ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

PI3K-Seeker: A Machine Learning-Powered Web Tool to Discover PI3K Inhibitors

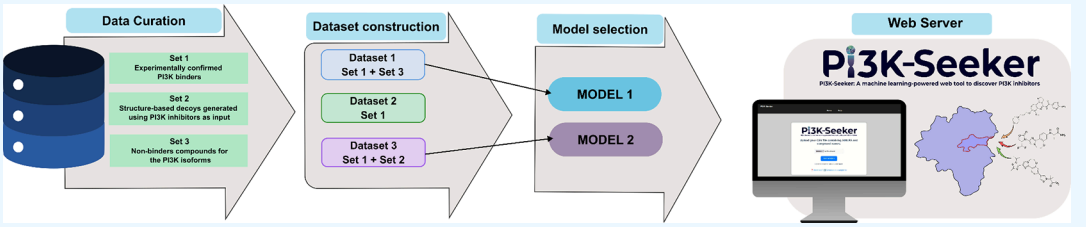
Published as part of ACS Omega special issue "Chemistry in Brazil: Advancing through Open Science".

Francisca Joseli Freitas de Sousa, Dinler Amaral Antunes, and Geancarlo Zanatta*

Cite This: *ACS Omega* 2025, 10, 57255−57266

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Phosphatidylinositol 3-kinases (PI3Ks) play a crucial role in human metabolism, and their dysregulation contributes to the development of several metabolic disorders, including cancer. Despite advances in experimental high-throughput screening, discovering new therapeutic agents remains challenging and costly. In this study, we developed PI3K-Seeker, a web server based on a two-stage prediction process to find new PI3K inhibitors. The first stage eliminates nonbinders, while the second refines the selection, leaving only molecules with a high probability of being potent inhibitors. Models were trained using the XGBoost algorithm and PubChem fingerprints extracted from distinct datasets. In the first stage of classification, the model showed impressive metrics (MCC: 0.917, AUC-ROC: 0.993, and ACC: 0.917). In the second stage, the data enhancement, the model trained also performed exceptionally well (MCC: 0.939, AUC-ROC: 0.956, and ACC: 0.994). The PI3K-Seeker is a user-friendly web server suitable for a large set of compounds, available at http://www.ufrgs.br/labec/pi3k-seeker/.

1. INTRODUCTION

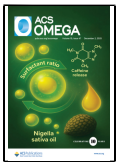
Phosphatidylinositol 3-kinase (PI3K) is an essential family of enzymes involved in cell signaling and is responsible for an extensive range of metabolic processes. Under normal physiological conditions, the PI3K/Akt/mTOR pathway regulates cell growth and controls the cell cycle. In abnormal situations, the overactivation of PI3K is associated with various metabolic disorders and is also implicated in oncological processes.[1−4]

Class IA PI3Ks (PI3Kα, PI3Kβ, and PI3Kδ) are heterodimers composed of a catalytic subunit (p110α, p110 β, or p110 δ) and a regulatory subunit (p85). The regulatory subunit binds to phosphorylated tyrosine residues on activated receptor tyrosine kinases (RTKs). This interaction recruits the enzyme to the plasma membrane and activates the catalytic subunit, which then phosphorylates phosphatidylinositol 4,5-bisphosphate (PIP2) to generate the second messenger phosphatidylinositol (3,4,5)-trisphosphate (PIP3).[5] Class IB PI3K (PI3Kγ) is activated downstream of G protein-coupled receptors (GPCRs). Their regulatory subunits (p101) bind directly to G$\beta\gamma$ subunits released from activated G-proteins, recruiting and activating the catalytic subunit at the membrane (Figure 1).[6]

PI3K class I isoforms are more studied due to their relationship with various diseases. For instance, changes in the activity of PI3Kα have been widely associated with multiple types of cancers, particularly solid tumors, as well as PIK3CA-related overgrowth spectrum (PROS) and activated phosphoinositide 3-kinase delta syndrome (APDS).[7−10] Meanwhile, PI3Kδ and PI3Kγ isoforms are more related to immune system pathologies.[11] Nonredundant functions associated with PI3Ks are also related to other processes directly or indirectly linked to malignancies and can have different impacts on specific types of cancers.[12,13]

Despite its importance, there is a limited number of PI3K inhibitors approved by Food and Drug Administration (FDA). Several factors contribute to this, including the incidence of serious adverse events associated with these compounds. Even among the approved inhibitors, poor tolerability, intrinsic and acquired drug resistance, and feedback signaling loops that counteract PI3K inhibition often led to treatment discontinuation.[14] In addition, the design of new PI3K inhibitors involves several challenges, with their highly conserved ATP-binding

Received: July 24, 2025
Revised: November 6, 2025
Accepted: November 10, 2025
Published: November 18, 2025



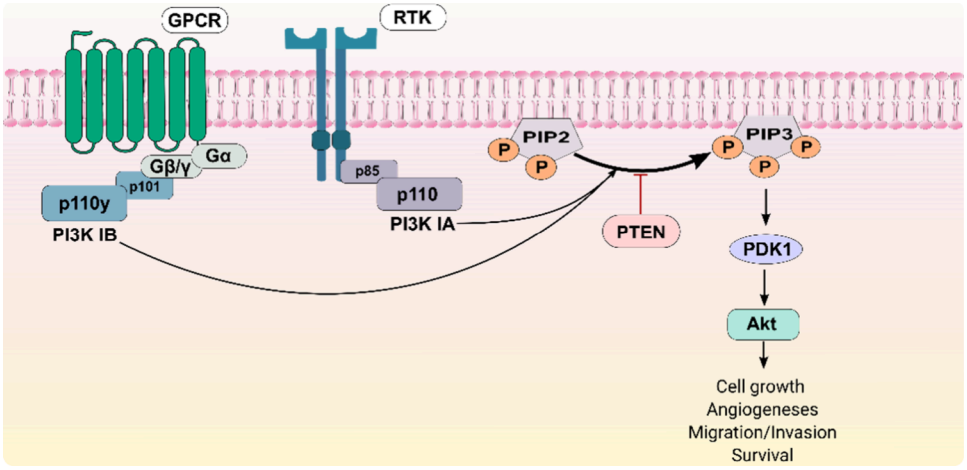


Figure 1. Overview of the PI3K class I pathway. The class IA PI3Ks (PI3Kα, PI3Kβ, and PI3Kδ) are activated downstream by Tyrosine Kinases (RTKs). In contrast, class IB (PI3Kγ) is activated downstream by G protein-coupled receptors (GPCRs). The catalytic subunit of PI3K phosphorylates phosphatidylinositol 4,5-bisphosphate (PIP2) to generate the second messenger phosphatidylinositol (3,4,5)-trisphosphate (PIP3). Additionally, the second messenger can be regenerated to PIP2 by the phosphatase and tensin homologue (PTEN).

pockets as one of the main obstacles. The PI3Kα, for example, can develop resistance to inhibition through functional compensation, or "feedback loops" via overexpression of tyrosine kinase receptors (RTKs). Other physiological changes resulting from the inhibition of this isoform may include disturbances in glucose metabolism. The use of PI3K inhibitors is associated with serious side effects resulting from their administration, such as hyperglycemia, diarrhea, nausea, pneumonia, fatigue, asthenia, skin rashes, among others, which lead to discontinuation of use or clinical trials.[15]

New technologies focused on drug discovery have been emerging, boosted by the increase in computational power and the advances in artificial intelligence (AI) approaches.[16] In addition, the increased number of public databases of chemical and biological information including PubChem and ChEMBL facilitates the training of new models, making them a valuable source for AI applications in drug discovery.[17,18] In this way, it is now possible to quickly scan large databases of various chemical components and use them as pharmacophores for drug design or repositioning.[19,20]

Among initiatives based on AI is the use of quantitative structure−activity relationship (QSAR) studies for the identification of ligands for specific targets. For instance, Bi et al. (2022) proposed machine learning models that utilize molecular fingerprinting to classify DYRK1A inhibitors, which are therapeutic targets for neurodegenerative diseases.[21] Similarly, Yu et al. (2023) applied a comparable approach to CYP17A1, a target for developing anticancer molecules.[22] The same approach was used by Yu and colleagues to identify LpxC inhibitors. In this work, they combined various fingerprints, such as MACCS and PubChem, to create classification models that predict the inhibitory activity of LpxC inhibitors.[23] Additionally, Srisongkram and colleagues (2023) employed a combination of fingerprints with an extreme gradient boosting (XGB)-QSAR model for high-throughput screening in drug design and for the identification of KRASg12c inhibitors.[21−24]

Attempts to develop models for drug discovery targeting PI3K have also been made. In this context, Zhu and colleagues[25] used a machine learning virtual screening approach to discover a novel selective inhibitor of PI3Kγ. Their study utilized a set of structures and ligands of the PI3Kγ isoform to construct a Naïve Bayesian Classification (NBC) model, employing a binary classification (1 for inhibitors and 0 for noninhibitors) and efficacy measured using the AUC-ROC curve (0.906). Although still an interesting approach to identify selective ligands, their approach was limited only to the isoform gamma. Recently, Kang and colleagues[26] have expanded the toolbox by developing a deep learning model, called MVGNet, which was reported to predict inhibitory activity by classifying molecules into two categories, active and inactive, across all four PI3K isoforms (α, β, γ, and δ). While their model represents a significant advancement in addressing isoform selectivity, it still faces limitations in effectively handling the early stages of virtual screening, particularly in rapidly and accurately distinguishing PI3K binders from nonbinders within large and diverse chemical libraries. Therefore, despite these promising advances, there is still room for improvement, especially at the early stages of virtual screening, where models that combine high accuracy, broad isoform coverage, and the ability to efficiently process large-scale compound libraries are needed. Developing such models would greatly enhance the identification and prioritization of novel PI3K-targeting compounds in drug discovery pipelines.

