A Compact QSAR Model for BRAF Inhibitors Based on Gaussian Process Regression

Elisa Y. Toma and Marcos G. Quiles(✉)

Institute of Science and Technology, Federal University of São Paulo, Av. Cesare Lattes, 1201, 12247-014 São José dos Campos, SP, Brazil
quiles@unifesp.br

Abstract. The dysregulated activation of the B-Raf proto-oncogene serine/threonine-protein kinase (BRAF) kinase acts as a primary driver in melanoma and several other malignancies. As a result, there is an ongoing need for computational approaches capable of precisely evaluating the inhibitory potency, typically expressed as the negative logarithm of the half-maximal inhibitory concentration(pIC50) of prospective drug candidates. Recently, the field has observed a trend of stacking-based ensemble methods for predicting this activity. Although these stack-based approaches deliver predictions with high accuracy, such performance comes with inherent drawbacks. Specifically, their highly complex architectures increase computational expenses and obscure model interpretability. Here, to address these drawbacks, we propose an uncertainty-aware Gaussian Process Regression (GPR) framework that rivals the predictive power of complex ensembles without the structural overhead. To construct a more comprehensive feature space, we integrate two distinct types of molecular data. Particularly, the model combines a minimally redundant set of physicochemical descriptors with latent molecular embeddings extracted from Simplified Molecular Input Line Entry System (SMILES) strings. Computer evaluation, spanning cross-validation, external testing, robustness checks, and a paired Wilcoxon signed-rank test, confirms that the proposed GPR model achieves an external predictive accuracy on par with stacking techniques. Beyond its compact design, the model also yields well-calibrated predictive uncertainties, which provide a rigorous, transparent mechanism to delimit the applicability domain concerning its predictions. Our findings suggest that advancing the state of the art in quantitative structure–activity relationship (QSAR) modeling does not, inherently, require complex architectures. As shown, a simpler probabilistic model can achieve competitive accuracy while providing superior transparency.

1 Introduction

Several cellular events, such as growth, survival, and proliferation, are regulated by the mitogen-activated protein kinase (MAPK) cascade. Mutations within

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 493–510, 2027.
https://doi.org/10.1007/978-3-032-30488-9_31

the BRAF (B-Raf proto-oncogene serine/threonine-protein kinase) frequently disrupt this network and might trigger persistent downstream signaling and sustained proliferative stimuli that ultimately contribute to tumorigenesis.

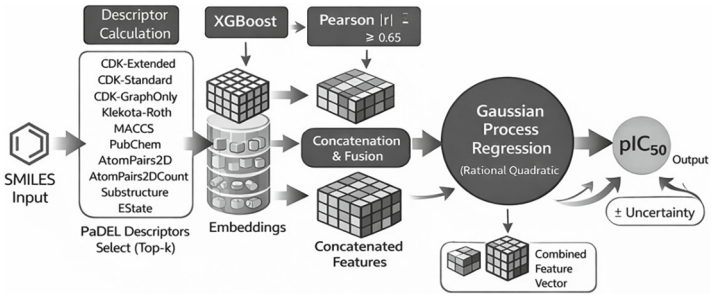

Fig. 1. Architecture of the proposed Gaussian Process Regression model.

Activating BRAF mutations present an overall frequency of approximately 7% across all human cancers. Among distinct cancers, melanoma exhibits the highest incidence of this genetic profile (40%), followed by colorectal adenocarcinoma (15%), glioblastoma (8%), and non-small cell lung cancer (4%). Given this pathogenic impact, researchers have searched for therapeutic interventions targeting BRAF and its downstream MAPK effectors. To date, these targeted approaches have yielded substantial clinical benefits in managing multiple BRAF-mediated tumors [4,15].

Predicting the half-maximal inhibitory concentration (pIC50) is therefore a critical objective in the discovery of novel BRAF inhibitors. To predict pIC50 values against BRAF, recent computational efforts have frequently relied on stacking-based ensemble models. This approach combines multiple base models trained on several molecular fingerprints. Some examples include StackBRAF and BRAFPred, both of which utilize complex ensemble architectures to integrate the predictive signals of heterogeneous algorithms [18,25].

StackBRAF [18], for instance, trains multiple base learners independently, including Support Vector Regression (SVR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP) neural networks on classical molecular fingerprints. The outputs of these models feed a secondary meta-model for final aggregation. Despite their efficacy, these frameworks impose severe architectural complexity. They force the predictive system to rely on indirect, model-level feature aggregation. Pooling multiple learners certainly allows architectures like StackBRAF [18] to reduce prediction variance. However, such performance improvements arise from synthesizing correlated models within a static descriptor space, which seldom yield novel chemical insights.

In contrast, our approach hypothesizes that high structural complexity is not necessary for robust predictions in this domain. Instead, we argue that predictive accuracy is primarily constrained by the depth of the underlying data rather

than by the sophistication of the ensemble technique. Following the principle of parsimony, if a simpler, more mathematically transparent formulation can match the accuracy of a stacked ensemble, it should be heavily preferred [16].

Here, we prioritize the enrichment of the input space itself. This strategy allows for the direct mapping of structure–activity relationships within a broader, more robust representation. We begin by calculating physicochemical descriptors with Pharmaceutical Data Exploration Laboratory (PaDEL) [24]. An XGBoost algorithm then evaluates and isolates the most informative features. To resolve the observed collinearity, we subject this selected subset of features to a Pearson correlation filter. Next, we merge the final set of features with the latent molecular embeddings generated from SMILES strings [22]. This synthesis of symbolic and learned representations ultimately underpins our unified Gaussian Process Regression (GPR) model. The proposed GPR framework is depicted in Fig. 1.

