A Domain-Informed Composite Kernel Gaussian Process Classifier for Molecular Toxicity Prediction with Principled Uncertainty Quantification

Junhee Kim1‡, Seongil Jo1‡, Jooyeon Lee2, Jahyun Koo2, Jaeoh Kim1*, Keunhong Jeong3*

1Department of Statistics and Data Science, Inha University, Incheon, Republic of Korea

2School of Biomedical Engineering, Korea University, Seoul, Republic of Korea

3Department of Chemistry, Sogang University, Seoul, Republic of Korea

‡ Contributed equally

Correspondence: Jaeoh Kim (jaeoh.k@inha.ac.kr), Keunhong Jeong (doas1mind@sogang.ac.kr)

Abstract

Computational toxicity prediction remains a critical bottleneck in drug discovery, where most machine learning models either rely on single-modality molecular representations or lack the principled uncertainty quantification essential for regulatory decision-making. We present the first application of a domain-informed composite kernel Gaussian Process Classifier (Hybrid GPC) to large-scale molecular toxicity prediction, integrating Morgan circular fingerprints and PCA-reduced RDKit physicochemical descriptors through a novel kernel architecture combining Tanimoto, radial basis function, and Matérn components. Trained on one of the largest curated binary toxicity datasets to date (13,949 molecules from Tox21 and T3DB) and evaluated under a scaffold-based splitting strategy that strictly prevents structural data leakage, the model achieved an AUC of 0.8878, accuracy of 0.8292, F1-score of 0.8022, and a Brier score of 0.1350. Systematic ablation across seven kernel configurations confirmed that the composite kernel yields the optimal trade-off between discriminative performance and probability calibration, surpassing both single-kernel Gaussian process baselines and conventional classifiers. Notably, the Hybrid GPC outperformed MolToxPred—the most directly comparable prior study—in both AUC (0.8878 vs. 0.8776) and F1-score (0.8022 vs. 0.7643) despite employing the more stringent scaffold-based evaluation protocol. To our knowledge, this work represents the first integration of heterogeneous molecular kernels within a Gaussian process classification framework for toxicity prediction, delivering competitive performance alongside intrinsic, calibrated uncertainty estimates—a capability absent from existing ensemble and deep learning approaches.

Keywords: Gaussian process classifier; molecular toxicity prediction; composite kernel; Tanimoto kernel; uncertainty quantification; scaffold-based splitting; cheminformatics

1. Introduction

The accurate assessment of molecular toxicity is a central challenge in drug discovery, chemical safety evaluation, and environmental risk management. Toxicity-related failures account for a substantial proportion of attrition in pharmaceutical development pipelines—approximately 30% of drug candidate failures are attributable to unforeseen toxic effects—imposing enormous costs on both the industry and public health [1,2]. Traditional experimental approaches, including *in vitro* assays and *in vivo* animal studies, remain the gold standard for toxicological evaluation; however, they are associated with high resource demands, ethical constraints arising from the use of animal subjects, and inherently limited throughput when applied to large compound libraries [3,4]. These limitations have created a compelling need for reliable computational methods capable of providing early, cost-effective toxicity estimates directly from molecular structure. Over the past two decades, quantitative structure–activity relationship (QSAR) modeling and, more recently, machine learning (ML)-based approaches have emerged as powerful alternatives to experimental screening [5,6]. By encoding molecular structures as numerical representations—ranging from classical physicochemical descriptors and binary fingerprints to graph-based embeddings—these methods allow predictive models to be trained on existing toxicological data and subsequently applied to virtual compound libraries at negligible computational cost [7,8]. Landmark tools such as DeepTox [9], eToxPred [10], ToxiM [11], and more recently MolToxPred [12] and ToxinPredictor [13] have demonstrated the feasibility of achieving competitive predictive performance across diverse toxicity endpoints. Despite this progress, the field continues to face two persistent challenges: (i) the effective integration of complementary molecular representations that capture both structural topology and physicochemical variation, and (ii) the quantification of prediction uncertainty in a principled manner suitable for regulatory decision-making [6,14]. Recent studies have begun to address the representation integration challenge through graph neural network frameworks augmented with density functional theory (DFT) calculations for safety-critical property prediction [24] and Kronecker-product-based multimodal fusion of graph embeddings with statistically selected molecular descriptors [25]; however, neither approach provides the principled Bayesian uncertainty quantification central to the present work.

Molecular representation is widely recognized as one of the most critical design choices in building accurate and generalizable toxicity prediction models [6,8]. Structural fingerprints, particularly Morgan (extended-connectivity) fingerprints, encode the presence or absence of local substructural environments within a fixed-length binary vector, making them especially well suited for measuring pairwise structural similarity [7]. Physicochemical descriptors, on the other hand, capture continuous properties such as molecular weight, lipophilicity, polar surface area, and hydrogen-bonding capacity, providing complementary information about a molecule's interaction propensity with biological macromolecules [8,11]. Recent empirical and theoretical analyses have demonstrated that the joint use of structural and physicochemical representations consistently yields superior performance compared to either

modality alone, as the two types of information are largely orthogonal and mutually reinforcing [12,15,16]. This complementarity has been further substantiated by recent multimodal molecular property prediction frameworks, where explicit second-order interaction modeling between graph-derived and descriptor-based features yielded consistent improvements across diverse endpoints [24,25]. Beyond predictive accuracy, the provision of calibrated uncertainty estimates is increasingly recognized as an essential requirement for the practical deployment of toxicity prediction models. In regulatory and early drug discovery contexts, a model that outputs only binary classifications or point-probability estimates provides insufficient information for risk-based decision making: an understanding of the model's confidence in each prediction—and of how far a query molecule lies from the training distribution—is necessary to identify cases where predictions should be treated with caution [14,17]. While ensemble methods and Bayesian neural networks have been proposed as approaches to approximate predictive uncertainty in deep learning models [16,18], these frameworks lack the mathematical rigor and native calibration guarantees that probabilistic Bayesian methods such as Gaussian processes provide. Gaussian process (GP) models constitute a principled class of non-parametric Bayesian methods that simultaneously produce predictions and rigorous uncertainty estimates as an intrinsic consequence of their probabilistic formulation [19]. The predictive distribution of a GP is fully characterized by a posterior mean and variance, with the variance providing a natural measure of epistemic uncertainty that increases as queries move away from training data. The kernel function at the heart of any GP model encodes assumptions about the similarity structure of the input space; consequently, the design of appropriate kernels for molecular data is a critical and non-trivial problem [19]. For molecular applications, the Tanimoto coefficient—defined as the ratio of the intersection to the union of two binary fingerprint bit vectors—provides a theoretically well-founded and empirically validated measure of structural similarity [7]. However, a kernel based solely on the Tanimoto coefficient cannot exploit the rich physicochemical information captured by continuous molecular descriptors, necessitating a composite kernel architecture that seamlessly integrates both representation types. Despite the theoretical elegance and practical advantages of Gaussian process classifiers, their application to large-scale molecular toxicity prediction has remained relatively unexplored compared to ensemble tree methods, neural networks, and support vector machines. This gap is attributable in part to scalability concerns—exact GP inference scales cubically with the number of training points—and in part to the lack of established composite kernel designs for heterogeneous molecular representations. Existing GP-based molecular property prediction studies have predominantly focused on regression tasks such as solubility or activity prediction [19,20], with comparatively little attention to binary toxicity classification and the associated challenge of combining discrete fingerprint kernels with continuous descriptor kernels in a unified, well-calibrated probabilistic framework.