In this work, we present PI3K-Seeker, a fast and user-friendly web server that classifies compounds as active or inactive against isoforms belonging to the PI3K class I. This tool was implemented using a two-stage pipeline based on XGB[27] machine learning classification models. Each model was evaluated in terms of accuracy, precision, sensitivity, the Matthews correlation coefficient (MCC), and ROC AUC. As proof-of-concept, each stage of the tool was tested using real data as input, showing strong predictive performance and demonstrating its potential to accelerate the identification of novel PI3K inhibitors.

2. METHODOLOGY

2.1. Datasets. To facilitate the development and evaluation of machine learning models for the identification of PI3K class I ligands, we compiled three curated molecular sets: (set 1) experimentally confirmed PI3K binders, (set 2) structure-based decoys generated using a set of active inhibitors of PI3K as input, and (set 3) compounds as nonbinders of PI3K isoforms.

https://doi.org/10.1021/acsomega.5c07315
ACS Omega 2025, 10, 57255−57266

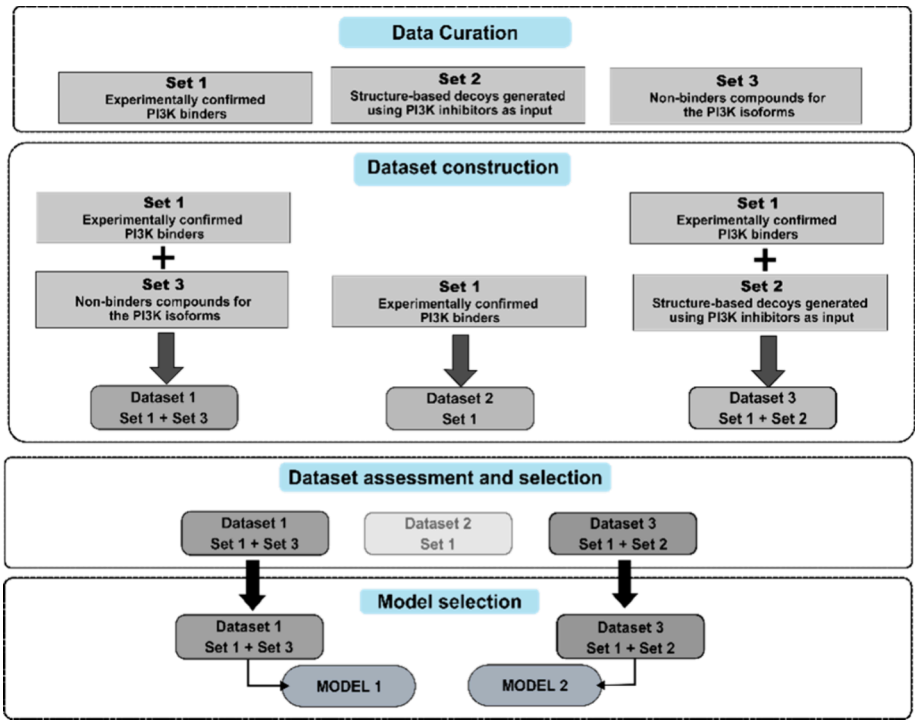


Figure 2. Flowchart showing the path from data curation to model selection.

Following these sets, three distinct datasets were constructed. Dataset1 comprises known binders (set 1) and nonbinders (set 3), providing a biologically grounded binary classification scenario. Dataset2 consists exclusively of the positive class (set 1) as a reference for supervised or one-class modeling approaches. Dataset3 integrates binders (set 1) and generated decoys (set 2), enabling evaluation of the capacity of the model to discriminate between actives and structurally plausible, yet presumed inactive compounds. Therefore, these datasets offer a robust framework for training and benchmarking ligand-based virtual screening algorithms across varying degrees of class separability and chemical diversity.

Set 1 was built based on data from the ChEMBL Data Web Services;[28] source IDs were 4005, 2111367, 3130, 2111432, 3145, 3038510, 3267, and 3559703. It comprised 22,175 molecules obtained by filtering based on IC50 values and limited by species (*Homo sapiens*) and with single protein and protein complex data. Bioactivity was retrieved using IC50 values (standard unit (nM)), and missing IC50 data or duplicate values were eliminated. For clarity, IC50 values were converted into pIC50 (pIC50 = −log 10(IC50)) and molecules were labeled as "active" when pIC50 > 6 or "inactive" when pIC50 < 5. Following data curation, 9028 unique compounds were obtained, comprising 7965 active and 1063 inactive molecules. Set 2 was built by generating decoys using the DUDE-Z server[29] In total, 11,312 new inactive molecules were generated using this approach based on 404 active molecules from set 1. Set 3 was made of 26,019 molecules using data from ChEMBL, comprising inhibitors of apoptosis protein 3 (IAP3/BIRC3), protein kinase C beta (PKCβ), C−C chemokine receptor type 5 (CCR3), MAP kinase ERK2 (MAPK/ERK2), vascular endothelial growth factor receptor 2 (VEGFR2), Janus kinase (JAK2), hexokinase type IV, and mechanistic target of rapamycin (mTOR). mTOR inhibitors with dual activity (e.g., PI3K/mTOR) were removed. The molecules obtained from ChEMBL and decoys were processed using a PaDEL descriptor before calculating the fingerprints, with salt removal, standardization of the nitro group, and tautomer standardization.

2.2. Scaffold Diversity and Chemical Space. We used the RDKit to process the chemical structures of the molecules and obtain Bemis−Murcko scaffolds for active and inactive compounds. Properties such as molecular weight (MW), number of hydrogen acceptors and donors (HBA/HBD), topological polar surface area (TPSA), rotatable bond count, carbon sp3 fraction (FracCsp3), ring count, aromaticity, and violations of Lipinski's rule of five were also calculated.[30,31]

2.3. Extracting Fingerprints. The molecular fingerprints used to extract features from all datasets were EState (79 bits), Molecular ACCess System (MACCS, 116 bits), and PubChem (881 bits), and PaDEL Descriptor software generated the fingerprints.[32]

2.4. Machine Learning Models. We compared the XGB[27] algorithm with three others: two machine learning algorithms (Support Vector Machine, SVM, and Random Forest, RF) and one deep learning algorithm (Graph Attention Network, GAT). SVM is effective for solving classification problems by finding a hyperplane that separates different classes.[33] Random Forest, on the other hand, utilizes an ensemble learning approach that combines multiple decision trees. In classification tasks, it determines the outcome based on a majority vote among those trees.[34] GAT, a subtype of Graph Neural Network, allows for the identification of important nodes beyond just their structural connections within the graph. This capability helps capture complex relationships based on the content of the nodes as well.[35]

We assessed metrics such as accuracy, precision, recall, F1-score, Matthews Correlation Coefficient (MCC), and ROC AUC (area under the curve of the receiver operating characteristic). We divided each dataset into a training set (80%) and a test set (20%). We performed a comprehensive 10-fold cross-validation to mitigate the possibility of overfitting.[36] The hyperparameter values for each method trained are

https://doi.org/10.1021/acsomega.5c07315
ACS Omega 2025, 10, 57255−57266

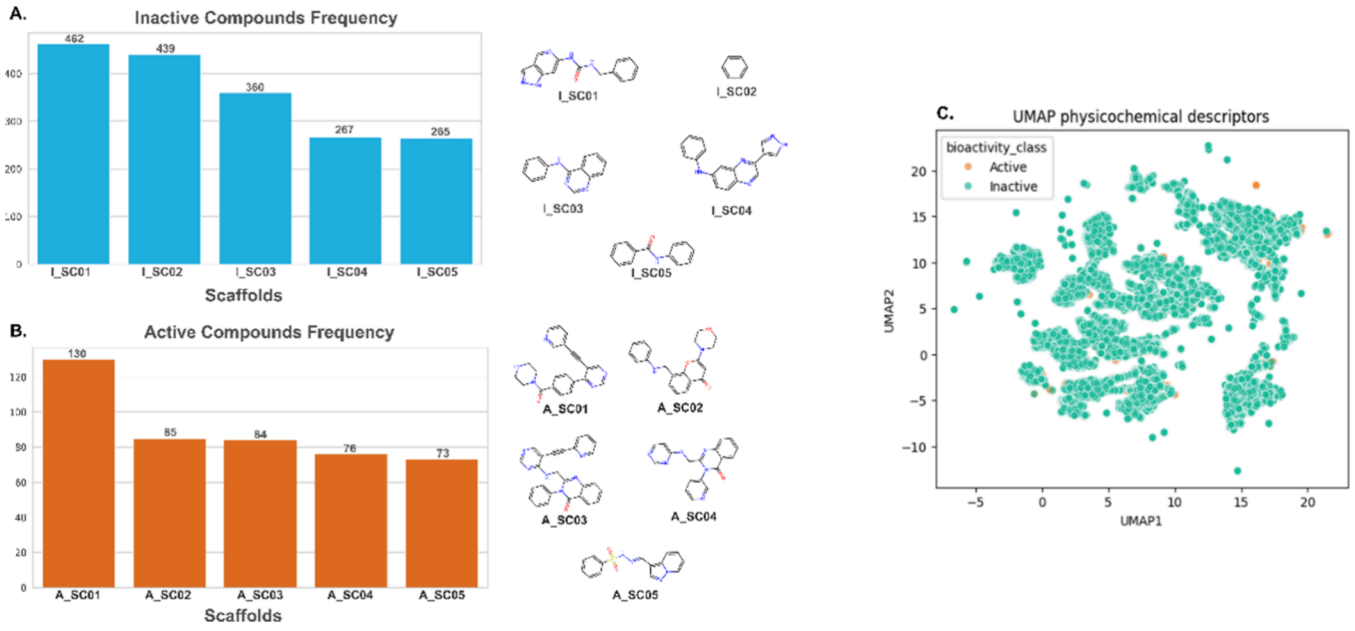


Figure 3. Scaffold diversity and chemical space. (A) Inactive compound frequency and (B) active compound frequency. (C) UMAP showing physicochemical descriptors in a 2D representation.

described in Supplementary Table S1, while the evaluation metrics are provided in Supplementary Table S2.

