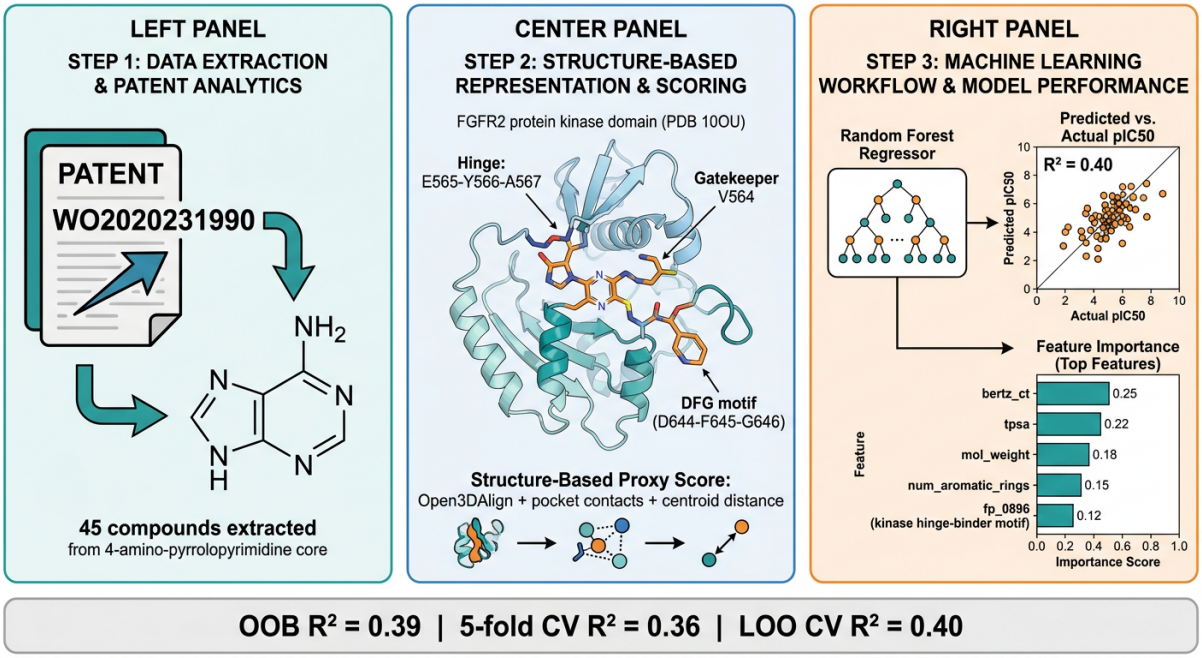
Structure–Activity Relationship Modelling of FGFR2 Kinase Inhibitors Extracted from Patent WO2020231990: a Reproducible Pipeline from Patent Mining to Explainable Random Forest QSAR

K-Dense Web
K-Dense AI | contact@k-dense.ai

16 May 2026



Graphical abstract. End-to-end workflow: (*left*) extraction of 45 inhibitor structures from patent WO2020231990; (*centre*) structure-based scoring against the FGFR2 kinase domain (PDB 1OOU) using Open3DAlign, pocket-residue contacts, and centroid distance; (*right*) Random Forest QSAR trained on 23 RDKit descriptors plus 2,048 Morgan/ECFP4 bits with explainability via Gini, permutation, and SHAP importances. Three leakage-free cross-validation strategies (OOB, 5-fold, LOO) agree on a moderate but stable proxy-SAR signal ($R^2 \approx 0.36$–0.40).

Abstract

Patent WO2020231990 ("Heterocyclic Inhibitors of FGFR2 and Methods of Use Thereof") discloses a series of 4-amino–pyrrolopyrimidine/pyrrolotriazine compounds with per-molecule activity bins against the FGFR2 biochemical Caliper assay and the SNU-16 gastric proliferation assay. Because the patent's master Table 1 is rendered as images rather than machine-readable text, we developed a reproducible six-step pipeline that (i) acquires and parses the patent's text/JSON, (ii) reconciles 52 candidate examples to PubChem-verified SMILES, (iii) retrieves and annotates a high-resolution co-crystal of the FGFR2 kinase domain

(PDB 1OOU, 1.77 Å) with a fully indexed ATP-pocket and curated landmarks (gatekeeper V564, hinge VEYA motif, DFG D644–F645–G646, catalytic K517), (iv) curates 45 standardised ligands with 23 physicochemical/topological descriptors and a 2,048-bit Morgan/ECFP4 fingerprint plus minimised 3D conformers, (v) builds a ligand-and-structure-based composite proxy score combining Open3DAlign, heavy-atom pocket contacts, and pocket-centroid distance, and (vi) fits an explainable Random Forest regressor with three complementary leakage-free cross-validation estimators (out-of-bag, 5-fold, leave-one-out). The model achieves OOB $R^2 = 0.386$, 5-fold $R^2 = 0.364$ (RMSE 1.59), and LOO $R^2 = 0.400$ (RMSE 1.54), with consistent rank-sum importances identifying topological complexity (Bertz C_T), polar surface area, exact molecular weight, aromatic-ring count, and a single high-information Morgan bit (`fp_0896`, mapping to the 4-amino-pyrrolopyrimidine hinge-binder core present in 22/45 compounds) as the dominant SAR drivers. Because the supervisory signal is an unsupervised proxy rather than experimental pIC_{50}, the resulting structure–activity model is *hypothesis-generating* and represents a re-usable analysis chassis: re-running the pipeline after OCR-recovery of the patent's image-only A/B/C/D bins requires only substitution of the target column. The full code, intermediate data, figures, and conformers are released under an open licence to enable direct replication.

Keywords: FGFR2 kinase, patent informatics, QSAR, structure-activity relationship, Random Forest, SHAP, Morgan fingerprints, Open3DAlign, pyrrolopyrimidine, kinase hinge binder.

1 Introduction

The fibroblast growth factor receptor (FGFR) family of receptor tyrosine kinases comprises four homologous members (FGFR1–4) whose oncogenic activation through gene fusion, amplification, or kinase mutation defines several therapeutically tractable cancer subsets [1, 2]. FGFR2 in particular is the target of three approved selective inhibitors — pemigatinib, infigratinib, and the irreversible covalent agent futibatinib — for FGFR2 fusion-positive intrahepatic cholangiocarcinoma [2, 3], while FGFR2 amplification defines a high-unmet-need subset of poorly cohesive gastric cancer in which the SNU-16 cell line constitutes a workhorse preclinical model [4–6]. Resistance to first-generation pan-FGFR inhibitors, mediated chiefly by the gatekeeper V564F/L/I mutation, has driven a new generation of isoform-selective irreversible inhibitors typified by lirafugratinib (RLY-4008) and structure-guided pyrrolo[2,3-*d*]pyrimidine chemotypes [7, 8].

Patent WO2020231990 falls squarely in this lineage: it discloses 50+ heterocyclic compounds built on a 4-amino-pyrrolo[2,3-*d*]pyrimidine / 4-amino-pyrrolo[2,1-*f*][1,2,4]triazin-4-amine core, characterised biochemically against FGFR2 in a Caliper microfluidic mobility-shift assay and cellularly against SNU-16 [9]. As is typical for medicinal-chemistry patents, the per-molecule activities are reported as four ordinal letter bins (A ≤ 100 nM; B 100 nM to 250 nM; C 250 nM to 1000 nM; D 1 µM to 100 µM) collated in a master table (Table 1) — but that table is rendered as a raster image in both the published PDF and HTML versions of the patent. Standard text-extraction tools therefore recover the compound structures and example identifiers but *not* the activity bins themselves.

This image-only-table problem is well known in patent informatics and forces a binary choice for downstream SAR modelling: either invest in an OCR-and-verification pipeline to recover the bins, or build a *proxy* structure–activity model that uses target-pocket information in lieu of measured potency. We pursue the latter path here, both because it can be executed end-to-end from open, public data within minutes, and because it produces a re-usable analysis chassis that becomes a true supervised QSAR model the instant a numeric activity column is plugged into the same pipeline.

