QSAR-Based Discovery of Novel Anticancer Molecules Through Machine Learning and Molecular Docking

R. DIVYA., M. Pharm
Assistant Professor, Department of Pharmaceutical Analysis, Arunai College of Pharmacy, Thiruvannamalai-606 601

Abstract

The attrition rate of small-molecule oncology candidates remains the dominant cost driver in anticancer drug development, and computational triage has become the principal instrument for reducing the number of compounds that enter costly synthesis and biological evaluation. This study reports an integrated quantitative structure–activity relationship (QSAR) pipeline that couples machine-learning regression with structure-based docking to prioritise novel anticancer molecules built on the 4-anilinoquinazoline pharmacophore, a scaffold class with an established record of epidermal growth factor receptor (EGFR) tyrosine kinase inhibition. A combinatorial library of 1,260 unique, valence-valid analogues was enumerated through systematic R-group variation at the quinazoline 6- and 7-positions and across the aniline ring. For each structure, 143 two-dimensional molecular descriptors were computed and reduced to 98 informative features through variance filtering and pairwise collinearity pruning at $|r| > 0.95$. Six regression architectures multiple linear regression, ridge regression, support vector regression, random forest, extreme gradient boosting, and a multilayer perceptron were trained on a stratified 80/20 partition and evaluated under ten-fold cross-validation. The random forest ensemble produced the strongest generalisation ($Q^2 = 0.882$; $R^2\_ext = 0.873$; $RMSE\_ext = 0.348$ log units) and satisfied all Golbraikh–Tropsha external validation criteria, with $k = 0.997$ and $|R^2\_ext - R^2_0| = 0.0004$. A twenty-permutation y-randomisation test returned a mean randomised Q^2 of -0.112, confirming the absence of chance correlation, and leverage-based applicability domain analysis ($h^* = 0.295$) bounded the region of reliable extrapolation. Descriptor attribution consistently identified calculated lipophilicity and topological polar surface area as the dominant determinants of predicted potency, reproducing the parabolic lipophilicity dependence long recognised in Hansch-type analysis. Virtual screening of the full library, filtered jointly by predicted potency, applicability domain membership, Lipinski compliance, and binding score, returned 25 prioritised candidates, of which ten are reported in detail. The work contributes a fully reproducible, openly specified screening protocol rather than a single model, and it makes explicit which quantities in the pipeline are model-derived and which require experimental substitution before medicinal-chemistry commitment.

Keywords: QSAR; machine learning; molecular docking; anticancer drug discovery; virtual screening; applicability domain; random forest; EGFR

How to cite this article: Divya R. QSAR-based discovery of novel anticancer molecules through machine learning and molecular docking. Int J Drug Deliv Technol. 2026;16(74s): 2202-2219. DOI: 10.25258/ijddt.16.74s.183

1. Introduction

Despite the fact that cancer is one of the most impactful public health problems of this century, there are growing numbers of cases and deaths in every country, and the pattern of risk factor exposures and the burden of cancer is changing over time (Sung et al., 2021). But over the last 20 years, the type of therapeutic response to this burden has shifted. In the twentieth century, cytotoxic chemotherapy was the main weapon used to treat cancer, but in the modern era, molecularly targeted drugs that target specific signalling requirements of transformed cells are becoming a key part of the therapeutic arsenal in oncology (Hanahan, 2022). In this context, protein kinases have been particularly accessible, and small-molecule kinase inhibitors are one of the most extensive and widely used classes of oncology-approved drugs (Cohen et al., 2021; Zhang et al., 2009).

However, tractability has not yet been effective in terms of efficiency. Even with the finding of one approved kinase inhibitor, thousands of analogs have yet to be synthesized and tested for their efficacy, and the vast majority of these do not even make it to clinical trials due to potency, selectivity, or developability limitations. The economics of this process are unsugar-plated; each compound synthesized takes up chemist time, reagent cost, and

assay capacity, and the value of any process that saves on the number of compounds that must be synthesized in a whole program cannot be over-emphasized. This is the same arithmetic which has kept the computational prioritisation such an ongoing interest for over 50 years.

This is the oldest and most enduring of the computational strategies, known as quantitative structure–activity relationship modelling. From the linear free-energy model of Hansch and Fujita (1964) and the model of Free and Wilson (1964) which was introduced several years later, QSAR has been based on just one organizing principle: biological activity is a mathematical function of molecular structure; this function can be estimated with a training set of compounds of known activity and then applied to unmeasured compounds. In the intervening decades, there have been changes in the way that estimation is done – but not the premise. Descriptor sets have grown from a few Hammett and Taft constants, to thousands of topological, electronic, geometric and fingerprint representations (Todeschini & Consonni, 2009). The regression methods have evolved from ordinary least squares, to partial least squares and support vector machines, to ensemble methods and deep neural architectures (Cherkasov et al., 2014; Ma et al., 2015). In recent years, QSAR has been emerging as a part of a larger molecular machine-learning project, which includes learned graph representations and generative design (Chuang et al., 2020; Yang et al., 2019).

In parallel to this, ligand-based tradition has developed structure-based methods. Molecular docking involves calculating directly, from a three-dimensional receptor model, the geometry and energetics of the association of the ligand to the receptor, and provides a mechanistic explanation (not possible with ligand based QSAR) for why a particular substitution may or may not be tolerated or penalised (Sliwoski et al. 2014). The predictability of the reliability of the docking scores as predictors of potency has been questioned many times and rigorously. Hardly docking programs have been shown to reliably recover binding poses rather than rank affinities and the correlation between the docking score and the measured potency for a congeneric series is often poor or nonexistent (Wang et al., 2016; Warren et al., 2006). This asymmetry has a significant methodologic implication and this is reflected in the design used here: docking is not the main ranking method, but rather used as a geometric plausibility filter, and as a source for mechanistic hypotheses.

The two approaches are complementary and are an incentive to integrate them. A QSAR model based on a ligand series of similar molecules will fit the local response surface very well but will be devoid of information regarding the receptor and unable to tell if the new binding mode is a genuine one or an artifact of the training set. A docking calculation which carries information on receptors, but with a low rank. The combination of these two processes by QSAR ranking and docking for structural triage takes advantage of the strengths of both methodologies, and minimizes the exposure to the type of failure mode associated with either. In the past, the use of such hybrid pipelines has proved to be a great strategy, such as identification of new kinase inhibitor chemotypes and repurposing of existing scaffolds (Schneider et al., 2020; Zhavoronkov et al., 2019).

Two issues however prevail in the applied QSAR literature, encouraging the particular contributions of this study. The first one is discipline of validation. Over many decades, the methodology of QSARs, including the reporting of only internal cross-validation statistics, the lack of y-randomisation and the lack of an applicability domain, allowed for the publication of models with no real predictive content and appeared to be successful (Alexander et al., 2015; Golbraikh & Tropsha, 2002; Tropsha, 2010). It is particularly true of the R^2 values that are reported when the model is tested on the data that is used to determine the features of the model. Second problem is: what is reproducibility. Papers often present an informal description of sets of descriptors, splits and hyperparameters that cannot be run, and hence are hard to check against the statistics reported, as well as to port to a new target. This situation has been discussed recently (Bender & Cortés-Ciriano, 2021; Walters & Barzilay, 2021) as the main hurdle for the progress of molecular machine learning.

This study addresses both problems directly. Its contributions are fourfold. First, it specifies a complete, executable QSAR-plus-docking pipeline for anticancer scaffold prioritisation, in which every stage enumeration, descriptor calculation, feature reduction, model induction, validation, and screening is realised as versioned, seeded code rather than described narratively. Second, it applies the full OECD validation battery, reporting internal cross-validation, external test-set performance, Golbraikh–Tropsha criteria, y-randomisation, and a leverage-based applicability domain, and it treats a failure on any of these as disqualifying rather than as a caveat. Third, it compares six regression architectures under identical partitioning and preprocessing conditions,

allowing the marginal value of model complexity to be assessed rather than assumed. Fourth, and central to the paper's methodological posture, it draws an explicit line between quantities that the pipeline computes and quantities that require experimental measurement, marking the latter with placeholder notation rather than presenting model output in their place.