2.5. Assessment of Machine Learning and Fingerprint Outcomes. Furthermore, the models were evaluated to assess their performance across various metrics, including accuracy, precision, recall, F1-score, ROC AUC, and the Matthews Correlation Coefficient (MCC). These metrics were computed for test sets to comprehensively gauge the efficacy of each model.[37,38] In addition, the SHapley Additive exPlanations (SHAP) tool helped to understand how the model interprets data.[39]

To define an applicability threshold, we define an applicability domain using the LOF (local outlier factor) method. LOF stands out for comparing a point with its immediate neighborhood, being more sensitive to outliers that global methods would not notice.[40]

2.6. Machine Learning-Powered Web Tool to Discover PI3K Inhibitors—PI3K-Seeker. The PI3K-Seeker server was built to analyze the input data in two stages. In the first stage, it employs ML model 1, which was trained with dataset1 and eliminates non-PI3K binders. In the second stage of the analysis, ML model 2, which was trained with dataset3, classifies the remaining compounds as weak or strong binders, making the final prediction. PI3K-Seeker initially works with a CSV file provided by the users that includes a column naming the molecules and the molecules themselves in the SMILES format. After submitting the input files, PI3K-Seeker does the predictions for each molecule and returns it as "active" or "inactive". All data used in the training of the models are available in the Supporting Information. The process of creating datasets leading up to the development of models is illustrated in Figure 2.

3. RESULTS AND DISCUSSION

3.1. Scaffold Diversity and Chemical Space. We used the RDKit package to process the chemical structures of the molecules and obtain Bemis–Murcko scaffolds for active and inactive compounds. Properties such as molecular weight (MW), number of hydrogen acceptors and donors (HBA/HBD), topological polar surface area (TPSA), rotatable bond count, sp3 carbon fraction (FracCsp3), ring count, aromaticity, and violations of Lipinski's rule of five were also calculated.

Scaffold analysis was conducted using the Bemis–Murcko framework on all molecules included in model training, categorized as active or inactive. Figure 3A details the frequency of the five most common scaffolds within the inactive compounds, with their respective structures on the right. In contrast, Figure 3B shows the most common scaffolds in the active compounds. Although fewer in number, the active scaffolds are structurally more complex than those in inactive compounds, including motifs such as morpholine, sulfonamide, and dihydroquinazolin-4-one. The uniform manifold approximation and projection (UMAP) analysis in Figure 3C illustrates clusters of molecules with similar physicochemical profiles. Despite the presence of outliers, active and inactive compounds exhibit considerable overlap, with no clear separation between the classes, highlighting the importance of applying machine learning models capable of capturing more subtle patterns. Details of physicochemical properties are shown in Supplementary Figure S1.

3.2. Machine Learning Models. We tested a different combination of fingerprints and algorithms. As Figure 4 shows, the best-performing fingerprint was PubChem for each dataset. The best-performing algorithms were XGB and SVM. The complete metrics are presented in Table S2 of the Supporting Information.

According to Figure 4, when models were trained with dataset1, XGB, RF, and SVM methods showed superior performance using PubChem as fingerprint. The use of dataset1 highlights the balance of metrics, all greater than 0.9. GAT did not perform satisfactorily when compared to the others, especially in terms of accuracy, F1-score, and MCC. MCC is a more comprehensive metric because it considers the four quadrants of the confusion matrix (TP, FN, TN, FP), measuring the correlation between actual labels and predicted labels.[38] The MACCS fingerprint had good metrics with the SVM and XGB

https://doi.org/10.1021/acsomega.5c07315
ACS Omega 2025, 10, 57255−57266

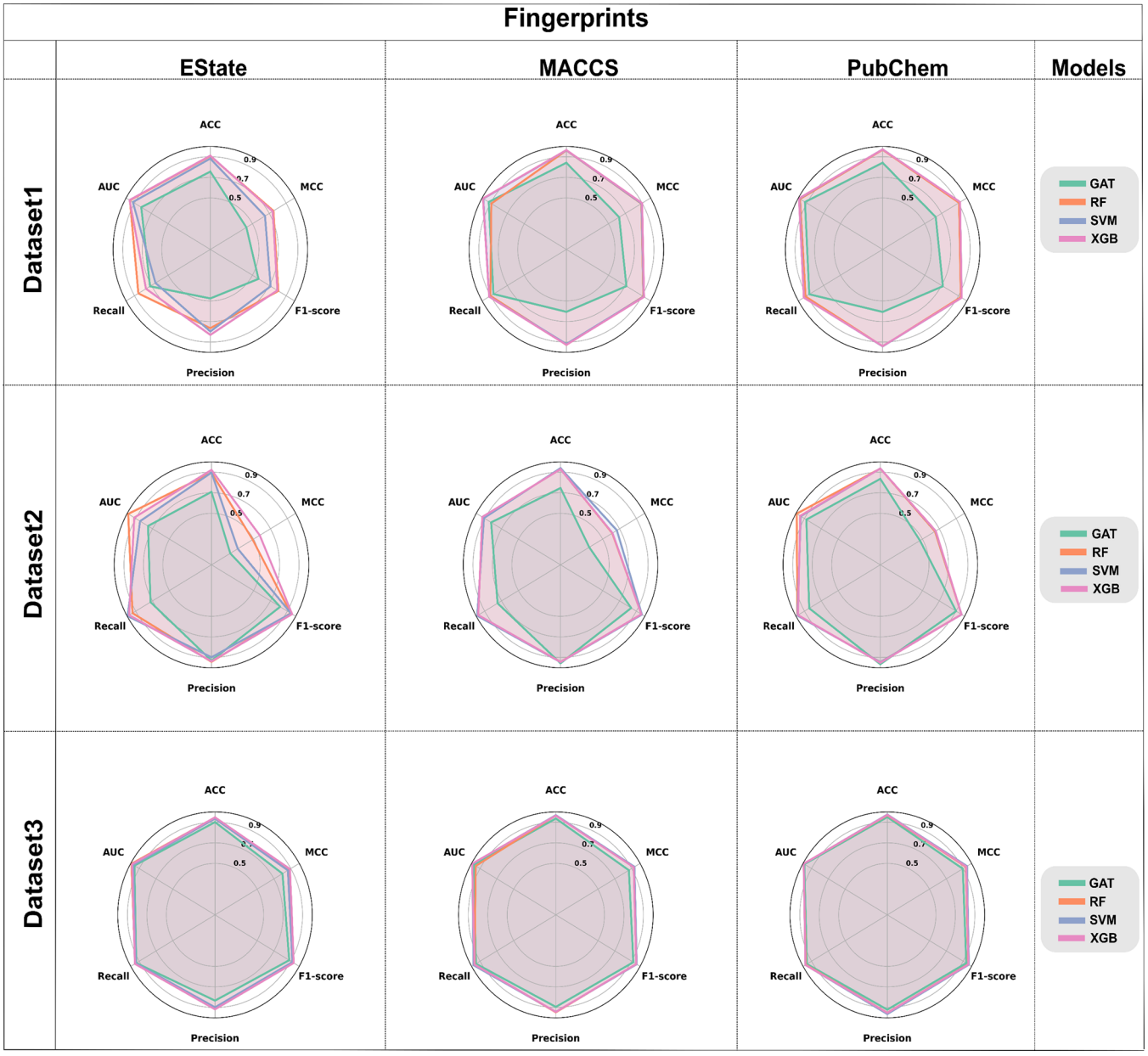


Figure 4. Metrics results of models tested. Legend: SVM (support vector machine), RF (random forest), XGB (extreme gradient boosting), GAT (graph attention networks), and three fingerprints (EState, MACCS, and PubChem).

algorithms, not surpassing PubChem, but better than EState. Overall, the algorithms performed worst on dataset2, especially when analyzing MCC values. For EState, from best to worst, the order was XGB > RF > SVM > GAT; for MACCS, SVM had a higher MCC value, and the order from best to worst was SVM > XGB > RF > GAT. Interestingly, all algorithms performed excellently with dataset3. Despite using approximate values, the PubChem fingerprint showed the best results overall. Among the algorithms, XGB and SVM performed exceptionally well. The results from the XGB were consistent across various datasets and fingerprints, consistently ranking among the best, particularly for MCC values. Due to this reliability, we opted to include XGB in the server pipeline.

