Combining bayesian and evidential uncertainty quantification for improved bioactivity modeling

Khalil, B.A.A.; Schweighofer, K.; Dyubankova, N.; Westen, G.J.P. van; Vlijmen, H. van

Citation
Khalil, B. A. A., Schweighofer, K., Dyubankova, N., Westen, G. J. P. van, & Vlijmen, H. van. (2025). Combining bayesian and evidential uncertainty quantification for improved bioactivity modeling. *Journal Of Chemical Information And Modeling*. doi:10.1021/acs.jcim.5c01597

Version:	Publisher's Version
License:	Creative Commons CC BY-NC-ND 4.0 license
Downloaded from:	https://hdl.handle.net/1887/4285672

Note: To cite this publication please use the final published version (if applicable).

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Open Access
This article is licensed under CC-BY-NC-ND 4.0

pubs.acs.org/jcim
Article

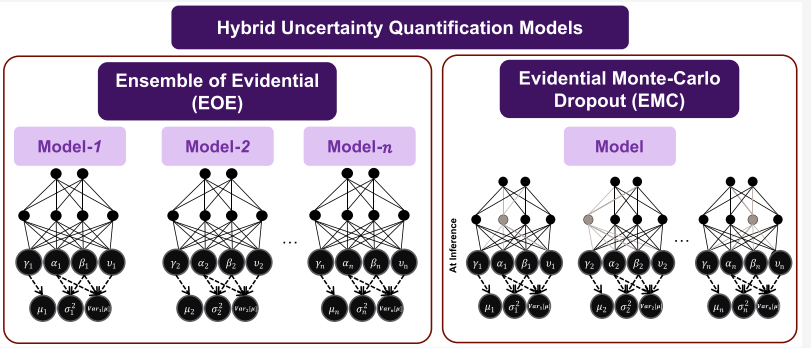
Combining Bayesian and Evidential Uncertainty Quantification for Improved Bioactivity Modeling

Bola Khalil, Kajetan Schweighofer, Natalia Dyubankova, Gerard J. P. van Westen,* and Herman van Vlijmen*

Cite This: https://doi.org/10.1021/acs.jcim.5c01597
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

Downloaded via LEIDEN UNIV on December 11, 2025 at 13:28:00 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.



ABSTRACT: Uncertainty quantification (UQ) has been recognized as a prerequisite for reliable and trustworthy computational modeling in drug discovery. Two widely considered paradigms, Bayesian methods (deep ensemble and MC dropout) and evidential learning, differ in their computational demands and expressivity of uncertainties, excelling in complementary settings. Here, we propose hybrid approaches that combine both paradigms and benchmark them on the Papyrus++ data set across two end points (xC_{50}, K_x) and multiple split strategies. Our ensemble of evidential models (EOE) consistently achieves the best overall performance, yielding the lowest RMSE and leading CRPS and interval scores, including under the most challenging distributional shifts. While large ensembles often excel in rejection-based utility, EOE matches or surpasses them at a fraction of the computational cost. Statistical tests confirm its advantage, and a hardware-agnostic compute analysis highlights favorable performance-efficiency trade-offs. These results demonstrate that combining evidential and Bayesian principles yields more accurate and informative uncertainties for bioactivity modeling, with **EOE** offering a robust—and computationally practical—default for uncertainty-aware decision-making in drug discovery.

INTRODUCTION

Machine learning (ML) has become integral to computational sciences, providing powerful tools for data-driven discovery across domains such as bioinformatics and cheminformatics.[1] In pharmaceutical research, ML models are widely applied to assess molecular interactions, screen compounds, and assist in the identification of candidates for drug development. While these models demonstrate high accuracy, their application in real-world settings demands an understanding of confidence in model outputs. In high-stakes decision-making, such as early-stage drug discovery, reliance on uncertain or poorly calibrated models can lead to costly experimental failures, resource misallocation, and erroneous conclusions. As a result, UQ has emerged as an important component of ML for scientific applications.[2−5]

Uncertainty in ML models falls into two principal categories: *aleatoric* and *epistemic* uncertainty.[6] Aleatoric uncertainty arises from inherent stochasticity, such as measurement noise or biological heterogeneity, and cannot be reduced by acquiring more data. In contrast, epistemic uncertainty reflects a lack of knowledge due to insufficient training data or structural biases and can be reduced by improving data coverage or model design. The ability to assess these uncertainties can determine the reliability of ML-driven assessments, guiding experimental validation efforts and optimizing resource allocation. This challenge is exacerbated by variability in public bioactivity data, where combining measurements from different assays can introduce substantial noise;[7] while Papyrus++[8] applies rigorous standardization, residual assay variability may still impact model reliability.

Received: July 12, 2025
Revised: November 4, 2025
Accepted: November 26, 2025

ACS Publications
© XXXX The Authors. Published by American Chemical Society

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

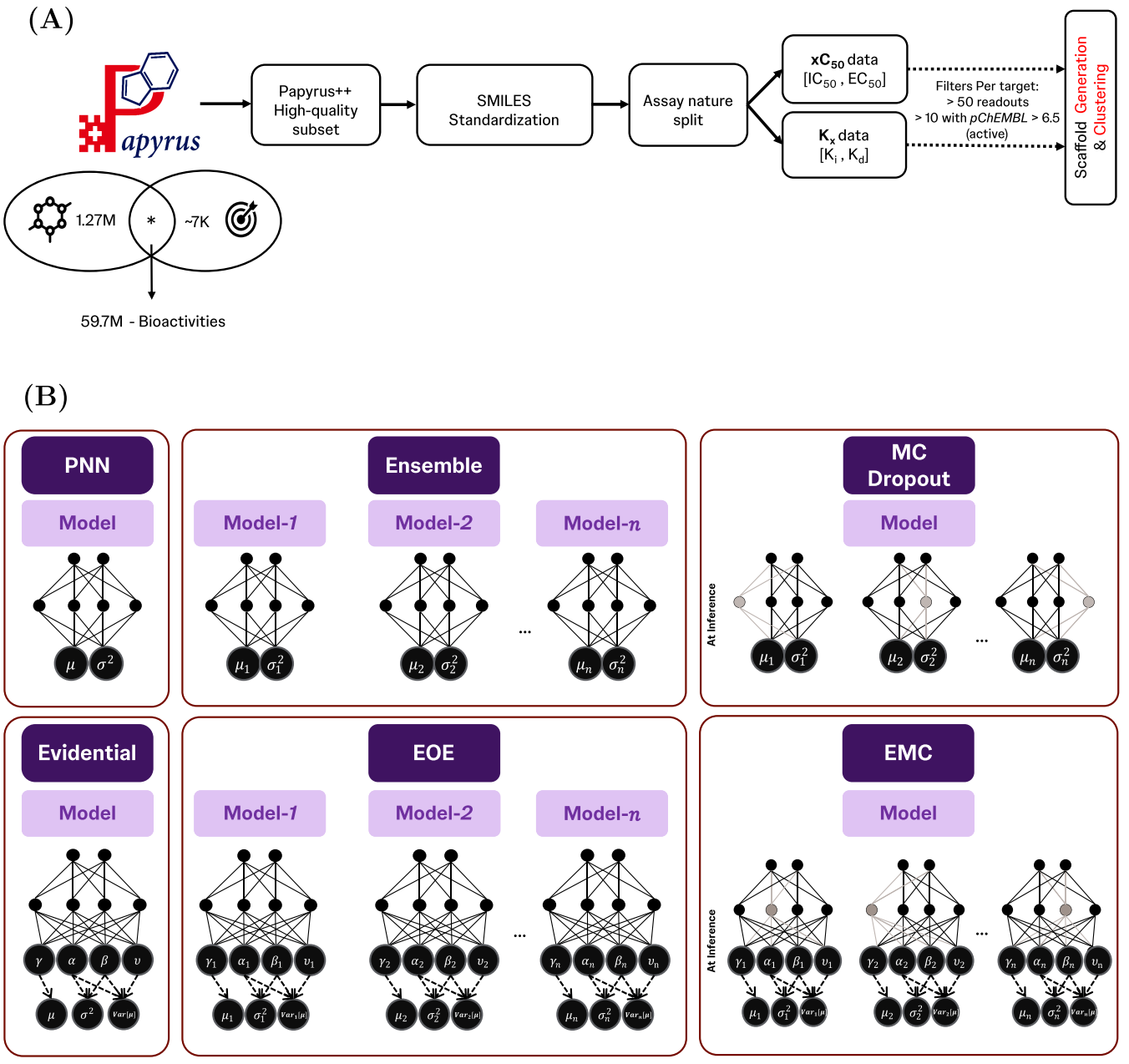
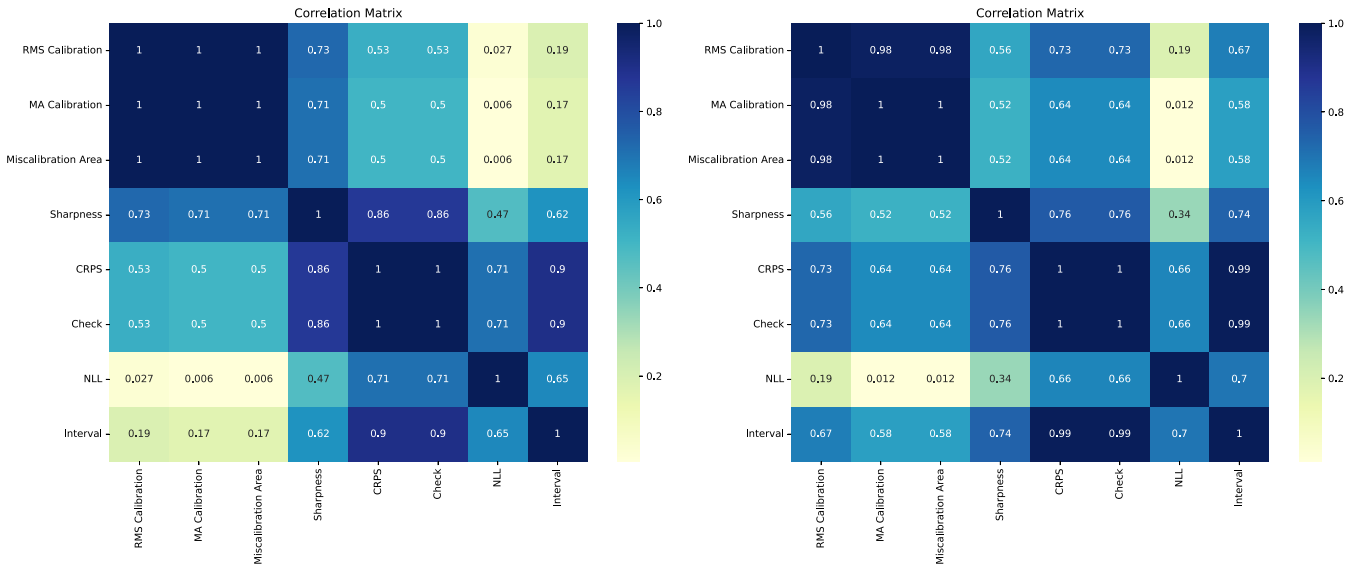