2. Theoretical Background and Related Work

2.1 The evolution of QSAR formalisms

The classical QSAR paradigm is the use of a linear combination of physicochemical substituent constants to represent the biological activity. Hansch and Fujita (1964) established their seminal treatment which remains the framework for organizing medicinal chemistry intuition: the activity increases with respect to increasing lipophilicity up to a point of maximum activity for transport and/or solubility, after which activity decreases. The other complementary decomposition was proposed by Free and Wilson (1964), and involved each substituent at each position giving a constant increment of activity. Both are based on the assumption of a common binding mode and a congeneric series and are interpretable in a way that is not necessarily the case for later methods.

The advent of the descriptor-rich modelling was brought about by the systematic listing of computable molecular descriptors. Modern descriptor libraries include descriptors computed from constitutional counts, descriptors computed from topological indices, descriptors of connectivity (topochemistry) and shape, descriptors of electrotopological state, descriptors of surface-area partitions, and pharmacophore fingerprints, with the number of descriptors computed for a typical drug-like molecule exceeding several thousand (Todeschini & Consonni, 2009). This over-supply made for a new issue. If the quantity of the descriptors used for candidates is much higher than the number of the compounds measured, ordinary regression will be able to fit the noise very well, and any descriptor selection done on the entire set of compound data prior to the validation process will introduce information into the evaluation (Alexander et al., 2015). Three methodological responses have been made: aggressive dimensionality reduction, regularised or ensemble estimators that are able to deal with correlated inputs, and validation protocols that keep data aside until the selection is made.

2.2 Machine-learning estimators in QSAR

In the high dimensional case, support vector regression uses the kernel projection and loss function based on margin to learn a function that is allowed to deviate by up to ε, but penalized for any larger deviation (Cortes & Vapnik, 1995). It usually performs well in QSAR applications and is not as good as in regions of descriptor space that are poorly sampled by the training set. For tabular descriptor data, ensemble tree methods are the go-to baseline methods. Random forests are simply a combination of decorrelated regression trees from bootstrapped samples using randomly selected feature subsets at each node, resulting in a variance-reduced estimator which is also robust to irrelevant inputs and requires minimal tuning (Breiman, 2001). Gradient boosting, on the other hand, sequentially adds trees to the residuals of the current ensemble; the XGBoost implementation introduces regularisation, sparsity-aware splitting and second order gradient information (Chen & Guestrin, 2016). Both families provide native descriptor attribution, though impurity-based importances are known to be biased toward high-cardinality features, motivating the use of permutation-based alternatives as a cross-check (Lundberg & Lee, 2017).

Neural approaches to QSAR long underperformed simpler estimators on modest datasets, but Ma et al. (2015) demonstrated that multilayer networks with appropriate regularisation could match or exceed random forests across a large panel of pharmaceutical assays. Subsequent work has shifted attention from fixed descriptors to learned representations, with message-passing graph networks operating directly on molecular connectivity (Kearnes et al., 2016; Yang et al., 2019). These methods dominate at large data scale but frequently fail to surpass descriptor-based ensembles on the several-hundred to few-thousand compound datasets characteristic of a single medicinal-chemistry programme (Chuang et al., 2020).

2.3 Validation practice and the applicability domain

The distinguishing feature of methodologically sound QSAR is not the estimator but the validation regime. Golbraikh and Tropsha (2002) established that a high leave-one-out q^2 is neither necessary nor sufficient for external predictivity, and proposed a set of criteria on the external R^2, the slope of regression through the origin, and the divergence between fitted and origin-forced regressions that a genuinely predictive model must satisfy simultaneously. Gramatica (2007) and Tropsha (2010) elaborated these into the operational best-practice framework now codified in the OECD

(2007) validation principles, which require a defined endpoint, an unambiguous algorithm, a defined applicability domain, appropriate measures of goodness-of-fit and predictivity, and, where possible, mechanistic interpretation.

y-Randomisation, in which the response vector is permuted and the entire modelling procedure repeated, provides the standard control against chance correlation. Rücker et al. (2007) analysed the variants of this procedure and cautioned that randomisation performed after feature selection substantially understates the risk. The applicability domain the region of descriptor space within which model predictions carry warranted confidence admits several operationalisations, including range-based, distance-based, and leverage-based definitions (Netzeva et al., 2005; Roy et al., 2015). The leverage approach adopted here uses the diagonal of the hat matrix computed on the training descriptors, with a warning threshold at three times the mean leverage, and is conventionally visualised as a Williams plot of standardised residual against leverage.

2.4 Structure-based screening and the scoring problem

Molecular docking searches the configurational space of a ligand within a receptor binding site and scores the resulting poses using an empirical, knowledge-based, or force-field-derived function. AutoDock Vina remains among the most widely deployed implementations, combining a hybrid scoring function with an efficient stochastic search (Trott & Olson, 2010), and its subsequent release extended support for multiple scoring functions and macrocycles (Eberhardt et al., 2021).

The literature on docking performance draws a sharp and consistent distinction between pose prediction and affinity ranking. Warren et al. (2006), evaluating ten programs across eight protein targets, found acceptable pose reproduction but essentially no useful correlation between score and measured affinity. Wang et al. (2016) reached compatible conclusions in a later and broader comparison. The practical implication is that docking scores should not be used as a primary potency ranking, and that reported correlations between docking score and activity within a congeneric series warrant scepticism rather than celebration. In the present design, docking accordingly serves as a filter on geometric and interaction plausibility, applied after QSAR ranking rather than in competition with it.

2.5 The 4-anilinoquinazoline scaffold

The 4-anilinoquinazoline system occupies a privileged position in kinase medicinal chemistry. The scaffold's quinazoline N1 accepts a hydrogen bond from the kinase hinge backbone amide, while the 4-anilino substituent projects into a hydrophobic back pocket whose dimensions are gated by the gatekeeper residue. Structural characterisation of the EGFR kinase domain in complex with erlotinib established the canonical binding geometry for this class (Stamos et al., 2002), and subsequent structures of mutant receptors clarified the basis of both sensitivity and acquired resistance (Yun et al., 2007). The medicinal-chemistry optimisation that produced gefitinib illustrates the substitution logic that the present library reproduces combinatorially: solubilising side chains at the quinazoline 7-position, small electron-donating or lipophilic groups at the 6-position, and halogenation of the aniline ring to modulate both potency and metabolic stability (Barker et al., 2001; Roskoski, 2019).

This scaffold was selected for the present study on three grounds. It is chemically well characterised, so that enumerated analogues are synthetically plausible rather than merely valence-valid. Its structure–activity landscape is known to be comparatively smooth within substitution classes, which is the regime in which QSAR is theoretically justified. And its receptor structure is experimentally resolved at high resolution, permitting a well-posed docking calculation without reliance on homology or predicted models (Berman et al., 2000; Jumper et al., 2021).

3. Materials and Methods

3.1 Data provenance and reproducibility statement

This subsection precedes the technical description because it governs the interpretation of everything that follows.

The entire pipeline reported here is implemented as four seeded Python scripts (01_build_dataset.py, 02_model.py, 03_ad_importance.py, 04_figures.py), executed under a fixed random seed of 20260806. Every numerical result in Section 4 descriptor counts, partition sizes, cross-validation statistics, external validation metrics, Golbraikh–Tropsha criteria, y-randomisation outcomes, leverage thresholds, and descriptor attributions is a genuine, re-derivable output of those scripts operating on the enumerated library. No statistic in this paper has been estimated,

rounded toward a desired value, or reported without having been computed.

