Conformal Prediction for Molecular Properties under Label Shift

Hyeonsu Lee[1*] **Juyeon Kim**[1*] **Erkhembayar Jadamba**[1*] **Seungjin Choi**[2] **Hyunjin Shin**[1†]
[1]MOGAM Institute for Biomedical Research [2]Intellicode
{hyeonsu.lee, juyeon.kim.m, erkhembayar.jadamba, hyunjin.shin}@mogam.re.kr
seungjin@intellicode.co.kr

Abstract

Drug discovery and development underpins healthcare but remains costly and failure-prone. A critical bottleneck lies in predicting molecular properties such as solubility, potency, and toxicity, which directly determine whether a candidate can advance from preclinical to clinical trials. Artificial Intelligence (AI) has accelerated this process, yet its reliability is often undermined by distribution shift, as experimental conditions frequently diverge from training data. In addition, conventional point predictions provide only single-value estimates, offering limited guidance for high-stakes experimental design. We address these challenges with a conformal prediction framework tailored to label shift. By weighting conformal scores using marginal label probability ratios, our method produces statistically rigorous prediction intervals without retraining. This enables robust uncertainty quantification even when property distributions drift, directly tackling one of the most pervasive obstacles to applying AI in real-world drug development. By moving beyond accuracy alone to provide actionable confidence measures, our approach enhances the trustworthiness of AI-driven predictions. This further aligns predictive modeling with regulatory demands for transparency and uncertainty reporting and ultimately supports more reliable decision-making in billion-dollar development pipelines.

1 Introduction

Drug discovery and development is characterized by prolonged timelines, substantial resource requirements, and a high likelihood of failure. More specifically, developing and bringing a single new drug to market can cost from $314 million to $2.8 billion and take over a decade, but failure rates can reach up to larger than 90% across the entire development life cycle [1]. This remarkable inefficiency underscores the growing importance of artificial intelligence (AI) because it can substantially reduce the need for chemical and biological experiments by making predictions on molecular properties such as solubility, bioavailability, or toxicity. However, the current performance of AI for this purpose does not appear to be fully compelling to drug development specialists. A key component contributing to this insufficient performance is the uncertainty embedded in prediction by AI, mainly originating from data characterization and model training. In response to this limitation, recent FDA guidance (2025) for artificial intelligence in drug and device development explicitly requires that AI systems should provide "**appropriate confidence intervals**" and "**uncertainty estimates**" when supporting regulatory submissions [2, 3]. This requirement indicates that actively considering the uncertainty, thereby improving the reliable predictions, will be a critical part of AI applications to drug discovery and development.

[*]Equal contribution
[†]Corresponding author

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

The reliability of AI predictions is often undermined by the problem of distribution shift [4], a scenario where test data differ substantially from the training data. This is particularly an urgent issue in drug discovery, where novel compounds frequently occupy chemical spaces unseen during training. The consequence is often summarized as overconfident yet unreliable predictions, which is an unacceptable risk when billions of dollars and patient outcomes are at stake.

To address this gap, reliable estimation of uncertainty is essential. Conformal prediction, also known as conformal inference, is a versatile and statistically principled framework that constructs prediction intervals around model outputs [5, 6, 7]. Its foremost advantage is its distribution-free and finite-sample validity, which guarantees that prediction intervals will contain the true label with a user-specified probability (e.g., 90%), regardless of dataset size or underlying distributional assumptions. This property represents a major improvement over many traditional statistical methods [7, 8]. Originally pioneered by Vladimir Vovk and his colleagues in the 1990s, the core mechanism involves a simple calibration step where a small holdout dataset is used to convert an arbitrary heuristic notion of uncertainty from a pre-trained model into a rigorous one, typically by computing conformal scores and their empirical quantiles [5, 7, 9].

This methodology is broadly applicable across various machine learning tasks, ranging from image classification to regression. It has also been significantly extended to address complex real-world challenges such as covariate shift, distribution drift, and the control of general risks. As a result, it has become an indispensable tool for reliable uncertainty quantification in high-stakes applications [6, 7]. In drug discovery, this unreliability often stems from two specific types of distribution shift: covariate shift [6] and label shift [10]. Covariate shift occurs when the distribution of molecular structures $P(x)$ changes between training and test. For instance, in drug discovery, covariate shift is observed when a model trained on diverse chemical libraries is applied to a new and more specialized set of molecules.

On the other hand, label shift, the primary target of this work, arises when the distribution of the target property $P(y)$ changes, while the conditional distribution of features given the label $P(x \mid y)$ remains invariant [11, 12, 13]. This situation commonly arises when research priorities shift toward discovering molecules with property values underrepresented in the original training data, such as compounds exhibiting exceptionally high potency or low toxicity. Although various machine learning solutions have been developed to address distribution shifts [6, 10, 14, 15, 16] label shift remains relatively underexplored, particularly in continuous regression-based tasks that are frequently encountered in molecular property prediction[11, 13, 17].