Figure 1. Overview of (A) data preparation workflow: Starting with the Papyrus++ data set, SMILES standardization and assay-based splitting yield xC_{50} and K_x subsets. Filtering ensures a minimum of 50 readouts per target and excludes under-represented targets, with at least 10 active data points, i.e., pChEMBL values >6.5. (B) UQ Approaches: Ensemble models aggregate outputs from multiple PNN models. MC dropout estimates uncertainty through stochastic forward passes. Evidential models output distribution parameters (α, β, γ, and υ) to derive mean, aleatoric, and epistemic uncertainty. Hybrid approaches (EOE and EMC) integrate ensemble diversity or MC dropout stochasticity with evidential uncertainty estimation. Papyrus logo is reproduced with permission from ref Béquignon et.al.,[8] Copyright 2021 Springer Nature.

Various UQ methodologies have been developed to address these challenges.[9−11] For modeling aleatoric uncertainty, **Probabilistic Neural Networks** (PNNs)[12] parametrize a Gaussian distribution over outputs, enabling the estimation of input-dependent variance. Other approaches include **deep ensembles**, which estimate epistemic uncertainty by aggregating outputs from multiple independently trained models,[13] and **Monte Carlo (MC) dropout**, which approximates Bayesian inference by randomly masking network weights during inference.[14] These approaches approximate Bayesian modeling of uncertainty by sampling different model parametrizations. While deep ensembles are widely regarded as effective, they incur significant computational costs. However, they have been shown to be very faithful to the Bayesian ideal in deep learning settings.[15,16] MC dropout provides a more computationally efficient alternative but may yield inconsistent uncertainty estimates depending on dropout rates and the number of stochastic passes.

A more recent alternative, **evidential deep learning**, reformulates uncertainty estimation by parametrizing a second-order distribution over possible distributions.[17−19] Unlike ensemble-based or sampling-based methods, evidential learning estimates aleatoric and epistemic uncertainty in a single deterministic forward pass, making it computationally attractive. However, empirical studies indicate that it may

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX



(a) xC_{50} Dataset (b) K_x Dataset

Figure 2. Correlation matrix of uncertainty metrics for models evaluated on the (a) xC_{50} and (b) K_x data sets. The figure presents the correlation structure among uncertainty metrics from the Uncertainty Toolbox.[26] A strong correlation is observed among calibration-related metrics as well as between CRPS and Check metrics. To ensure a nonredundant selection of uncertainty metrics, we retain five representative metrics: miscalibration area, sharpness, CRPS, NLL, and interval, which capture distinct aspects of the model uncertainty.

struggle in extrapolative settings where data sparsity limits generalization.[20]

To address these limitations, we introduce two novel *hybrid* UQ architectures that integrate evidential deep learning with ensemble and dropout-based techniques: (1) an **ensemble of evidential models (EOE)**, which aggregates multiple evidential networks to improve epistemic uncertainty estimation, and (2) an **evidential MC dropout model (EMC)**, which combines evidential learning with stochastic dropout sampling to capture variance in model outputs. These hybrid approaches aim to enhance the robustness of uncertainty estimation while balancing computational efficiency. Figure 1B summarizes the UQ models considered in this study, including Bayesian, evidential, and hybrid approaches.

Previous work has explored combining evidential and ensemble models. Shen et al.[21] trained an evidential model on outputs from physiological and regression models, while Shao et al.[22] used a multimodal ensemble of evidential models aggregated via Dempster-Shafer rules. In contrast, we perform Bayesian estimation of evidential model parameters, yielding a unified Bayesian-evidential hybrid approach.

We evaluate these UQ methodologies in the context of proteochemometric modeling (PCM), a framework that jointly models interactions between ligands and protein targets.[23−25] PCM provides a structured approach to bioactivity assessment by integrating information from both ligand and protein domains, offering a broader context for ML-based assessments. By systematically comparing deep ensembles, MC dropout, evidential deep learning, and our proposed hybrid models, we assess their ability to estimate uncertainty, improve calibration, and provide well-calibrated confidence measures in bioactivity assessment. In addition to accuracy and calibration, we assess the statistical significance of model differences and quantify training cost in hardware-agnostic GPU-hours to provide a rigorous and practically relevant comparison. The broader significance of this study lies in its potential to refine ML methodologies for drug discovery in the chemical sciences. By identifying limitations in current UQ techniques and proposing hybrid approaches, we provide a framework for enhancing the reliability of ML-driven decision-making in cheminformatics. These findings contribute to ongoing efforts in uncertainty-aware modeling, helping to establish more robust and interpretable ML frameworks for real-world scientific applications.

RESULTS

We evaluate the performance and uncertainty quantification capabilities of all considered models across two bioactivity end points; xC_{50} (e.g., IC_{50}, EC_{50}) and K_x (e.g., K_i, K_d), using both stratified and scaffold-cluster data splits. Root Mean Squared Error (RMSE) is used as the primary performance metric, with other values, e.g., Coefficient of Determination (R^2) and Pearson Correlation Coefficient (PCC), provided in the Supporting Information.

To assess uncertainty quality, we adopt five nonredundant metrics from the Uncertainty Toolbox[26]—miscalibration area, negative log-likelihood (NLL), continuous ranked probability score (CRPS), interval score, and sharpness. These were selected based on their low redundancy (Figure 2) and collectively reflect calibration, probabilistic accuracy, and dispersion of uncertainty estimates. To evaluate the practical utility of uncertainty estimates for selective outputs, we also use the RMSE-rejection curve area under the curve (RRC-AUC). All metrics are computed on the aleatoric uncertainty component, except RRC-AUC, which utilizes the total uncertainty. Detailed information on all evaluation metrics can be found in the Supporting Information.

Table 1 summarizes the quantitative results, and a complementary visual summary is provided in Figure 3. Detailed results can be found in the Supporting Information. Across all experimental conditions, EOE_{10} consistently achieves the lowest RMSE, despite using only 10 ensemble

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Performance, Uncertainty, and Ranking Metrics for Evaluated Models on xC_{50} and K_x Data Sets across Stratified and Scaffold Cluster Splits[a]

Model	Activity	Performance Metrics	Uncertainty Metrics					Rank Metrics
		RMSE ↓	Miscalib. Area ↓	NLL ↓	CRPS ↓	Interval ↓	Sharpness ↓	RRC-AUC ↓
				Stratified				
PNN_1	xC_{50}	0.732 (0.004)	0.133 (0.009)	1.478 (0.063)	0.409 (0.003)	2.264 (0.031)	0.465 (0.012)	0.986 (0.003)
$Ensemble_{100}$	xC_{50}	0.728 (0.003)	0.129 (0.011)	1.441 (0.065)	0.406 (0.000)	2.234 (0.030)	0.469 (0.021)	**.825 (0.009)**
MC $Dropout_{100}$	xC_{50}	0.763 (0.008)	**.026 (0.016)**	1.140 (0.014)	0.418 (0.004)	2.111 (0.015)	0.669 (0.052)	0.920 (0.004)
$Evidential_1$	xC_{50}	0.682 (0.010)	0.103 (0.035)	1.005 (0.015)	.378 (0.009)	2.247 (0.121)	.414 (0.053)	0.907 (0.011)
EOE_{10}	xC_{50}	**.646 (0.002)**	.074 (0.008)	**.943 (0.003)**	**.352 (0.002)**	**2.028 (0.029)**	0.422 (0.012)	.900 (0.006)
EMC_{10}	xC_{50}	0.693 (0.007)	0.125 (0.034)	1.033 (0.011)	0.389 (0.008)	2.335 (0.125)	**.402 (0.052)**	0.909 (0.009)
Random Rejection	xC_{50}							0.945 (0.008)
PNN_1	K_x	0.639 (0.011)	0.127 (0.005)	1.757 (0.030)	0.341 (0.013)	2.059 (0.084)	0.699 (0.383)	0.982 (0.011)
$Ensemble_{100}$	K_x	0.595 (0.001)	0.079 (0.001)	1.261 (0.010)	.311 (0.001)	1.809 (0.005)	.655 (0.037)	**.607 (0.001)**
MC $Dropout_{100}$	K_x	0.641 (0.004)	0.063 (0.011)	1.230 (0.083)	0.330 (0.001)	1.874 (0.025)	**.439 (0.022)**	0.679 (0.005)
$Evidential_1$	K_x	0.624 (0.003)	**.051 (0.001)**	.531 (0.005)	0.313 (0.001)	1.754 (0.006)	0.787 (0.005)	0.615 (0.003)
EOE_{10}	K_x	**.573 (0.001)**	.054 (0.001)	**.428 (0.002)**	**.291 (0.000)**	**1.673 (0.005)**	0.749 (0.002)	.613 (0.001)
EMC_{10}	K_x	0.626 (0.003)	0.086 (0.004)	0.560 (0.005)	0.317 (0.002)	1.842 (0.011)	0.725 (0.004)	0.620 (0.002)
Random Rejection	K_x							0.949 (0.005)
				Scaffold Cluster				
PNN_1	xC_{50}	0.907 (0.006)	0.049 (0.010)	1.403 (0.024)	0.509 (0.004)	2.608 (0.030)	.777 (0.018)	1.008 (0.003)
$Ensemble_{100}$	xC_{50}	0.878 (0.001)	**.031 (0.008)**	**1.320 (0.009)**	.490 (0.000)	2.474 (0.009)	0.795 (0.023)	**.836 (0.006)**
MC $Dropout_{100}$	xC_{50}	0.928 (0.006)	0.085 (0.008)	1.522 (0.031)	0.525 (0.004)	2.735 (0.030)	**.738 (0.017)**	0.938 (0.004)
$Evidential_1$	xC_{50}	0.897 (0.003)	0.049 (0.005)	1.360 (0.004)	0.502 (0.002)	2.514 (0.006)	0.992 (0.013)	0.906 (0.005)
EOE_{10}	xC_{50}	**.871 (0.001)**	0.062 (0.001)	1.323 (0.001)	**.488 (0.001)**	**2.455 (0.005)**	1.000 (0.004)	.900 (0.003)
EMC_{10}	xC_{50}	0.909 (0.003)	.042 (0.008)	1.383 (0.004)	0.508 (0.001)	2.545 (0.007)	0.989 (0.022)	0.909 (0.006)
Random Rejection	xC_{50}							0.948 (0.006)
PNN_1	K_x	1.102 (0.012)	**.034 (0.006)**	1.531 (0.014)	0.614 (0.006)	3.060 (0.034)	1.059 (0.022)	0.982 (0.001)
$Ensemble_{100}$	K_x	1.097 (0.002)	0.038 (0.001)	1.527 (0.002)	0.611 (0.001)	3.045 (0.005)	1.038 (0.004)	0.842 (0.005)
MC $Dropout_{100}$	K_x	1.101 (0.003)	0.044 (0.010)	1.596 (0.022)	0.620 (0.002)	3.092 (0.017)	1.136 (0.031)	0.851 (0.004)
$Evidential_1$	K_x	1.061 (0.016)	.035 (0.009)	1.539 (0.028)	.589 (0.010)	3.072 (0.041)	0.939 (0.018)	0.844 (0.008)
EOE_{10}	K_x	**.998 (0.007)**	0.047 (0.004)	**1.464 (0.009)**	**.553 (0.004)**	**2.933 (0.017)**	**.889 (0.006)**	**.831 (0.002)**
EMC_{10}	K_x	1.064 (0.023)	0.056 (0.009)	1.555 (0.027)	0.591 (0.013)	3.133 (0.065)	.904 (0.011)	.837 (0.007)
Random Rejection	K_x							0.950 (0.003)