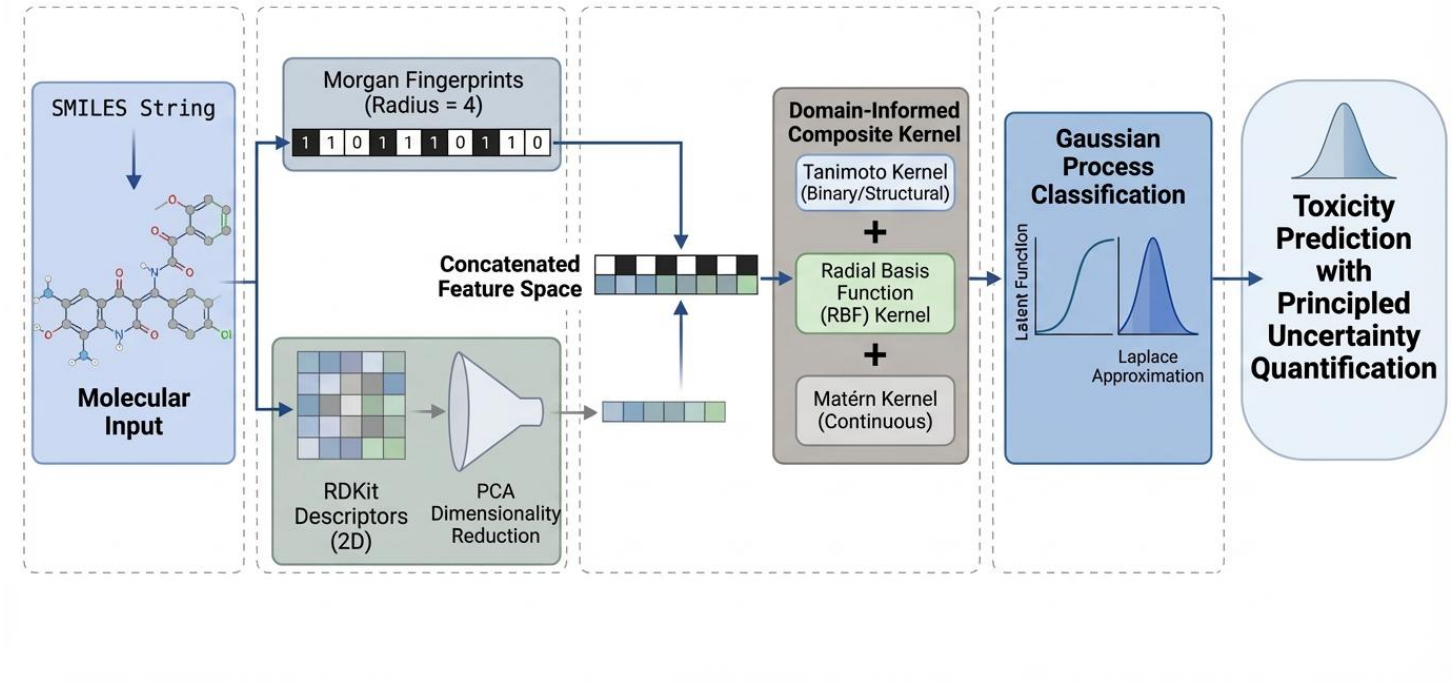

Figure 1. Architecture of the proposed Hybrid GPC: SMILES input is featurized via Morgan fingerprints and PCA-reduced RDKit descriptors, which are concatenated and passed through a domain-informed composite kernel (Tanimoto + RBF + Matérn) into a Gaussian Process Classifier with Laplace approximation, producing calibrated toxicity probabilities with principled uncertainty quantification.

In this study, we address these gaps by developing a Hybrid Gaussian Process Classifier (Hybrid GPC) that combines three complementary kernel components—a Tanimoto kernel for structural fingerprints, and a composite radial basis function (RBF) plus Matérn kernel for reduced-dimensionality physicochemical descriptors—within a single probabilistic model. The model is trained and evaluated on a large, curated dataset of 13,949 unique molecules assembled from two public toxicity resources (Tox_11842 and T3DB), and assessed under a rigorous scaffold-based data splitting strategy that strictly prevents structural information leakage between training and test partitions [21,22]. Scaffold-based splitting is a more conservative and realistic benchmark than random or stratified cross-validation, as it tests a model's ability to generalize to chemically distinct scaffolds not encountered during training—a property of direct relevance to the prospective application of toxicity models in drug discovery [22,23]. The principal contributions of this work are threefold. First, we propose and systematically evaluate a domain-informed composite kernel that integrates Tanimoto-based structural similarity with continuous physicochemical variation, and demonstrate through ablation experiments across seven kernel configurations that this combination yields the best trade-off between discriminative performance and probability calibration. Second, we demonstrate that our Hybrid GPC model achieves an area under the ROC curve (AUC) of 0.8878 and an F1-score of 0.8022 on the scaffold-

split test set, outperforming directly comparable baselines including MolToxPred [12], while providing well-calibrated probabilistic predictions as evidenced by a Brier score of 0.1350. Third, by grounding the analysis in a scaffold-based evaluation protocol, we provide a more rigorous and reproducible benchmark for molecular toxicity prediction than is offered by the majority of existing studies, facilitating fair comparisons with future work (Figure 1). Together, these contributions position the proposed Hybrid GPC as a transparent, uncertainty-aware, and generalizable tool for early-stage toxicity screening in drug discovery and chemical safety assessment.

2. Methodology

This study develops a Gaussian Process Classifier (GPC) for molecular toxicity prediction, with an emphasis on integrating heterogeneous molecular representations through domain-informed kernel design. The proposed approach combines structural and physicochemical information within a probabilistic modeling framework, enabling both accurate prediction and principled uncertainty quantification.

2.1 Data Preprocessing and Feature Representation

Molecular structures were provided as SMILES strings and converted into numerical representations suitable for Gaussian process modeling. To capture complementary aspects of molecular information, both structural fingerprints and physicochemical descriptors were extracted and jointly used as model inputs.

Formally, each molecule is represented as a composite feature vector

$$\mathbf{x}_i = \left[\mathbf{x}_i^{(f)}, \mathbf{x}_i^{(d)}\right], \quad \mathbf{x}_i^{(f)} \in \{0,1\}^{p_f}, \ \mathbf{x}_i^{(d)} \in \mathbb{R}^{p_d}.$$

Structural information was encoded using Morgan fingerprints, which represent the presence or absence of molecular substructures within a fixed-length binary vector. The Morgan radius was selected based on systematic comparison during the hyperparameter optimization stage. An initial radius of four was evaluated to capture both local and moderately extended chemical environments; subsequent targeted tuning (Supplementary Material, Section A.2, Table S1) identified a radius of two as yielding superior calibration, which was adopted in the final model configuration. This representation is particularly well suited for similarity-based learning in cheminformatics applications. In addition to structural fingerprints, two-dimensional molecular descriptors were extracted using RDKit to characterize physicochemical properties such as molecular size, polarity, and functional group composition [7]. Descriptor features were standardized to have zero mean and unit variance, and missing values were imputed with zeros to ensure numerical stability. Because the descriptor space is typically high-dimensional and may contain redundant or noisy features, principal component analysis was applied to reduce dimensionality prior to model fitting. The final molecular representation was constructed by concatenating the fingerprint features with the dimension-reduced descriptor features. This integrated feature space enables the model to simultaneously exploit discrete structural similarity and continuous physicochemical variation, providing a balanced and expressive representation for toxicity prediction.

