



NAS COMPUTATIONAL RESEARCH

When should a BBB model abstain?

Controlled comparisons, audited external records and frozen prediction policies

All three computational follow-up arms were completed. Equal-size fixed-test comparisons showed that model rankings depend on the training condition. On 428 external records retained after conservative molecular-overlap filtering, a simple confidence policy reduced mean accepted error from 12.34% to 5.55% at 74.95% coverage. Part of that reduction came from altered class composition, and accepted negative-class performance remained weak. No policy met every prespecified usefulness requirement on the most chemically distant external subset. These results support further study of selective prediction, not a clinical or first-discovery claim.

DATE / VERSION / AUTHOR

27 September 2026 UTC / Follow-up v1 / NaS Research

COMPUTATIONALLY CHECKED; NOT PEER REVIEWED

Private research report. No wet-lab work, clinical deployment or public release. Prior BBBP audit preserved. All aggregate results refer to the frozen follow-up unless stated otherwise.

A frozen computational follow-up

NAS-BBB-002 extends the completed BBBP audit with three questions: does model ordering persist under controlled test composition; does it transfer to external records after molecular-overlap filtering; and can validation-frozen abstention reduce error while retaining both classes? All three arms ran. No wet-lab work or public release occurred.

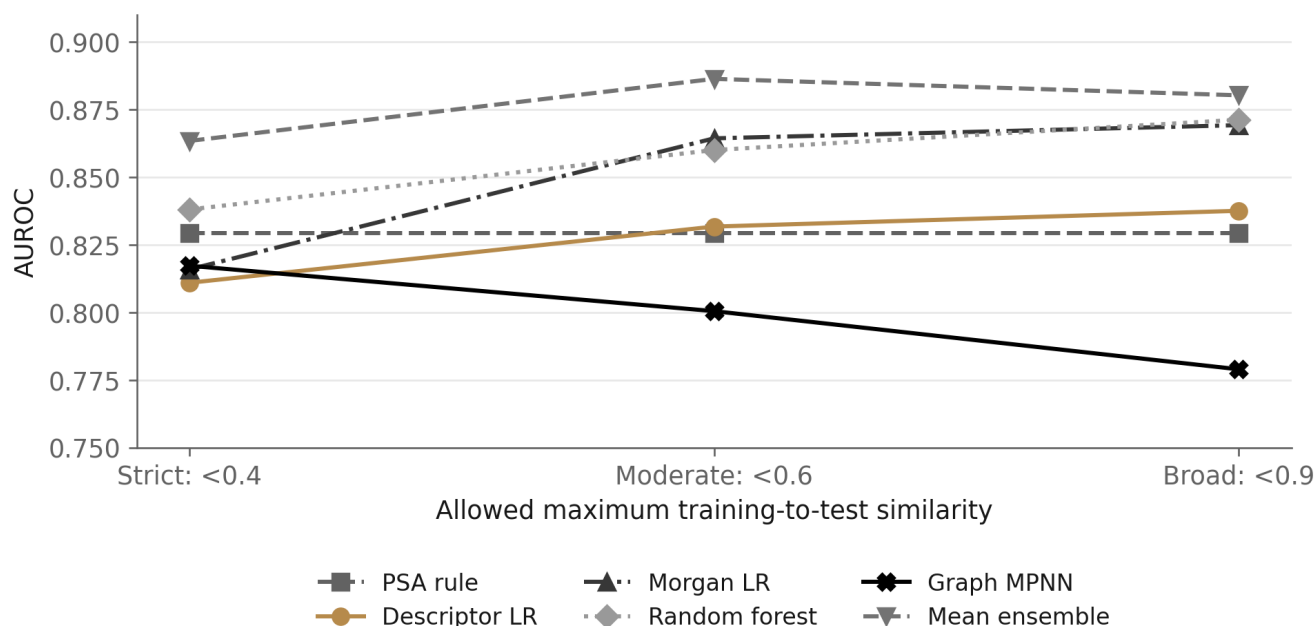
Arm	Design	Completed scope
1. Controlled comparison	Five seeds, three similarity restrictions; identical test and validation rows within seed	75 model runs; 300 training and 100 test records per condition
2. External evaluation	Five original-data training splits; three fixed external cohorts	25 model runs; no external labels used to fit or calibrate
3. Abstention	Four retention scores plus random baseline; three validation coverage levels	Six predictors including ensemble; all 810 population/policy settings reported

The protocol was frozen on 27 September 2026 at 02:57 UTC before follow-up model scoring. Prior main-study results were known. Seeds 21-25, model definitions and calibration were fixed. Each controlled training set has 150 records per class; test sets have 50 per class. Model-validation and separate policy-validation sets each contain 20 records per class.

