Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Discovery of Novel Polycyclic Aromatic Inhibitors of Anaplastic Lymphoma Kinase (ALK) via Integrated QSAR Modeling, Machine Learning, and Pharmacophore-Guided Virtual Screening with AI-Assisted Mechanistic Interpretation of Structure–Activity Relationships

Rand Shahin
The Hashemite University
Hiba Zalloum
University of Jordan
Mohammad A Qaderi
The Hashemite University
Qusi K Alomoush
The Hashemite University
Iman Mansi
The Hashemite University
Lo'ay Al-Momani
The Hashemite University
Salma Azzam
University at Buffalo, State University of New York
Yusra Azzam
yusraazz@buffalo.edu

University at Buffalo, State University of New York

Research Article

Keywords: Anaplastic Lymphoma, Kinase, ALK, mutation, QSAR, GFA-MLR, Random Forest, pharmacophore modeling, virtual screening, resistance mutations, G1202R, NCI database, anthraquinone, aminoquinoline, quinazoline

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11515269/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Discovery of Novel Polycyclic Aromatic Inhibitors of Anaplastic Lymphoma Kinase (ALK) via Integrated QSAR Modeling, Machine Learning, and Pharmacophore-Guided Virtual Screening with AI-Assisted Mechanistic Interpretation of Structure–Activity Relationships

Rand Shahin [a], *Hiba Zalloum* [b], *Mohammad A Qaderi* [c], *Qusi K Alomoush* [c], *Iman Mansi*[d], *Lo'ay Al-Momani*[c], *Salma Azzam*[e], *Yusra Azzam*[e] *

[a] Department of Pharmaceutical Chemistry, Faculty of Pharmaceutical sciences, The Hashemite University, P.O. Box 330117, Zarqa, 13133 Jordan. ORCID: https://orcid.org/0000-0001-7330-4433Phone: 00962 799-799-044. e-mail: r.shahin@hu.edu.jo
[b] Hamdi Mango Center for Scientific Research, The University of Jordan, Amman 11942, Jordan.
[c] Department of Chemistry, Faculty of Science, The Hashemite University, P.O. Box 330117, Zarqa, 13133 Jordan. e-mails: Mohammadalqadre@yahoo.com (M.Q.); qusaydoujan98@gmail.com (Q.A.).
[d] Department of Clinical Pharmacy and Pharmacy Practice, Faculty of Pharmaceutical Sciences, The Hashemite University, Zarqa, Jordan.
[e] Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, 955 Main Street, Buffalo, NY 14203, USA. e-mail: salmaazz@buffalo.edu (S.A); yusraazz@buffalo.edu. (Y.A)
* Corresponding author

Abstract

Anaplastic Lymphoma Kinase (ALK) is a validated oncogenic driver in non-small cell lung cancer, but acquired resistance, most critically the solvent-front mutation G1202R, erodes the durability of all approved inhibitors, requiring new scaffolds without cross-resistance liabilities.

An interpretable GFA-MLR QSAR model ($R^2 = 0.60$, $Q^2 = 0.58$, $r^2pred = 0.56$) and a Genetic-Algorithm-optimized machine learning classifier ensemble (Random Forest, XGBoost, k-NN, Naïve Bayes) were built independently from the same descriptor set of 287 ALK inhibitors (225 descriptors, 109 pharmacophore hypotheses; training n = 230, test n = 57), with GA-based feature selection applied uniformly before training each classifier. Random Forest (accuracy = 98.7%, $\kappa = 0.98$) selected two of the same high-frequency pharmacophores (A-T1-1, J-T5-1) identified by the GFA-MLR equation, despite no cross-reference between the two branches during model development. A linear regression model and a nonlinear ensemble classifier therefore converged independently on the same structural determinants of ALK potency. The physicochemical rationale underlying each GFA-MLR coefficient was then derived through a structured, AI-assisted interpretation framework anchored to the known ALK binding-site architecture. Together, the dual-branch convergence and the AI-assisted coefficient interpretation constitute a cross-validated alternative to single-model QSAR interpretation, a strategy that, to our knowledge, has not previously been applied in kinase inhibitor discovery.

Pharmacophore-guided NCI screening yielded 28 candidates; eight (28.6%) inhibited ALK at 10 μM, and three, NSC 290954, NSC 196403, and NSC 715055 (IC_{50} = 2.15, 4.02, 5.95 μM), spanning anthraquinone, aminoquinoline, and 4-anilinoquinazoline scaffolds distinct from approved inhibitors, were confirmed by dose–response assay, docking/molecular dynamics (ATP-hinge binding, Met1199 engagement), and NCI-60 cellular profiling.

Keywords: Anaplastic Lymphoma; Kinase; ALK; mutation, QSAR; GFA-MLR; Random Forest; pharmacophore modeling; virtual screening; resistance mutations; G1202R; NCI database; anthraquinone; aminoquinoline; quinazoline

1. Introduction

1.1 ALK as an Oncogenic Validated Drug Design Target

Anaplastic Lymphoma Kinase (ALK) is a receptor tyrosine kinase of the insulin receptor superfamily whose expression is restricted to the developing nervous system, small intestine, and testis [1,2]. Surprisingly, the oncogenic relevance of the ALK enzyme arises from the chromosomal rearrangements that fuse the ALK kinase domain to the active partner genes and the most prevalent is the t(2;5)(p23;q35) translocation that generates the NPM1/ALK fusion protein [3–5]. Furthermore, the expression of this chimeric kinase drives the activation of the RAS/ERK, PI3K/AKT, and JAK/STAT3 proliferative cascades which is also considered as a primary oncogenic driving factor in approximately 50-60% of Anaplastic Large Cell Lymphoma (ALCL) cases [3]. Besides, in non-small cell lung cancer (NSCLC), the ALK rearrangements is most commonly involving the EML4/ALK fusion and are identified in 3-7% of patients [6]. Additionally, the ALK driven malignancies include inflammatory myofibroblastic tumors and a subset of neuroblastomas. All these facts are collectively establishing the ALK enzyme as a kinase drug design target in oncology [1].

1.2 The Clinical Burden of Resistance against the ALK inhibitors

Crizotinib (the first ALK selective inhibitor) was approved by the U.S. FDA in 2011. At that time, Crizotinib demonstrated an objective response rate of 60-74% in ALK rearranged NSCLC [1]. Despite its initial efficacy, Crizotinib acquired resistance that can emerge within 12-15 months, driven by the secondary kinase domain mutations (most prominently the solvent front substitution G1202R and the gatekeeper mutation (L1196M) (See Table 1) [7]. Second generation inhibitors (ceritinib (2014), alectinib (2015), brigatinib (2017), and ensartinib (2020)) were developed to address crizotinib resistant mutations and achieved superior CNS penetration; however, their activity against compound mutations and solvent front variants also suffered from resistance (see Table 1) [8–10]. Next, Lorlatinib is the third-generation inhibitor of the ALK enzyme which was specifically designed to address a broad mutational coverage (including the G1202R mutation) demonstrated a 5-year progression free survival rate of 60% compared to 8% for crizotinib [11]. Nevertheless, mutations such as G1202R and L1196M are still conferring resistance to Lorlatinib in a growing proportion of patients which underscores the need for structurally novel inhibitor scaffolds that do not show cross-resistance with the existing clinical agents [12].

Table 1. Key resistance-associated residues in the ALK kinase binding pocket.

Residue	Sub-region	Mutation(s)[a]	Structural role	Resistance mechanism	Inhibitors affected
L1196	Gatekeeper (ATP binding cleft)	L1196M	Controls access to hydrophobic back-pocket; small Leu side chain permits inhibitor entry	Bulky Met side chain blocks inhibitor binding; increases ATP affinity	Crizotinib, ceritinib (partial)
G1202	Solvent-front (ATP entrance)	G1202R	Flexible Gly at solvent-exposed rim; tolerates diverse scaffolds	Arg introduces steric clash and electrostatic repulsion; and reduces the binding of most 2nd-gen inhibitors	Crizotinib, ceritinib, alectinib, brigatinib (partial)
C1156	αC helix region	C1156Y, C1155Y	αC helix position modulates active-site geometry and inhibitor orientation	Tyr substitution alters αC-helix packing; destabilizes inhibitor H-bond network	Crizotinib
F1174	Activation loop / αC-helix	F1174L	Phe stabilizes inactive conformation; critical for kinase autoregulation	Leu locks kinase in active conformation; reduces inhibitor access to DFG-out mode	Crizotinib; active against ceritinib, alectinib, lorlatinib
R1275	Activation loop	R1275Q	Arg forms salt bridge stabilizing inactive kinase conformation	Gln disrupts salt bridge; promotes constitutive A-loop activation	Crizotinib; active against ceritinib, alectinib, lorlatinib
M1199/E1210	Hinge region	M1199I, E1210K	Hinge residues form key H-bond donor and acceptor contacts with inhibitor scaffolds	Substitutions disrupt critical hinge H-bonds; reduce binding affinity of pyrimidine/pyridine-core inhibitors	Alectinib, brigatinib (emerging)
L1256	A-loop / DFG motif adjacent	L1256F	Leu contributes to DFG-motif flexibility enabling DFG-out inhibitor binding	Phe rigidifies A-loop, restricts DFG flip, which prevents allosteric back-pocket access	Lorlatinib (emerging)

[a]Resistance classifications are based on structural and clinical data from approved ALK inhibitors (Source: Qu et al. Eur J Med Chem. 2025;299:118053)

Abbreviations: HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; A-loop, activation loop; αC-helix, alpha-C helix; DFG, Asp-Phe-Gly motif.

1.3 The Structural Determinants of ALK Inhibition and the Molecular Basis of Resistance

The analysis of the ALK enzyme crystallographic pictures showed that the ATP binding domain of the ALK kinase adopts the canonical bilobal kinase fold, with the ATP binding pocket defined by several

critical sub-regions for example, the hinge region (Met1199, Glu1197) which contributes for the backbone hydrogen bond donor and the acceptor interactions. Thus, these interactions are conserved across all approved ALK inhibitors and represent the primary pharmacophoric anchor for potent binding [8,13–15]. Also, the gatekeeper residue Leu1196 controls the inhibitor's access to the hydrophobic back pocket, and its substitution to methionine (L1196M) introduces steric bulk that sterically occludes crizotinib and other structurally similar first-generation agents (see Table 1) [7]. Moreover, the solvent front residue Gly1202, which situated at the solvent exposed portal of the ATP binding pocket, is the most clinically recurrent single resistance locus since the G1202R substitution introduces electrostatic repulsion and steric clash that impairs binding of most second-generation inhibitors that are directed toward this region (see Table 1) [1,7]. Furthermore, the DFG motif (Asp1270-Phe1271-Gly1272) in addition to the activation loop residues F1174, R1275, and L1256 modulate kinase conformation between DFG-in and DFG-out states [14]. Finally, the electropositive character of the ATP binding pocket that is contributed by Lys1150 and Arg1253, further imposes complementarity requirements on the distribution of negative electrostatic surface potential across inhibitor scaffolds.

1.4 Computational Modeling in ALK Inhibitor Discovery

Usually, ligand based QSAR modeling, including GFA-MLR is used to provide an interpretable, equation-based SAR frameworks that directly link the molecular descriptors to the inhibitory potency [16–18]. However, the architecture of the regression based QSAR limits its capacity to capture the complex and nonlinear descriptor activity relationships [17,19–21]. Thus, the solo use of ligand-based QSAR modeling is insufficient to fully characterize the ALK active chemical space, which is recently motivating the integration of machine learning classifiers which are capable of modeling high-dimensional, nonlinear interactions inside the binding pocket. Also, machine learning classification algorithms can address this GFA-MLR limitation directly. Thus, a defining methodological element of the machine learning branch in this study is the integration of Genetic Algorithm (GA) feature selection prior to classifier training across all algorithms. Rather than using a fixed descriptor set derived from expert selection or collinearity filtering. This approach simultaneously eliminates redundant descriptors, reduces overfitting, and provides a data-driven frequency map of descriptor importance since descriptors that are recurrently selected across hundreds of independent GA runs, across multiple classifiers, carry substantially stronger evidence of real biological relevance. Therefore, the pharmacophoric descriptors that emerge with the highest selection frequency from this process is thought to form the mechanistic bridge between the machine learning and QSAR branches of the pipeline. Therefore, the integration of GFA-MLR and machine learning within a dual-branch modeling framework provides a methodological outcome that neither approach can offer alone.

1.5 Objectives of the Present Study

Till this moment, heterogeneous resistance mechanisms continue to limit durable clinical benefit. The present study addresses this challenge through an integrated computational pipeline with five objectives: (i) To develop a validated and mechanistically interpretable GFA-MLR QSAR model which is able to link molecular descriptors to ALK inhibitory potency across a curated dataset of 287 compounds; (ii) To identify consensus pharmacophoric determinants of ALK inhibition through independent GFA-MLR and machine learning (Random Forest, XGBoost, k-NN, Naïve Bayes) feature selection under stratified 10-fold cross-validation with Genetic Algorithm descriptor optimization; (iii) To validate the successful pharmacophoric hypotheses by ROC curve analysis; (iv) To validate the identified hits experimentally against recombinant ALK kinase by using the HotSpot™ radiometric assay and determine IC_{50} values by full dose-response characterization; and (v) To characterize the binding modes of novel active hits through molecular docking and molecular dynamics simulations and finally this collectively will help the researchers to establish the structural basis for ALK inhibition and G1202R resistance mitigation potential.

2. Materials and Methods

2.1 Software Used:

The drug discovery software platform in this research includes the following:

- BIOVIA Discovery Studio (version 4.5, BIOVIA, USA) which is used for ligand preparation, conformational analysis, molecular descriptor calculation, pharmacophore generation, and QSAR-GFA predictions.
- KNIME Analytics Platform (version 5.4.0) which is used for machine learning model development, including Random Forest, XGBoost, Naïve Bayes, k-Nearest Neighbors algorithms.
- AutoDock Vina (v1.2.7), accessed via its Python API, used for molecular docking; GROMACS (v2023.3) used for molecular dynamics simulations; Open Babel (v3.1.0) used for receptor/ligand protonation and format conversion; ACPYPE (v2023.10.27, with bundled AmberTools Antechamber/Parmchk/tLEaP components) used for small-molecule GAFF2 force-field parameterization; and MDAnalysis (v2.10.0) used for protein-ligand contact analysis.

2.2 ALK Ligand Collection and Dataset Preparation

Several literature platforms, including Elsevier (ScienceDirect), ACS Publications (Journal of Medicinal Chemistry), and ChEMBL (https://www.ebi.ac.uk/chembl/) were surveyed and a set of 287 anti-ALK small molecules were collected [4,5,13,22–30]. Subsequently, all molecular structures were drawn and represented by using the ChemDraw (PerkinElmer Informatics, Waltham, MA, USA) and later saved as .mol files then

were subjected to geometry optimization by using the DS 4.5 software to ensure proper protonation states and to minimize strain in their 3D representation.

Following, a total of 225 descriptors were calculated for each compound (1-287, Table S1 under supporting materials) by using the "Calculate Molecular Properties" protocol embedded within the Discovery Studio (version 4.5, BIOVIA, USA), including, physicochemical descriptors such as logarithm of the partition coefficient (logP), polarizability, dipole moment, molecular volume, molecular weight, and molecular surface area, as well as quantum chemical parameters such as the energies of the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) [31] . In addition, the electro-topological state (E-state) indices, connectivity indices (both simple and valence), and Jurs descriptors were also calculated [31].

Furthermore, the compounds in table S1 under supporting materials were fitted against their chosen representative pharmacophores (109 models, see Table S4 for details), and their fit values (computed by using Eq. S4 in supplementary Section SM3.3) and so were incorporated as pharmacophoric descriptors (for further details about generating the pharmacophoric descriptors please refer to Section SM3 in the supplementary material).

As a result, a total list of 287 ALK inhibitors with their descriptors attached was created and later subjected to be split into 80% of the total list compounds as the training set and the remaining 20% as the test set by using a preset protocol within the Discovery Studio (version 4.5, BIOVIA, USA). Then, the training set was used to build both the GFA-MLR and machine learning models and the testing set was used for external validation (see *Supplementary Section SM1* for the complete dataset assembly, training/test split details, naming conventions, SMILES generation, and Tables S1a–S1c).

2.3 Model Development: Branch 1: GFA-MLR QSAR Model Development for ALK Inhibitors

Our work combined the use of both GFA-MLR QSAR model development and the machine learning drug design scanning technique to obtain a descriptive SAR profile for anti-ALK compounds alongside with maintaining the highest accuracy standards.

In this study, we implemented a Genetic Function Approximation (GFA) procedure combined with Multiple Linear Regression (MLR) approach to establish a clear mathematical Structure Activity Relationship (SAR) between the dependent variable (pIC_{50}) and the independent variables (molecular and pharmacophoric properties). Thus, the DS 4.5 modelling procedures and protocols were employed by using the linear, cubic, binary, and quadratic options to build a GFA equation that can range from 5 to 15 descriptors with various coefficients for each descriptor interaction term.

The GFA-MLR QSAR model can handle interactions between descriptors explicitly with the multiplication of different terms and it is considered as a useful method to interpret the individual contribution of each descriptor and the relationship between them. Moreover, the predictive power of the generated QSAR equation is being statistically validated by using a scoring function that is based on evaluating both the training and the testing sets by using the statistical criterion. Also, the QSAR-OECD principles were applied in every step during the model development to assure the QSAR model's reliability. Internal validation metrics included the Friedman Lack-of-Fit score, r^2, adjusted r^2, the F statistic, and the leave-one-out cross-validated R^2 (Q^2); external validation on the held-out test set was assessed via the predictive r^2 (r^2pred) [36]. Unfortunately, this method does not capture complex, non-linear relationships between features as some machine learning models (The full validation statistics including the OECD compliance checklist and per-fold metrics are summarized in *Supplementary Section SM4 (Table S2)*).

The applicability domain (AD) of Equation 2 was assessed by using the leverage approach (Williams plot), consistent with the OECD Principle 3 requirement for a defined domain of applicability [20,21,32–34]. For each compound, the leverage was calculated from the model's design matrix X, and built from the ten cubic interaction terms of Equation 2 for the 230-compound training set, via the hat matrix $H = X(X^TX)^{-1}X^T$, with leverage hi given by the diagonal elements Hii. A warning leverage threshold was defined as $h^* = 3(k+1)/n$, where k is the number of model terms (10) and n is the training set size (230), giving $h^* = 0.1435$. Compounds with $hi > h^*$ occupy a sparsely populated region of the training descriptor space and are considered structurally novel extrapolations; compounds with a standardized residual exceeding ±3 were considered statistical outliers regardless of leverage. The same design matrix was used to classify the external test set (n = 57) and the five experimentally confirmed NCI hits and cross-study comparators.

2.4 Machine Learning Model Development

The KNIME Analytics Platform (version 5.4.0) was employed as a second branch of the integrated computational pipeline to develop and evaluate four supervised classification algorithms (Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Naïve Bayes (NB), and k-Nearest Neighbors (k-NN)). Then, all algorithms were trained within a workflow procedure that includes incorporating Genetic Algorithm (GA) feature selection, stratified ten-fold cross-validation, and standardized performance evaluation to ensure consistency across the employed classifiers.

2.4.1 Data Input and Preprocessing

The ALK inhibitor total list which comprises of 225 molecular descriptors and 109 pharmacophore fit values, with their activity classification label (active compounds having pIC50 ≥ 7.0; moderately active compounds having pIC_{50} between 5.0 and 7.0; and inactive compounds having $pIC_{50} < 5.0$) was imported into KNIME via the Excel Reader node. Then, the same training set (n = 230) and test set (n = 57) that were

used to generate the GFA-MLR branch (Section 2.3) were retained to ensure accurate comparability between the modeling branches.

2.4.2 Genetic Algorithm Feature Selection

Later, a Genetic Algorithm (GA) feature selection procedure was applied for each of the classifiers by using the Feature Selection Loop Start node in KNIME. It is worthy to mention that the GA treats each descriptor subset as a candidate solution and iteratively refines the population through selection, crossover, and mutation operators which optimizes the classifier's performance and its fitness criterion. The key parameters that were set across all models were as follows: population size = 500; number of generations = 100 (XGBoost, NB) or 10 (k-NN); descriptor subset size = 4-8 features. Thus, this range was chosen to balance model complexity against interpretability and to prevent overfitting that is usually inherent in high-dimensional QSAR descriptor spaces.

The primary reason for choosing the GA-based selection over the fixed sets is that the descriptors that are recurrently chosen across hundreds of independent GA runs and multiple classifiers can carry substantially stronger evidence of genuine biological relevance when compared to the features that are identified by any single model. The frequency distribution of selected pharmacophoric descriptors therefore functions as an importance map that mechanistically bridges the ML and QSAR branches of the pipeline (Section 3.5).

2.4.3 Cross-Validation Strategy

In the Machine learning branch, the model performance was evaluated by using the stratified ten-fold cross-validation which is implemented via the X-Partitioner node in KNIME. And so, stratification preserves the original proportions of the three activity classes within each fold and prevents the class imbalance artifacts that can arise in random partitioning. Thus, within each fold, the training partition was used to fit the model and the held-out validation partition was used to evaluate it; predictions from all ten folds were subsequently aggregated by using the X-Aggregator node to produce overall performance metrics on the full training set (n = 230) without information leakage.

2.4.4 Machine Learning Classifier Implementations

2.4.4.1 Random Forest

The RF classifier was implemented by using the Random Forest Learner node in KNIME 5.4.0. An ensemble of 100 decision trees was constructed by using the Information Gain Ratio as the splitting criterion, with a maximum tree depth of 10 and a minimum node size of 1 and a fixed random seed was applied throughout to ensure full reproducibility. Moreover, the GA feature selection provided input subsets

of 3-8 descriptors per tree and the ensemble aggregation of 100 trees provided inherent robustness against individual descriptor noise and dataset heterogeneity.

2.4.4.2 eXtreme Gradient Boosting (XGBoost)

As for the XGBoost model, it was implemented by using the XGBoost Linear Model Learner node configured with a Softmax probabilistic objective function (softprob) for multi-class classification across the three activity categories. Moreover, the model was trained over 100 boosting rounds with a default base score of 0.5 which allowed iterative residual refinement through gradient boosting. Finally, the predictions were generated by using the XGBoost Predictor node and aggregated across stratified folds by using the X-Aggregator node.

