Proceedings of the 6th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/2026.37473

Scaffold-Aware Molecular Machine Learning for Drug Candidate Prioritization: A SARS-CoV-2 Main Protease Case Study

Yulin Pan

Troy High School, Fullerton, USA
Muyu1501@gmail.com

Abstract. Molecular machine learning can prioritize compounds before experimental follow-up, but its value depends on generalization to chemically distinct molecules rather than performance on randomly divided analogues. This study evaluated a scaffold-aware classification workflow using severe acute respiratory syndrome coronavirus 2 main protease (Mpro) as a case-study target. A harmonized multisource dataset assembled from ChEMBL-derived activities, the COVID Moonshot, and peer-reviewed biochemical screens contained 345 curated compounds accepted for modeling: 264 development compounds balanced between active and inactive/weak classes, and an 81-compound external benchmark with no Bemis-Murcko scaffolds shared with the development set. Molecules were represented by 21 RDKit physicochemical descriptors and a 1,024-bit Morgan fingerprint. A descriptor-only comparison of six algorithms showed that nonlinear models produced the three highest cross-validated F1-scores, led by Gradient Boosting (0.749). A prespecified Random Forest was subsequently tuned with grouped inner cross-validation using the combined 1,045-feature representation. On a development scaffold holdout, the model achieved accuracy 0.869, F1-score 0.852, and receiver operating characteristic area under the curve (ROC-AUC) 0.938. After refitting on all development compounds, performance on the external scaffold-disjoint benchmark was accuracy 0.852, F1-score 0.829, and ROC-AUC 0.925. Fingerprint bits and global size, surface-area, and complexity descriptors contributed to predictions. The results support scaffold-aware candidate prioritization while remaining insufficient to establish binding mechanisms or therapeutic activity.

Keywords: molecular machine learning, Morgan fingerprints, scaffold split, virtual screening, SARS-CoV-2 main protease

1. Introduction

1.1. Background and objectives

Computational screening is used to reduce large chemical collections to smaller sets for experimental or structure-based follow-up. In ligand-based modeling, molecular structures are

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.37473

converted into numerical representations and related to experimentally assigned activity labels. The output is therefore a statistical prioritization score rather than direct evidence of protein binding, cellular efficacy, safety, or clinical utility [1].

SARS-CoV-2 Mpro, also termed 3CLpro, was selected as a defined case study because it is required for viral polyprotein processing and has a well-characterized catalytic site [2]. Open-science inhibitor development and recent second-generation Mpro programs further demonstrate that the enzyme remains a current, experimentally tractable antiviral target [3, 4]. The research focus was the modeling framework: whether molecular descriptors and substructure fingerprints could classify active versus inactive/weak compounds when validation explicitly separated chemical scaffolds. This distinction addresses a common source of optimism in molecular machine learning, because random splits can place close structural analogues in both training and test sets [1, 5, 6].

1.2. Research question and hypotheses

The primary research question was: How effectively can supervised molecular machine learning classify and prioritize SARS-CoV-2 Mpro inhibitors when evaluated on scaffold-disjoint compounds? H1 predicted discrimination above non-informative ranking. H1a predicted that combined physicochemical and fingerprint features would retain predictive information across unseen scaffolds. H1b predicted that nonlinear models would outperform linear baselines. H1c predicted that both substructure-level and global physicochemical features would contribute. H1d predicted that model probabilities would support ranking in an external scaffold-disjoint benchmark, without being interpreted as measured binding probabilities.

2. Literature review

Simplified molecular-input line-entry system (SMILES) strings encode molecular graphs as text [7]. RDKit can standardize these structures, calculate interpretable physicochemical descriptors, and generate circular Morgan fingerprints [7, 8]. Descriptors summarize properties such as molecular size, lipophilicity, polarity, hydrogen bonding, flexibility, ring composition, surface area, and graph complexity. Morgan fingerprints complement these global summaries by recording local atom environments as hashed binary features. The two representations are therefore chemically distinct but complementary; neither explicitly encodes a protein-ligand pose, assay conditions, protonation state, or protein flexibility. Recent large-scale benchmarking has reinforced that representation choice must be interpreted together with dataset composition, label quality, and the intended form of chemical-space generalization [6].