By preserving only essential structural features and eliminating redundancy, the proposed framework better controls the bias–variance trade-off. It also yields a streamlined, highly interpretable framework equipped for robust uncertainty quantification. It is worth noting that its predictive accuracy remains statistically equivalent to that of heavily resourced stacking models.

2 Materials and Methods

2.1 Benchmark Dataset

Chemical European Molecular Biology Laboratory (ChEMBL) is an open-access bioactivity database with several properties, such as binding, functional, and pharmacokinetic information. It covers absorption, distribution, metabolism, excretion, and toxicity (ADMET) for millions of compounds with measured activity data across thousands of protein targets, including inhibitory concentration values such as IC_{50} and derived pIC_{50} values. To focus on compounds within a physicochemically relevant space for intracellular targets, molecules with a mass greater than 700 Da and a octanol–water partition coefficient (logP) larger than 8 were ignored from the dataset [18].

To maintain rigorous comparability with current benchmarks, we utilized the same dataset as the StackBRAF [18] study. This corpus contains 3,854 compounds characterized by their BRAF pIC_{50} values. Following the original experimental configuration in [18], the data were partitioned into a training set of 2,697 molecules and an independent test set of 1,157 compounds.

In addition to the ChEMBL dataset, a dataset of Food and Drug Administration, FDA-approved compounds was also considered [18]. The FDA dataset consists of 2,123 compounds represented by canonical isomeric SMILES. This dataset was used exclusively for external prediction and qualitative analysis. These molecules were not considered in the training or validation steps.

2.2 Molecular Descriptors

Molecules were represented via canonical SMILES, and descriptors were extracted using the PaDEL-Descriptor software, following the same protocol

adopted in the StackBRAF [18] study to ensure direct methodological comparability.

The molecular representation employed in this work consisted of twelve PaDEL fingerprint types, which can be grouped according to the type of structural information they encode. Altogether, these fingerprints result in approximately 16,000 molecular features per compound prior to feature selection. All descriptors were computed using default PaDEL settings, providing a consistent and reproducible structural representation aligned with previously published benchmark models [20,24].

2.3 Supervised Feature Selection

To reduce dimensionality and retain only activity-relevant information, supervised feature selection was applied to the PaDEL descriptors. Feature selection was first performed using an Extreme Gradient Boosting (XGBoost) [8] regressor trained exclusively on the training set, thereby preventing information leakage from the external test set. We set the XGBoost model with 600 trees, a maximum depth of 12, a subsampling ratio of 0.9, a learning rate of 0.03, a column subsampling ratio of 0.8, and a squared error loss function. We assessed feature importance through the gain metric, which captures how much each individual variable minimizes prediction error throughout the ensemble.

Based on the ranked gain values, the top 1,800 PaDEL descriptors were initially selected. To further remove redundancy and mitigate multicollinearity, a second filtering step based on pairwise Pearson correlation was applied to the selected subset. Descriptor pairs presenting a correlation coefficient were considered highly correlated, and for each such pair, the descriptor with lower importance according to the XGBoost gain ranking was discarded, ensuring the retention of the most informative and non-redundant features.

Importantly, feature selection was not applied to the learned molecular embeddings, which were retained in their original latent dimensionality to preserve their data-driven representation capacity.

2.4 Molecular Embeddings

SMILES strings were tokenized at the character level and padded to the maximum sequence length. Molecular embeddings were learned through a self-supervised contrastive learning framework that exploits the intrinsic non-uniqueness of SMILES representations as a source of sequence invariance. Random SMILES augmentation was generated using RDKit, an open-source cheminformatics toolkit [1], two augmented SMILES views were produced using non-canonical atom ordering.

The encoder architecture consisted of an embedding layer with 56 dimensions, followed by a bidirectional Gated Recurrent Unit (GRU) [3,14] with 136 hidden units per direction and a unidirectional GRU with 98 hidden units. The encoder outputs a 98-dimensional latent representation of the molecule. During contrastive training, a projection head (98-136-136) with Rectified Linear Unit

(ReLU) activation and L2 normalization was attached to the encoder, and the projected representations were optimized using the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss, a contrastive learning objective designed to encourage similar samples to have closer representations in latent space while pushing dissimilar samples apart.

The embedding network was optimized using the NT-Xent contrastive loss with a temperature parameter of $\tau = 0.1$.

Training was executed over 200 epochs with a batch size of 128, utilizing the Adam optimizer at a learning rate of 0.001. Once complete, the projection head was removed, leaving the encoder to generate invariant 98-dimensional molecular embeddings for all compounds.

2.5 Feature Fusion

For each molecule, the final input representation was constructed by concatenating the 98-dimensional SMILES-based molecular embedding with the selected PaDEL descriptors. All features were standardized using z-score normalization parameters computed from the training set only. No feature selection or dimensionality reduction was applied to the molecular embeddings.