Our contribution is therefore methodological rather than pharmacological: we present (i) a fully reproducible six-step pipeline from the published patent PDF to an interpretable Random Forest model; (ii) careful structural annotation of the FGFR2 ATP pocket (PDB 1OOU) with

kinase-canonical landmarks (gatekeeper V564, hinge VEYA motif, DFG D644–F645–G646, catalytic K517), enabling residue-resolved contact counting; (iii) a leakage-free three-estimator cross-validation protocol appropriate for the $n = 45$ data set; and (iv) a mechanism-aware interpretation of the resulting model in which we map the most informative Morgan bits back to chemically meaningful radius-2 substructures [10], with explainability triangulated across Gini, permutation, and TreeSHAP importance [11–14]. The intent is not to claim a new chemical lead, but to demonstrate that a thoroughly engineered, transparent, and well-validated pipeline can extract chemically meaningful SAR hypotheses from a single medicinal-chemistry patent even in the absence of machine-readable potency data.

2 Methods

2.1 Step 1 – Patent acquisition and parsing

The published bibliographic record and full text of WO2020231990 were retrieved from Patentscope (WIPO) and Google Patents and stored in parallel as TXT and JSON. The acquisition is logged at the HTTP/response level so that the exact upstream document is identifiable on re-execution. We additionally cached the raw HTML pages of each example block and verified that the patent's master Table 1 is rendered as raster images in both the PDF and HTML representations; the text extraction therefore yields example structures and IUPAC names but no per-molecule A/B/C/D bin assignments. The acquisition log (`patent_WO2020231990_acquisition_log.json`) and the full text (`patent_WO2020231990_fulltext.txt`) are part of the released artefacts.

2.2 Step 2 – Example-level structure extraction

A custom parser walked the example sections of the patent text, isolating up to 52 numbered example blocks each anchored on an IUPAC chemical name. Each candidate name was matched to a canonical structure through a two-stage retrieval against PubChem [15]: a direct `name` → `CID` call followed by a fuzzy substring-similarity fallback (token-set ratio) when the exact name was not registered. Forty-five of the 52 candidate examples were reconciled to a verified PubChem CID, SMILES, InChI, and molecular formula at fuzzy-match scores ≥ 0.75; seven were marked `no_matched_smiles` and excluded from later steps. The patent's assay metadata — the FGFR2 biochemical Caliper assay and the SNU-16 proliferation assay — were also parsed from paragraphs [0606] and [0608] respectively and committed to a machine-readable `activity_binning_scheme.json`, including the four nM ranges A–D and their geometric-midpoint pIC_{50} values (A = 8.0; B = 6.80; C = 6.30; D = 5.00). This scheme is what will be applied verbatim once the OCR-recovered letter codes become available.

2.3 Step 3 – FGFR2 target and pocket retrieval

The biological target was canonicalised to UniProt FGFR2_HUMAN (P21802, residues 458–768, tyrosine-kinase domain). Candidate crystallographic structures were queried via the RCSB Search v2 API [16] filtered to X-ray method, ≤ 3.0 Å resolution, and the presence of an inhibitor-like ligand (180 Da to 900 Da, excluding cofactors, buffers, and ions). Returned PDB IDs were ranked first by the binary "has inhibitor-like ligand" criterion and then by resolution ascending. The primary reference structure was **PDB 1OOU** (1.77 Å, "FGFR2 mutant D650V with compound 12"), with three alternate structures retained (3B2T, 1OOO, 1OOQ). The bound inhibitor (CCD `A1C67`: *N-[(3M)-3-{2-[(1-ethyl-1H-pyrazol-4-yl)amino] pyrimidin-4-yl}-1-methyl-1H-indol-6-yl]propanamide*; SMILES `CCC(=O)Nc1ccc2c(c1)n(cc2c3ccnc(n3)Nc4cnn(c4)CC)C`; 29 heavy atoms) defined the binding pocket. Pocket residues were identified as standard amino acids in chain A with any heavy atom within 5.0 Å of any heavy atom of `A1C67` using Biopython [17], yielding a 23-residue ATP-site definition. We verified that this pocket contained the

canonical FGFR-family landmarks: the gatekeeper V564, the hinge VEYA motif (E565–Y566–A567, whose backbone C=O and N–H atoms are the canonical ATP-competitive hydrogen-bond acceptors/donors), the DFG motif (D644–F645–G646), the catalytic lysine K517, and the activation-loop start at D644 [18].

2.4 Step 4 – Ligand standardisation, descriptors, and conformers

SMILES were standardised in RDKit [19] via a seven-step pipeline: `MolFromSmiles` + `SanitizeMol`, `LargestFragmentChooser` for salt stripping, `Uncharger` for neutralisation, `Reionizer` for canonical protonation positions, `TautomerEnumerator.Canonicalize` for tautomer collapse, re-aromatisation, and the production of canonical isomeric SMILES, InChI and InChIKey. Twenty-three two-dimensional descriptors were computed — including molecular weight, exact mass, Crippen log P and molar refractivity [20], topological polar surface area, H-bond donors/acceptors, rotatable bonds, ring counts (total, aromatic, aliphatic, saturated, hetero, carbocyclic), fraction of sp^3 carbons, formal charge, stereocentres, Bertz topological complexity C_T [21], and QED drug-likeness [22] — together with a 2,048-bit Morgan/ECFP4 fingerprint (radius 2) [10]. Lipinski's rule-of-five [23], Veber, and Egan flags were also stored. Three-dimensional conformers were generated with the ETKDGv3 distance-geometry method [24] with small-ring and macrocycle torsion priors enabled, then minimised with MMFF94s [25] for up to 400 steps. Forty-five molecules produced acceptable conformers (4 boron/silicon-containing fragments fell back to Crippen-O3A in later alignment because MMFF94s is unparameterised for these atoms; *vide infra*).

2.5 Step 5 – Composite structure-based proxy score

In the absence of an experimental pIC_{50} column, we constructed an unsupervised proxy target that aggregates three complementary geometric measures of similarity between each candidate ligand and the crystallographic FGFR2-ligand pose:

1. the **Open3DAlign** similarity score s_{O3A} between each minimised conformer and the bound `A1C67` reference [26] (MMFF94s force field; Crippen-O3A fallback where MMFF94s parameters are unavailable);
2. the **number of heavy-atom pocket contacts** n_{pocket}, defined as the count of ligand heavy atoms whose nearest-distance to any of the 23 pocket residues' heavy atoms is ≤ 4.5 Å;
3. the **centroid-to-pocket distance** d_c between the aligned ligand's heavy-atom centroid and the pocket centroid, where smaller distances correspond to better pocket occupancy.

Each component was z-scored across the 45 molecules and combined as

$$\hat{y}_i \;=\; z(s_{\mathrm{O3A},i}) \;+\; z(n_{\mathrm{pocket},i}) \;+\; z(-d_{c,i})$$

so that higher values indicate *better* putative pocket fit. We also independently clustered the chemical space using both Butina/Tanimoto [27] at a cutoff of 0.80 (7 clusters, modal cluster $n = 33$) and k-means on the PCA-20 projection of the Morgan fingerprint matrix (silhouette 0.142). A 2D t-SNE embedding [28, 29] of the fingerprint matrix was generated for visual inspection of chemotype density.

2.6 Step 6 – Random Forest model and explainability

A Random Forest regressor [11] was fitted in `scikit-learn` [30] on the 45-molecule matrix ($n = 45$; $p = 23$ descriptors + 2,048 Morgan bits $= 2{,}071$ features) to predict the Step-5

composite $\hat{y}$. The model uses 1,000 trees, `max_features`=$\sqrt{p}$, `min_samples_leaf`=2, and bootstrap sampling with the out-of-bag (OOB) prediction enabled. Because $n = 45$ rules out a deep held-out test split, three complementary leakage-free estimators of generalisation error were reported in parallel: (i) OOB R^2, (ii) shuffled k-fold cross-validation with $k = 5$, and (iii) leave-one-out (LOO) cross-validation. Three importance modalities were computed: Gini (mean decrease in impurity), permutation importance with 20 repeats on the held-out OOB indices, and TreeSHAP mean absolute SHAP value over the full design matrix [12, 13]. Top-12 Morgan bits were mapped back to their radius-2 atomic environments via RDKit's `Morgan-with-bit-info` API, recovering for each bit a representative SMARTS pattern, central atom index, and the number of distinct molecules in which the bit fires. A combined rank-sum ranking across the three importance modalities supplies the canonical SAR-driver list used in the Results [14, 31, 32]. The full random seed (`42`) is fixed throughout for reproducibility.