Linear classifiers estimate comparatively simple decision boundaries, while K-nearest neighbors, Gradient Boosting, and Random Forest can capture nonlinear structure-property relationships. Random Forest combines randomized decision trees and is compatible with mixed continuous and sparse binary features [9]. Greater flexibility, however, can increase overfitting, particularly in high-dimensional molecular datasets. Reliable evaluation consequently requires grouped cross-validation, class-sensitive metrics, and a scaffold-disjoint test that more closely approximates extrapolation to novel chemistry [1, 5, 6]. Scaffold separation reduces analogue leakage but primarily tests inter-scaffold generalization; activity cliffs and assay-dependent labels can remain difficult even when scaffolds are separated [6, 10].

DOI: 10.54254/2753-8818/2026.37473

3. Methodology

3.1. Dataset curation and molecular representation

The source workbook contained 503 records. After structure validation, activity filtering, duplicate control, and application of the model-use rules, 345 compounds were accepted for modeling. The final curated collection comprised 250 ChEMBL-derived SARS-CoV-2 Mpro records (125 active and 125 inactive/weak), 75 COVID Moonshot compounds (41 active and 34 inactive/weak), 17 compounds from the Fu et al. biochemical screen (1 active and 16 inactive/weak), and 3 compounds from the Jin et al. screen (1 active and 2 inactive/weak). The 250 ChEMBL-derived records were imported from the public MPro_compounds_200325.csv file in the Mayorcool-lab/Covid-Mpro-ChEMBL repository; the source compilation identified the measurements as ChEMBL-derived and linked each accepted record to a ChEMBL molecule identifier and target entry CHEMBL4523582 [11-13]. The Moonshot records were obtained from the openly released submission and assay data and are documented by the COVID Moonshot dataset and its published open-science campaign [3, 14]. The remaining literature-derived records came from peer-reviewed fluorescence-based Mpro assays [15, 16].

Compound eligibility was based on experimental Mpro evidence rather than clinical status or general drug-likeness. A record required a parseable molecular structure and a documented Mpro activity measurement or a clearly reported single-point inactivity result. Untested medicines, compounds included only because they were used clinically, and molecules lacking a defensible target-specific activity value were not converted into negative examples. This criterion was intended to make the classification task distinguish Mpro activity rather than distinguish research inhibitors from ordinary drugs. All 345 final records were assigned high label confidence under the harmonized curation procedure.

Labels followed a prespecified potency gap. Active compounds had IC50 values no greater than 1.0 micromolar or pIC50 values at least 6.0. Inactive/weak compounds had IC50 values at least 10 micromolar, pIC50 values no greater than 5.0, or no detectable inhibition at 100 micromolar. Intermediate 1-10 micromolar records were excluded rather than forced into either class. For the ChEMBL-derived subset, the available extreme-potency selection was retained with a stricter active threshold of pIC50 at least 6.5 and an inactive threshold of pIC50 no greater than 5.0. Where repeated comparable activity measurements were available for one structure, values were placed on the pIC50 scale and summarized by the median; the original measurement count and observed pIC50 range were retained in the curated record. These rules prevented absence of evidence from being treated as evidence of inactivity.

SMILES strings were parsed with RDKit and converted to canonical SMILES before duplicate detection. Exact duplicate rows and records resolving to the same canonical structure were consolidated to one compound-level entry. When multiple compatible measurements described the same structure, the median potency summary was retained rather than allowing repeated assays to appear as independent molecules. Cross-source duplicates were reviewed using canonical structure, experimental evidence, and label consistency; unresolved structural or activity conflicts were not accepted for modeling. Following this process, every accepted row had RDKit status OK and the 345 model records represented 345 unique canonical SMILES, with no canonical-structure duplicate remaining.

Only after compound-level curation was complete were records assigned to model-development and external-evaluation pools. Scaffold keys were generated as RDKit Bemis-Murcko scaffolds without chirality. Because an acyclic molecule returns an empty Murcko scaffold, its canonical non-

DOI: 10.54254/2753-8818/2026.37473

isomeric SMILES was used as the grouping key; this rule affected three accepted compounds. The development pool contained 264 compounds, balanced at 132 active and 132 inactive/weak, and spanned 175 scaffold groups. The fixed external benchmark contained 81 compounds (36 active and 45 inactive/weak), and the intersection of development and benchmark scaffold keys was zero. Because ChEMBL and Moonshot contain medicinal-chemistry analogue series, this separation reduced inflation from close structural relatives appearing on both sides of the evaluation.

