

A Reproducible Scaffold-Aware AI Workflow for Prioritizing Candidate Ligands for KRAS G12C Switch-II Pocket

Selvaraj Sengottian

selvaraj.sengottian@gmail.com

Ewha Womans University

Research Article

Keywords: KRAS G12C, QSAR, ChEMBL, RDKit, BRICS, AutoDock Vina, molecular docking, virtual screening

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11190700/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

KRAS remains one of oncology's key targets, yet clinically proven inhibition is still constrained by mutant-specific biology, limited pocket depth, and common routes to resistance. This work describes a reproducible, open-source computational workflow meant to sort and prioritize KRAS G12C candidate molecules before anyone commits to synthesis or wet-lab testing. The pipeline ties together the ChEMBL bioactivity collection, RDKit curation, Morgan fingerprints with physicochemical descriptors, random-forest QSAR, BRICS analog creation, medicinal-chemistry filters, AutoDock Vina docking, and a final consensus ranking step. Model building relied on 400 curated KRAS IC₅₀ entries spanning 132 Murcko scaffolds. When evaluation was performed with a scaffold split, performance was $R^2 = 0.512$, RMSE = 0.514, and MAE = 0.381 in pIC₅₀ units, which supports rank-ordering use cases but not precise potency calls. BRICS recombination produced 400 analogs. After structural checks and drug-likeness triage, 57 remained; the top 20 were then docked into the KRAS G12C switch-II pocket. The highest consensus-ranked compound, K1, had a predicted pIC₅₀ of 8.440, a Vina score of -8.737 kcal/mol, and a docked geometry suggesting polar contacts with GLN61, ASP92, and THR58. Though experimental inhibition is unclaimed, instead, the output is a transparent prioritization recipe, an auditable list of candidates, and structure-based ideas to carry into biochemical and cellular follow-up.

1. Introduction

KRAS belongs to the RAS family of small GTPases. It acts as a molecular switch that helps govern proliferation, differentiation, migration, and survival, toggling between an inactive GDP-bound state and an active GTP-bound one^{1,2}. In normal physiology, that cycling is regulated by guanine nucleotide exchange factors, which drive GTP loading, and by GTPase-activating proteins, which speed up GTP hydrolysis and return KRAS to the inactive form³. Mutations that weaken intrinsic hydrolysis or disrupt GAP-mediated hydrolysis can trap KRAS in an on-state, keeping signaling flowing through pathways such as RAF-MEK-ERK and PI3K-AKT-mTOR³⁻⁶, thereby supporting oncogenic transformation. KRAS mutations are frequent in pancreatic ductal adenocarcinoma, colorectal adenocarcinoma, and non-small-cell lung cancer, so KRAS remains one of the most clinically significant oncogenes in precision oncology^{2,4,6}. For years, it was treated as difficult to drug. The reasons are familiar: very tight GDP/GTP binding, a protein surface that often appears smooth and uninviting, and flexible conformations that complicate ligand design^{2,3,7,8}. The identification of a cryptic switch-II pocket in KRAS G12C shifted that outlook and opened the door to covalent inhibitors that trap the mutant in its inactive GDP-bound conformation^{5,9}. Clinical approvals of sotorasib and adagrasib have confirmed KRAS G12C as a tractable target^{10,11}.

The clinical validation of KRAS G12C has been driven by the development of covalent inhibitors that selectively target the mutant cysteine residue within the switch-II pocket. Sotorasib (AMG 510), the first FDA-approved KRAS G12C inhibitor, demonstrated clinically meaningful activity in patients with KRAS G12C-mutant non-small-cell lung cancer and established KRAS as a druggable therapeutic target¹².

Adagrasib (MRTX849) subsequently showed durable antitumour activity and received regulatory approval for selected KRAS G12C-driven malignancies. Several next-generation inhibitors, including divarasib (GDC-6036), JDQ443, olomorasib (LY3537982), and D-1553, are currently under clinical evaluation to improve potency, selectivity, and overcome acquired resistance. Representative KRAS G12C inhibitors, together with their reported biochemical activities (IC₅₀ values), development status, and corresponding chemical structures, are summarised in Table 1 and Supporting Information Figure S1.

Table 1
Representative KRAS G12C inhibitors reported in the literature.

Inhibitor	Developer	Reported biochemical IC ₅₀ [*]	Binding mechanism	Development status [†]	Primary reference(s)
Sotorasib (AMG 510)	Amgen	47 nM	Covalent inhibitor targeting the switch-II pocket (Cys12)	FDA approved	Canon <i>et al.</i> ¹³ (2019); Hong <i>et al.</i> ¹² (2020)
Adagrasib (MRTX849)	Mirati Therapeutics (Bristol Myers Squibb)	89 nM	Covalent inhibitor targeting the switch-II pocket (Cys12)	FDA approved	Hallin <i>et al.</i> ¹⁴ (2020); Jänne <i>et al.</i> ¹⁵ (2022)
Divarasib (GDC-6036)	Genentech/Roche	Low nanomolar	Covalent inhibitor targeting the switch-II pocket (Cys12)	Clinical development	Fell <i>et al.</i> ¹⁶ (2023)
JDQ443	Novartis	Low nanomolar	Covalent inhibitor targeting the switch-II pocket (Cys12)	Clinical development	Bange <i>et al.</i> ¹⁷ (2022)
Olomorasib (LY3537982)	Eli Lilly	Low nanomolar	Covalent inhibitor targeting the switch-II pocket (Cys12)	Clinical development	Fakih <i>et al.</i> ¹⁸ (2023)
D-1553	InventisBio	Low nanomolar	Covalent inhibitor targeting the switch-II pocket (Cys12)	Clinical development	Wang <i>et al.</i> ¹⁹ (2022)

*Reported biochemical IC₅₀ values are representative values obtained from the cited primary literature. Because different studies employ different biochemical assays, protein constructs, and experimental conditions, the values should be interpreted as approximate and are intended for general comparison rather than direct quantitative ranking.

† Development status reflects the status of each compound at the time of manuscript preparation.

Although the generated analogues described in this study originate from BRICS recombination of active KRAS-associated chemistry, they should not be interpreted as direct analogues of clinically approved KRAS G12C inhibitors such as sotorasib or adagrasib. Instead, they represent computationally generated variants intended to explore nearby regions of chemically relevant KRAS inhibitor space while preserving medicinal-chemistry plausibility. Experimental studies will ultimately be required to determine whether these compounds represent genuinely distinct chemotypes or modified derivatives with similar binding characteristics. Even with that progress, early discovery around KRAS remains hard. The continued need for improved KRAS G12C inhibitors has stimulated extensive efforts in structure-based drug design, virtual screening, molecular docking, and artificial intelligence-assisted discovery workflows.²⁰ G12C is only one slice of the broader KRAS-mutant landscape; resistance still emerges, and medicinal chemistry groups must decide what to make next under real-time constraints and budget limits. Modern computer-aided drug discovery (CADD) integrates complementary computational techniques, including ligand-based virtual screening, quantitative structure–activity relationship (QSAR) modelling, molecular docking, pharmacokinetic and toxicity prediction (ADMET), and molecular simulations, to accelerate the identification and prioritization of promising drug candidates²¹. Rather than relying on a single computational method, integrated CADD workflows combine multiple sources of evidence to improve decision-making during early-stage drug discovery, reduce experimental cost and time, and increase the likelihood of selecting chemically tractable candidates for subsequent biological evaluation²². Recent studies have demonstrated that combining ligand-based prediction, structure-based docking, and in silico ADMET assessment provides a more reliable framework for compound prioritization than individual computational methods alone²³.