To enhance the efficiency and predictive performance of the platform, we implemented a two-stage virtual screening strategy. In the first stage, a fast and accurate classification model was employed to eliminate compounds with a low likelihood of binding to PI3K, thereby significantly reducing the chemical search space. This initial filter was built using an XGB model trained on dataset1, which included both known PI3K binders and nonbinders. As shown in Table 1, the model exhibited strong predictive performance and generalizability, making it well-suited for the initial filtering step.

In the second stage, more computationally intensive analyses were performed on the subset of compounds retained from the initial filtering, allowing for a more refined identification of high-confidence PI3K binders. This stage required a model able to distinguish not only between binders and nonbinders but also to capture differences in the predicted binding affinity. To achieve this, we tested the performance of two XGB models. The first model was trained on dataset2, which includes only known PI3K binders, categorized as "active" or "inactive" based on their pIC50 values. The second was trained on dataset3, which expanded the negative class by incorporating a large number of

https://doi.org/10.1021/acsomega.5c07315
ACS Omega 2025, 10, 57255−57266

Table 1. Metrics for Dataset1 and PubChem Fingerprints Are Implemented in the First Step of the Virtual Screening Pipeline

parameter	metric	value
training	MCC_train	0.987
	CV_mean	0.971
	CV_SD	0.003
test	MCC_test	0.919
	precision	0.939
	accuracy	0.971
	recall	0.936
	F1-score	0.937
	AUC-ROC	0.993

decoy compounds alongside the active ligands from dataset2, thereby enhancing the capacity of the model to minimize the prediction of false positives.[41]

As the model trained on dataset3 outperformed the one trained on dataset2, it was implemented in the second stage of the server's analysis. The improvement obtained with this model is reflected across multiple performance metrics, as shown in Table 2. Although MCC was considered for model selection, other metrics, such as precision, recall, and AUC-ROC, were also critical for evaluating the ability of the model to generalize and correctly classify both active and inactive compounds. A comparative overview of the results is presented in Table 2.

Table 2. Metrics for XGB and PubChem Fingerprints Models Trained for the Second Step of the Virtual Screening Pipeline

dataset split	metrics	XGB (dataset2)	XGB (dataset3)
training	CV_mean	0.939	0.972
	CV_SD	0.006	0.005
	MCC_train	0.985	0.996
test	MCC_test	0.658	**0.937**
	precision	0.951	**0.952**
	accuracy	0.935	**0.971**
	recall	**0.977**	0.972
	F1-score	**0.964**	0.962
	AUC-ROC	0.943	**0.993**

3.3. Interpreting the Predictions Made by the Models.

Subsequently, to enhance interpretability and elucidate the decision-making process of the model, we employed SHAP (SHapley Additive exPlanations), a widely recognized framework for interpreting complex machine learning models.[39,42] SHAP analysis enabled the identification of key molecular substructures that drive prediction outcomes, thereby increasing confidence in both the interpretability and practical applicability of the model.

As shown in Figure 5, the XGB model trained on dataset1 (model 1), the descriptor with the most significant impact was PubchemFP392, which corresponds to a secondary amine (R-NH-R′). The SHAP analysis of the impact of this feature showed that its high frequency is strongly associated with positive SHAP values. This indicates that the model has learned to recognize this group as a strong predictor of inactivity. The model showed a tendency to penalize simple fragments, which is reinforced by the features PubchemFP635 (hydrazine, for example, Figure 5) and PubchemFP791 (1,4-diaminacyclohexane, for example, Figure 5), which shift the prediction to the inactive class. Taken together, these results suggest that the model has learned to filter out molecules containing nitrogenous and low-density groups that are atypical in PI3K inhibitors. Other fingerprint features, such as PubChemFP193, FP728, FP368, and FP335, also contributed to the prediction of nonbinders, further reinforcing the discriminatory capacity of the model for early stage filtering of inactive molecules.

In contrast, the XGB model trained on dataset3 (model 2), which aims to differentiate between weak and strong PI3K binders in the second stage of analysis performed by the server, exhibited a set of features with high values in the left (associated with active compounds). To illustrate how model 2 identifies active compounds, we used the known PI3K inhibitor, Buparlisib, as a case study for interpreting SHAP results (Figure

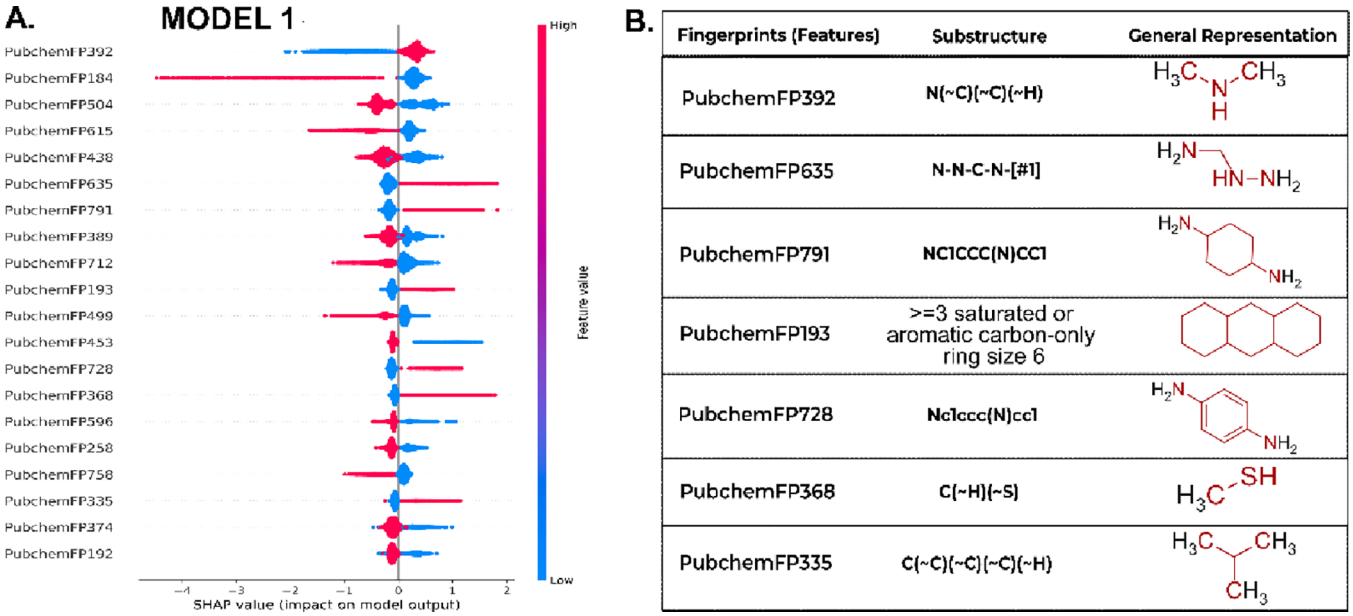


Figure 5. SHAP plot of model 1. In panel (A), the features were ranked according to the extent of their influence on the decisions of the models. Panel (B) shows the patterns associated with the prediction of PI3K noninhibitors.

https://doi.org/10.1021/acsomega.5c07315
ACS Omega 2025, 10, 57255−57266

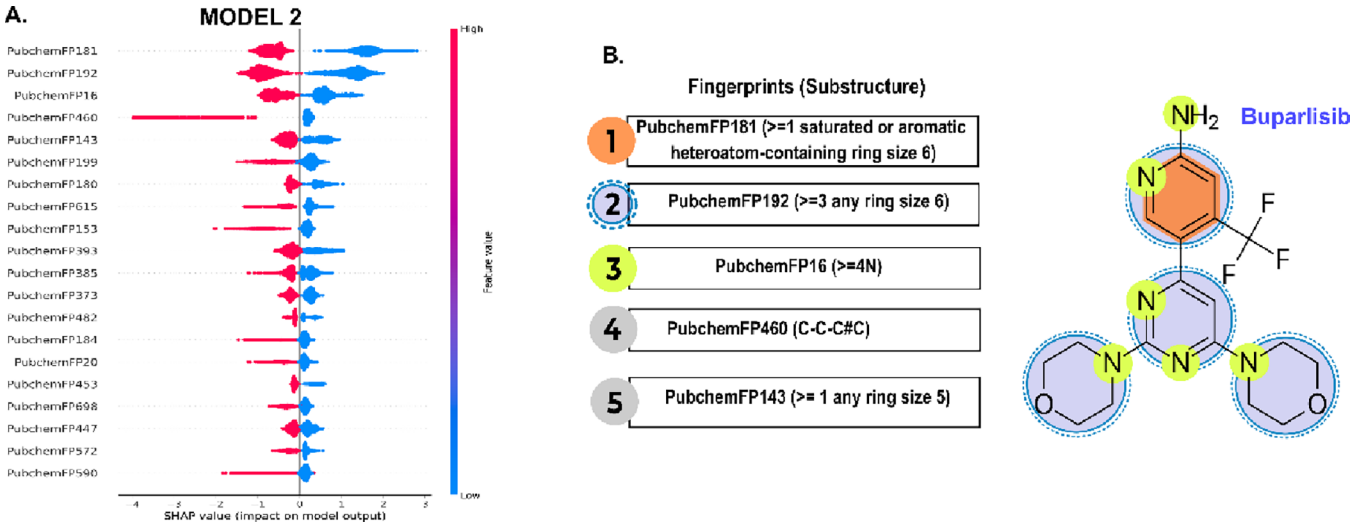


Figure 6. Fingerprints and substructure patterns. (A) SHAP plot associated with the active classification of model 2. (B) Fingerprints and the substructures that they represent are numbered 1−5, along with a description. In compound Buparlisib, a pan-inhibitor of PI3K, the substructures are highlighted. Each number corresponds to a color shown in Buparlisib, and they are linked to the classification of active compounds.