Canonical SMILES were converted into 21 RDKit descriptors and a binary 1,024-bit Morgan fingerprint generated with radius 2. Three matrices were constructed: descriptors only, Morgan fingerprints only, and their 1,045-feature concatenation. The six-model algorithm comparison used the descriptor-only matrix. The combined matrix was the prespecified representation for the final Random Forest because it retained both global physicochemical variables and local substructure indicators; the external benchmark was not consulted when specifying this representation. Molecular names, source fields, assay notes, and target labels were excluded from every feature matrix. Saved model artifacts identify scikit-learn version 1.7.2. The external benchmark was scaffold-disjoint but not fully source-disjoint because records from the same source ecosystems occurred in both pools.

3.2. Model comparison and scaffold-aware validation

First, six classifiers were compared using the same 21-descriptor representation: Logistic Regression, Ridge Classifier, Linear Support Vector Classifier, K-Nearest Neighbors, Gradient Boosting, and Random Forest. Five-fold StratifiedGroupKFold preserved class balance while grouping molecules by scaffold. F1-score was the primary comparison metric because both missed actives and unsupported positive predictions were relevant. A separate sensitivity analysis tested 30 previously unused random seeds; for each seed, one test fold was taken from a four-fold scaffold-grouped split of the 264-compound development pool. These runs quantified split sensitivity and were not used to choose a favorable seed, representation, threshold, or external-benchmark result.

Second, the prespecified descriptor-plus-Morgan Random Forest was evaluated using a locked scaffold-grouped development holdout containing 61 compounds (33 inactive/weak and 28 active), leaving 203 compounds for model development. Thirty randomized hyperparameter configurations were evaluated by five-fold grouped inner cross-validation with F1 scoring. The selected forest used 300 trees, the Gini criterion, no maximum depth, two minimum samples per split, two minimum samples per leaf, 10% of features considered per split, balanced class weights, bootstrap sampling, and random state 42. Class labels were generated at the default probability threshold of 0.50. After one development-holdout evaluation, the same parameters were refitted on all 264 development compounds and applied once to the 81-compound external benchmark.

Performance was quantified by accuracy, balanced accuracy, precision, recall, F1-score, ROC-AUC, Matthews correlation coefficient (MCC), specificity, and confusion-matrix counts. Feature interpretation used impurity-based Random Forest importance. Because hashed Morgan bits were not accompanied by atom-environment mappings in the saved output, individual bit identifiers were treated as anonymous substructure indicators rather than assigned mechanistic interpretations.

4. Results

4.1. Descriptor-only model comparison

The descriptor-only comparison favored nonlinear learning. Mean cross-validated F1-scores were 0.749 for Gradient Boosting, 0.746 for K-Nearest Neighbors, and 0.730 for Random Forest,

DOI: 10.54254/2753-8818/2026.37473

compared with 0.721 for both Ridge Classifier and Linear SVC and 0.690 for Logistic Regression (Figure 1a). Across 30 scaffold-aware seed trials, Gradient Boosting also had the highest mean F1-score (0.812), followed by Random Forest (0.789) and K-nearest neighbors (0.785); Logistic Regression averaged 0.752. These results supported H1b at the family level but did not identify Random Forest as the best descriptor-only algorithm.