Models are a calibrated PSA rule, logistic regression on eight physicochemical descriptors, logistic regression on Morgan fingerprints, a 300-tree random forest, and a 99,521-parameter message-passing network trained from scratch. The ensemble is the unweighted mean of their calibrated probabilities. External training uses grouped original-data splits with approximately 64% training, 16% model validation and 20% policy validation.

AUROC measures ranking. Classification uses probability 0.5. Lower Brier score indicates better squared probability error. Results are means over five fitted seeds, not five independent external datasets. Intervals resample molecular groups 1,000 times and are conditional on these fitted models; they do not include full retraining uncertainty.

1. Holding the test set fixed



Five-seed mean AUROC. Each seed uses the same 100 test records, 300 balanced training records and common validation records in all three conditions. Lines connect conditions; they are not confidence intervals.

Allowing closer training chemistry raised Morgan LR AUROC from 0.816 to 0.869. The paired broad-minus-strict difference was +0.053 (pointwise 95% interval +0.023 to +0.087). Random forest rose from 0.838 to 0.871. The PSA rule remained exactly 0.829 across conditions, as required by the design.

The graph model moved in the opposite direction: 0.817 under the strict restriction and 0.779 under the broad restriction. In the broad condition it ranked below descriptor LR by 0.059; the interval adjusted for the three planned model comparisons was -0.105 to -0.009. This is evidence about this fixed small-data implementation, not all graph networks.

The mean of per-seed median nearest-training similarities was 0.290, 0.363 and 0.384 for strict, moderate and broad. Equal size and prevalence remove two confounders; sampled training chemistry still differs. Descriptor distributions and all seeds are retained. Broad-minus-strict contrasts are exploratory diagnostics; the frozen primary comparisons are between models within each condition.

2. What counts as external evidence

Cohort	Candidate rows	Retained + / -	Interpretation
Tong S-data	527	353 / 75 (428)	Primary molecular holdout; heterogeneous endpoints
B3DB recent set	175	113 / 2 (115)	Insufficient negatives for comparative discrimination
B3DB quantitative	175 source-filtered	120 / 19 (139)	Below frozen 20-per-class requirement

All external candidates were compared with all 2,030 original source rows, including excluded original records. Canonical, neutral-parent, tautomer and stereo-free representation matches were excluded, followed by Morgan similarity ≥ 0.9 . Conflicts, duplicates, ambiguous organic mixtures and failures were audited. Retained structures passed a cleanup-idempotence check.

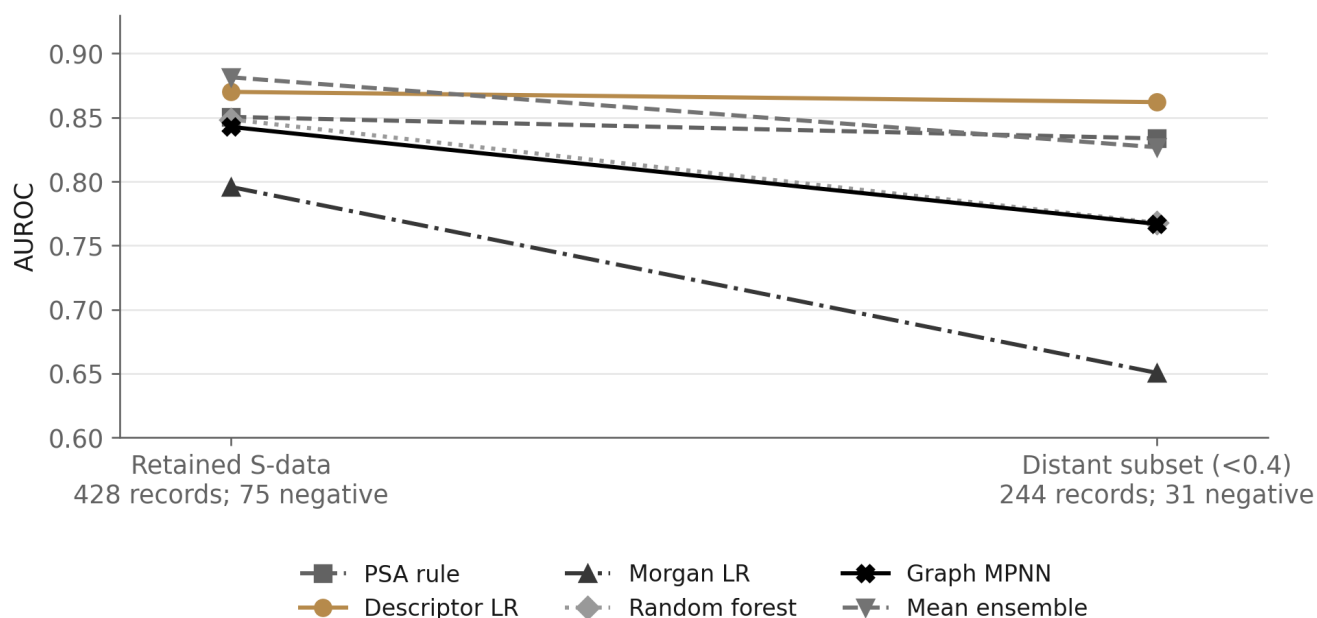
For S-data, exclusions were 74 representation matches, 14 additional near matches, nine ambiguous organic mixtures, one duplicate and one standardization failure. The 74 are not 74 proven identical molecules: 45 matched a tautomer key, and 29 additional records matched only the stereo-free tautomer key. No canonical or neutral key matched. These are deliberately conservative representation exclusions.