Computational screening can narrow options, as long as the results are read with restraint. ChEMBL offers curated activity data that can support model development;²⁴ RDKit supplies the core chemistry toolset for cleaning, descriptors, fingerprints, and scaffold handling²⁵; QSAR can learn ligand-based trends across related series^{26,27}, and docking can add a second perspective by checking whether shapes and functional groups plausibly fit the pocket²⁸. Docking scores, though, are not direct stand-ins for binding affinity or cellular response. They depend on how the receptor was prepared, protonation choices, protein conformation, solvent treatment, and what assumptions²⁹. Recent computational studies have therefore emphasized combining ligand-based machine learning with structure-based methods to improve prioritization reliability while maintaining interpretability.³⁰ The scoring function makes. So docking works best as one component of a broader, clearly documented prioritization strategy that also includes careful data curation, scaffold-aware validation, medicinal-chemistry filters,

and strict limits on interpretation. This is an executable Python-based workflow for KRAS G12C candidate prioritization. The point is practical integration, not a standalone demonstration: data retrieval, structure cleanup, descriptor generation, validation, analog generation, filtering, docking, consensus scoring, and file-level reporting are integrated into a single pipeline. The emphasis is on a human-auditable decision trail, where it is possible to see why a molecule survived triage, how far the model can be trusted beyond the training data, and what structural contacts are worth testing experimentally. The output should be read as hypotheses for KRAS lead discovery, not validated inhibitors. A key contribution is the full disclosure of the prioritization route for KRAS G12C candidates. Rather than presenting only docking ranks or a generic AI screen, the workflow combines scaffold-split QSAR validation with BRICS-based analog expansion and a weighted ligand- and structure-based scoring scheme. This setup speaks to a recurring early problem in medicinal chemistry: selecting a small, chemically interpretable subset of compounds for follow-up while preserving sufficient provenance for independent reproduction and critique. The disease focus is KRAS-mutant cancer, especially tumors carrying the G12C substitution, where target validation exists, but resistance and incomplete responses still push the need for next-generation chemical matter, which remains a challenge. As a result, the approach prioritizes the early identification of compounds that are chemically interpretable, computationally traceable, and suitable for further testing in KRAS G12C biochemical and cellular models. The following explanations supplement the study's data-to-decision logic (Supporting Information, Section S1). Each dataset, model output, plot, and ranking file is interpreted in accordance with its role in the workflow: raw bioactivity data form the ligand-based evidence base; curated and scaffold-aware model validation defines the reliability envelope; BRICS-generated analogs investigate nearby chemical space; medicinal-chemistry filters eliminate unsuitable candidates; and docking plus consensus scoring converts the candidate set into a practical experimental queue.

The present work should therefore be understood as a target-focused application and reproducibility study, rather than as the introduction of a new chemoinformatics algorithm. Its central question is whether a sequence of familiar computational steps can be assembled into a transparent, scaffold-aware decision process that produces a manageable experimental queue. That distinction matters. The ChEMBL records provide an empirical starting point, but they may span assay formats, protein constructs, and KRAS mutation states; the ligand-based model is consequently described here as learning from KRAS-associated activity data. The receptor-based stage is the portion specifically directed toward the KRAS G12C switch-II pocket. The two evidence streams are combined to organize follow-up work, not to convert computational scores into proof of inhibition. The study's value lies in the disclosed data trail, scaffold-conscious evaluation, candidate attrition record, and the explicit connection between model uncertainty and decisions about which molecules to test next. Compared with many previously reported KRAS virtual-screening studies that primarily emphasize either molecular docking or machine-learning prediction alone, the present workflow integrates curated ChEMBL bioactivity data, scaffold-aware QSAR validation, BRICS-based analogue generation, medicinal chemistry filtering, molecular docking, consensus scoring, and comprehensive reporting of computational outputs within a single reproducible framework. Therefore, the principal contribution of this study is not the introduction

of a new predictive algorithm, but the transparent integration of established computational methods into an auditable workflow that enables each prioritization step to be independently reproduced and critically evaluated.

1.1 Research Rationale

Despite the clinical success of KRAS G12C inhibitors such as sotorasib and adagrasib, the discovery of next-generation inhibitors remains challenging because of acquired resistance and the need to efficiently explore new chemical space. Computational approaches can help prioritize compounds before experimental testing, but many reported workflows lack transparent validation and reproducibility. The present study was designed to establish a reproducible, scaffold-aware computational workflow that integrates curated bioactivity data, machine-learning prediction, BRICS-based analogue generation, medicinal chemistry filtering, molecular docking, and consensus scoring. The objective is not to identify experimentally validated inhibitors, but to provide a practical and transparent framework for prioritizing KRAS G12C candidate molecules for subsequent biochemical and cellular evaluation.

2. Results and Discussion

2.1. Workflow and model validation

The pipeline began with KRAS IC₅₀ records from ChEMBL. Structures were canonicalized, invalid entries were removed, IC₅₀ values were converted to pIC₅₀, and duplicate records were consolidated. The resulting curated dataset was intentionally limited to 400 molecules to provide a transparent, computationally manageable, and fully reproducible benchmark dataset for the present workflow rather than to maximize the size of the training set. Molecules were encoded with radius-2 Morgan fingerprints and RDKit descriptors. Performance was checked under both a random split and a Murcko-scaffold split. The scaffold split was treated as the main estimate of generalization, since it better resembles the situation of predicting on new analog series.

Table 2
Random-forest QSAR validation metrics.

Split	R ²	RMSE	MAE	n_test
Random split	0.536	0.472	0.358	80
Scaffold split	0.512	0.514	0.381	108

The scaffold-split R² of 0.512 indicates that the model retained a meaningful ranking signal when evaluated on structurally separated compounds. The MAE of 0.381 pIC₅₀ units also defines an important boundary: compounds with very similar predicted values should not be overinterpreted as having reliably different potencies. For this reason, the model was used to enrich and prioritize candidates rather than to make definitive activity claims for individual molecules.