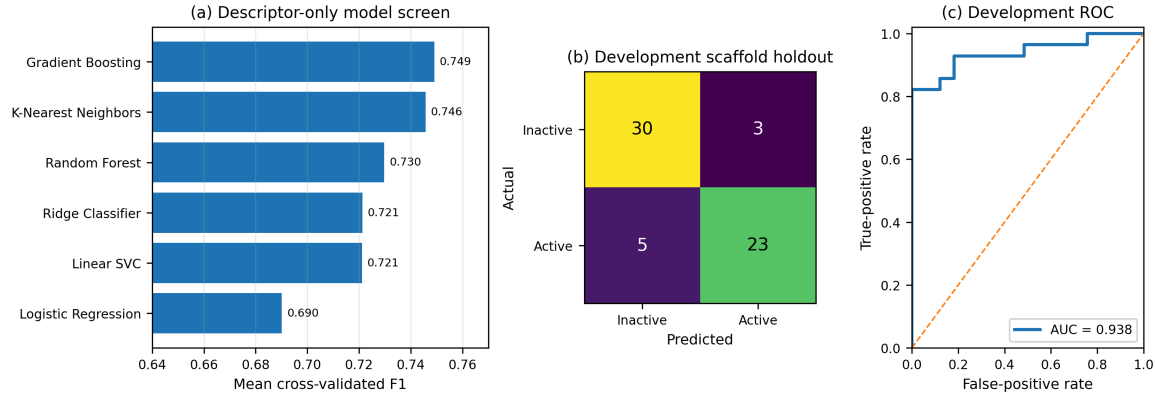

Figure 1. Development evaluation. (a) Descriptor-only cross-validated F1-scores; (b) confusion matrix and (c) ROC curve for the descriptor-plus-Morgan Random Forest on the scaffold holdout

4.2. Final Random Forest performance

The best inner cross-validation F1-score for the expanded Random Forest was 0.868 ± 0.067. On the development scaffold holdout, the model correctly classified 53 of 61 compounds: 30 true negatives, 3 false positives, 5 false negatives, and 23 true positives (Figure 1b). Accuracy was 0.869, precision 0.885, recall 0.821, F1-score 0.852, ROC-AUC 0.938, and MCC 0.736. The inner cross-validation and scaffold-holdout F1-scores were similar.

After refitting on all development compounds, the external benchmark yielded 40 true negatives, 5 false positives, 7 false negatives, and 29 true positives. Accuracy was 0.852, precision 0.853, recall 0.806, specificity 0.889, F1-score 0.829, ROC-AUC 0.925, and MCC 0.699 (Figure 2). Performance therefore remained similar when all benchmark scaffolds were absent from development, although the external set was assembled under the same harmonized labeling rule and was not source-disjoint.

Table 1. Final scaffold-aware Random Forest performance

Evaluation set	#	Accuracy	Balanced accuracy	Precision	Recall	F1	ROC-AUC
Development scaffold holdout	61	0.869	0.865	0.885	0.821	0.852	0.938
External scaffold-disjoint benchmark	81	0.852	0.847	0.853	0.806	0.829	0.925

DOI: 10.54254/2753-8818/2026.37473

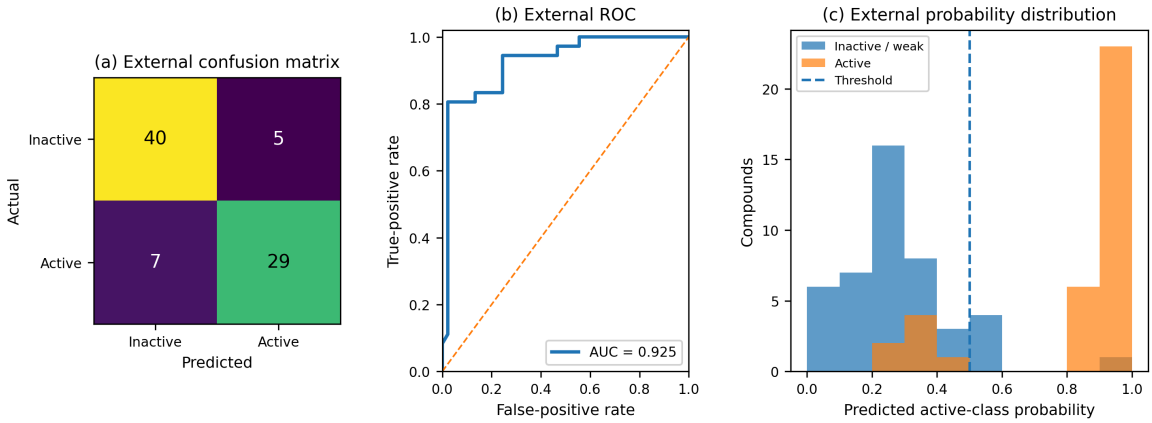

Figure 2. External scaffold-disjoint benchmark. (a) Confusion matrix, (b) receiver operating characteristic curve, and (c) predicted-probability distributions by experimental class

4.3. Feature contributions and ranking behavior

The two highest impurity-importance features were Morgan_848 (0.0437) and Morgan_327 (0.0363), followed by molecular weight (0.0325), Labute approximate surface area (0.0287), Morgan_659 (0.0235), valence-electron count (0.0235), and Morgan_10 (0.0235). Heavy-atom count, Bertz complexity, fraction Csp3, TPSA, molar refractivity, and heteroatom count also appeared among the top 15 features. The mixture of fingerprint bits and global descriptors supported H1c, while the distributed importance indicated that predictions were not controlled by a single property.