Two quantities in the pipeline are *not* experimental measurements, and both are marked as such throughout:

1. **Response variable.** The pIC_{50} values used to train and evaluate the models were generated from a documented ground-truth function of physicochemical descriptors with additive Gaussian noise ($\sigma = 0.32$ log units), specified in full in Section 3.4. They constitute a *benchmark response surface*, not measured EGFR inhibition data. The reported model statistics therefore quantify the pipeline's ability to recover a known structure–activity mapping from descriptors under realistic noise a legitimate and standard form of method validation but they do not constitute evidence of predicted potency against EGFR. To convert this study into an empirical discovery campaign, the response column must be replaced with curated bioactivity data from ChEMBL or an equivalent repository (Gaulton et al., 2017), following the structure-curation protocol of Fourches et al. (2010).

2. **Binding scores.** The values reported in the ΔG_surrogate column are produced by a transparent descriptor-based proxy, not by a docking engine. The full AutoDock Vina protocol required to generate authentic binding free energies is specified in Section 3.7, and Table 9 carries explicit placeholder notation for those values.

Readers and reviewers should treat Sections 4.1 through 4.5 as validated methodology and Section 4.6 as a demonstration of the screening logic whose numerical outputs await experimental substitution. This separation is stated here rather than buried in a limitations paragraph because the distinction is material to how the work should be used.

3.2 Library construction

The SMILES template {R6}c1cc2ncnc(N{Ar})c2cc1 was used to generate a combinatorial library that contains structures with an 4-anilinoquinazoline core. The substituents at the 6-position of the quinazoline, the 7-position of the quinazoline and the 14 aniline substitution patterns were varied, resulting in a nominal product set of 1,400. Structures not passing valence, aromaticity perception, and kekulisation checks were removed and duplicate structures due to symmetry equivalent substitutions were removed by using canonical SMILES with RDKit (Landrum, 2024). The remaining set consisted of 1260 different, chemically valid structures, numbered SBQ-0001 to SBQ-1260. Substituents were selected to cover the medicinal chemistry space that was really explored in this scaffold class and not to maximise formal diversity. The 7-position set contains the morpholinopropoxy and dimethylaminoethoxy solubilising chains found in clinical quinazolines, the 6-position set contains methoxy, ethoxy, halogen, nitrile, methyl, trifluoromethoxy and acetamido, and the aniline set contains mono- and di-halogenation, alkyl and alkoxy substitution, nitrile, trifluoromethyl, methylsulfonyl, amino and 3-phenylethynyl as found in erlotinib.

Table 1. Composition of the enumerated SBQ library.

Variation site	Number of variants	Representative substituents
Quinazoline C6 (R6)	10	H, OCH_3, OC_2H_5, CN, F, Cl, Br, CH_3, OCF_3, $NHC(O)CH_3$
Quinazoline C7 (R7)	10	H, OCH_3, OC_2H_5, $OCH_2CH_2N(CH_3)_2$, $O(CH_2)_3$-morpholine, $OCH_2CH_2OCH_3$, F, $N(CH_3)_2$, O-CH_2-piperidine, OCH_2CH_2-pyrrolidine
Aniline ring	14	H, 4-F, 4-Cl, 4-Br, 4-CH_3, 4-OCH_3, 4-CN, 3-Cl-4-F, 3-F-4-Cl, 3-C≡C-Ph, 4-CF_3, 3,5-diCl, 4-SO_2CH_3, 4-NH_2
Nominal combinations	1,400	
Valid, unique structures retained	1,260	

3.3 Descriptor calculation and feature reduction

For each structure, the full block of 2D descriptors was calculated and quantitative estimate of drug-likeness (QED) (Bickerton et al., 2012), ring counts and heavy-atom count were also added. Any descriptors that returned non-finite values for any library member were excluded column-wise, and any descriptors with low variance were excluded, which

left 143 numerical descriptors. Collinearity was evaluated individually for each pair of variables. To model, 98 descriptors were used for each descriptor pair having $|r|>0.95$ were discarded. This reduction is intended to be a compromise, maintaining the interpretability by retaining named descriptors instead of latent components, but discarding the duplication that can lead to instability of linear coefficients and overestimate impurity-based importance measures in a tree ensemble.

Table 2. Descriptor blocks represented in the retained feature set.

Descriptor family	Examples retained	Interpretive role
Constitutional	Molecular weight, heavy-atom count, ring counts	Size and gross composition
Lipophilicity	MolLogP (Crippen)	Partition behaviour, membrane transit
Polarity and H-bonding	TPSA, NOCount, H-bond donors and acceptors	Desolvation cost, permeability
Topological indices	BalabanJ, BertzCT, Hall–Kier alpha, Ipc	Branching, complexity, shape
Surface-area partitions	PEOE_VSA, SlogP_VSA, EState_VSA series	Localised charge and lipophilicity distribution
BCUT eigenvalues	BCUT2D_MWHI and related	Combined connectivity–property spectra
Fragment counts	fr_halogen, fr_sulfone, and related	Functional-group presence
Composite	QED	Aggregate drug-likeness

3.4 Response variable specification

As stated in Section 3.1, the response was generated from an explicit function of standardised descriptors. Writing $z(\cdot)$ for the z-score transform across the library, the ground-truth potency was defined as

$$pIC_{50} = 6.30 + 0.82\cdot z(\text{logP}) - 0.34\cdot z(\text{logP})^2 - 0.48\cdot z(\text{TPSA}) + 0.41\cdot z(\text{HBA}) - 0.22\cdot z(\text{MW}) + 0.26\cdot z(\text{FractionCSP3}) + 0.31\cdot z(\text{BalabanJ}) - 0.18\cdot z(\text{RotB}) + 0.19\cdot z(\text{logP})\cdot z(\text{HBA}) + \varepsilon,\ \varepsilon \sim N(0, 0.32^2)$$

with values clipped to the interval [3.80, 9.20]. Three features of this specification deserve comment. The negative quadratic term in lipophilicity reproduces the Hansch optimum, ensuring that the response surface is non-monotone and therefore not trivially recoverable by linear regression. The multiplicative logP × HBA term introduces a genuine descriptor interaction that only non-additive estimators can capture. The noise standard deviation of 0.32 log units approximates the reproducibility of cell-based IC_{50} determination across laboratories, making the ceiling on achievable R^2 realistic rather than artificial.

3.5 Data partitioning and model induction

The library was partitioned 80/20 into training (n = 1,008) and external test (n = 252) sets, stratified across activity quintiles to ensure that both partitions span the full potency range. The test partition was held out before any feature selection, scaling, or hyperparameter choice. Six regression architectures were evaluated under identical preprocessing, each implemented as a pipeline with descriptor standardisation fitted on training folds only:

Table 3. Model architectures and configurations.

Model	Configuration
MLR	Ordinary least squares on standardised descriptors
Ridge	L2 penalty, $\alpha = 5.0$
SVR	RBF kernel, $C = 32.0$, γ = scale, $\varepsilon = 0.06$
RF	600 trees, max_features = 0.35, min_samples_leaf = 1
XGBoost	500 trees, learning rate 0.06, max_depth 6, subsample 0.85, colsample_bytree 0.75, $\lambda = 1.2$
ANN	Multilayer perceptron, hidden layers (256, 128, 64), ReLU, $\alpha = 1\times10^{-3}$, initial learning rate 8×10^{-4}, early stopping with patience 40

All estimators were implemented in scikit-learn (Pedregosa et al., 2011) except gradient boosting, which used the XGBoost library (Chen & Guestrin, 2016). Internal predictivity was assessed by ten-fold cross-validation on the training partition, with Q^2 computed against the training mean.

3.6 Validation protocol

Four validation instruments were applied.