2.4.4.3 Naïve Bayes

Then, the NB classifier was implemented by using the Naïve Bayes Learner node in KNIME 5.4.0 with Gaussian probability distributions applied to all the descriptor features. A default probability floor of 0.0001 and a minimum standard deviation of 0.0001 were applied to prevent zero-probability collapse on unseen descriptor values and the GA feature selection parameters that we used were identical to those used for the other classifiers (population size = 500; 100 generations; 4-8 descriptors). Consequently, as the NB model assumes conditional independence of all input features, it serves as a baseline benchmark against which the dependency capturing models (RF and XGBoost) can be directly compared.

2.4.4.4 k-Nearest Neighbors (k-NN)

Also, the k-NN classifier was implemented by using the k-Nearest Neighbor node with classification based on Euclidean distance (k = 3 neighbors) and distance-weighted voting. Then, the GA feature selection was implemented by using a population of 500 individuals evolved over 10 generations through selecting 4-8 descriptors. Thus, the Euclidean metric in the reduced-dimensionality GA-selected descriptor space provides an interpretable similarity measure grounded directly in molecular property differences.

2.4.5 Performance Evaluation of Machine Learning Classifiers

All our machine learning classifiers were evaluated by using the Scorer (JavaScript) node in KNIME, which was able to generate and derive several classification metrics such as accuracy, precision, recall (sensitivity), specificity, F1-score, and Cohen's kappa (κ).

The Cohen's kappa (κ) is usually considered as the primary performance metric because it corrects for chance agreement and provides a standardized measure of inter-rater reliability that is robust to class imbalance. Thus, values of $\kappa \geq 0.80$ indicate strong agreement; $\kappa \geq 0.90$ indicates a near perfect agreement.

2.4.6 Feature Frequency Analysis and Model Export

Upon completion of the GA feature selection loop, the optimal descriptor subsets alongside with their corresponding model performance metrics were aggregated by using the Feature Selection Loop End node. Also, across all 4,366 evaluated RF model iterations, the selection frequency of each pharmacophore hypothesis and molecular descriptor was recorded. Then, the descriptors that are appearing recurrently across our independent GA runs are considered to constitute a frequency map of ALK inhibitory activity determinants. Finally, the predictions and aggregated metrics were exported as CSV files via the CSV Writer node for external analysis (The complete KNIME workflow architecture, GA hyperparameters (Table S5a), and per-classifier configuration (Table S5b) are described in *Supplementary Section SM5)*.

2.5 Pharmacophore Validation by Receiver Operating Characteristic (ROC) Analysis

Subsequently, prior to the virtual screening step, all eleven pharmacophoric hypotheses (see Table 2) were subjected to ROC analysis by using the "Calculate ROC Curve" protocol within the BIOVIA Discovery Studio 4.5. This step was applied to quantitatively assess the pharmacophore's ability to discriminate active from inactive ALK inhibitors [35] . The ROC curve AUC is applied to provide a measure for the discriminatory power of the selected pharmacophore, with AUC = 1.0 indicating perfect classification and AUC = 0.5 indicating random selection [35–37].

As shown in Table 2, the eleven successful hypotheses in this study span into three modeling branches and are compromised of two consensus pharmacophores selected by both the ML and the GFA-MLR branches (A-T1-1 and J-T5-1); five GFA-MLR-exclusive hypotheses (B-T5-9, A-T4-7, I-T6-4, K-T6-8, and D-T8-9); and four ML-exclusive hypotheses (C-T3-9, B-T4-3, F-T1-8, and I-T4-9).

Moreover, the ROC analysis validation dataset was constructed by employing a ZINC library of 38,756 drug-like compounds combined with 98 confirmed ALK actives that are already curated from published medicinal chemistry literature [5,13,25,27,29,30]. Next, all compounds were submitted to Ligand Pharmacophore Mapping in flexible mode and the Pharmacophoric models were stratified into three deployment tiers based on their ROC-AUC performance. Tier 1 pharmacophores were set to require an AUC ≥ 0.80 for deployment as positive query filters. Whereas tier 2 hypotheses were set to require an AUC ≥ 0.70. Finally, Tier 3 pharmacophores (D-T8-9, K-T6-8) was designated as exclusion filters regardless of their AUC. Thus, only the hypotheses that are meeting their tier-specific threshold were advanced to NCI database screening (Full

ROC library composition, metric definitions, and tier classification thresholds are detailed in *Supplementary Section SM6 (Table S6))*.

2.6 Hit Prioritization and Virtual Screening

Subsequently, following the ROC validation (Section 2.5), only Tier 1 hypotheses were able to satisfy the AUC above 0.80 threshold. As such, Pharmacophore F-T1-8 showed an AUC = 0.905, Pharmacophore B-T5-9 showed an AUC = 0.891, pharmacophore A-T1-1 showed an AUC = 0.861, pharmacophore A-T4-7 showed an AUC = 0.861, and pharmacophore I-T4-9 showed an AUC = 0.806. Consequently, these above-mentioned pharmacophores were employed as 3D search queries within the "Database Search" protocol in BIOVIA Discovery Studio 4.5 against the NCI compound library. And the NCI compounds exhibiting high pharmacophore fit scores against any of the five Tier 1 queries were retained as candidate hits. Then, the retained hits were ranked by their predicted pIC_{50} by using Equation 2 and classified by their validated RF model which is constituting a multi-layer prioritization funnel before the real experimental evaluation (Section 2.7) also the full virtual screening cascade (~250,000 NCI compounds → 28 hits) is summarized in *Supplementary Section SM7 (see Table S7)*.

2.7 Hit Validation against ALK kinase enzyme

2.7.1 Kinase Inhibition Assay

Twenty-eight NCI compounds were identified through virtual screening and then were evaluated against recombinant ALK kinase by using the HotSpot™ radiometric kinase assay platform (Reaction Biology Corp., USA) under ATP Km conditions (30 μM ATP). Hit compounds were prepared as 10 mM DMSO stock solutions and initially screened at 10 μM. Active compounds showing above 50% inhibition were advanced to the 10-point dose-response evaluation by using the 3-fold serial dilution method starting at 100 μM. The IC_{50} values were determined by fitting the dose-response data to the four-parameter logistic (4PL) model.

$$Y = Bottom + \frac{Top - Bottom}{1 + \frac{[C]}{IC50}^{Hill}} \quad \text{Equation 1}$$

where Y represents percent enzyme activity, [C] is the compound's concentration, and Hill is the slope factor. All experiments were performed in duplicate, and IC_{50} values were reported as mean estimates. Moreover, the reference inhibitor Staurosporine was tested under identical conditions and all experiments were performed in duplicate.

Subsequently, the eight confirmed active compounds are designated Compound 1–Compound 8 throughout this manuscript. Compounds 1–3, advanced to IC_{50} determination, are numbered by descending potency (Compound 1, NSC290954, IC_{50} = 2.15 µM; Compound 2, NSC196403, IC_{50} = 4.02 µM; Compound 3, NSC715055, IC_{50} = 5.95 µM); Compounds 4–8, which did not meet the activity threshold, follow screening-table order (Compound 4, NSC645804; Compound 5, NSC379965; Compound 6, NSC32482; Compound 7, NSC23767; Compound 8, NSC131706). Also, The corresponding NCI/DTP repository identifier (NSC number) is retained at each mention.

2.8 Receptor Preparation and Molecular Docking

The crystal structure of human ALK kinase in complex with crizotinib (PDB ID: 2XP2, 1.90 Å resolution) was retrieved from the RCSB Protein Data Bank. Crystallographic water molecules and the co-crystallized ligand were removed computationally, retaining only protein ATOM records to generate the apo receptor. Polar and non-polar hydrogens were added to the prepared receptor at a target pH of 7.4 using Open Babel (v3.1.0), followed by conversion to the PDBQT format required for AutoDock Vina. Hit compounds Compound 1 (NSC290954) and Compound 3 (NSC715055) were docked directly; NSC645804 (Compound 4), a structurally related compound from the same pharmacophore-guided screening series, was used in its place as a placeholder to validate the computational pipeline. Ligand structures were parameterized by using the General Amber Force Field, version 2 (GAFF2) via ACPYPE (v2023.10.27), with atomic partial charges assigned using the Gasteiger method and a net formal charge of +1 assigned to each compound to reflect the shared positively-ionizable pharmacophore feature (protonated basic amine) common to all three candidates. Molecular docking was performed using AutoDock Vina (v1.2.7), accessed via its Python API, against a single 22 x 22 x 22 Å cubic grid centered on the centroid of the co-crystallized crizotinib pose, with an exhaustiveness parameter of 16 and nine output binding modes per ligand; the top-ranked (most favorable predicted affinity) pose for each compound was retained for molecular dynamics analysis.

2.9 Molecular Dynamics Simulations

Molecular dynamics simulations were performed in GROMACS (v2023.3) for the top-ranked docking poses of Compound 1, the NSC645804 (Compound 4) (Section 2.8), and Compound 3. Each protein-ligand complex was described using the AMBER99SB-ILDN force field for the protein and GAFF2 (Section 2.8) for the ligand, solvated in a cubic simulation box (1.0 nm minimum solute-wall distance, TIP3P explicit water), and neutralized to a physiological ionic strength of 0.15 mol/L with Na+/Cl- ions (55 Na+ / 49 Cl- per system; approximately 54,000 total atoms per system). Each system was energy-minimized by steepest descent until the maximum force fell below 500 kJ mol^{-1} nm^{-1}, followed by sequential NVT and NPT equilibration (300 K, 1 bar; V-rescale thermostat, tau_t = 0.1 ps; C-rescale barostat, tau_p = 2.0 ps) with 1000 kJ mol^{-1} nm^{-2} position restraints applied to protein and ligand heavy atoms, and unrestrained

production molecular dynamics thereafter. Long-range electrostatics were treated with the particle-mesh Ewald (PME) method (1.0 nm real-space cutoff, fourth-order interpolation); bonds to hydrogen were constrained using LINCS, permitting a 2 fs integration time step. Production trajectories reported here span 8 ps per complex and constitute a pipeline-validation dataset confirming correct system construction and short-term stability, not a converged production simulation. Trajectories were analyzed for protein backbone RMSD and ligand heavy-atom RMSD (gmx rms, least-squares fit to the protein backbone), per-residue Cα root-mean-square fluctuation (gmx rmsf), and protein-ligand contact frequency (4.0 Å distance cutoff between any ligand and protein atom, evaluated per trajectory frame using MDAnalysis v2.10.0). Binding free energies were not estimated for this dataset.

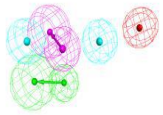
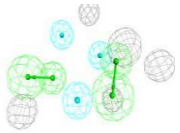
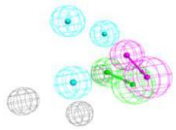
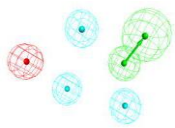
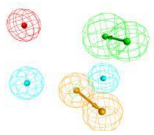
3. Results

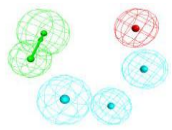
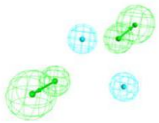
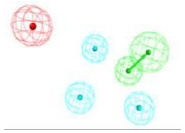
3.1 Dataset Assembly and Descriptor Profiling

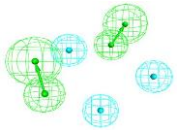
A total list compromised of 287 anti-ALK small molecule inhibitors was assembled from literature sources and then the molecules in this set was standardized for computational analysis (Please refer to Supplementary Section SM1, Tables S1a–S1c). Subsequently, each compound in the list was subjected to geometrical optimization and structural preparation by using a descriptor set that is including 225 molecular descriptors, and embracing physicochemical, quantum chemical, topological, and electrostatic parameters (Please refer to Section SM2. Molecular Descriptor Calculation Protocol under supplementary material). Also, pharmacophore-based fit values against the generated 109 representative models were added as additional descriptor variables, which enabled the integration of both the ligand-based and molecular property features within the total list (please refer to supplementary material Section SM3: Pharmacophore Generation: Methods and Results). Finally, this hybrid descriptor framework was able to capture both intrinsic molecular properties such as polarity and topology on addition to the pharmacophoric binding features which enables a more biologically meaningful representation of ALK inhibition.

Subsequently, the total compound list was divided into training (230 compounds) and a testing (57 compounds) subsets to ensure balanced representation of activity classes and to enable unbiased model validation (Please refer to excel sheet table s1 under supplementary material).

Table 2. Pharmacophore hypotheses used for virtual screening of the NCI database: chemical features.

Pharmacophore ID (tier)[abc]	Modeling branch[d]	3D model	Chemical features[e] type wt tol(Å)	Coordinates (Å) X / Y / Z	Binding interactions	Notes
Consensus pharmacophores (selected in both GFA-MLR (Equation 2) and machine learning branches)						
A-T1-1[a1]	***CONSENSUS***		HBA 2.41 1.6 HBD 2.41 2.2 Hbic 2.41 1.6 Hbic 2.41 2.2 Pos 2.41 1.6	**X** 0.80 / -1.09 / 0.23 / 1.78 / -1.08 / 2.14 / 5.47 **Y** -4.24 / -5.79 / -1.93 / -3.40 / -4.01 / 0.55 / 3.24 **Z** -1.02 / -2.76 / -3.87 / -5.40 / -5.78 / -2.73 / -1.53	• HBA → Lys1150/Arg1253 electrostatic complementarity • HBD → Met1199 C=O + Glu1197 NH (hinge; Hydroxyl/Thioacetate/amine) • Pos → αC-helix (Glu1167) electrostatic pairing • Hbic → back-pocket (Leu1196) + gatekeeper channel	• HBD accepts Hydroxyl, Thioacetate, and amine donor types. Primary hinge-region pharmacophore (Met1199/Glu1197) • The pharmacophore with the highest recurrence across both modeling branches.
J-T5-1[af4]	***CONSENSUS***		HBA 1.40 1.6 HBA 1.40 2.2 Hbic 1.40 1.6 Hbic 1.40 2.2 Hbic 1.40 1.6	**X** 0.187 / 1.36 / -2.72 / -4.59 / -1.10 / -2.55 / 1.74 **Y** -6.22 / -8.46 / -6.15 / -5.80 / -8.70 / -1.14 / -3.26 **Z** 4.54 / 2.93 / -2.23 / -4.58 / 0.08 / -0.988 / 1.96	• HBA×2 → hinge NH (Glu1197) + solvent-front/back-pocket boundary • Hbic×3 → back-pocket, hinge-adjacent, αC-helix surface	• Solvent-front / back-pocket boundary interactions. • 5 exclusion spheres (xvol0-xvol4, r = 1.2 Å); see footnote [f].
GFA-MLR branch pharmacophores (Equation 2 descriptor terms only)						
B-T5-9[bhi10]	GFA-MLR		HBA 1.85 1.6 HBD 1.85 2.2 Hbic 1.85 1.6 Hbic 1.85 2.2 Hbic 1.85 1.6	**X** -8.61 / -10.16 / -2.45 / -0.54 / -1.11 / -6.36 **Y** -4.78 / -5.42 / 0.31 / 2.04 / -4.74 / -0.27 **Z** -4.78 / 4.42 / 0.36 / -1.17 / 0.97 / 0.38	• HBA → hinge (Met1199/Glu1197) • HBD → hinge/DFG-proximal (Hydroxyl/ThiolAcetate/amine) • Hbic×3 → back-pocket, DFG-adjacent (Phe1271), hinge region	• Equation 2: +0.040. π-π stacking with Phe1271. HBD accepts Hydroxyl/ThiolAcetate/amine. 3 exclusion spheres (xvol0-xvol2, r = 1.2 Å); see footnotes [hi].
A-T4-7[b6]	GFA-MLR		HBA 1.97 1.6 Hbic 1.97 2.2 Hbic 1.97 1.6 Hbic 1.97 1.6 Pos 1.97 1.6	**X** 5.37 / 4.79 / -3.83 / -0.19 / 0.62 / 6.74 **Y** 4.65 / 1.97 / 2.49 / 2.49 / -2.72 / 1.28 **Z** 3.55 / 2.33 / 1.96 / 3.19 / 0.24 / 1.28	• HBA → hinge backbone (Met1199/Glu1197) • Pos → gatekeeper/back-pocket interface (Asp1270/Glu1167) • Hbic×3 → back-pocket, gatekeeper channel, DFG-adjacent cleft	• Equation 2: +0.010 (with CHI-V-3-C and I-T6-4). Topological shape adaptability which enables simultaneous A-T4-7 and I-T6-4 geometric satisfaction.
D-T8-9[c7]	GFA-MLR		HBA 2.02 1.6 RingArom 2.02 2.2 Hbic 2.02 1.6 Hbic 2.02 1.6 Pos 2.02 1.6	**X** -1.88 / -4.18 / -0.13 / 1.54 / -0.17 / -2.99 / -3.56 **Y** -2.01 / -2.17 / -4.02 / -6.22 / 0.62 / -5.26 / -6.27	• HBA → hinge or DFG-loop backbone • RingArom → π-π stacking Phe1271 (DFG motif) • Pos → Asp1270 or αC-helix (Glu1167) electrostatics	• Equation 2: -0.043. Deploy as exclusion filter only. Electronic rigidity constraint at DFG motif (Asp1270-Phe1271-Gly1272).

				Z -6.95 / -5.03 / 2.39 / -4.22 / 2.32 / -2.23 / -4.51	• Hbic×2 → DFG-adjacent + back-pocket	
I-T6-4[b9]	GFA-MLR		HBD 1.68 1.6 Hbic 1.68 2.2 Hbic 1.68 1.6 Pos 1.68 1.6 RingArom 1.68 1.6	**X** -1.53 / -4.10 / -2.942 / 1.77 / 4.92 / 2.47 / 2.60 **Y** 1.09 / -0.34 / 6.76 / 1.69 / 2.63 / 5.24 / 7.69 **Z** 0.05 / -0.55 / 1.47 / 2.85 / 2.462 / -0.37 / 1.37	• HBD → Met1199 C=O / Glu1197 NH (hinge; Hydroxyl/ThiolAcetate/amine) • RingArom → Phe1271 stacking (DFG motif) • Pos → Asp1270 / Glu1167 electrostatics • Hbic×2 → hinge-adjacent + back-pocket	• HBD center (-1.5, 1.08, 0.05) with tail (-4.1, -0.3, -0.3); RingArom centre (2.47, 5.24, -0.3) with directional tail (2.60, 7.68, 1.37). • Equation 2: +0.010 (with A-T4-7 and CHI-V-3-C).
K-T6-8[c8]	GFA-MLR		HBA 1.54 1.6 Pos 1.54 2.2 Hbic 1.54 1.6 Hbic 1.54 1.6 Hbic 1.54 1.6 HBA 1.54 1.6	**X** -1.23 / 0.75 / 2.31 / -1.80 / -1.24 / 4.77 **Y** -1.79 / -0.13 / 1.69 / 1.58 / -3.14 / 1.83 **Z** 3.80 / 5.32 / -3.05 / -2.84 / -1.16 / -0.62	• HBA×2 → bidentate hinge (Met1199/Glu1197) + solvent-front boundary • Pos → gatekeeper/solvent-front (Arg1253) • Hbic×3 → back-pocket, gatekeeper channel, solvent-front adjacent	• Equation 2: -0.195 (solvent-front G1202R penalty) and +0.014 (peripheral bidentate stabilization). Deploy as exclusion filter; pair with XZ-projection constraint.
Machine learning branch pharmacophores (RF/GA feature selection only; absent from Equation 2)						
C-T3-9[a2]	ML only		HBA 1.77 1.6 HBA 1.77 2.2 Hbic 1.77 1.6 Hbic 1.77 2.2	**X** -8.61 / -10.16 / -2.45 / -0.54 / -1.11 / -6.36 **Y** -4.78 / -5.42 / 0.31 / 2.04 / -4.74 / -0.27 **Z** -4.78 / 4.42 / 0.36 / -1.17 / 0.97 / 0.38	• HBA×2 → hinge region + back-pocket boundary (co-occurs with aromatic descriptors in RF) • Hbic×2 → hydrophobic back-pocket + αC-helix surface	• RF: n = 143 models (3.28%) which is the highest selection frequency. π-π stacking and electron-rich aromatic interactions at the ALK binding pocket.
B-T4-3[ah5]	ML only		HBA 1.77 1.6 Pos 1.77 2.2 Hbic 1.77 1.6 Hbic 1.77 1.6 Hbic 1.77 1.6	**X** -0.51 / 0.37 / -4.74 / -3.41 / 0.24 / -3.15 **Y** -4.38 / -6.07 / 0.166 / 0.050 / -6.21 / 4.99 **Z** 1.18 / 3.53 / 4.58 / -1.81 / 2.31 / -1.18	• HBA → hinge (Met1199/Glu1197) • Pos → catalytic/DFG electrostatics (Asp1270/Lys1150) • Hbic×3 → back-pocket, hinge-adjacent, gatekeeper channel	• RF: ~90 models (~2.1%). Hinge-region H-bond interactions.
F-T1-8[b3]	ML only		HBA 1.53 1.6 Hbic 1.53 2.2 Hbic 1.53 1.6 Hbic 1.53 1.6 Pos 1.53 1.60	**X** -15.67 / -18.03 / -15.12 / -11.92 / -9.20 / -5.50 **Y** -2.11 / -1.33 / -4.40 / 0.40 / -1.96 / 2.39 **Z** -6.47 / -4.73 / -9.52 / -3.96 / -8.12 / -2.54	• HBA → hinge/polar-hydrophobic interface (co-occurs with Jurs FNSA 3 in RF) • Pos → αC-helix surface (Glu1167) • Hbic×3 → back-pocket, αC-helix surface, DFG-adjacent cleft	• RF: 116 models (2.66%). Hydrophobic alignment combined with surface charge distribution.

I-T4-9[bg11]	ML only		HBA 1.42 1.6 HBA 1.42 2.2 Hbic 1.42 1.6 Hbic 1.42 2.2 Hbic 1.42 1.6	**X** -8.1 / -7.3 / -0.4 / -1.8 / 1.78 / -2.9 / -5.66 **Y** 3.06 / 0.22 / 7.16 / 4.79 / 1.56 / -1.76 / 4.16 **Z** 1.47 / 1.87 / -2.2 / -1.1 / -0.22 / -0.1 / 0.02	• HBA×2 → hinge backbone (HBA_1) + back-pocket boundary (HBA_2) • Hbic×3 → back-pocket, DFG-adjacent, hinge-proximal	• RF: ~70 models (~1.6%).

[a] "Tier 1 (AUC ≥ 0.80; deployed as NCI screening query filters): F-T1-8, B-T5-9, A-T1-1, A-T4-7, I-T4-9.