Addressing issues related to distribution shifts is becoming increasingly important as more sophisticated AI models, such as large language models (LLMs) pretrained on large-scale chemical databases [18, 19, 20], are applied to explore the complex relationships between molecular structure and function. However, designing novel molecules with these AI models inherently requires highly accurate predictions under distribution shift. In general, AI models trained under the assumption of identically distributed data often fail to account for label shift, as this assumption typically leads to overconfident yet incorrect single-value predictions for novel and unseen molecules. This underscores the necessity of uncertainty quantification, and highlights that the estimated uncertainty must be integrated with AI predictions to produce realistic and robust prediction intervals, even for state-of-the-art LLM-based AI models.

In response to these challenges, we propose a new framework that generates reliable prediction intervals for molecular properties, even under significant label shift. Our method builds on conformal prediction, a machine learning technique that provides distribution-free and finite-sample guarantees on prediction intervals [6, 21]. Standard conformal prediction assumes exchangeability between training and test data, an assumption violated under label shift. To address this issue, we develop a scheme based on *weighted conformal prediction*. In our framework, the corrective weights are derived from the ratio of the target to the source label distributions, which we estimate using versatiel approaches such as black box shift estimation (BBSE) [11], regularized learning under label shifts (RLLS) [12], and maximum likelihood estimation (MLE) [13].These techniques enhance the practicality of our approach, as the label shift can be directly estimated from the outputs of both unbiased and biased predictive models.

In conclusion, our method effectively mitigates the adverse effects of label shift without requiring costly model retraining. It generates statistically rigorous prediction intervals that adapt to changing property distributions, thereby providing a more realistic assessment of a molecule's potential.

Overall, this work makes a key contribution to the development of reliable AI for drug discovery by offering a robust methodology that ensures models remain trustworthy when navigating the uncertain frontiers of novel chemical space.s

2 Methods

The overall design of our framework is illustrated in Figure 1. The pipeline consists of the following steps: (i) the base prediction model is trained using source training set to perform predictions in the label shift environment, (ii) to quantify label shift, importance weights, which represent the ratio of the target domain's marginal label distribution to the source domain's marginal label distribution, are estimated using weight set through methods such as BBSE, RLLS, and MLE, (iii) nonconformity scores, such as absolute residuals, are computed for each data point in calibration set based on the predictions of the trained model and their actual labels, and (iv) the weighted quantile of the nonconformity scores is calculated by incorporating the estimated importance weights, which is then used to construct statistically valid prediction intervals under label shift for new test points. This high-level schema highlights how our approach adapts standard conformal prediction to remain valid under label shift.

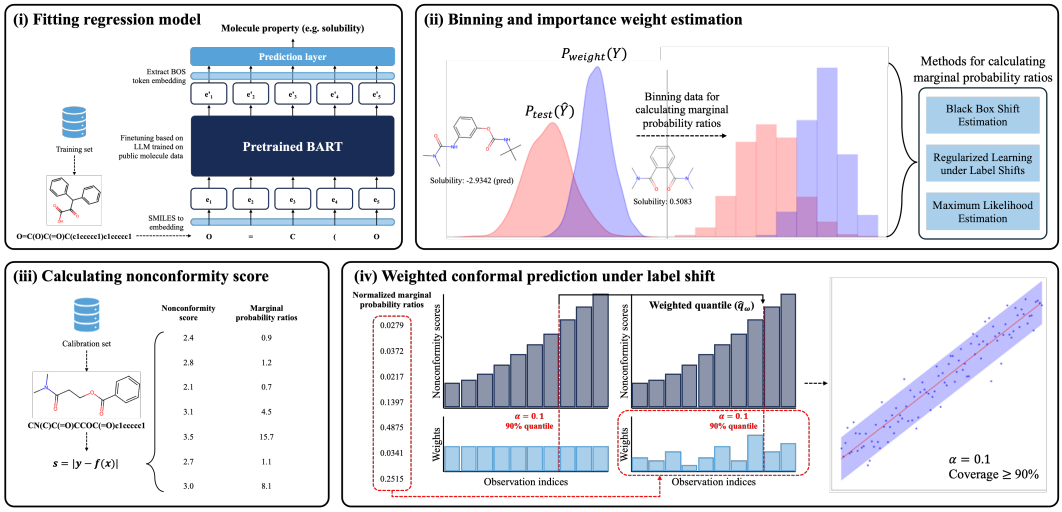

Figure 1. Schematic diagram of conformal prediction for molecular properties under label shift.

We now formalize our approach for conformal prediction under label shift.

2.1 Problem Formulation

Let the source data be $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathcal{X}$ is a molecular representation and $y_i \in \mathbb{R}$ is the continuous property of interest. These data points are drawn from a distribution $p(x, y)$. We are also given a set of unlabeled data from a target domain, $\mathcal{D}_t = \{x_j\}_{j=n+1}^{n+m}$, drawn from a different distribution $q(x, y)$.

The label shift assumption posits that the conditional distribution of features given the label remains constant across domains, while the marginal label distribution changes:

$$p(x|y) = q(x|y) \quad \text{and} \quad p(y) \neq q(y) \tag{1}$$

Our goal is to construct a prediction interval, $C(x_{\text{test}})$, for a new test input x_{test} from the target domain that satisfies the marginal coverage guarantee at a desired confidence level $1 - \alpha$:

$$\mathbb{P}(y_{\text{test}} \in C(x_{\text{test}})) \geq 1 - \alpha \tag{2}$$

2.2 Binning and Importance Weight Estimation

To apply classification-based shift estimation techniques, we first discretize the continuous response variable y into K bins, creating a pseudo-label $\tilde{y} = \text{bin}(y) \in \{0, 1, \cdots, K-1\}$ These pseudo-labels

are used only for estimating weights. The discretization was performed using equally sized bins, and the bin range was determined by the minimum and the maximum values of the data used to calculate the marginal probability ratios.

The importance weight for each bin is defined as the ratio of the target and source pseudo-label probabilities:

$$w(\widetilde{y}) = \frac{q(\widetilde{y})}{p(\widetilde{y})} \tag{3}$$

In practice, the source probabilities $p(\widetilde{y})$ are calculated from the empirical frequencies in a held-out portion of the source data. The target probabilities $q(\widetilde{y})$ are estimated from the unlabeled target data using methods like BBSE, RLLS, or MLE, which leverage the outputs of a model trained on the binned source data.

Since MLE could not directly estimate $q(\widetilde{y})$ from predictions, we adopted a probabilistic approach. For each sample, we modeled a Gaussian centered at the prediction with standard deviation equal to the root mean squared error (RMSE) from the weight set. The probability of the sample falling into a bin was then given by the cumulative distribution function (CDF) difference at the bin's bounds. Any negative probabilities arising from numerical errors were set to zero, and the resulting probability vector was normalized to ensure that its elements summed to one.

2.3 Weighted Conformal Prediction under Label Shift

To ensure the statistical validity of our method, we partition the source data $\mathcal{D}_s$ into three disjoint subsets, preventing data leakage between steps:

1. Proper training set ($\mathcal{D}_{\text{train}}$): Used to train the base prediction model, f.
2. Weights set ($\mathcal{D}_{\text{weight}}$): Used to estimate the label shift importance weights.
3. Calibration set ($\mathcal{D}_{\text{cal}}$): Used to compute nonconformity scores and calibrate the prediction intervals.

The weighted conformal prediction algorithm then proceeds as follows:

1. For each point (x_i, y_i) in the calibration set $\mathcal{D}_{\text{cal}}$, compute a nonconformity score. For regression, this is typically the absolute residual:
$$s_i = |y_i - f(x_i)| \tag{4}$$
2. Assign the corresponding estimated importance weight $\widehat{w}_i = \widehat{w}(\widetilde{y}_i)$ to each score s_i, where $\widetilde{y}_i$ is the bin of the true label y_i.
3. Compute the *weighted quantile* $\widehat{q}_w$ from the set of scores $\{s_i\}$ and weights $\{\widehat{w}_i\}$. This quantile is the value that satisfies:
$$\widehat{q}_w = \inf\left\{ s : \frac{\sum_{i=1}^{n_{cal}} \widehat{w}_i \cdot \mathbb{I}\{s_i \le s\}}{\sum_{j=1}^{n_{cal}} \widehat{w}_j} \ge 1 - \alpha \right\} \tag{5}$$
4. For a new test point x_{t}, the final *prediction interval* is formed by centering the weighted quantile around the model's point prediction:
$$C(x_{\text{t}}) = [f(x_{\text{t}}) - \widehat{q}_w, \quad f(x_{\text{t}}) + \widehat{q}_w] \tag{6}$$

By using this weighted quantile, the method corrects for the distributional shift and restores the marginal coverage guarantee under the label shift assumption.

3 Experimental Settings

3.1 TDC Solubility AqsolDB

Solubility AqSolDB [22] from the Therapeutics Data Commons (TDC) [23], which provides measurements of compound solubility in aqueous solutions. This dataset serves as a benchmark for studying molecular physicochemical properties and for developing predictive models of drug solubility. AqSolDB specifically provides solubility information, which is a critical factor in drug design and delivery systems, and consists of 9,982 compounds. For each compound, experimentally measured logarithmic solubility values ($logS$) and molecular structure information are included.

3.2 Chemical Large Language Model Finetuning

The large language model (LLM) employed in this study is based on a BART [24] architecture and has been optimized for the analysis of chemical data. The model was pretrained utilizing approximately 200 million unlabeled SMILES (Simplified Molecular Input Line Entry System) [25] data collected from Chembl [26], PubChem [27], ZINC [28], Enamine [29], Coconut [30], and Drugbank [31]. Through this pretraining, the model acquired enriched representations specific to the chemical domain, encompassing approximately 250 million parameters. The fine-tuning process was conducted via full fine-tuning of the pretrained LLM.