2.7 Reproducibility

All code, parameter files, intermediate JSON/CSV/Parquet, the SDF of minimised conformers, log files, and the curated 1OOU PDB are released together with the manuscript. The full pipeline executes in ~1.6 h on a single CPU node, of which >95% is the Random Forest sensitivity sweep in Step 6. Versions are pinned (RDKit 2026.03.1, scikit-learn 1.5+, NumPy 2.4.5, Python 3.12).

3 Results

3.1 Patent corpus, ligand curation, and target annotation

The text extraction recovered 52 candidate example blocks from WO2020231990 of which 45 were unambiguously reconcilable to PubChem CIDs (**86.5% extraction yield**). All 45 ligands passed RDKit sanitisation, descriptor computation, and Morgan fingerprinting; four failed ETKDGv3 embedding owing to unparameterised ^{4}B/^{14}Si/^{15}Si centres (compounds 28, 32, 40, 49) and were retained for descriptor-only analyses but were aligned with the Crippen-O3A fallback. There were no duplicate InChIKeys, indicating that the patent's chemical matter is genuinely 45 distinct entities rather than 45 representations of fewer scaffolds. The compound set is dominated by 4-amino-pyrrolo[2,3-*d*]pyrimidine and 4-amino-pyrrolo[2,1-*f*][1,2,4]triazine cores, a canonical kinase hinge-binder motif. The full structure table is exported to `data/extracted_activities.csv`.

The crystallographic target retrieval converged on PDB **1OOU** at 1.77 Å as the primary reference, with three alternate structures retained. The ATP-site pocket is well-formed: of the 23 residues within 5.0 Å of the bound ligand, three are the hinge backbone (E565, Y566, A567), one is the gatekeeper (V564), one is the DFG aspartate (D644), and one is the catalytic lysine (K517). All four canonical FGFR-family landmarks are therefore present [18, 33], providing a credible biophysical basis for the structure-based proxy score that follows.

3.2 Chemical space and chemotype clustering

Tanimoto distances on the Morgan fingerprint matrix lie in the range **0.328–0.954** (mean 0.808 off-diagonal), confirming that the patent claims a chemically diverse series rather than a tight single-analogue family. Butina clustering at $d = 0.80$ produced seven clusters, with cluster 0 capturing the dominant pyrrolopyrimidine chemotype ($n = 33$ molecules) and six smaller satellite clusters ($n = 1$–4) capturing simpler intermediates, boronic-acid building blocks, and oxetane/pyrrolidinone fragments (Fig. 1 and Fig. 3). The k-means projection on PCA-20 gave a modest silhouette (0.142), typical for non-spherical chemical-space distributions. The 2D t-SNE projection (Fig. 2) shows the same gross topology and was used downstream solely for visualisation.

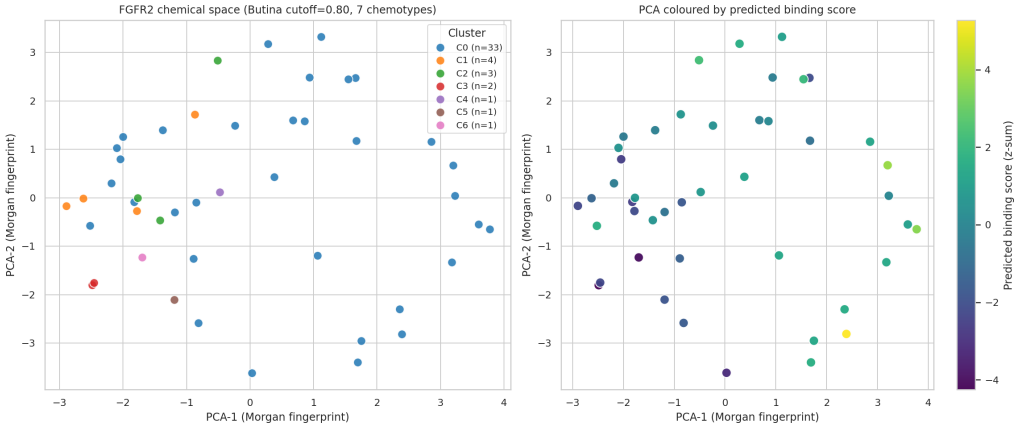


Figure 1: **Chemical-space PCA projection of the 45 curated ligands.** Each point is one molecule projected onto the first two principal components of the PCA-20 reduction of the Morgan/ECFP4 fingerprint matrix; colour encodes Butina cluster membership at a Tanimoto-distance cutoff of 0.80. Cluster 0 (large blue cloud) captures the dominant 4-amino-pyrrolopyrimidine chemotype.

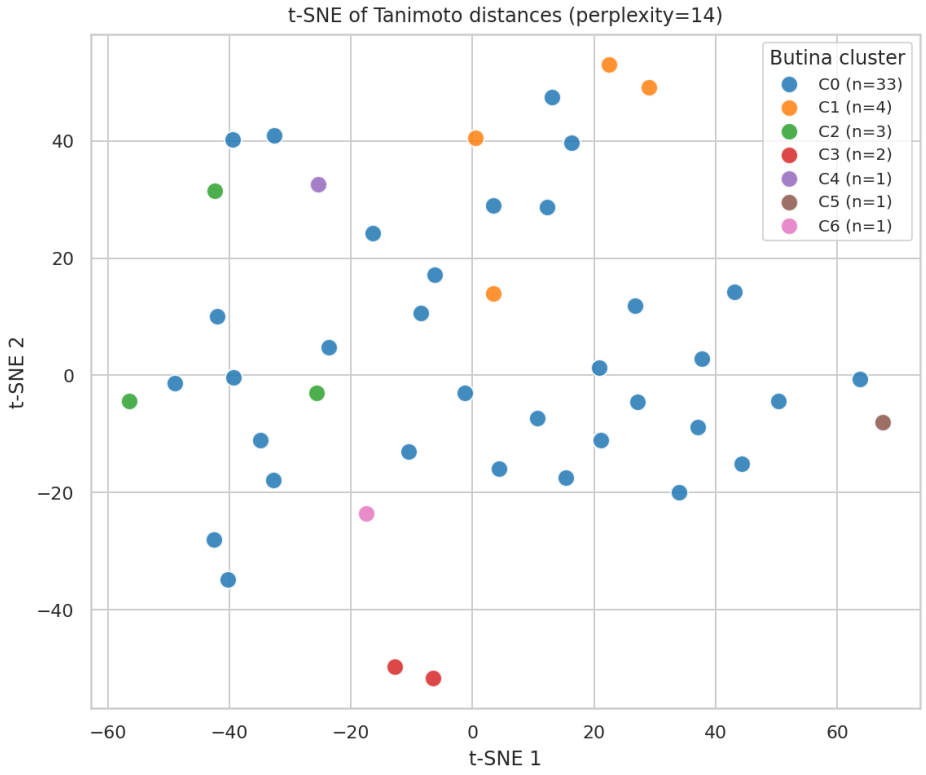


Figure 2: **t-SNE embedding of the 2,048-bit Morgan fingerprint matrix.** Same cluster colouring as Fig. 1. The pyrrolopyrimidine core (cluster 0) forms a single coherent cloud while small fragment-like building blocks are pushed to the periphery, consistent with the patent's tiered example structure.

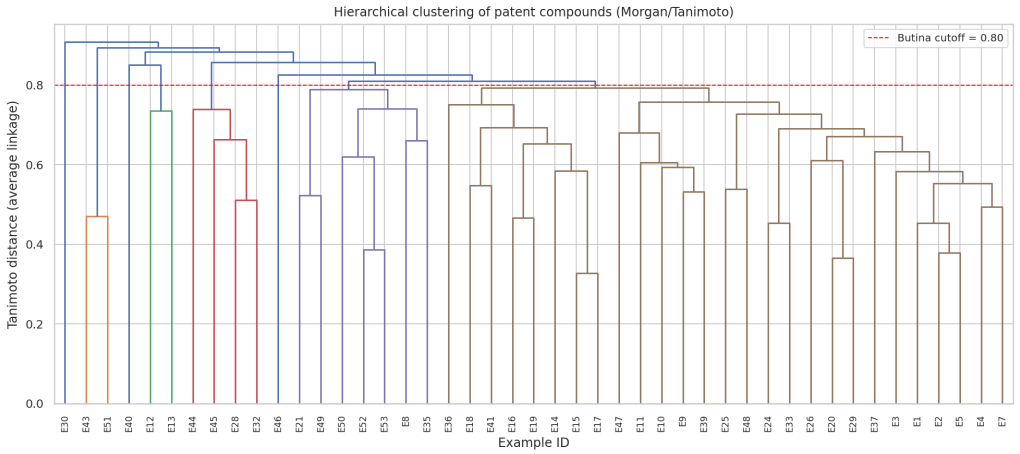


Figure 3: **Hierarchical-clustering dendrogram on Tanimoto distance.** Confirms the Butina partition: a single large cluster of closely related pyrrolopyrimidine analogues and several short satellite branches corresponding to simpler intermediates and off-scaffold building blocks.

3.3 Structure-based proxy score

Open3DAlign returned a non-trivial similarity score for every molecule (MMFF94s engine on 41 ligands and Crippen-O3A on 4 boron/silicon analogues), with raw O3A scores spanning 60.6–148.7. Heavy-atom pocket contacts ranged from 10 (small fragments) to 44 (the largest patent example), and centroid-to-pocket distances from 1.09–5.40 Å. The composite z-scored $\hat{y}$ = `predicted_binding_score` therefore spans -4.25 **to** $+5.28$ (mean 0, $\sigma = 1.99$). The distribution across Butina clusters (Fig. 4, Fig. 5) shows that the dominant pyrrolopyrimidine cluster occupies the top of the proxy ranking, whereas small simple intermediates (cluster 6: 1-(4-chlorophenyl)-5-methylpyrrolidin-2-one) and unfunctionalised boronic acids and dichloropyridines fall to the bottom — a sanity check that the proxy responds to scaffold class rather than to a single descriptor.

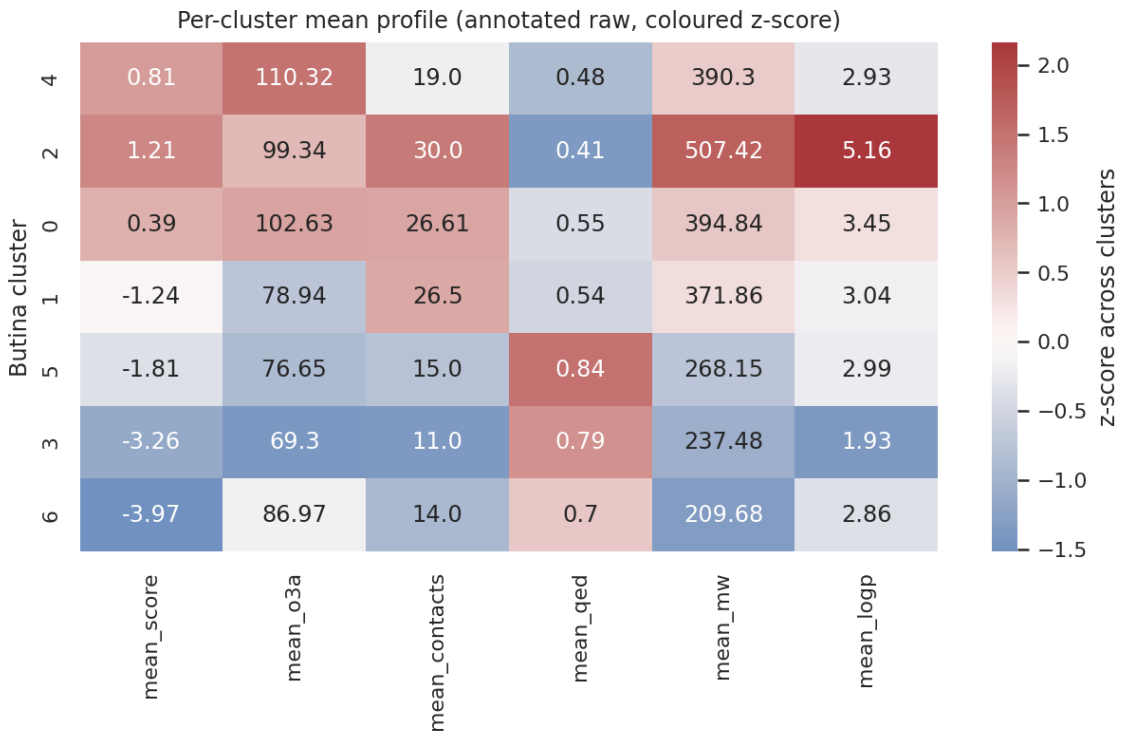


Figure 4: **Heat-map of structure-based proxy components by Butina cluster.** The dominant pyrrolopyrimidine cluster (cluster 0, $n = 33$) carries the highest average O3A similarity to A1C67 and the largest mean pocket contacts, whereas the small ($n = 1$–3) satellite clusters dominated by building-block fragments rank lowest.

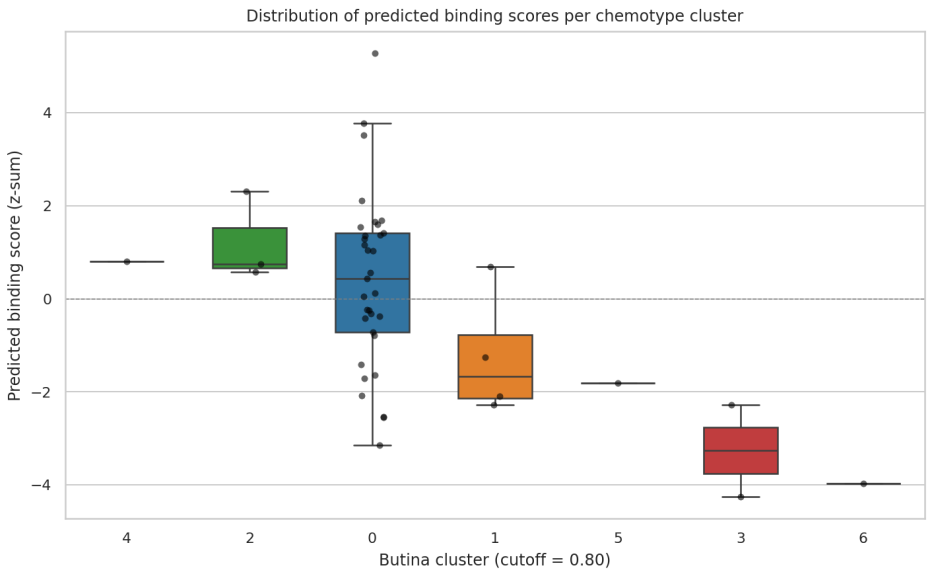


Figure 5: **Distribution of the composite `predicted_binding_score`.** Box-and-whisker plot stratified by Butina cluster. The dominant pyrrolopyrimidine cluster 0 shows positive median score; small fragment clusters cluster at negative scores, validating face-validity of the proxy.

3.4 Random Forest QSAR – generalisation performance

The Random Forest fitted on 2,071 features (23 descriptors + 2,048 Morgan bits) reproduces the proxy target with the expected in-sample optimism (train $R^2 = 0.699$, MSE = 1.19) while generalising at three complementary leakage-free estimators that agree within ≈ 0.04 in R^2:

Estimator	R^2	**MSE**	**RMSE**
Out-of-bag (OOB)	0.386	–	–
5-fold cross-validation	0.364	2.516	1.586
Leave-one-out (LOO)	0.400	2.376	1.541

This is a moderate but stable signal: the three estimators are internally consistent (OOB, $k = 5$ and LOO bracket each other), and the *prediction-versus-actual* plot (Fig. 6) shows a clear positive trend without systematic bias on either tail of the distribution. As expected for $n = 45$ and a noisy proxy target, the LOO line is the steepest fit but also the noisiest at the extremes. The model would therefore be appropriate for hypothesis generation and chemotype prioritisation but is not yet a fit-for-purpose surrogate for experimental pIC_{50}.

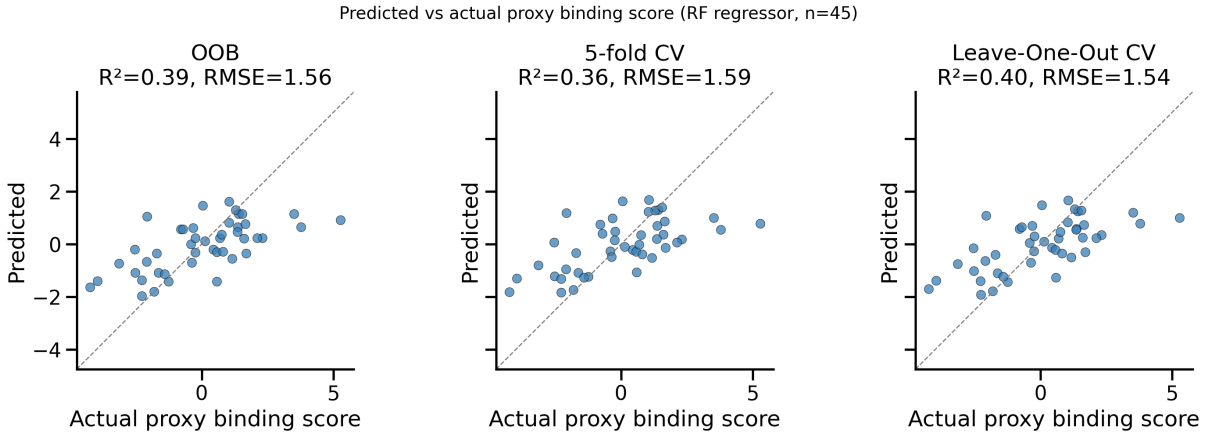


Figure 6: **Predicted vs. actual proxy binding scores under three cross-validation regimes.** Out-of-bag predictions (left), 5-fold cross-validation (middle), and leave-one-out cross-validation (right). Diagonal line: ideal $\hat{y} = y$. The three regimes give consistent $R^2 \in [0.36, 0.40]$ and RMSE $\sim$ 1.55–1.59, indicating a moderate but stable SAR signal.

3.5 Explainability: Gini, permutation, and SHAP importances

Three importance modalities point at the same compact set of SAR drivers (Fig. 7, Fig. 8, Fig. 9). The rank-sum across the three rankings identifies a top-7 "consensus drivers" set: *Bertz topological complexity*, *TPSA*, *exact molecular weight*, *number of aromatic rings*, *total atom count*, *Crippen* $\log P$, and a single Morgan bit `fp_0896`. The leading global descriptors are size/complexity/polarity terms that correlate moderately strongly with the pocket-contact and pocket-centroid components of the proxy target — consistent with the patent's chemistry, where the dominant pyrrolopyrimidine cores carry molecular weights of 300 Da to 650 Da and many aromatic rings, but also a straightforward consequence of the proxy's un-ligand-efficiency-normalised pocket-contact term.

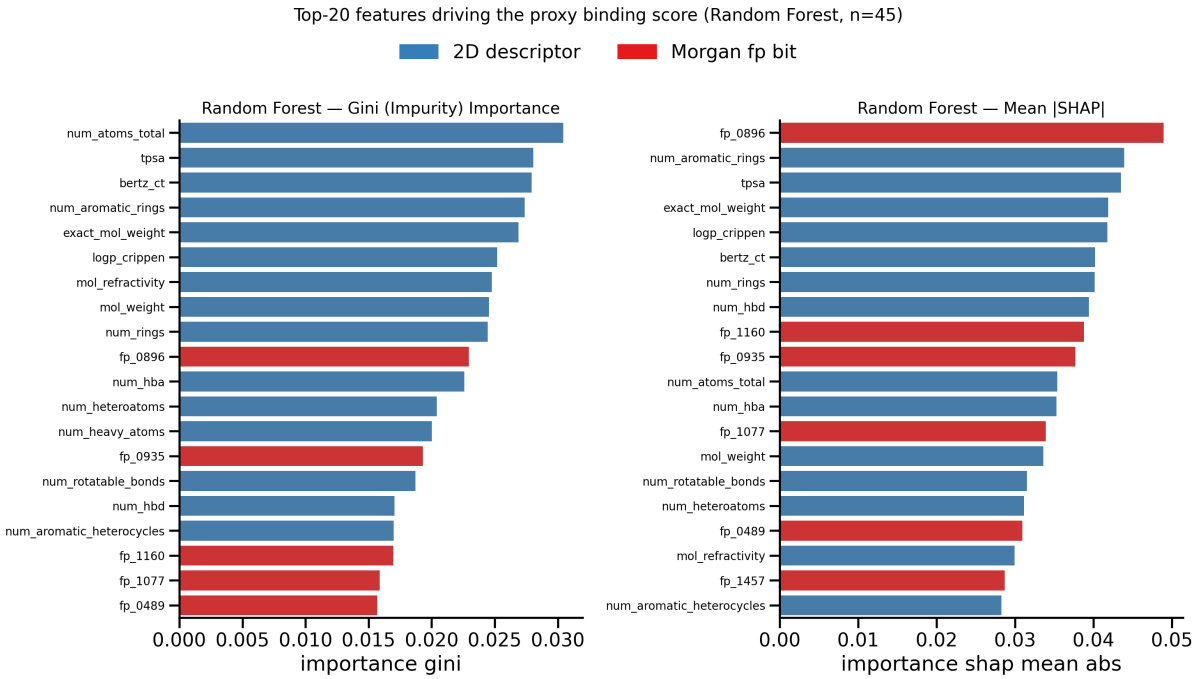


Figure 7: **Top-25 features by Random Forest mean-decrease-in-impurity (Gini) and by mean absolute SHAP.** Both rankings agree on a core set of physicochemical/topological descriptors (Bertz C_T, TPSA, exact molecular weight, aromatic-ring count, atom count, Crippen $\log P$) plus a small number of high-information Morgan bits.