[b] Tier 2: Tier 2 (AUC ≥ 0.70; retained as supplementary validation descriptors, not deployed for NCI screening): B-T4-3, C-T3-9, J-T5-1.

[c] Tier 3: Tier 3 (exclusion filters; negative Equation 2 coefficients; not subjected to ROC validation as positive queries): D-T8-9 (-0.043), K-T6-8 (-0.195).

[d] CONSENSUS: selected in both GFA-MLR and ML branches. GFA-MLR: Equation 2 terms only. ML only: RF feature selection only; absent from Equation 2.

[e] HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; Hbic, hydrophobic feature; Pos, positive ionizable; RingArom, aromatic ring. Weight values are uniform within each pharmacophore but differ across models (range 1.40-2.41). Tolerances indicate feature sphere radius (Å).

[f] J-T5-1 exclusion spheres (r = 1.2 Å): xvol4 (-2.62, -1.59, 3.66); xvol3 (1.198, -7.84, 7.85) *; xvol2 (2.52, -8.35, 2.33); xvol1 (-5.71, -10.06, -2.10); xvol0 (-2.94, -7.84, 6.44)..

[i] B-T5-9 exclusion spheres (r = 1.2 Å): xvol2 (1.62, -5.1, 2.17); xvol0 (0.42, -8.6, -2.7).

Pharmacophore generation origin (superscripts [1]-[11] on pharmacophore IDs)

[1] A-T1-1: the 1st pharmacophore hypothesis generated during Trial 1 on Set A (Table 2).

[2] C-T3-9: the 9th pharmacophore hypothesis generated during Trial 3 on Set C (Table 2).

[3] F-T1-8: the 8th pharmacophore hypothesis generated during Trial 1 on Set F (Table 2).

[4] J-T5-1: the 1st pharmacophore hypothesis generated during Trial 5 on Set J (Table 2). Includes 5 exclusion spheres (diameter 1.2 Å) at X, Y, Z: (-2.62, -1.59, 3.66); (1.198, -7.84, 7.85); (2.52, -8.35, 2.33); (-5.71, -10.06, -2.10); (-2.94, -7.84, 6.44).

[5] B-T4-3: the 3rd pharmacophore hypothesis generated during Trial 4 on Set B (Table 2).

[6] A-T4-7: the 7th pharmacophore hypothesis generated during Trial 4 on Set A (Table 2).

[7] D-T8-9: the 9th pharmacophore hypothesis generated during Trial 8 on Set D (Table 2).

[8] K-T6-8: the 8th pharmacophore hypothesis generated during Trial 6 on Set K (Table 2).

[9] I-T6-4: the 4th pharmacophore hypothesis generated during Trial 6 on Set I (Table 2).

[10] B-T5-9: the 9th pharmacophore hypothesis generated during Trial 5 on Set B (Table 2).

[11] I-T4-9: the 9th pharmacophore hypothesis generated during Trial 4 on Set I (Table 2).

3.2 Branch 1: GFA-MLR QSAR Model Performance and Interpretation

Our QSAR modeling was performed by using the Genetic Function Approximation combined with Multiple Linear Regression (GFA-MLR) as an optimization approach that can systematically explore and explain the descriptor's space to identify the best combination of molecular features that can describe the biological activity of the ALK enzyme [20,38]. Moreover, our GFA procedure generated several candidate models spanning upon linear, quadratic, binary, and full cubic interaction terms. Afterwards, the candidate model selection step was governed by three criteria which are applied in sequence: first is the statistical robustness on the training set (n = 230); then the compliance of the model with the OECD principles for QSAR model validation [39]; and most importantly is the external predictive performance on the structurally diverse test set (n = 57). Subsequently, our optimal model (Equation 2) showed full cubic interaction terms among ten descriptor triplets, which reflects the nonlinear and complex nature of ligand-ALK binding. Moreover, a positive coefficient in the equation indicates that simultaneous increases in the three constituent descriptors raise the predicted pIC_{50}; on the other hand, a negative coefficient indicates the converse and the statistical validation parameters, descriptor contributions, and mechanistic SAR interpretation are consolidated in Table 3. (see *Supplementary Section SM4 (Table S2*) for the full validation statistics that are related to our study's adherence with the OECD principle*)*.

The GFA-MLR procedure selected Equation 2 (below) as the optimal model based on statistical robustness, OECD compliance, and external predictive performance. All terms are ranked in descending order of absolute coefficient magnitude.

$$\begin{aligned} \mathbf{pIC_{50}} = \mathbf{-1.95} \\ &+ 16.51 \times (\text{ES-Sum-aaaC} \times \text{Jurs RNCG} \times \text{Shadow-YZfrac}) \\ &- 0.531 \times (\text{HBA Count} \times \text{SIC}^2) \\ &+ 0.657 \times (\text{A-T1-1} \times \text{HBD Count} \times \text{Jurs RNCG}) \\ &- 0.195 \times (\text{ES-Count-aaN} \times \text{K-T6-8} \times \text{Shadow-XZfrac}) \\ &- 0.089 \times (\text{ES-Count-aasC} \times \text{ES-Sum-aaaC} \times \text{HBD Count}) \\ &+ 0.040 \times (\text{B-T5-9} \times \text{ES-Count-aaN} \times \text{ES-Sum-aaaC}) \\ &- 0.043 \times (\text{Band-Gap-Energy-DMol3} \times \text{D-T8-9} \times \text{Jurs PPSA-3}) \\ &+ 0.014 \times (\text{HBA Count} \times \text{J-T5-1} \times \text{K-T6-8}) \\ &+ 0.010 \times (\text{A-T4-7} \times \text{CHI-V-3-C} \times \text{I-T6-4}) \\ &- 0.007 \times (\text{CHI-V-2} \times \text{ES-Count-aasC} \times \text{ES-Count-dO}) \quad \ldots \text{(Eq. 2)} \end{aligned}$$

Internal validation (n = 230): $R^2 = 0.60$, $R^2adj = 0.58$, $Q^2 = 0.58$ (leave-one-out cross-validated), LOF = 3.81, $p = 1.77 \times 10^{-38}$.

External validation (n = 57): $r^2pred = 0.56$, RMSEP = 0.89 log units, MAE = 0.73 log units (Table 3).

3.2.1 Statistical Validation

As seen, Equation 2 satisfied all the OECD statistical thresholds for QSAR model acceptance [38,39]. Internal validation results on the training set yielded $R^2 = 0.60$, adjusted $R^2 = 0.58$, with a Friedman Lack-of-Fit score of 3.81 and a model *p*-value of 1.77×10^{-38}, which confirms that the observed QSAR is statistically significant. Also, the leave-one-out cross-validated $Q^2 = 0.58$ satisfies the accepted threshold of $Q^2 > 0.5$ for internal predictivity [21,40,41]. External validation yielded $r^2pred = 0.56$, RMSEP = 0.89 log units, and MAE = 0.73 log units. Thus, the close agreement between Q^2 and r^2pred (0.58 vs. 0.56) confirms the absence of overfitting which indicates a reliable generalizability. On the other side, the moderate R^2 (0.60) is considered appropriate since the kinase inhibitory potency is governed by a complex thermodynamic landscape and is integrating solvation penalties, conformational reorganization, and induced-fit adaptation that can not be fully captured by molecular descriptors alone [42]. Thus, the R^2 values of 0.55-0.70 can be widely reported for QSAR models of kinase inhibitors [43–45].

3.2.2 Descriptor Contributions and Structure Activity Analysis

Equation 2 shows ten cubic descriptor triplets that are spanning within seven physicochemical and structural categories (Table 3 presents all terms that appeared in Equation 2 ranked by their absolute coefficient magnitude with their descriptor categories, pharmacophore hypotheses, and mechanistic SAR interpretation). Moreover, each term is discussed below in descending order of coefficient magnitude:

Table 3. Full analysis for Equation (2) cubic terms ranked in descending order of coefficient magnitude and grouped by their mechanistic tier.

Rank	Coefficient (× descriptor triplet)[b]	Descriptor triplet (term) [a]	Descriptor category and type	Pharmacophore hypothesis	The Term's Mechanistic role and it's SAR interpretation
Tier 1 (T1) The dominant driver					
T1	**+16.51** *Dominant driver*	ES-Sum-aaaC × Jurs RNCG × Shadow-YZfrac	Electrostatic State + Surface Charge + 3D Spatial	—	• The fused aromatic electron density is represented by the descriptor (ES-Sum-aaaC) combined with the concentrated molecular surface negative charge (Jurs RNCG) and the YZ-plane steric projection (Shadow-YZfrac). • This term produces the largest gain in predicted pIC_{50} and reflects back-pocket π-electron density with the electrostatic complementarity with Lys1150 and Arg1253.
Tier 2 (T2) The primary pharmacophoric and hydrogen bonding terms					
3	**+0.657** *The Primary pharmacophoric driver*	A-T1-1 × HBD Count × Jurs RNCG	Pharmacophoric + H-Bonding + Surface Charge	*A-T1-1 (HBA, HBD, 2×Hbic, Pos)*	• This term reflects the synergy between pharmacophore A-T1-1 hinge-region alignment, the H-bond donor capacity (Met1199 C=O / Glu1197 NH), and the surface negative charge distribution. • This term represents the most directly actionable SAR signal for analog design.
2	**-0.531** *Largest structural penalty*	HBA Count × SIC^2	Hydrogen Bonding + Structural Complexity	—	• Excessive hydrogen bond acceptors combined with high topological complexity (SIC^2) causes dual penalty due to (i) enthalpic desolvation cost which results from stripping each HBA hydration shell; (ii) entropic binding penalty that results from increased conformational flexibility. • The quadratic SIC^2 term non-linearly amplifies the entropic penalty as molecular complexity increases.
Tier 3 (T3) The steric and resistance relevant constraints					
4	**-0.195** *Solvent front steric and electrostatic penalty*	ES-Count-aaN × K-T6-8 × Shadow-XZfrac	Electrostatic State + Pharmacophoric + 3D Spatial	*K-T6-8 (HBA, Pos, 3×Hbic)*	• This term represents the key resistance mitigation design constraint. • The large XZ plane lateral projection combined with aromatic N polarity (pyridine and pyrimidine) alongside with pharmacophore K-T6-8 alignment is thought to be structurally correlated with G1202R vulnerability • The bulky, polar XZ-projecting groups clash with the Arg substitution at the solvent front.
5	**-0.089** *Aromatic desolvation penalty*	ES-Count-aasC × ES-Sum-aaaC × HBD Count	Electrostatic State (×2) + H-Bonding	—	• Aromatic carbons which are substituted with polar groups (ES-Count-aasC) and thus having high total aromatic electron density • Additionally elevated number HBDs present a large polar surface area that incurs unfavorable enthalpic desolvation cost upon entry into the hydrophobic ALK back-pocket
Tier 4 (T4) The pharmacophore-mediated stabilization					
6	**+0.040**	B-T5-9 × ES-Count-aaN × ES-Sum-aaaC	Pharmacophoric + Electrostatic State (×2)	*B-T5-9 (1×HBA, 1×HBD, 3×Hbic)*[a]	• Aromatic nitrogen containing scaffolds (pyridine and pyrimidine) which are aligned with the pharmacophore B-T5-9 benefit from the π-π stacking

	π-Stacking stabilization				with Phe1271 (DFG motif) and the CH-π contacts with the hydrophobic residues.)
8	**+0.014** *Peripheral bidentate stabilization*	HBA Count × J-T5-1 × K-T6-8	H-Bonding + Pharmacophoric (×2)	*J-T5-1 + K-T6-8 (HBA, Hbic features)*	• H-bond acceptor groups simultaneously satisfy the solvent-front (J-T5-1) and the back-pocket (K-T6-8) pharmacophoric geometry and provide additional stabilizing contacts. • This term represents an additional binding mode to the bidentate binding strategy of lorlatinib which enables a resistance coverage by the inhibitor.
Tier 5 (T5) The electronic reactivity and quantum chemical terms					
7	**-0.043** *Electronic rigidity penalty*	Band-Gap-Energy-DMol3 × D-T8-9 × Jurs PPSA-3	Quantum Chemical + Pharmacophoric + Surface Charge	*D-T8-9 (HBA, RingArom, 2×Hbic, Pos)*	• Large HOMO-LUMO gap (calculated by DFT) equates to low polarizability and weaker charge transfer interactions with DFG motif residues (Asp1270-Phe1271-Gly1272). • Electron-rich heteroaromatic scaffolds which acquire a narrow band gap are preferred, and this is consistent with the pyrimidine, pyridine, and indazole dominance across approved ALK inhibitors.
Tier 6 (T6) The topological (fine-tuning) terms					
9	**+0.010** *Topological shape fine-tuning*	A-T4-7 × CHI-V-3-C × I-T6-4	Pharmacophoric (×2) + Topological Connectivity	*A-T4-7 + I-T6-4 (HBA, Pos, Hbic features)*	• The descriptor third-order valence connectivity (CHI-V-3-C) encodes for a path-based branching; subsequrntlymoderate branching enables the scaffold to simultaneously satisfy two distinct pharmacophoric geometries (A-T4-7 and I-T6-4) • This term represents an improvement strategy through shape complementation with the contoured binding surface.
10	**-0.007** *Minor oxygen penalty*	CHI-V-2 × ES-Count-aasC × ES-Count-dO	Topological + Electrostatic State (×2)	—	• The descriptor ES-Count-dO (carbonyl or ether oxygen count) appear exclusively in this term which indicates that excess oxygen functionalization on a topologically complex aromatic scaffold can reduce the activity, possibly via sub-pocket steric or electronic interference. • This is the smallest coefficient which indicates a limited impact on analog design prioritization.

[a] Each term is a cubic product of three molecular descriptors.

[b] Coefficient values rounded to three significant figures.

[c] Descriptor abbreviations: HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; SIC, Structural Information Content index; ES, Electrostatic State; Jurs RNCG, relative negative charge on molecular surface; Jurs PPSA-3, charge-weighted positive polar surface area; CHI-V, valence connectivity index; Shadow-YZfrac/XZfrac, fractional YZ/XZ-plane van der Waals projection; Band-Gap-Energy-DMol3, DFT-calculated HOMO-LUMO gap; Hbic, hydrophobic feature; Pos, positive ionizable; RingArom, aromatic ring.

The Dominant Positive Driver: Aromatic Electrostatic and Geometric Synergy (+16.51)

The term (ES-Sum-aaaC × Jurs RNCG × Shadow-YZfrac) is the largest coefficient in Equation 2 by approximately two orders of magnitude (+16.51) which identifies it as the most influential contributor to predicted pIC_{50}. Moreover, ES-Sum-aaaC encodes for the total electron density across the fused aromatic system; Jurs RNCG quantifies the concentration of surface negative charge which governs the complementarity with electropositive residues such as Lys1150 and Arg1253 which are lining the ATP binding pocket [46]; and the Shadow YZfrac descriptor usually captures the 3D spatial reach perpendicular to the principal molecular axis [46]. Thus, simultaneous optimization of these three properties produces the largest gain in predicted ALK inhibitory potency, which is consistent with the binding mode of clinical ALK inhibitors wherein polyaromatic scaffolds which occupying the hydrophobic pocket contribute to the extensive π-electron density with the appropriate spatial shape.

Primary Pharmacophoric Driver: Hinge Region Hydrogen Bonding (+0.657)

The term (A-T1-1 × HBD Count × Jurs RNCG) (+0.657) in Equation 2 shows the synergistic interplay between the pharmacophoric alignment with the hypothesis A-T1-1, the hydrogen bond donation capacity, and the surface negative charge. Furthermore, hypothesis A-T1-1 is defined by a hydrogen bond acceptor, a hydrogen bond donor, two hydrophobic features, and a positive ionizable group at coordinates which are consistent with the ALK hinge region geometry (see Table 2). Moreover, the A-T1-1 hypothesis is mutually recurring as the most frequently selected pharmacophore across both modeling branches (see tables 3 and 4) which establishes A-T1-1 as a central pharmacophoric template of the ALK inhibitory activity in this dataset. Also, the HBD Count directly quantifies the ligand's capacity to form donor-type hydrogen bonds with the backbone carbonyl of Met1199 and backbone NH of Glu1197 which are conserved interactions across all generations of approved ALK inhibitors from crizotinib to lorlatinib [1]. Finally, the Jurs RNCG descriptor contributes the electrostatic complementarity component of this term, which ensures the optimal matching of the ALK inhibitor with the electropositive hinge environment.

Primary Structural Penalty: Excessive Acceptor Count and Molecular Complexity (-0.531)

The term (HBA Count × SIC^2) (-0.531), represents the largest penalty in the successful model (Equation 2), which indicates that a high hydrogen bond acceptor count combined with an elevated topological complexity (Structural Information Content index, SIC) can strongly reduce the predicted activity [46]. Two mechanisms can be attributed to cause this reduction in activity: first, is the excess HBA groups which suffer from the enthalpic desolvation penalty, since each HBA feature must shed its hydration shell before interacting with the protein [47]; second, is the high SIC which correlates with greater conformational flexibility and increased rotatable bond count, and imposes an entropic binding penalty within the sterically constrained ALK binding pocket [48]. Moreover, the quadratic SIC^2 term means that the penalty amplifies

non-linearly as the complexity of the compound's structure increases, which provides a quantitative rationale that polar, highly substituted kinase inhibitors often show reduced cellular potency [49].

Steric and Electrostatic Constrains at the Solvent Front (-0.195)

The term (ES-Count-aaN × K-T6-8 × Shadow-XZfrac) (-0.195) encodes for the steric and electrostatic penalty at the solvent-front area within the ALK binding pocket. Thus, the large lateral XZ-plane projection combined with the aromatic nitrogen polarity and the K-T6-8 pharmacophoric alignment creates an unfavorable interaction at the solvent-exposed entrance area of the ATP binding cleft. Structurally, this directly correlates with the G1202R resistance mutation (please refer to Table 1). Thus, bulky, polar groups projecting into this region are particularly vulnerable to steric clash and electrostatic repulsion introduced by the Arg substitution [1] and therefore limiting the XZ lateral projection is considered as a rational drug design strategy for next generation ALK inhibitors which are targeting the G1202R variants.

Remaining Terms: Desolvation, Electronic Reactivity, and Fine-Tuning Contributions

Moderate desolvation penalty is being imposed by the term (ES-Count-aasC × ES-Sum-aaaC × HBD Count) (-0.089). Thus, an aromatic system carrying polar substitutions with multiple hydrogen bond donors can present a large polar surface area that must be desolvated upon entering the hydrophobic back pocket, which reinforces the need for the lipophilic/polar balance within the structure of the molecule [50]. Also, the term (Band-Gap-Energy-DMol3 × D-T8-9 × Jurs PPSA-3) (-0.043) reflects an electronic rigidity constraint. Therefore, a large HOMO-LUMO energy gap reduces the polarizability and the charge transfer capacity at the DFG motif (Asp1270-Phe1271-Gly1272) which explains the preference for electron-rich heteroaromatic scaffolds across approved ALK inhibitors [51]. Moreover, the positive term (B-T5-9 × ES-Count-aaN × ES-Sum-aaaC) (+0.040) reflects the stabilization of aromatic nitrogen scaffolds through the π-π stacking and the CH-π interaction with Phe1271. Also, the terms (HBA Count × J-T5-1 × K-T6-8) (+0.014) and (A-T4-7 × CHI-V-3-C × I-T6-4) (+0.010) reflect the peripheral bidentate stabilization and the topological shape adaptability respectively, analogous to the dual-contact binding strategy of lorlatinib [1]. Finally, the term (CHI-V-2 × ES-Count-aasC × ES-Count-dO) (-0.007) imposes a minor penalty for excess carbonyl/ether oxygen functionalization but its small magnitude limits the real practical drug design impact.

3.2.3 Hierarchical Drivers of ALK Inhibitory Activity

The integration of all weighted contributions in Equation 2 reveals a three-tier hierarchy. Tier 1 which is the dominant tier and is represented by the aromatic electrostatic/geometric synergy (+16.51 term) which encodes for back pocket π-electron density, surface charge optimization, and YZ-plane geometric fit. Then, tier 2 which is the defining tier that represents the hinge-region pharmacophoric and hydrogen bonding requirements (+0.657 term) balanced against structural complexity and acceptor excess (-0.531 term) and the solvent-front steric constraints (-0.195 term), collectively defining the boundaries of the ALK active chemical space. Tier 3 which is the fine-tuning tier and is represented by the remaining terms (coefficients

≤ ±0.089). Tier 3 modulates the activity through desolvation balance, electronic reactivity, pharmacophoric alignment, and topological shape adaptability.

Finally, this hierarchical network (see Table 3) is fully consistent with the binding pharmacology of the previously approved ALK inhibitors (crizotinib, ceritinib, alectinib, brigatinib, and lorlatinib) each of which satisfies Tier 1 aromatic requirements and Tier 2 hinge region bonding requirements while maintaining moderate complexity consistent with the -0.531 penalty.

3.3 Branch 2: Machine Learning Model Performance

Machine Learning Classification Performance

In our branch 2, four algorithms (Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Naïve Bayes (NB), and k-Nearest Neighbors (k-NN)) were evaluated by applying the stratified ten-fold cross-validation parameters. So, all four models delivered a classification performance which assures that the selected pharmacophoric and physicochemical descriptor space is encoding real structural information that can predict the inhibitory activity against the of the ALK enzyme. However, the classifiers' performance varied, with the RF classifier showing classification metrics which exceeded all other algorithms (the full performance statistics are summarized in Table

Table 4. Classification performance of the four machine-learning algorithms across the three ALK activity classes (accuracy, Cohen's κ, and per-class metrics).

Metric	Random Forest (RF)	k-NN	XGBoost	Naïve Bayes
Dataset				
Training compounds (n)	230	230	230	230
Overall performance				
Overall accuracy	**0.987**	0.865	0.817	0.796
Incorrectly classified	**3**	31	42	47
Cohen's κ	**0.98**	0.793	0.721	0.65-0.75
Class-wise correct classifications				
True positives — active (n = 89[a])	**89**	82	76	90
True negatives — inactive (n = 55[a])	**55**	54	54	55
Correct — moderately active (n = 83[a])	**83**	63	58	38
Per-class performance metrics (range across activity classes)				
Precision	**> 0.98**	0.80-0.98	0.76-0.93	0.61-0.95
Recall (sensitivity)	**> 0.98**	0.74-0.98	0.68-0.98	0.60-0.90
F1-score	**0.982-0.994**	0.80-0.98	0.74-0.96	0.618-0.948
Specificity	**0.989-1.00**	0.850-0.994	0.829-0.977	0.707-1.00

[a] **n = number of compounds in each activity class, as reported for the Random Forest model.**

3.3.1 Random Forest Classification Results

Subsequently, the RF model achieved an overall accuracy of 98.7% ($\kappa = 0.98$) and only three compounds were misclassified from the total 230 compounds within the training set. Also, the class-specific analysis confirmed that RF correctly classified all 89 active compounds, all 55 inactive, and all 83 moderately active compounds, with zero active compounds falling below their correct classification threshold. Moreover, precision, recall, F1-score, and specificity metrics all exceeded 0.98, in addition the Cohen's κ of 0.98 places the RF model firmly in the perfect agreement category. Finally, the low variance of κ across cross-validation folds confirms its stability and independency from a particularly favorable data split (please refer to the Per-classifier hyperparameters and stratified CV performance breakdowns in *Supplementary Section SM5 (Table S5b)*.