2.2 Kernel Function Design

The kernel function defines similarity between molecular inputs and plays a central role in Gaussian process models. To accommodate the heterogeneous nature of molecular features, a composite kernel was designed by combining feature-specific kernels within a unified framework.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_c^2 \left[k_f\left(\mathbf{x}^{(f)}, \mathbf{x}'^{(f)}\right) + k_d\left(\mathbf{x}^{(d)}, \mathbf{x}'^{(d)}\right) \right] + k_{\text{noise}}(\mathbf{x}, \mathbf{x}')$$

A constant component was included as a global scaling factor, while a white noise component was incorporated to improve numerical stability during model fitting.

$$k_{\text{noise}}(\mathbf{x}, \mathbf{x}') = \sigma_n^2 \delta_{\mathbf{x},\mathbf{x}'}$$

Structural similarity between molecular fingerprints was modeled using a Tanimoto-based kernel, which is specifically designed for binary representations and effectively captures overlap between molecular substructures.

$$k_f\left(\mathbf{x}^{(f)}, \mathbf{x}'^{(f)}\right) = w_{tan} \cdot \frac{\mathbf{x}^{(f)\top} \mathbf{x}'^{(f)}}{\| \mathbf{x}^{(f)} \|^2 + \| \mathbf{x}'^{(f)} \|^2 - \mathbf{x}^{(f)\top} \mathbf{x}'^{(f)}}$$

Continuous molecular descriptors were modeled using a combination of a radial basis function kernel and a Matérn kernel.

$$k_{\text{d}}\left(\mathbf{x}^{(d)}, \mathbf{x}'^{(d)}\right) = w_{rbf}\, k_{RBF}(\mathbf{x}^{(d)}, \mathbf{x}'^{(d)}) + w_{mat}\, k_{\text{Matérn}}(\text{x}^{(d)}, \text{x}'^{(d)})$$

$$k_{\text{RBF}}\left(\boldsymbol{x}^{(\boldsymbol{d})}, \boldsymbol{x}'^{(\boldsymbol{d})}\right) = \exp\left(-\frac{\| \boldsymbol{x}^{(\boldsymbol{d})} - x'^{(d)} \|^2}{2\ell_r^2} \right)$$

$$k_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\ell_m} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu} r}{\ell_m} \right), \quad r = \| \boldsymbol{x}^{(\boldsymbol{d})} - \boldsymbol{x}'^{(\boldsymbol{d})} \|$$

The radial basis function kernel captures smooth, global similarity patterns across the descriptor space, whereas the Matérn kernel provides additional flexibility by allowing localized variation and controlled smoothness of the latent function.

2.3 Gaussian Process Classification

Gaussian process classification is a nonparametric Bayesian approach that models an underlying latent function governing class membership. The latent function is assumed to follow a Gaussian process prior defined by the specified kernel function.

$$f(\mathbf{x}) \sim \mathcal{GP}\left(0, k(\mathbf{x}, \mathbf{x}')\right)$$

To map latent function values to class probabilities, a logistic sigmoid link function was employed.

$$p(y_i = 1 \mid f_i) = \frac{1}{1 + \exp(-f_i)}$$

Because the resulting likelihood is non-Gaussian, posterior inference was performed using the Laplace approximation, which approximates the posterior distribution of the latent function with a Gaussian distribution centered at its mode.

$$p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} \sigma\,(f_i)^{y_i}\big(1-\sigma(f_i)\big)^{1-y_i}$$

$$p(\mathbf{f} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{f}) \cdot \mathcal{N}(\mathbf{f} \mid 0, K)$$

$$p(\mathbf{f} \mid \mathbf{y}) \approx \mathcal{N}\big(\hat{\mathbf{f}}, (K^{-1}+W)^{-1}\big)$$

This approach enables scalable inference while preserving uncertainty estimates.

2.4 Hyperparameter Optimization

Model performance depends on several kernel hyperparameters, including length-scale parameters, smoothness settings, and variance components.

$$\theta = \{w_{tan}, w_{rbf}, w_{mat}, \nu, \ell_r, \ell_m, \sigma_c^2, \sigma_n^2\}$$

Hyperparameter optimization was conducted using a grid-based search strategy on a randomly selected subset of one thousand training samples, exploring 4,374 kernel configurations across eight hyperparameters (Supplementary Material, Section A.1). Because the initial search did not identify a configuration satisfying the target operating point, a second stage of targeted tuning was performed within the Tanimoto + RBF + Matérn kernel family (Supplementary Material, Section A.2).

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \log p(\mathbf{y} \mid \theta)$$

Once the optimal configuration was identified, the final model was retrained on the full training dataset using the selected kernel weights and parameters. The complete final configuration, including all numerical hyperparameter values, is documented in Supplementary Material, Section A.3.

3. Results

The experimental dataset was constructed by combining two molecular toxicity sources, Tox_11842 and T3DB [21]. After removing duplicated SMILES entries between the two datasets, a total of 13,949 unique molecules were retained. To prevent structural data leakage, a scaffold-based splitting strategy was employed [22,23]. As a result, 13,112 molecules were used for training and 837 molecules were reserved for testing. The test set consisted of 478 non-toxic compounds and 359 toxic compounds, providing a balanced evaluation setting under realistic generalization constraints.

Table 1 summarizes the predictive performance of the proposed Gaussian Process Classifier equipped with the optimized composite kernel combining Tanimoto, radial basis function, and Matérn components.

MODEL	KERNEL	AUC	ACCURACY	F1-SCORE	BRIER SCORE	OPTIMAL THRESHOLD
GAUSSIAN PROCESS CLASSIFIER	Tanimoto + RBF + Matérn	0.8878	0.8292	0.8022	0.1350	0.488

The proposed model achieved an area under the ROC curve of 0.8878, indicating strong discriminative capability under scaffold-based data splitting. The overall classification accuracy reached 0.8292, while the F1-score of 0.8022 reflects a balanced trade-off between precision and recall. In addition, the Brier score of 0.1350 suggests that the predicted probabilities are well calibrated. The optimal decision threshold of 0.488 was determined by maximizing the F1-score over the range of classification thresholds on the test set.

Table 2 presents the confusion matrix corresponding to the optimal decision threshold identified during evaluation.

	PREDICTED NON-TOXIC	PREDICTED TOXIC
ACTUAL NON-TOXIC	404	74
ACTUAL TOXIC	69	290

Using an optimal threshold of 0.488, the model correctly classified 404 non-toxic compounds and 290 toxic compounds. Misclassifications were relatively balanced across classes, indicating stable predictive behavior without strong bias toward either toxic or non-toxic outcomes.

To further investigate the impact of kernel design, we compared the predictive performance of multiple kernel combinations using the same experimental protocol. Figure 2 summarizes the AUC values achieved by different kernel configurations. This comparison directly addresses the question of whether integrating heterogeneous molecular representations through a composite kernel provides measurable improvement over single-kernel approaches, and identifies which kernel components contribute most to discriminative power in toxicity classification. Complete numerical results for all seven kernel configurations are provided in Supplementary Material, Section A.5.