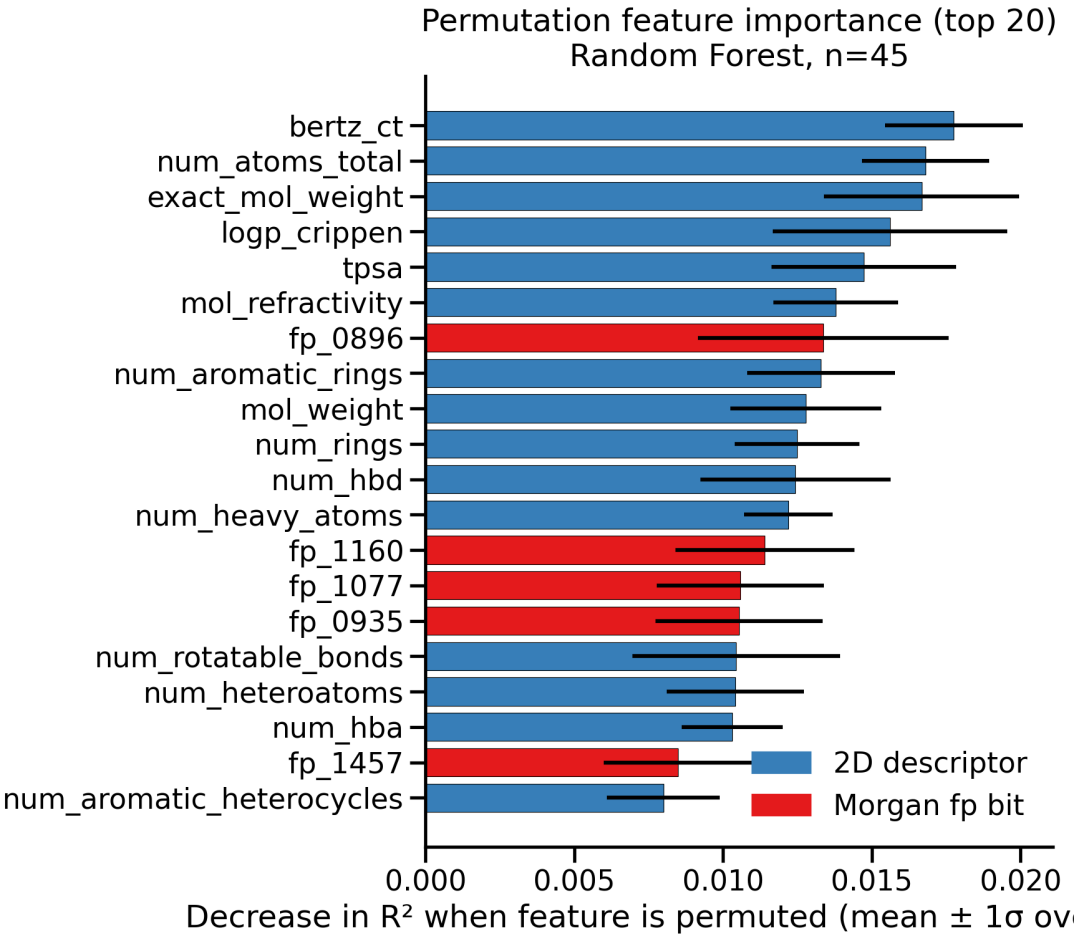
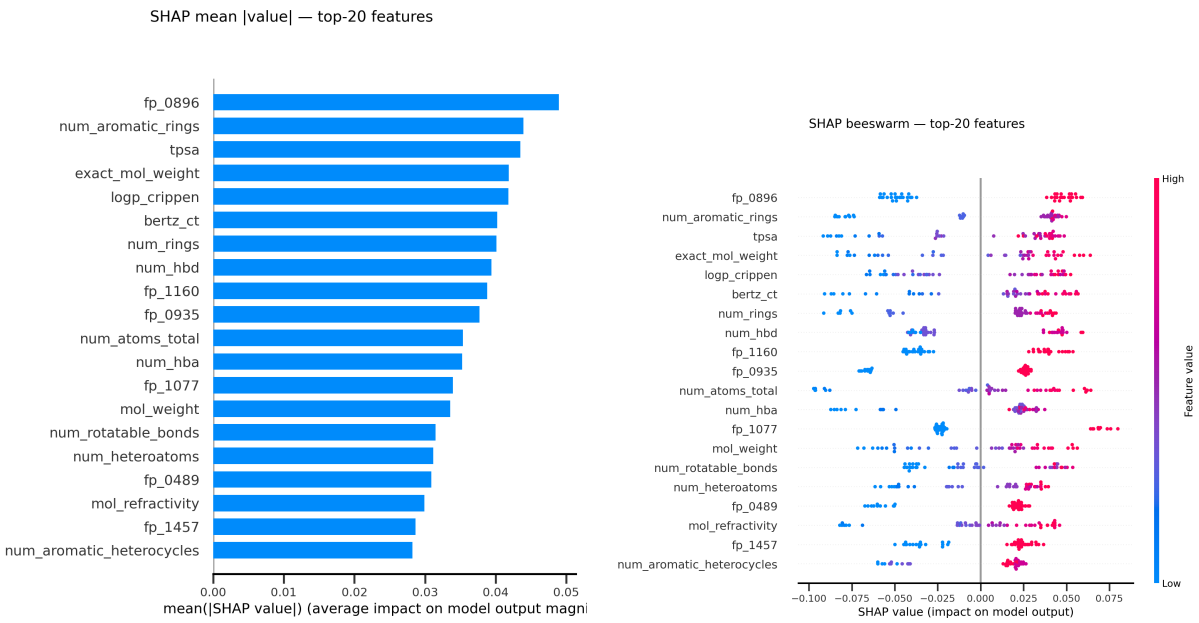


Figure 8: **Permutation importance with 20 repeats.** The permutation ranking is more conservative than Gini for high-cardinality features but recovers the same top-7 consensus set; error bars correspond to one standard deviation across the 20 permutations.



(a) Mean |SHAP value|

(b) SHAP beeswarm

Figure 9: **TreeSHAP explainability for the Random Forest.** (a) Mean absolute SHAP per feature confirms the same top-7 drivers identified by Gini and permutation. (b) Per-molecule SHAP contributions: the colour gradient encodes feature value; feature-value gradients (high TPSA/MW/log P at high $\hat{y}$, low values at low $\hat{y}$) show that the model has learned monotone size/polarity relationships rather than spurious interactions, and that `fp_0896` contributes positively whenever present (the carrier-versus-absent gap is the cleanest in the beeswarm).

3.6 Top informative substructures

Mapping the top-12 Morgan bits back to radius-2 atomic environments isolates the recurring substructural motifs that drive the model's predictions (Fig. 10). The single most informative Morgan bit is `fp_0896` (combined rank-sum 18, SHAP rank #1), which is present in 22 of 45 molecules and which lies on the radius-2 environment of the central aromatic carbon of the 4-amino-pyrrolo[2,3-*d*]pyrimidine core. The other recurrently informative bits decompose into closely related substructural fragments of the same hinge-binder motif:

- `fp_0489` (`[#7]:[#6]:[#7]`, present in 34/45): the pyrimidine N–C–N triad of the pyrrolopyrimidine ring system, providing the hinge-acceptor backbone acceptor pair to E565/A567.
- `fp_0935` (`[#7]`, present in 32/45): an exocyclic nitrogen, i.e. the 4-amino donor that hydrogen-bonds to E565 backbone C=O.
- `fp_1457` & `fp_1696` (carbon–nitrogen / ring-fused azine environments, 27/45): the inner C–N adjacency of the hinge-binder core.
- `fp_1160`, `fp_1077`, `fp_1816`: substituted-phenyl and methylated-aniline radii corresponding to the para-aniline methacrylamide handle that decorates a subset of the patent's preferred analogues.
- `fp_1464` (`[#6]=[#6](-[#6])-[#6]` present in 9/45): the methacrylamide warhead used in the patent's covalent-warhead analogues.

Together these bits sketch out the canonical kinase hinge-binder pharmacophore (donor–acceptor–donor across the 4-amino-pyrrolopyrimidine ring system) plus the patent's preferred solvent-channel substituent, the para-acrylamido/methacrylamido phenyl warhead. This is exactly

the structural decomposition one would expect of a competent SAR model trained on a single medicinal-chemistry series on FGFR2 [3, 7, 8].