External probabilities separated most active and inactive/weak compounds, consistent with the ROC-AUC of 0.925. The probability distribution nevertheless contained overlap near the 0.50 threshold, and one inactive/weak compound received a high predicted active-class probability, demonstrating that probability magnitude should not be treated as measured affinity. H1d was therefore supported for ranking under this benchmark but not for probability calibration or novel-drug discovery.

5. Discussion

5.1. Hypothesis evaluation and interpretation

H1 and H1a were supported within the defined curation framework. Both scaffold-holdout and external ROC-AUC values exceeded 0.92, and positive-class F1 remained above 0.82. The external results are especially relevant because no Bemis-Murcko scaffold was shared with development. H1b was also supported in the descriptor-only experiment: all three nonlinear models exceeded all three linear models in mean F1. However, Gradient Boosting, not Random Forest, was the strongest descriptor-only model. The final Random Forest results therefore demonstrate the effectiveness of a prespecified high-dimensional ensemble and combined representation rather than proving that Random Forest is universally superior.

H1c was supported statistically because both Morgan bits and physicochemical descriptors ranked highly. The fingerprint contribution is consistent with substructure-specific signal, whereas molecular weight, surface area, atom counts, and complexity describe broader chemical context. These importances characterize the fitted model but do not establish Mpro interaction mechanisms.

DOI: 10.54254/2753-8818/2026.37473

Correlated descriptors may divide importance, impurity importance can favor variables with many candidate split points, and unmapped hashed bits cannot be assigned to specific functional groups without reconstructing fingerprint bit environments.

H1d was supported for external ranking but should be interpreted narrowly. The external benchmark was scaffold-disjoint and experimentally labeled, yet several data sources and the curation thresholds were shared with development. This design tests chemical extrapolation more strongly than a random split, but it does not test transfer to a new laboratory, assay format, or target. The model is consequently suitable as a research prioritization filter, not as a stand-alone determination of activity.

5.2. Limitations and future work

The principal limitations are heterogeneous assay sources, a moderate sample size, and the use of a binary potency gap that discards intermediate compounds. Morgan fingerprints and global two-dimensional descriptors do not represent stereochemical ensembles, protonation under assay conditions, covalent reaction mechanisms, active-site water, or protein flexibility. In addition, the external benchmark, although scaffold-disjoint, was curated through the same project pipeline and partly from the same source ecosystems. Recent molecular-learning studies show that split design, label noise, and activity cliffs can materially affect apparent generalization [6, 10]. A second independently curated benchmark would provide stronger evidence against source-specific artifacts.

Future refinement should prioritize stronger data validation and model characterization. Quantitative IC50 values could support regression after assay harmonization; repeated or nested scaffold validation could quantify uncertainty; and permutation importance, fingerprint bit mapping, probability calibration, and applicability-domain analysis could improve interpretation. Docking, interaction fingerprints, and three-dimensional conformer descriptors may add target-specific information, but they should be evaluated as additional predictors rather than treated as experimental truth.

6. Conclusion

This study evaluated molecular machine learning as a candidate-prioritization framework using SARS-CoV-2 Mpro as a case study. Descriptor-only comparisons supported the hypothesis that nonlinear models capture useful chemical relationships, while a scaffold-aware Random Forest using the prespecified combination of 21 descriptors and 1,024 Morgan bits maintained strong performance on both a development scaffold holdout and an external scaffold-disjoint benchmark. External accuracy was 0.852, F1-score 0.829, and ROC-AUC 0.925. The results indicate that this combined representation can prioritize experimentally labeled compounds across unseen scaffolds. They do not establish direct binding mechanisms, therapeutic effectiveness, or performance on unrelated targets. The strongest next steps are independently sourced validation, assay-consistent quantitative modeling, and target-specific structural features.

Acknowledgement

The author gratefully acknowledges Li Zhu for mentorship and guidance throughout this research project.