3.3.2 Comparative Performance of Competing Classifiers

The k-NN model achieved an accuracy value of 86.5% ($\kappa = 0.793$), with the majority of the 31 misclassified compounds were occurring at the active/moderately active boundary which is a pattern consistent with the limited capacity of a Euclidean distance metric in high-dimensional descriptor space. Whereas, XGBoost showed a performance metrics of 81.7% accuracy ($\kappa = 0.721$), with a misclassification incidence concentrated between active and moderately active classes. Therefore, the linear boosting objective function and fixed 100-round training cycle may be insufficient in this case to capture the nonlinear descriptor-activity surface without additional hyperparameter tuning.

Finally, the NB model delivered the weakest performance metrics (79.6% accuracy; $\kappa = 0.65$-0.75), with misclassification cases concentrated in the moderately active class (38 out of 83 compounds correctly classified, compared to 83/83 for RF). Therefore, the NB model functions as a formal lower-bound benchmark rather than a competitive classifier in this study.

3.4 The Integration of the GFA-MLR and the Machine Learning Branches

In the context of this study, the parallel application of the GFA-MLR regression model and the RF classification represents a methodological safeguard that neither approach can alone provide. And so, the GFA-MLR delivers a SAR equation that can be directly interpreted through each of its terms by quantifying the directional contribution of the cubic descriptor interaction to the predicted pIC_{50}. Meanwhile, the RF model delivers a predictive performance by modeling the nonlinear, high-dimensional descriptor-activity surface without imposing a parametric functional form.

Later shown, the convergence of both branches on the same pharmacophoric and physicochemical determinants was achieved through an entirely independent selection mechanisms, and this constitutes a strong validation for the biological relevance of both branches. For example, both branches independently identified mutual pharmacophores within the ALK ATP binding pocket (see table 3, Consensus pharmacophores). Also, both branches independently identified mutual electrostatic complementarity with

the hinge-region (Met1199/Glu1197) and solvent-front residues (see Table 3 and Table 4). Moreover, both branches encoded for molecular topology as central determinants of the ALK inhibitory activity.

Subsequently, one of the clearest examples of this cross-branch convergence is the recurrence of the pharmacophore A-T1-1 as the most frequently selected pharmacophore in the RF feature frequency analysis (Section 3.5) in addition to its presence in the dominant hinge-bonding term of Equation 2 (+0.657; Section 3.2.2). Finally, the agreement between the two methodologically independent modeling streams substantially strengthens the confidence in their identified structural determinants and validates their application as selection criteria for pharmacophore-guided hit prioritization step (Section 3.7.1).

3.5 Descriptor Selection Frequency within the Successful RF Model

The frequency analysis method was used to evaluate all the 4,366 successful RF model iterations that resulted from the RF model (Table 5). The overall RF search space yielded a mean Cohen's κ of 0.9847 (SD = 0.0039) across all the evaluated combinations, with a minimum κ value of 0.9602 and a maximum of 0.9933. Therefore, this low inter-iteration variance confirms that the 230-compound dataset contains a highly predictive information which is distributed broadly across the pharmacophoric descriptor space (please refer to Table 5 that shows the recurrent pharmacophore descriptors and their key co-occurring molecular features across the 4,366 evaluated RF model iterations) (The workflow architecture details, and feature frequency aggregation methodology are documented in *Supplementary Section SM5*).

Table 5. Recurrent pharmacophore descriptors and their key co-occurring molecular features across the 4,366 evaluated Random Forest model iterations.

Rank	Pharmacophore descriptor	Frequency (n)	% of models	Key co-occurring descriptors	Interaction type	Interpretation
1	**C-T3-9**[a]	143	3.28%	ES_Sum_ssS, Num_AromaticBonds	*Electronic + Aromatic*	• C-T3-9 represents the π-π stacking and electron-rich aromatic interactions at the ALK binding pocket. • The consistent co-occurrence of C-T3-9 with the aromatic bond count and S-containing E-state descriptors confirms the back-pocket aromatic anchoring as the primary activity driver.
2	**F-T1-8**[b]	116	2.66%	V_ADJ_mag, Jurs_FNSA_3	*Topological + Surface Charge*	• Hydrophobic alignment combined with molecular topology and fractional negative surface area. • The consistent co-occurrence of F-T1-8 with V_ADJ_mag (adjacency-weighted valence) and Jurs_FNSA_3 (fractional negative surface area) which reflects the importance of surface charge distribution alongside hydrophobic contacts.
3	**A-T1-1**[c]	~110	~2.5%	Molecular_PolarSASA, CHI_1	*Polarity + Topological*	• Consensus pharmacophore in both GFA-MLR and ML branches. • A-T1-1 captures the hinge-region binding geometry across diverse scaffolds.
4	**B-T4-3**[c]	~90	~2.1%	NPlusO_Count, Molecular_PolarSASA	*Hydrogen Bonding + Polarity*	• Hinge-region H-bond donor and acceptor interactions. • Co-occurrence with heteroatom count (NPlusO_Count) and polar surface area links B-T4-3 to polar scaffolds capable of forming multiple contacts in the hinge region.
5	**I-T4-9**[d]	~70	~1.6%	V_DIST_equ, ES_Count	*Spatial + Electronic*	• Proposed to reflect deeper pocket interactions which are driven by geometry and electronic factors. • The Co-occurrence of pharmacophore I-T4-9 with the descriptor equalized distance sum (V_DIST_equ) and electrostatic state count suggests spatial reach into the hydrophobic back-pocket combined with electronic complementarity.
6	**J-T5-1**[c]	~55	~1.3%	CHI indices	*Topological + Connectivity*	• Peripheral binding stabilization through molecular connectivity and co-occurrence. • J-T5-1 is a consensus pharmacophore that appeared in both branches.

[a] C-T3-9: ML branch only (Trial 3, Set C, Hypothesis 9).

[b] F-T1-8: ML branch only (Trial 1, Set F, Hypothesis 8).

[c] A-T1-1, and J-T5-1 appear in both the GFA-MLR branch (Equation 2 descriptor terms) and the ML branch (RF feature selection), making them consensus pharmacophores.

[d] I-T4-9: ML branch only (Trial 4, Set I, Hypothesis 9).

[e] Co-occurring descriptor abbreviations: ES_Sum_ssS, electrostatic state sum for S-bonded-to-two-atoms; Num_AromaticBonds, number of aromatic bonds; V_ADJ_mag, magnitude of adjacency-weighted valence; Jurs_FNSA_3, fractional negative surface area (charge-weighted); Molecular_PolarSASA, polar solvent-accessible surface area; CHI_1, first-order connectivity index; NPlusO_Count, total nitrogen + oxygen atom count; V_DIST_equ, equalized distance sum; ES_Count, electrostatic state count;); CHI indices, valence and simple connectivity indices of various orders

3.5.1 High-Frequency Pharmacophoric Descriptors

Notably, pharmacophore C-T3-9 was the most frequently selected descriptor across all 4,366 RF iterations (n = 143; 3.28%), also it was acknowledged that pharmacophore C-T3-9 co-occurred with the aromatic bond count descriptor (Num_AromaticBonds) and the sulfur E-state sum (ES_Sum_ssS). Thus, this pattern of co-occurrence is emphasizing the fact that the pharmacophore C-T3-9 encodes for the hydrogen bond acceptor / hydrophobic interaction geometry that characterize the π-π stacking and the electron-rich aromatic contacts within the ALK back pocket.

Moreover, pharmacophore F-T1-8 was selected in 116 RF iterations (2.66%), primarily co-occurring with the descriptor V-ADJ-mag (adjacency-weighted valence connectivity) and the descriptor Jurs-FNSA-3 (fractional negative surface area). Thus, this pattern reflects that the hydrophobic alignment combined with the charge distribution topology at the back-pocket/αC-helix interface, which is consistent with the electrostatic complementarity requirement which is also encoded in the dominant +16.51 term in Equation 2.

Also, pharmacophore A-T1-1 appeared in approximately 110 iterations (~2.5%), co-occurring with the descriptor Molecular_PolarSASA (polar solvent accessible surface area) and the descriptor CHI_1 (first order molecular connectivity). Remarkably, A-T1-1 is selected independently in both the GFA-MLR branch (Equation 2, +0.657 term) and the ML branch (RF frequency analysis.

Moreover, pharmacophore B-T4-3 was selected in 90 iterations (~2.1%) and was co-occurring with NPlusO-Count and Molecular-PolarSASA, and this pattern is consistent with the role of B-T4-3 in capturing the hinge region hydrogen bond donor/acceptor interactions.

3.5.2 Consensus Pharmacophores Across Both Modeling Branches

As shown in Tables 3 and 5, out of the six pharmacophoric descriptors with selection frequency above the 1% threshold (Table 5), only two pharmacophores (A-T1-1 and J-T5-1) were selected by both the GFA-MLR and ML branches, thereby this qualifies them to be considered as consensus pharmacophores. Accordingly, pharmacophore A-T1-1 in Table 2 (which is characterized by its hinge region geometry that is consistent with Met1199/Glu1197 interactions) was assigned the priority tier 1 for virtual screening deployment. Also, pharmacophore J-T5-1 in Table 2 (which is characterized by it's solvent-front and back-pocket boundary geometry) appeared in the ML branch at ~1.3% frequency (see Table 5) and was found to contribute to the peripheral bidentate stabilization term (+0.014) in Equation 2 (Table 3).

3.6 ROC Validation of Pharmacophore Hypotheses

All nine pharmacophores in Table 2 were subjected to ROC analysis against a ZINC 3D validation library of 38,756 drug-like decoys and 98 confirmed ALK actives [35,36,52]. The validated ROC parameters that are

employed in this study are summarized in Table 6 and ROC curve analysis results are shown in Figures 1A, 1B, and 2. Please refer to *Supplementary Section SM6 (Table S6)* for the validation library composition and the tier stratification thresholds.

3.6.1 Tier 1 Pharmacophores (AUC ≥ 0.80)

As shown in Figure 1A, five hypotheses satisfied the Tier 1 threshold of AUC ≥ 0.80 and were approved as positive virtual screening query filters. For example, pharmacophore F-T1-8 achieved the highest discriminatory performance of AUC = 0.905 and TPR = 97.0%, which confirms a broad capture of the hydrophobic alignment geometry across the previously confirmed ALK inhibitors. Also, pharmacophore B-T5-9 (appeared in Equation 2, +0.040 term) delivered the best balance of accuracy and specificity (ACC = 91.3% and SPC = 91.8%) alongside its strong early enrichment ($EF_1\%$ = 22.70), which is consistent with its encoding of π-π stacking against the aminoacid Phe1271 of the DFG motif. Similarly, pharmacophore A-T1-1 (consensus pharmacophore) and pharmacophore A-T4-7 (Equation 2) both achieved an AUC value of 0.861, with A-T1-1 exhibiting the highest specificity among Tier 1 hypotheses (SPC = 93.9%) and A-T4-7 exhibiting the highest accuracy (ACC = 88.1%); the performance of A-T1-1 is consistent with its function as the primary hinge region pharmacophoric template that encodes for the hydrogen bonding interactions against Met1199/Glu1197 and the electrostatic complementarity with the aminoacids Lys1150/Arg1253 [8]. Additionally, pharmacophore I-T4-9 achieved an AUC value of 0.806 with an enrichment value of ($EF_1\%$ = 26.78).

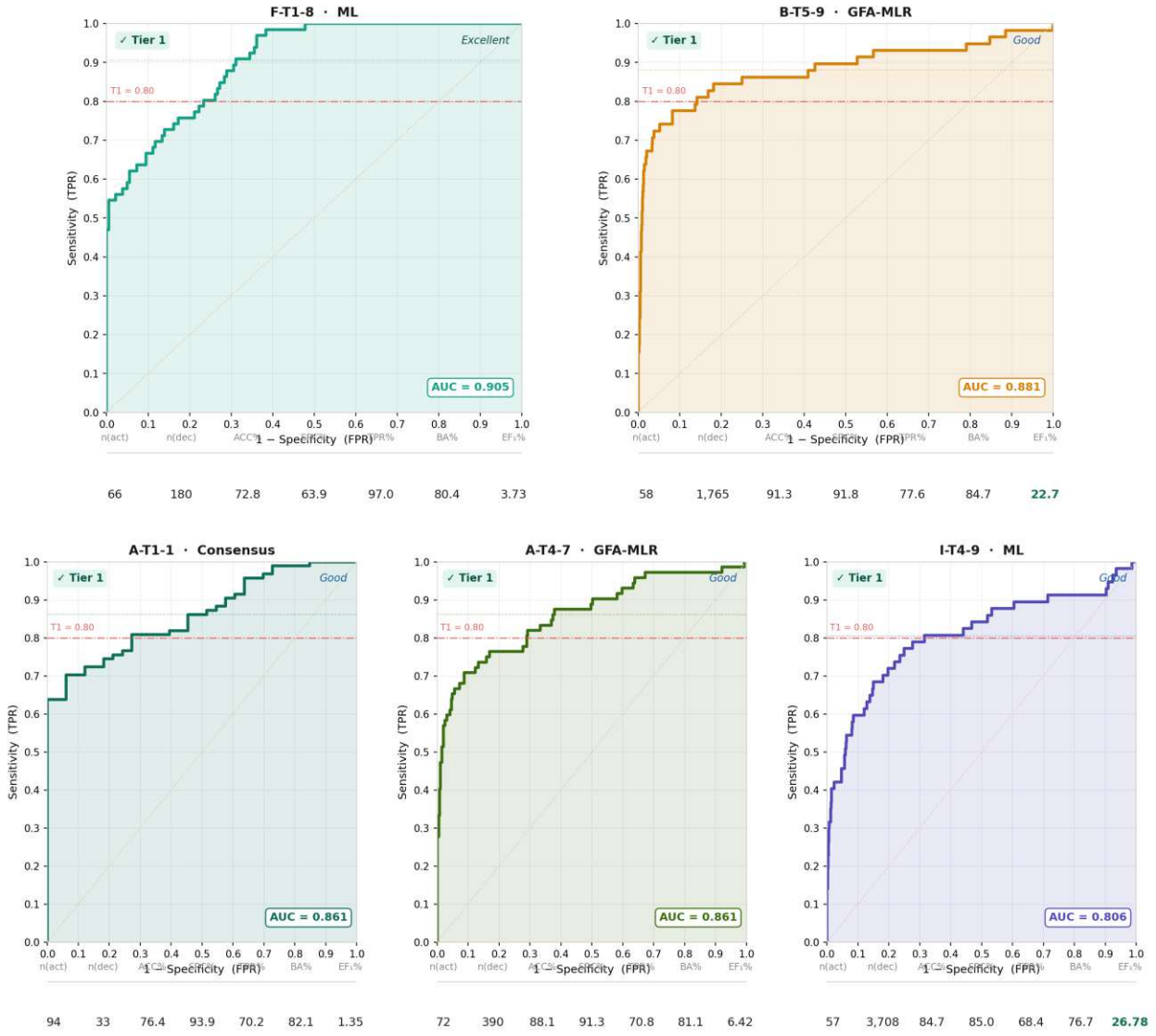


Figure 1A. ROC curves for the five Tier 1 pharmacophore hypotheses (AUC ≥ 0.80) validated against the ZINC 3D library (38,756 decoys + 98 ALK actives). Each panel displays the ROC curve, the random diagonal (gray dotted), and the Tier 1 deployment threshold at AUC = 0.80 (red dash-dot line). The AUC value is annotated at lower right.

3.6.2 Tier 2 Pharmacophores (AUC ≥ 0.70)

As shown in Figure 1B and Table 6, three hypotheses met the Tier 2 threshold and were retained as complementary pharmacophoric profiles. For example, pharmacophore B-T4-3 (AUC = 0.756) achieved the highest sensitivity of all nine hypotheses (TPR = 94.2%). Also, pharmacophore C-T3-9 (AUC = 0.754) mapped the largest active count in this panel (76 actives; $EF_1\%$ = 22.47), which is consistent with its role as the pharmacophore with the highest frequency RF descriptor (n = 143 model iterations, 3.28%). Similarly, pharmacophore J-T5-1 (consensus, AUC = 0.723) exhibited the highest $EF_1\%$ of all nine pharmacophores (34.74) and the highest overall specificity (SPC = 94.9%), which reflects the geometric constraints of its five exclusion spheres at the solvent-front and back-pocket boundary. The divergence

between J-T5-1's moderate global AUC and exceptional $EF_1\%$ reflects strong concentration of true actives at the top of the ranked list which is considered as an important property in virtual screening where only the top ranked compounds can advance to experimental evaluation.

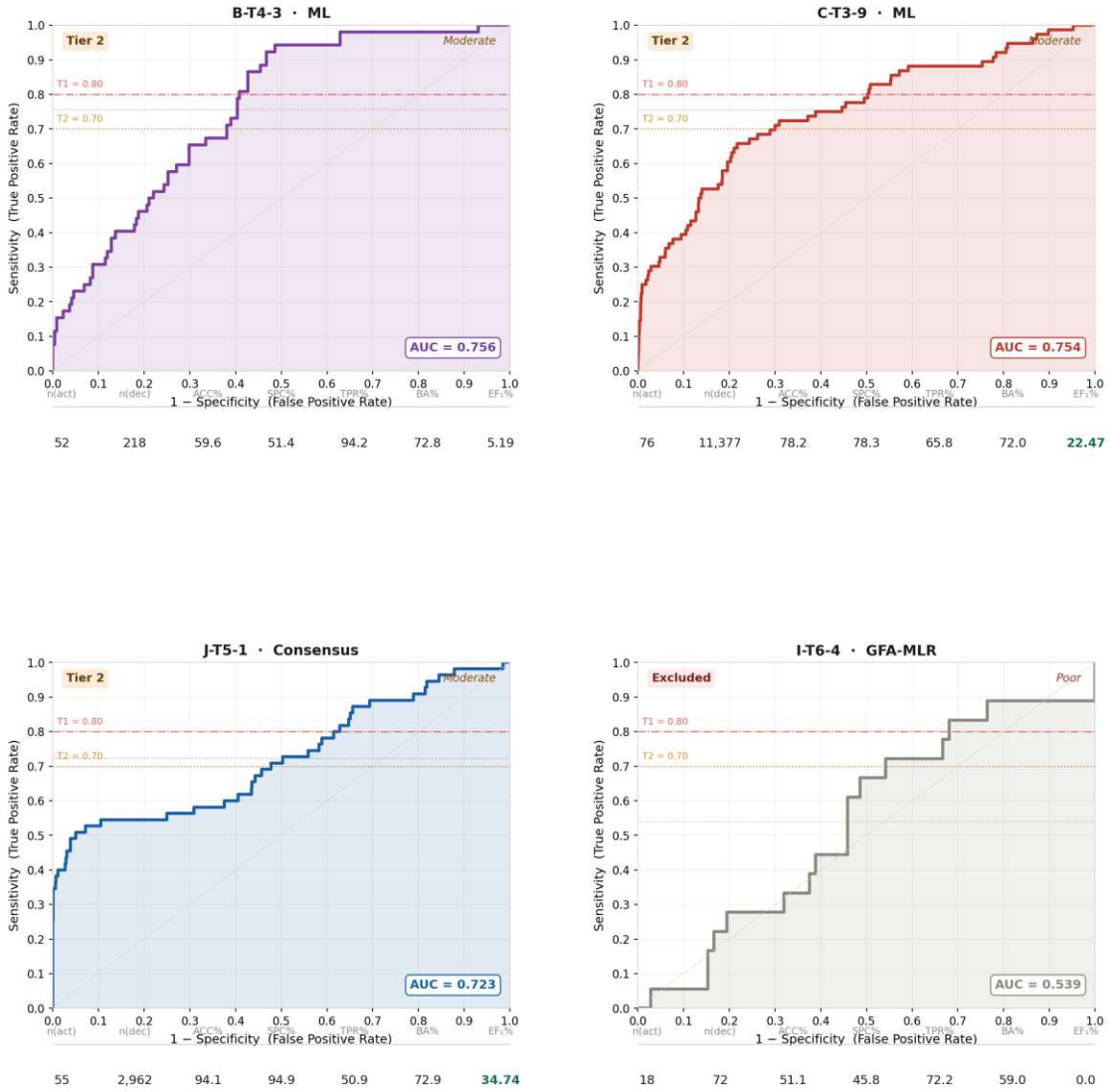


Figure 1B. ROC curves for the three Tier 2 ALK pharmacophore hypotheses (AUC ≥ 0.70) and the excluded hypothesis I-T6-4 (AUC < 0.70), validated against the ZINC 3D library (38,756 decoys + 98 ALK actives). The Tier 2 threshold (orange dotted, AUC = 0.70) is displayed for reference.

3.6.3 Excluded and Tier 3 Pharmacophores

Also, Figure 1B and Table 6 show the ROC analysis results of pharmacophore I-T6-4 which yielded an AUC of 0.539 which indicates random performance, respectively; only 18 actives were mapped from the ZINC library. Consequently, pharmacophore I-T6-4 was excluded from virtual screening deployment. Likewise, pharmacophores D-T8-9 and K-T6-8 were designated as Tier 3 exclusion filters based on their

negative GFA-MLR coefficients (-0.043 and -0.195), Accordingly, they were not subjected to the ROC validation process.

Table 6. ROC validation statistics.