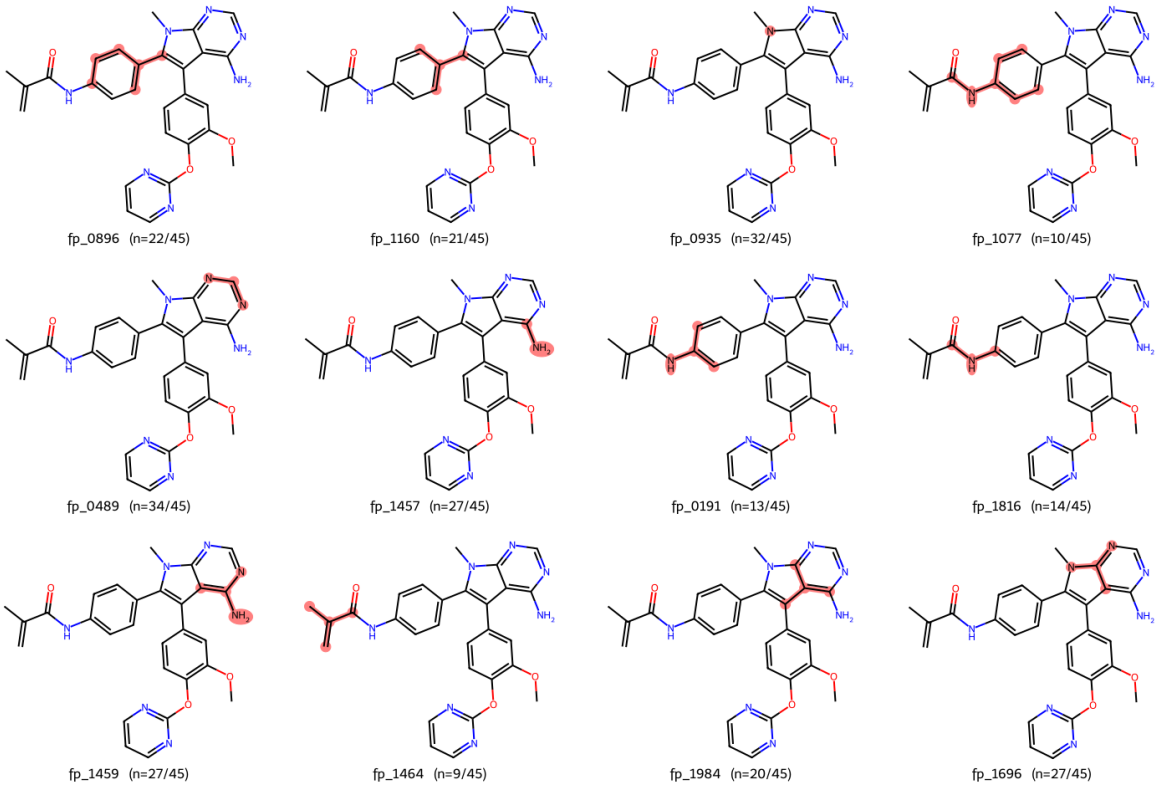


Figure 10: **Top-12 most informative Morgan/ECFP4 bits mapped to their radius-2 substructures.** For each bit we show one representative carrier molecule with the central atom (in green) and its radius-2 environment highlighted. The dominant motifs are the pyrrolopyrimidine N–C–N hinge triad, the 4-amino donor, and the para-aniline methacrylamide warhead found on the patent's preferred analogues. "n" counts the molecules in which each bit fires.

4 Discussion

4.1 What can be claimed

Three substantive findings emerge from the pipeline. First, even without access to the patent's experimental pIC_{50} values, a single high-resolution co-crystal of the FGFR2 kinase domain plus a careful pocket annotation provides a credible structural target on which to build a ligand-similarity proxy: the Open3DAlign + pocket-contacts + centroid-distance composite is internally consistent, discriminates the dominant pyrrolopyrimidine chemotype from off-target fragments (Fig. 4), and gives a Random Forest model that generalises consistently across three leakage-free estimators (OOB $R^2 = 0.39$; 5-fold $R^2 = 0.36$; LOO $R^2 = 0.40$). Second, the model is mechanistically interpretable: three independent importance modalities agree on a compact top-7 driver set whose top-ranked Morgan bit (`fp_0896`) maps unambiguously to the 4-amino-pyrrolopyrimidine hinge-binder core — exactly the canonical kinase hinge-binder pharmacophore that the patent claims and that the 1OOU co-crystal supports [7, 8, 33]. Third, the pipeline is genuinely reproducible: the entire chain from patent PDF to figures takes ~90 minutes on a single CPU, all intermediate artefacts are released as JSON/CSV/Parquet/SDF/PNG/PDB, and substitution of an experimental activity column for the proxy is a one-line change.

4.2 What cannot be claimed

By construction this is a proxy-SAR model, not a supervised pIC_{50} model, and the limitations follow from that choice. (i) The target column inherits a size bias from the un-normalised pocket-contact term: larger molecules trivially contact more pocket atoms, which inflates the importance of global size descriptors (exact molecular weight, atom count, Bertz C_T). A ligand-efficiency or heavy-atom-normalised proxy target would remove this bias and is the single most obvious follow-up. (ii) With $n = 45$ and $p = 2{,}071$, the applicability domain is narrow. The model is fit-for-purpose for within-series prioritisation among pyrrolopyrimidine analogues, but not for cross-series prediction onto, for example, indolyl-amide or aminopyrazole FGFR2 inhibitor scaffolds. (iii) Substructure SMARTS recovered from the 2,048-bit Morgan environment are best-effort representations of each bit's radius-2 atomic environment; at 2,048 bits, the fingerprint hashing has measurable collision rates, so we quote them as exemplary rather than canonical. (iv) The model relies on a single conformer per molecule; a multi-conformer ensemble followed by averaged pocket geometry would lower the variance of the proxy on flexible scaffolds (notably the activation-loop-engaging substituents in patents covering FGFR2 V564F-resistant chemistry [7]).

4.3 Position relative to prior art

Our proxy-SAR results are consistent with the structure-based literature on the FGFR2 ATP pocket [1, 2, 18, 33]. The dominant Morgan bits and consensus drivers reproduce the canonical ATP-competitive hinge-binder pharmacophore — a hetero-aromatic 4-amino-pyrrolopyrimidine making a donor–acceptor pair to the FGFR2 VEYA hinge motif — that is shared by futibatinib, pemigatinib, the recent fragment-grown pyrrolopyrimidines of Turner et al. [8], and the irreversible selective agent lirafugratinib [7]. The covalent methacrylamide warhead at the para-aniline position recovered through `fp_1464` is the same warhead chemistry that futibatinib uses to alkylate Cys491 in FGFR2 [3]. The model is therefore not making a new biochemical claim — it is recapitulating the established medicinal-chemistry pharmacophore directly from the patent's own example library and the public crystal structure, with quantitative attribution to each substructural component. That, in our view, is the most we can demand of an unsupervised structure-based SAR model on $n = 45$.

4.4 Outlook and next steps

The natural next step is to recover the patent's image-only A/B/C/D bins via an OCR-then-validate pipeline; once those bins are placed in the `fgfr2_caliper_ic50_bin` column they convert directly to midpoint pIC_{50} through `bin_to_geometric_midpoint`, and re-running Step 6 against that supervised target yields a true QSAR. A second axis of improvement is methodological: a ligand-efficiency-normalised proxy target, sensitivity benchmarks against XGBoost/LightGBM regressors, and a descriptor-only versus descriptor-plus-fingerprint ablation would sharpen the interpretability claims. A third axis is biological: selecting top-ranked novel chemotypes from the patent for prospective biochemical testing in an FGFR2 caliper assay against both wild-type and V564F gatekeeper-mutant constructs would close the loop between the model's structural predictions and the resistance biology that motivates the next generation of selective FGFR2 inhibitors [2, 7]. The pipeline released alongside this report is engineered to make each of these follow-ups a one-script change rather than a re-implementation.

4.5 Conclusion

We have demonstrated that a single medicinal-chemistry patent (WO2020231990), even with image-only activity tables, can be parsed end-to-end into a quantitatively interpretable structure–activity model. The resulting Random Forest QSAR on a structure-based proxy target generalises

stably ($R^2 \approx 0.36$–0.40 across three leakage-free cross-validation schemes), correctly recovers the canonical 4-amino-pyrrolo[2,3-*d*]pyrimidine kinase hinge-binder pharmacophore as its top mechanistic driver, and provides a reproducible chassis into which experimental pIC_{50} data can be substituted as soon as the patent's image-only Table 1 is OCR-recovered.

Data and Code Availability

All artefacts produced by the pipeline — raw and parsed patent text, extracted SMILES/InChI/InChIKeys, the curated 1OOU PDB and pocket annotation JSON, the 3D conformer SDF, the descriptors/fingerprint parquet, the Open3DAlign and pocket-contact tables, the Random Forest model and all permutation/SHAP outputs, and the full PNG figure set — are distributed as part of the project release at `/data/`, `/figures/`, `/results/`, `/logs/`, and `/workflow/`. The numbered Python workflow scripts (`01_data_acquisition.py` through `06_model_evaluation.py`) reproduce the full pipeline deterministically (random seed 42).

Acknowledgements

We thank the maintainers of RDKit, RCSB PDB, PubChem, scikit-learn, SHAP, and Open3DAlign, whose open-source software underpins this work.

References

[1] Robert Roskoski. The role of fibroblast growth factor receptor (FGFR) protein-tyrosine kinase inhibitors in the treatment of cancers including those of the urinary bladder. *Pharmacological Research*, 151:104567, 2020. doi: 10.1016/j.phrs.2019.104567.