6). The functional group feature with the greatest impact, PubchemFP181, represents the presence of six-membered rings, saturated or aromatic, containing heteroatoms. Buparlisib exemplifies this characteristic very well, as its structure features pyridine and pyrimidine rings, along with two morpholine units (Figure 6B). SHAP analysis confirms that the presence of these fragments results in a strong negative SHAP value, validating them as key indicators in the prediction of active compounds. This structural signature is complemented by other important functional group feature, such as PubchemFP192 (three or more rings of size 6), reflecting the polycyclic nature of inhibitors in the classification of active compounds. Similarly, the high nitrogen count, captured by PubchemFP16 (four or more N atoms), also drives the prediction to "active", which is consistent with the profile of Buparlisib. Taken together, the analysis of Buparlisib reveals that the model has learned to associate activity with a more complex chemical structure, characterized by nitrogen-rich systems with multiple heteroatoms in six-membered rings.

Interestingly, the two models shared a few features, and their contributions varied significantly in both direction and magnitude. This divergence underscores the complementary roles of models 1 and 2 within the classification pipeline, enhancing both the precision and efficiency of PI3K inhibitor identification across chemically diverse compound libraries.

3.4. Applicability Domain (AD). Our models underwent a two-stage validation process by applying AD in both model 1 and model 2. First, we demonstrated its robustness and generalization ability in a chemically diverse dataset (dataset1, model 1), as shown in Table 3, maintaining high performance even for samples outside the domain of applicability. Next, we proved its remarkable sensitivity and discriminatory efficiency in a rigorous test with decoys (dataset3, model 2), confirming its effectiveness for the critical task of refining PI3K ligands. This dual approach with AD ensures that the model is not only a reliable generalist but also a high-performance specialist, validating it for practical application.

3.5. Web Server Deployment. The PI3K-Seeker web server is a user-friendly and computationally efficient platform designed to accelerate the identification of new inhibitors targeting class I PI3Ks. Implemented in Python, the application

Table 3. Applying the Applicability Domain (AD) in Model 1 and Model 2

	XGB-dataset1 (test)		XGB-dataset3 (test)	
metrics	in AD	out AD	in AD	out AD
MCC	**0.920**	0.906	**0.941**	0.856
accuracy	0.972	0.969	0.972	0.931
precision	0.939	0.944	0.955	0.917
recall	0.937	0.907	0.972	0.976
F1-score	0.938	0.925	0.964	0.945

employs the XGB algorithm to predict the ability of small-molecule ligands to bind to the PI3K isoforms α, β, γ, and δ. The only required input is a list of SMILES strings representing the chemical structures of the candidate molecules to be evaluated (Figure 7A).

Initially, the algorithm performs a validation step to ensure that the input file contains exclusively SMILES representations of small-molecule ligands. Subsequently, the platform computes molecular fingerprints using the PubChem fingerprinting methodology, executes predictive modeling, and generates the corresponding output files (Figure 7B). The results page provides access to a CSV file containing detailed information for each compound (Figure 7C), including the molecule name, SMILES string, predicted activity, molecular weight (MW), LogP, number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), and Topological Polar Surface Area (TPSA). The entire workflow is computationally efficient, requiring only a fraction of a second per molecule.

The high processing speed of PI3K-Seeker comes mainly from the use of the XGB algorithm as the core predictive model. Designed to efficiently handle high-dimensional and sparse datasets, such as those encountered in biological and cheminformatics applications, XGB supports parallelization across CPUs and GPUs, enabling rapid execution at scale.[27] Its scalability and predictive accuracy have been extensively validated across various biomedical contexts, including quantitative structure−activity relationship (QSAR) modeling, classification of early and late-stage cancers, and the prediction of clinical treatment outcomes, acute kidney injury prediction,

https://doi.org/10.1021/acsomega.5c07315
ACS Omega 2025, 10, 57255−57266

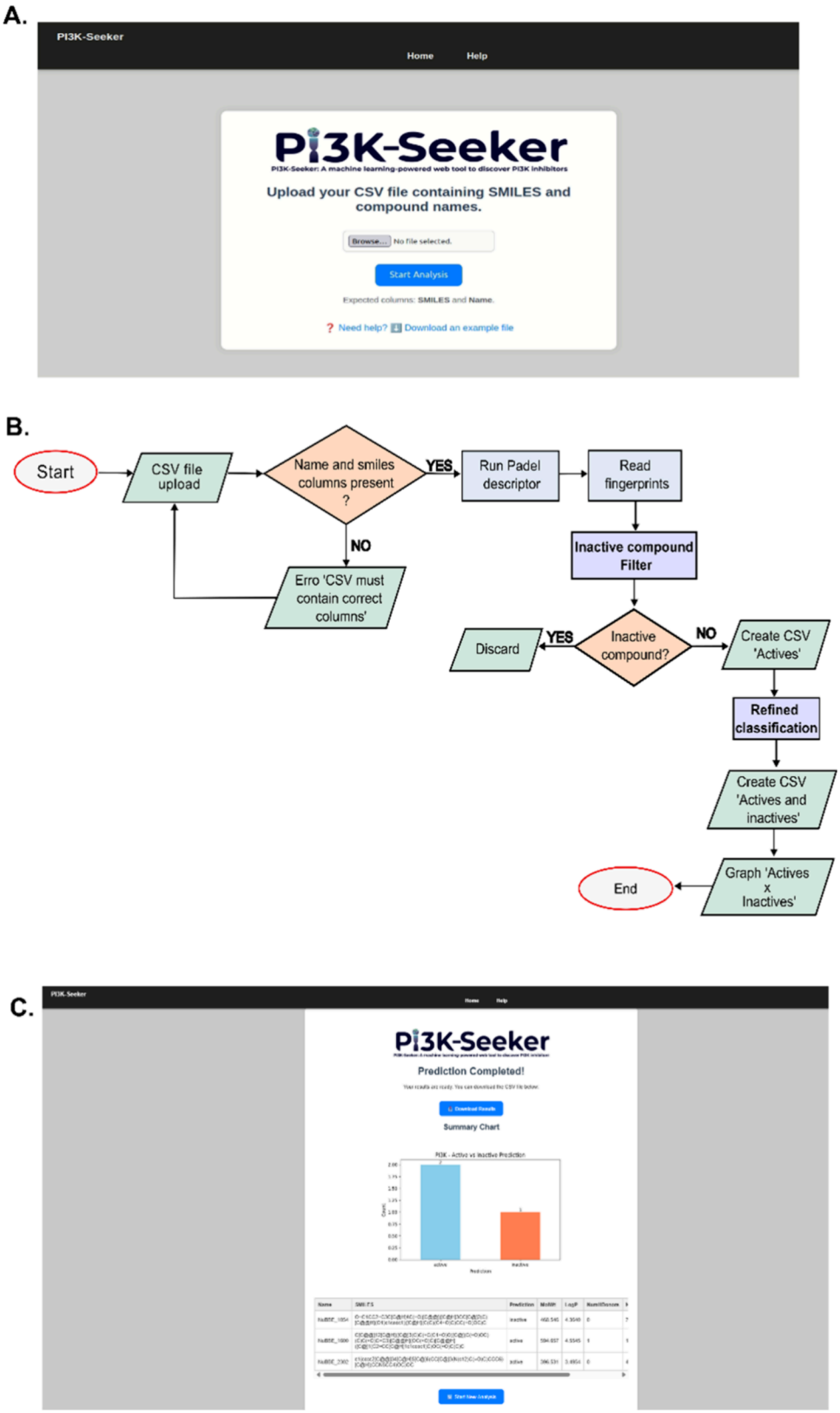


Figure 7. General overview of the PI3K-Seeker web server. (A) Input screen; (B) workflow including data processing, classification, and results generation; and (C) the output page showing a graphical representation of results and links to download CSV files with result data.

blood−brain barrier drug classification, and biomarker discovery in Alzheimer's disease transcriptomic data.[43−46]

3.6. Proof-of-Concept Validation. To better evaluate the performance of PI3K-Seeker, we subjected it to two distinct and challenging real-world scenarios. In the first case study, we assessed the server using PI3K ligands extracted from crystallographic structures, also absent from the training set.

https://doi.org/10.1021/acsomega.5c07315
ACS Omega 2025, 10, 57255−57266

Table 4. Ligands Found in the Crystallographic Structures of PI3K and Tested with PI3K-Seeker versus SVM