DOI: 10.54254/2753-8818/2026.37473

References

[1] Wu, Z., Ramsundar, B., Feinberg, E. N., et al. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science, 9*, 513–530. https://doi.org/10.1039/C7SC02664A

[2] Zhang, L., Lin, D., Sun, X., et al. (2020). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. *Science, 368*, 409–412. https://doi.org/10.1126/science.abb3405

[3] Boby, M. L., Fearon, D., Ferla, M., et al.; COVID Moonshot Consortium. (2023). Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. *Science, 382*(6671), eabo7201. https://doi.org/10.1126/science.abo7201

[4] Allerton, C. M. N., Arcari, J. T., Aschenbrenner, L. M., et al. (2024). A second-generation oral SARS-CoV-2 main protease inhibitor clinical candidate for the treatment of COVID-19. *Journal of Medicinal Chemistry, 67*(16), 13550–13571. https://doi.org/10.1021/acs.jmedchem.3c02469

[5] Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry, 39*, 2887–2893. https://doi.org/10.1021/jm9602928

[6] Deng, J., Yang, Z., Wang, H., et al. (2023). A systematic study of key elements underlying molecular property prediction. *Nature Communications, 14*, 6395. https://doi.org/10.1038/s41467-023-41948-6

[7] Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences, 28*, 31–36. https://doi.org/10.1021/ci00057a005

[8] Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling, 50*, 742–754. https://doi.org/10.1021/ci100050t

[9] Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

[10] van Tilborg, D., Alenicheva, A., & Grisoni, F. (2022). Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling, 62*(23), 5938–5951. https://doi.org/10.1021/acs.jcim.2c01073

[11] Mendez, D., Gaulton, A., Bento, A. P., et al. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research, 47*, D930–D940. https://doi.org/10.1093/nar/gky1075

[12] Zdrazil, B., Felix, E., Hunter, F., et al. (2024). The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research, 52*(D1), D1180–D1192. https://doi.org/10.1093/nar/gkad1004

[13] Mayorcool-lab. (n.d.). MPro_compounds_200325.csv. *Covid-Mpro-ChEMBL GitHub repository*. https://github.com/Mayorcool-lab/Covid-Mpro-ChEMBL/blob/main/MPro_compounds_200325.csv

[14] COVID Moonshot Consortium. (n.d.). COVID Moonshot submissions dataset. *GitHub repository*. https://github.com/postera-ai/COVID_moonshot_submissions

[15] Fu, L., Ye, F., Feng, Y., et al. (2020). Both boceprevir and GC376 efficaciously inhibit SARS-CoV-2 by targeting its main protease. *Nature Communications, 11*, 4417. https://doi.org/10.1038/s41467-020-18233-x

[16] Jin, Z., Du, X., Xu, Y., et al. (2020). Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature, 582*, 289–293. https://doi.org/10.1038/s41586-020-2223-y

DOI: 10.54254/2753-8818/2026.37473

Appendix

Appendix A - Pseudocode for Scaffold-Aware Mpro Molecular Machine Learning. This material describes the filtering, feature construction, scaffold grouping, descriptor-only model comparison, seed sensitivity analysis, final Random Forest tuning, and external evaluation.

Appendix B - Curated SARS-CoV-2 Mpro Dataset. This material contains the 503-record audit trail, the 345 compounds accepted for modeling, source information, label rules, duplicate handling, scaffold assignments, and development/external split decisions.

Appendix A. Embedded compact workflow pseudocode

The following compact pseudocode embeds the substantive scaffold-aware modeling workflow in the manuscript.