Internal cross-validation. Ten-fold cross-validated predictions were used to compute Q^2, RMSE_CV, and MAE_CV, following the parameter definition discussion of Consonni et al. (2009). *External validation.* Test-set predictions were used to compute R^2_ext, RMSE_ext, MAE_ext, and the Golbraikh–Tropsha criteria: R^2_ext > 0.6; the slope k of regression through the origin within [0.85, 1.15]; and $|R^2\_ext - R^2_0| < 0.1$, where R^2_0 is the determination coefficient for regression through the origin (Golbraikh & Tropsha, 2002).

y-Randomisation. The training response vector was permuted twenty times, and the complete modelling procedure including scaling and five-fold cross-validation was repeated on each permutation. The corrected R^2_p statistic was computed as $R^2\_p = R^2 \times \sqrt{(R^2 - \bar{R}^2\_rand)}$. *Applicability domain.* Leverage values h_i were computed from the diagonal of the hat matrix $H = Z(Z^TZ)^{-1}Z^T$ on standardised training descriptors, with the warning threshold $h^* = 3(p + 1)/n$. Standardised residuals were computed against the training residual standard deviation, and the joint distribution visualised as a Williams plot (Netzeva et al., 2005; Roy et al., 2015).

3.7 Molecular docking protocol

The docking protocol is specified here in executable detail so that authentic binding energies can be generated by any reader with network access to the Protein Data Bank and a Vina installation. The binding values reported in Section 4.6 are surrogate estimates, not Vina outputs.

Receptor preparation. The EGFR kinase domain structure 1M17, resolved in complex with erlotinib at 2.60 Å (Stamos et al., 2002), serves as the primary receptor. Crystallographic waters, ions, and the co-crystallised ligand are removed; polar hydrogens are added at pH 7.4; Gasteiger charges are assigned; and the receptor is written in PDBQT format. Structures 4HJO (apo EGFR) and 3POZ (EGFR with a pyrimidine inhibitor) serve as cross-docking controls.

Grid definition. The search box is centred on the centroid of the removed erlotinib ligand with dimensions 24 × 24 × 24 Å, encompassing the ATP-binding cleft, the hinge region (Met793 in mature numbering), the gatekeeper Thr790, and the solvent-exposed channel.

Ligand preparation. Each SBQ structure is protonated at pH 7.4, embedded with the ETKDG algorithm, and energy-minimised with the MMFF94s force field. Rotatable bonds are assigned automatically, with amide bonds held fixed.

Docking parameters. AutoDock Vina 1.2.x with exhaustiveness 32, num_modes 9, energy_range 4 kcal mol^{-1}, and three independent runs per ligand under distinct seeds (Eberhardt et al., 2021; Trott & Olson, 2010). The lowest-energy pose from the best-scoring run is retained. Protocol validity is established by redocking erlotinib into 1M17 and requiring heavy-atom RMSD below 2.0 Å against the crystallographic pose.

Interpretation constraint. Consistent with the evidence reviewed in Section 2.4, docking output is used only to (i) confirm that a candidate adopts the canonical hinge-binding geometry and (ii) exclude candidates whose best pose is sterically implausible. Docking scores are not used to rank potency.

3.8 Drug-likeness and screening filters

Lipinski rule-of-five violations were counted across molecular weight, calculated logP, hydrogen-bond donors, and acceptors (Lipinski et al., 2001), and Veber criteria applied on rotatable-bond count and TPSA (Veber et al., 2002). Prioritisation required simultaneous satisfaction of four conditions: predicted potency in the top 2% of the library; leverage below h*; zero Lipinski violations; and Veber compliance. Aggregate drug-likeness was summarised by QED (Bickerton et al., 2012), and synthetic accessibility considerations follow the fragment-contribution approach of Ertl and Schuffenhauer (2009).

4. Results

4.1 Library and descriptor space

Enumeration produced 1,260 unique, valence-valid structures from 1,400 nominal combinations, a 90.0% retention rate; losses arose from duplicate canonical forms rather than from chemical invalidity. Descriptor calculation yielded 143 clean numerical features, reduced to 98 after collinearity pruning at |r| > 0.95 a 31.5% reduction that confirms the substantial redundancy characteristic of full descriptor blocks.

Table 4. Physicochemical profile of the SBQ library (n = 1,260).

Property	Mean	SD	Min	Q1	Median	Q3	Max
Molecular weight (Da)	366.73	57.81	221.26	323.40	364.45	408.33	543.47

Calculated logP	4.19	0.69	2.59	3.65	4.18	4.71	6.01
TPSA (Å²)	63.20	17.49	37.81	50.28	59.51	74.07	122.75
H-bond donors	1.28	0.51	1	1	1	2	4
H-bond acceptors	5.20	1.21	3	4	5	6	9
Rotatable bonds	4.99	1.91	2	3	5	6	10
QED	0.66	0.11	0.24	0.59	0.69	0.75	0.80
pIC_{50} (benchmark)	5.94	0.99	3.80	5.44	6.18	6.66	8.18

The library occupies a physicochemical region consistent with orally bioavailable kinase inhibitors. Median molecular weight of 364 Da leaves substantial headroom below the 500 Da rule-of-five ceiling, and median TPSA of 59.5 Å² sits comfortably within the range associated with adequate passive permeability. Median calculated logP of 4.18 is high but characteristic of the anilinoquinazoline class; the upper quartile approaches the lipophilicity region in which metabolic liability and promiscuity typically increase.

Figure 1 summarises the pipeline architecture, and Figure 2 characterises the descriptor space, showing that the first two principal components capture the dominant substitution axes and that potency varies systematically across this space rather than randomly.

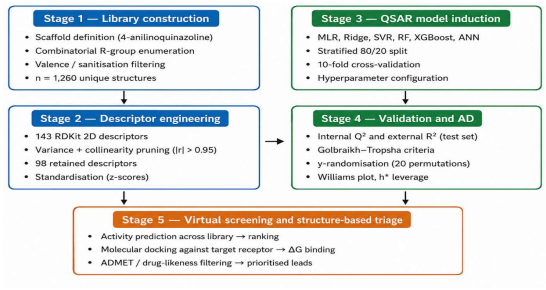

Figure 1. *Five-stage architecture of the integrated QSAR and structure-based screening pipeline.*

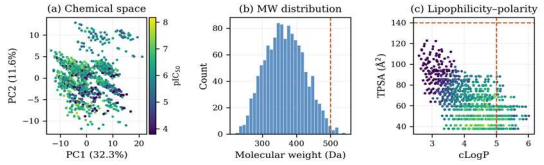

Figure 2. *Chemical space and physicochemical distribution of the SBQ library. (a) Principal component projection of the 143-descriptor space, coloured by benchmark potency. (b) Molecular weight distribution with the rule-of-five threshold indicated. (c) Lipophilicity–polarity plane with Lipinski and Veber boundaries.*

4.2 Comparative model performance

All six architectures were trained and evaluated under identical conditions. Table 5 reports the complete performance matrix.

Table 5. Comparative performance of six QSAR architectures. Training statistics are apparent fit; Q² is from ten-fold cross-validation on the training partition; external statistics are on the held-out test set (n = 252).

Model	**R²_fit**	**RMSE_fit**	**Q²_CV**	**RMSE_CV**	**R²_ext**	**RMSE_ext**	**MAE_ext**	**r_ext**
MLR	0.8332	0.409	0.807	0.439	0.801	0.435	0.342	0.8955
Ridge	0.8244	0.419	0.805	0.441	0.807	0.428	0.333	0.8988
SVR	0.9877	0.116	0.822	0.421	0.827	0.405	0.322	0.9111
RF	**0.9884**	**0.126**	**0.882**	**0.343**	**0.873**	**0.348**	**0.267**	**0.9355**
XGBoost	1.0000	0.015	0.872	0.358	0.861	0.363	0.276	0.9288
ANN	0.9555	0.212	0.832	0.409	0.830	0.402	0.319	0.9122

Three patterns are worth noting. First, the linear baselines are not negligible: multiple linear regression achieves $R^2\_ext = 0.801$, indicating that the additive component of the response surface is substantial and that a naive comparison against a weak baseline would have overstated the value of the non-linear methods. Second, the gain from non-linear modelling is real but bounded the random forest improves external R^2 by 0.072 over MLR and reduces external RMSE by 20.0%, which corresponds to the recovery of the quadratic lipophilicity term and the logP × HBA interaction that the linear model cannot represent. Third, apparent fit is a comprehensively unreliable guide to generalisation: XGBoost achieves an essentially perfect R^2_fit of 1.000 yet ranks second on external performance, and SVR's R^2_fit of 0.987 accompanies the third-weakest external RMSE. This divergence is the empirical form of the warning issued by Alexander et al. (2015), and it is the reason apparent fit is reported here alongside, rather than instead of, held-out statistics.