Table 2 defines the reliability envelope for the computational screen (Supporting Information, Section S4). The random split estimates interpolation within the available KRAS chemical series, whereas the scaffold split provides a stricter test of performance on less familiar scaffolds. Supporting Figures S2-S4 summarize the pIC₅₀ distribution and parity plots for random- and scaffold-split validation. Together, these analyses support the model's use as an early-stage triage tool and underscore that experimental testing remains necessary before any compound can be described as an active KRAS inhibitor. The pIC₅₀ distribution plot explains the activity range available to the model after curation (Supporting Information, Section S3 and Figure S2). The curated training set contains 400 molecules with pIC₅₀ values ranging from 7.208 to 10.000, with a median of 7.854 and a mean of 8.003. This distribution shows that the model was trained mainly on moderately to highly active KRAS compounds rather than on a broad inactive-active classification set. Consequently, the model is best interpreted as a local rank-ordering tool for KRAS-like analogs, not as a universal classifier for all chemical matter. The parity plots provide a visual check on how predicted pIC₅₀ values track observed pIC₅₀ values (Supporting Information, Section S4 and Figures S3-S5). The random-split parity plot reflects interpolation within related chemistry. In contrast, the scaffold-split parity plot is the stricter test because the test molecules belong to scaffolds separated from the training scaffolds. The scaffold-split R² of 0.512, RMSE of 0.514, and MAE of 0.381 pIC₅₀ units show useful enrichment ability but also justify the cautious interpretation used throughout the manuscript.

Several features of the validation results deserve emphasis before the generated compounds are considered. The curated response values occupy a relatively narrow, active interval, and the predictions for the leading generated molecules are much more tightly clustered than those of the scaffold-split MAE. The model can therefore support coarse enrichment, but it cannot justify a precise potency order among candidates whose predictions differ by only a few hundredths of a pIC₅₀ unit. In practical use, those candidates should be treated as a high-priority group and then separated by chemical diversity, synthetic feasibility, availability, and experimental design. A future prospective implementation should also report a similarity-based applicability domain, repeated scaffold splits, and prediction intervals or another calibrated uncertainty measure. These additions would show when the training chemistry sufficiently represents a generated structure to warrant model-based prioritization.

2.2. Candidate Generation, Filtering, and Docking

BRICS fragmentation and recombination were used to generate analogs from active seed molecules because the method preserves chemically meaningful disconnections while exploring nearby analog space (Supporting Information, Section S5).^{31,32} Generated candidates were filtered for common atom types, charge state, halogen valence, PAINS liability, physicochemical ranges, ring count, rotatable bonds, fraction sp³ carbon, and QED. These filters were applied as practical medicinal-chemistry triage criteria, not as absolute measures of safety, selectivity, or developability.^{33,34}

Filtered candidates were prepared with Open Babel and docked with AutoDock Vina (Supporting Information, Section S6).³⁵⁻³⁷ The docking box was inferred from the local KRAS G12C structure and

held constant across all 20 prioritized candidates to ensure comparable rankings. Final ranking combined standardized predicted pIC_{50} and standardized docking affinity, assigning 60% weight to the QSAR prediction and 40% to docking. This weighting intentionally favored the experimentally trained ligand-based signal while still allowing docking to penalize candidates with less favorable predicted pocket accommodation. The candidate-generation data also define the attrition profile of the screen (Supporting Information, Section S5). BRICS recombination produced 400 analogs, all labeled as BRICS-derived structures in generated.csv. After sanitization, structural checks, and medicinal-chemistry filters, 57 candidates remained in generated_predictions.csv. This reduction from 400 to 57 candidates is important because it shows that the workflow did not simply rank every generated molecule; it first removed structures that were less suitable for meaningful prediction or practical follow-up. The final top_hits.csv file contains the 20 docked candidates and records; for each molecule, the predicted pIC_{50} , Vina docking score, positive docking-affinity transform, standardized ligand-based score, standardized docking score, and final consensus score (Supporting Information, Section S6). Among the top 20, predicted pIC_{50} values are tightly clustered between 8.380 and 8.446, whereas Vina scores span -9.143 to -7.748 kcal/mol. Because these predicted activities differ by substantially less than the scaffold-split MAE (0.381 pIC_{50} units), the ligand-based component primarily functions as a coarse prioritization filter within the top-ranked candidates, whereas differences in docking scores provide most of the discrimination within this prioritized experimental tier.

The docking component is intentionally used as a secondary line of evidence. A single prepared receptor, a single search region, and a common Vina protocol make scores comparable within this run. Still, they do not account for the full conformational behavior of KRAS, explicit solvent effects, alternative protonation states, or induced fit. In addition, the current protocol was not benchmarked by redocking a co-crystallized ligand and did not explicitly model covalent bond formation with CYS12. The reported poses and residue contacts are consequently geometric hypotheses. They are useful for choosing compounds and planning mutational, binding, or structural experiments, but they should not be interpreted as measured binding modes or free energies. Consequently, compounds lacking an appropriate electrophilic warhead for covalent engagement with Cys12 should not be interpreted as having the same therapeutic potential as clinically approved covalent KRAS G12C inhibitors, and experimental validation will be required to determine their actual binding mechanism and biological activity. The 60:40 QSAR-docking weighting likewise represents a transparent prioritization rule rather than an empirically optimized relationship with biological response. The selected 60:40 weighting should therefore be viewed as an interpretable prioritization rule rather than an optimized predictive model, and alternative weighting schemes should be evaluated in future studies.

2.3. KRAS G12C Binding Poses and Interaction Profiles

K1 occupied the switch-II pocket region and positioned its aromatic core near GLN99, GLU63, GLN61, ASP92, TYR96, THR58, MET72, and VAL9. The predicted polar contacts were GLN61 OE1-K1 N at 3.07 Å, ASP92 OD1-K1 N at 3.07 Å, and THR58 OG1-K1 O at 3.29 Å. Close hydrophobic or van der Waals contacts were observed with GLU63, GLN99, HIS95, and CYS12. These interactions are consistent with a

plausible docked pose, but they should be interpreted as geometry-derived hypotheses until confirmed experimentally.

Figure 1 links the numerical ranking to a chemically interpretable binding hypothesis. K1 spans a shallow switch-II pocket region rather than a deep enclosed cavity, which is consistent with the known difficulty of KRAS ligand discovery. The interaction panel is intentionally descriptive: the listed contacts are distance-based interactions from the docked coordinates. They should guide follow-up experiments rather than be treated as validated binding determinants. K1 is emphasized because it is the highest-consensus-ranked candidate, rather than the single strongest docking hit. Its predicted pIC_{50} of 8.440 places it near the top of the filtered analog set, and its Vina score of -8.737 kcal/mol indicates favorable predicted pocket accommodation. The combination of GLN61, ASP92, and THR58 polar contacts with nearby hydrophobic or van der Waals contacts provides a reasonable structural hypothesis for testing, while still avoiding any claim that the molecule is experimentally active.

Figure 2 compares the 20 prioritized candidates using the same receptor representation; complementary two-dimensional structures are provided in the Supporting Information (Section S7 and Figure S6). The figure is retained in the main manuscript because it allows direct visual comparison of how the analog series samples the KRAS G12C pocket. Several ligands maintain related aromatic placement, whereas substituent orientation and residue proximity vary across the series; these differences provide a rational starting point for selecting compounds for synthesis, purchase, or focused redesign.