PDB	ligand	IUPAC	state	prediction PI3K-Seeker	prediction SVM
3PRZ[50]	3RZ	4-amino-2-methyl-*N*-(1*H*-pyrazol-3-yl)quinazoline-8-carboxamide	active	**active**	**active**
3PS6[50]	3PS	4-amino-*N*-(6-methoxypyridin-3-yl)-2-methylquinazoline-8-carboxamide	active	**active**	**inactive**
4OVV[48]	PBU	[(2*R*)-2-butanoyloxy-3-[hydroxy-[(1*R*,2*R*,3*S*,4*R*,5*R*,6*S*)-2,3,6-trihydroxy-4,5-diphosphonooxycyclohexyl]oxyphosphoryl]oxypropyl] butanoate	inactive	**inactive**	**inactive**
5JHA[51]	6K7	[1-{4-[6-amino-4-(trifluoromethyl)pyridin-3-yl]-6-(morpholin-4-yl)pyrimidin-2-yl}-3-(chloromethyl)azetidin-3-yl]methanol	active	**active**	**inactive**
5UK8[49]	8DV	(*R*)-4-(6-(1-(cyclopropylsulfonyl)cyclopropyl)-2-(1*H*-indol-4-yl)pyrimidin-4-yl)-3-methylmorpholine	inactive	**inactive**	**inactive**
5XGH[52]	84U	3-[(4-fluorophenyl)methylamino]-5-(4-morpholin-4-ylthieno[3,2-*d*]pyrimidin-2-yl)phenol	active	**active**	**active**
6EYZ[53]	C5W	2-methoxy-5-[4-[5-[(4-propan-2-ylpiperazin-1-yl)methyl]-1,3-oxazol-2-yl]-2∼{*H*}-indazol-6-yl] pyridine-3-carboxylic acid	active	**active**	**active**
6GY0[54]	FGE	∼{*N*}-[4-methyl-5-(1-oxidanylidene-7-sulfamoyl-isoindol-5-yl)-1,3-thiazol-2-yl]ethanamide	active	**active**	**active**
6ZAA[55]	QD2	4-[6-methoxy-5-(methylsulfamoyl)pyridin-3-yl]-∼{*N*}-(1-methylpiperidin-4-yl)-2,3-dihydro-1,4-benzoxazine-6-carboxamide	active	**active**	**active**
8AM0[56]	MWF	(2*R*)-2-[[2-[(4*S*)-4-[bis(fluoranyl)methyl]-2-oxidanylidene-1,3-oxazolidin-3-yl]-5,6-dihydroimidazo[1,2-*d*][1,4]benzoxazepin-9-yl]amino]propanamide	active	**inactive**	**inactive**
8ILV[47]	L2V	*N*-[(2*R*)-1-(ethylamino)-1-oxidanylidene-3-[3-(2-quinoxalin-6-ylethynyl)phenyl]propan-2-yl]-2,3-dimethyl-quinoxaline-6-carboxamide	inactive	**inactive**	**inactive**
8SC8[57]	D0D	*N*-[(5*P*)-2-chloro-5-(4-{[(1*R*)-1-phenylethyl]amino}quinazolin-6-yl)pyridin-3-yl] methanesulfonamide	active	**active**	**active**
9GCF[58]	A1IJ5	3-[(1*S*)-1-[4-azanyl-3-(5-oxidanylpyridin-3-yl)pyrazolo[3,4-*d*]pyrimidin-1-yl]ethyl]-4-[3-[(4-methylpiperazin-1-yl)methyl]phenyl]isochromen-1-one	active	**active**	**active**
9GDI[58]	A1IJ1	3-[(1*S*)-1-[4-azanyl-3-(3-fluoranyl-5-oxidanyl-phenyl)pyrazolo[3,4-*d*]pyrimidin-1-yl]ethyl]-4-(1-methyl-3,6-dihydro-2*H*-pyridin-4-yl)isochromen-1-one	active	**active**	**active**

Also, we compared it with SVM, which performed very similarly to XGB using the parameters applied in PI3K-Seeker. In the second case study, the PI3K-Seeker was tested with a dataset composed exclusively of non-PI3K binders retrieved from the ChEMBL database, which were not included in the training data of the model.

As shown in Table 4, the PI3K-Seeker server was evaluated for its ability to classify ligands from crystallographic PI3K structures not used during model training. A total of 14 ligands were tested: 11 with confirmed inhibitory activity against at least one PI3K isoform and three classified as noninhibitors. As the primary objective of this test was to assess the capacity of the model to distinguish between high- and low-affinity binders, crystallographic noninhibitors were included as controls based on their well-characterized interactions with key residues in the PI3K binding site. Among them, ligand L2 V interacts with residues such as R770α and W780α, critical for PI3Kα isoform selectivity, without interfering with kinase activity, thereby validating its classification as a noninhibitor.[47] Similarly, PBU, a di-C4-phosphatidylinositol-4,5-bisphosphate (diC4-PIP2) lipid substrate mimetic used to study PI3Kα catalysis, was correctly identified as a noninhibitor.[48] The ATR inhibitor 8DV, structurally distinct and inactive against PI3K, was also correctly classified.[49]

As shown in Table 4, PI3K-Seeker achieved correct classification for 13 of the 14 compounds. The only misclassified ligand was MWF (Inavolisib), a molecule that does not act through conventional kinase inhibition but instead induces proteasome-dependent degradation of the mutant p110α protein.[56] This mechanistic divergence from the training data likely accounts for the misclassification. Inavolisib was approved by the U.S. Food and Drug Administration in October 2024, in combination with palbociclib and fulvestrant, for treating endocrine-resistant, PIK3CA-mutated, hormone receptor-positive, HER2-negative advanced breast cancer.[27]

When dealing with external data with PI3K inhibitors, SVM ended up incorrectly classifying active inhibitors as inactive. Therefore, the underperformance below expectations is a bottleneck that prevents the expected improvement in phase two. In this regard, XGB proved to be superior, capable of satisfactorily separating inactive ligands and identifying active ones. However, both algorithms failed to classify Inavolisib. The final choice of pipeline was based on the balance between robustness and accuracy. In this scenario, we have XGB (PI3K-Seeker) as the alternative that best balances these aspects.

In the second case study, the server demonstrated excellent performance in correctly identifying non-PI3K binders across a diverse set of protein targets, with predictive accuracy exceeding 95% in the vast majority of datasets (Table 5). This high accuracy confirms the model's ability to effectively discriminate PI3K-specific ligands from those active against unrelated molecular targets, even within large and structurally diverse compound libraries. Notably, the server maintained high discriminatory efficiency for other kinase targets, such as EGFR (97.9%), FGFR1 (94.8%), RAF (96.0%), and insulin receptor (98.5%), reinforcing its robustness in distinguishing PI3K inhibitors from other ATP-competitive ligands. Overall, these findings highlight the high specificity of the model and its potential applicability in virtual screening workflows aimed at identifying selective PI3K inhibitors.

4. CONCLUSIONS

In this work, we present the PI3K-Seeker server, a powerful and freely accessible computational tool designed to speed up drug discovery by narrowing down compounds with the ability to bind to the ATP-binding pocket of PI3K class I enzymes. The server integrates the use of two machine learning models based on the XGB algorithm to differentiate between PI3K binders and nonbinders. The user can easily interrogate the server by solely supplying a list of SMILES strings. Compared to existing approaches, the PI3K-Seeker server offers superior performance, delivering high-accuracy predictions in under seconds. This study highlights the predictive power of machine learning

https://doi.org/10.1021/acsomega.5c07315
ACS Omega 2025, 10, 57255−57266

Table 5. List of Datasets with Non-PI3K Binders, Selected from Experimental Assays

protein	compounds	inactive (%)	active (%)
acetylcholinesterase	8127	99.8	0.2
adenosine A2a receptor	1977	98.1	1.9
aldose reductase	1163	99.7	0.3
androgen receptor	3707	97.7	2.3
angiotensin converting enzyme	876	100	0.0
caspase 3	2434	98.8	1.2
cytochrome P450 2C19 (CYP2C19)	3397	96.1	3.9
cytochrome P450 2C9 (CYP2C9)	5366	99.7	0.3
cytochrome P450 3A4 (CYP3A4)	11165	95.5	4.5
copamine D3	508	99.8	0.2
epidermal growth factor receptor (EGFR)	16711	97.9	2.1
estrogen receptor alpha	4251	99.6	0.4
fibroblast growth factor receptor 1 (FGFR1)	3222	94.8	5.2
gamma-aminobutyric acid receptor alpha-1	40	100	0.0
HIV-1 reverse transcriptase	10104	99.0	1.0
insulin receptor	1364	98.5	1.5
RAS	688	89.7	10.3
renin	3083	99.0	1.0
serine/threonine protein kinase B RAF	4930	96.0	4.0
xanthine oxidase	602	99.0	1.0

algorithms in virtual screening protocols, contributing to speeding up the discovery of PI3K inhibitors.

■ ASSOCIATED CONTENT

Data Availability Statement

The PI3K-Seeker server is freely available at http://www.ufrgs.br/labec/pi3k-seeker/.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.5c07315.

Machine learning algorithms and hyperparameters (Table S1); metrics for all datasets, fingerprints, and machine learning algorithms in the test set (Table S2); and physicochemical properties (Figure S1) (PDF)
Dataset1.csv (2.74 MB) (ZIP)
Dataset2.csv (714 KB) (ZIP)
Dataset3.csv (1.42 MB) (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Geancarlo Zanatta − *Postgraduate Programme in Biochemistry, Department of Biochemistry at Federal University of Ceará, Fortaleza 60440-554 CE, Brazil; Department of Biophysics, Federal University of Rio Grande do Sul, Porto Alegre 91501-970, Brazil;* orcid.org/0000-0003-0111-5347; Email: geancarlo.zanatta@gmail.com

Authors

Francisca Joseli Freitas de Sousa − *Postgraduate Programme in Biochemistry, Department of Biochemistry at Federal University of Ceará, Fortaleza 60440-554 CE, Brazil*

Dinler Amaral Antunes − *Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.5c07315

Author Contributions

F.J.F.d.S., D.A.A., and G.Z. designed the research; D.A.A. participated in the planning of the project; F.J.F.d.S. and G.Z. performed the research; F.J.F.d.S. analyzed the data; and F.J.F.d.S., D.A.A., and G.Z. wrote the paper.