2.6 Gaussian Process Regression

A single Rational Quadratic kernel [12] was adopted to model the combined feature space, as this kernel can be interpreted as a scale mixture of squared exponential kernels and is therefore well suited to capturing variability across features with heterogeneous characteristic length scales. pIC_{50} prediction was performed using Gaussian Process Regression (GPR) with a kernel defined as a constant-scaled Rational Quadratic component plus a white noise term. Gaussian Process Regression models the target property as a distribution over functions:

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}),\, k(\mathbf{x}, \mathbf{x}')\big) \tag{1}$$

where $m(x)$ denotes the mean function and $k(x, x')$ represents the covariance kernel defining similarity between input vectors. A zero-mean prior was assumed, and the target variable was centered prior to model fitting, i.e., transformed to have zero empirical mean. This is equivalent to modeling deviations from the empirical mean.

The Rational Quadratic (RQ) kernel is defined as:

$$k_{RQ}(\mathbf{x}, \mathbf{x}') = \sigma^2 \left(1 + \frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{2}$$

where σ^2 is the signal variance, ℓ is the characteristic length scale, and controls the relative weighting of large- and small-scale variations. This formulation allows the kernel to model functions with multiple characteristic length scales, making it particularly suitable for heterogeneous feature spaces.

The initial hyperparameters and bounds were set as follows: constant scale (initial value 10^{-3}, bounds $[10^{-6}, 10^{3}]$), length scale (initial value 10^{-3}, bounds $[10^{-6}, 10^{3}]$), shape parameter α (initial value 1.0, bounds $[10^{-3}, 10^{4}]$), and noise level (initial value 10^{-3}, bounds $[10^{-6}, 10^{2}]$).

Final model assessment relied on an external test set, evaluated through the coefficient of determination (R^2), root mean squared error (RMSE), mean squared error (MSE), and the Pearson correlation coefficient (r). Beyond standard metrics, we recorded and scrutinized predictive uncertainties extracted from the GPR posterior distribution.

2.7 Model Validation and Performance Metrics

The performance of the proposed regression models was evaluated using the coefficient of determination (R^2) and the predictive squared correlation coefficient (Q^2), which are standard metrics in QSAR modeling. Model fit was assessed on the training set using R^2_train, which quantifies the proportion of variance in experimental pIC_{50} values explained by the model [19].

Internal predictive performance was evaluated by 5-fold cross-validation, reported as Q^2_cv, which measures the model's stability and generalization ability within the training domain. External predictivity was assessed using an independent test set not involved in model training or hyperparameter optimization, reported as Q^2_ext, providing an estimate of predictive performance on novel compounds. Model robustness was assessed by jointly analyzing R^2_train, Q^2_cv, and Q^2_ext.

2.8 Non-parametric Model Comparison via the Wilcoxon Signed Rank Test

Differences in predictive performance between competing models were assessed using the non-parametric statistical test for paired samples, Wilcoxon signed-rank test. This statistical test is specifically designed for paired comparisons and evaluates whether the median difference between two related samples is significantly different from zero. [6] Formally, let $d_i = e_i^{(1)} - e_i^{(2)}$ denote the paired differences between absolute prediction errors of the two models for compound i. The Wilcoxon signed-rank test ranks the absolute values $|d_i|$, assigns signs according to the direction of the difference, and computes the test statistic:

$$W = \sum_{i=1}^{n} \text{sign}(d_i)\, R_i \tag{3}$$

where R_i is the rank of $|d_i|$ among non-zero differences. Under the null hypothesis that the median difference is zero, W follows a known reference distribution used to compute the p-value.

Here, we used the Wilcoxon signed-rank test to compare the predictive errors of the StackBRAF [18] baseline model and the proposed GPR-based model on the same external test set comprising 1,157 compounds. For each compound,

absolute prediction errors were computed for both models, and these paired absolute errors were used as input to the statistical test.

We employed a two-sided formulation for all statistical evaluations, maintaining a significance threshold (α) of 0.05.

Unless otherwise stated, all architectural choices, model configurations, and hyperparameter values reported in this section were empirically determined based on preliminary experiments conducted exclusively on the training data.

2.9 Applicability Domain

The applicability domain (AD) of a predictive model defines the region of the descriptor space in which the model can be expected to produce reliable predictions. In this work, the applicability domain of the proposed GPR model was quantitatively defined based on the uncertainty distribution of the training set. Specifically, the maximum admissible uncertainty was defined as the 90th percentile of the predictive standard deviation obtained from cross-validated predictions on the training set.

Beyond the definition of the threshold, complementary uncertainty-based diagnostics were employed to characterize and validate the applicability domain. First, the relationship between absolute prediction error and predicted uncertainty was quantified using the Spearman rank correlation coefficient on the external test set.

Second, a coverage-based analysis was performed by ranking test samples according to increasing predictive uncertainty and progressively evaluating model performance as a function of retained coverage. The resulting coverage versus RMSE curve provides insight into how predictive accuracy degrades as predictions extend toward more uncertain regions of the descriptor space. A stable RMSE at high coverage, followed by a pronounced increase at lower coverage levels, is indicative of a well-defined applicability domain in which reliable predictions are concentrated within low-uncertainty regions.

Finally, probabilistic calibration of the GPR model was assessed by comparing nominal and empirical coverage of predictive confidence intervals at multiple confidence levels (e.g., 68%, 95%, and 99%). Agreement between expected and observed coverage supports the statistical validity of the predictive uncertainty estimates and reinforces their suitability as a criterion for defining the applicability domain [12].

2.10 Y-randomization

The y-randomization (also referred to as y-scrambling) test is a widely accepted validation procedure in QSAR modeling, employed to assess whether the observed predictive performance of a model arises from genuine structure–activity relationships or from chance correlations. In this procedure, the response variable (pIC_{50} values) is randomly permuted while the molecular representations are kept unchanged, and new models are trained using the same regression modeling procedure applied to the original dataset [13].