[a]Comparative analysis of different uncertainty quantification models, reporting one performance metric; RMSE, five uncertainty quantification metrics; miscalibration area, NLL, CRPS, interval score, and sharpness, and one ranking-based metric; RRC-AUC. Lower values indicate better outcomes for all listed metrics. The best-performing model for each metric is highlighted in bold, while the second-best is underlined. Random rejection is included as a baseline reference for ranking performance. The results show that EOE_{10} achieves consistently competitive performance, excelling across most uncertainty metrics and demonstrating strong overall accuracy, with significance confirmed across most metrics, highlighting the benefit of combining evidential and Bayesian approaches for bioactivity modeling tasks.

members—significantly fewer than $Ensemble_{100}$, which comprises 100 models. The choice of 10 members for the EOE strikes a deliberate balance between computational efficiency and diversity among ensemble members. This configuration was empirically determined to provide sufficient variance reduction while keeping training and inference costs manageable, particularly given the added complexity of the evidential parameter estimation. Notably, EOE_{10} also performs best on CRPS and interval score across all splits and tasks, underscoring its effective trade-off between accuracy and uncertainty expressiveness.

Stratified Splits. In the more familiar stratified data splits, models were generally able to learn well-calibrated estimations.

xC_{50} End Point. EOE_{10} shows strong performance across uncertainty metrics, obtaining the best NLL, CRPS, and interval score, while being the second-best method for miscalibration area and RRC-AUC. EMC_{10} has the narrowest estimated distribution with the lowest Sharpness. $Ensemble_{100}$ outperforms all other methods in the RRC-AUC metric, reflecting a strong practical ranking utility.

K_x End Point. Similarly, EOE_{10} leads in NLL, CRPS, and interval score and is again second-best in the miscalibration area and RRC-AUC. MC $Dropout_{100}$ achieves the sharpest uncertainty estimates, while $Ensemble_{100}$ ranks highest in ranking utility, as measured by RRC-AUC. EMC_{10} shows robust sharpness and competitive probabilistic calibration.

Scaffold Cluster Splits. This setting represents a more challenging scenario with a substantial distributional shift between training and test scaffolds.

xC_{50} End Point. $Ensemble_{100}$ achieves the lowest miscalibration area, NLL, and best RRC-AUC, and also performs second-best on CRPS and interval score. EOE_{10} obtains the best CRPS and interval score, and is second-best on NLL and RRC-AUC. EMC_{10} achieves the lowest miscalibration, suggesting potential underconfidence. MC $Dropout_{100}$ again produces the sharpest distributions.

K_x End Point. This setting represents one of the most challenging evaluation scenarios, as evidenced by the higher RMSE values observed across all models. The increased difficulty stems from the combination of the K_x end point and

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

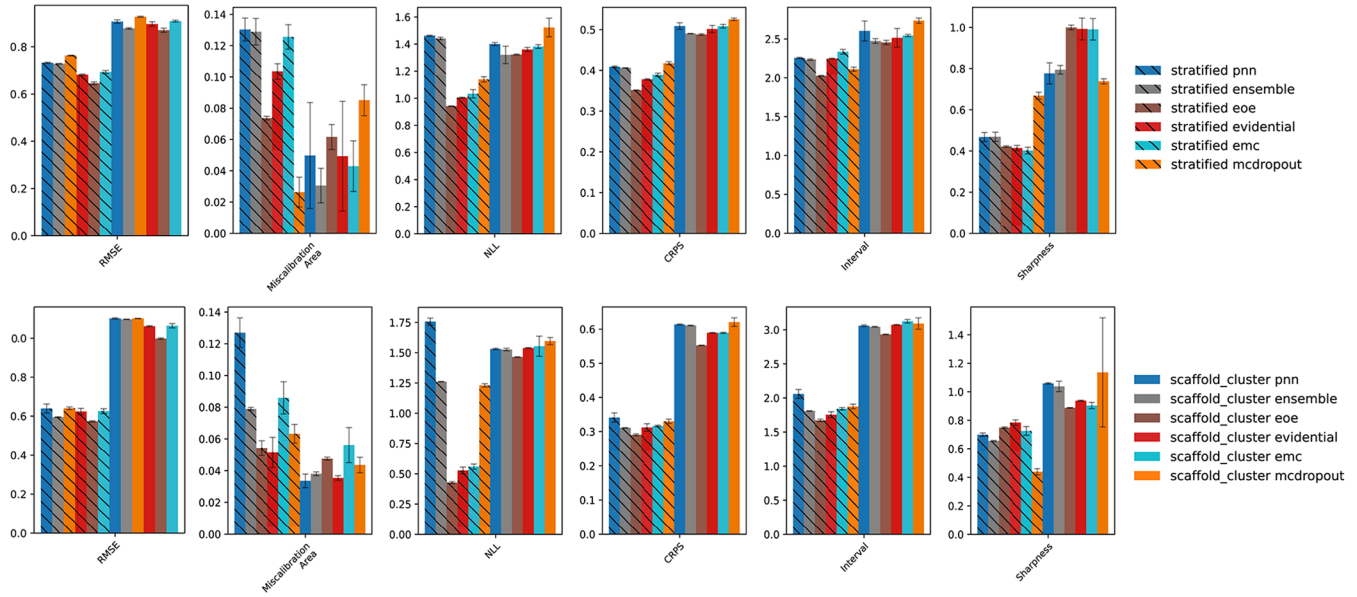


Figure 3. RMSE and uncertainty metrics for xC_{50} and K_x. Bar plots summarizing model performance (RMSE) and uncertainty estimation metrics (miscalibration area, NLL, CRPS, interval, and sharpness) across the xC_{50} (top row) and K_x (bottom row) data sets. All metrics are plotted such that lower values indicate a better performance. This visualization complements the numerical results reported in Table 1 by illustrating trends and differences in model behavior across metrics and data sets.

scaffold cluster-based splitting, which introduces substantial distributional shift and limits the model's ability to generalize to unseen chemical scaffolds. In this context, **EOE_{10}** demonstrates robust performance, outperforming all other models on NLL, CRPS, interval score, sharpness, and RRC-AUC. **EMC_{10}** ranks second on both sharpness and RRC-AUC. These findings highlight the effectiveness of the hybrid evidential-Bayesian modeling approach under distributional stress and low-data generalization scenarios.

Uncertainty-Based Rejection Performance. RMSE-rejection curves (RRC), shown in Figure 4, provide a practical measure of uncertainty effectiveness by quantifying how well models identify uncertain instances. The area under the curve (AUC) (lower is better) is reported in Table 1. $Ensemble_{100}$ generally achieves the lowest AUCs in stratified settings and in xC_{50} under scaffold split. However, in the most challenging setting-K_x with scaffold splitting- EOE_{10} outperforms all others, achieving the best RRC-AUC, indicating superior capability for filtering low-confidence instances in high-uncertainty regimes. Other rank-based metrics from ref 9 were also computed and are available in the Supporting Information. However, we observed that they did not yield consistent conclusions across tasks and were, therefore, not included in the main body of results.

Miscalibration Direction Analysis. To complement the numerical evaluation of calibration error, we analyzed the direction of miscalibration by using calibration curves (Figure 5). While the miscalibration area metric captures the magnitude of deviation from perfect calibration, the curves reveal whether models are overconfident (below the diagonal) or underconfident (above the diagonal). Across settings, EOE_{10}, Evidential, and EMC tend to exhibit slight over-confidence, particularly under scaffold-based splits, whereas MC $Dropout_{100}$ is consistently underconfident. Ensemble-based models generally have better calibration, especially in the scaffold split, benefiting from model averaging. These findings illustrate that directionality of miscalibration varies across models and settings and should be considered alongside quantitative scores to fully assess the calibration behavior.