Table 3
Summary of predicted interactions for 20 prioritized KRAS G12C candidates.

ID	Pred. pIC50	Vina	Final	Polar contacts	Hydrophobic/VDW contacts
K1	8.440	-8.737	1.506	GLN61 OE1-N 3.07 Å; ASP92 OD1-N 3.07 Å; THR58 OG1-O 3.29 Å	GLU63 3.28 Å; GLN99 3.57 Å; HIS95 3.71 Å; CYS12 3.94 Å
K2	8.446	-8.044	1.055	ASP92 OD2-O 3.09 Å; THR58 OG1-O 3.11 Å; TYR96 OH-O 3.11 Å	GLU63 3.27 Å; HIS95 3.46 Å; GLN99 3.61 Å; GLN61 3.63 Å; ALA11 3.70 Å; CYS12 3.85 Å
K3	8.441	-7.984	0.820	GLY60 O-O 2.87 Å; ARG68 NH1-O 3.00 Å; GLN61 OE1-O 3.25 Å	GLN99 3.33 Å; TYR96 3.34 Å; HIS95 3.40 Å; ALA11 3.58 Å; GLU63 3.72 Å; THR58 3.84 Å
K4	8.406	-8.773	0.470	GLY60 O-O 3.08 Å; ARG102 NH1-N 3.12 Å; GLU98 OE1-N 3.16 Å; GLN99 NE2-N 3.37 Å; GLN61 O-N 3.40 Å	TYR96 3.49 Å; MET72 3.49 Å; HIS95 3.86 Å
K5	8.400	-8.931	0.404	THR58 OG1-O 3.02 Å	GLU63 3.28 Å; GLN61 3.42 Å; HIS95 3.43 Å; GLN99 3.46 Å; TYR96 3.56 Å; GLU62 3.80 Å
K6	8.410	-8.519	0.330	CYS12 N-N 3.15 Å; THR58 OG1-O 3.21 Å; TYR96 OH- N 3.23 Å	GLU63 3.13 Å; GLN99 3.37 Å; HIS95 3.45 Å; GLU62 3.75 Å
K7	8.413	-8.365	0.299	THR58 OG1-O 3.12 Å; ASP92 OD2-O 3.22 Å; GLU62 O-N 3.42 Å	GLN99 2.94 Å; TYR96 3.40 Å; HIS95 3.66 Å; VAL103 3.95 Å; MET72 3.98 Å
K8	8.411	-8.405	0.282	THR58 OG1-O 3.10 Å; ASP92 OD2-N 3.25 Å; GLU62 O-N 3.45 Å	GLN99 2.91 Å; TYR96 3.46 Å; HIS95 3.64 Å; VAL103 3.90 Å; GLU63 3.97 Å
K9	8.411	-8.411	0.270	ARG102 NH1-O 2.99 Å; GLY60 O-O 3.21 Å; GLN61 O-N 3.43 Å; GLU62 O-N 3.44 Å	GLN99 3.15 Å; TYR96 3.44 Å; HIS95 3.65 Å; GLU63 3.78 Å; MET72 3.87 Å
K10	8.403	-8.602	0.188	THR58 OG1-O 3.03 Å	GLU63 3.24 Å; GLN99 3.36 Å; TYR96 3.48 Å; HIS95 3.56 Å; GLN61 3.72 Å; GLU62 3.82 Å
K11	8.386	-8.972	0.001	ARG102 NH1-O 3.08 Å; GLY60 O-O 3.13 Å; GLU98 OE1-O 3.15 Å; GLN99 NE2-O 3.30 Å; GLN61 O-N 3.46 Å	TYR96 3.51 Å; MET72 3.57 Å; GLU63 3.68 Å; HIS95 3.92 Å
K12	8.380	-9.143	-0.012	GLY60 O-O 3.22 Å; GLN99 NE2-O 3.32 Å; GLN61 O-N 3.48 Å	TYR96 3.40 Å; MET72 3.53 Å; ARG102 3.56 Å; GLU98 3.64 Å; HIS95 3.75 Å; GLU63 3.93 Å

ID	Pred. pIC ₅₀	Vina	Final	Polar contacts	Hydrophobic/VDW contacts
K13	8.408	-7.956	-0.254	THR58 OG1-O 3.05 Å	GLU63 3.22 Å; TYR96 3.48 Å; HIS95 3.61 Å; GLN99 3.70 Å; GLU62 3.90 Å
K14	8.404	-7.839	-0.489	GLY60 O-O 3.17 Å; ASP92 OD1-O 3.24 Å; GLN61 O-N 3.27 Å; GLU62 O-N 3.40 Å	MET72 3.29 Å; HIS95 3.63 Å; GLN99 3.67 Å; TYR96 3.85 Å
K15	8.404	-7.748	-0.576	GLN99 NE2-O 3.04 Å; GLU63 OE1-O 3.09 Å; GLY60 O-O 3.19 Å; GLN61 O-N 3.29 Å	TYR96 3.52 Å; HIS95 3.71 Å
K16	8.384	-8.260	-0.725	THR58 OG1-O 3.09 Å; GLU62 O-N 3.33 Å; GLN61 O-O 3.49 Å	GLN99 2.83 Å; TYR96 3.46 Å; HIS95 3.55 Å; GLU63 3.72 Å; VAL103 3.88 Å
K17	8.380	-8.287	-0.820	GLN61 NE2-O 3.16 Å; THR58 OG1-O 3.20 Å; ASP92 OD1-O 3.24 Å; TYR96 OH-N 3.49 Å	GLU63 3.08 Å; LYS88 3.30 Å; HIS95 3.62 Å; GLN99 3.83 Å
K18	8.394	-7.789	-0.839	GLY60 O-O 3.17 Å; GLN61 O-N 3.32 Å	GLN99 3.09 Å; GLU63 3.25 Å; TYR96 3.54 Å; HIS95 3.72 Å
K19	8.392	-7.872	-0.854	GLY60 O-O 2.92 Å; ARG68 NH1-O 2.97 Å; GLN61 OE1-O 3.14 Å	TYR96 3.25 Å; HIS95 3.35 Å; GLN99 3.39 Å; ALA11 3.62 Å; GLU63 3.80 Å; THR58 3.89 Å
K20	8.386	-7.838	-1.056	GLN99 NE2-O 2.97 Å; GLY60 O-O 3.17 Å; GLN61 O-N 3.30 Å	GLU63 3.25 Å; TYR96 3.49 Å; HIS95 3.61 Å

Across the 20 candidates, docked poses repeatedly placed ligand scaffolds near GLN61, GLU62, GLU63, THR58, TYR96, ASP92, GLN99, and VAL9, suggesting that the generated analogs explored related regions of the KRAS G12C switch-II pocket. K12 showed the strongest Vina score among the candidates, whereas K1 ranked highest by consensus because it balanced high predicted potency with favorable docking. This difference illustrates why consensus ranking is more informative than relying on a single docking score. The compound-level trends suggest that the generated analogs repeatedly occupy a related switch-II pocket region but differ in substituent orientation, polar-contact pattern, and docking-score contribution. K12 provides the most favorable Vina score (-9.143 kcal/mol) but ranks lower in the consensus because its predicted pIC₅₀ is lower than those of K1-K3. This behavior is consistent with the intended workflow design: docking is used to refine and contextualize ligand-based prioritization, not to override the bioactivity-trained model. Although K1 achieved the highest consensus score, the predicted pIC₅₀ values of the leading candidates fall within the scaffold-split prediction error. Therefore, K1-K20 should be regarded as a prioritized experimental set rather than a precisely ranked potency series. Table 3 is the central compound-level result in the main manuscript. It reports the consensus identifiers

K1-K20, predicted activity, docking score, final ranking score, and principal distance-based contacts. The Supporting Information (Table S2) provides the data provenance, validation plots, and output files needed to audit or repeat the analysis.

2.4. ADMET profiling of the top-ranked compounds

To further evaluate the drug-likeness and developability of the prioritized compounds, in silico ADMET profiling was performed for the twenty highest-ranked candidates obtained from the integrated machine-learning and molecular docking workflow. Drug-likeness and pharmacokinetic properties, including Lipinski's rule of five, gastrointestinal (GI) absorption, blood-brain barrier (BBB) permeability, and cytochrome P450 (CYP)-related information, were obtained using SwissADME³⁸. Ames mutagenicity and hERG cardiotoxicity were evaluated using ADMETlab 3.0³⁹. As summarized in Table 4, all twenty compounds satisfied Lipinski's rule of five, suggesting favorable drug-like physicochemical properties. Most compounds exhibited either high or low predicted GI absorption, whereas only one compound was predicted to permeate the blood-brain barrier. CYP inhibition information was reported as not determined (n/d) in the current SwissADME output and therefore was not interpreted further in this study. ADMETlab predictions further provided an initial assessment of Ames mutagenicity and hERG-related cardiotoxicity, supplying additional information for prioritizing candidates for future experimental evaluation. Overall, the ADMET analysis complements the machine-learning predictions and molecular docking results by providing an early assessment of pharmacokinetic behavior and potential toxicity, thereby improving confidence in prioritizing candidate KRAS G12C inhibitors.

Table 4
In silico ADMET profiling of the top 20 prioritized KRAS G12C inhibitor candidates.

Compound	Lipinski	GI Absorption	BBB Permeant	CYP Inhibition	Ames	hERG
K1	Pass	Low	No	n/d	Negative	Low risk
K2	Pass	High	No	n/d	Negative	Low risk
K3	Pass	High	No	n/d	Positive	High risk
K4	Pass	High	No	n/d	Positive	High risk
K5	Pass	Low	No	n/d	Negative	High risk
K6	Pass	Low	No	n/d	Negative	High risk
K7	Pass	Low	No	n/d	Negative	High risk
K8	Pass	High	No	n/d	Positive	High risk
K9	Pass	Low	No	n/d	Negative	High risk
K10	Pass	Low	No	n/d	Positive	High risk
K11	Pass	Low	No	n/d	Negative	High risk
K12	Pass	Low	No	n/d	Negative	Low risk
K13	Pass	High	No	n/d	Negative	High risk
K14	Pass	High	Yes	n/d	Positive	High risk
K15	Pass	Low	No	n/d	Negative	High risk
K16	Pass	Low	No	n/d	Negative	High risk
K17	Pass	Low	No	n/d	Negative	High risk
K18	Pass	High	No	n/d	Positive	High risk
K19	Pass	Low	No	n/d	Positive	High risk
K20	Pass	High	No	n/d	Negative	High risk

K1–K20 correspond to the twenty highest-ranked KRAS G12C inhibitor candidates listed in top_hits.csv in descending order of the integrated final score. Drug-likeness and pharmacokinetic properties (Lipinski's rule of five, GI absorption, BBB permeability, and CYP inhibition) were predicted using SwissADME. Ames mutagenicity and hERG cardiotoxicity were predicted using ADMETlab 3.0. The corresponding SMILES structures and prediction scores are provided in Supporting Information Table S1.

2.5. Validation and Scope of Interpretation

The present study uses computational validation appropriate for an early virtual-screening workflow: duplicate-aware data curation, scaffold-aware train/test splitting, parity-plot inspection, medicinal-

chemistry filtering, and structure-based pose analysis. These steps reduce the risk of overinterpreting training-set memorization or docking-score artifacts, but they do not replace biochemical validation. The prioritized molecules should be advanced only after orthogonal testing, including KRAS G12C binding assays, nucleotide-exchange or pathway-readout assays, cellular antiproliferative assays in matched KRAS-dependent models, selectivity profiling, and, where possible, structural confirmation of the binding mode. The Supporting Information therefore reports the data and explanations of the plots needed to audit the computational claims (Sections S1-S8). It identifies the raw ChEMBL-derived input, the curated 400-molecule modeling set, the generated BRICS analog table, the 57 filtered prediction table, the 20-candidate docking and consensus table, the model metrics file, the run-summary JSON file, the docking logs, and the plot files. This organization allows the reader to trace each numerical result in the manuscript back to a specific machine-readable output.

Accordingly, the manuscript frames K1-K20 as a ranked experimental queue rather than a list of confirmed inhibitors. This distinction is important for both scientific accuracy and translational relevance: the workflow is intended to save time in candidate selection, not to bypass the experimental evidence required for a therapeutic claim.

The most informative experimental extension would begin with a diversity-aware subset rather than only the first few numerical ranks. A primary KRAS G12C biochemical binding or nucleotide-exchange assay should be paired with appropriate positive and negative controls. Compounds showing reproducible activity could then be examined in matched KRAS G12C-dependent and control cell lines using pathway readouts such as pERK suppression, followed by cytotoxicity and selectivity counterscreens. For candidates designed to engage CYS12, direct evidence of covalent engagement would be especially important. Comparing those prespecified experimental outcomes with the current ranking would provide a genuine prospective test of the workflow and reveal which components of the consensus score are informative. Until such evidence is available, K1-K20 remain candidate structures proposed for testing; they are not experimentally confirmed inhibitors or therapeutic leads.

Another limitation of the present workflow is the absence of a formal applicability-domain analysis. Consequently, prediction confidence for individual generated compounds cannot be quantitatively assessed. Future work should incorporate similarity-based applicability-domain methods, prediction uncertainty estimation, and confidence intervals to determine whether newly generated analogues fall within the chemical space represented by the training dataset.

3. Conclusion

A transparent, reproducible computational pipeline is described for prioritizing KRAS G12C candidates by combining public bioactivity curation, scaffold-aware machine-learning validation, medicinal-chemistry filtering, BRICS analog generation, and molecular docking. With 400 curated KRAS activity records, the random-forest model retained a useful predictive signal under scaffold-split validation, supporting its role as a ranking tool rather than a precise potency predictor. The workflow reduced 400 generated

analogues to 20 docked candidates, placing K1 at the top of the consensus list while K12 produced the strongest Vina score. Recurrent predicted contacts involving residues such as GLN61, THR58, ASP92, TYR96, and GLN99 give a readable structural basis for follow-up experiments. The main value here is not a claim of immediate therapeutic discovery; it is a reproducible route from public KRAS data to a focused set of candidate compounds for testing. Next steps should include biochemical and cellular assays for the prioritized compounds, selectivity and covalent engagement checks where relevant, and iterative refinement of the QSAR and docking parts of the pipeline as new data arrive.

Taken together, the analysis offers a reproducible computational starting point and a clearly defined set of next experiments. Its strongest claim is procedural: public bioactivity data, scaffold-aware validation, chemically constrained analogue generation, and structure-based inspection can be connected in a way that another researcher can audit and rerun. Its biological claims remain deliberately provisional. This separation between computational prioritization and experimental confirmation is essential for interpreting the candidate list responsibly and for deciding how to improve the workflow after prospective results become available.

4. Materials and methods

KRAS activity records were collected from ChEMBL via the REST API⁴⁰ and limited to IC₅₀ values reported in nanomolar units with an equality relation. SMILES strings were canonicalized with RDKit; invalid structures were removed; IC₅₀ values were transformed to pIC₅₀; and duplicates were merged by median pIC₅₀. Molecules were represented using 2048-bit radius-2 Morgan fingerprints plus RDKit descriptors. Random-forest regression was run with scikit-learn^{41,42}. Performance was assessed with both random and Murcko-scaffold splits, after which a final model was trained on the full curated dataset. BRICS analogues were generated from the selected seed molecules with the highest experimental pIC₅₀ values after dataset curation. Seed molecules were selected to represent chemically distinct Murcko scaffolds, thereby preserving structural diversity while retaining high bioactivity. A total of X seed molecules meeting the selection criterion of pIC₅₀ ≥ Y were used for BRICS recombination. Ligands were converted from SMILES to PDBQT with Open Babel and docked using AutoDock Vina^{35–37}. For each candidate, the best Vina pose was mapped back into top_hits.csv using exact SMILES matching and docking score agreement. Protein-ligand contacts were computed from docked coordinates using a 4.0 Å cutoff for nearby residues, a 3.5 Å N/O/S cutoff for polar contacts, and a 4.0 Å carbon/halogen cutoff for hydrophobic or van der Waals contacts. All thresholds and outputs were kept to support independent review. Key files are organized as follows: raw_chembl.csv contains the original ChEMBL KRAS records; cleaned_data.csv holds the curated 400-molecule QSAR set; generated.csv lists the 400 BRICS analogues; generated_predictions.csv contains the 57 filtered candidates with predicted pIC₅₀ values; top_hits.csv contains the 20 docked and consensus-ranked molecules; metrics.txt records validation metrics; and run_summary.json logs the docking box, receptor inputs, candidate counts, and execution status. The Supporting Information explains how to read each file and plot, as well as the study's limitations.

For reproducibility, the numerical outputs should be read together with the released code and machine-readable files. The reported random and scaffold results each arise from a single held-out split, so they do not capture the variation expected across alternative partitions. No external prospective set, Y-randomization analysis, formal applicability-domain threshold, or calibrated uncertainty model was used in the present run. Likewise, docking scores were not converted to binding affinities, and the contact analysis used fixed-distance criteria rather than energetic decomposition. These points define the boundary of reproducibility: the calculations and ranking can be repeated from the archived inputs, while biological validity must be established independently.

Associated Content

Declarations

Supporting Information

The Supporting Information is free of charge and includes the dataset pIC₅₀ distribution, random- and scaffold-split validation plots, candidate-ranking files, docking outputs, supplemental graphics, and section-level explanations S1-S8 needed to evaluate the workflow.

Data and Code Availability

The source code and generated output files are available at <https://github.com/selvacbm/ai-drug-discovery-pipeline>. The repository includes the CSV, JSON, TXT, PNG, and docking-log files used to reproduce the candidate-prioritization results.

Author Information

Corresponding Author: Selvaraj Sengottian.

Global AI Drug Discovery Center, College of

Pharmacy and Graduate School of Pharmaceutical Sciences,

Ewha Woman's University, Seoul 03760, Republic of Korea

Orcid ID: <https://orcid.org/0000-0003-1264-6901>

Email: Selvaraj.sengottian@gmail.com

Funding: No funding was received for this study.

Ethics, Consent to Participate, and Consent to Publish declarations: not applicable.

Competing interests: The author declares no competing financial interests.

Acknowledgments

The author acknowledges ChEMBL, RDKit, scikit-learn, Open Babel, AutoDock Vina, and PyMOL for public data and open-source software resources used in this computational study.

References

1. Downward J, Targeting. RAS Signalling Pathways in Cancer Therapy. *Nat Rev Cancer*. 2003;3:11–22.
2. Cox AD, Fesik SW, Kimmelman AC, Luo J, Der CJ. Drugging the Undruggable RAS: Mission Possible? *Nat Rev Drug Discov*. 2014;13:828–51.
3. Simanshu DK, Nissley DV, McCormick FRAS. Proteins and Their Regulators in Human Disease. *Cell*. 2017;170:17–33.
4. Prior IA, Lewis PD, Mattos C. A Comprehensive Survey of Ras Mutations in Cancer. *Cancer Res*. 2012;72:2457–67.
5. Canon J, Rex K, Saiki AY, et al. The Clinical KRAS(G12C) Inhibitor AMG 510 Drives Anti-Tumor Immunity. *Nature*. 2019;575:217–23.
6. Ryan MB, Corcoran RB. Therapeutic Strategies to Target RAS-Mutant Cancers. *Nat Rev Clin Oncol*. 2018;15:709–20.
7. Ostrem JML, Shokat KM. Direct Small-Molecule Inhibitors of KRAS: From Structural Insights to Mechanism-Based Design. *Nat Rev Drug Discov*. 2016;15:771–85.
8. McCormick F. KRAS as a Therapeutic Target. *Clin Cancer Res*. 2015;21:1797–801.
9. Ostrem JM, Peters U, Sos ML, Wells JA, Shokat K, M.. K-Ras(G12C) Inhibitors Allosterically Control GTP Affinity and Effector Interactions. *Nature*. 2013;503:548–51.
10. Hong DS, Fakih MG, Strickler JH, et al. KRASG12C Inhibition with Sotorasib in Advanced Solid Tumors. *N Engl J Med*. 2020;383:1207–17.
11. Jänne PA, Riely GJ, Gadgeel SM, et al. Adagrasib in Non-Small-Cell Lung Cancer Harboring a KRASG12C Mutation. *N Engl J Med*. 2022;387:120–31.
12. Hong DS, Fakih MG, Strickler JH, et al. KRASG12C inhibition with sotorasib in advanced solid tumors. *N Engl J Med*. 2020;383:1207–17. <https://doi.org/10.1056/NEJMoa1917239>.
13. Canon J, Rex K, Saiki AY, et al. The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature*. 2019;575:217–23. <https://doi.org/10.1038/s41586-019-1694-1>.
14. Hallin J, Engstrom LD, Hargis L, et al. The KRASG12C inhibitor MRTX849 provides insight toward therapeutic susceptibility of KRAS-mutant cancers. *Cancer Discov*. 2020;10(1):54–71. <https://doi.org/10.1158/2159-8290.CD-19-1167>.
15. Jänne PA, Riely GJ, Gadgeel SM, et al. Adagrasib in non-small-cell lung cancer harboring a KRASG12C mutation. *N Engl J Med*. 2022;387:120–31.
16. Fell JB, Fischer JP, Baer BR, et al. Identification of the clinical development candidate MRTX849, a covalent KRASG12C inhibitor for the treatment of cancer. *J Med Chem*. 2020;63(13):6679–93.