A total of 20 independent random permutations were performed. This procedure ensures that any differences in performance arise solely from the disruption of the structure–activity relationship rather than from changes in model configuration or training conditions.

We evaluate the model's robustness by comparing its predictive performance against the original model trained on randomized responses. The evaluation focused on the coefficient of determination (R^2). Robust QSAR models are expected to yield R^2 values close to zero when trained on randomized responses, indicating the absence of systematic predictive patterns.

Additionally, the cRp^2 metric, commonly used in Y-randomization tests to assess model robustness, was also considered as a complementary criterion to assess the statistical significance.

3 Results

3.1 Descriptor Selection and Correlation Filtering Analysis

The supervised feature selection procedure was performed in two sequential stages: (i) importance-based filtering using XGBoost and (ii) redundancy removal using Pearson correlation analysis.

After XGBoost ranking, we retain only the top 1,800 descriptors. The selected features were derived from CDK-based descriptor families, particularly CDK_Extended (544), CDK_Standard (458), and CDK_GraphOnly (336), which together accounted for approximately 74% of the retained variables and nearly 69% of the cumulative importance score. Subsequently, a Pearson correlation filter ($r \geq 0.65$) was applied to eliminate redundant information. This last selection step reduced the descriptor set from 1,800 to 1,319 features.

The most significant reduction occurred within the CDK_GraphOnly descriptors, signaling significant internal correlation among these topological feature subsets. As shown in Table 1, the distribution of descriptor origins confirms that the fundamental structural profile of the feature space remained stable despite the redundancy elimination.

3.2 Model Performance and Comparison with StackBRAF

Model performance was assessed using the coefficient of determination (R^2/Q^2) and mean absolute error (MAE) for training, cross-validation, and external test sets, following the same evaluation protocol adopted for the StackBRAF [18] model. StackBRAF [18] showed excellent training performance (R^2_train = 0.9718, MAE_train = 0.1628) and strong external prediction (R^2_test = 0.8399, MAE_test = 0.3946).

The proposed GPR-based model achieved similarly strong results, with high training fit (R^2_train = 0.9814, MAE_train = 0.1307), robust internal validation (Q^2_cv = 0.8003, MAE_cv = 0.4352), and strong external performance (Q^2_ext = 0.8401, MAE_ext = 0.3978). These values closely match those obtained for StackBRAF [18].

Table 1. Distribution of molecular descriptors retained after filtering

Origin	Number of Features	Sum of Importance
CDK Extended	491	0.2139
CDK Standard	400	0.1946
CDK GraphOnly	204	0.1822
PubChem	54	0.0562
Klekota Roth	58	0.0372
Klekota Roth Count	35	0.0241
Substructure	7	0.0135
Atom Pairs 2D Count	24	0.0109
Substructure Count	5	0.0096
MACCS	10	0.0087
Atom Pairs 2D	27	0.0083
EState	4	0.0042

According to commonly adopted QSAR validation criteria, Q^2 values above approximately 0.6 are indicative of predictive models, although this threshold may vary depending on dataset size and modeling context. In addition, the difference between training and validation metrics ($\Delta(R^2_{\text{train}} - Q^2)$), whether cross-validated or external) is frequently used as an indicator of overfitting, with small differences ($\leq\sim 0.2 - 0.3$) reflecting consistency between training and validation performance. Overall, both models demonstrate stable and coherent performance across training, validation, and external test sets, supporting their robustness and generalizability [11,19].

3.3 Applicability Domain and Uncertainty Analysis

The applicability domain (AD) of the proposed GPR model was evaluated using uncertainty-aware diagnostics derived from its predictive variance (σ). In this work, the AD was quantitatively defined based on the uncertainty distribution of the training set, where the maximum admissible uncertainty was set as the 90th percentile of the predictive standard deviation obtained from cross-validated predictions on the training set. Predictions associated with uncertainty values below this threshold were therefore considered to lie within the model's applicability domain [7].

A statistically significant positive Spearman rank correlation ($\rho = 0.2990$, $p = 2.50 \times 10^{-25}$) was observed between the predictive standard deviation (σ) and the absolute prediction error on the external test set, indicating that higher uncertainty values are systematically associated with larger prediction errors. Although the magnitude of the correlation is moderate, its high statistical significance supports the consistency and practical utility of the uncertainty estimates for reliability assessment.

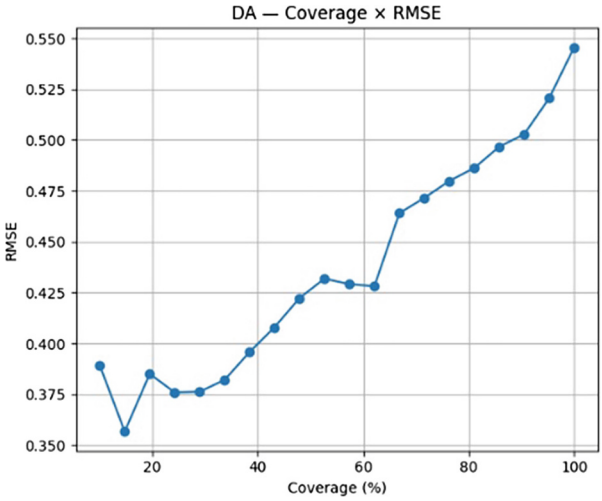

Fig. 2. Root mean squared error (RMSE) as a function of coverage level, with samples ordered by predictive uncertainty.