Overall, EOE_{10} consistently demonstrates a top-tier performance in both RMSE and uncertainty estimation metrics. It strikes a particularly strong balance between accuracy, calibration, and uncertainty informativeness, despite being significantly more computationally efficient than the full $Ensemble_{100}$. EMC_{10}, while slightly less dominant overall, shows favorable sharpness and robustness in certain high-difficulty scenarios. These findings support the effectiveness of combining evidential and Bayesian principles for practical and scalable bioactivity modeling.

Statistical Significance of Model Comparisons. To assess whether observed performance differences between uncertainty quantification methods are statistically meaningful, we first evaluated the distributional assumptions underlying repeated-measures analyses. Figure 6 shows the normality diagnostics for the xC_{50} stratified split, where residuals across seeds appear broadly symmetric and bell-shaped, although formal Shapiro–Wilk tests[27] reject strict normality. This aligns with previous findings,[28] which caution against over-reliance on normality tests in repeated-measures settings and recommend complementing parametric analyses with robust nonparametric alternatives. Accordingly, we applied both repeated-measures ANOVA with Tukey HSD and Friedman tests with Conover–Holm post hoc analysis. Full diagnostic plots for the other data splits (xC_{50} scaffold cluster, K_x stratified, K_x scaffold cluster) are provided in the Supporting Information.

To complement the distributional diagnostics, we assessed the statistical significance of pairwise model differences using the nonparametric Friedman test with Conover-Holm post hoc correction.[28−30] Figure 7 shows the resulting sign plot for the xC_{50} stratified split, as a representative example. Three robust patterns emerge. *First*, differences in performance/error (RMSE) are statistically significant across *all* model pairs. *Second*, the same holds for the probabilistic scoring metrics

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

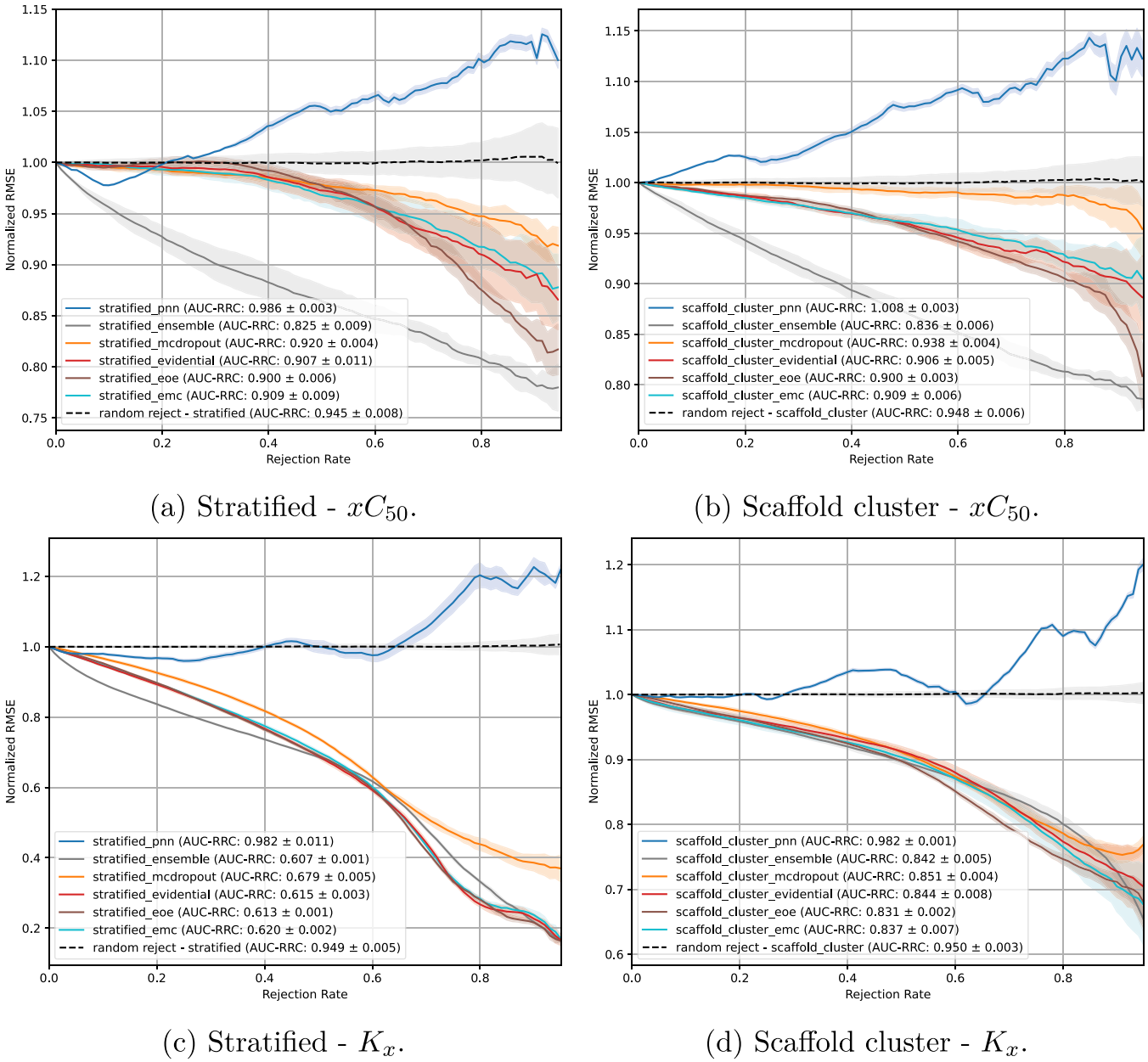


Figure 4. RRC for xC_{50} and K_x under stratified and scaffold cluster splits. (a) Stratified-xC_{50}. (b) Scaffold cluster-xC_{50}. (c) Stratified-K_x. (d) Scaffold cluster-K_x. RRC illustrates the utility of uncertainty estimates in identifying unreliable instances. For each model, the RMSE is computed after progressively removing the samples with the highest estimated total uncertainty. A lower curve indicates that uncertain outputs are effectively ranked and filtered, leading to an improved overall estimation quality. EOE_{10} and $Ensemble_{100}$ demonstrate consistently strong rejection performance, with EOE_{10} particularly excelling in the most challenging setting (K_x, scaffold cluster split).

NLL and CRPS, where every pairwise contrast is significant. *Third*, for the miscalibration area, most model pairs do not reach significance, consistent with the qualitative conclusions from the miscalibration direction analysis, and the difference between EOE and ensemble remains statistically significant. Importantly, EOE also demonstrated statistically significant improvements compared to the full ensemble model in nearly all metrics. While the strength of evidence varied (with *p*-values ranging from very strong, $p < 0.001$, to more moderate, $p < 0.01$), the consistency of significance highlights the robustness of EOE's advantage.

Beyond the sign plots, we employed multiple comparison of scores (MCS) plots and critical difference (CD) diagrams to provide complementary perspectives on statistical significance. MCS plots summarize mean differences and adjusted *p*-values in a heatmap-style format, enabling direct inspection of effect sizes. CD diagrams, in contrast, depict average ranks across models and highlight cliques of methods that cannot be statistically distinguished. Together, these complementary statistical visualizations emphasize not only the detectability but also the robustness and practical significance of EOE's advantage. Full sets of MCS and CD plots for all metrics and splits are provided in the Supporting Information.

Compute Budget and Training Cost Summary. To contextualize performance and UQ results with computational demand, training cost was quantified in hardware-agnostic units as GPU-hours, defined as wall-clock hours multiplied by the number of concurrently used GPUs. In our setup, each run