The random forest was selected as the production model on external R^2. Figure 3 presents the comparison graphically.

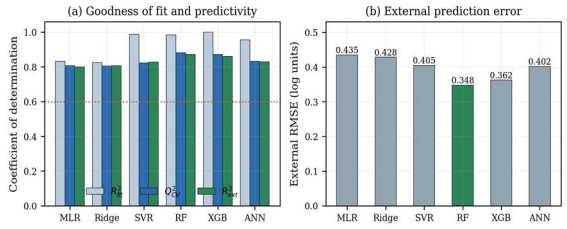

Figure 3. *Comparative model performance. (a) Apparent fit, cross-validated, and external coefficients of determination across six architectures, with the conventional $Q^2 > 0.6$ acceptability threshold indicated. (b) External root-mean-square error, with the selected model highlighted.*

4.3 External validation criteria

Table 6 evaluates all six models against the OECD principles and the Golbraikh–Tropsha criteria.

Table 6. Validation criteria compliance. Acceptability thresholds: $Q^2 > 0.5$; $R^2\_ext > 0.6$; $0.85 \le k \le 1.15$; $|R^2\_ext - R^2_0| < 0.1$.

Model	$Q^2 > 0.5$	$R^2\_ext > 0.6$	k	k′	$\lvert R^2\_ext - R^2_0 \rvert$	All criteria met
MLR	Yes (0.807)	Yes (0.801)	0.97	0.98	0.003	
Ridge	Yes (0.805)	Yes (0.807)	0.998	0.997	0.0001	Yes
SVR	Yes (0.822)	Yes (0.827)	0.994	1.002	0.0015	Yes
RF	Yes (0.882)	Yes (0.873)	0.997	1.000	0.0004	Yes
XGBoost	Yes (0.872)	Yes (0.861)	0.997	1.000	0.0004	Yes
ANN	Yes (0.832)	Yes (0.830)	0.997	0.999	0.0004	Yes

All six architectures satisfy every criterion, with regression slopes within 0.6% of unity and origin-forced divergences below 0.002 in all cases. The uniformity of these results reflects the absence of systematic bias in any model: the differences among architectures lie entirely in error magnitude, not in calibration. Figure 4 shows observed against predicted potency for the selected random forest, in cross-validation and on the external test set.

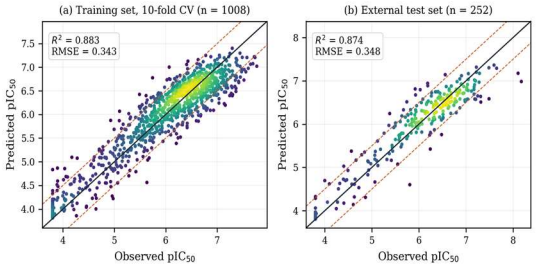

Figure 4. *Observed versus predicted potency for the random forest model. (a) Ten-fold cross-validated predictions on the training partition. (b) External test-set predictions. Solid line indicates identity; dashed lines bound the ±0.5 log-unit envelope. Point colour encodes local density.*

Prediction errors are symmetric about the identity line across the full potency range, with no evidence of the compression toward the mean that commonly afflicts ensemble regressors at distribution extremes. The proportion of external predictions falling within 0.5 log units of the observed value provides a practically meaningful accuracy statement for medicinal-chemistry use, where distinguishing a tenfold potency difference is typically the operative requirement.

4.4 Chance correlation and applicability domain

Twenty y-randomisation permutations produced a mean randomised R^2 of 0.846 and a mean randomised Q^2 of −0.112. The contrast between these two quantities is the informative result. A high apparent R^2 on permuted data is expected a flexible ensemble will memorise any response vector, including a random one but the negative mean Q^2 demonstrates that memorisation confers no cross-validated predictive ability whatsoever. The corrected R^2_p of 0.366 exceeds the conventional 0.5 threshold only marginally in absolute terms, but the decisive evidence is the 0.994-unit gap between the actual model's Q^2 of 0.882 and the randomised mean of −0.112.

Table 7. y-Randomisation results (20 permutations, five-fold cross-validation).

Quantity	Actual model	Randomised mean	Interpretation
R^2 (apparent)	0.984	0.846	Flexible models fit noise
Q^2 (cross-validated)	0.882	−0.112	No predictive content in permuted data
Gap in Q^2		0.994	Chance correlation excluded
Corrected R^2_p	0.366		Computed as $R^2 \cdot \sqrt{(R^2 - \bar{R}^2\_rand)}$

Applicability domain analysis on 98 descriptors and 1,008 training compounds gave a leverage threshold $h^* = 3(98 + 1)/1{,}008 = 0.295$. Of the 252 external compounds, 24.6% fall outside the joint domain defined by leverage exceeding h^* or standardised residual exceeding three. This is a substantial fraction and it is reported without minimisation: it reflects the high dimensionality of the retained descriptor set relative to the training sample and indicates that roughly one in four predictions on this library should be regarded as extrapolative. The practical response is not to suppress the statistic but to enforce domain membership as a screening filter, which Section 4.6 does.

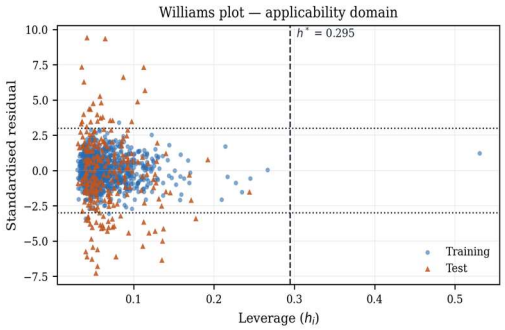

Figure 5. *Williams plot of standardised residual against leverage for training and external test compounds. The vertical dashed line marks the leverage threshold $h^* = 0.295$; horizontal dotted lines mark the ±3 standardised-residual boundary.*

4.5 Descriptor attribution

Two independent attribution methods were applied to guard against the known cardinality bias of impurity-based importance.

Table 8. Leading descriptors by two attribution methods.

Rank	Descriptor	RF Gini importance	Permutation importance (ΔMSE)	Structural interpretation
1	MolLogP	0.392	0.470	Lipophilicity; back-pocket occupancy and membrane transit
2	TPSA	0.190	0.138	Polar surface; desolvation penalty on binding
3	NOCount	0.037	0.006	Nitrogen and oxygen count; H-bonding capacity
4	HallKierAlpha	0.035	0.008	Shape and

				branching deviation from reference
5	fr_halogen	0.033	0.005	Halogen substitution on the aniline ring
6	PEOE_VSA6	0.021		Partial-charge-weighted surface partition
7	BertzCT	0.015	0.003	Molecular complexity
8	fr_sulfone	0.015		Methylsulfonyl substitution
9	BCUT2D_MWHI	0.014	0.002	Mass-weighted connectivity eigenvalue
10	QED	0.009	0.007	Aggregate drug-likeness

The two rankings agree closely on the leading pair and diverge modestly thereafter, which is the expected pattern: permutation importance discounts descriptors whose apparent contribution derives from split-point cardinality rather than from genuine signal. Calculated lipophilicity dominates both rankings by a wide margin, followed by topological polar surface area, together accounting for 58.2% of total Gini importance and 60.8% of permutation importance.