A coverage-based analysis further demonstrates the practical relevance of uncertainty estimates. As predictions are progressively restricted to compounds with lower uncertainty (Fig. 2), the RMSE decreases monotonically, indicating that excluding high-uncertainty predictions leads to improved accuracy. This behavior indicates that σ provides a meaningful basis for defining confidence-based applicability thresholds.

Model calibration was assessed by comparing expected and observed coverage probabilities for confidence intervals of 68%, 95%, and 99%. As depicted in Table 2, the observed coverages (approximately 74.85%, 95.51%, and 98.53%, respectively) nearly match their theoretical expectations, suggesting well-calibrated uncertainty estimates. [21]

The parity plot between experimental and predicted pIC_{50} values (Fig. 3) shows a high linear correlation across the full activity range, with no systematic bias at low or high potency values. Deviations from the ideal trend are primarily associated with predictions of higher uncertainty, reinforcing the role of σ as an indicator of extrapolation beyond the model's effective domain.

Table 2. Calibration of predictive uncertainty for the external test set

Confidence-Interval	Expected-Coverage	Observed-Coverage
68%	0.68	0.7485
95%	0.95	0.9551
99%	0.99	0.9853

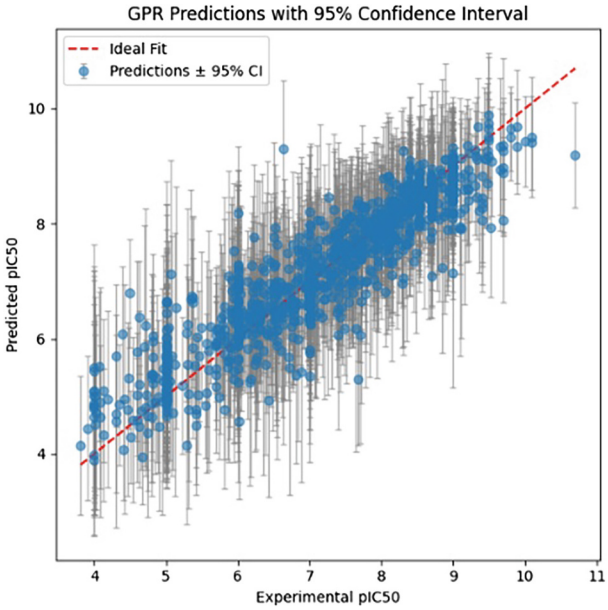

Fig. 3. Regression plot of predicted versus experimental pIC_{50} values for the external test set using the Gaussian Process Regression (GPR) model.

3.4 External Performance Evaluation on FDA-Approved Drugs

We assessed the external performance of the GPR-based QSAR model using a curated collection of FDA-approved drugs. The FDA dataset was used as provided in the Supporting Information of the StackBRAF [18] study, containing curated canonical isomeric SMILES representations of approved compounds.

To evaluate model behavior on chemically relevant compounds, a structural similarity filter retained only those with a Tanimoto coefficient exceeding 0.4 relative to the training set. Structural similarity was computed using Morgan fingerprints (radius = 2, 2048 bits), and the maximum Tanimoto similarity between each FDA compound and the training molecules was used as the filtering criterion. This threshold was intentionally set lower than that used in StackBRAF [18] to allow for a broader exploration of the chemical space while still avoiding extreme extrapolation [9].

To ensure compatibility with the trained model pipeline, the feature-selection procedure based on XGBoost importance ranking was reconstructed using the original training dataset. This step allowed the same subset of PaDEL descriptors used during model development to be consistently applied to the FDA compounds.

Predictions of pIC_{50} values were generated using the trained Gaussian Process Regression (GPR) model, which provides both point estimates and predictive uncertainties. In addition, an Applicability Domain (AD) analysis was

Table 3. Predicted pIC_{50} values for selected FDA approved drugs compared with experimental literature ranges

Compound	ChEMBL ID	Predicted pIC_{50} ($\mu \pm \sigma$)	LCB_{95}	Experimental pIC_{50}	Reference
Regorafenib	1946170	6.85 ± 1.08	4.73	∼6.0–7.0	[23]
Ponatinib	1171837	6.78 ± 1.06	4.69	∼7–9	[5]
Sorafenib	1336	6.67 ± 1.03	4.66	∼6.5–6.7	[2]
Vemurafenib	1229517	6.61 ± 1.06	4.53	∼7.0	[26]
Cabozantinib	2105717	6.58 ± 1.06	4.49	∼6.3–7.1	[17]

performed to evaluate the reliability of external predictions. The AD threshold was defined as the 90th percentile of the predictive standard deviation (σ) calculated on the training set, and FDA compounds with σ values below this threshold were considered to be within the model's reliable prediction space. This procedure ensured consistency with the internal validation strategy and prevented overinterpretation of predictions associated with unusually high uncertainty. To enable a conservative and statistically robust ranking, predictions were analyzed using the 95% Lower Confidence Bound (LCB_{95}), defined as:

$$\mathrm{LCB}_{95} = \mu_{\mathrm{pred}} - 1.96\,\sigma_{\mathrm{pred}} \tag{4}$$

where is the predicted pIC_{50}, and what is the associated predictive standard deviation. This uncertainty-aware criterion penalizes highly uncertain predictions, ensuring that compounds are prioritized based on both predicted potency and reliability.