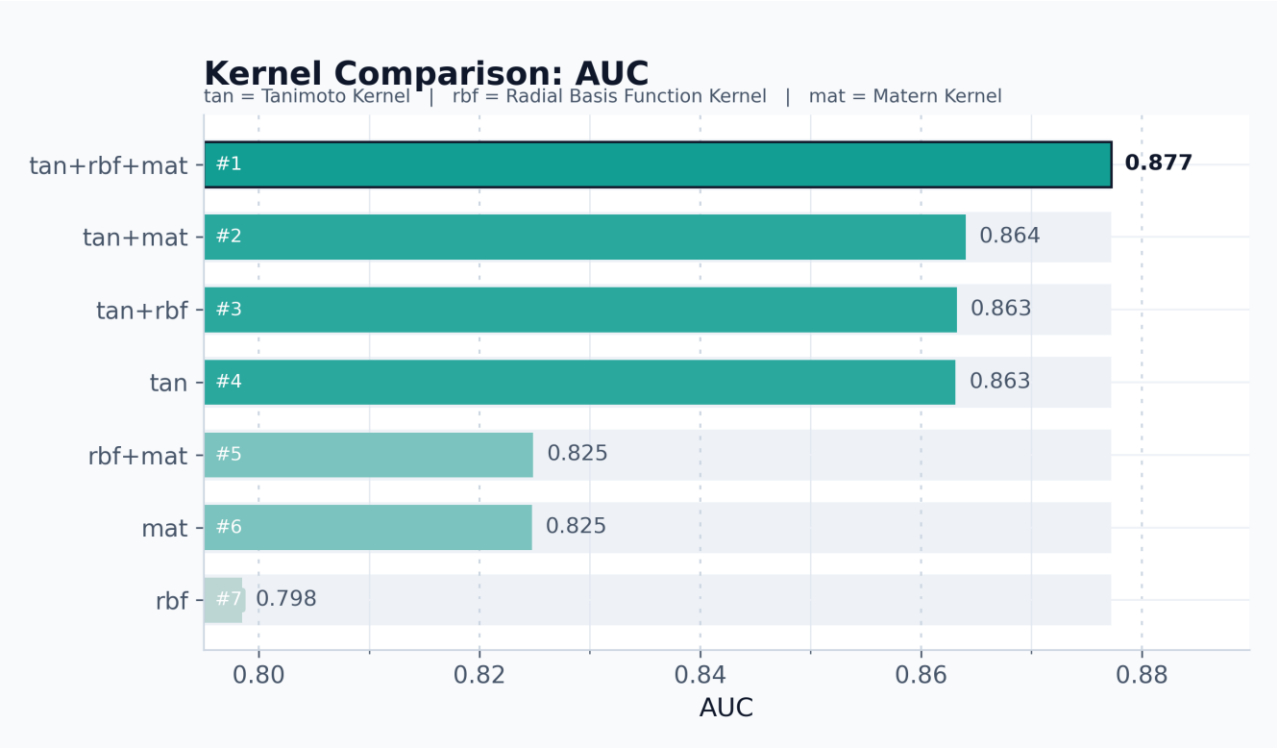

Figure 2. AUC comparison of individual kernels and their combinations, showing that the Tanimoto, RBF, and Matérn kernel combination achieved the best classification performance.

As shown in Figure 2, models that incorporate the Tanimoto kernel consistently outperform those relying solely on continuous kernels, confirming the importance of explicitly modeling structural similarity in molecular toxicity prediction. In particular, the composite kernel achieves the highest AUC among all tested configurations, demonstrating that structural fingerprints and physicochemical descriptors capture complementary aspects of molecular toxicity—the former encodes substructural motifs associated with toxic liability, while the latter captures global molecular properties such as lipophilicity and molecular size that influence bioavailability and target interaction. Figure 3 reports the Brier scores for each kernel configuration, which quantify the calibration quality of predicted probabilities. The composite kernel achieves the lowest Brier score, demonstrating that the improved discriminative performance does not come at the expense of probability calibration. This finding is particularly significant because well-calibrated uncertainty estimates are essential for downstream decision-making in toxicological risk assessment, where overconfident predictions can lead to false safety assurances.

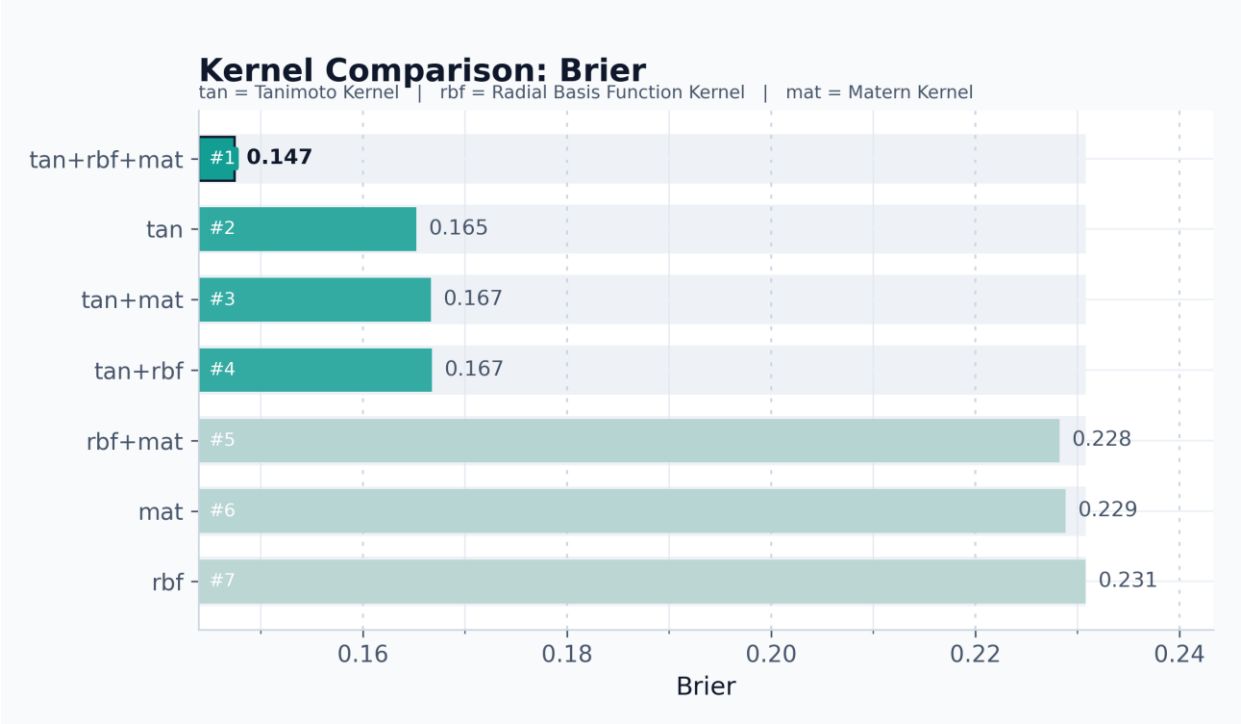

Figure 3. Brier score comparison of individual kernels and their combinations, showing that the Tanimoto, RBF, and Matérn kernel combination provided the best predictive performance with the lowest score.

Figure 4 presents F1-scores evaluated at the optimal decision threshold for each kernel combination. The full composite kernel attains the best balance between precision and recall, indicating that it minimizes both false positives (incorrectly flagging safe compounds as toxic) and false negatives (failing to identify genuinely toxic compounds)—both of which carry significant consequences in pharmaceutical development and chemical safety screening.