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

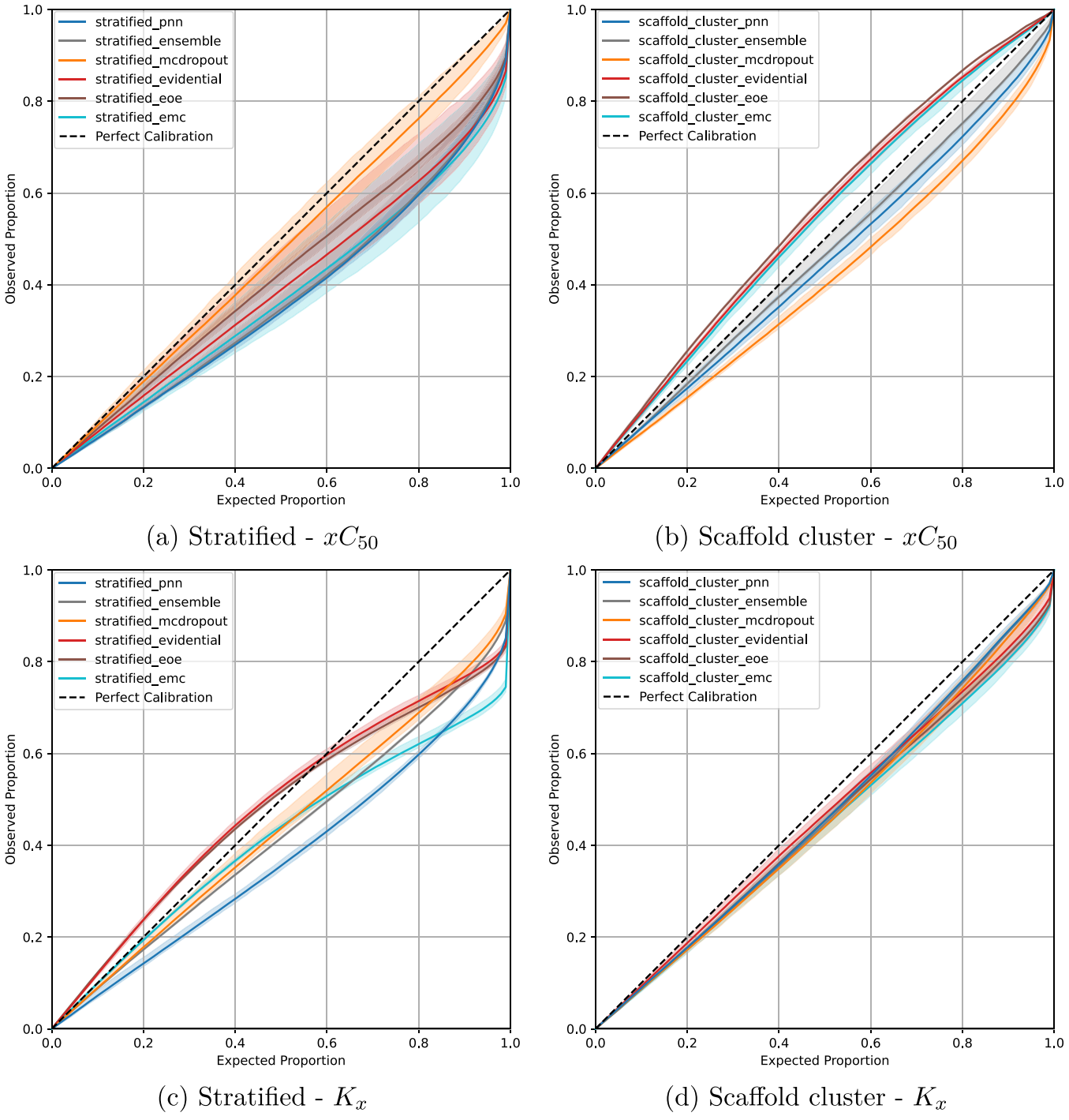


Figure 5. Calibration curves of models evaluated on xC_{50} and K_x data sets, assessing the alignment between estimated confidence levels and observed accuracy. (a) Stratified−xC_{50}. (b) Scaffold cluster−xC_{50}. (c) Stratified−K_x. (d) Scaffold cluster−K_x. The dashed line represents a perfect calibration, where the observed proportion matches the expected proportion.

exclusively occupied a single GPU; therefore, GPU-hours are numerically equal to wall-clock hours. Runtimes were exported from Weights & Biases logs and aggregated across all seeds, end points, and split types.

Interpretation. Training costs varied markedly across methods (Table 2). The full ensemble baseline required by far the largest compute (mean 111.44 GPU-h per run; 4457.58 GPU-h total), reflecting the multiplicative expense of training many independent members. The proposed hybrids were substantially more economical: EOE averaged 8.75 GPU-h per run (350.15 GPU-h total) and EMC 7.05 GPU-h per run (281.88 GPU-h total). Single-model approaches were naturally most compute efficient (Evidential: 4.74 GPU-h; MC Dropout: 2.45 GPU-h; PNN: 1.34 GPU-h).

DISCUSSION

This study presents a systematic evaluation of UQ approaches for bioactivity modeling with a focus on combining evidential and Bayesian paradigms. The proposed hybrid models, **EOE** and **EMC**, aim to synergize the complementary strengths of

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

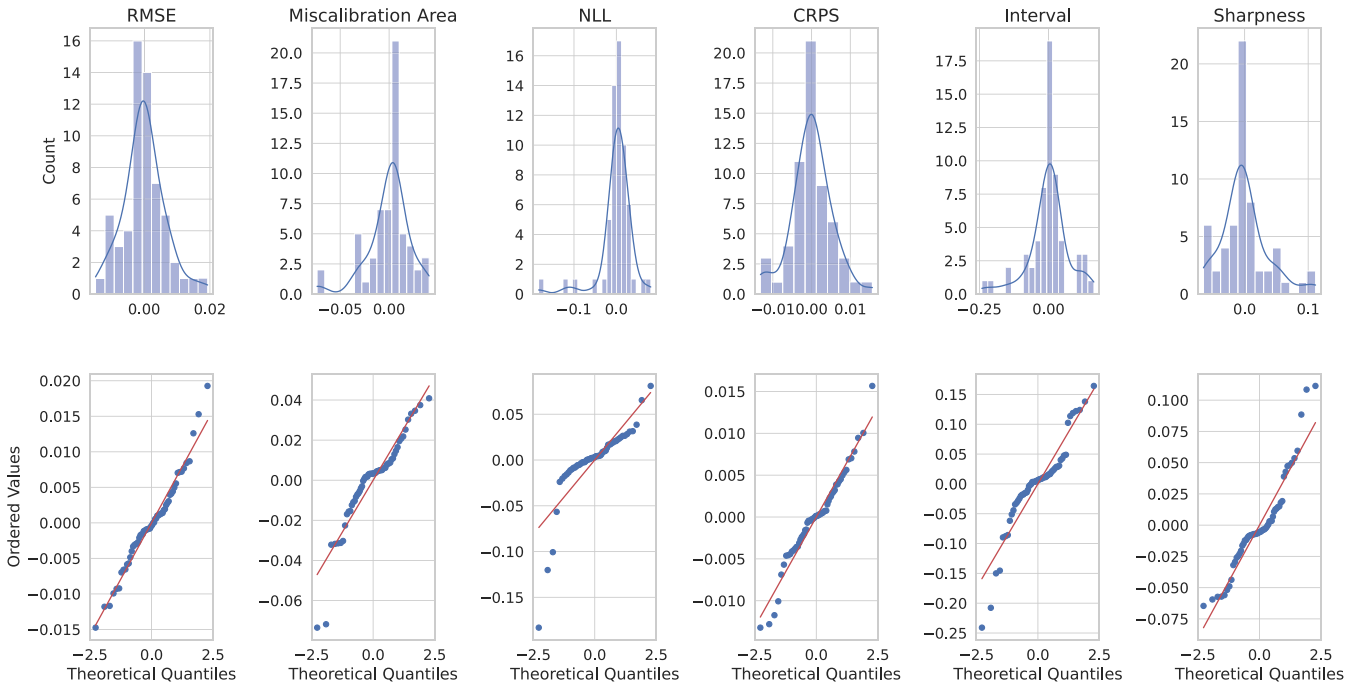


Figure 6. Normality diagnostics for repeated-measures residuals on xC_{50} (stratified split). Residual histograms and Q–Q plots for RMSE, Miscalibration Area, NLL, CRPS, Interval, and Sharpness, used to guide the choice between parametric (RM-ANOVA + Tukey HSD) and nonparametric (Friedman + Conover–Holm) tests. Residuals appear broadly symmetric, though Shapiro tests reject normality, underscoring the importance of complementary nonparametric tests.

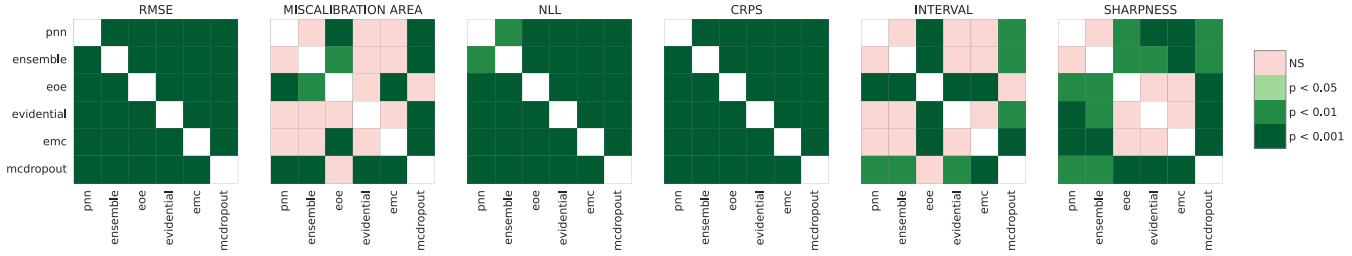


Figure 7. Conover-Holm sign plot for xC_{50} (stratified split). Pairwise post hoc comparisons after Friedman tests across seeds for RMSE and uncertainty metrics. Overall, the EOE versus ensemble difference is significant across *all* metrics in this setting, indicating a consistent advantage for EOE under stratified conditions.

Table 2. Training Cost Aggregated by Model Family

Model	Mean GPU-hours	Total GPU-hours
PNN	1.34	53.57
Ensemble	111.44	4457.58
EOE	8.75	350.15
Evidential	4.74	189.74
EMC	7.05	281.88
MC Dropout	2.45	97.86

these frameworks: the data-driven regularization and expressivity of evidential methods, and the principled variance estimation offered by Bayesian techniques.

Efficacy of Hybrid Approaches. Across all experimental settings, **EOE_{10}** consistently achieved the lowest RMSE values, while requiring nearly an order of magnitude fewer GPU-hours than $Ensemble_{100}$. Statistical tests confirmed that these improvements were significant across most metrics, underscoring that the gains are not only empirical trends but also robust to variation across seeds. This balance of accuracy, significance, and efficiency highlights the hybrid approach as a scalable alternative to large ensembles. Furthermore, EOE_{10} achieved the best scores on CRPS and Interval across all conditions, suggesting that it not only performs well but also provides uncertainty estimates with high informativeness and tight estimated distributions, key attributes in real-world drug discovery workflows.

Uncertainty Quality and Calibration. Uncertainty metrics from the Uncertainty Toolbox revealed important differences in the quality and characteristics of the estimated uncertainties. Following correlation analysis to minimize redundancy, we focused on five complementary metrics, enabling a balanced evaluation across calibration, probabilistic accuracy, and dispersion. This selection, combined with statistical tests, ensures that the reported trends are both interpretable and statistically sound. Although **$Ensemble_{100}$** and **$Evidential_1$** occasionally outperformed EOE_{10} on specific metrics (e.g., Miscalibration Area or NLL), EOE_{10} remained consistently competitive across all axes. Calibration curves further revealed that **MC $Dropout_{100}$** tended to be underconfident, while **EOE_{10}**, **$Evidential_1$**, and **EMC_{10}** exhibited slight overconfidence, particularly under scaffold-cluster splits. These observations highlight the importance of jointly

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

assessing the calibration directionality and magnitude to understand model behavior more comprehensively.