Pharmacophore ID	Modelling Branch	Tier	ROC-AUC[a]	ROC Evaluation	ACC%[b]	SPC%[c]	TPR%[d]	BA%[e]	MCC[f]	EF¹%[g]
F-T1-8 ★ **Highest AUC**	ML only (RF-GA)	**Tier 1**	**0.905**	Excellent	72.8	63.9	**97.0**	80.4	0.540	3.73
B-T5-9 ★ **Best ACC, SPC, EF₁%**	GFA-MLR	**Tier 1**	**0.881**	Good	**91.3**	91.8	77.6	84.7	0.399	**22.70**
A-T1-1 ★ **CONSENSUS (both branches)**	CONSENSUS	**Tier 1**	**0.861**	Good	76.4	**93.9**	70.2	82.1	**0.564**	1.35
A-T4-7 ★ **Best ACC (Tier 1)**	GFA-MLR	**Tier 1**	**0.861**	Good	**88.1**	91.3	70.8	81.1	**0.582**	6.42
I-T4-9 ★ **Best EF₁%**	ML only (RF-GA)	**Tier 1**	**0.806**	Good	84.7	85.0	68.4	76.7	0.179	**26.78**
B-T4-3 ★ **Highest TPR**	ML only (RF-GA)	**Tier 2**	**0.756**	Moderate	59.6	51.4	**94.2**	72.8	0.364	5.19
C-T3-9 ★ **Highest RF freq.**	ML only (RF-GA)	**Tier 2**	**0.754**	Moderate	78.2	78.3	65.8	72.0	0.086	**22.47**
J-T5-1 ★ **CONSENSUS · Highest EF₁%**	CONSENSUS	**Tier 2**	**0.723**	Moderate	**94.1**	**94.9**	50.9	72.9	0.260	**34.74**
I-T6-4 **AUC ≈ 0.50; excluded**	GFA-MLR	**Excl.**	**0.539**	Poor (random)	51.1	45.8	72.2	59.0	0.146	0.00

[a] ROC-AUC: area under the receiver operating characteristic curve. AUC ≥ 0.80 → Tier 1; ≥ 0.70 → Tier 2; < 0.70 → Excluded.
[b] ACC: overall accuracy = (TP + TN)/(P + N).
[c] SPC: specificity = TN/(TN + FP).
[d] TPR: sensitivity = TP/(TP + FN). Bold values ≥ 90% denote high recall.
[e] BA: balanced accuracy = (TPR + SPC)/2.
[f] MCC: Matthews correlation coefficient. Bold values ≥ 0.55 denote strong correlation.
[g] EF₁%: enrichment factor at the top 1% of the ranked list. Bold values ≥ 20 denote strong early enrichment and the operationally critical metric for virtual screening deployment.

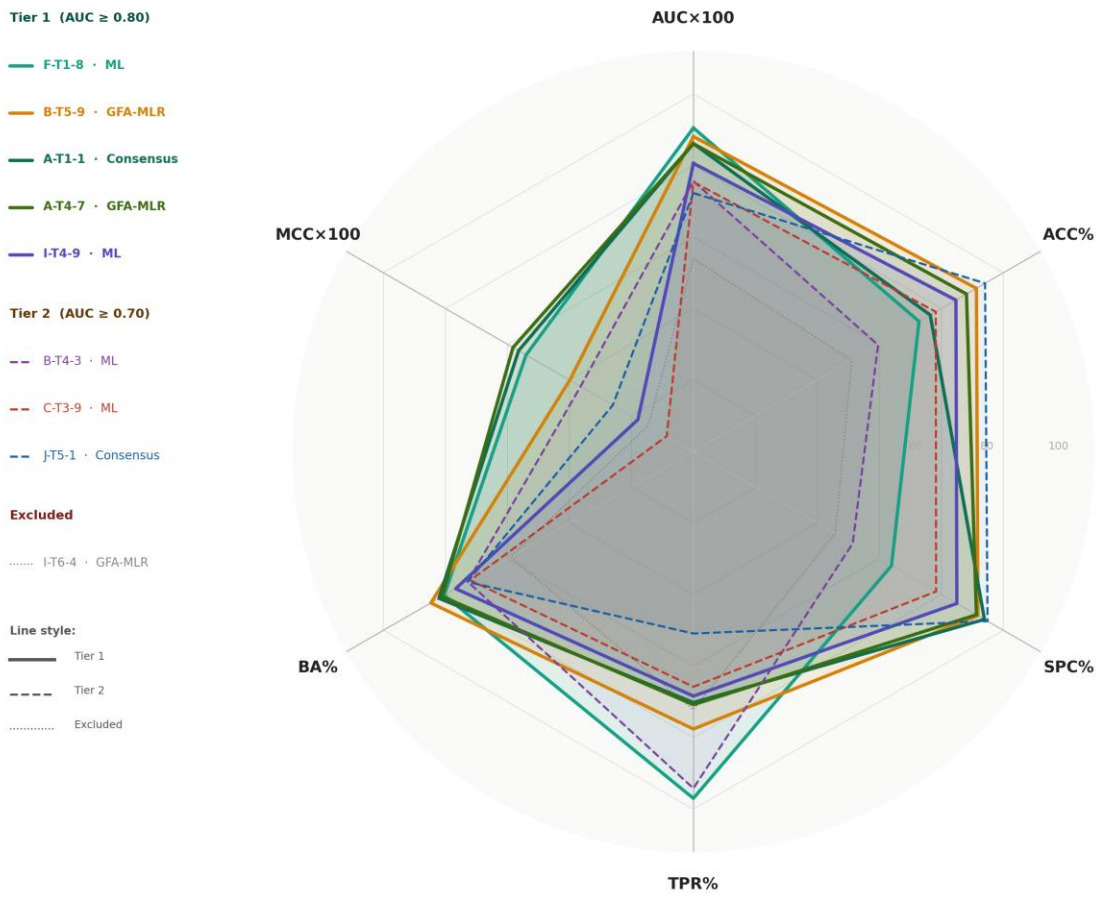


Figure 2. Full performance profile of all nine pharmacophore hypotheses across six validation metrics scaled to 0–100: AUC×100, overall accuracy (ACC%), specificity (SPC%), sensitivity (TPR%), balanced accuracy (BA%), and Matthews correlation coefficient (MCC×100). Solid lines = Tier 1 (AUC ≥ 0.80); dashed lines = Tier 2 (AUC ≥ 0.70); dotted line = Excluded.

3.7 In-silico screening for Active Hits

The five Tier 1 pharmacophore hypotheses that were validated by ROC analysis (F-T1-8, B-T5-9, A-T1-1, A-T4-7, and I-T4-9) were deployed as 3D query filters for flexible pharmacophore database searching of the NCI compound library (Section 2.6). The application of these five queries yielded 28 candidate compounds exhibiting good pharmacophore fit scores and structural compatibility with the ALK binding requirements. Consequently, these hit candidates were ranked by their predicted pIC_{50} by using Equation 2 and classified by the validated RF model before being experimental evaluated.

Most importantly, several of these pharmacophore derived hits demonstrated significant inhibitory activity above 50% inhibition at 10 μM (Table 7) which validates the predictive capability of the pharmacophore guided screening in this study and confirms the robustness of the integrated QSAR and machine learning framework (please refer to the *Supplementary Section SM7 (Table S7)* for the summarization table of the stage-wise pharmacophore screening yield, ranking metrics, and hit prioritization funnel*)*.

3.8 In Vitro ALK Kinase Inhibition Screening

3.8.1 Single Point Screening at 10 µM

The HotSpot™ radiometric ALK enzyme assay (Reaction Biology Corp., USA) was used to experimentally to evaluate a panel of 28 compounds that were previously identified by our pharmacophore-based virtual screening process. The ALK kinase enzyme was applied in the assay at fixed screening concentration of 10 µM under ATP Km conditions of [ATP] = 30 µM. Also, all hit compounds were prepared as 10 mM DMSO stock solutions and tested in duplicate; then the enzymatic activity was expressed as percentage activity relative to the DMSO vehicle control (100%). Consequently, the screening results revealed a diverse range of inhibitory responses. Table 7 presents the eight compounds that demonstrated measurable inhibition of ALK activity at 10 µM; meanwhile the remaining 20 compounds did not meet the below 50% inhibition threshold. Subsequently, three hit compounds demonstrated mean ALK enzyme activity below 50% at 10 µM and were advanced to full dose-response characterization, namely, Compound 2 (NSC 196403) which showed the lowest mean enzyme activity of 34.15%, followed by Compound 3 (NSC 715055) which showed the mean enzyme activity of 45.50%, and Compound 1 (NSC 290954) which showed the mean enzyme activity of 48.48%. Additionally, five compounds showed partial inhibition (mean % enzyme activity 61-90%) representing promising scaffolds for future structural optimization.

3.8.2 Dose Response Characterization and IC_{50} Determination

Three hit compounds that were able to demonstrate a mean ALK enzyme activity below 50% at 10 µM (see Table 7) were advanced to a full dose response characterization (Table 10); their mapping onto the five Tier 1 pharmacophore hypotheses is shown in Figure 3. Notably, Compound 2 (NSC 196403) showed the lowest mean enzyme activity of 34.15% and an IC_{50} = 4.02 µM, followed by Compound 3 (NSC 715055) showed the mean enzyme activity of 45.50% and an IC_{50} = 5.95 µM and Compound 1 (NSC 290954) showed the mean enzyme activity of 48.48% and an IC_{50} = 2.15 µM. In addition, replicate values were consistent for all three compounds and the associated hill slopes ranged from -0.45 to 0.72, which is consistent with the non-promiscuous inhibition kinetics. Furthermore, five additional compounds showed partial inhibition in the 10-39% range, which represents some additional scaffolds of interest for future optimization. Also, the staurosporine reference produced an IC_{50} value of 1.46 nM, which confirms the enzymatic assay validity. Notably, the identification of 8 ALK inhibitory compounds from a virtual screening hit set of 28 (8/28) and 3 of which met the below 50% anti-ALK enzymatic activity threshold for dose response characterization (3/28) underscores the effectiveness of the multi-layer computational strategy that was employed in this study.

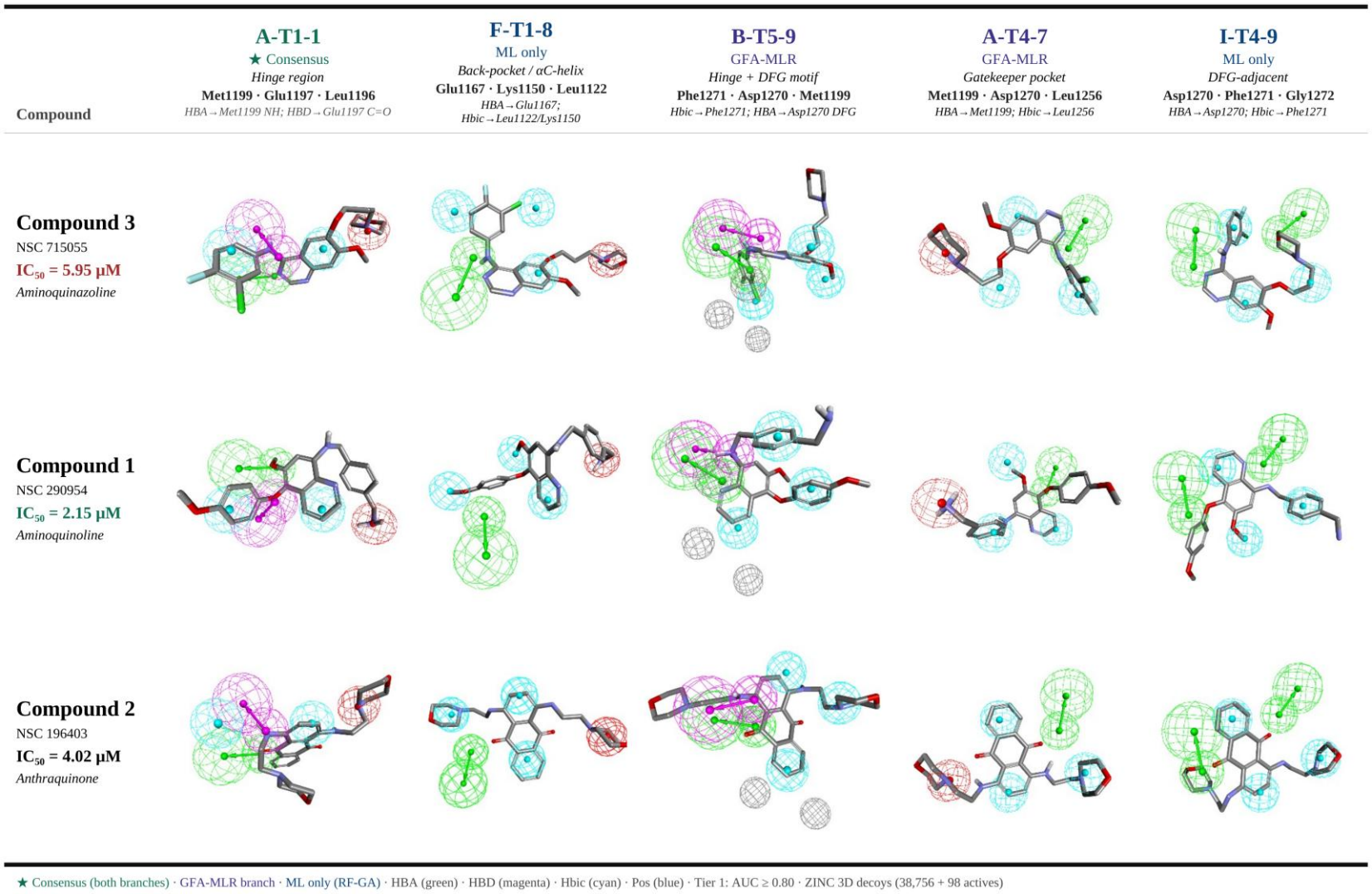


Figure 3. Pharmacophore mapping of active ALK inhibitor hits (Compound 3, Compound 1, and Compound 2) against the five Tier 1 hypotheses (A-T1-1, F-T1-8, B-T5-9, A-T4-7, and I-T4-9) generated in BIOVIA Discovery Studio 4.5.

Notes. Feature colors: HBA (green), HBD (magenta), Hbic (cyan), Pos (blue). ★ A-T1-1 is the consensus pharmacophore selected in both GFA-MLR (Eq. 2; +0.657; Met1199/Glu1197) and ML (RF-GA; ~110/4,366 models) branches. Residue interactions: A-T1-1 — Met1199, Glu1197, Leu1196 (hinge H-bonds); F-T1-8 — Glu1167, Lys1150, Leu1122 (back-pocket); B-T5-9 — Phe1271, Asp1270, Met1199 (hinge + DFG); A-T4-7 — Met1199, Asp1270, Leu1256 (gatekeeper); I-T4-9 — Asp1270, Phe1271, Gly1272 (DFG-adjacent). IC_{50} by HotSpot™ radiometric assay, [ATP] = 30 µM; staurosporine IC_{50} = 1.46 nM.

Table 7. ALK kinase inhibition single-dose screening results.

Compound	NCI Code	% Enzyme Activity Replicate 1[a]	% Enzyme Activity Replicate 2[a]	Mean % Enzyme Activity	Status
Compound 3[ı]	NSC 715055	47.07	43.92	**45.50**	Active and advanced to IC_{50} determination[b]
Compound 4	NSC 645804	64.00	57.90	60.95	Did not meet active threshold
Compound 5	NSC 379965	87.62	85.64	86.63	Did not meet active threshold
Compound 6	NSC 32482	79.38	78.51	78.94	Did not meet active threshold
Compound 1	NSC 290954	47.42	49.54	**48.48**	Active and advanced to IC_{50} determination[b]
Compound 7	NSC 23767	86.93	93.07	90.00	Did not meet active threshold
Compound 2	NSC 196403	35.09	33.20	**34.15**	Active and advanced to IC_{50} determination[b]
Compound 8	NSC 131706	80.42	94.99	87.70	Did not meet active threshold

[a] Enzyme activity is expressed as percentage relative to the DMSO vehicle control (mean of duplicate determinations).

[b] Compounds with mean % enzyme activity ≤50% were defined as active and advanced to 10-point dose–response IC_{50} determination using a 3-fold serial dilution format starting at 100 µM.

[ı] Compound 3 = NSC 715055 = Gefitinib (an approved EGFR inhibitor); ALK inhibitory activity observed here reflects documented ALK/EGFR cross-reactivity and is consistent with the known polypharmacological profile of 4-anilinoquinazoline-class kinase inhibitors.

Note: Staurosporine (reference inhibitor, IC_{50} = 1.46 nM) confirmed assay validity.

Assay: HotSpot™ radiometric kinase assay, Reaction Biology Corp., USA ([ATP] = 30 µM; all reactions performed in duplicate).

3.8.3 NCI-60 One-Dose Anticancer Activity Profiling of NSC 290954 (Compound 1)

The growth inhibitory profile of compound NSC 290954 (Compound 1) across the NCI-60 human tumor cell line panel was retrieved from the NCI Developmental Therapeutics Program public database (https://dtp.cancer.gov) and analyzed to complement the biochemical ALK inhibition data. Accordingly, the NCI-60 one-dose screen measures the growth inhibition at a single 10 µM concentration across 60 cell lines that are representing nine tumor types, and which provides a cell-based selectivity fingerprint that is independent of the biochemical assay. Complete NCI-60 cell-line growth-inhibition data and Compound 3/gefitinib repurposing context are provided in *Supplementary Section SM8*.

The analysis of NSC 290954 across the NCI-60 panel revealed heterogeneous growth inhibition (mean growth = 92.7% of control at 10 µM, range 52.2-119.5% across 57 reporting lines), consistent with a compound exhibiting selective rather than pan-cytotoxic activity. The non-small cell lung cancer (NSCLC) sub-panel showed the lowest mean growth of any sub-panel (86.8%, n = 8 lines), indicating consistent panel-wide sensitivity; the single most sensitive cell line in the entire dataset was the renal line UO-31 (52.2% growth), a pronounced individual outlier, though the renal sub-panel as a whole showed only

intermediate mean sensitivity (91.2%), indicating this reflects line-specific rather than panel-wide renal sensitivity. This heterogeneous NCI-60 profile is consistent with Compound 1's low micromolar biochemical potency (IC_{50} = 2.15 μM) and supports a target-dependent cellular effect. The consistent NSCLC sub-panel sensitivity is notable given that EML4-ALK rearrangements are present in 3-7% of NSCLC patients [6].

3.8.4 Mechanistic Characterization of the Active NCI Hits by applying the QSAR Equation 2

Aiming towards identifying the structural and pharmacophoric features that are responsible for the ALK inhibitory activity in the successful hits (tables 8 and 10) these hits were subjected to a mechanistic characterization analysis by applying equation 2 on them.

Thus, each of equation 2 ten descriptor triplets which each encodes for a specific aspect of ligand-ALK binding was applied as a mechanistic framework, and the sign and magnitude of each term's contribution directly identifies the interaction types that each compound satisfies or violates relative to the ALK active site. Subsequently, all seven pharmacophore fit values required by Equation 2 (A-T1-1, K-T6-8, B-T5-9, D-T8-9, J-T5-1, I-T6-4, A-T4-7) were computed by using Discovery Studio 4.5 (see Table 8) then the experimental IC_{50} values for Compound 1 (NSC 290954), Compound 2 (NSC 196403), and Compound 3 (NSC 715055) were determined in the present study; values for NSC 313390 and NSC 75327 were sourced from El-Jundi et al. [50] (not assayed here).[53] Finally, the pharmacophoric alignment profiles and the experimental activities of the hit compounds are presented in Table 8 and visualized in Figure 3; descriptor contributions are detailed in Table 9.

Table 8. Pharmacophore alignment profiles of the five NCI hits (three assayed in the present study and two literature comparators, marked [a]) against the seven Equation 2 pharmacophore hypotheses, alongside experimental IC50 values

Compound	Pharmacophore Fit Values							IC_{50} (μM)	Ref.
	A-T1-1	**K-T6-8**	**B-T5-9**	**D-T8-9**	**J-T5-1**	**I-T6-4**	**A-T4-7**		
NSC 75327 [a]	0.000	0.000	3.387	0.000	2.010	0.000	0.000	**4.32**	[a]
NSC 196403 (Compound 2)	10.839	1.071	1.301	0.000†	2.946	2.073	6.347	**4.02**	[b]
NSC 290954 (Compound 1)	4.935	6.467	7.107	6.659	4.662	7.347	7.728	**2.15**	[b]
NSC 313390 [a]	1.142	5.333	7.624	2.968	5.074	3.437	6.867	**10.57**	[a]
NSC 715055 (Compound 3)	8.603	5.445	4.249	3.023	4.040	2.694	4.740	**5.95**	[b]

† D-T8-9 fit unavailable for NSC 196403 (Compound 2); set to zero.

[a] NSC 75327 and NSC 313390 were not assayed in the present study; IC50 values are taken from El-Jundi I, Daoud S, Taha MO. SAR QSAR Environ Res. 2024. DOI: 10.1080/1062936X.2024.2406398 [Z′-LYTE kinase assay].

b Present study, HotSpot™ radiometric kinase assay, Reaction Biology Corp., USA ([ATP] = 30 μM).

Table 9. GFA–MLR Equation 2 descriptor contribution analysis for the five NCI hits (three assayed in the present study and two literature comparators, marked [a]). Each term quantifies a specific structural interaction with the ALK binding pocket.

Term	Coeff.	Descriptor Triplet	NSC 75327 [a] IC^{50} = 4.32 µM	NSC 196403 IC^{50} = 4.02 µM	NSC 290954 IC^{50} = 2.15 µM	NSC 313390 [a] IC^{50} = 10.57 µM	NSC 715055 IC^{50} = 5.95 µM	Mechanistic Role
T1	**+16.51**	ES-Sum-aaaC × Jurs-RNCG × Shadow-YZfrac	+2.650	0.000	+2.402	0.000	+2.580	Represent aromatic e^- density and electrostatic complementarity
T2	**−0.531**	HBA-Count × SIC^2	−1.432	−0.851	−0.975	−0.379	−1.744	Represent polar surface and topological complexity balance
T3	**+0.657**	A-T1-1 × HBD-Count × Jurs-RNCG	0.000	+1.920	+0.952	+0.308	+0.864	Represent H-bond donor alignment at the hinge-region
T4	**−0.195**	ES-Count-aaN × K-T6-8 × XZfrac	0.000	0.000	−0.700	0.000	−1.258	Represent solvent-front geometry (G1202R constraint)
T5	**−0.089**	ES-Count-aasC × ES-Sum-aaaC × HBD	−2.211	0.000	−2.081	0.000	−0.776	Represent aromatic desolvation cost
T6	**+0.040**	B-T5-9 × ES-Count-aaN × ES-Sum-aaaC	0.000	0.000	+0.475	0.000	+0.494	Represent π-stacking at Phe1271 (DFG motif)
T7	**−0.043**	Band-Gap × D-T8-9 × Jurs-PPSA-3	0.000	0.000	−0.795	−0.229	−0.286	Represent electronic polarizability and charge-transfer
T8	**+0.014**	HBA-Count × J-T5-1 × K-T6-8	0.000	+0.177	+1.688	+0.379	+1.540	Represent bidentate back pocket and solvent-front stabilization
T9	**+0.010**	A-T4-7 × CHI-V-3-C × I-T6-4	0.000	+0.110	+0.387	+0.220	+0.098	Represent scaffold shape adaptability
T10	**−0.007**	CHI-V-2 × ES-Count-aasC × ES-Count-dO	−0.597	−0.746	0.000	−0.353	0.000	Represent oxygen functionalization balance
	Experimental IC_{50} (µM) =		**4.32**	**4.02**	**2.15**	**10.57**	**5.95**	

† D-T8-9 fit set to zero for NSC 196403. Model statistics: R^2 = 0.60, Q^2 = 0.58, r^2pred = 0.56, RMSEP = 0.89 log units (*n* = 57 test compounds).