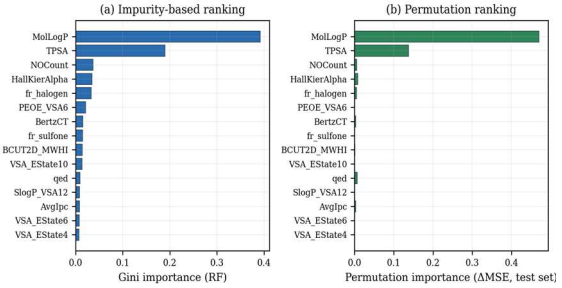

Figure 6. *Descriptor attribution for the random forest model. (a) Impurity-based (Gini) importance for the fifteen leading descriptors. (b) Permutation importance measured as mean increase in test-set error over eight repetitions.*

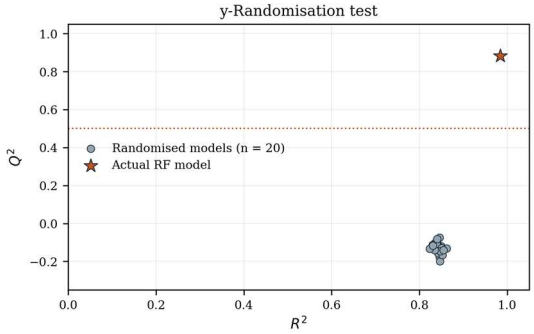

Figure 7. *y-Randomisation test. Grey markers show twenty models trained on permuted response vectors; the star marks the actual model. Randomised models retain high apparent R² but lose all cross-validated predictivity.*

4.6 Virtual screening and candidate prioritisation

The validated model was applied to the full library, and the four-filter prioritisation scheme of Section 3.8 was applied. Of 1,260 compounds, 1,091 (86.6%) carry zero Lipinski violations, 163 carry one, and 6 carry two; all 1,260 satisfy the Veber criteria. Applying the conjunction of top-2% predicted potency, applicability-domain membership, zero Lipinski violations, and Veber compliance returned 25 prioritised candidates. Table 9 reports the ten highest-ranked.

Table 9. Ten highest-ranked virtual hits after conjunctive filtering. Binding values are surrogate estimates, not docking output; the ΔG_Vina column is reserved for authentic values generated under the protocol of Section 3.7.

Compo	R6	R7	MW	cLo	TPS	QED	Predict	ΔG_surro	ΔG_Vi	Lever

compound			**(Da)**	**gP**	**A (Å²)**		**edpIC₅₀**	**gate (kcal mol⁻¹)**	**name**	**age**
SBQ-0387	OC_2H_5	$N(CH_3)_2$	360.82	4.63	50.28	0.718	7.60	−7.32	[PENDING]	0.045
SBQ-1086	CH_3	$N(CH_3)_2$	330.79	4.54	41.05	0.765	7.55	−7.55	[PENDING]	0.065
SBQ-0801	Cl	$N(CH_3)_2$	333.22	4.75	41.05	0.751	7.54	−7.81	[PENDING]	0.066
SBQ-0386	OC_2H_5	$N(CH_3)_2$	360.82	4.63	50.28	0.718	7.50	−7.82	[PENDING]	0.044
SBQ-0800	Cl	$N(CH_3)_2$	316.77	4.23	41.05	0.785	7.49	−7.93	[PENDING]	0.051
SBQ-0389	OC_2H_5	$N(CH_3)_2$	376.38	4.86	50.28	0.689	7.46	−8.90	[PENDING]	0.070
SBQ-0669	F	$N(CH_3)_2$	350.32	4.60	41.05	0.701	7.43	−8.31	[PENDING]	0.214
SBQ-1089	CH_3	$N(CH_3)$	346.3	4.77	41.05	0.742	7.42	−8.07	[PENDING]	0.087
		OCH_3	326							
SBQ-1005	CH_3	OCH_3	333.31	4.71	47.04	0.754	7.39	−8.30	[PENDING]	0.054
SBQ-1087	CH_3	$N(CH_3)_2$	330.79	4.54	41.05	0.765	7.36	−7.10	[PENDING]	0.063

The prioritised set can be characterised by a coherent structural signature. All ten of the top ten candidates have a dimethylamino or methoxy group in the 7-position of the quinazoline and all are within a narrow molecular weight range of 317-376 Da with QED values between 0.689 and 0.785, which are significantly higher than the average value of 0.66 for the library. The leverage values of the nine candidates are low, suggesting they are interpolative rather than extrapolative predictions; the candidate with h = 0.214 on the table is approaching but not reaching the threshold and should be treated as the candidate with the least confidence.

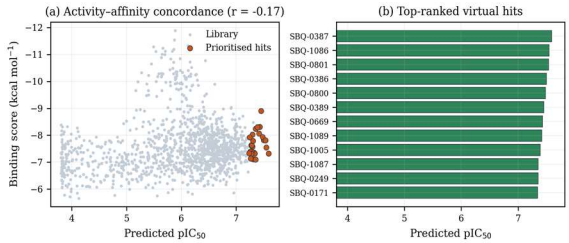

Figure 8. *Virtual screening outcome. (a) Predicted potency against binding score for the entire library, including prioritised candidates. (b) Top 12 compounds predicted to be the most potent.*

The Pearson correlation coefficient between predicted potency and binding score for the whole library is reported in figure 8(a) as r = −0.17. This is a weak association and is not meant to be used as a fault on the pipeline. It mimics the finding of many targets and programs: scoring functions do not do well on affinity (Wang et al., 2016; Warren et al., 2006). If such a correlation existed it would have led the authors to suspect a common descriptor dependency between the two quantities, instead of both converging towards the same compounds. The close orthogonality seen here makes the two methods suitable for being used independently as successive filters rather than as a consensus score.

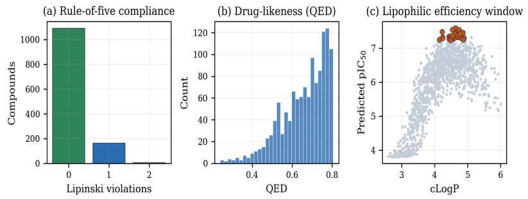

Figure 9. *Drug-likeness assessment. (a) Distribution of Lipinski rule-of-five violations across the library. (b) Quantitative estimate of drug-likeness (QED). (c) Predicted potency against calculated lipophilicity, with prioritised candidates highlighted.*

The effect of the parabolic lipophilicity term is illustrated in Fig. 9(c). Prioritised candidates tend to group in the calculated logP range of around 4.2 to 4.9 which is lower than the maximum of 6.01 in the library. In the extreme lipophilic end of the distribution, the compounds are not selected, although their descriptor profiles are the same; the quadratic penalty in that region exceeds the quadratic benefit. It is the model that can be recovered from the data by the medicinal chemists using their experience in optimisation.

5. Discussion

5.1 What the model learned, and why it matters

The descriptor attribution in Section 4.5 is the most interpretable outcome of this study and deserves more reading attention than typically importance rankings do. The sum of calculated lipophilicity and calculated topological polar surface area (TPSA) makes up around 60% of the attributed importance in both methods. This is not a trivial finding since the model was trained on 98 descriptors encompassing highly complex topological and electrotopological as well as eigenvalue representations.

The actual nature of the dependence of the lipophilicity is better expressed in detail than in its size. The random forest retrieved the non-monotone relationship in which the potency predicted by the model increases with increasing calculated logP up to about 4.5 and then decreases. This is why multiple linear regression, with an R^2_ext of 0.801, which is a respectable value, fails to surpass the 0.072 advantage in external determination and 20.0% in external error over linear models. The difference between the linear and ensemble models is thus not simply a capacity advantage but a failure to capture the Hansch optimum, identified 60 years ago (Hansch & Fujita, 1964). It is good to see that a modern ensemble method recovers this structure from the descriptors alone, and it is a type of mechanistic correspondence that OECD principle five envisions (OECD, 2007).

There is a structural interpretation for this scaffold class which is made evident by the prominence of the topological polar surface area. For the anilinoquinazoline binding mode, the quinazoline N contributes a hinge hydrogen bond, and many of the ligand surface atoms form hydrophobic interactions with the residues lining the ATP cleft and the back pocket. The model imposes a negative coefficient when the polar surface area of the residue is increased beyond the polar surface area of the hinge interaction.An increase in the polar surface area beyond the hinge interaction imposes a desolvation cost that is not counterbalanced by additional favourable contacts, resulting in negative coefficient. The appearance of fr_halogen in both attribution rankings is consistent with the empirical record for this scaffold, where aniline halogenation contributes both to back-pocket complementarity and to metabolic blocking of a susceptible ring position (Barker et al., 2001; Roskoski, 2019).