Practical Utility and Significance of Uncertainty Estimates. From a practical perspective, an uncertainty-based rejection analysis demonstrated that **EOE_{10}** and **$Ensemble_{100}$** were particularly effective at identifying unreliable outputs. RRC-AUC values showed that **EOE_{10}** achieved the best rejection performance in the most challenging experimental setting—K_x with scaffold cluster split—outperforming the full 100-member ensemble. While **$Ensemble_{100}$** consistently delivered strong ranking utility across other scenarios, its training cost was more than 12-fold higher than **EOE_{10}**. Under realistic compute budgets, **EOE_{10}** therefore provides a more favorable cost-benefit profile, retaining competitive ranking performance while substantially reducing computational demand.

Although the numerical differences in RMSE, calibration, or RRC-AUC may appear modest, they are practically consequential in drug discovery pipelines, where models are used to triage *large* candidate sets. Even incremental gains in calibration or rejection utility can reduce false positives admitted to assay, improve hit enrichment among prioritized compounds, and lower experimental costs. Reliable UQ further enables practitioners to set decision thresholds (e.g., accept, defer, or confirm) that preferentially retain high-confidence outputs and route uncertain cases to follow-up testing. Prior studies confirm that principled UQ frameworks improve success rates in virtual screening and active learning campaigns by coupling probabilistic confidence with ranking.[9−11] Consequently, the statistically significant advantages observed for **EOE_{10}** imply *operational* gains: at fixed experimental budgets, a better-calibrated UQ admits fewer low-quality candidates and concentrates resources on higher quality molecules.

Insights into Model Behavior. Although **EOE_{10}** demonstrated the most robust overall performance, the analysis revealed that no single method was uniformly optimal across all metrics. For example, the **MC $dropout_{100}$** consistently provided sharp uncertainties, albeit with reduced calibration and ranking utility. In contrast, **EMC_{10}** exhibited strong sharpness and favorable calibration in more difficult settings despite being slightly behind the EOE in RMSE and probabilistic scoring metrics.

The results underscore the promise of hybrid UQ architectures in computational drug discovery. While deep ensembles remain a strong baseline, their computational overhead can be prohibitive. Hybrid models like EOE_{10} offer a compelling alternative, achieving competitive or superior performance at lower computational cost.

CONCLUSION

In this study, we investigated the integration of Bayesian and evidential uncertainty quantification techniques for bioactivity modeling. We introduced two hybrid models, **EOE_{10}** and **EMC_{10}**, and benchmarked them against established baselines across multiple data sets and experimental splits. Our results demonstrate that **EOE_{10}** achieves superior performance in both estimation accuracy and uncertainty expressiveness, outperforming traditional ensembles with significantly fewer members. Statistical tests confirmed the robustness of these improvements across multiple uncertainty metrics, while benchmarking in GPU-hours revealed that hybrids achieve this performance at a fraction of the computational cost. Notably, **EOE_{10}** excelled under challenging distributional shifts, highlighting its robustness and practical utility.

Overall, this study demonstrates that EOE provides the best balance of performance, uncertainty calibration, and rank-based reliability, making it the most suitable model for bioactivity assessment tasks. MC dropout is a strong alternative when uncertainty calibration is the primary concern, while ensembles remain effective for robust uncertainty ranking at substantially higher computational expense. These insights guide model selection depending on application priorities and highlight hybrid approaches as scalable and trustworthy tools for decision-making in computational drug discovery.

METHODS

Data Preprocessing and Splitting. In computational drug discovery, selecting and processing high-quality data sets is important to developing models with improved generalizability and robustness.[8] Proper data curation mitigates noise and biases, enhancing model reliability and facilitating informed decision-making. This study leverages Papyrus++, a refined subset of the Papyrus data set,[8] which aggregates standardized bioactivity data from multiple sources, including ChEMBL.[31] Papyrus++ retains only high-confidence measurements, exhibiting agreement across multiple experimental assays, ensuring consistency and reproducibility in bioactivity assessments. The data preprocessing pipeline is illustrated in Figure 1A.

To ensure compatibility with machine learning workflows, we performed rigorous data standardization. Chemical structures were curated using RDKit[32] to correct stereochemical inconsistencies, neutralize formal charges with SMARTS patterns,[33] and unify isotopic representations. Only wild-type human proteins (UniProt-mapped) were retained. The data set was filtered to include only IC_{50} and EC_{50} measurements, referred to as the xC_{50} data set, with targets restricted to those having at least 50 bioactivity measurements, including a minimum of 10 above a threshold of 6.5 on the pChEMBL scale.

To transform molecular information into machine-readable representations, we employed featurization techniques optimized for deep learning. Protein sequences were encoded using Ankh embeddings,[34] a transformer-based protein language model demonstrating state-of-the-art performance in structural and functional tasks. Chemical structures were represented using Extended Connectivity Fingerprints (ECFP),[35] a widely used molecular fingerprinting approach based on the Morgan algorithm. ECFP was computed with a radius of 2 and a vector length of 2048, ensuring a fixed-length representation capturing substructural patterns relevant for bioactivity modeling.

Robust data partitioning strategies were employed to evaluate the uncertainty estimation techniques under different generalization scenarios. Two splitting strategies were applied: (1) a **stratified split**, ensuring proportional representation of scaffold-based clusters across training (70%), validation (15%), and test (15%) sets to evaluate model performance in a controlled setting, and (2) a **scaffold cluster split**, where entire scaffold clusters were assigned to distinct subsets to assess model generalization to novel chemical scaffolds.

To implement scaffold clustering, we followed the methodology of He et al.[36] Scaffolds were extracted using the Bemis-Murcko algorithm,[37] which isolates core cyclic frameworks of molecules. Similarity between scaffolds was computed using

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Maximum Common Substructure (MCS) similarity,[38] quantified by the Tanimoto coefficient. Hierarchical clustering with Ward linkage[39] was then applied to assign scaffolds into clusters, optimizing within-cluster variance. The optimal number of scaffold clusters was determined using the silhouette coefficient,[40] ensuring meaningful chemical partitioning. These strategies provide a robust framework for evaluating uncertainty quantification methods under both interpolation and extrapolation settings.

Hyperparameter Optimization. For hyperparameter optimization, we employed a Bayesian Optimization and Hyperband (BOHB) search approach.[41] Specifically, we leveraged the sweep functionalities provided by the "Weights and Biases" (`Wandb`) package[42] to efficiently explore the hyperparameter space and identify the configurations that yielded the best performance on the validation set. Bayesian optimization utilizes a probabilistic model to guide the selection of hyperparameter values through an iterative process of testing on a surrogate function before evaluating the actual objective function.

The hyperparameters we optimized included the number and size of hidden layers and featurizers including protein layer (for protein embeddings) and chemical layers (for chemical features), the type of optimizer, weight decay, learning rate (LR), dropout rate, batch size, and the maximum norm for gradients' clipping. Supporting Table S2 outlines the various values we tested per each of the hyperparameters. The final chosen hyperparameters are provided in the run configurations of the accompanying code.

Uncertainty Estimation Models. Uncertainty estimation is used in deep learning models for bioactivity assessment, enabling a more informed decision-making process in drug discovery. In this study, we investigate five uncertainty estimation approaches built upon a shared neural network model. These include Bayesian approaches, i.e., Deep Ensembles and MC dropout, as well as Evidential Deep Learning and two hybrid models that integrate Bayesian modeling with evidential learning. While Bayesian approaches capture epistemic uncertainty through sampling different models according to a posterior distribution, evidential learning aims to directly learn a second-order distribution of possible estimated distributions.

Each method extends the core model architecture to provide distinct approaches to quantifying aleatoric and epistemic uncertainties. While the classical PNN serves as the foundation for ensemble and MC dropout models, the evidential models (including the hybrid approaches) utilize a modified last layer that outputs different distribution parameters. Nevertheless, all models share the same architecture for feature extraction and regression, ensuring consistency in the hyperparameter tuning.

PNN. PNNs[12] assume a Gaussian distribution for the target variable, thus the model outputs both a mean $\mu(\boldsymbol{x};\boldsymbol{\theta})$ and variance $\sigma^2(\boldsymbol{x};\boldsymbol{\theta})$.

The estimated distribution is then given by

$$p(y|\mu(\boldsymbol{x};\boldsymbol{\theta}),\sigma^2(\boldsymbol{x};\boldsymbol{\theta})) = \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x};\boldsymbol{\theta})}}\exp\left\{-\frac{(y-\mu(\boldsymbol{x};\boldsymbol{\theta}))^2}{2\sigma^2(\boldsymbol{x};\boldsymbol{\theta})}\right\} \tag{1}$$

Training is performed by minimizing the negative log-likelihood of eq 1.[12] The PNN model architecture comprises separate feature extractors for chemical structures and protein sequences, each feeding into a shared regression layer that estimates the output distribution. This approach allows for the estimation of the aleatoric uncertainty in deep learning models for molecular property assessments.

Deep Ensemble. Deep Ensembles[13] leverage multiple independently trained models to enhance accuracy and uncertainty estimation. Each model is initialized with different random weights and trained independently. Then, the final output is obtained by aggregating outputs across the ensemble. Deep Ensembles have been shown to be very faithful to the Bayesian ideal[15] and are often considered as gold standard for approximate Bayesian inference.[43] The expected mean output is computed as

$$\mathbb{E}[\mu(\boldsymbol{x})] \approx \bar{\mu}(\boldsymbol{x}) = \frac{1}{M}\sum_{i=1}^{M}\mu(\boldsymbol{x};\boldsymbol{\theta}_i) \tag{2}$$

where $M = 100$ is the number of ensemble members. Following ref 44, the uncertainty in the regression setting is usually captured as the variance of the outputs $\mathrm{Var}[y]$, which can the decomposed using the law of total variance

$$\mathrm{Var}[y] = \mathbb{E}[\sigma^2(\boldsymbol{x})] + \mathrm{Var}[\mu(\boldsymbol{x})] \tag{3}$$

Here, the first term is usually associated with aleatoric uncertainty, and the second term with epistemic uncertainty.