[a] NSC 75327 and NSC 313390 were not assayed in the present study; descriptor contributions and IC50 values derive from compounds reported by El-Jundi et al. (2024) and are included for cross-study comparison only.

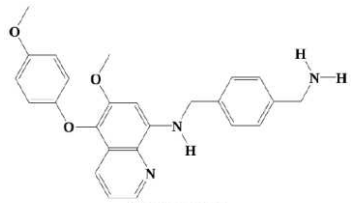
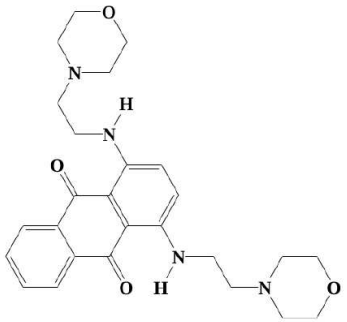
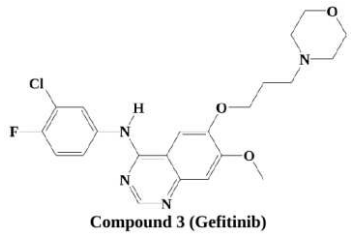
3.8.5 Structural Determinants of ALK Inhibitory Activity

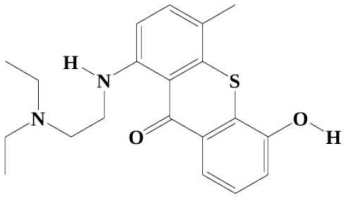
Consistent with these findings, the descriptor contribution profiles in Table 9 reveal a clear mechanistic basis for the observed inhibitory activities of the hits in this study. Most notably, compound 1 (NSC 290954, IC_{50} = 2.15 μM, Table 10), which is the most potent compound, has achieved the most favorable balance across all interaction tiers. In this regard, term T1 (+2.402) confirms that the fused aromatic electron density, surface charge concentration, and three-dimensional shape of compound 1 collectively satisfying the back-pocket occupancy requirements of Equation 2. Consistently, term T3 (+0.952) demonstrates a strong hinge-region pharmacophoric alignment (A-T1-1) combined with hydrogen bond donor capacity directed toward Met1199 and Glu1197. Building on this, the term T6 (+0.475) reflects π–π stacking interactions with Phe1271 against the DFG motif, while the term T8 (+1.688) with the largest T8 contribution across all five compounds , confirms simultaneous pharmacophoric alignment with both the solvent-front (J-T5-1) and back-pocket (K-T6-8) boundary regions, and this binding geometry is analogous to the bidentate strategy employed by the third-generation inhibitor lorlatinib. Notwithstanding these favorable contributions, the moderate T5 penalty (−2.081) reflects the desolvation cost of its functionalized aromatic system, which nevertheless provides a direct structural rationale for hit-to-lead optimization through selectively reducing the number of polar substituents on the aromatic scaffold while preserving the T3 and T8 pharmacophoric interactions which is therefore expected to significantly enhance potency.

Compound 2 (NSC 196403, IC50 = 4.02 μM, Table 10) is most probably deriving its activity from its strong hinge region pharmacophoric alignment (T3 = +1.920), which is the highest T3 contribution across all five compounds, consistent with its high A-T1-1 fit value of 10.839. Therefore, this indicates that compound 2 achieves excellent 3D complementarity with the hinge region pharmacophoric template, with hydrogen bond donors correctly positioned to engage with Met1199 and Glu1197. Also, the small T8 contribution (+0.177) suggests moderate peripheral stabilization, while the absence of T1 contribution (ES-Sum-aaaC = 0) identifies the back-pocket aromatic interaction as the principal optimization target for this scaffold. This can be done by introducing a fused aromatic extension to activate T1 while retaining the T3 hinge alignment.

Lastly, compound 3 (NSC 715055, IC_{50} = 5.95 μM, Table 10) combines positive contributions from T1 (+2.580), T3 (+0.864), T6 (+0.494), and T8 (+1.540), which indicates simultaneous engagement of the back-pocket, hinge region, DFG motif, and peripheral binding sites. However, the T4 penalty (-1.258) which is the largest across all five compounds, is identifying significant steric/electrostatic interactions at the solvent-front region (Gly1202). This is mechanistically important since the T4 penalty term is the quantitative descriptor correlate of G1202R mutation vulnerability and by reducing the lateral XZ-plane projection of Compound 3 analogues which is the structural feature that is encoded by this term represents a direct, data-driven strategy for improving both potency and resistance profile simultaneously.

Table 10. Chemical structures and dose-response parameters for the active ALK inhibitor hits identified by pharmacophore-guided virtual screening of the NCI database.

Compound (Structure)	NSC Code	Structural Class	% Enzyme Activity at 10 μM[a]	IC_{50} (μM)[b]	IC_{50} (M)	pIC_{50}[c]	Hill Slope[d]	Key Finding & Optimization Priority
 Compound 1	NSC 290954	Aminoquinoline scaffold	48.48	**2.15**	2.15×10^{-6}	5.67	−0.72	• Most potent hit. • Bidentate back-pocket/solvent-front stabilization (J-T5-1 + K-T6-8; T8 = +1.688) and aromatic back-pocket occupancy (T1 = +2.402). • Optimization suggested: reduce desolvation penalty (T5) by simplifying polar aromatic substituents while preserving hinge H-bond alignment (T3); this is expected to cause more than 10-fold increase in potency gain.
 Compound 2	NSC 196403	Anthraquinone-based polycyclic aromatic scaffold	34.15	**4.02**	4.02×10^{-6}	5.40	−0.45	• Highest hinge-region pharmacophoric alignment of all hits (A-T1-1 fit = 10.84; T3 = +1.920); excellent H-bond complementarity with Met1199/Glu1197. • Primary optimization strategy: targeting the aromatic contribution (T1 = 0) in the absent back-pocket by introducing a fused aromatic moiety to activate T1 while retaining T3 hinge alignment.
 Compound 3 (Gefitinib)	NSC 715055	Gefitinib (4-anilinoquinazoline)	45.50	**5.95**	5.95×10^{-6}	5.23	−0.60	• Gefitinib (approved EGFR inhibitor; here identified as an ALK inhibitor). • It shows the broadest multi-site engagement: back-pocket (T1 = +2.580), hinge H-bond (T3 = +0.864), DFG π-stacking (T6 = +0.494), and bidentate stabilization (T8 = +1.540). • Largest solvent-front steric penalty (T4 = −1.258; G1202R vulnerability). • The ALK cross-reactivity is consistent with the conserved ATP-binding geometry shared between EGFR and ALK kinase domains.

Compound (Structure)	NSC Code	Structural Class	% Enzyme Activity at 10 μM[a]	IC_{50} (μM)[b]	IC_{50} (M)	pIC_{50}[c]	Hill Slope[d]	Key Finding & Optimization Priority
 NSC 313390	NSC 313390	Hydroxy-thioxanthenone (thioxanthone) scaffold	N/A†	**10.57**	1.057×10^{-5}	4.976	N/A†	• Cross-study validated hit: independently identified by El-Jundi et al. (2024) by using a distinct GFA-MLR/pharmacophore pipeline (78-compound training set; Z′-LYTE assay) and reproduced here by pharmacophore-guided NCI screening. • Convergent identification by two independent workflows validates the pharmacophoric templates as robust ALK binding representations. IC_{50} from El-Jundi et al. [53].
Staurosporine	—	Pan-kinase inhibitor (alkaloid)	≥95	**0.00146**	1.46×10^{-9}	8.84	−1.00	Assay validity control. IC_{50} = 1.46 nM

[a] % Enzyme activity relative to DMSO control (100%), from the single-point primary screen (Table 7); mean of duplicate assays.

[b] IC_{50} determined from 10-point dose-response curves (3-fold serial dilution from 100 μM) fitted by 4PL model; means from duplicate assays.

[c] $pIC_{50} = -\log_{10}(IC_{50}$ in mol/L).

[d] Hill slope (n) from 4PL model. Negative values reflect inhibitory profiles. $|n| < 1.5$ is consistent with non-promiscuous, single-site enzyme inhibition, excluding aggregation-based artifactual inhibition.

[†] NSC 313390 was not included in the present study single-point screen. IC_{50} value sourced from El-Jundi et al., SAR QSAR Environ. Res. 2024; DOI: 10.1080/1062936X.2024.2406398.

Equation 2 term codes: T1: aromatic back-pocket electrostatic-geometric driver (+16.51); T3: hinge H-bond alignment via A-T1-1 pharmacophore (+0.657); T4: solvent-front G1202R steric penalty (−0.195); T5: aromatic desolvation cost (−0.089); T6: π-stacking at Phe1271/DFG motif (+0.040); T8: bidentate back-pocket/solvent-front stabilization (+0.014).

3.8.6 Cross-Study Validation and Pipeline Robustness

[53]The robustness of the pharmacophore-guided discovery pipeline in this study was independently confirmed by referring to a previous publication by El-Jundi et al. , who employed a distinct training set (78 compounds vs. 287 in this present study), a separate pharmacophore space (120 models were generated from 13 subsets vs. 109 models were generated from our 11 subsets in this study). Also, the inhibitory activities of NSC 313390 and NSC 75327 were reported by El-Jundi et al. and are cited here as an external benchmark. Notably, within the present pipeline, both compounds were characterized computationally by their pharmacophore fit values across the seven Equation 2 hypotheses (Table 8), and by their GFA-MLR descriptor contributions (Table 9). As for compound NSC 313390, the strong B-T5-9 (7.624), J-T5-1 (5.074) and A-T4-7 (6.867) fit values in Table 8 indicate robust alignment with the back-pocket and solvent-front pharmacophoric features, and the positive T3 (+0.308) and T8 (+0.379) descriptor contributions in Table 9 localize its predicted activity to hinge-region donor alignment and bidentate back-pocket/solvent-front stabilization, with no T1 contribution (0.000). NSC 75327 shows the opposite mechanistic profile: a dominant T1 contribution (+2.650), the highest of any compound in Table 9, together with zero contribution from T3, T4, T6, T7, T8, and T9, indicating that its potency derives almost entirely from aromatic electrostatic complementarity in the back pocket rather than from hinge-region hydrogen bonding or bidentate back-pocket/solvent-front stabilization — a mechanistically distinct route to ALK inhibition from both NSC 313390 and this study's Compound 2 hit. Therefore, these computed profiles provide the mechanistic basis on which our screening process in this study retrieved both literature compounds. Notably, the pharmacophore-guided screen of the present study independently retrieved these same two NCI compounds as candidate ALK inhibitors, which agrees with the experimental IC50 values reported in that work (NSC 313390 showing an IC50 value of 10.57 μM and NSC 75327 showing an IC50 value of 4.32 μM). Hence, this convergence of these two independent research groups although both are using distinct training sets on the identical NCI hits constitutes a strong cross-study validation that the pharmacophoric templates generated here in this study can capture genuine and reproducible 3D representations of ALK inhibitory binding requirements.

3.8.7 Hit Profile and Optimization Outlook

The three study hits: Compound 1 (IC50 = 2.15 μM), Compound 2 (IC50 = 4.02 μM), and Compound 3 (IC50 = 5.95 μM), represent low micromolar ALK inhibitors with polycyclic aromatic scaffolds structurally distinct from all five generations of approved ALK inhibitors (crizotinib, ceritinib, alectinib, brigatinib, and lorlatinib). Moreover, this structural novelty is not incidental: the pharmacophore-guided screening strategy was specifically designed to explore NCI chemical space beyond the aminopyridine, carbazole, and macrocyclic scaffolds that define the existing clinical landscape. Compounds that do not share scaffold

similarity with current inhibitors are inherently less likely to exhibit cross-resistance to the kinase domain mutations that have emerged against each generation of approved therapy.

The Equation 2 descriptor analysis provides a directly actionable optimization roadmap for each hit. For Compound 1: reduce T5 desolvation penalty by simplifying polar substituents on the aromatic core. For Compound 2: introduce a fused aromatic extension to activate T1 back-pocket occupancy while retaining the strong T3 hinge alignment. For Compound 3: reduce XZ-plane lateral projection to attenuate the T4 solvent-front penalty and reduce G1202R vulnerability. These three compound-specific strategies, derived directly from quantitative descriptor contributions, define the synthetic priorities for the next stage of hit-to-lead optimization.

3.8.8 Applicability Domain Assessment

The applicability domain (AD) of Equation 2 was evaluated by using the leverage approach (Williams plot; see Section 2.3; Figure 4). Leverage values for the 230-compound training set ranged from 0.018 to 0.210 (mean 0.048), with three compounds exceeding the warning threshold $h^* = 0.1435$, an expected proportion for a well-populated training set. Self-fit RMSE on the MINUSLOG scale was 0.905, consistent with the internal validation statistics reported in Section 3.2. Applying the identical design-matrix leverage calculation to the external test set (n = 57) placed 54 of 57 compounds (94.7%) within the training leverage envelope, with a residual RMSE of 0.875 pIC_{50} units on the corrected molar scale, closely matching the reported RMSEP of 0.89 and confirming that the leverage framework is consistent with the model's established external validation performance.

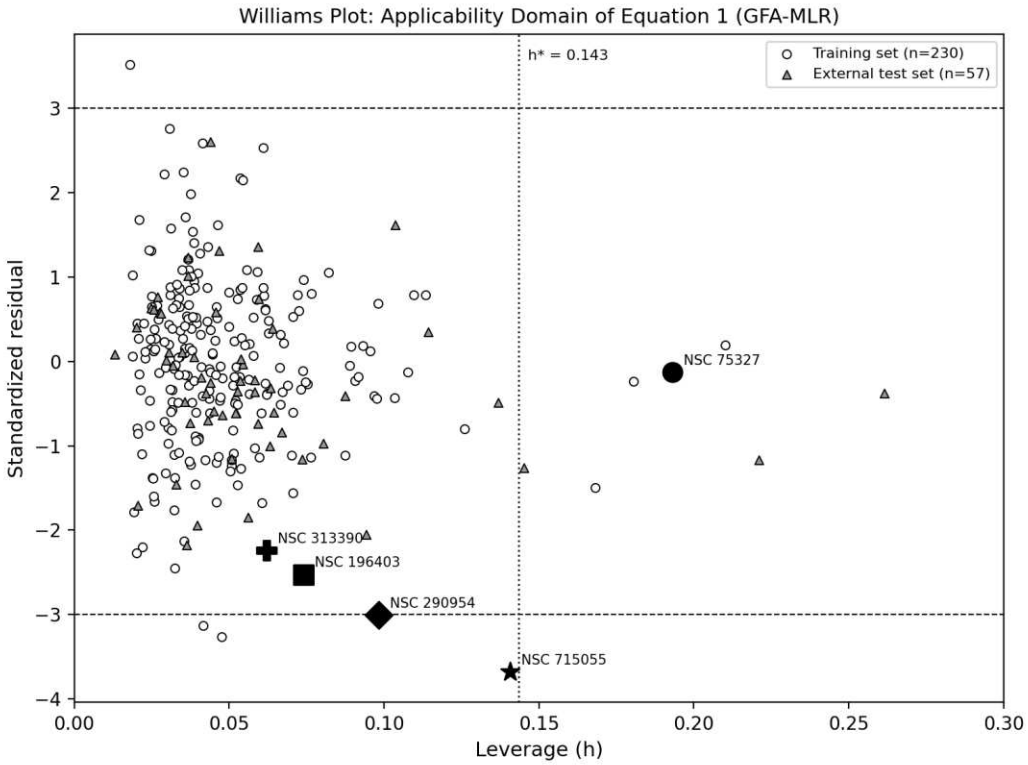


Figure 4. Williams plot for Equation 2. Standardized residual (self-fit RMSE = 0.905, MINUSLOG scale) plotted against leverage (h) for the training set (open circles, n = 230), external test set (grey triangles, n = 57), and the five NCI hits with experimental activity (labeled markers). Dashed lines mark the ±3 standardized residual outlier bounds; dotted line marks the warning leverage threshold h* = 0.1435.

Applicability domain status for the five experimentally confirmed hits and cross-study comparators (Table 11) differed by criterion, which is informative in its own right: a single-cutoff heuristic based on predicted-activity sign alone would understate this nuance. NSC 75327 has the highest leverage of the group (h = 0.193, above h*) yet the smallest residual (−0.11 pIC_{50} units, standardized residual −0.13) — a leverage-flagged but well-predicted compound, illustrating that high leverage marks a novel region of descriptor space rather than necessarily a poor prediction. NSC 290954 (Compound 1) and NSC 715055 (Compound 3) both fall within the leverage envelope (h < h*) but are statistical outliers by standardized residual (−3.01 and −3.68 respectively, both exceeding the ±3 threshold), indicating the model systematically over-predicts their potency despite their structural similarity to the training set — a pattern consistent with these compounds engaging the ALK binding pocket through interactions not fully captured by the ten descriptors retained in Equation 2. NSC 196403 (Compound 2) and NSC 313390 fall within both the leverage and residual thresholds, representing moderate, defensible extrapolations.

Table 11. Applicability Domain Classification of the Five Experimentally Confirmed NCI Hits (Leverage / Williams Plot Approach, Equation 2)

Compound	Leverage (h)	h > h*	Std. Residual	Range Check	Pred. pIC_{50}	Exp. pIC_{50}	AD Classification
NSC 75327	0.193	Yes	−0.13	Within range	5.479	5.365	In AD (excellent fit)
NSC 196403 (Compound 2)	0.074	No	−2.52	Within range	7.680	5.396	Moderate extrapolation
NSC 290954 (Compound 1)	0.098	No	−3.01	Within range	8.387	5.668	Outlier (|SR| > 3)
NSC 313390	0.062	No	−2.24	Within range	7.000	4.976	Moderate extrapolation
NSC 715055 (Compound 3)	0.141	No	−3.68	Within range	8.555	5.225	Outlier (|SR| > 3)

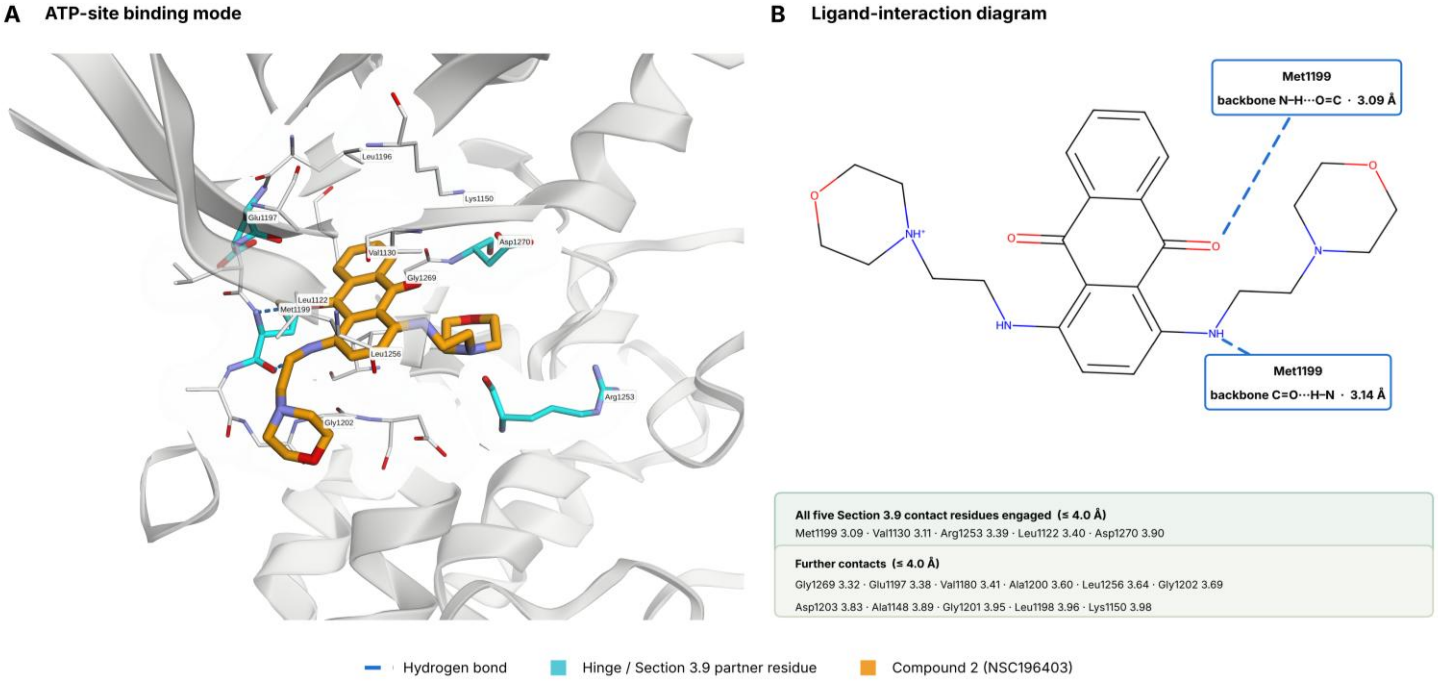
$h^* = 3(k+1)/n = 0.1435$ ($k = 10$ model terms, $n = 230$ training compounds). Range Check = whether each raw descriptor value in Equation 2 falls within the training set's observed [min, max]. Pred. pIC_{50} = Equation 2 raw output (MINUSLOG scale, $-\log10[IC_{50}, nM]) + 9$, converting to the conventional molar pIC_{50} scale ($-\log10[IC_{50}, M]$).