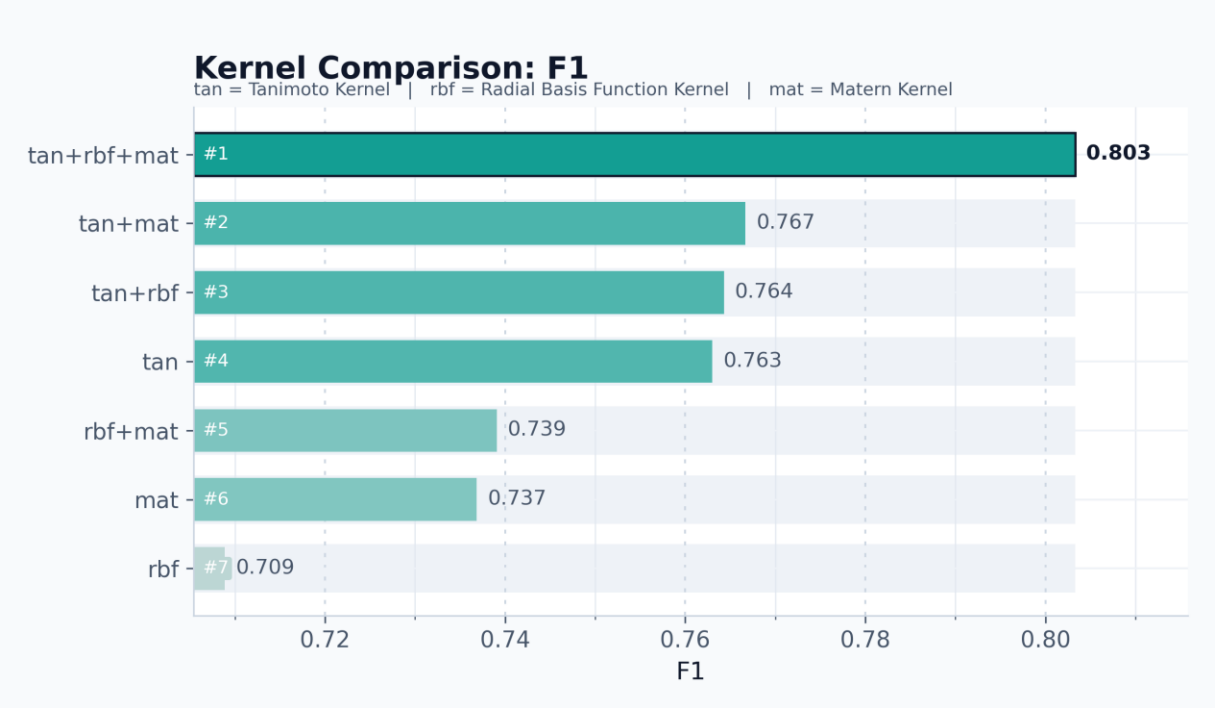

Figure 4. F1-score comparison of individual kernels and their combinations, showing that the Tanimoto, RBF, and Matérn kernel combination achieved the best classification performance.

Finally, Figure 5 visualizes the relationship between F1-score and Brier score across kernel configurations, providing a unified view of the trade-off between classification accuracy and probability calibration. The composite Tanimoto + RBF + Matérn kernel occupies the most favorable position in this space, simultaneously achieving the highest F1-score and the lowest Brier score. This result confirms that the proposed composite kernel does not sacrifice calibration quality for discriminative performance—a common failure mode in complex models—but instead achieves Pareto-optimal behavior across both evaluation dimensions.

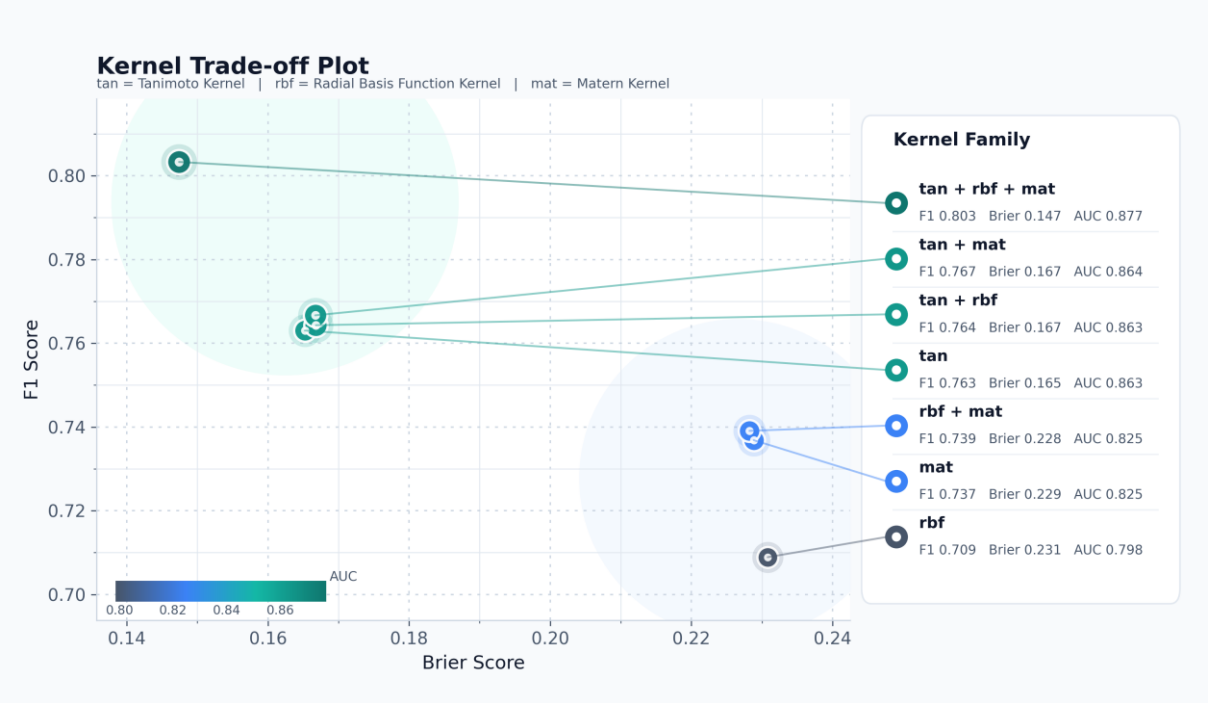

Figure 5. Trade-off between classification performance and probability calibration for individual kernels and their combinations. The Tanimoto, RBF, and Matérn kernel achieves the most favorable balance, achieving the highest F1-score and the lowest Brier score simultaneously.

4. Discussion

Having established the performance characteristics of the proposed composite kernel through ablation experiments in Section 3, we now contextualize these results through systematic comparison with established baseline methods and state-of-the-art approaches. To demonstrate the efficacy of the proposed Hybrid Gaussian Process (Hybrid GP) model, we compared its predictive performance against a suite of widely used statistical Bayesian classifiers and Gaussian Process baselines utilizing single-kernel configurations. The baseline models included Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Naïve Bayes, and Gaussian Processes using exclusively physicochemical (RBF) or structural (Tanimoto) kernels.

Table 3 summarizes the final performance metrics for all evaluated models on the test set.