The aleatoric uncertainty is given by

$$\mathbb{E}[\sigma^2(\boldsymbol{x})] \approx \bar{\sigma}^2(\boldsymbol{x}) = \frac{1}{M}\sum_{i=1}^{M}\sigma^2(\boldsymbol{x};\boldsymbol{\theta}_i) \tag{4}$$

while the epistemic uncertainty is given by the variance of the output means

$$\mathrm{Var}[\mu(\boldsymbol{x})] \approx \frac{1}{M}\sum_{i=1}^{M}(\mu(\boldsymbol{x};\boldsymbol{\theta}_i) - \bar{\mu}(\boldsymbol{x}))^2 \tag{5}$$

It is worth noting that recent work has proposed Bayesian Optimized Deep Ensembles (BODE), where each ensemble member is hyperparameter-optimized separately to increase diversity and improve calibration.[45] In our approach, the ensemble members were also trained with hyperparameters obtained via Bayesian Optimization; however, the same optimized configuration was applied to all members rather than optimizing each one individually. This strategy provides a computationally efficient compromise while still yielding a strong model performance.

MC Dropout. MC dropout[14] is a Bayesian approximation technique that introduces stochasticity into deep neural networks by activating dropout layers during both training and inference. Multiple forward passes yield a distribution of outputs, from which the mean, aleatoric uncertainty, and epistemic uncertainty are estimated using eqs 2−5. We employ $M = 100$ stochastic passes in our implementation.

Evidential Deep Learning. Evidential deep learning[10,18] models estimate the parameters of a Normal Inverse-γ (NIG) distribution, allowing for direct estimation of aleatoric and epistemic uncertainties. The network outputs four parameters: γ, υ, α, and β (we omit expressing them as functions of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ for brevity), which define the NIG distribution

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$$p(\mu, \sigma^2|\gamma, \upsilon, \alpha, \beta) = \frac{\beta^\alpha \sqrt{\upsilon}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left\{-\frac{2\beta + \upsilon(\gamma-\mu)^2}{2\sigma^2}\right\} \quad (6)$$

The expected mean, aleatoric, and epistemic uncertainties can be directly estimated from the parameters of the NIG distribution and are given by

$$\mathbb{E}[\mu(\boldsymbol{x})] \approx \gamma,\ \mathbb{E}[\sigma^2(\boldsymbol{x})] \approx \frac{\beta}{\alpha-1},\ \mathrm{Var}[\mu(\boldsymbol{x})] \approx \frac{\beta}{\upsilon(\alpha-1)} \quad (7)$$

Hybrid Models: EOE and EMC. To further enhance uncertainty quantification, we offer two novel hybrid approaches: **EOE** (Ensemble of Evidential) and **EMC** (Evidential MC dropout). Both methods combine the expressivity of evidential learning with the variance reduction benefits of Bayesian sampling.

The **EOE model** extends ensemble learning by replacing each individual model with an evidential network. This approach retains ensemble diversity while leveraging evidential learning's capacity to jointly model aleatoric and epistemic uncertainty. In contrast to the standard ensemble using M = 100 members, our EOE implementation employs only M = 10 evidential models for computational efficiency. The **EMC model** applies evidential learning within the MC dropout framework. Each forward pass with activated dropout layers outputs the parameters of the NIG distribution. As in EOE, we use M = 10 stochastic passes.

In both EOE and EMC, the outputs of the M networks are first aggregated by computing the mean of each NIG parameter $\overline{\alpha}$, $\overline{\beta}$, $\overline{\gamma}$, and $\overline{\upsilon}$. These averaged parameters define a new Normal Inverse-γ distribution, from which we compute the final output and uncertainty estimates using the same closed-form expressions as in standard evidential learning (same as in eq 7)

$$\mathbb{E}[\mu(\boldsymbol{x})] = \overline{\gamma}(\boldsymbol{x}),\ \mathbb{E}[\sigma^2(\boldsymbol{x})] = \frac{\overline{\beta}(\boldsymbol{x})}{\overline{\alpha}(\boldsymbol{x})-1},\ \mathrm{Var}[\mu(\boldsymbol{x})] = \frac{\overline{\beta}(\boldsymbol{x})}{\overline{\upsilon}(\boldsymbol{x})(\overline{\alpha}(\boldsymbol{x})-1)} \quad (8)$$

Metrics for Model Evaluation. To rigorously evaluate both the accuracy and reliability of uncertainty estimation, we adopt a focused set of metrics that reflect the key objectives of our study: reliable estimation quality, well-calibrated uncertainty, and practical utility for selective decision-making. All metrics used in the main analysis are computed on the test sets and averaged across multiple seeds. Formal definitions and additional metrics not discussed here are provided in the Supporting Information.

Performance Metric. We report **RMSE** as the primary performance metric, quantifying the average squared deviation between estimated and observed values. RMSE provides a robust measure of estimation error and is particularly sensitive to large deviations. Additional performance metrics such as $\mathbf{R^2}$ and **PCC** are presented in the Supporting Information.

Uncertainty Metrics. To assess the quality of the estimated uncertainty distributions, we employ five nonredundant metrics from the Uncertainty Toolbox,[26] selected for their complementarity and empirical relevance. The **miscalibration area** quantifies the discrepancy between estimated and observed confidence levels using calibration curves. The **NLL** evaluates the likelihood of the observed values under the estimated distributions, penalizing overconfident or misaligned estimates. The **CRPS** assesses the compatibility between the estimated cumulative distribution and the empirical distribution. The **interval score** jointly evaluates the width and coverage of output intervals, penalizing both overconfident and underconfident outputs. Finally, **sharpness** captures the concentration or narrowness of the estimated distribution regardless of its calibration. Additional metrics, including RMS Calibration, MA Calibration, and Check Score, are reported in the Supporting Information.

Uncertainty-Based Rejection Performance. To evaluate the practical utility of uncertainty estimates for decision-making, we use the **RRC-AUC**. This metric summarizes how RMSE changes when samples with the highest estimated total uncertainty are iteratively removed. Lower RRC-AUC values indicate that the model's uncertainty estimates are effective in identifying unreliable outputs. Other rank-based metrics from ref 9, such as Spearman's Rank Correlation between uncertainty and error, or are reported in the Supporting Information for completeness.

Statistical Significance Testing. To determine whether observed differences in performance and uncertainty quantification metrics were statistically significant, we employed a two-step procedure combining normality diagnostics with both parametric and nonparametric repeated-measures tests, following the methodology outlined by Ash et al.[28]

First, we assessed the distribution of residuals across seeds using a visual inspection of histograms and Q-Q plots, complemented by Shapiro-Wilk tests. As noted by Ash et al.,[28] such tests can be overly sensitive in small-sample regimes, and normality assessment should not rely solely on hypothesis tests but also on graphical diagnostics. Given the inconclusive results from Shapiro-Wilk across some metrics, we applied both parametric and nonparametric approaches to ensure robustness.

When approximate normality could be assumed, we used a repeated-measures ANOVA with seed as the blocking factor, followed by Tukey's honest significant difference (HSD) post hoc test to obtain pairwise adjusted p-values. This analysis also quantified the magnitude and direction of mean differences, which we summarized in multiple comparison of score (MCS) plots.

In parallel, we applied the Friedman test with seeds as blocks to account for potential non-normality, followed by Conover post hoc pairwise comparisons with Holm correction for multiple testing.[29,30] These nonparametric results were visualized using sign plots, which directly encode significant pairwise contrasts, and critical difference (CD) diagrams, which summarize average ranks across models and identify cliques of methods without significant differences.[46]

For each bioactivity end point (xC_{50} and K_x) and each split type (stratified and scaffold cluster), we conducted tests separately across models and across the selected performance and uncertainty metrics (RMSE, miscalibration area, NLL, CRPS, interval, sharpness). This dual-testing strategy ensured that conclusions did not depend on a single assumption about distributional form, and aligns with recent recommendations for practically significant model comparisons in molecular machine learning.[28] Importantly, this ensures that observed advantages, such as those of **EOE** over ensembles, are statistically robust and not artifacts of a single analysis choice.

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

■ ASSOCIATED CONTENT

Data Availability Statement

All bioactivity data used in this study originate from the publicly available Papyrus data set (version 05.6, 10.5281/zenodo.7821775). Our code automatically downloads, filters, and processes this data set for modeling, ensuring full reproducibility. The complete software used for model training, evaluation, and uncertainty quantification is provided as a.zip archive in the Supporting Information and is publicly available on GitHub at https://github.com/CDDLeiden/uqdd/.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c01597.

Detailed formulations and extended experimental results; **Mathematical Notation; Metric Formulations; Hyperparameter Optimization; Statistical Significance Analyses**; visual illustrations such as pairplots showing intermetric relationships, bar plots comparing RRC-AUC values, and recalibration comparisons using isotonic regression to assess post hoc model calibration; and extensive tables reporting performance and uncertainty scores across models, end points (xC_{50}, K_x), and splitting strategies (stratified, scaffold-cluster) (PDF)

UQDD_code (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Gerard J. P. van Westen − *Computational Drug Discovery (CDD), Division of Medicinal Chemistry, Leiden University, 2333 CC Leiden, The Netherlands;* orcid.org/0000-0003-0717-1817; Email: gerard@lacdr.leidenuniv.nl

Herman van Vlijmen − *Computational Drug Discovery (CDD), Division of Medicinal Chemistry, Leiden University, 2333 CC Leiden, The Netherlands; In Silico Discovery (ISD), Johnson & Johnson, 2340 Beerse, Belgium;* orcid.org/0000-0002-1915-3141; Phone: +32 473 55 59 89; Email: h.van.vlijmen@lacdr.leidenuniv.nl

Authors

Bola Khalil − *Computational Drug Discovery (CDD), Division of Medicinal Chemistry, Leiden University, 2333 CC Leiden, The Netherlands; In Silico Discovery (ISD), Johnson & Johnson, 2340 Beerse, Belgium;* orcid.org/0000-0002-9137-992X

Kajetan Schweighofer − *ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning, Johannes Kepler University Linz, 4040 Linz, Austria;* orcid.org/0009-0001-0482-2699

Natalia Dyubankova − *In Silico Discovery (ISD), Johnson & Johnson, 2340 Beerse, Belgium;* orcid.org/0000-0002-5892-3778

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c01597

Author Contributions

B.K. conceptualized the study, implemented the models, conducted the experiments, and analyzed the results. B.K. wrote the manuscript with parts contributed by K.S. K.S. and N.D. contributed to the mathematical foundations of the uncertainty quantification framework and supported the methodological design. N.D. further contributed to cheminformatics preprocessing, including molecular standardization techniques. G.J.P.v.W. and H.v.V. provided guidance on manuscript structure and alignment with computational drug discovery applications, contributed to scientific discussions, and reviewed and edited the manuscript. All authors reviewed and approved the final manuscript.