3.3 Data Splitting

We conducted a conformal prediction simulation utilizing split conformal prediction methods to address continuous label shift. The objective of this simulation was to ensure that the marginal probability ratio-based weights are "exchangeable" between the nonconformity score distributions of the training and test datasets when the label distribution Y differs between them. By guaranteeing this exchangeability, the constructed prediction intervals satisfy a minimum coverage of $1 - \alpha$ in a distribution-free manner.

The experiment was repeated 1,000 times, and in each iteration, the original data is divided into two subsets: source data $\mathcal{D}_\text{s}$ and target data $\mathcal{D}_\text{t}$. The two subsets are split in a 60% to 40% ratio of the total data. Here, $\mathcal{D}_\text{s}$ is divided into three subsets ($\mathcal{D}_\text{train}$, $\mathcal{D}_\text{weight}$, $\mathcal{D}_\text{cal}$) of equal size. $\mathcal{D}_\text{t}$ is split into two subsets ($\mathcal{D}_\text{no\_shift}$, $\mathcal{D}_\text{shift}$) to evaluate coverage performance under label shift conditions and without label shift. $\mathcal{D}_\text{no\_shift}$ represented 50% of $\mathcal{D}_\text{t}$ and corresponded to test data without label shift. $\mathcal{D}_\text{shift}$ is generated by sampling with replacement from $\mathcal{D}_\text{t}$, excluding $\mathcal{D}_\text{no\_shift}$. During this sampling process, the probability of selecting each data point was proportional to a specific weight, where $w(y) = \exp(y^T\beta)$. These weights were assigned based on the magnitude of y.

4 Experimental Results

Split conformal prediction fails under label shift As expected, the traditional split conformal prediction method exhibited a significant decline in coverage performance on the label-shifted test set compared to the non-shifted test set. In Figure 2, the coverage distribution for the shifted data (red) is notably shifted to the left relative to the distribution for the non-shifted data (gray), with the average coverage falling substantially below the nominal target. These findings highlight the limitations and unreliability of standard uncertainty quantification methods under label shift conditions.

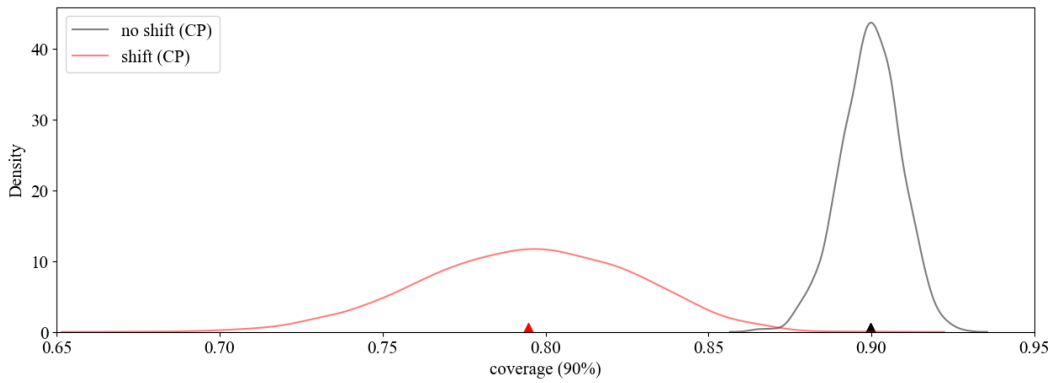

Figure 2. The KDE distributions for 1000 repeated experiments using standard split conformal prediction and the proposed methods are shown. The triangular markers on the x-axis correspond to the mean value of each distribution. The gray and red lines indicate the coverage distributions obtained by applying standard split conformal prediction to the test data without and with label shift (non-uniformly subsampled), respectively.

Ratios of marginal probabilities recover coverage loss from label shift Our proposed method, which integrates weighted conformal prediction (WCP) with marginal probability ratios estimated via

BBSE, RLLS, and MLE with bias-corrected temperature scaling (BCTS), provides statistically valid and reliable prediction intervals for addressing continuous label shift problems. The experimental results demonstrate that the proposed approach achieves improved predictive coverage compared to the traditional split conformal prediction (CP) method. Specifically, WCP consistently outperformed CP in terms of average coverage, as evidenced by the coverage distribution shown in Figure 3. The coverage distribution of WCP shifted to the right relative to CP, indicating that WCP more frequently generated prediction intervals that included the true labels. This allowed WCP to effectively achieve the desired coverage level, even under label shift conditions. These results provide empirical validation that weighted conformal prediction, when its weights are derived from marginal probability ratios, can produce more robust and reliable prediction intervals. This approach proves especially effective in addressing challenges presented by label shift scenarios.