We selected the top-5 compounds from the FDA dataset based on LCB_{95} scores, prioritizing compounds with known BRAF-inhibitory activity. We also checked both the global rankings (post-Tanimoto filtering) and the AD-restricted rankings, confirming that the top inhibitors fall within the model's reliable prediction space. Table 3 details the predicted pIC_{50} values, their uncertainties, and the corresponding LCB_{95} scores.

Overall, these outcomes show that the proposed model can produce biologically meaningful and statistically consistent pIC_{50} predictions for FDA-approved drugs with known BRAF inhibitory activity. Notably, by combining confidence–based ranking strategy with an explicit applicability-domain assessment, we prevents overinterpretation of high but uncertain predictions and provides a more reliable framework for external performance evaluation. Such an uncertainty-conscious perspective represents a core strength of the GPR framework, reinforcing its utility for practical decision-making in kinase inhibitor discovery.

3.5 Robustness and Statistical Validation

To evaluate the model's robustness, we used a y-randomization test. The training-set pIC_{50} values were randomly permuted while preserving molecular descriptors and representations. After generating 20 random permutations, the

regression modeling workflow was repeated for each instance using the exact configuration from the original dataset.

For randomized responses, the model produced R^2 values near zero, which contrasted with the original R^2 of about 0.84. This result indicates that the model relies on genuine structure-activity relationships instead of spurious correlations. We observe the same behavior for the cR_p^2 value (≈ 0.64), which exceeds the standard validation threshold of 0.5. This finding further substantiates the overall robustness of the proposed framework.

A paired Wilcoxon signed-rank test was conducted to compare the absolute prediction errors of StackBRAF [18] and the proposed GPR-based model on the same external test set (n = 1157). Both models achieved an identical mean MAE of 0.395. The resulting p-value (0.570) indicates that there is no statistically significant difference between their error distributions.

As shown in Fig. 4, the boxplot distributions of absolute prediction errors for both models show substantial overlap, with comparable medians and interquartile ranges. While StackBRAF [18] depends on a stacking ensemble, our model achieves a level of predictive performance equivalent to that of StackBRAF using a less complex probabilistic regression framework. This approach offers a simpler, interpretable alternative, demonstrating that state-of-the-art accuracy can be preserved without the structural overhead of complex ensemble methods.

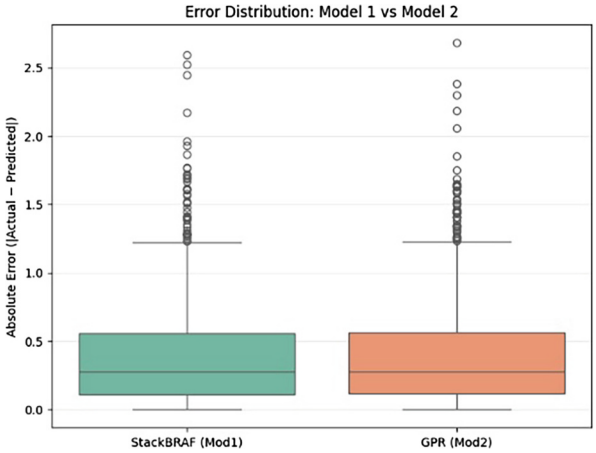

Fig. 4. Mean absolute error (MAE) distribution for both models on the external test set (n = 1157).

4 Ablation and Sensitivity Analysis

To quantify the contributions of distinct molecular representations and kernel formulations, we run a combined feature-level and model-structure ablation study. This ablation study follows a rigorous component-wise evaluation in

machine learning [10], in which the influence of isolated components are evaluated.

We have considered two distinct molecular representations: (i) physicochemical descriptors computed using PaDEL, and (ii) learned molecular embeddings derived from a recurrent neural network trained on SMILES strings. Specifically, four configurations were defined: (A) PaDEL descriptors only, (B) molecular embeddings only, (C) a combined feature space using a single shared kernel applied to the concatenated feature vector, and (D) a combined feature space modeled through the additive composition of separate kernels assigned to each representation. This experimental design allows the contributions of individual and combined feature representations to be examined under controlled conditions.

All ablation experiments were conducted under strictly controlled conditions. Identical train–test splits, target variables, random seeds, scaling procedures, kernel hyperparameter bounds, initial values, and optimization protocols were employed across all configurations.

An ablation study was also performed to evaluate the individual and combined contributions of PaDEL molecular descriptors and learned molecular embeddings to model performance. Four configurations were considered: (A) PaDEL descriptors only, (B) molecular embeddings only, (C) the combined PaDEL–embedding representation within a single-kernel Gaussian Process Regression (GPR) framework, and (D) a combined representation modeled using additive separate kernels.

Using only the PaDEL descriptors selected by XGBoost feature importance ranking, followed by Pearson correlation-based redundancy filtering (Experiment A), the model achieved strong external predictive performance ($R^2 = 0.8379$, MAE = 0.3987, r = 0.9156), confirming that established handcrafted fingerprints capture highly informative structural features relevant to BRAF inhibition.

In contrast, the models with embeddings-only representation (Experiment B) demonstrated lower performance ($R^2 = 0.6169$, $MAE = 0.6205$, $r = 0.7860$). While these results indicate that latent embeddings are insufficient as a standalone representation for this task, they maintain a significant correlation with experimental pIC_{50} values. This suggests that the contrastive learning strategy successfully encodes relevant, albeit incomplete, chemical features, which likely serves as a high-level complement to the more granular information provided by traditional descriptors.