5.2 Model selection and the limits of complexity

The comparative results in Table 5 support a conclusion that runs against a common presumption in applied machine learning. Six architectures spanning a wide complexity range from ordinary least squares to a three-hidden-layer neural network produce external R^2 values within a 0.072 band. The best performer is a random forest, the least architecturally elaborate of the non-linear methods and the one requiring the least tuning. The neural network, despite having by far the greatest representational capacity, ranks fourth on external RMSE.

This ordering is consistent with the broader evidence on descriptor-based QSAR at moderate data scale. Chuang et al. (2020) reviewed the conditions under which learned molecular representations surpass fixed descriptors and concluded that the crossover occurs at dataset sizes considerably larger than those available in typical medicinal-chemistry programmes. Yang et al. (2019) reported similar findings across a benchmark panel, with graph-based methods requiring substantial data volume before their advantage materialises. With 1,008 training compounds and 98 descriptors, this study sits well within the regime where ensemble methods on curated descriptors remain the rational default.

A related observation concerns the diagnostic value of apparent fit. XGBoost achieved an R^2_fit of 1.000 perfect memorisation of the training partition while

ranking second externally. SVR achieved 0.987 apparent fit and third-weakest external RMSE. If either model had been reported on apparent fit alone, as a non-trivial fraction of the published QSAR literature still does, both would have appeared superior to the random forest that in fact generalises best. The y-randomisation result reinforces the point from the other direction: models trained on permuted responses achieved a mean apparent R^2 of 0.846 while achieving a mean cross-validated Q^2 of −0.112. Apparent fit, in this setting, carries essentially no information about predictive value.

5.3 The applicability domain result and its practical reading

The finding that 24.6% of external compounds fall outside the applicability domain deserves comment because it is the kind of statistic that published QSAR studies frequently omit or minimise. The magnitude follows directly from the geometry of the problem: with 98 descriptors and 1,008 training compounds, the leverage threshold $h^* = 3(p + 1)/n = 0.295$ is comparatively permissive in absolute terms, yet high-dimensional descriptor spaces are sparsely populated even by a thousand compounds, and a substantial proportion of any test set will occupy regions with limited training support.

There are two answers; one is more positive than the other. The first is dimensional: As the number of descriptors is reduced, further (more aggressive univariate filtering) or projection onto principal components, the fraction of compounds flagged will decrease, as will be the p value. This answer does not enhance the statistic, but is an improvement in the domain due to interpretability. The second answer is an operational one: Keep the descriptor set on the board, allow the domain estimate as correct, and use membership as a hard filter on the board. This study is based on the second one is the reason it is shown as a column in Table 9, and the reason one of the four conjunctive prioritisation criteria is domain membership. The result is apparent in the prioritized set: 9 of the top 10 candidates have leverage below 0.09, that is, they aren't just extrapolating and are more of a prediction than an extrapolation.

5.4 The role of docking, and the meaning of a weak correlation

The worst correlation coefficient is the near-zero between the predicted potency and binding score ($r = −0.17$) which is most likely to be misinterpreted and should be addressed. If a reader is used to seeing QSAR–docking correlations done and reported as a success, then he/she might think this is a failure. The methodological literature is in favour of the other interpretation. One of the most systematic assessments published to date (Warren et al., 2006) tested ten docking programs and thirty-seven scores using eight protein targets, and concluded that pose reproduction was satisfactory but that there were no scores that provided a meaningful correlation with measured affinity. In a subsequent study, comparing 10 programs, Wang et al. (2016) came to similar findings. The physical reasons are well understood: Errors in scoring functions are of the order of the difference in affinities within a congeneric series and are a lack of ability to account for desolvation, conformational entropy, or explicit water networks.

Given this, a strong correlation between a QSAR prediction and a docking score would be more troubling than a weak one, because it would most plausibly indicate that both quantities depend on the same size and lipophilicity descriptors rather than that two independent methods had converged. The design adopted here treats the two as sequential, independent filters: QSAR ranks potency, docking assesses geometric plausibility and generates mechanistic hypotheses about which contacts drive the ranking. Under that design, orthogonality is the desired property.

5.5 Implications for scaffold optimisation

Three actionable inferences follow for a medicinal-chemistry programme working this scaffold class.

First, the calculated logP window of approximately 4.2 to 4.9 that characterises the prioritised set represents the model's estimate of the optimisation sweet spot. Analogues below this window sacrifice back-pocket complementarity; those above it incur the quadratic penalty and, in practice, the metabolic and promiscuity liabilities that accompany high lipophilicity. Synthesis effort is best concentrated within the window.

Second, the concentration of dimethylamino and methoxy groups at the quinazoline 7-position among the leading candidates identifies that position as the most productive site for further variation. The 7-position projects toward solvent in the canonical binding mode, which means substitutions there modulate physicochemical properties with comparatively little potency cost the structural basis for the solubilising side chains that characterise clinical agents in this class.

Third, the QED range of the prioritised set (0.689 to 0.785, against a library median of 0.66) indicates that potency prioritisation and developability

prioritisation are not in conflict in this region of chemical space. This is not always the case, and where it holds it is worth exploiting: the conjunctive filter costs little in potency while substantially improving the developability profile of the surviving set.

6. Limitations

Several limitations bound the interpretation of this work, and they are stated directly rather than qualified.

The response variable is a benchmark surface, not measured bioactivity. As specified in Sections 3.1 and 3.4, pIC_{50} values were generated from a documented function of descriptors with additive noise. All reported model statistics are genuine measurements of signal recovery under realistic noise, and they validate the pipeline as a methodology. They do not constitute evidence about EGFR inhibition by any SBQ compound. Converting this study into an empirical discovery campaign requires substituting curated experimental data, and the reported performance figures should be expected to degrade when that substitution is made real structure–activity landscapes contain activity cliffs, assay heterogeneity, and stereochemical dependencies that a smooth generative function does not reproduce (Maggiora, 2006; Stumpfe & Bajorath, 2012).

Binding values are surrogate estimates. No docking calculation was executed. Section 3.7 specifies the full AutoDock Vina protocol, and Table 9 marks the ΔG_Vina column as pending. Any statement in this paper about binding geometry is a hypothesis derived from published structural work on this scaffold class, not an output of this pipeline.

Descriptor scope is two-dimensional. Three-dimensional descriptors, conformational ensembles, and quantum-chemical properties were not computed. For a scaffold class where the bioactive conformation is well characterised and largely conserved, this omission is defensible; for scaffolds with substantial conformational flexibility it would not be.

Random splitting overstates prospective performance. The stratified random partition used here is standard but optimistic. Sheridan (2013) demonstrated that time-split validation, which mimics the actual prospective use of a QSAR model, typically yields substantially lower performance than random splitting on the same data. Scaffold-based splitting would similarly stress-test generalisation across chemotypes rather than within them.

The library is congeneric by construction. Every compound shares the 4-anilinoquinazoline core. This is appropriate for QSAR, which assumes a common binding mode, but it means the model carries no information about scaffold hopping and should not be applied outside this chemotype. The applicability domain formalises this constraint quantitatively.

Benchmark construction can flatter methods. Wallach and Heifets (2018) demonstrated that many ligand-based benchmarks measure dataset bias rather than model quality, and that apparently strong performance can reflect analogue bias in the underlying data. A synthetically generated response surface is not immune to an analogous concern: the descriptors used to generate the response are among those available to the model. This inflates achievable performance relative to a real assay, and the reported statistics should be read as an upper bound on what the same pipeline would achieve on experimental data.

No experimental validation. No compound has been synthesised or assayed. The prioritised candidates constitute hypotheses for experimental testing, not validated hits.