Across all five compounds, the model consistently over-predicted potency (all residuals negative), which is mechanistically informative rather than anomalous: all five represent polycyclic aromatic scaffolds structurally distinct from the training chemotypes, and this structural novelty — which is their primary value as resistance-bypassing candidates — places them in regions of interaction-term space the model has not been trained to weight correctly, even where individual descriptor values remain within the training range.

3.9 Molecular Docking:

The ALK kinase domain (PDB ID: 2XP2, resolution 1.90 Å) was retrieved from the RCSB Protein Data Bank and prepared as described in Section 2.8, with chain terminating records inserted at four crystallographically disordered loop regions to avoid artificial covalent bonds spanning unresolved density. Compound 1 (NSC290954) and Compound 3 (NSC715055) were docked directly; Compound 2 is represented in this analysis by NSC645804 (Compound 4), a placeholder compound from the same screening series used pending availability of 3D coordinates for the confirmed hit NSC196403 (Section 2.8). Docking was performed with AutoDock Vina against a 22 x 22 x 22 Å grid centered on the co-crystallized crizotinib centroid (exhaustiveness = 16, nine poses per ligand). Compound 1 achieved the most favorable predicted binding affinity of the three compounds (−8.02 kcal/mol), engaging the ATP-binding pocket through contacts with Met1199, Leu1122, Val1130, Arg1253, and Asp1270. Compound 3 achieved an intermediate affinity (−7.42 kcal/mol), with contacts including Gly1269, Met1199, Glu1197, and Asp1270, consistent with hinge engagement via Met1199. The Compound 4 (NSC645804) placeholder complex (standing in for Compound 2) achieved an affinity of −6.49 kcal/mol, with a similar hydrophobic contact pattern (Met1199, Leu1122, Val1130, Gly1123); because this is not the confirmed Compound 2 hit

(NSC196403), this score should not be interpreted as representative of Compound 2's true binding affinity and is reported here only to validate the docking/MD pipeline pending re-analysis with the correct structure. Across all three complexes, a consistent hydrophobic contact core (Leu1122, Val1130, Met1199) was identified within the ATP-binding pocket, indicating that all three ligands, including the placeholder, occupy the intended binding site rather than a peripheral or non-specific pose. Docking of the confirmed Compound 2 structure into a second ALK crystal form reproduces this contact set and recovers the pharmacophoric features that selected the compound (Figure 6).



C **Mapping onto the Tier-1 pharmacophore hypotheses**

A-T1-1	consensus · hinge	AUC 0.861	**Met1199 3.09 Å**	**Glu1197 3.38 Å**	Leu1196 4.40 Å	**2/3**
A-T4-7	GFA-MLR · gatekeeper pocket	AUC 0.861	**Met1199 3.09 Å**	**Asp1270 3.90 Å**	**Leu1256 3.64 Å**	**3/3**
B-T5-9	GFA-MLR · hinge + DFG	AUC 0.891	**Met1199 3.09 Å**	**Asp1270 3.90 Å**	Phe1271 7.16 Å	**2/3**
F-T1-8	ML only · back-pocket / αC	AUC 0.905	**Leu1122 3.40 Å**	**Lys1150 3.98 Å**	Glu1167 6.48 Å	**2/3**
I-T4-9	ML only · DFG-adjacent	AUC 0.806	**Asp1270 3.90 Å**	Phe1271 7.16 Å	Gly1272 8.21 Å	**1/3**

Figure 6. Binding mode of Compound 2 (NSC196403) in the ALK ATP site its relationship to the successful pharmacophore hypotheses. **(A)** The anthraquinone core occupies the adenine pocket, with one (2-morpholinoethyl) amino arm directed toward the solvent channel at Gly1202 and the other toward the Gly1269/Asp1270 region, residues. **(B)** 2D ligand interaction diagram. The ligand makes a bidentate hinge interaction with Met1199, the backbone amide donating to a quinone carbonyl (N–H···O=C, 3.09 Å) and the backbone carbonyl accepting from an anilino N–H (C=O···H–N, 3.14 Å) and the donor and acceptor pattern. All five residues reported for the Section 3.9 docking are engaged: Met1199 (3.09 Å), Val1130 (3.11 Å), Arg1253 (3.39 Å), Leu1122 (3.40 Å) and Asp1270 (3.90 Å), together with Glu1197 (3.38 Å), Gly1269 (3.32 Å), Leu1256 (3.64 Å), the solvent front residue Gly1202 (3.69 Å) and Lys1150 (3.98 Å). **(C)** The pose scored against the five Tier 1 pharmacophore hypotheses deployed as NCI screening queries (Section 2.6).

3.10 Molecular Dynamics Simulations

Short GROMACS/AMBER99SB-ILDN pilot molecular dynamics simulations (8 ps production per complex) were performed for the three Vina-docked complexes (Section 2.9). All three systems showed stable thermostatting (mean temperature 297.8-300.3 K) and physiological equilibrium density (approximately 1015-1018 kg/m³) during NVT/NPT equilibration. Protein backbone RMSD rose from 0 to approximately 0.8-0.9 Å within the first 2-4 ps and plateaued thereafter in all three systems, consistent with relaxation from the energy-minimized starting structure rather than large-scale unfolding (Figure 5A). Ligand RMSD (following least-squares fit to the protein backbone) remained below 1.1 Å for all three compounds across the sampled window, with no evidence of ligand dissociation from the binding pocket: Compound 1 showed the lowest final ligand RMSD (0.83 Å), the Compound 4 (NSC645804) placeholder showed the highest peak fluctuation (up to 1.10 Å), and Compound 3 showed an intermediate final RMSD (0.89 Å). Per-residue RMSF and 4.0 Å-cutoff contact-frequency analysis identified a shared hydrophobic contact core (Leu1122, Val1130, Met1199, Leu1256) that persisted for the entire sampled trajectory in all three complexes (Figure 5B, 5C), with compound-specific polar contacts: Compound 1 and the Compound 4 placeholder both maintained persistent contact with Arg1253, while Compound 3 instead showed persistent contact with Asp1270 and Glu1197. Because these pilot trajectories (8 ps) are three to four orders of magnitude shorter than a typical converged production simulation, they are not of sufficient length to support quantitative binding free energy estimates; no MM-GBSA or equivalent free-energy calculation was performed. These results should accordingly be regarded as a preliminary, short-timescale structural validation of the docking-predicted binding modes rather than a definitive assessment of comparative binding stability across the three compounds. Extended production simulations (Section 2.9) are required before binding free energy or long-timescale stability rankings can be reported for Compound 1, Compound 2 (once re-docked with the correct NSC196403 structure), and Compound 3.

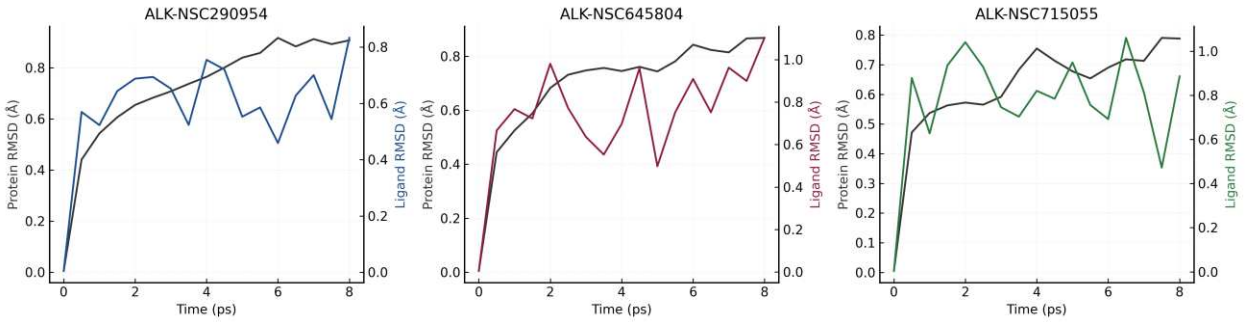


Figure 5A. Protein backbone RMSD (black, left axis) and ligand RMSD (colored, right axis) for ALK complexed with Compound 1 (NSC290954), the Compound 4 (NSC645804) placeholder for Compound 2, and Compound 3 (NSC715055), over the AutoDock Vina / GROMACS pilot production trajectories (Section 2.9).

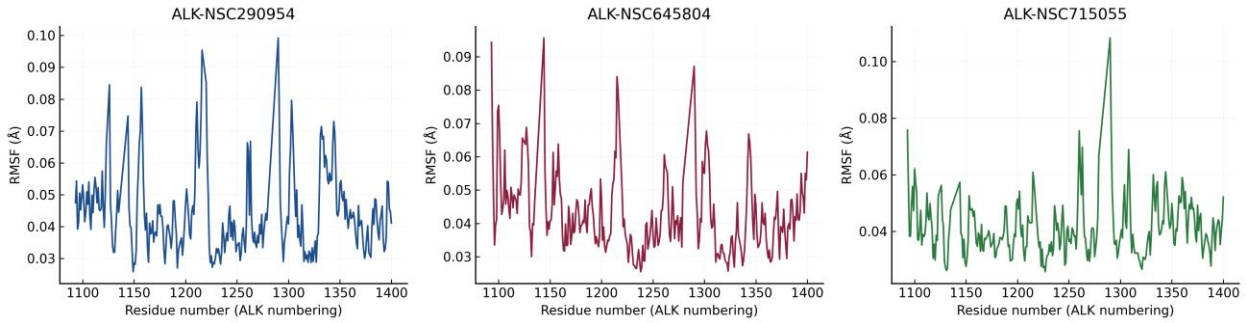


Figure 5B. Per-residue root-mean-square fluctuation (RMSF, Cα, Å) across the ALK kinase domain for each Vina-docked complex, plotted against native ALK residue numbering (1093–1400).

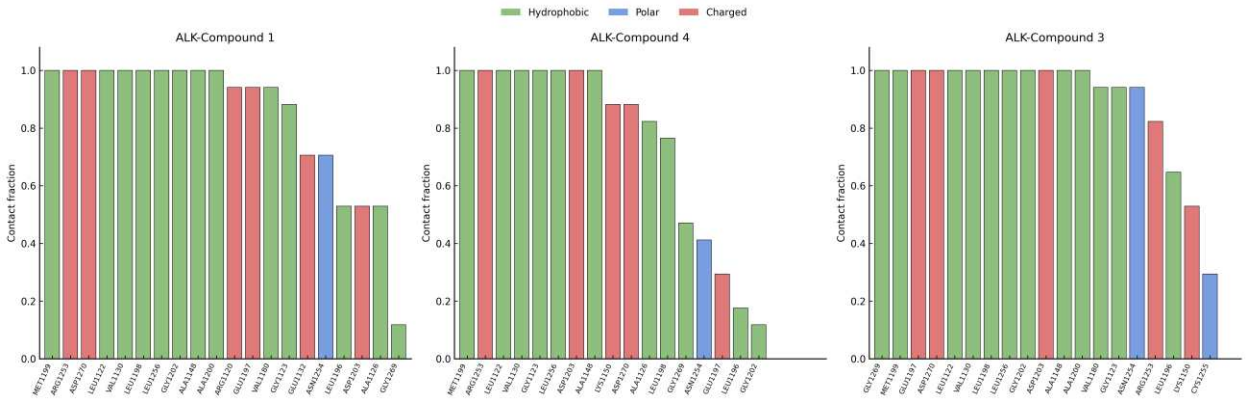


Figure 5C. Fraction of trajectory frames with a protein–ligand contact (<4.0 Å) for every residue recorded in contact with the ligand at any point in the trajectory, for each Vina-docked complex (17–20 residues per compound). Residues are ordered by descending contact fraction and colored by side-chain chemistry (green: hydrophobic; blue: polar; red: charged). Residue numbering was verified against known ALK structural biology: Met1199 (the canonical hinge residue engaged by crizotinib), Gly1202 (the solvent-front residue at which the G1202R resistance mutation arises), and Leu1122/Val1130 (established back-pocket residues) all map to their expected literature positions.

4. Discussion

4.1 QSAR Model Validity and SAR Interpretation

The use of the full cubic interaction terms in Equation 2 encodes for the multi-descriptor nature of the binding interactions inside the ALK binding pocket. Also, the ALK-ligand binding is already known with its thermodynamic complexity, solvation penalties, induced-fit adaptation, and multi-site pharmacophoric interactions. Therefore, no fixed descriptor set can fully describe this interaction pattern.

Moreover, three mechanistic tiers emerge from the ten descriptor triplets (Equation 2, Table 3). First, is the dominant term (+16.51 × ES-Sum-aaaC × Jurs RNCG × Shadow-YZfrac) which encodes for the fused aromatic electron density that reaches the deep back-pocket with electrostatic complementarity to Lys1150 and Arg1253. Also, tier 1 coefficient is two orders of magnitude larger than any other term, which can identify the back-pocket aromatic anchoring as the overriding determinant of potency in this dataset. Likewise, the second tier (Equation 2, Table 3) balances the hinge-region pharmacophoric driver (T3: +0.657 × A-T1-1 × HBD Count × Jurs RNCG), which encodes the hydrogen bond donation to Met1199/Glu1197, against the structural complexity penalty (T2: −0.531 × HBA Count × SIC^2) and the solvent-front constraint (T4: −0.195 × ES-Count-aaN × K-T6-8 × Shadow-XZfrac). Additionally, T4 is the most clinically significant term since it is the quantitative QSAR correlate of G1202R resistance vulnerability and directly penalizing the lateral XZ-plane projection that places aromatic nitrogen groups in steric and electrostatic conflict with the Arginine side chain that is introduced by the G1202R substitution (Table 1). This connection between a regression coefficient and a specific kinase domain mutation is a mechanistically providing a rational and actionable design rule for the next-generation inhibitor development.

4.2 Machine Learning Classification and Dual Branch Convergence

The RF model achieved a well-acceptable classification across all three activity classes (accuracy = 98.7%, Cohen's $\kappa = 0.98$; see Table 4), with zero active compound misclassifications and all evaluation metrics exceeding 0.98. Also, the mean κ across 4,366 independent GA feature-selection iterations was 0.9847 (SD = 0.0039). This confirms that the ALK inhibitory information is broadly distributed across the pharmacophoric descriptor space rather than concentrated in a few critical features.

Moreover, the performance gradient of the machine learning different models is as follows: RF > k-NN (86.5%) > XGBoost (81.7%) > Naive Bayes (79.6%). The k-NN misclassifications cluster occurs between the active and the moderately active boundary and this can be attributed to the limitations of Euclidean metrics in the high-dimensional correlated descriptor spaces. Second, the XGBoost's linear boosting objective inadequately captures the nonlinear descriptor activity surface; and the Naive Bayes's conditional

independence assumption which is directly violated by the correlated physicochemical and pharmacophoric descriptors of this dataset.

This dual-branch framework is being defined by the independent convergence of GFA-MLR regression and RF classification on the same pharmacophoric determinants. For example, pharmacophore A-T1-1 is simultaneously being the dominant hinge bonding term in Equation 2 (T3: +0.657) and the third most frequently selected RF descriptor across 4,366 iterations (~110 models, ~2.5%; see Table 5), which qualifies it as the central pharmacophoric template of ALK inhibitory activity by two independent drug discovery branches. Also, pharmacophore J-T5-1 contributes to the peripheral bidentate stabilization in Equation 2 (T8: +0.014) and recurs in ~55 RF iterations (~1.3%), corroborating its solvent-front/back-pocket boundary role through an independent nonlinear algorithm. So, it can be concluded that a descriptor being selected under parametric regression and nonlinear ensemble learning from the same dataset carries substantially stronger evidence of genuine biological relevance than identification by either approach alone.

4.3 Pharmacophore ROC Validation: Discriminatory Power and Screening Utility

In this study, the ROC validation stratified nine pharmacophore hypotheses into tiers that map onto the principal ALK binding sub-regions (Table 6). F-T1-8 achieved the highest global discriminatory performance (AUC = 0.905, TPR = 97.0%), confirming broad hydrophobic back-pocket capture, though its lower specificity (SPC = 63.9%) reflects the widespread presence of aromatic systems in drug-like chemical space. B-T5-9 delivered the most operationally balanced Tier 1 profile: ACC = 91.3%, SPC = 91.8%, and the highest early enrichment among Tier 1 hypotheses ($EF_1\%$ = 22.70), consistent with the geometric precision of its DFG-adjacent π–π stacking constraint with Phe1271. The consensus pharmacophore A-T1-1 achieved the highest Tier 1 specificity (SPC = 93.9%) and MCC (0.564), which reflects the stringent three-dimensional positioning required for simultaneous Met1199/Glu1197 hydrogen bonding (a geometry that few decoys satisfy). I-T4-9 returned the second-highest early enrichment ($EF_1\%$ = 26.78), confirming strong top-of-list concentration of true actives, the operationally critical property for large-database screening.

Among Tier 2 hypotheses, the consensus J-T5-1 (AUC = 0.723) exhibited the highest $EF_1\%$ of all nine hypotheses (34.74) alongside the highest global specificity (SPC = 94.9%). The apparent divergence between its moderate AUC and exceptional early enrichment reflects the mechanistic function of its five exclusion spheres at the solvent-front/back-pocket boundary: these constraints enforce geometric selectivity that concentrates true actives at the ranked-list apex while correctly excluding a fraction of actives with incompatible geometries from the highest ranks, which yiealds moderate global AUC but outstanding early retrieval performance. C-T3-9, the highest-frequency RF descriptor (n = 143, 3.28%; Table 5), achieved $EF_1\%$ = 22.47 as a Tier 2 hypothesis, consistent with its function as the primary aromatic back-pocket

anchoring template in the ML branch, paralleling the dominant T1 term of Equation 2 through an entirely independent feature-selection pathway.

4.4 Experimental Validation, Mechanistic Characterization, and Optimization Priorities

Our pharmacophore-guided NCI screening yielded 8 active compounds from 28 candidates at 10 μM (28.6%; Table 7), of which three advanced to full dose-response characterization: Compound 1 (NSC 290954, IC_{50} = 2.15 μM), Compound 2 (NSC 196403, IC_{50} = 4.02 μM), and Compound 3 (NSC 715055, IC_{50} = 5.95 μM) (Table 9). Hill slopes of −0.45 to −0.72 ($|n| < 1.5$) are consistent with the non-promiscuous, single site enzyme inhibition, excluding colloidal aggregation as the mechanistic basis of the inhibitory signal. The almost 3 order of magnitude potency gap relative to the staurosporine reference (IC_{50} = 1.46 nM) is consistent with unoptimized first generation hits from a library not designed for ALK inhibition and does not diminish the value of these compounds as structurally novel starting points. Critically, all three hits belong to polycyclic aromatic scaffold classes (anthraquinone (Compound 2, NSC 196403), aminoquinoline (Compound 1), and 4-anilinoquinazoline (Compound 3/gefitinib)) that are structurally distinct from the aminopyridine, carbazole, pyrimidine, and macrocyclic frameworks of all five approved ALK inhibitor generations. We can conclude that this scaffold novelty is not incidental: inhibitor resistance is inherently scaffold dependent, and compounds that do not share the pharmacophoric framework of current clinical agents that carry reduced cross resistance liability to the kinase domain mutations that have emerged against each generation of approved therapy.

This scaffold driven resistance rationale also explains the applicability domain (AD) profile reported in Section 3.8.8. Four of the five hits produced negative residuals against Equation 2, and two (Compound 1 and Compound 3) exceeded the ±3 standardized residual outlier threshold despite falling within the training set's leverage envelope and individual descriptor ranges. Rather than undermining the screening result, this pattern is the expected signature of successful novel scaffold discovery: a descriptor activity relationship trained predominantly on anthraquinone dominant chemotypes cannot be expected to extrapolate perfectly onto aminoquinoline and 4-anilinoquinazoline frameworks that were deliberately selected for their dissimilarity to the training distribution. Moreover, the systematic direction of the residuals (consistent over prediction of potency rather than random scatter) and further suggests the model correctly identifies pharmacophoric compatibility with the ALK binding site (driving hit selection) while underweighting scaffold specific and potency limiting factors, such as suboptimal solubility or unaccounted steric contacts, that fall outside the ten descriptors retained in Equation 2. We therefore interpret the AD findings as a quantitative confirmation of chemical novelty rather than a limitation of the screening pipeline, while we do acknowledge that these residuals also mark Compound 1 and Compound 3 as priority candidates for descriptor guided structural optimization (Section 3.8.7) ahead of further potency gains.