Funding

B.K. acknowledges funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 955879. K.S. acknowledges support from the ELLIS Unit Linz, the LIT AI Lab, and the Institute for Machine Learning at Johannes Kepler University Linz. The ELLIS Unit Linz, the LIT AI Lab, and the Institute for Machine Learning are supported by the Federal State Upper Austria. We thank the projects FWF AIRI FG 9-N (10.55776/FG9), AI4GreenHeatingGrids (FFG-899943), Stars4Waters (HORIZON-CL6−2021-CLIMATE-01−01), and FWF Bilateral Artificial Intelligence (10.55776/COE12). We thank NXAI GmbH, Audi AG, Silicon Austria Laboratories (SAL), Merck Healthcare KGaA, GLS (Univ. Waterloo), TÜV Holding GmbH, Software Competence Center Hagenberg GmbH, SPACE GmbH, TRUMPF SE + Co. KG.

Notes

The authors declare the following competing financial interest(s): B.K., N.D., and H.v.V. are affiliated with Johnson and Johnson Innovative Medicine.

■ REFERENCES

(1) Mervin, L. H.; Johansson, S.; Semenova, E.; Giblin, K. A.; Engkvist, O. Uncertainty quantification in drug design. *Drug Discovery Today* **2021**, *26*, 474−489, DOI: 10.1016/j.drudis.2020.11.027.

(2) Gawlikowski, J.; Tassi, C. R. N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; Shahzad, M.; Yang, W.; Bamler, R.; Zhu, X. X. A Survey of Uncertainty in Deep Neural Networks. 2021 arXiv:2107.03342. arXiv.org e-Printarchive. https://arxiv.org/abs/2107.03342.

(3) Yu, J.; Wang, D.; Zheng, M. Uncertainty quantification: Can we trust artificial intelligence in drug discovery? *iScience* **2022**, *25*, No. 104814.

(4) Svensson, E.; Friesacher, H. R.; Winiwarter, S.; Mervin, L.; Arany, A.; Engkvist, O. Enhancing Uncertainty Quantification in Drug Discovery with Censored Regression Labels 2024 arXiv:2409.04313. arXiv.org e-Printarchive. https://arxiv.org/abs/2409.04313.

(5) Roth, J. P.; Bajorath, J. Relationship between prediction accuracy and uncertainty in compound potency prediction using deep neural networks and control models. *Sci. Rep.* **2024**, *14*, No. 6536.

(6) Hüllermeier, E.; Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning* **2021**, *110*, 457−506.

(7) Landrum, G. A.; Riniker, S. Combining IC50 or K i Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **2024**, *64*, 1560−1567.

(8) Béquignon, O. J. M.; Bongers, B. J.; Jespers, W.; IJzerman, A. P.; Water, B. v. d.; Westen, G. J. P. v. Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *J. Cheminf.* **2023**, *15*, No. 3.

(9) Rasmussen, M. H.; Duan, C.; Kulik, H. J.; Jensen, J. H. Uncertain of uncertainties? A comparison of uncertainty quantification metrics for chemical data sets. *J. Cheminf.* **2023**, *15*, No. 121.

(10) Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. Evidential Deep Learning for Guided Molecular Property prediction and Discovery. *ACS Cent. Sci.* **2021**, *7*, 1356−1367.

(11) Hirschfeld, L.; Swanson, K.; Yang, K.; Barzilay, R.; Coley, C. W. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 3770−3780.
(12) Nix, D.; Weigend, A. In *Estimating the mean and variance of the target probability distribution*, Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94); IEEE, 1994; pp 55−60.
(13) Lakshminarayanan, B.; Pritzel, A.; Blundell, C. In *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*, Advances in Neural Information Processing Systems; NIPS, 2017.
(14) Gal, Y.; Ghahramani, Z. In *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, International Conference on Machine Learning; PMLR, 2016; pp 1050−1059.
(15) Wilson, A. G.; Izmailov, P. In *Bayesian Deep Learning and a Probabilistic Perspective of Generalization*, Advances in Neural Information Processing Systems; NIPS, 2020; pp 4697−4708.
(16) Izmailov, P.; Vikram, S.; Hoffman, M. D.; Wilson, A. G. G. In *What Are Bayesian Neural Network Posteriors Really Like?*, Proceedings of the 38th International Conference on Machine Learning; PMLR, 2021; pp 4629−4640.
(17) Sensoy, M.; Kaplan, L.; Kandemir, M. In *Evidential deep learning to quantify classification uncertainty*, Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY, USA; NIPS, 2018; pp 3183−3193.
(18) Amini, A.; Schwarting, W.; Soleimany, A.; Rus, D. In *Deep Evidential Regression*, Advances in Neural Information Processing Systems; NIPS, 2020; pp 14927−14937.
(19) Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Cent. Sci.* **2021**, *7*, 1356−1367.
(20) Ye, K.; Chen, T.; Wei, H.; Zhan, L. In *Uncertainty regularized evidential regression*, Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence;PKP, 2024.
(21) Shen, Z.; Chakraborti, T.; Wang, W.; Yao, S.; Fu, Z.; Chen, Y.; Ding, X. Uncertainty quantification of cuffless blood pressure estimation based on parameterized model evidential ensemble learning. *Biomed. Signal Process. Control* **2024**, *92*, No. 106104.
(22) Shao, Z.; Dou, W.; Pan, Y. Dual-level Deep Evidential Fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning. *Inf. Fus.* **2024**, *103*, No. 102113.
(23) Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta, Gen. Subj.* **2001**, *1525*, 180−190.
(24) Bongers, B. J.; IJzerman, A. P.; Westen, G. J. V. Proteochemometrics - recent developments in bioactivity and selectivity modeling. *Drug Discovery Today:Technol.* **2019**, *32-33*, 89−98.
(25) van Westen, G. J. P.; Wegner, J. K.; IJzerman, A. P.; Vlijmen, H. W. T. v.; Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *MedChemComm* **2011**, *2*, 16−30.
(26) Chung, Y.; Char, I.; Guo, H.; Schneider, J.; Neiswanger, W. Uncertainty Toolbox: an Open-Source Library for Assessing, Visualizing, and Improving Uncertainty Quantification. 2021 arXiv:2109.10254. arXiv.org e-Printarchive. https://arxiv.org/abs/2109.10254.
(27) Shapiro, S. S.; Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591−611.
(28) Ash, J. R.; Wognum, C.; Rodríguez-Pérez, R.; Aldeghi, M.; Cheng, A. C.; Clevert, D.-A.; Engkvist, O.; Fang, C.; Price, D. J.; Hughes-Oliver, J. M.; Walters, W. P. Practically Significant Method Comparison Protocols for Machine Learning in Small Molecule Drug Discovery. *J. Chem. Inf. Model.* **2025**, *65*, 9398−9411.
(29) Conover, W. J.; Iman, R. L. Conover-Iman post-hoc test for rank-based multiple comparisons following Kruskal-Wallis. *On Multiple-Comparisons Procedure*, Technical Reports LA-7677-MS;**1979**.
(30) Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65−70.
(31) Anna, G.; Bellis, J. L.; Bento, P. A.; Jon, C.; Mark, D.; Anne, H.; Yvonne, L.; Shaun, M.; David, M.; Bissan, A.-L.; P, J. Overington ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(32) Landrum, G.; Tosco, P.; Kelley, B.et al. rdkit/rdkit: 2023_09_6 (Q3 2023) Release. 2024 DOI: 10.5281/zenodo.10793672.
(33) James, C.; Weininger, D.; Delany, J. 2006 SMARTShttps://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
(34) Elnaggar, A.; Essam, H.; Salah-Eldin, W.; Moustafa, W.; Elkerdawy, M.; Rochereau, C.; Rost, B. Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. 2023 arXiv:2301.06568. arXiv.org e-Printarchive. https://arxiv.org/abs/2301.06568.
(35) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(36) He, K. Pharmacological affinity fingerprints derived from bioactivity data for the identification of designer drugs. *J. Cheminf.* **2022**, *14*, No. 35.
(37) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(38) Zhang, B.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures. *JJ. Comput.-Aided Mol. Des.* **2015**, *29*, 937−950.
(39) Müllner, D. Modern hierarchical, agglomerative clustering algorithms. 2011 arXiv:1109.2378. arXiv.org e-Printarchive. https://arxiv.org/abs/1109.2378.
(40) Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53−65.
(41) Falkner, S.; Klein, A.; Hutter, F. BOHB: Robust and Efficient Hyperparameter Optimization at Scale. 2018 arXiv:1807.01774. arXiv.org e-Printarchive. https://arxiv.org/abs/1807.01774.
(42) Biewald, L. Experiment Tracking with Weights and Biases 2020 https://www.wandb.com/, Software available from wandb.com.
(43) Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; Snoek, J. In *Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift*, Advances in Neural Information Processing Systems; NIPS, 2019.
(44) Depeweg, S.; Hernandez-Lobato, J.-M.; Doshi-Velez, F.; Udluft, S. In *Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning*, Proceedings of the 35th International Conference on Machine Learning; PMLR, 2018; pp 1184−1193.
(45) Abulawi, Z.; Hu, R.; Balaprakash, P.; Liu, Y. Bayesian Optimized Deep Ensemble for Uncertainty Quantification of Deep Neural Networks: a System Safety Case Study on Sodium Fast Reactor Thermal Stratification Modeling. *Reliab. Eng. Syst. Saf.* **2025**, *264*, No. 111353.
(46) Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1−30.

https://doi.org/10.1021/acs.jcim.5c01597
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX