Supporting Information

Scaffold-Aware QSAR, BRICS Analog Design, and Docking Consensus for Prioritizing KRAS G12C-Targeted Cancer Therapeutic Leads

Selvaraj Sengottiyan*

*Ewha Womans University, Seoul, Republic of Korea

S1. Purpose and Organization of the Supporting Information

This Supporting Information provides the data and plot explanations required to audit the computational workflow reported in the main manuscript. The files support a transparent decision-support process for KRAS G12C lead prioritization in cancer therapy. They do not constitute experimental evidence of KRAS inhibition; instead, they document how public bioactivity records were curated, how the QSAR model was validated, how BRICS analogs were generated and filtered, and how the final 20 candidates were selected for docking and consensus ranking.

S2. Data and Output File Inventory

File or folder	Count	Explanation
raw_chembl.csv	2400	Initial KRAS ChEMBL records with SMILES and pIC_{50} values before final curation.
cleaned_data.csv	400	Curated QSAR training set after canonicalization, pIC_{50} transformation, invalid-structure removal, and duplicate handling.
generated.csv	400	BRICS-derived analogs generated from active seed chemistry.
generated_predictions.csv	57	Generated candidates that passed quality filters and received random-forest predicted pIC_{50} values.
top_hits.csv	20	Docked candidates with predicted pIC_{50}, Vina score, standardized scores, and final consensus score.
metrics.txt	1	Human-readable model-validation summary.
run_summary.json	1	Machine-readable run summary including counts, receptor paths, docking box,

		exhaustiveness, and completion status.
plots/*.png	4	Model and dataset plots: pIC_{50} distribution, combined parity, random-split parity, and scaffold-split parity.
molecules/mol_0.png to mol_19.png	20	Two-dimensional depictions of the top 20 prioritized candidates.
docking_logs/log_0.txt to log_19.txt	20	AutoDock Vina logs for the 20 docked candidates.

S3. Bioactivity Data Curation

The raw ChEMBL-derived input contains 2400 KRAS activity rows with SMILES and pIC50 values. The curated modeling table contains 400 unique training molecules. The curation step is important because QSAR models are sensitive to duplicated structures, inconsistent salts, invalid SMILES strings, and mixed activity units. The final curated set was intentionally limited to a transparent, auditable training set rather than an opaque maximum-size collection.

The curated pIC50 values range from 7.208 to 10.000, with a median of 7.854, a mean of 8.003, a first quartile of 7.469, and a third quartile of 8.472. This distribution shows that the model was trained in a high-activity region of KRAS chemical space. Therefore, small differences among top predicted candidates should be interpreted cautiously.

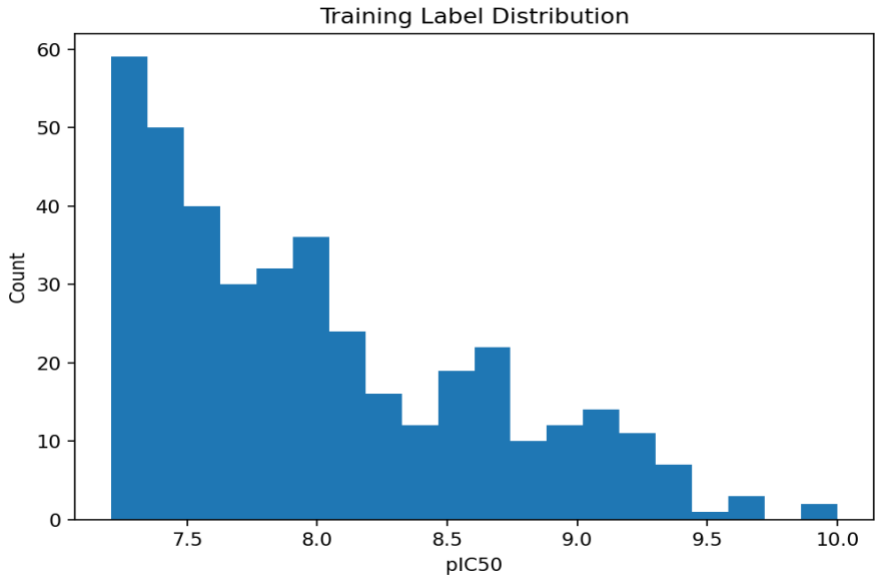

Figure S1. Distribution of curated KRAS pIC_{50} training labels. The plot shows the activity range available after curation and helps define the model's applicability domain. Because the values are concentrated among active compounds, the model is most suitable for rank-ordering related KRAS analogs rather than separating inactive compounds from active compounds across all chemical space.

S4. QSAR Model Validation

The random-forest QSAR model was evaluated by both random and Murcko-scaffold splits. The random split estimates performance when train and test molecules can share related chemistry. The scaffold split is a stricter and more relevant estimate for analog prioritization because test molecules are separated by scaffold from the training molecules.

Validation split	R^2	RMSE	MAE	n_test	Interpretation
Random split	0.5364	0.4721	0.3584	80	Interpolation within related chemistry.
Scaffold split	0.5124	0.5135	0.3809	108	Primary reported estimate for less familiar scaffolds.

The scaffold-split MAE of 0.381 pIC_{50} units is a practical uncertainty boundary. Candidates whose predicted pIC_{50} values differ by less than this amount should not be treated as confidently separated in potency. This is why the manuscript uses the model for enrichment and prioritization rather than exact potency assignment.

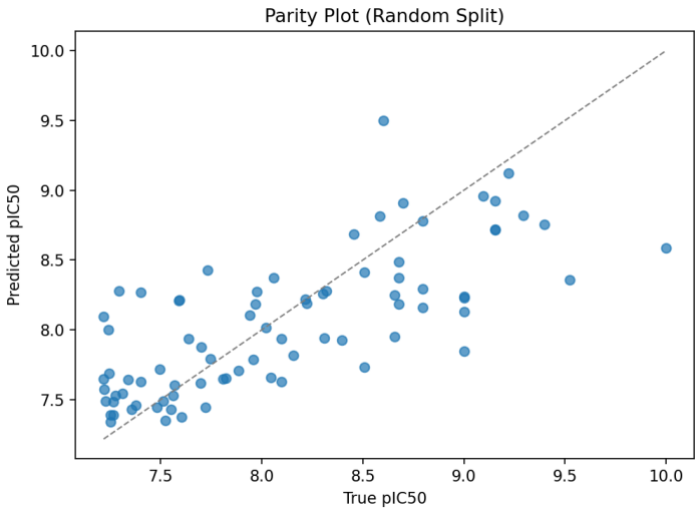

Figure S2. Random-split parity plot. Points near the diagonal indicate agreement between observed and predicted pIC_{50} values for a split that primarily tests interpolation within the available KRAS chemical series.

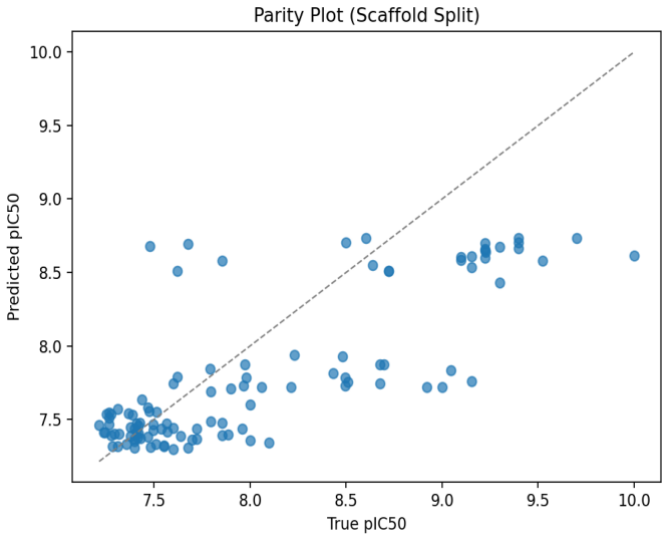

Figure S3. Scaffold-split parity plot. This plot is the key generalization check because test molecules are structurally separated from training molecules by Murcko scaffold. The broader scatter relative to the random split supports cautious ranking rather than overconfident potency claims.

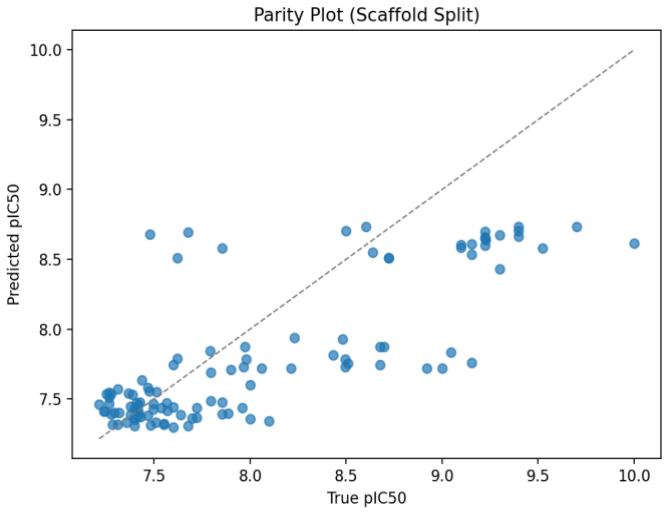

Figure S4. Combined parity summary retained from the pipeline output. The plot provides an additional visual check of observed versus predicted pIC_{50} behavior and should be interpreted together with the split-specific parity plots.

The metrics file records the following validation summary:

n_molecules: 400
n_unique_scaffolds: 132
random_split: R^2=0.5364, RMSE=0.4721, MAE=0.3584, n_test=80
scaffold_split: R^2=0.5124, RMSE=0.5135, MAE=0.3809, n_test=108
recommended_metric_for_reporting: scaffold_split

S5. BRICS Analog Generation and Filtering

BRICS recombination generated 400 analogs. After structural and medicinal-chemistry filtering, 57 candidates remained for QSAR prediction and ranking. The filtered candidates had predicted pIC_{50} values from 8.076 to 8.446, with a median of 8.367. The relatively narrow prediction range shows that the retained analogs are chemically related and should be prioritized by a combination of QSAR score, docking behavior, and synthetic or purchasing practicality.

Workflow stage	Molecule count	Interpretation
Generated BRICS analogs	400	Starting analog pool from chemically meaningful BRICS recombination.
Filtered predicted candidates	57	Candidates passing structural and drug-likeness filters.
Docked candidates	20	Highest-priority candidates taken forward to AutoDock Vina docking.

S6. Docking Setup and Consensus Ranking

Docking used receptor protein/kras.pdbqt derived from protein/KRAS_G12C.pdb. The docking box was centered at x = 2.743, y = -6.323, z = 3.292, with dimensions 26.000 by 26.000 by 19.248 A. The exhaustiveness value was 8. The box source is recorded as kras-site-residues-approx, indicating that the site was approximated from local KRAS site residues.

The final score combines standardized predicted pIC50 and standardized docking affinity, with 60% weight assigned to the ligand-based QSAR signal and 40% weight assigned to docking. This weighting keeps the ranking anchored in experimentally derived bioactivity data while allowing structure-based pocket accommodation to influence final prioritization.

ID	Predicted pIC50	Vina score (kcal/mol)	Final consensus score
K1	8.440	-8.737	1.506
K2	8.446	-8.044	1.055

K3	8.441	-7.984	0.820
K4	8.406	-8.773	0.470
K5	8.400	-8.931	0.404
K6	8.410	-8.519	0.330
K7	8.413	-8.365	0.299
K8	8.411	-8.405	0.282
K9	8.411	-8.411	0.270
K10	8.403	-8.602	0.188
K11	8.386	-8.972	0.001
K12	8.380	-9.143	-0.012
K13	8.408	-7.956	-0.254
K14	8.404	-7.839	-0.489
K15	8.404	-7.748	-0.576
K16	8.384	-8.260	-0.725
K17	8.380	-8.287	-0.820
K18	8.394	-7.789	-0.839
K19	8.392	-7.872	-0.854
K20	8.386	-7.838	-1.056

K1 is the top consensus-ranked molecule because it balances a high predicted pIC50 with a favorable docking score. K12 has the strongest Vina score in the docked set but a lower consensus rank because the QSAR component is less favorable. This difference demonstrates why consensus ranking is preferable to single-score selection in an early virtual-screening workflow.

S7. Two-Dimensional Structures of Prioritized Candidates

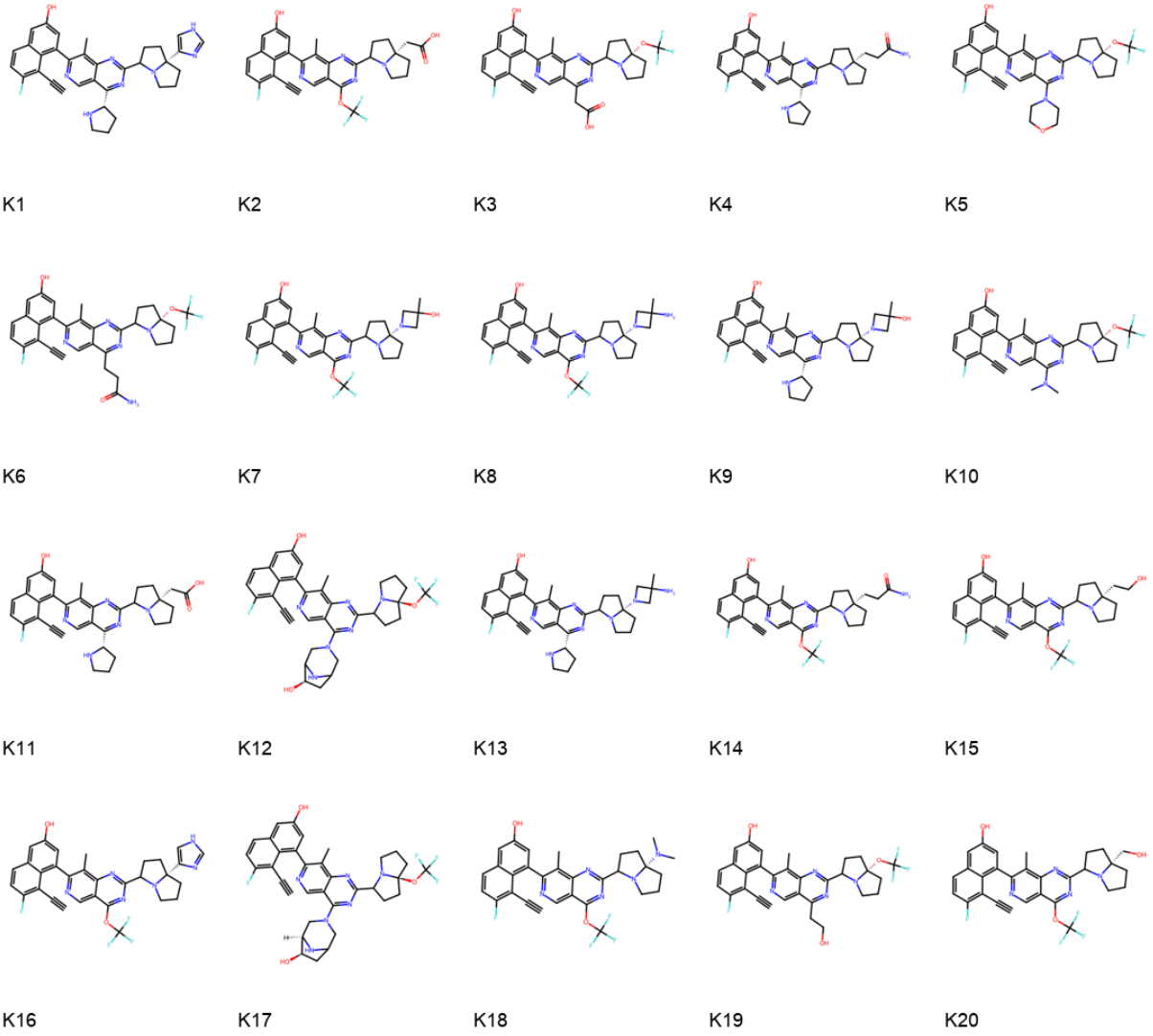

Figure S5. Two-dimensional depictions of K1-K20. The structures provide a rapid visual audit of the prioritized analog set and help readers compare scaffold conservation, substituent changes, and chemical diversity before experimental follow-up.

S8. Run Summary and Reproducibility Notes

The pipeline status was recorded as completed-with-docking, with data source chembl-rest. The run summary reports 400 training molecules, 132 unique scaffolds, 400 generated candidates, 57 ranked candidates after filtering, and 20 docked candidates. These values match the counts reported in the main manuscript.

The docking logs and output PDBQT files are retained to support independent inspection of the docking stage. The molecule images and CSV files are retained to support chemical review and re-ranking. The workflow remains a prioritization framework; biochemical binding assays, cellular pathway assays, selectivity profiling, and structural validation are required before any candidate can be described as a confirmed KRAS G12C inhibitor.