The combined representation (Experiment C) consistently improved predictive performance relative to the PaDEL-only model, yielding $R^2 = 0.8418$, MAE = 0.3949, and r = 0.9178**.** While the numerical gain in R^2 is modest, the improvement is consistently observed across all evaluated metrics, demonstrating that molecular embeddings provide complementary information that enhances model generalization on unseen compounds when integrated with handcrafted descriptors.

In Experiment D, the combined representation was modeled using an additive kernel structure, with independent kernels applied to PaDEL descriptors

and molecular embeddings. This configuration produced predictive performance similar to the shared-kernel model ($R^2 = 0.8416$, MAE = 0.3954, r = 0.9177), showing that the model is able to capture the contribution of each representation separately without degrading predictive accuracy. The similarity in performance between Experiments C and D suggests that the predictive signal extracted from the combined feature space is robust to kernel decomposition. At the same time, the additive formulation provides a more structured representation of how different molecular feature groups contribute to the prediction. This result further supports the complementary nature of handcrafted descriptors and learned embeddings within the GPR framework.

5 Discussion

Stacking ensembles are designed to improve predictive performance by combining multiple heterogeneous base learners, often yielding gains through variance reduction rather than through fundamentally deeper learning of structure–activity relationships. In this context, ensemble learning primarily exploits redundancy across models trained on the same descriptor space, smoothing prediction errors via aggregation. While effective in many scenarios, such gains do not necessarily imply the extraction of new chemical information, but rather a stabilization of predictions through model averaging.

The present results suggest that, for the BRAF inhibition dataset, the performance ceiling is largely determined by the available data rather than by the expressive power of increasingly complex architectures. Both the proposed GPR-based model and StackBRAF achieve similar accuracy, supported by comparable R^2/Q^2 values, successful y-randomization tests, robust applicability domain analyses, and a paired Wilcoxon signed-rank test indicating no statistically significant difference between their prediction error distributions.

The proposed framework differs conceptually from stacking-based approaches in how information is learned and integrated. In StackBRAF-like architectures [18], multiple models are trained independently on the same molecular descriptor space, and their predictions are subsequently used as inputs to a meta-model. This approach transforms model outputs into proxy features, providing an indirect representation of the original molecular information. On the contrary, our approach learns a single continuous function from an enriched input space that combines selected PaDEL descriptors with learned molecular embeddings. This architectural design preserves chemical information, removes redundancy, and integrates symbolic and learned representations within a unified learning process.

Furthermore, the feature selection approach employed here plays a central role in controlling the bias–variance trade-off. By identifying and retaining the most informative PaDEL descriptors prior to model training, the approach reduces feature redundancy and noise, enabling the model to focus on chemically meaningful variation. When combined with learned molecular embeddings, this strategy enriches the input space with additional information rather than relying on ensemble diversity.

Taken together, these findings support the notion that, for this specific prediction task, stacking ensembles does not confer a fundamental advantage over well-designed simple approaches.

6 Conclusions

This study proposed a principled and parsimonious framework for predicting BRAF inhibitor potency, demonstrating that state-of-the-art QSAR performance can be achieved without reliance on stacking-based ensemble architectures. By integrating selected PaDEL molecular descriptors with learned molecular embeddings within a single Gaussian Process Regression (GPR) model, our proposed approach achieved predictive performance statistically equivalent to that of state-of-the-art stacking-based models.

The equivalence in predictive accuracy between the proposed model and stacking-based approaches is supported by consistent performance across training, internal validation, and external test sets, as well as by multiple robustness analyses, including y-randomization, applicability domain evaluation, and a paired Wilcoxon signed-rank test, which indicated no statistically significant difference in prediction errors. These results suggest that, for this specific prediction task, additional architectural complexity does not yield meaningful gains in predictive performance, and that the achievable performance ceiling is largely constrained by the underlying data rather than by model expressiveness.

The proposed framework differs fundamentally from stacking-based strategies in how information is learned and utilized. While stacking ensembles aggregate predictions from multiple models trained on the same descriptor space, primarily reducing variance through model averaging, the present approach focuses on enriching the input representation. By preserving chemically interpretable descriptors, removing redundancy through informed feature selection, and integrating complementary latent information from molecular embeddings, the model learns a single continuous function directly in an enriched feature space. This design enables efficient use of available information without relying on ensemble diversity as a surrogate for representational completeness.

Lastly, from a theoretical and practical perspective, Gaussian Process Regression is particularly well-suited to this scenario. GPR naturally captures smooth structure–activity relationships, operates effectively in moderate-data settings typical of QSAR studies, and provides calibrated uncertainty estimates alongside point predictions. The strong correspondence observed between predictive uncertainty and error, together with well-calibrated confidence intervals, highlights the value of uncertainty-aware modeling for defining applicability domains and supporting reliable decision-making in virtual screening and lead optimization.

Acknowledgment. The authors gratefully acknowledge support from FAPESP (São Paulo Research Foundation), project No. 2022/09285-7, and the support from the Brazilian National Council for Scientific and Technological Development (CNPq), grant 310377/2025-0.

Data and Code Availability Statement. The dataset used in this study corresponds to the dataset reported in the StackBRAF [18]. The source code is publicly available at https://github.com/CIDAG/GPR-BRAF-Model.