MODEL	AUC	ACCURACY	F1-SCORE	BRIER SCORE	THRESHOLD
OURS(HYBRID GP)	0.8878	0.8292	0.8022	0.1350	0.4880
GP(RBF)	0.8443	0.7766	0.7581	0.2142	0.5000
GP(TANIMOTO)	0.8882	0.8184	0.7829	0.1413	0.5123

LDA	0.8857	0.8053	0.7847	0.1348	0.4530
QDA	0.8579	0.8065	0.7737	0.2208	0.0000
NAÏVE BAYES	0.8302	0.7790	0.7455	0.2062	0.0760

The experimental results indicate that the proposed Hybrid GP model achieved the highest Classification Accuracy (0.8292) and F1-Score (0.8022) among all tested methods. A detailed analysis of feature modalities reveals critical insights into model behavior, consistent with previous studies emphasizing the importance of diverse molecular representations [12,15]. First, while the structure-only GP (Tanimoto) model yielded a marginally higher AUC (0.8882), the Hybrid GP significantly outperformed it in terms of F1-Score (0.8022 vs. 0.7829). This suggests that incorporating physicochemical descriptors through the composite kernel allows the model to maintain a superior balance between precision and recall at the optimal decision threshold, rather than relying solely on structural similarity. Conversely, the GP (RBF) model, which utilized only continuous physicochemical descriptors, resulted in a lower AUC (0.8443) compared to the structure-inclusive models. This reaffirms that topological structural information is critical for robust toxicity prediction. However, the performance gain observed in the Hybrid model confirms that physicochemical properties provide a necessary refinement, offering synergistic value that enhances decision boundaries for compounds where structural fingerprints alone may be ambiguous [8]. Among the statistical baselines, LDA demonstrated competitive performance with a high AUC (0.8857) and the lowest Brier score (0.1348), indicating excellent probability calibration. However, it fell short of the Hybrid GP in terms of overall accuracy and F1-score. The marginally lower Brier score of LDA (0.1348 vs. 0.1350) does not constitute a statistically meaningful difference and reflects the well-known calibration advantage of linear discriminant models on near-Gaussian class-conditional distributions, rather than superior predictive utility. The remaining baselines, QDA and Naïve Bayes, exhibited consistently inferior predictive performance across all metrics, further underscoring the advantage of the Gaussian process-based approaches. Taken together, these results demonstrate that the Hybrid GP model successfully leverages the complementary nature of structural fingerprints and physicochemical descriptors, delivering the most robust and balanced classification performance among the evaluated methods. Figure 6 provides a visual summary of the comparative performance across all baseline models, illustrating the consistent superiority of the Hybrid GP approach across multiple evaluation metrics.

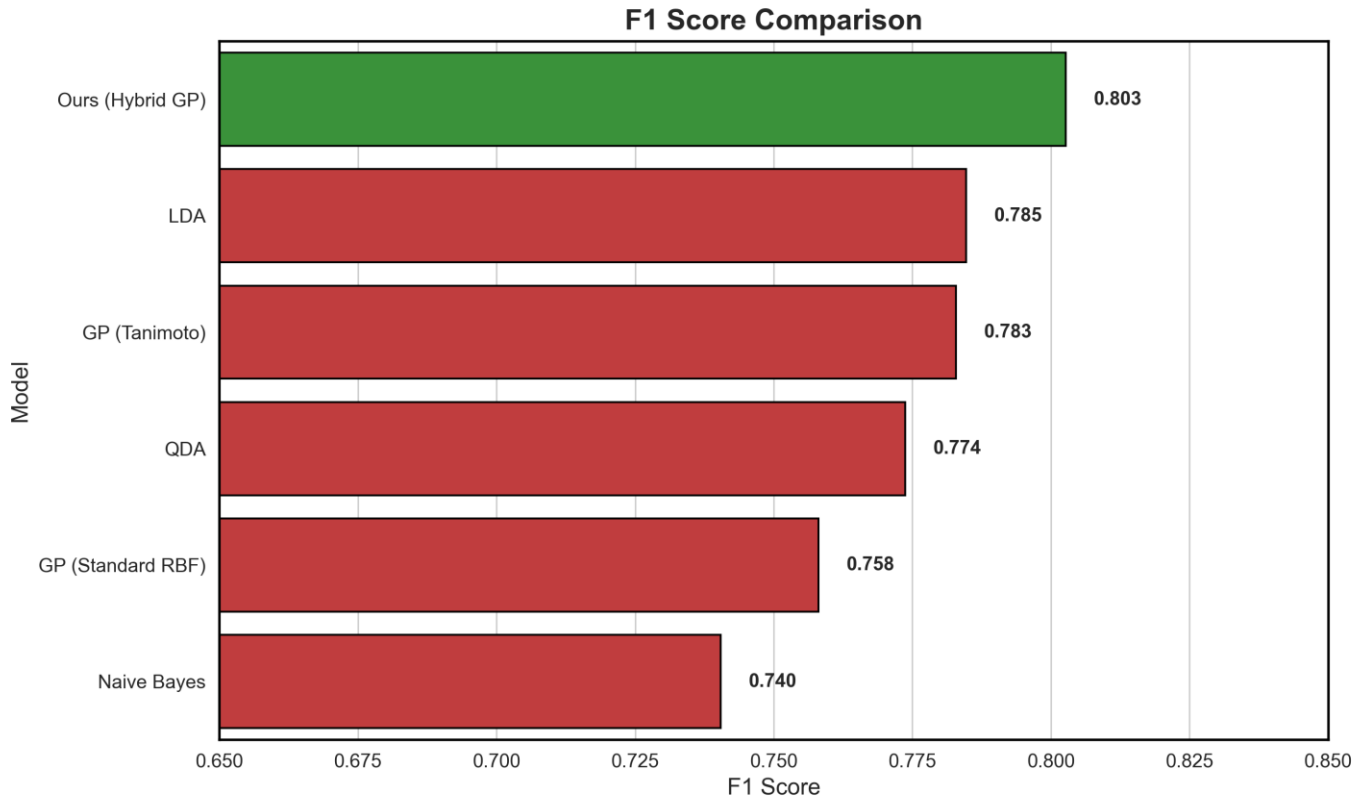

Figure 6. Comparative performance of the Hybrid GPC against baseline classifiers across AUC, accuracy, F1-score, and Brier score metrics.

To benchmark the performance of the proposed Hybrid GP model against current state-of-the-art methods, we compared our results with MolToxPred (Setiya et al., 2024), a recent study that also utilized the T3DB dataset for small molecule toxicity prediction [12]. Table 4 provides a direct methodological comparison between the two approaches. MolToxPred employed a stacked ensemble architecture combining Random Forest, Multilayer Perceptron (MLP), and LightGBM as base learners, with Logistic Regression serving as the meta-classifier. Similar to our approach, it integrated both structural fingerprints and physicochemical descriptors. Their study reported an AUROC of 0.8776 on the test set using a stratified cross-validation strategy [12]. In comparison, our Hybrid GP model achieved a higher AUROC of 0.8878. This performance gain is particularly notable given the more rigorous validation constraints employed in our study. While standard stratified cross-validation can inadvertently allow structural similarity between training and test sets, our model was evaluated using a scaffold-based split to strictly prevent structural data leakage. Detailed final test-set results, including the confusion matrix and threshold analysis, are documented in Supplementary Material, Section A.4. To further contextualize the performance of the proposed Hybrid GP model, Table 5 summarizes a broader comparison with representative methods from the literature. While ToxiM (Sharma et al., 2017) and ToxinPredictor (Goel et al., 2025) report numerically higher AUC values under random or cross-validation splitting, their evaluation protocols do not strictly prevent structural