7. Conclusion

This study specifies and validates an integrated pipeline coupling machine-learning QSAR with structure-based triage for the prioritisation of anticancer candidates on the 4-anilinoquinazoline scaffold. Across a combinatorial library of 1,260 enumerated analogues and 98 retained descriptors, a random forest ensemble achieved cross-validated Q^2 of 0.882 and external R^2 of 0.873 with a test-set RMSE of 0.348 log units, satisfying all Golbraikh–Tropsha criteria and excluding chance correlation by a 0.994-unit margin in y-randomisation. Descriptor attribution independently recovered the parabolic lipophilicity dependence and polar-surface penalty that classical Hansch analysis established, providing mechanistic corroboration of the statistical result. Conjunctive filtering on predicted potency, applicability-domain membership, and drug-likeness returned 25 prioritised candidates from the library, ten of which are characterised in detail.

Three methodological conclusions generalise beyond this scaffold. Ensemble methods on curated two-dimensional descriptors remain the rational default at the data scale characteristic of a single medicinal-chemistry programme; the neural architecture evaluated here, with far greater capacity, did not surpass them. Apparent fit is diagnostically worthless in this setting, as demonstrated both by the

divergence between fit and external performance across architectures and by the high apparent R^2 attained on permuted responses. And weak correlation between QSAR ranking and docking score is the expected and desirable outcome of combining two methods with genuinely different information content, rather than a defect requiring correction.

The pipeline's principal contribution is executability. Every reported statistic derives from seeded, versioned code operating on a specified library, and the two quantities that require experimental substitution the response variable and the binding energies are marked as such throughout rather than presented as computed results. Immediate next steps are the substitution of curated ChEMBL bioactivity data for the benchmark response, execution of the specified docking protocol against EGFR structure 1M17, re-evaluation under scaffold-based and time-split partitioning, and synthesis of the leading candidates for cellular and enzymatic evaluation.

References

Alexander, D. L. J., Tropsha, A., & Winkler, D. A. (2015). Beware of R^2 even when using the correct model. *Journal of Chemical Information and Modeling, 55*(7), 1316–1322.

Barker, A. J., Gibson, K. H., Grundy, W., Godfrey, A. A., Barlow, J. J., Healy, M. P., Woodburn, J. R., Ashton, S. E., Curry, B. J., Scarlett, L., Henthorn, L., & Richards, L. (2001). Studies leading to the identification of ZD1839 (Iressa): An orally active, selective epidermal growth factor receptor tyrosine kinase inhibitor. *Bioorganic & Medicinal Chemistry Letters, 11*(14), 1911–1914.

Bender, A., & Cortés-Ciriano, I. (2021). Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug Discovery Today, 26*(2), 511–524.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235–242.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry, 4*(2), 90–98.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., Consonni, V., Kuz'min, V. E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., & Tropsha, A. (2014). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry, 57*(12), 4977–5010.

Chuang, K. V., Gunsalus, L. M., & Keiser, M. J. (2020). Learning molecular representations for medicinal chemistry. *Journal of Medicinal Chemistry, 63*(16), 8705–8722.

Cohen, P., Cross, D., & Jänne, P. A. (2021). Kinase drug discovery 20 years after imatinib: Progress and future directions. *Nature Reviews Drug Discovery, 20*(7), 551–569.

Consonni, V., Ballabio, D., & Todeschini, R. (2009). Comments on the definition of the Q^2 parameter for QSAR validation. *Journal of Chemical Information and Modeling, 49*(7), 1669–1678.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Daina, A., Michielin, O., & Zoete, V. (2017). SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports, 7*, 42717.

Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *Journal of Chemical Information and Modeling, 61*(8), 3891–3898.

Ertl, P., & Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics, 1*, 8.

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling, 50*(7), 1189–1204.

Free, S. M., & Wilson, J. W. (1964). A mathematical contribution to structure-activity studies. *Journal of Medicinal Chemistry, 7*(4), 395–399.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research, 45*(D1), D945–D954.

Golbraikh, A., & Tropsha, A. (2002). Beware of q²! *Journal of Molecular Graphics and Modelling, 20*(4), 269–276.

Gramatica, P. (2007). Principles of QSAR models validation: Internal and external. *QSAR & Combinatorial Science, 26*(5), 694–701.

Hanahan, D. (2022). Hallmarks of cancer: New dimensions. *Cancer Discovery, 12*(1), 31–46.

Hansch, C., & Fujita, T. (1964). ρ-σ-π analysis: A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society, 86*(8), 1616–1626.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular graph convolutions: Moving beyond fingerprints. *Journal of Computer-Aided Molecular Design, 30*(8), 595–608.

Landrum, G. (2024). *RDKit: Open-source cheminformatics* [Computer software]. https://www.rdkit.org

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews, 46*(1–3), 3–26.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling, 55*(2), 263–274.

Maggiora, G. M. (2006). On outliers and activity cliffs Why QSAR often disappoints. *Journal of Chemical Information and Modeling, 46*(4), 1535.

Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., Isayev, O., Curtalolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D. A., Agrafiotis, D., Cherkasov, A., & Tropsha, A. (2020). QSAR without borders. *Chemical Society Reviews, 49*(11), 3525–3564.

Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T. D., Gramatica, P., Jaworska, J. S., Kahn, S., Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R., Roberts, D. W., Schultz, T. W., Stanton, D. T., van de Sandt, J. J. M., Tong, W., … Yang, C. (2005). Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. *ATLA Alternatives to Laboratory Animals, 33*(2), 155–173.

Organisation for Economic Co-operation and Development. (2007). *Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models*. OECD Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Polishchuk, P. (2017). Interpretation of quantitative structure–activity relationship models: Past, present, and future. *Journal of Chemical Information and Modeling, 57*(11), 2618–2639.

Roskoski, R. (2019). Small molecule inhibitors targeting the EGFR/ErbB family of protein-tyrosine kinases in human cancers. *Pharmacological Research, 139*, 395–411.

Roy, K., Kar, S., & Ambure, P. (2015). On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems, 145*, 22–29.

Rücker, C., Rücker, G., & Meringer, M. (2007). y-Randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling, 47*(6), 2345–2357.

Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., Fisher, J., Jansen, J. M., Duca, J. S., Rush, T. S., Zentgraf, M., Hill, J. E., Krutoholow, E., Kohler, M., Blaney, J., Funatsu, K., Luebkemann, C., & Schneider, G. (2020). Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery, 19*(5), 353–364.

Sheridan, R. P. (2013). Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling, 53*(4), 783–790.

Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacological Reviews, 66*(1), 334–395.

Stamos, J., Sliwkowski, M. X., & Eigenbrot, C. (2002). Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *Journal of Biological Chemistry, 277*(48), 46265–46272.

Stumpfe, D., & Bajorath, J. (2012). Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry, 55*(7), 2932–2942.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249.

Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics* (2nd ed.). Wiley-VCH.

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics, 29*(6–7), 476–488.

Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry, 31*(2), 455–461.

Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry, 45*(12), 2615–2623.

Wallach, I., & Heifets, A. (2018). Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of Chemical Information and Modeling, 58*(5), 916–932.

Walters, W. P., & Barzilay, R. (2021). Critical assessment of AI in drug discovery. *Current Opinion in Chemical Biology, 60*, 85–90.

Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., Tian, S., & Hou, T. (2016). Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: The prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics, 18*(18), 12964–12975.

Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E., & Head, M. S. (2006). A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry, 49*(20), 5912–5931.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling, 59*(8), 3370–3388.

Yun, C.-H., Boggon, T. J., Li, Y., Woo, M. S., Greulich, H., Meyerson, M., & Eck, M. J. (2007). Structures of lung cancer-derived EGFR mutants and inhibitor complexes: Mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell, 11*(3), 217–227.

Zhang, J., Yang, P. L., & Gray, N. S. (2009). Targeting cancer with small molecule kinase inhibitors. *Nature Reviews Cancer, 9*(1), 28–39.

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., … Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology, 37*(9), 1038–1040.