References

1. Bento, A.P., et al.: An open source chemical structure curation pipeline using RDKit. J. Cheminform. **12**(1), 1–16 (2020). https://doi.org/10.1186/s13321-020-00456-1
2. Bollag, G., Hirth, P., Tsai, J., et al.: Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. Nature **467**, 596–599 (2010). https://doi.org/10.1038/nature09454
3. Bonaccorso, G.: Mastering Machine Learning Algorithms: Expert Techniques for Implementing Popular Machine Learning Algorithms, Fine-Tuning Your Models, and Understanding How They Work. Packt Publishing (2018)
4. Bouchè, V., et al.: BRAF signaling inhibition in glioblastoma: which clinical perspectives? Front. Oncol. **11** (2021). https://doi.org/10.3389/fonc.2021.772052
5. Cotto-Rios, X.M., et al.: Inhibitors of BRAF dimers using an allosteric site. Nat. Commun. **11**, 4370 (2020). https://doi.org/10.1038/s41467-020-18123-2
6. Hollander, M., Wolfe, D.A., Chicken, E.: Nonparametric Statistical Methods. Wiley (2015). https://doi.org/10.1002/9781119196037
7. Jaworska, J., Nikolova-Jeliazkova, N., Aldenberg, T.: QSAR applicability domain estimation by projection of the training set in descriptor space: a review. Altern. Lab. Anim. **33**, 445–459 (2005). https://doi.org/10.1177/026119290503300508
8. Kuhn, M., Johnson, K.: Feature Engineering and Selection. Chapman and Hall/CRC (2019). https://doi.org/10.1201/9781315108230
9. Leach, A.R., Gillet, V.J.: An Introduction to Chemoinformatics. Springer (2007). https://doi.org/10.1007/978-1-4020-6291-9
10. Lipton, Z.C., Steinhardt, J.: Troubling trends in machine learning scholarship. Queue **17**, 45–77 (2019). https://doi.org/10.1145/3317287.3328534
11. Nakarin, F., Boonpalit, K., Kinchagawat, J., et al.: Assisting multitargeted ligand affinity prediction of receptor tyrosine kinases associated nonsmall cell lung cancer treatment with multitasking principal neighborhood aggregation. Molecules **27**, 1226 (2022). https://doi.org/10.3390/molecules27041226
12. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press (2005). https://doi.org/10.7551/mitpress/3206.001.0001
13. Rücker, C., Rücker, G., Meringer, M.: y-randomization and its variants in QSPR/QSAR. J. Chem. Inf. Model. **47**, 2345–2357 (2007). https://doi.org/10.1021/ci700157b
14. Salem, F.M.: Recurrent Neural Networks. Springer (2022). https://doi.org/10.1007/978-3-030-89929-5
15. Shan, K.S., Rehman, T.U., Ivanov, S., et al.: Molecular targeting of the BRAF proto-oncogene/mitogen-activated protein kinase (MAPK) pathway across cancers. Int. J. Mol. Sci. **25** (2024). https://doi.org/10.3390/ijms25010624
16. Bargagli Stoffi, F.J., Cevolani, G., Gnecco, G.: Simple models in complex worlds: Occam's razor and statistical learning theory. Mind. Mach. **32**(1), 13–42 (2022). https://doi.org/10.1007/s11023-022-09592-z
17. Su, K.W., Ou, D.L., Fu, Y.H., et al.: Repurposing cabozantinib with therapeutic potential in KIT-driven t(8;21) acute myeloid leukaemias. Cancer Gene Therapy **29**, 519–532 (2022). https://doi.org/10.1038/s41417-021-00329-1

18. Syahid, N.F., Weerapreeyakul, N., Srisongkram, T.: StackBRAF: a large-scale stacking ensemble learning for BRAF affinity prediction. ACS Omega **8**, 20881–20891 (2023). https://doi.org/10.1021/acsomega.3c01641
19. Todeschini, R., Ballabio, D., Grisoni, F.: Beware of unreliable Q^2 ! a comparative study of regression metrics for predictivity assessment of QSAR models. J. Chem. Inf. Model. **56**, 1905–1913 (2016). https://doi.org/10.1021/acs.jcim.6b00277
20. Todeschini, R., Consonni, V.: Molecular descriptors for chemoinformatics. Wiley-VCH; John Wiley [distributor] (2009). https://doi.org/10.1002/9783527628766
21. Tropsha, A.: Best practices for QSAR model development, validation, and exploitation. Mol. Inform. **29**, 476–488 (2010). https://doi.org/10.1002/minf.201000061
22. Weininger, D.: SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. **28**(2), 31–36 (1988). https://doi.org/10.1021/ci00057a005
23. Wilhelm, S.M., et al.: Regorafenib (BAY 73-4506): a new oral multikinase inhibitor of angiogenic, stromal and oncogenic receptor tyrosine kinases with potent preclinical antitumor activity. Int. J. Cancer **129**, 245–255 (2011). https://doi.org/10.1002/ijc.25864
24. Yap, C.W.: PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. **32**, 1466–1474 (2011). https://doi.org/10.1002/jcc.21707
25. Zhang, M., Zhang, C., Liu, K., et al.: BRAFPred: a novel approach for accurate prediction of the B-type rapidly accelerated fibrosarcoma inhibitor. ACS Omega **10**, 12170–12184 (2025). https://doi.org/10.1021/acsomega.4c10367
26. Zhong, L., Li, Y., Xiong, L., et al.: Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. Signal Transduct. Target. Ther. **6**, 201 (2021). https://doi.org/10.1038/s41392-021-00572-w