Funding

The Article Processing Charge for the publication of this research was funded by the Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior (CAPES), Brazil (ROR identifier: 00x0ma614).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

G.Z. was funded by grant nos. 440412/2022-6 and 408135/2023-9 from CNPq. G.Z. has a CNPq fellowship (306190/2025-7). F.J.F.d.S. has a FUNCAP fellowship.

■ REFERENCES

(1) Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* **2022**, *12* (1), 31−46.

(2) Haselmayer, P.; Camps, M.; Muzerelle, M.; Bawab, S. El; Waltzinger, C.; Bruns, L.; Abla, N.; Polokoff, M. A.; Jond-Necand, C.; Gaudet, M.; Benoit, A.; Meier, D. B.; Martin, C.; Gretener, D.; Lombardi, M. S.; Grenningloh, R.; Ladel, C.; Petersen, J. S.; Gaillard, P.; Ji, H. Characterization of Novel PI3K*δ* Inhibitors as Potential Therapeutics for SLE and Lupus Nephritis in Pre-Clinical Studies. *Front. Immunol.* **2014**, *5* (May), 233.

(3) Pillinger, G.; Loughran, N. V.; Piddock, R. E.; Shafat, M. S.; Zaitseva, L.; Abdul-Aziz, A.; Lawes, M. J.; Bowles, K. M.; Rushworth, S. A. Targeting PI3K*δ* and PI3K*γ* Signalling Disrupts Human AML Survival and Bone Marrow Stromal Cell Mediated Protection. *Oncotarget* **2016**, *7* (26), 39784−39795.

(4) Rao, V. K.; Webster, S.; Dalm, V. A. S. H.; Šedivá, A.; van Hagen, P. M.; Holland, S.; Rosenzweig, S. D.; Christ, A. D.; Sloth, B.; Cabanski, M.; Joshi, A. D.; de Buck, S.; Doucet, J.; Guerini, D.; Kalis, C.; Pylvaenaeinen, I.; Soldermann, N.; Kashyap, A.; Uzel, G.; Lenardo, M. J.; Patel, D. D.; Lucas, C. L.; Burkhart, C. Effective, "Activated PI3K*δ* Syndrome"−Targeted Therapy with the PI3K*δ* Inhibitor Leniolisib. *Blood* **2017**, *130* (21), 2307−2316.

(5) Vanhaesebroeck, B.; Perry, M. W. D.; Brown, J. R.; André, F.; Okkenhaug, K. PI3K Inhibitors Are Finally Coming of Age. *Nat. Rev. Drug Discov* **2021**, *20* (10), 741−769.

(6) Rathinaswamy, M. K.; Dalwadi, U.; Fleming, K. D.; Adams, C.; B Stariha, J. T.; Pardon, E.; Baek, M.; Vadas, O.; DiMaio, F.; Steyaert, J.; Hansen, S. D.; Yip, C. K.; Burke, J. E. Structure of the Phosphoinositide 3-Kinase (PI3K) P110*γ*-P101 Complex Reveals Molecular Mechanism of GPCR Activation. *Sci. Adv.* **2021**, *7* (35), No. eabj4282.

(7) Adam, M. P.; Feldman, J.; Mirzaa, G. M. *PIK3CA-Related Overgrowth Spectrum Synonym: PROS*; **2013**.

(8) Martínez-Saéz, O.; Chic, N.; Pascual, T.; Adamo, B.; Vidal, M.; González-Farré, B.; Sanfeliu, E.; Schettini, F.; Conte, B.; Brasó-Maristany, F.; Rodríguez, A.; Martínez, D.; Galván, P.; Rodríguez, A. B.; Martinez, A.; Munõz, M.; Prat, A. Frequency and Spectrum of PIK3CA Somatic Mutations in Breast Cancer. *Breast Cancer Res.* **2020**, *22* (1), 45.

(9) Takeda, A. J.; Zhang, Y.; Dornan, G. L.; Siempelkamp, B. D.; Jenkins, M. L.; Matthews, H. F.; McElwee, J. J.; Bi, W.; Seeborg, F. O.; Su, H. C.; Burke, J. E.; Lucas, C. L. Novel PIK3CD Mutations Affecting N-Terminal Residues of P110*δ* Cause Activated PI3K*δ* Syndrome (APDS) in Humans. *J. Allergy Clin. Immunol.* **2017**, *140* (4), 1152−1156.

(10) von Keudell, G.; Moskowitz, A. J. The Role of PI3K Inhibition in Lymphoid Malignancies. *Curr. Hematol Malig Rep* **2019**, *14* (5), 405−413.
(11) Batlevi, C. L.; Younes, A. Revival of PI3K Inhibitors in Non-Hodgkin's Lymphoma. *Annals of Oncology* **2017**, *28* (9), 2047−2049.
(12) Geering, B.; Cutillas, P. R.; Nock, G.; Gharbi, S. I.; Vanhaesebroeck, B. Class IA Phosphoinositide 3-Kinases Are Obligate P85-P110 Heterodimers. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (19), 7809−7914.
(13) Wee, S.; Wiederschain, D.; Maira, S.-M.; Loo, A.; Miller, C.; deBeaumont, R.; Stegmeier, F.; Yao, Y.-M.; Lengauer, C. PTEN-Deficient Cancers Depend on PIK3CB. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (35), 13057−13062.
(14) Fruman, D. A.; Chiu, H.; Hopkins, B. D.; Bagrodia, S.; Cantley, L. C.; Abraham, R. T. The PI3K Pathway in Human Disease. *Cell* **2017**, *170* (4), 605−635.
(15) Sapon-Cousineau, V.; Sapon-Cousineau, S.; Assouline, S. PI3K Inhibitors and Their Role as Novel Agents for Targeted Therapy in Lymphoma. *Curr. Treat. Options Oncol.* **2020**, *21* (6), 51.
(16) Ekins, S.; Puhl, A. C.; Zorn, K. M.; Lane, T. R.; Russo, D. P.; Klein, J. J.; Hickey, A. J.; Clark, A. M. Exploiting Machine Learning for End-to-End Drug Discovery and Development. *Nat. Mater.* **2019**, *18* (5), 435−441.
(17) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, *51* (D1), D1373−D1380.
(18) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; Magarinos, M. P.; Bosc, N.; Arcila, R.; Kizilören, T.; Gaulton, A.; Bento, A. P.; Adasme, M. F.; Monecke, P.; Landrum, G. A.; Leach, A. R. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2024**, *52* (D1), D1180−D1192.
(19) Napolitano, F.; Zhao, Y.; Moreira, V. M.; Tagliaferri, R.; Kere, J.; Greco, D. Drug Repositioning: A Machine-Learning Approach through Data Integration. *J. Cheminf.* **2013**, *5* (1), 30.
(20) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119* (18), 10520−10594.
(21) Bi, M.; Guan, Z.; Fan, T.; Zhang, N.; Wang, J.; Sun, G.; Zhao, L.; Zhong, R. Identification of Pharmacophoric Fragments of DYRK1A Inhibitors Using Machine Learning Classification Models. *Molecules* **2022**, *27* (6), 1753.
(22) Chiang, Y. C.; Ren, R.; Piacham, T.; Anuwongcharoen, N.; Nantasenamat, C.; Yu, L.; Huang, T.; Yu, T. Exploring the Chemical Space of CYP17A1 Inhibitors Using Cheminformatics and Machine Learning. *Molecules* **2023**, *28* (4), 1679.
(23) Yu, T.; Chong, L. C.; Nantasenamat, C.; Anuwongcharoen, N.; Piacham, T. Machine Learning Approaches to Study the Structure-Activity Relationships of LpxC Inhibitors. *EXCLI J.* **2023**, *22*, 975−991.
(24) Srisongkram, T.; Khamtang, P.; Weerapreeyakul, N. Prediction of KRASG12C Inhibitors Using Conjoint Fingerprint and Machine Learning-Based QSAR Models. *J. Mol. Graphics Modell.* **2023**, *122*, No. 108466.
(25) Zhu, J.; Li, K.; Xu, L.; Cai, Y.; Chen, Y.; Zhao, X.; Li, H.; Huang, G.; Jin, J. Discovery of Novel Selective PI3Kγ Inhibitors through Combining Machine Learning-Based Virtual Screening with Multiple Protein Structures and Bio-Evaluation. *J. Adv. Res.* **2022**, *36*, 1−13.
(26) Kang, Y.; Xia, Q.; Jiang, Y.; Li, Z. MVGNet: Prediction of PI3K Inhibitors Using Multitask Learning and Multiview Frameworks. *ACS Omega* **2024**, *9* (45), 45159−45168.
(27) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery, **2016**; Vol. *13*-17-August-2016, pp 785−794. .
(28) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Res.* **2015**, *43* (W1), W612−W620.
(29) Stein, R. M.; Yang, Y.; Balius, T. E.; O'Meara, M. J.; Lyu, J.; Young, J.; Tang, K.; Shoichet, B. K.; Irwin, J. J. Property-Unmatched Decoys in Docking Benchmarks. *J. Chem. Inf Model* **2021**, *61* (2), 699−714.
(30) Landrum, G. *RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling*; http://rdkit.sourceforge.net.
(31) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−2893.
(32) Yap, C. W. PaDEL-descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466−1474.
(33) Cortes, C.; Vapnik, V.; Saitta, L. *Support-Vector Networks Ed.*; Kluwer Academic Publishers, 1995; *Vol. 20*.
(34) Zhang, L.; Ai, H.; Chen, W.; Yin, Z.; Hu, H.; Zhu, J.; Zhao, J.; Zhao, Q.; Liu, H. CarcinoPred-EL: Novel Models for Predicting the Carcinogenicity of Chemicals Using Molecular Fingerprints and Ensemble Learning Methods. *Sci. Rep.* **2017**, *7* (1), 2118.
(35) Vrahatis, A. G.; Lazaros, K.; Kotsiantis, S. Graph Attention Networks: A Comprehensive Review of Methods and Applications. *Future Internet* **2024**, *16* (9), 318.
(36) Arlot, S.; Celisse, A. A Survey of Cross-Validation Procedures for Model Selection. *Stat Surv* **2010**, *4*, 40−79.
(37) Chicco, D.; Jurman, G. The Matthews Correlation Coefficient (MCC) Should Replace the ROC AUC as the Standard Metric for Assessing Binary Classification. *BioData Min.* **2023**, *16* (1), 4.
(38) Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics* **2020**, *21* (1), 6.
(39) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach Intell* **2020**, *2* (1), 56−67.
(40) Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. *LOF: Identifying Density-Based Local Outliers*; **2000**.
(41) Cáceres, E. L.; Mew, N. C.; Keiser, M. J. Adding Stochastic Negative Examples into Machine Learning Improves Molecular Bioactivity Prediction. *J. Chem. Inf Model* **2020**, *60* (12), 5957−5970.
(42) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should i Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery, **2016**; Vol. *13*-17-August-2016, pp 1135−1144. https://doi.org/10.1145/2939672.2939778.
(43) Yi, Y.; Tae, M.; Shin, S.; Choi, S. I. Predicting Acute Kidney Injury in Trauma Using an Extreme Gradient Boosting Model. *Clin. Kidney J.* **2025**, *18* (4), sfaf002.
(44) Subha Ramakrishnan, M.; Ganapathy, N. Extreme Gradient Boosting Based Improved Classification of Blood-Brain-Barrier Drugs. In *Studies in Health Technology and Informatics*; IOS Press BV, 2022; Vol. *294*, pp 872−873. .
(45) Zhang, Y.; Shen, S.; Li, X.; Wang, S.; Xiao, Z.; Cheng, J.; Li, R. A Multiclass Extreme Gradient Boosting Model for Evaluation of Transcriptomic Biomarkers in Alzheimer's Disease Prediction. *Neurosci. Lett.* **2024**, *821*, No. 137609.
(46) Mateo, J.; Rius-Peris, J. M.; Maraña-Pérez, A. I.; Valiente-Armero, A.; Torres, A. M. Extreme Gradient Boosting Machine Learning Method for Predicting Medical Treatment in Patients with Acute Bronchiolitis. *Biocybern Biomed Eng.* **2021**, *41* (2), 792−801.
(47) Zhou, Q.; Liu, X.; Neri, D.; Li, W.; Favalli, N.; Bassi, G.; Yang, S.; Yang, D.; Vogt, P. K.; Wang, M. W. Structural Insights into the Interaction of Three Y-Shaped Ligands with PI3Kα. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120* (34), No. e2304071120.
(48) Miller, M. S.; Schmidt-Kittler, O.; Bolduc, D. M.; Brower, E. T.; Chaves-Moreira, D.; Allaire, M.; Kinzler, K. W.; Jennings, I. G.; Thompson, P. E.; Cole, P. A.; Amzel, L. M.; Vogelstein, B.; Gabelli, S. B.