Table A1. Compact scaffold-aware modeling workflow pseudocode

Stage	Compact pseudocode
A1. Ingest and partition	LOAD the curated workbook. FILTER rows with RDKit Status = OK, Use in Model = Yes, binary Binding Label, and complete 21 descriptors; ASSERT 345 accepted compounds. ASSIGN 264 compounds to D_dev and 81 to D_ext from the prespecified split; SET y and y_ext from Binding Label.
A2. Scaffold and features	FOR each compound, COMPUTE an achiral RDKit Bemis-Murcko scaffold; IF the scaffold is empty for an acyclic molecule, USE canonical non-isomeric SMILES as the group key. ASSERT no development/external scaffold overlap. GENERATE 21 descriptors and 1,024-bit binary Morgan fingerprints (radius = 2); CONCATENATE to X_comb with 1,045 features.
A3. Six-model screen	ON X_desc only, COMPARE Logistic Regression, Ridge Classifier, Linear SVC, KNN, Gradient Boosting, and Random Forest by five-fold StratifiedGroupKFold grouped by scaffold (shuffle = True, random_state = 42); standardize LR, Ridge, Linear SVC, and KNN inside pipelines; USE F1 as the primary metric and report the remaining classification metrics.
A4. Seed sensitivity	FOR 30 previously unused seeds, CREATE a four-fold scaffold-grouped split and EVALUATE the first train/test fold for all six descriptor-only models; ASSERT zero scaffold overlap; SUMMARIZE metric variability. DO NOT use these trials to select the final seed, representation, threshold, or external result.
A5. Final RF tuning	CREATE the locked first fold of four-fold StratifiedGroupKFold at seed 42: 203 development-training and 61 holdout compounds. ON X_comb, RUN RandomizedSearchCV for 30 Random Forest configurations with five-fold grouped inner CV and F1 scoring. SELECT 300 trees, Gini, max_depth=None, min_samples_split=2, min_samples_leaf=2, max_features=0.10, bootstrap=True, class_weight=balanced, random_state=42.

DOI: 10.54254/2753-8818/2026.37473

Table A1. (continued)

A6. Holdout and external	PREDICT the locked holdout at threshold 0.50 and CALCULATE accuracy, balanced accuracy, precision, recall, F1, ROC-AUC, MCC, and confusion matrix. REFIT the selected estimator on all 264 development compounds; EVALUATE the 81-compound external benchmark once at the same threshold and additionally calculate specificity.
A7. Reporting controls	COMPUTE impurity-based feature importance over all 1,045 features; keep hashed Morgan bit identifiers anonymous unless atom environments are reconstructed. EXPORT predictions, metrics, feature importance, and model artifacts. INTERPRET model probabilities only as candidate-prioritization scores, not measured affinity or therapeutic efficacy.

Appendix B. Embedded curated-dataset documentation

Table B1. Composition of the curated SARS-CoV-2 Mpro dataset by source and activity class

Source	Accepted	Active (1)	Inactive/weak (0)
ChEMBL-derived Mpro compilation	250	125	125
COVID Moonshot	75	41	34
Fu et al. biochemical screen	17	1	16
Jin et al. biochemical screen	3	1	2
Total accepted	345	168	177

Curation and labels. The source workbook contained 503 records. A record required a parseable molecular structure and documented Mpro activity evidence. General active labels used IC50 <= 1.0 micromolar or pIC50 >= 6.0; inactive/weak labels used IC50 >= 10 micromolar, pIC50 <= 5.0, or no detectable inhibition at 100 micromolar. Intermediate 1-10 micromolar records were excluded. The ChEMBL-derived active subset retained the stricter median pIC50 >= 6.5 rule. Repeated comparable measurements were summarized by median pIC50; untested medicines were not converted into negative examples.

Duplicate control. RDKit canonical SMILES were used for structure identity. Exact duplicates and records resolving to the same canonical structure were consolidated, compatible repeated measurements were summarized rather than treated as independent molecules, and cross-source conflicts were reviewed before inclusion. The 345 accepted rows represent 345 unique canonical structures.

Scaffold-aware partitions. The development set contains 264 compounds (132 active, 132 inactive/weak) across 175 scaffold groups. The external benchmark contains 81 compounds (36 active, 45 inactive/weak). Achiral Bemis-Murcko scaffolds were used, with canonical non-isomeric SMILES as the grouping key for the three accepted acyclic compounds; development/external scaffold overlap is zero.

Model-feature fields. The 21 RDKit descriptors are molecular weight, LogP, TPSA, valence-electron count, H-bond acceptors, H-bond donors, rotatable bonds, ring count, amide bonds, heavy-atom count, fraction Csp3, aromatic ring count, aliphatic ring count, saturated ring count, heteroatom count, formal charge, molar refractivity, Labute ASA, Bertz complexity, Lipinski

DOI: 10.54254/2753-8818/2026.37473

violations, and drug-likeness pass. The curated dataset also preserves structure identifiers, activity values/units, evidence type, source URLs, curation decisions, label confidence, repeated-measurement summaries, and model-dataset version.