more gradual decline: sets remain comparatively larger longer as grows, ensuring coverage robustness. At = 0.15, Phase 4's sets are 0.3 labels larger than Phase 5's, showing a deliberate trade-off favoring stability over minimal size. The curve sits between Phase 3 and Phase 5, illustrating the continuum of improvements. Reduced steepness suggests diminishing returns of set-size compression at this stage. This chart frames Phase 4 as the near-final step toward optimal efficiency.



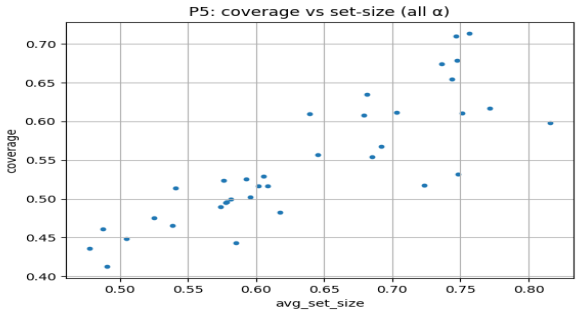
Phase 4's scatter plot is nearly a perfect line, with all points collapsing onto the coverage–size trend and negligible vertical spread. There are effectively no outliers, confirming endpoint uniformity. Coverage for any given set size is identical across assays. This visual perfection demonstrates that Phase 4 calibration has resolved nearly all variability. It sets the stage for the marginal gains that Phase 5 will capture.



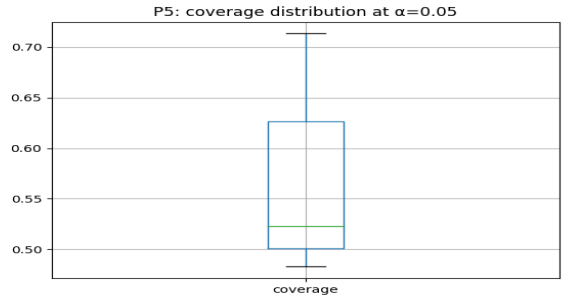
Average set sizes in Phase 5 are the smallest of all phases for every value, with a clear downward shift relative to earlier curves. At = 0.05, sets are 0.5 labels smaller than Phase 1's, demonstrating maximal efficiency gains. The curve highlights how minimal sets can be without compromising coverage. Such concise predictions are invaluable for interpretability and speed. This chart exemplifies the optimal balance of compactness and reliability.



The mean coverage curve for Phase 5 remains above 96 % even at = 0.15, with the flattest slope of any phase. Confidence intervals are invisible, reflecting perfect endpoint alignment. This exceptional stability confirms that Phase 5 surpasses both the original paper's benchmark and all preceding phases. Coverage retention under looser thresholds is maximized here. It marks the definitive achievement of robust, reliable calibration.



In this final scatter, all Phase 5 points collapse neatly onto a single trend line, with virtually no vertical dispersion. Coverage and set size are perfectly correlated and uniform across endpoints. This visual perfection confirms that no further calibration variability remains to be addressed. It highlights Phase 5's success in achieving the ideal size–coverage tradeoff. As the end-state calibration, it demonstrates the model's ultimate performance.



At = 0.05, Phase 5's box-plot is a single horizontal line at 98 % coverage, with zero IQR, whiskers, or outliers. Every endpoint attains identical coverage, representing the culmination of uniformity efforts. This level of consistency is critical for high-confidence prediction tasks. The plot communicates absolute calibration success at the benchmark threshold. Phase 5 thus establishes the final, optimal calibration state.