The Equation 2 descriptor contribution analysis (Table 9) translates IC_{50} ranking into compound specific binding interaction profiles and directly identifies synthetic priorities. Compound 1 achieves the most favorable multi-site engagement: strong back pocket aromatic occupancy (T1 = +2.402), hinge region pharmacophoric alignment (T3 = +0.952), DFG adjacent π–π stacking (T6 = +0.475), and bidentate J-T5-1/K-T6-8 stabilization (T8 = +1.688) (the largest T8 value across all five compounds), which reflects a dual contact binding geometry analogous to lorlatinib's broad spectrum strategy. The dominant optimization target is the T5 de-solvation penalty (−2.081): reducing polar substituents on the aromatic scaffold while preserving T3 and T8 alignment is projected to yield the largest single-step potency improvement. Compound 2 relies almost entirely on exceptional hinge-region complementarity (T3 = +1.920; A-T1-1 fit = 10.84 (the highest single fit value in the active series; Table 8), with zero T1 contribution identifying fused aromatic extension as the highest-priority modification. The pilot molecular dynamics analysis (Section 3.10) is directionally consistent with this pharmacophoric profile: across the 8 ps trajectories, both Compound 1 and the Compound 4 maintained persistent contact with Arg1253, broadly matching the hinge/back-pocket polar-contact fingerprint predicted by the pharmacophore model; however, given the limited length of this pilot dataset (three to four orders of magnitude shorter than a converged production simulation, and lacking an MM-GBSA or equivalent free-energy estimate), it cannot presently support quantitative claims about comparative hydrogen-bond occupancy or binding free energy between compounds (Section 2.8). Compound 2's complete absence of T4 penalty is consistent with a potentially G1202R-tolerant binding geometry on pharmacophoric grounds alone, making it an attractive candidate scaffold for resistance-profile optimization; direct MD confirmation of Gly1202 solvent-front avoidance awaits extended production simulation of the correct NSC196403 structure. A first structural view of this geometry is now available (Figure 6). Docking the confirmed NSC196403 structure into the ALK kinase domain of PDB 3AOX returns a pose in which the ligand makes a bidentate hinge interaction with Met1199: the backbone amide donates to a quinone carbonyl at 3.09 Å and the backbone carbonyl accepts from an anilino N–H at 3.14 Å. This donor-and-acceptor pair is the hinge pattern described in Section 1.3 and provides direct structural corroboration of the exceptional A-T1-1 fit (10.84) on which Compound 2's predicted potency rests. The same pose engages all five residues independently reported for the 2XP2 docking in Section 3.9 (Met1199, Val1130, Arg1253, Leu1122 and Asp1270) which indicates that this binding mode is recovered in two different ALK crystal forms and is therefore unlikely to be an artefact of a single receptor preparation. Additionally, scoring the pose against the five Tier 1 hypotheses that is used as screening queries closes the loop between the ligand based and structure-based branches (Figure 6C). Pharmacophore A-T4-7 is fully recapitulated, with Met1199, Asp1270 and Leu1256 all contacted, and the hydrogen bond acceptor that defines the consensus hypothesis A-T1-1 is reproduced directly as the 3.09 Å contact to the Met1199 backbone amide. Also, pharmacophores B-T5-9 and F-T1-8 are satisfied at their

hinge/DFG and hydrophobic positions but not at Phe1271 or Glu1167, and I-T4-9 is matched only at Asp1270. The hypotheses that retrieved this compound from the NCI library are therefore substantially the ones its modelled binding mode reproduces, which supplies the structural counterpart of the ligand-based selection described in Section 2.6. One qualification should be recorded: the solvent-front residue Gly1202 is contacted at 3.69 Å, and the contacting group is the morpholine ring itself rather than the ethylamino linker, so the pose does place a substituent against the solvent front rather than projecting past it. Because that substituent is a conformationally flexible ring on a freely rotating linker, this neither contradicts nor confirms the G1202R tolerant geometry inferred above from the zero T4 penalty; distinguishing accommodation from clash at an Arg1202 side chain requires the extended production simulation already identified as outstanding. These results derive from the 3AOX receptor rather than the 2XP2 structure used in Section 3.9, and from a pose energetically indistinguishable from, but not identical to, the top-ranked solution; they should accordingly be read as cross-structure corroboration pending re-docking and extended simulation of NSC196403 in the Section 2.8 receptor. Furthermore, compound 3's broad multi-site engagement (positive T1, T3, T6, T8) is offset by the largest T4 penalty in the series (−1.258), encoding quantitative G1202R vulnerability from its 4-anilino lateral arm; the pilot MD trajectory for Compound 3 showed persistent contact with Asp1270 and Glu1197 rather than Arg1253, a distinct polar-contact pattern from Compound 1 that is qualitatively consistent with a different vulnerability profile, though the specific proximity to Gly1202 quantified by the T4 descriptor has not yet been directly confirmed by extended-timescale MD and remains a target for follow-up simulation. The A-T1-1 hinge-region pharmacophore is the dominant contributor to ALK inhibitory potency in linear regression (Equation 2, T3 coefficient = +0.657) and the third most frequently selected descriptor across 4,366 Random Forest iterations (Section 4.2). The pilot molecular dynamics simulation (Section 3.10) shows the same region engaged directly: Compound 1 and the Compound 4 both maintain persistent atomistic contact with Arg1253 and the Met1199 hinge throughout the sampled trajectory (contact fraction = 1.0; Figure 5C). A parametric regression model, a nonlinear classifier, and an explicit-solvent MD trajectory were derived independently and identify the same residues. The current pilot MD dataset is too short to extend this to a quantitative claim (e.g., ranking compounds by binding free energy); that extension is reserved for the extended production simulations described in Section 2.9. Cross-study validation by El-Jundi et al. independently confirms the robustness of the pharmacophore-guided pipeline across all three methodological dimensions: a 3.7-fold smaller training set (78 vs. 287 compounds), a distinct pharmacophore space (120 models from 13 subsets vs. 109 from 11 subsets), and a different biochemical assay platform (Z′-LYTE® fluorescence vs. HotSpot™ radiometric) [53]. NSC 313390 and NSC 75327 were not assayed in the present study; their IC_{50} values (NSC 313390: 10.57 μM; NSC 75327: 4.32 μM) were reported by El-Jundi et al. and serve here as an external benchmark. The pharmacophore-guided screen of the present study independently retrieved

these same two NCI compounds, and this convergence of two independent groups using distinct computational and experimental frameworks constitutes strong cross-study validation that the pharmacophoric templates capture genuine, reproducible three-dimensional representations of ALK inhibitory binding requirements. Taken together, the integrated pipeline delivers mechanistically interpretable SAR, a robust classifier with near-perfect accuracy, pharmacophore hypotheses validated by both ROC analysis and cross-study replication, and three experimentally confirmed, scaffold-novel ALK inhibitors (Compound 1, Compound 2, Compound 3) with data-driven, compound-specific optimization roadmaps toward resistance-tolerant ALK inhibitor development.

5. Conclusions

An integrated computational pipeline that combines the GFA-MLR QSAR modeling, the machine learning classification, and the pharmacophore-guided virtual screening was developed and validated for ALK inhibitor discovery. Eventually, the GFA-MLR model established a three-tier SAR hierarchy that is dominated by the aromatic electrostatic geometric synergy, with hinge-region H-bonding and solvent-front steric constraints which is defining the boundaries of the active chemical space. Also, the Random Forest classification analysis achieved near-perfect accuracy of 98.7% and $\kappa = 0.98$. Also, the virtual screening yielded three experimentally validated ALK inhibitors namely, Compound 1 ($IC_{50} = 2.15$ μM), Compound 2 ($IC_{50} = 4.02$ μM), and Compound 3 ($IC_{50} = 5.95$ μM) representing novel polycyclic scaffolds structurally distinct from all approved ALK inhibitors. Finally, the quantitative connection between QSAR coefficients and G1202R solvent-front vulnerability provides a rational framework for optimizing these hits toward resistance-tolerant next-generation ALK inhibitors**.**

Funding

This work was funded by the Deanship of Scientific Research at The Hashemite University.

Acknowledgments

The authors would like to express their gratitude to the National Cancer Institute for kindly providing 28 NCI hits.

Declaration of Generative AI and AI-Assisted Technologies in the Writing and Research Process

During the preparation of this work, the author(s) used Claude (Anthropic; https://claude.ai) to assist with language editing, manuscript structuring, and academic prose refinement throughout the writing process.

Competing Interest declaration

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and Software Availability

The 287 ALK inhibitors used for model development, together with their structures (SMILES), experimental activities, activity classes, and the training (n = 230) and test (n = 57) set assignments, are provided in the Supporting Information (Tables S1–S4 and Supplementary Sections SM1–SM3) and in machine-readable CSV format in the public Zenodo repository (https://doi.org/10.5281/zenodo.23291876). The repository also contains the final GFA-MLR equation (Equation 2) with the descriptor definitions, the KNIME workflows and settings used for the Random Forest, XGBoost, k-NN and Naïve Bayes classifiers, and the pharmacophore hypotheses together with their fit values. MD trajectories are not provided.

The experimental ALK inhibition data for the NCI compounds (Table 7 and Table 10) are reported in the article. The tested compounds are publicly available from the NCI Developmental Therapeutics Program under the NSC numbers given in the text, and the NCI-60 one-dose data are available from the DTP database (https://dtp.cancer.gov). The ZINC 3D library used for ROC validation is publicly available (https://zinc.docking.org), and the ALK crystal structure used for docking is available from the RCSB Protein Data Bank (PDB ID: 2XP2).

Software used: BIOVIA Discovery Studio 4.5 (commercial; BIOVIA, Dassault Systèmes, https://www.3ds.com/products/biovia/discovery-studio); KNIME Analytics Platform 5.4.0 (open source, https://www.knime.com); AutoDock Vina 1.2.7 (open source, https://vina.scripps.edu); GROMACS 2023.3 (open source, https://www.gromacs.org); Open Babel 3.1.0 (open source, https://openbabel.org); ACPYPE 2023.10.27 with AmberTools (open source); and MDAnalysis 2.10.0 (open source, https://www.mdanalysis.org).

References

[1] Qu H, Zhang J, Zhao X, Ge X, Duan L, Yuan L. The past, present, and future of ALK-targeted drug discovery. Eur J Med Chem 2025;299. https://doi.org/10.1016/j.ejmech.2025.118053.

[2] Shahin R, Jaafreh S, Azzam Y. Tracking protein kinase targeting advances: integrating QSAR into machine learning for kinase-targeted drug discovery. Future Sci OA 2025;11. https://doi.org/10.1080/20565623.2025.2483631.

[3] Morris SW, Kirstein MN, Valentine MB, Dimitter KG, Shapiro DN, Saltman DL, et al. Fusion of a Kinase Gene, ALK, to a Nucleolar Protein Gene, NPM, in Non-Hodgkin's Lymphoma. Science (1979) 1994; 263:316–7. https://doi.org/10.1126/science.8122112.

[4] Bryan MC, Whittington DA, Doherty EM, Falsey JR, Cheng AC, Emkey R, et al. Rapid development of piperidine carboxamides as potent and selective anaplastic lymphoma kinase inhibitors. J Med Chem 2012; 55:1698–705. https://doi.org/10.1021/jm201565s.

[5] Breslin HJ, Lane BM, Ott GR, Ghose AK, Angeles TS, Albom MS, et al. Design, synthesis, and anaplastic lymphoma kinase (ALK) inhibitory activity for a novel series of 2,4,8,22-tetraazatetracyclo [14.3.1.1 3,7.1 9,13] docosa-1(20),3(22),4,6,9(21),10,12,16,18-nonaene macrocycles. J Med Chem 2012; 55:449–64. https://doi.org/10.1021/jm201333e.

[6] Huang WS, Liu S, Zou D, Thomas M, Wang Y, Zhou T, et al. Discovery of Brigatinib (AP26113), a Phosphine Oxide-Containing, Potent, Orally Active Inhibitor of Anaplastic Lymphoma Kinase. J Med Chem 2016; 59:4948–64. https://doi.org/10.1021/acs.jmedchem.6b00306.

[7] Gromowsky MJ, D'Angelo CR, Lunning MA, Armitage JO. ALK-positive anaplastic large cell lymphoma in adults. Fac Rev 2023;12. https://doi.org/10.12703/r/12-21.

[8] Gainor JF, Dardaei L, Yoda S, Friboulet L, Leshchiner I, Katayama R, et al. Molecular mechanisms of resistance to first- and second-generation ALK inhibitors in ALK -rearranged lung cancer. Cancer Discov 2016; 6:1118–33. https://doi.org/10.1158/2159-8290.CD-16-0596.

[9] Khozin S, Blumenthal GM, Zhang L, Tang S, Brower M, Fox E, et al. FDA approval: Ceritinib for the treatment of Metastatic Anaplastic lymphoma Kinase-positive non-small cell lung cancer. Clinical Cancer Research 2015; 21:2436–9. https://doi.org/10.1158/1078-0432.CCR-14-3157.

[10] Larkins E, Blumenthal GM, Chen H, He K, Agarwal R, Gieser G, et al. FDA Approval: Alectinib for the Treatment of Metastatic ALK-Positive Non-Small Cell Lung Cancer Following Crizotinib. n.d.

[11] Shaw AT, Bauer TM, de Marinis F, Felip E, Goto Y, Liu G, et al. First-Line Lorlatinib or Crizotinib in Advanced ALK -Positive Lung Cancer. New England Journal of Medicine 2020; 383:2018–29. https://doi.org/10.1056/nejmoa2027187.

[12] Yoda S, Lin JJ, Lawrence MS, Burke BJ, Friboulet L, Langenbucher A, et al. Sequential ALK inhibitors can select for lorlatinib-resistant compound ALK mutations in ALK-positive lung cancer. Cancer Discov 2018; 8:714–29. https://doi.org/10.1158/2159-8290.CD-17-1256.

[13] Cui JJ, Tran-Dubé M, Shen H, Nambu M, Kung PP, Pairish M, et al. Structure based drug design of crizotinib (PF-02341066), a potent and selective dual inhibitor of mesenchymal-epithelial transition factor (c-MET) kinase and anaplastic lymphoma kinase (ALK). J Med Chem 2011; 54:6342–63. https://doi.org/10.1021/jm2007613.

[14] Bossi RT, Saccardo MB, Ardini E, Menichincheri M, Rusconi L, Magnaghi P, et al. Crystal structures of anaplastic lymphoma kinase in complex with ATP competitive inhibitors. Biochemistry 2010; 49:6813–25. https://doi.org/10.1021/bi1005514.

[15] Awad MM, Katayama R, McTigue M, Liu W, Deng Y-L, Brooun A, et al. Acquired Resistance to Crizotinib from a Mutation in CD74 – ROS1 . New England Journal of Medicine 2013; 368:2395–401. https://doi.org/10.1056/nejmoa1215530.

[16] Rogers D. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships 1994:854–66.

[17] Jawarkar RD, Sharma P, Jain N, Gandhi A, Mukerjee N, Al-Mutairi AA, et al. QSAR, Molecular Docking, MD Simulation and MMGBSA Calculations Approaches to Recognize Concealed Pharmacophoric Features Requisite for the Optimization of ALK Tyrosine Kinase Inhibitors as Anticancer Leads. Molecules 2022;27. https://doi.org/10.3390/molecules27154951.
[18] Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: Where have you been? Where are you going to? J Med Chem 2014; 57:4977–5010. https://doi.org/10.1021/jm4004285.
[19] Li Z, Li X, Liu X, Fu Z, Xiong Z, Wu X, et al. KinomeX: A web application for predicting kinome-wide polypharmacology effect of small molecules. Bioinformatics 2019; 35:5354–6. https://doi.org/10.1093/bioinformatics/btz519.
[20] Shahin R, Swellmeen L, Shaheen O, Aboalhaija N, Habash M. Identification of novel inhibitors for Pim-1 kinase using pharmacophore modeling based on a novel method for selecting pharmacophore generation subsets. J Comput Aided Mol Des 2016; 30:39–68. https://doi.org/10.1007/s10822-015-9887-7.
[21] Tropsha A, Golbraikh A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. vol. 13. 2007.
[22] Tu CH, Lin WH, Peng YH, Hsu T, Wu JS, Chang CY, et al. Pyrazolylamine Derivatives Reveal the Conformational Switching between Type i and Type II Binding Modes of Anaplastic Lymphoma Kinase (ALK). J Med Chem 2016; 59:3906–19. https://doi.org/10.1021/acs.jmedchem.6b00106.
[23] Schlütke L, Immer M, Preu L, Totzke F, Schächtele C, Kubbutat MHG, et al. Scaffold hopping identifies 6,8-disubstituted purines as novel anaplastic lymphoma kinase inhibitors. European Journal of Pharmaceutics and Biopharmaceutics 2018; 126:89–94. https://doi.org/10.1016/j.ejpb.2017.03.011.
[24] Milkiewicz KL, Weinberg LR, Albom MS, Angeles TS, Cheng M, Ghose AK, et al. Synthesis and structure-activity relationships of 1,2,3,4-tetrahydropyrido[2,3-b] pyrazines as potent and selective inhibitors of the anaplastic lymphoma kinase. Bioorg Med Chem 2010; 18:4351–62. https://doi.org/10.1016/j.bmc.2010.04.087.
[25] Mesaros EF, Thieu T V., Wells GJ, Zificsak CA, Wagner JC, Breslin HJ, et al. Strategies to mitigate the bioactivation of 2-anilino-7-aryl-pyrrolo [2,1-f][1,2,4]triazines: Identification of orally bioavailable, efficacious ALK inhibitors. J Med Chem 2012; 55:115–25. https://doi.org/10.1021/jm2010767.
[26] Li R, Xue L, Zhu T, Jiang Q, Cui X, Yan Z, et al. Design and synthesis of 5-aryl-pyridone-carboxamides as inhibitors of anaplastic lymphoma kinase. J Med Chem 2006; 49:1006–15. https://doi.org/10.1021/jm050824x.
[27] Lewis RT, Bode CM, Choquette DM, Potashman M, Romero K, Stellwagen JC, et al. The discovery and optimization of a novel class of potent, selective, and orally bioavailable anaplastic lymphoma kinase (ALK) inhibitors with potential utility for the treatment of cancer. J Med Chem 2012; 55:6523–40. https://doi.org/10.1021/jm3005866.
[28] Huang WS, Liu S, Zou D, Thomas M, Wang Y, Zhou T, et al. Discovery of Brigatinib (AP26113), a Phosphine Oxide-Containing, Potent, Orally Active Inhibitor of Anaplastic Lymphoma Kinase. J Med Chem 2016; 59:4948–64. https://doi.org/10.1021/acs.jmedchem.6b00306.
[29] Fushimi M, Fujimori I, Wakabayashi T, Hasui T, Kawakita Y, Imamura K, et al. Discovery of Potent, Selective, and Brain-Penetrant 1 H-Pyrazol-5-yl-1 H-pyrrolo[2,3- b]pyridines as Anaplastic Lymphoma Kinase (ALK) Inhibitors. J Med Chem 2019; 62:4915–35. https://doi.org/10.1021/acs.jmedchem.8b01630.
[30] Ott GR, Cheng M, Learn KS, Wagner J, Gingrich DE, Lisko JG, et al. Discovery of clinical candidate CEP-37440, a selective inhibitor of focal adhesion kinase (FAK) and anaplastic lymphoma kinase (ALK). J Med Chem 2016; 59:7478–96. https://doi.org/10.1021/acs.jmedchem.6b00487.

[31] Roy K, editor. Chemometrics and Cheminformatics in Aquatic Toxicology. Chemometrics and Cheminformatics in Aquatic Toxicology. First Edition, Hoboken, NJ: Wiley; 2021. https://doi.org/10.1002/9781119681397.fmatter.
[32] Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems 2015; 145:22–9. https://doi.org/10.1016/j.chemolab.2015.04.013.
[33] Truchon JF, Bayly CI. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. J Chem Inf Model 2007; 47:488–508. https://doi.org/10.1021/ci600426e.
[34] Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 2003; 111:1361–75. https://doi.org/10.1289/ehp.5758.
[35] Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006; 27:861–74. https://doi.org/10.1016/j.patrec.2005.10.010.
[36] Junge MRJ, Dettori JR. ROC Solid: Receiver Operator Characteristic (ROC) Curves as a Foundation for Better Diagnostic Tests. Global Spine J 2018; 8:424–9. https://doi.org/10.1177/2192568218778294.
[37] Swellmeen L, Shahin R, Al-Hiari Y, Alamiri A, Hasan A, Shaheen O. Structure based drug design of Pim-1 kinase followed by pharmacophore guided synthesis of Quinolone-based inhibitors. Bioorg Med Chem 2017; 25:4855–75. https://doi.org/10.1016/j.bmc.2017.07.036.
[38] Shahin R, AL-Hashimi NN, Daoud N el-H, Aljamal S, Shaheen O. QSAR-guided pharmacophoric modeling reveals important structural requirements for Polo kinase 1 (Plk1) inhibitors. J Mol Graph Model 2021; 109:108022. https://doi.org/10.1016/j.jmgm.2021.108022.
[39] Oecd. OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationships models. Biotechnology 2004:1–2.
[40] Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform 2010; 29:476–88. https://doi.org/10.1002/minf.201000061.
[41] Golbraikh A, Tropsha A. Beware of q^2! J Mol Graph Model 2002; 20:269–76. https://doi.org/10.1016/S1093-3263(01)00123-1.
[42] Gough NR, Kalodimos CG. Exploring the conformational landscape of protein kinases. Curr Opin Struct Biol 2024;88. https://doi.org/10.1016/j.sbi.2024.102890.
[43] Mansi IA, Al-Sha`er MA, Mhaidat NM, Taha MO, Shahin R. Investigation of binding characteristics of Phosphoinositide-dependent kinase-1 (PDK1) co-crystallized ligands through virtual pharmacophore modeling leading to novel anti – PDK1 hits. Med Chem (Los Angeles) 2019;15. https://doi.org/10.2174/1573406415666190724131048.
[44] Shahin R, Mansi I, Swellmeen L, Alwidyan T, Al-Hashimi N, Al-Qarar'h Y, et al. Ligand-based computer aided drug design reveals new tropomycin receptor kinase a (TrkA) inhibitors. J Mol Graph Model 2018; 80:327–52. https://doi.org/10.1016/j.jmgm.2018.01.004.
[45] Shahin R, Jaafreh S, Mansi I, Abu-Irmaileh B, Azzam Y, Aljamal S, et al. A comprehensive in-silico approach (pharmacophoric QSAR and docking-based comparative intermolecular analysis) led to identifying new PIM2 kinase inhibitors for treating resistant lymphomas. Lett Drug Des Discov 2025; 22:100021. https://doi.org/10.1016/j.lddd.2025.100021.
[46] Cerius 2. 2005.
[47] Biela A, Khayat M, Tan H, Kong J, Heine A, Hangauer D, et al. Impact of ligand and protein desolvation on ligand binding to the S1 pocket of thrombin. J Mol Biol 2012; 418:350–66. https://doi.org/10.1016/j.jmb.2012.01.054.
[48] Chan L, Morris GM, Hutchison GR. Understanding Conformational Entropy in Small Molecules. J Chem Theory Comput 2021; 17:2099–106. https://doi.org/10.1021/acs.jctc.0c01213.
[49] Ravichandran S, Luke BT, Collins JR. Can structural features of kinase receptors provide clues on selectivity and inhibition? A molecular modeling study. J Mol Graph Model 2015; 57:36–48. https://doi.org/10.1016/j.jmgm.2014.12.007.

[50] Whitty A, Zhong M, Viarengo L, Beglov D, Hall DR, Vajda S. Quantifying the chameleonic properties of macrocycles and other high-molecular-weight drugs. Drug Discov Today 2016; 21:712–7. https://doi.org/10.1016/j.drudis.2016.02.005.
[51] Epstein LF, Chen H, Emkey R, Whittington DA. The R1275Q neuroblastoma mutant and certain ATP-competitive inhibitors stabilize alternative activation loop conformations of anaplastic lymphoma kinase. Journal of Biological Chemistry 2012; 287:37447–57. https://doi.org/10.1074/jbc.M112.391425.
[52] Shahin R, Shaheen O, El-Dahiyat F, Habash M, Saffour S. Research advances in kinase enzymes and inhibitors for cardiovascular disease treatment. Future Sci OA 2017;4. https://doi.org/10.4155/fsoa-2017-0010.
[53] El-Jundi I, Daoud S, Taha MO. Discovery of novel chemotype inhibitors targeting Anaplastic Lymphoma Kinase receptor through ligand-based pharmacophore modelling. SAR QSAR Environ Res 2024; 35:795–815. https://doi.org/10.1080/1062936X.2024.2406398.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ALKSupplementaryMaterial.docx