<https://doi.org/10.1021/acs.jmedchem.9b02052>.

17. Weiss A, Lorthiois E, Barys L, et al. Discovery, preclinical characterization, and early clinical activity of JDQ443, a structurally novel, potent, and selective covalent oral inhibitor of KRASG12C. *Cancer Discov.* 2022;12(6):1500–17. <https://doi.org/10.1158/2159-8290.CD-22-0158>.
18. Ammakkanavar NR, Call J, Shimizu T et al. A first-in-human phase I study of LY3537982, a novel, highly selective and potent KRASG12C inhibitor in patients with KRASG12C-mutant advanced solid tumors (trial in progress). *Cancer Res.* 2022;82(12 Suppl):CT015. (Conference abstract.).
19. Shi Z, Weng J, Niu H, et al. D-1553: A novel KRASG12C inhibitor with potent and selective cellular and in vivo antitumor activity. *Cancer Sci.* 2023;114(7):2951–60. <https://doi.org/10.1111/cas.15829>.
20. Ostrem JM, Shokat KM. Direct Small-Molecule Inhibitors of KRAS: From Structural Insights to Mechanism-Based Design. *Nat Rev Drug Discovery.* 2016;15:771–85.
21. Zellagui DR, Mokrani EH, Allam A, Ozturk M, Bensouici C, Zellagui A. Unravelling chemical profile, antioxidant, anti-Alzheimer and antimicrobial potentials of three propolis from northeastern regions of Algeria: in vitro and in silico evaluation. *Chem Biodivers.* 2025;22:e202402684. [10.1002/cbdv.202402684](https://doi.org/10.1002/cbdv.202402684).
22. Benlatreche T, Mokrani EH, Zerizer MA, Bensouici C, Golhen S, Zouchoune B, Denes G, Merazig H. Synthesis, crystal structure, in vitro anticholinesterase activity, molecular docking and DFT investigation of three 1,10-phenanthroline-tin(IV) complexes. *J Mol Struct.* 2024;1313:138595. [10.1016/j.molstruc.2024.138595](https://doi.org/10.1016/j.molstruc.2024.138595).
23. Moussaoui S, Mokrani EH, Kabouche Z, Guendouze A, Laribi A, Bradai N, Bensouici C, Yilmaz MA, Cakir O, Tarhan A. Evaluation of polyphenolic profile, antioxidant, anti-cholinesterase, and anti-alpha-amylase activities of Pistacia lentiscus L. leaves. *Nat Prod Res.* 2026;40:3415–28. [10.1080/14786419.2025.2471836](https://doi.org/10.1080/14786419.2025.2471836).
24. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* 2017;45:D945–54.
25. Landrum G, RDKit. Open-Source Cheminformatics Software. 2023. <https://www.rdkit.org/>
26. Cherkasov A, Muratov EN, Fourches D, et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J Med Chem.* 2014;57:4977–5010.
27. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J Chem Inf Comput Sci.* 2003;43:1947–58.
28. Trott O, Olson AJ. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J Comput Chem.* 2010;31:455–61.
29. Warren GL, Andrews CW, Capelli AM, et al. A Critical Assessment of Docking Programs and Scoring Functions. *J Med Chem.* 2006;49:5912–31.
30. Vamathevan J, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discovery.* 2019;18:463–77.

31. Lionta E, Spyrou G, Vassilatis DK, Cournia Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr Top Med Chem*. 2014;14:1923–38.
32. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M. On the Art of Compiling and Using Drug-Like Chemical Fragment Spaces. *ChemMedChem* 2008, 3, 1503–1507.
33. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the Chemical Beauty of Drugs. *Nat Chem*. 2012;4:90–8.
34. Baell JB, Holloway GA. New Substructure Filters for Removal of Pan Assay Interference Compounds from Screening Libraries and for Their Exclusion in Bioassays. *J Med Chem*. 2010;53:2719–40.
35. O’Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An Open Chemical Toolbox. *J Cheminform*. 2011;3:33.
36. Eberhardt J, Santos-Martins D, Tillack AF, Forli S AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and, Bindings P. *J. Chem. Inf. Model*. 2021, 61, 3891–3898.
37. Trott O, Olson AJ. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J Comput Chem*. 2010;31:455–61.
38. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep*. 2017;7:42717. 10.1038/srep42717.
39. Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, Yin M, et al. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res*. 2021;49(W1):W5–14. 10.1093/nar/gkab255.
40. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL Database in 2017. *Nucleic Acids Res*. 2017;45:D945–54.
41. Breiman L. Random Forests. *Mach Learn*. 2001;45:5–32.
42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–30.