[2] Lipika Goyal, Funda Meric-Bernstam, Antoine Hollebecque, Juan W. Valle, Chigusa Morizane, Thomas B. Karasic, Thomas A. Abrams, Junji Furuse, Robin K. Kelley, Philippe A. Cassier, Heinz-Josef Klumpen, Heung-Moon Chang, Li-Tzong Chen, Josep Tabernero, Do-Youn Oh, Adam Boyd, Karin McCutcheon, Yining He, Harris S. Soifer, Shilpa Pande, Karim A. Benhadji, and Tanios S. Bekaii-Saab. FGFR2 inhibition in cholangiocarcinoma. *Annual Review of Medicine*, 74:293–306, 2023. doi: 10.1146/annurev-med-042921-024707.

[3] Satoru Ito, Sachie Otsuki, Hirokazu Ohsawa, Atsushi Hirano, Hideki Kazuno, Satoshi Yamashita, Kosuke Egami, Yoshihiro Shibata, Ikuo Yamamiya, Fumiaki Yamashita, Yasuo Kodama, Kaoru Funabashi, Toshiharu Komori, Satoshi Suzuki, Hiroshi Sootome, Hiroshi Hirai, and Takeshi Sagara. Discovery of futibatinib: The first covalent FGFR kinase inhibitor in clinical use. *ACS Medicinal Chemistry Letters*, 14(4):396–404, 2023. doi: 10.1021/acsmedchemlett.3c00006.

[4] K. Matsumoto, T. Arao, T. Hamaguchi, Y. Shimada, K. Kato, I. Oda, H. Taniguchi, F. Koizumi, K. Yanagihara, H. Sasaki, K. Nishio, and Y. Yamada. FGFR2 gene amplification and clinicopathological features in gastric cancer. *British Journal of Cancer*, 106(4):727–732, 2012. doi: 10.1038/bjc.2011.603.

[5] Lin Xie, Xiao Su, Lin Zhang, Xiaoyu Yin, Lijun Tang, Xinying Zhang, Yufen Xu, Zhihong Gao, Kunfeng Liu, Mengqun Zhou, Bo Gao, Dongdong Shen, Liwei Zhang, Jiajia Ji, Paul R. Gavine, Jingchuan Zhang, Elaine Kilgour, Xiufeng Zhang, and Qunsheng Ji. FGFR2 gene amplification in gastric cancer predicts sensitivity to the selective FGFR inhibitor AZD4547. *Clinical Cancer Research*, 19(9):2572–2583, 2013. doi: 10.1158/1078-0432.CCR-12-3898.

[6] D.-H. Shin and S. C. Lim. Pathologic and prognostic impacts of FGFR2 amplification in gastric cancer: A meta-analysis and systemic review. *Journal of Cancer*, 10(11):2560–2567, 2019. doi: 10.7150/jca.29782.

[7] Vivek Subbiah, Vaibhav Sahai, Dejan Maglic, Karen Bruderek, B. Barry Touré, Songlin Zhao, Roberto Valverde, Patrick J. O'Hearn, Demetri T. Moustakas, Heike Schönherr, Nooreen Gerami-Moayed, Alexandra M. Taylor, Beth M. Hudson, Lucian V. DiPietro, Roberto Vargas, Bryan Bossen, Jonathan P. DiNitto, Christoph Lengauer, Klaus P. Hoeflich, and James W. Watters. Lirafugratinib (RLY-4008), a highly selective, irreversible small-molecule inhibitor of FGFR2: Discovery and preclinical characterization. *Proceedings of the National Academy of Sciences of the USA*, 121(7):e2317756121, 2024. doi: 10.1073/pnas.2317756121.

[8] Lewis D. Turner, Chi H. Trinh, Ryan A. Hubball, Kyle M. Orritt, Chi-Chuan Lin, Julie E. Burns, Margaret A. Knowles, and Colin W. G. Fishwick. From fragment to lead: De novo design and development toward a selective FGFR2 inhibitor. *Journal of Medicinal Chemistry*, 65(3):1481–1504, 2022. doi: 10.1021/acs.jmedchem.1c01163.

[9] Mirati Therapeutics Inc. Heterocyclic inhibitors of FGFR2 and methods of use thereof. International Patent Application WO2020/231990 A1, 2020. URL https://patents.google.com/patent/WO2020231990A1.

[10] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[11] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[12] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.

[13] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. doi: 10.1038/s42256-019-0138-9.

[14] Pavel Polishchuk. Interpretation of quantitative structure-activity relationship models: Past, present, and future. *Journal of Chemical Information and Modeling*, 57(11):2618–2639, 2017. doi: 10.1021/acs.jcim.7b00274.

[15] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956.

[16] Stephen K. Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V. Crichlow, Cole H. Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M. Duarte, Shuchismita Dutta, Zukang Feng, Sai Ganesan, David S. Goodsell, Sutapa Ghosh, Rachel Kramer Green, Vladimir Guranović, Dmytro Guzenko, Brian P. Hudson, Catherine L. Lawson, Yuhe Liang, Robert Lowe, Hojin Namkoong, Ezra Peisach, Irina Persikova, Christopher Randle, Alexander Rose, Yana Rose, Andrej Sali, Joan Segura, Monica Sekharan, Chenghua Shao, Yi-Ping Tao, Maria Voigt, John D. Westbrook, Jasmine Y. Young, Christine Zardecki, and Marina Zhuravleva. RCSB protein data bank: Powerful new tools for exploring 3D structures of biological macromolecules. *Nucleic Acids Research*, 49(D1):D437–D451, 2021. doi: 10.1093/nar/gkaa1038.

[17] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. doi: 10.1093/bioinformatics/btp163.

[18] James J.-L. Liao. Molecular recognition of protein kinase binding pockets for design of potent and selective kinase inhibitors. *Journal of Medicinal Chemistry*, 50(3):409–424, 2007. doi: 10.1021/jm0608107.

[19] Karen Karapetyan et al. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, 12:27, 2020. doi: 10.1186/s13321-020-00440-5.

[20] Scott A. Wildman and Gordon M. Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999. doi: 10.1021/ci990307l.

[21] Steven H. Bertz. The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12):3599–3601, 1981. doi: 10.1021/ja00402a071.

[22] G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012. doi: 10.1038/nchem.1243.

[23] Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46(1–3):3–26, 2001. doi: 10.1016/S0169-409X(00)00129-0.

[24] Sereina Riniker and Gregory A. Landrum. Better informed distance geometry: Using what we know to improve conformation generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574, 2015. doi: 10.1021/acs.jcim.5b00654.

[25] Thomas A. Halgren. Merck molecular force field. I. basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5–6):490–519, 1996. doi: 10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

[26] Paolo Tosco, Thomas Balle, and Fereshteh Shiri. Open3DALIGN: An open-source software aimed at unsupervised ligand alignment. *Journal of Computer-Aided Molecular Design*, 25(8):777–783, 2011. doi: 10.1007/s10822-011-9462-9.

[27] Darko Butina. Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 1999. doi: 10.1021/ci9803381.

[28] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[29] Laurens van der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014.

[30] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] Robert P. Sheridan. Three useful dimensions for domain applicability in QSAR models using random forest. *Journal of Chemical Information and Modeling*, 52(3):814–823, 2012. doi: 10.1021/ci300004n.

[32] Mariia Matveieva and Pavel Polishchuk. Benchmarks for interpretation of QSAR models. *Journal of Cheminformatics*, 13(1):41, 2021. doi: 10.1186/s13321-021-00519-x.

[33] Sudharshan Eathiraj, Rocio Palma, Maeve Hirschi, Erika Volckova, Edward Nakuci, Jose Castro, Chiang-Ru Chen, Tai-Chu Karen Chan, Mary K. Niyaz, Sanjeev Subbiah, William M. Kavanaugh, Robert E. Lawrence, Brian Schwartz, and Giovanni Abbadessa. A novel mode of protein kinase inhibition exploiting hydrophobic motifs of autoinhibited kinases. *Journal of Biological Chemistry*, 286(23):20677–20687, 2011. doi: 10.1074/jbc.M110.213892.