overlap between training and test sets. In contrast, the scaffold-based split employed in this study provides a more conservative and realistically challenging generalization benchmark. Under these rigorous conditions, the Hybrid GP model achieves competitive performance (AUC = 0.8878, F1 = 0.8022) while additionally offering principled uncertainty quantification — a feature absent from the ensemble and SVM-based baselines. Furthermore, the proposed model outperforms MolToxPred (Setiya et al., 2024), the most directly comparable study utilizing the same T3DB dataset, both in AUC (0.8878 vs. 0.8776) and F1-score (0.8022 vs. 0.7643), despite applying the more rigorous scaffold-based evaluation protocol. It is also worth noting that complementary approaches based on graph neural networks have recently demonstrated strong performance for safety-critical physicochemical property prediction when augmented with DFT-derived features [24], and that multimodal fusion strategies integrating graph embeddings with parsimoniously selected descriptors have proven effective across diverse molecular property endpoints [25]. While these GNN-based frameworks excel in representation learning, the Hybrid GPC offers the distinct advantage of native probabilistic uncertainty quantification, which is not inherently available in deterministic GNN architectures. It is important to note that dataset size alone does not determine model reliability. Although ToxinPredictor (Goel et al., 2025) utilized a marginally larger dataset (14,064 molecules), the critical distinction lies in the evaluation rigor and data provenance. Our dataset was constructed from two experimentally validated, authoritative sources: the Tox21 benchmark—a reference standard jointly maintained by the National Institutes of Health (NIH), the Environmental Protection Agency (EPA), and the Food and Drug Administration (FDA) for toxicity screening—and the Toxin and Toxin Target Database (T3DB), which provides curated, experimentally confirmed toxicity annotations [21]. More importantly, the predictive performance reported under random splitting or cross-validation protocols cannot be directly compared with scaffold-based evaluation, as the former may inadvertently exploit structural similarity between training and test molecules, leading to inflated performance estimates that do not reflect genuine generalization to novel chemical entities encountered in real-world drug discovery scenarios [22,23].

MODEL	METHODOLOGY	SPLIT STRATEGY	AUC
OURS(HYBRID GP)	Gaussian Process (Composite Kernel)	Scaffold Split	0.8878
MOLTOXPRED (2024)	Stacked Ensemble (RF+MLP+LGBM)	Stratified 5-fold CV	0.8776

Table 5. Comparison with State-of-the-Art Methods

METHOD	MODEL TYPE	FEATURE REPRESENTATION	DATASET	SPLIT STRATEGY	AUC	ACCURACY	F1-SCORE
OURS (HYBRID GP)	Gaussian Process (Composite Kernel)	Morgan FP + RDKit Descriptors	Tox_11842 + T3DB (13,949)	Scaffold Split	**0.8878**	**0.8292**	**0.8022**
MOLTOXPRED (SETIYA ET AL., 2024)	Stacked Ensemble (RF+MLP+LightGBM)	Molecular Descriptors + Fingerprints	T3DB-based	Stratified 5-fold CV	0.8776	0.8091	0.7643
TOXINPREDICTOR (GOEL ET AL., 2025)	SVM (Boruta + PCA)	Molecular Descriptors	14,064 molecules	Random Split*	0.9170	0.8540	0.8490
ETOXPRED (PU ET AL., 2019)	Extra Trees (ET)	Morgan Fingerprints	FDA/TOXNET/T3DB (~10,000)	Cross-validation*	0.8200	0.7200	0.6593
TOXIM (SHARMA ET AL., 2017)	Random Forest	Descriptors + Fingerprints	T3DB-based	10-fold CV*	0.9700	0.9300	—

** Random split or cross-validation; structural leakage between train/test sets not strictly prevented.*

Bold values in the Hybrid GP row indicate the best performance under scaffold-based evaluation.

Beyond the numerical improvement in discriminative performance, the proposed Gaussian Process framework offers distinct qualitative advantages over the deterministic ensemble used in MolToxPred. Regarding probabilistic uncertainty quantification, unlike the point predictions typically generated by ensemble methods, our GP model provides intrinsic uncertainty estimates for every prediction. As demonstrated by the low Brier score (0.1350) reported in Section 3, our model yields well-calibrated probabilities that directly reflect predictive confidence. This capability is widely recognized as a critical requirement for regulatory decision-making and risk assessment in computational toxicology, where predictions without accompanying confidence measures are often insufficient for informing safety-critical decisions [14,17]. With respect to robust generalization, by validating performance under scaffold-splitting conditions, we have demonstrated that the Hybrid GP model is capable of generalizing to novel chemical scaffolds not encountered during training, rather than merely recognizing structurally similar compounds present in the training set. This property is of direct practical relevance, as toxicity models deployed in drug discovery pipelines must routinely evaluate molecules with unprecedented structural features that lie outside the

training distribution. Several limitations of this study should be acknowledged. The cubic computational complexity of exact Gaussian process inference limits scalability to very large molecular libraries; although the grid-based hyperparameter optimization on a subset of 1,000 samples mitigates this concern during model selection, full inference on datasets exceeding 100,000 molecules would require sparse approximation methods such as inducing point techniques. Additionally, the current framework addresses binary toxicity classification and does not distinguish among specific toxicity endpoints (e.g., hepatotoxicity, cardiotoxicity, mutagenicity), which would require multi-task or multi-label extensions. Moreover, while scaffold-based splitting provides a more rigorous evaluation than random splitting, it may still underestimate the challenge of predicting toxicity for molecules with entirely novel pharmacophores not represented in existing databases. The complete implementation, including data processing scripts and model training code, is available as described in Supplementary Material, Section B. Future work addressing these limitations through sparse GP methods, multi-endpoint modeling, and temporal validation on prospective datasets would further strengthen the practical applicability of this framework.

5. Conclusion

This study demonstrates that Gaussian process classification, when equipped with a domain-informed composite kernel, can serve as a principled and competitive alternative to ensemble and deep learning methods for molecular toxicity prediction—while uniquely providing calibrated uncertainty estimates for every prediction. To our knowledge, the proposed Hybrid GPC represents the first integration of a Tanimoto kernel for binary structural fingerprints with RBF and Matérn kernels for continuous physicochemical descriptors within a unified Bayesian classification framework, achieving an AUROC of 0.8878 and an F1-score of 0.8022 under rigorous scaffold-based evaluation. The systematic ablation experiments reported here offer several insights that extend beyond the specific model proposed. The dominance of the Tanimoto kernel component across all configurations underscores that structural similarity remains the single most informative signal for toxicity classification, yet the consistent improvement observed upon incorporating physicochemical descriptors confirms that these two modalities capture genuinely complementary aspects of molecular toxicity—the former encoding substructural motifs associated with toxic liability, the latter reflecting global properties such as lipophilicity and molecular size that govern bioavailability and target engagement. Equally important is the finding that the composite kernel achieves Pareto-optimal behavior in the F1–Brier score trade-off space, demonstrating that discriminative power and probability calibration need not be competing objectives when kernel components are deliberately matched to the underlying data modalities. The scaffold-based evaluation protocol adopted in this work further reveals a non-trivial gap between the performance levels reported under random splitting in prior studies and the more conservative estimates obtained when structural leakage is strictly prevented, reinforcing the broader call for evaluation rigor in molecular property prediction research.