Structural Basis of NSH2 Regulation and Lipid Binding in PI3Kα. *Oncotarget* **2014**, *5* (14), 5198−5208.
(49) Lu, Y.; Knapp, M.; Crawford, K.; Warne, R.; Elling, R.; Yan, K.; Doyle, M.; Pardee, G.; Zhang, L.; Ma, S.; Mamo, M.; Ornelas, E.; Pan, Y.; Bussiere, D.; Jansen, J.; Zaror, I.; Lai, A.; Barsanti, P.; Sim, J. Rationally Designed PI3Kα Mutants to Mimic ATR and Their Use to Understand Binding Specificity of ATR Inhibitors. *J. Mol. Biol.* **2017**, *429* (11), 1684−1704.
(50) Liu, K. K. C.; Huang, X.; Bagrodia, S.; Chen, J. H.; Greasley, S.; Cheng, H.; Sun, S.; Knighton, D.; Rodgers, C.; Rafidi, K.; Zou, A.; Xiao, J.; Yan, S. Quinazolines with Intra-Molecular Hydrogen Bonding Scaffold (IMHBS) as PI3K/MTOR Dual Inhibitors. *Bioorg. Med. Chem. Lett.* **2011**, *21* (4), 1270−1274.
(51) Bohnacker, T.; Prota, A. E.; Beaufils, F.; Burke, J. E.; Melone, A.; Inglis, A. J.; Rageot, D.; Sele, A. M.; Cmiljanovic, V.; Cmiljanovic, N.; Bargsten, K.; Aher, A.; Akhmanova, A.; Díaz, J. F.; Fabbro, D.; Zvelebil, M.; Williams, R. L.; Steinmetz, M. O.; Wymann, M. P. Deconvolution of Buparlisib's Mechanism of Action Defines Specific PI3K and Tubulin Inhibitors for Therapeutic Intervention. *Nat. Commun.* **2017**, *8*, 8.
(52) Yang, X.; Zhang, X.; Huang, M.; Song, K.; Li, X.; Huang, M.; Meng, L.; Zhang, J. New Insights into PI3K Inhibitor Design Using X-Ray Structures of PI3Kα Complexed with a Potent Lead Compound. *Sci. Rep.* **2017**, *7* (1), 14572.
(53) Dalton, S. E.; Dittus, L.; Thomas, D. A.; Convery, M. A.; Nunes, J.; Bush, J. T.; Evans, J. P.; Werner, T.; Bantscheff, M.; Murphy, J. A.; Campos, S. Selectively Targeting the Kinome-Conserved Lysine of PI3Kδ as a General Approach to Covalent Kinase Inhibition. *J. Am. Chem. Soc.* **2018**, *140* (3), 932−939.
(54) Gangadhara, G.; Dahl, G.; Bohnacker, T.; Rae, R.; Gunnarsson, J.; Blaho, S.; Öster, L.; Lindmark, H.; Karabelas, K.; Pemberton, N.; Tyrchan, C.; Mogemark, M.; Wymann, M. P.; Williams, R. L.; Perry, M. W. D.; Papavoine, T.; Petersen, J. A Class of Highly Selective Inhibitors Bind to an Active State of PI3Kγ. *Nat. Chem. Biol.* **2019**, *15* (4), 348−357.
(55) Spencer, J. A.; Baldwin, I. R.; Barton, N.; Chung, C. W.; Convery, M. A.; Edwards, C. D.; Jamieson, C.; Mallett, D. N.; Rowedder, J. E.; Rowland, P.; Thomas, D. A.; Hardy, C. J. Design and Development of a Macrocyclic Series Targeting Phosphoinositide 3-Kinase δ. *ACS Med. Chem. Lett.* **2020**, *11* (7), 1386−1391.
(56) Hanan, E. J.; Braun, M. G.; Heald, R. A.; Macleod, C.; Chan, C.; Clausen, S.; Edgar, K. A.; Eigenbrot, C.; Elliott, R.; Endres, N.; Friedman, L. S.; Gogol, E.; Gu, X. H.; Thibodeau, R. H.; Jackson, P. S.; Kiefer, J. R.; Knight, J. D.; Nannini, M.; Narukulla, R.; Pace, A.; Pang, J.; Purkey, H. E.; Salphati, L.; Sampath, D.; Schmidt, S.; Sideris, S.; Song, K.; Sujatha-Bhaskar, S.; Ultsch, M.; Wallweber, H.; Xin, J.; Yeap, S.; Young, A.; Zhong, Y.; Staben, S. T. Discovery of GDC-0077 (Inavolisib), a Highly Selective Inhibitor and Degrader of Mutant PI3Kα. *J. Med. Chem.* **2022**, *65* (24), 16589−16621.
(57) Whitehead, C. E.; Ziemke, E. K.; Frankowski-McGregor, C. L.; Mumby, R. A.; Chung, J.; Li, J.; Osher, N.; Coker, O.; Baladandayuthapani, V.; Kopetz, S.; Sebolt-Leopold, J. S. A First-in-Class Selective Inhibitor of EGFR and PI3K Offers a Single-Molecule Approach to Targeting Adaptive Resistance. *Nat. Cancer* **2024**, *5* (8), 1250−1266.
(58) Bruno, P.; Micoli, A.; Corsi, M.; Pala, D.; Guariento, S.; Fiorelli, C.; Ronchi, P.; Fioni, A.; Gallo, P. M.; Marenghi, G.; Bertolini, S.; Capacchi, S.; Mileo, V.; Biagetti, M.; Capelli, A. M. Discovery and Optimization of Pyridazinones as PI3Kδ Selective Inhibitors for Administration by Inhalation. *J. Med. Chem.* **2024**, *67* (13), 11103−11124.

https://doi.org/10.1021/acsomega.5c07315
ACS Omega 2025, 10, 57255−57266