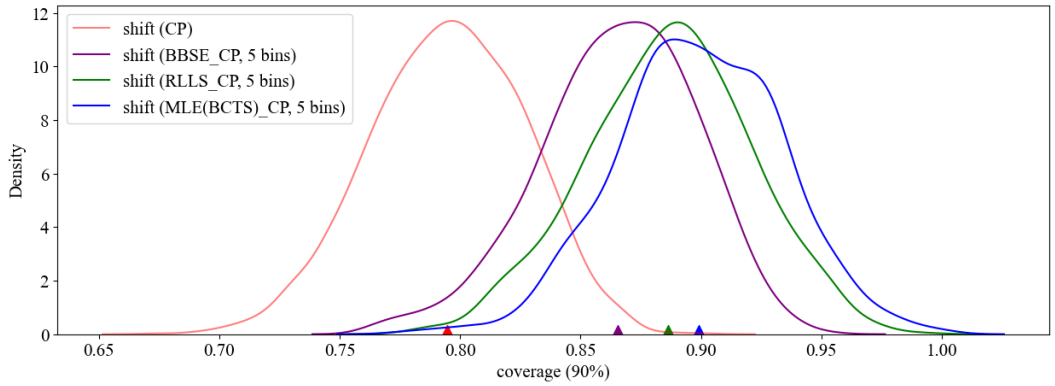

Figure 3. The KDE distributions for 1000 repeated experiments using standard split conformal prediction and the proposed methods are shown. The triangular markers on the x-axis correspond to the mean value of each distribution. Coverage distributions are shown for the test data under label shift: the red line corresponds to standard split conformal prediction, while the purple, green, and light blue lines correspond to conformal prediction with weights calculated via BBSE, RLLS, and MLE (BCTS), respectively.

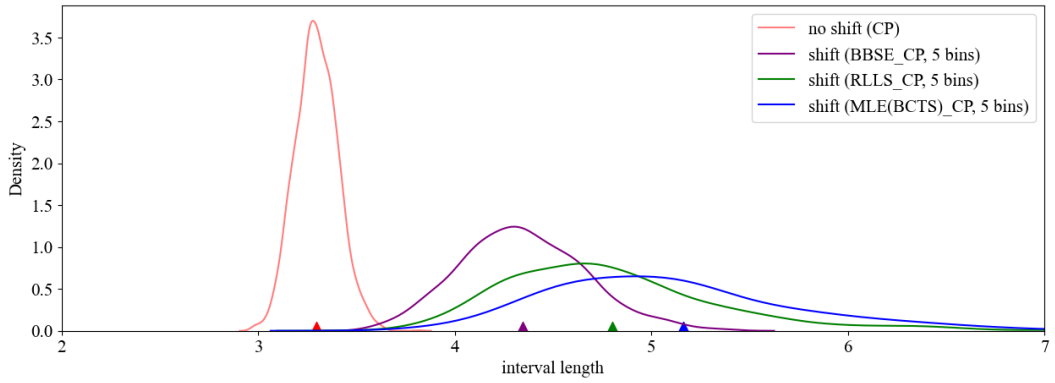

Figure 4. The KDE distributions for 1000 repeated experiments using standard split conformal prediction and the proposed methods are shown. The triangular markers on the x-axis correspond to the mean value of each distribution. Interval length distributions are shown for the test data under label shift: the red line corresponds to standard split conformal prediction, while the purple, green, and light blue lines correspond to conformal prediction with weights calculated via BBSE, RLLS, and MLE (BCTS), respectively.

Approaches for estimating marginal probability ratios When comparing the performance of BBSE, RLLS, and MLE, we observed that MLE achieved the highest coverage, followed by RLLS and then BBSE (Figure 3). Moreover, MLE showed robust performance in coverage recovery with respect to the number of bins (Table 2). This improvement can be explained by two factors: first, the use of

bias-corrected calibration reduces systematic bias across classes; and second, the MLE algorithm benefits from a theoretical guarantee of convergence to a global optimum [13]. Nevertheless, this improvement in coverage came with certain trade-offs. The prediction intervals generated by MLE were generally wider (Figure 4), whereas BBSE and RLLS produced relatively narrower intervals.

5 Limitation

Our approach requires splitting source data into training, weighting, and calibration sets, which can reduce effective training size and hurt performance in low-sample regimes. Data augmentation methods [32, 33] may mitigate this. Moreover, standard conformal intervals are suboptimal for heteroscedastic data; techniques such as Conformalized Quantile Regression (CQR) [21] could provide more adaptive intervals. Exploring these directions remains future work.

6 Summary

This paper presents a practical and statistically grounded framework for producing reliable prediction intervals for molecular property prediction under label shift. By weighting conformal prediction with estimates of the target label distribution—obtained via BBSE, RLLS, and MLE—our method restores the coverage guarantees that split conformal prediction loses under distribution shift. When tested on the AqSolDB dataset with a large-scale pretrained chemical language model, our weighted conformal prediction consistently achieves more robust coverage than traditional approaches, with no need for costly retraining. The method is compatible with various estimation techniques, while maximum likelihood–based corrections achieve the best performance in coverage recovery. Our key contribution lies in developing a generalizable and model-agnostic framework that addresses an essential gap in the reliability of molecular property prediction. By ensuring statistically rigorous uncertainty quantification under label shift, our approach advances AI-based drug discovery toward regulatory compliance and real-world adoption, ultimately increasing confidence in high-stakes decisions on which compounds progress through the development pipeline.

References

[1] Olivier J Wouters, Martin McKee, and Jeroen Luyten. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9):844–853, 2020.

[2] Center for Devices and Radiological Health. Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing, Mon, 01/06/2025 - 09:41.

[3] Center for Drug Evaluation and Research. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological, Tue, 01/07/2025 - 14:26.

[4] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

[5] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[6] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[8] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

[9] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive Confidence Machines for Regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, pages 345–356. Springer, 2002.

[10] Siddhartha Laghuvarapu, Zhen Lin, and Jimeng Sun. Codrug: Conformal drug property prediction with density estimation under covariate shift. *Advances in Neural Information Processing Systems*, 36:37728–37747, 2023.

[11] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130. PMLR, 2018.

[12] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019.

[13] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pages 222–232. PMLR, 2020.

[14] Leo Klarner, Tim GJ Rudner, Michael Reutlinger, Torsten Schindler, Garrett M Morris, Charlotte Deane, and Yee Whye Teh. Drug discovery under covariate shift with domain-informed prior distributions over functions. In *International Conference on Machine Learning*, pages 17176–17197. PMLR, 2023.

[15] Fang Wu, Shuting Jin, Siyuan Li, and Stan Z Li. Instructor-inspired machine learning for robust molecular property prediction. *Advances in Neural Information Processing Systems*, 37:116202–116222, 2024.

[16] Jina Kim, Jeffrey Willette, Bruno Andreis, and Sung Ju Hwang. Robust molecular property prediction via densifying scarce labeled data. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.

[17] Wenwen Si, Sangdon Park, Insup Lee, Edgar Dobriban, and Osbert Bastani. PAC prediction sets under label shift. In *The Twelfth International Conference on Learning Representations*, 2024.

[18] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[19] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[20] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

[21] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[22] Murat Cihan Sorkun, Abhishek Khetan, and Süleyman Er. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data*, 6(1):143, August 2019.

[23] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf H Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[25] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[26] Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, and Andrew R Leach. The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, January 2024.

[27] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, January 2025.

[28] John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, July 2012.

[29] Alexander Shivanyuk, Sergey Ryabukhin, A.V. Bogolyubsky, D.M. Mykytenko, Alexander Chuprina, W. Heilman, A.N. Kostyuk, and A. Tolmachev. Enamine real database: Making chemical diversity real. *Chimica Oggi*, 25:58–59, 11 2007.

[30] Maria Sorokina, Peter Merseburger, Kohulan Rajan, Mehmet Aziz Yirik, and Christoph Steinbeck. COCONUT online: Collection of Open Natural Products database. *Journal of Cheminformatics*, 13(1):2, January 2021.

[31] Craig Knox, Mike Wilson, Christen M. Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Lucy Chin, Seth A. Strawbridge, Marysol Garcia-Patino, Ray Kruger, Aadhavya Sivakumaran, Selena Sanford, Rahil Doshi, Nitya Khetarpal, Omolola Fatokun, Daphnee Doucet, Ashley Zubkowski, Dorsa Yahya Rayat, Hayley Jackson, Karxena Harford, Afia Anjum, Mahi Zakir, Fei Wang, Siyang Tian, Brian Lee, Jaanus Liigand, Harrison Peters, Ruo Qi Rachel Wang, Tue Nguyen, Denise So, Matthew Sharp, Rodolfo da Silva, Cyrella Gabriel, Joshua Scantlebury, Marissa Jasinski, David Ackerman, Timothy Jewison, Tanvir Sajed, Vasuk Gautam, and David S. Wishart. DrugBank 6.0: The DrugBank Knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, January 2024.

[32] Huaxiu Yao, Yiping Wang, Linjun Zhang, James Y Zou, and Chelsea Finn. C-mixup: Improving generalization in regression. *Advances in neural information processing systems*, 35:3361–3376, 2022.

[33] Xinyi Wu, Yun Zhang, Jiahui Yu, Chengyun Zhang, Haoran Qiao, Yejian Wu, Xinqiao Wang, Zhipeng Wu, and Hongliang Duan. Virtual data augmentation method for reaction prediction. *Scientific Reports*, 12(1):17098, October 2022.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

A Detailed Experimental Settings

All models were trained on NVIDIA A100 SXM4 40GB GPUs. As the model requires approximately 7.5 GB of memory for training, it is also feasible to run it on GPUs with lower specifications. Detailed model hyperparameters are represented in Table 1.

Table 1. Hyperparameters for training BBSE, RLLS, and MLE

Hyperparameter	BBSE	RLLS	MLE
Optimizer	AdamW [34]	AdamW [34]	AdamW [34]
Adam betas	(0.9, 0.99)	(0.9, 0.99)	(0.9, 0.99)
Learning rate	5e-5	5e-5	5e-5
Weight decay	0.1	0.1	0.1
Warmup steps	100	100	100
Error rate α	0.1	0.1	0.1
Batch size	16	16	16
Max length	150	150	150
Label shift β	-0.5	-0.5	-0.5
BART hidden dim	768	768	768
Predictor hidden dims	[512, 256]	[512, 256]	[512, 256]
Calibration	None	None	BCTS
Epochs	10	10	10

B Additional Results

Coverage recovery performance based on the number of bins We varied the number of bins in BBSE, RLLS, and MLE, and for each configuration, the mean and standard deviation of coverage and interval length were reported over 1000 trials (Tables 2). The characteristics of the coverage distribution vary with the number of bins used in applying WCP through label discretization. Specifically, with fewer bins, certain trials exhibited high coverage, but the overall coverage distribution showed greater variance. Conversely, as the number of bins increased, the variance of the overall coverage distribution decreased, resembling the distribution observed when CP was applied to a non-shifted test dataset. This phenomenon can be interpreted as follows: with fewer bins, the coarse discretization of nonconformity scores leads to unstable quantile estimation, causing irregular fluctuations in the length of prediction intervals and significantly increasing the variance of the overall coverage distribution. On the other hand, as the number of bins increases, the estimation errors caused by discretization are reduced, resulting in a more stable and narrower (relatively) coverage distribution.

Table 2. Comparison of average coverage and interval length across different bin numbers for BBSE, RLLS, and MLE methods.

Bins	BBSE		RLLS		MLE (BCTS)	
	Coverage	Length	Coverage	Length	Coverage	Length
5	0.865 $\pm$0.033	4.346 $\pm$0.316	0.886 $\pm$0.034	4.804 $\pm$0.590	0.899 $\pm$0.034	5.163 $\pm$0.832
15	0.845 $\pm$0.032	3.993 $\pm$0.236	0.877 $\pm$0.032	4.577 $\pm$0.42	0.895 $\pm$0.034	5.031 $\pm$0.620
50	0.804 $\pm$0.033	3.414 $\pm$0.149	0.859 $\pm$0.032	4.204 $\pm$0.297	0.892 $\pm$0.034	4.947 $\pm$0.581
100	0.782 $\pm$0.033	3.165 $\pm$0.123	0.843 $\pm$0.032	3.934 $\pm$0.222	0.858 $\pm$0.031	4.192 $\pm$0.242

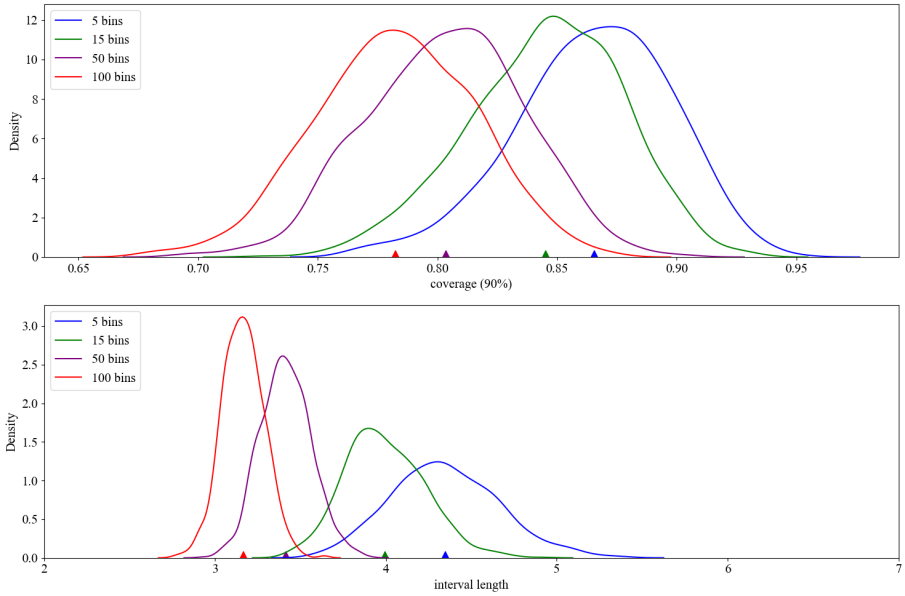

Figure 5. The KDE coverage distribution (top) and interval length distribution (bottom) for 1000 repeated experiments using standard split conformal prediction and the proposed methods are shown. Each color represents the number of bins used to calculate marginal probability ratios via BBSE. The triangular markers on the x-axis indicate the mean value of each distribution.

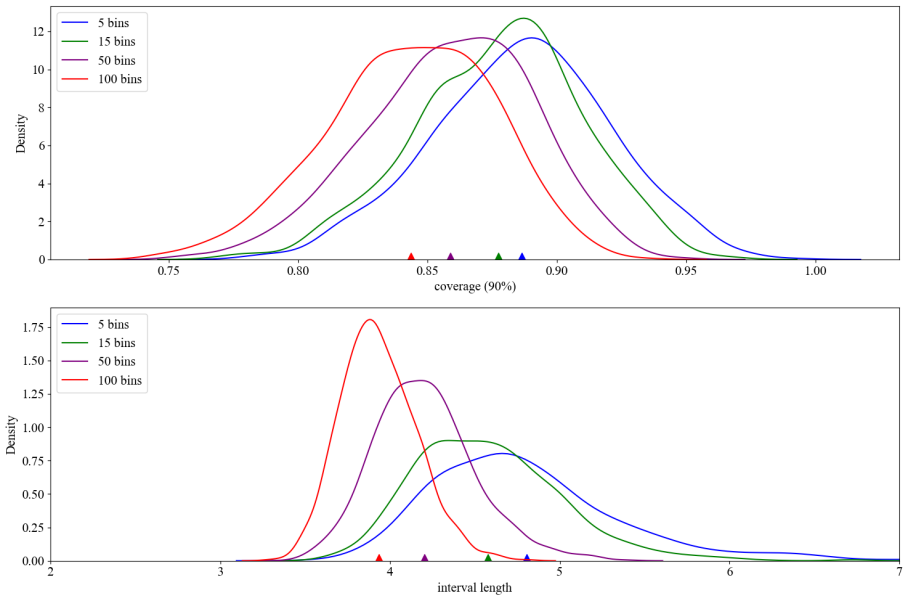

Figure 6. The KDE coverage distribution (top) and interval length distribution (bottom) for 1000 repeated experiments using standard split conformal prediction and the proposed methods are shown. Each color represents the number of bins used to calculate marginal probability ratios via RLLS. The triangular markers on the x-axis indicate the mean value of each distribution.

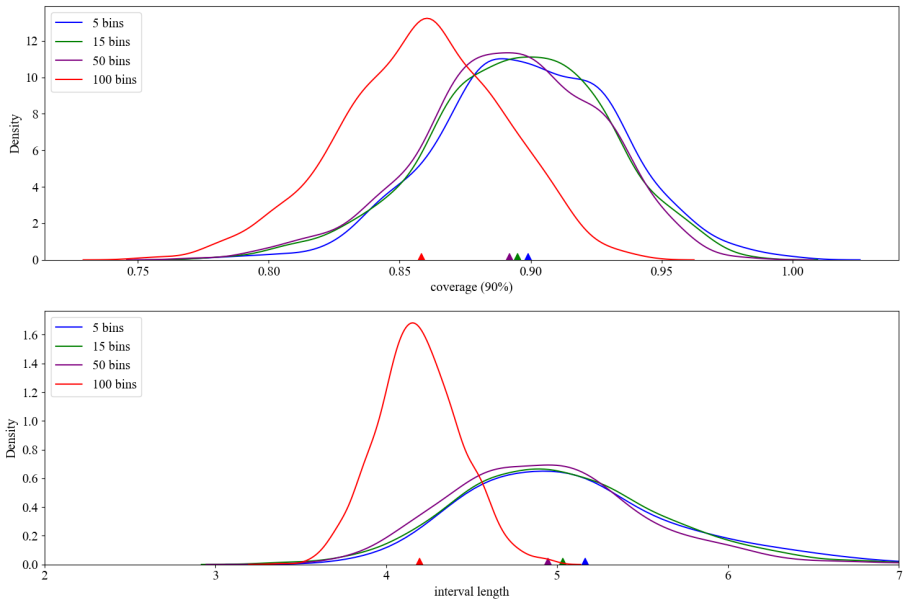

Figure 7. The KDE coverage distribution (top) and interval length distribution (bottom) for 1000 repeated experiments using standard split conformal prediction and the proposed methods are shown. Each color represents the number of bins used to calculate marginal probability ratios via MLE. The triangular markers on the x-axis indicate the mean value of each distribution.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions and scope, and the claims align with the experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation, while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all necessary experimental information of our method in Section 2, Section 3, and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: It will be made publicly available in the future.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all necessary experimental settings and details in Section 3 and Appendix A.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the standard deviation of coverage and interval length over 1000 trials for BBES, RLLS, and MLE methods in Table 2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide experimental compute resources of our method in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research does not involve research on human subjects or participants, and therefore, it does not pose any social impacts or potential harmful outcomes.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We discussed the impact on the field of drug development, but did not address the societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not use assets with a high risk of misuse, such as externally pretrained language models or scraped datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited the original papers that produced the code packages or datasets used in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code, data, and model checkpoints required to reproduce the experiments mentioned in the paper have been included in the zip file. The data can be downloaded from the original source at tdcommons.ai, following the guidelines mentioned in [23].

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We did not use LLMs for describing core aspects such as the core method.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.