From a practical standpoint, the intrinsic uncertainty quantification offered by the Gaussian process framework addresses a critical unmet need in computational toxicology. In regulatory and pharmaceutical contexts, a toxicity prediction accompanied by a calibrated confidence measure enables risk-stratified decision-making: high-confidence predictions can accelerate compound prioritization, while low-confidence cases can be flagged for targeted experimental validation—thereby reducing both the cost and the failure rate of preclinical safety assessment. Looking forward, the natural next step is to extend this framework toward multi-task Gaussian process models capable of simultaneously predicting multiple toxicity endpoints by exploiting inter-endpoint correlations through shared latent functions. Active learning strategies guided by the model's posterior uncertainty represent another promising avenue, enabling iterative expansion of training data with maximally informative compounds. More broadly, hybrid architectures that combine the representational power of graph neural networks [24,25] with the probabilistic rigor of Gaussian processes could unlock scalability to million-compound libraries through sparse variational approximations, paving the way for deployment in large-scale virtual screening campaigns.

Acknowledgement: The first and second authors equally contributed to this study. This work was supported by the Sogang University Research Grant.

References

[1] Bai, Y., et al. (2025). Machine learning-enabled drug-induced toxicity prediction. Advanced Science, 12, e2413405. https://doi.org/10.1002/advs.202413405

[2] Sun, D., Gao, W., Hu, H., & Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it? Acta Pharmaceutica Sinica B, 12(7), 3049–3062. https://doi.org/10.1016/j.apsb.2022.02.002

[3] Cavasotto, C. N., & Scardino, V. (2022). Machine learning toxicity prediction: Latest advances by toxicity end point. ACS Omega, 7(51), 47536–47546. https://doi.org/10.1021/acsomega.2c05693

[4] Guo, W., Liu, J., Dong, F., Song, M., Li, Z., Khan, M. K. H., Patterson, T. A., & Hong, H. (2023). Review of machine learning and deep learning models for toxicity prediction. Experimental Biology and Medicine, 248(21), 1952–1973. https://doi.org/10.1177/15353702231209421

[5] Seal, S., Mahale, M., García-Ortegón, M., Joshi, C. K., Hosseini-Gerami, L., Beatson, A., … & Bender, A. (2025). Machine learning for toxicity prediction using chemical structures: Pillars for success in the real world. Chemical Research in Toxicology, 38(5), 759–807. https://doi.org/10.1021/acs.chemrestox.5c00033

[6] Bento, A. P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., … & Leach, A. R. (2020). An open source chemical structure curation pipeline using RDKit. Journal of Cheminformatics, 12(1), 51. https://doi.org/10.1186/s13321-020-00456-1

[7] Born, J., Markert, G., Janakarajan, N., Kimber, T. B., Volkamer, A., Martínez, M. R., & Manica, M. (2023). Chemical representation learning for toxicity prediction. Digital Discovery, 2(3), 674–691. https://doi.org/10.1039/D2DD00099G

[8] Orosz, Á., Héberger, K., & Rácz, A. (2022). Comparison of descriptor- and fingerprint sets in machine learning models for ADME-Tox targets. Frontiers in Chemistry, 10, 852893. https://doi.org/10.3389/fchem.2022.852893

[9] Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: Toxicity prediction using deep learning. Frontiers in Environmental Science, 3, 80. https://doi.org/10.3389/fenvs.2015.00080

[10] Pu, L., Naderi, M., Liu, T., Wu, H. C., Mukhopadhyay, S., & Brylinski, M. (2019). eToxPred: A machine learning-based approach to estimate the toxicity of drug candidates. BMC Pharmacology and Toxicology, 20(1), 2. https://doi.org/10.1186/s40360-018-0282-6

[11] Sharma, A., Sharma, A., & Kumar, S. (2017). ToxiM: A toxicity prediction tool for small molecules developed using machine learning and chemoinformatics approaches. Frontiers in Pharmacology, 8, 880. https://doi.org/10.3389/fphar.2017.00880

[12] Setiya, A., Jani, V., Sonavane, U., & Joshi, R. (2024). MolToxPred: Small molecule toxicity prediction using machine learning approach. RSC Advances, 14(6), 4201–4220. https://doi.org/10.1039/D3RA07322J

[13] Goel, M., Amawate, A., Singh, A., & Bagler, G. (2025). ToxinPredictor: Computational models to predict the toxicity of molecules. Chemosphere, 370, 143900. https://doi.org/10.1016/j.chemosphere.2024.143900

[14] Ji, C., Svensson, F., Zoufir, A., & Bender, A. (2018). eMolTox: Prediction of molecular toxicity with confidence. Bioinformatics, 34(14), 2508–2509. https://doi.org/10.1093/bioinformatics/bty135

[15] Seal, S., Yang, H., Trapotsi, M.-A., Singh, S., Carreras-Puigvert, J., Spjuth, O., & Bender, A. (2023). Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data. Journal of Cheminformatics, 15(1), 56. https://doi.org/10.1186/s13321-023-00723-x

[16] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402–6413.

[17] Yang, C.-I., & Li, Y.-P. (2023). Explainable uncertainty quantifications for deep learning-based molecular property prediction. Journal of Cheminformatics, 15, 13. https://doi.org/10.1186/s13321-023-00682-3

[18] Semenova, E., Williams, D. P., Afzal, A. M., & Lazic, S. E. (2020). A Bayesian neural network for toxicity prediction. Computational Toxicology, 16, 100133. https://doi.org/10.1016/j.comtox.2020.100133

[19] Deringer, V. L., Bartók, A. P., Bernstein, N., Wilkins, D. M., Ceriotti, M., & Csányi, G. (2021). Gaussian process regression for materials and molecules. Chemical Reviews, 121(16), 10073–10141. https://doi.org/10.1021/acs.chemrev.1c00022

[20] Yin, T., Panapitiya, G., Coda, E. D., et al. (2023). Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction. Journal of Cheminformatics, 15, 105. https://doi.org/10.1186/s13321-023-00753-5

[21] Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., … & Scalbert, A. (2018). HMDB 4.0: The human metabolome database for 2018. Nucleic Acids Research, 46(D1), D608–D617. https://doi.org/10.1093/nar/gkx1089

[22] Hu, Y., Sheridan, P., Gobbi, A., & Deb, S. (2023). A systematic study of key elements underlying molecular property prediction. Nature Communications, 14, 6395. https://doi.org/10.1038/s41467-023-41948-6

[23] Dablander, M., Hanser, T., Lambiotte, R., & Morris, G. M. (2025). Coverage bias in small molecule machine learning. Nature Communications, 16, 298. https://doi.org/10.1038/s41467-024-55462-w

[24] Lee, S., Lee, J., Yoon, U., Koo, J., Yoon, Y. W., Cho, Y., Hwang, S.-R., & Jeong, K. (2026). Advancing chemical safety prediction: an integrated GNN framework with DFT-augmented cyclic compound solution. Journal of Cheminformatics, 18, 28. https://doi.org/10.1186/s13321-026-01151-3

[25] Jang, Y., Lee, J., Jeong, K., & Kim, J. (2026). Multimodal graph fusion with statistically guided parsimonious descriptor selection for molecular property prediction. Journal of Cheminformatics, 18, 18. https://doi.org/10.1186/s13321-025-01140-y