Figures

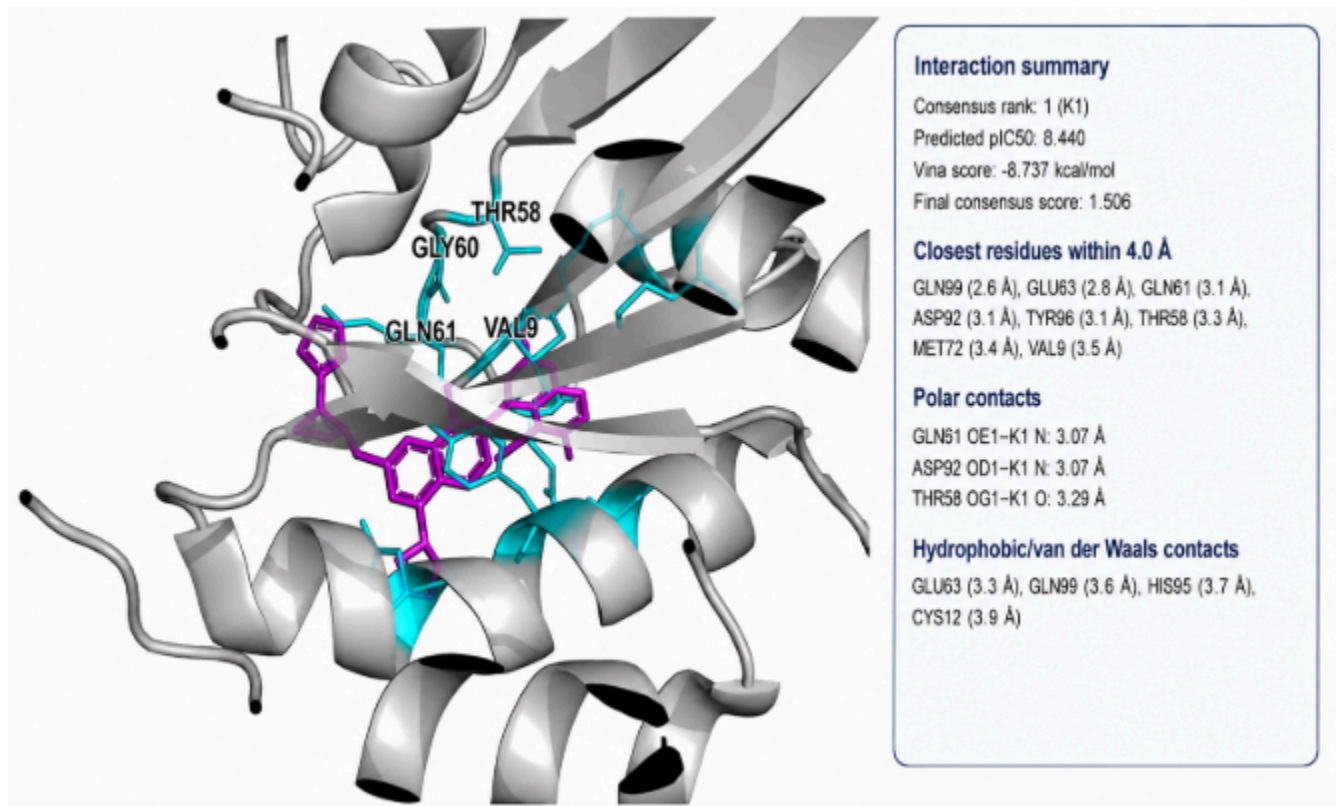


Figure 1

Docked binding pose and predicted protein contacts of K1 in the KRAS G12C pocket. KRAS is shown in gray cartoon, K1 as magenta sticks, and nearby pocket residues in cyan sticks. The panel summarizes predicted polar and hydrophobic/van der Waals contacts derived from the docked coordinates.

KRAS G12C docked poses for 20 prioritized candidates

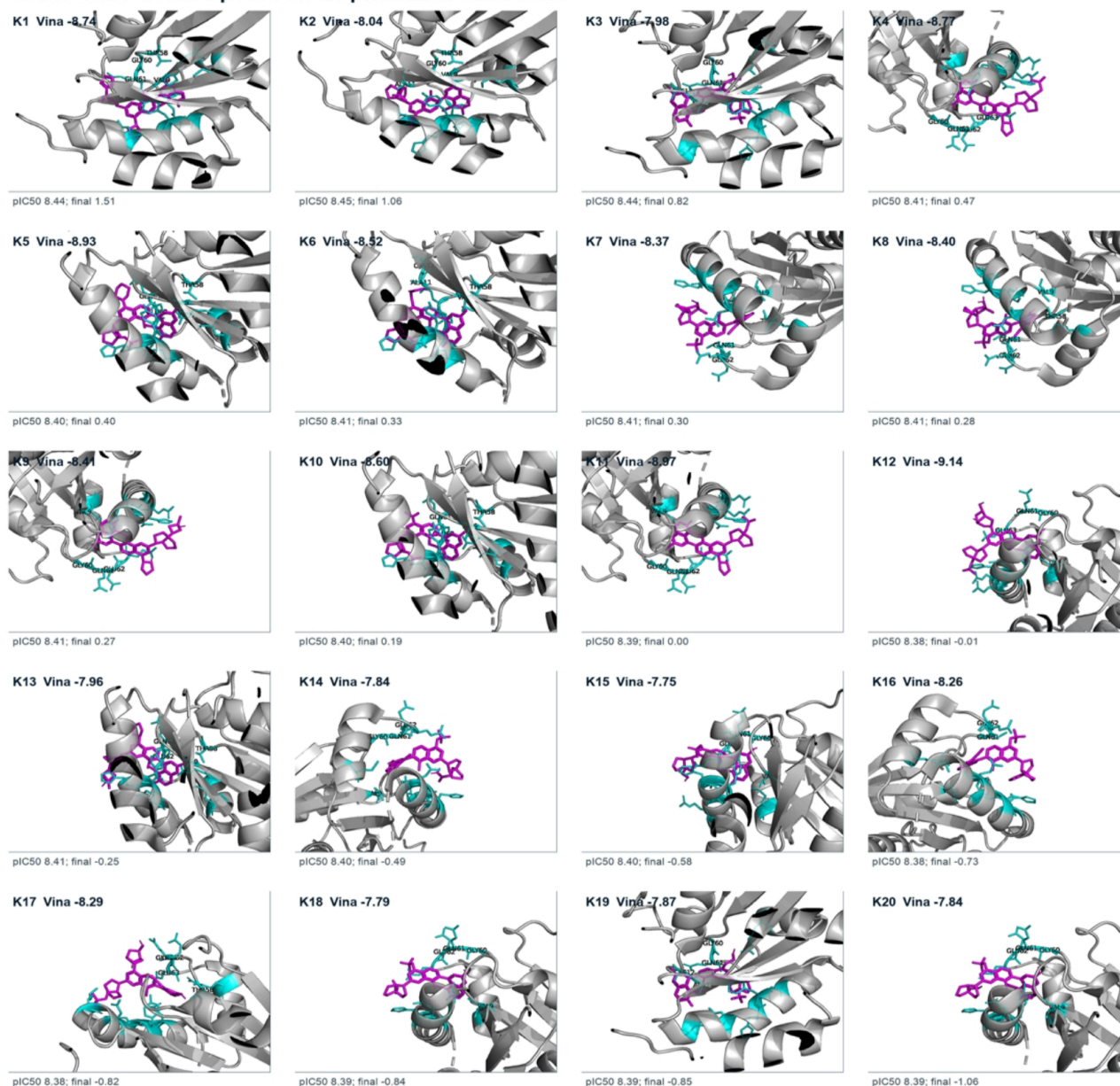


Figure 2

KRAS G12C docking poses for all 20 prioritized candidates (K1-K20). Each panel uses the same rendering style: KRAS cartoon in gray, candidate ligand in magenta, and nearby pocket residues in cyan. Vina score, predicted pIC₅₀, and final consensus score are shown for each candidate.

Supplementary Files

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

- [ManuscriptSI.pdf](#)