The most distant S-data subset contains 244 records (213 positive, 31 negative), each with maximum Morgan similarity < 0.4 to the entire original pool. It is a nested sensitivity analysis, not an additional independent validation dataset.

S-data was independently assembled but combines endpoint types and inferred annotations; its released two-column file lacks row-specific assay mapping. B3DB aggregates literature sources and includes direct Martins/TDC lineage. The quantitative sensitivity excludes R1/R30 and requires all listed source years after 2012, but this does not establish independently measured assays. Data names and post-2012 citation years are not proof of experimental independence.

The recent B3DB set and quantitative sensitivity remain reported secondary checks. Their two and 19 negative records respectively prevent the stronger balanced-performance conclusions required by the frozen protocol. Neither threshold was relaxed after observing outcomes.

External ranking and classification



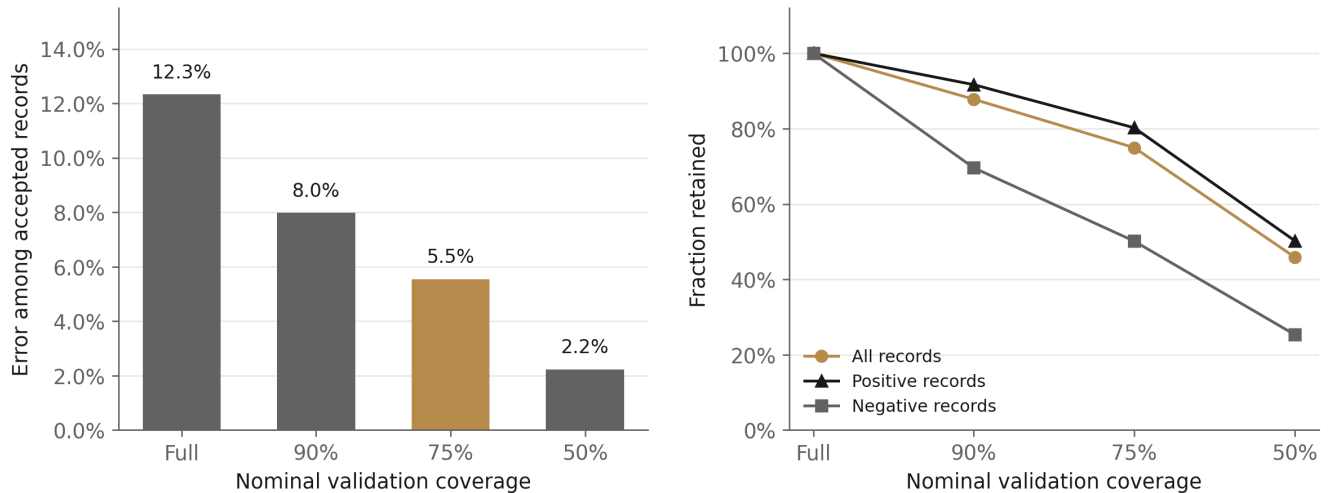
Five-seed mean AUROC on the retained primary external cohort and its distant subset. The subset changes composition as well as distance; this is not a controlled causal distance experiment.

Model	AUROC	Brier	Sensitivity	Specificity
PSA rule	0.851	0.096	0.981	0.365
Descriptor LR	0.870	0.094	0.985	0.368
Morgan LR	0.796	0.113	0.925	0.563
Random forest	0.849	0.099	0.960	0.499
Graph MPNN	0.843	0.101	0.926	0.547
Mean ensemble	0.881	0.089	0.981	0.467

On all 428 external records, none of the three planned AUROC model differences excluded zero after adjustment. Descriptor LR versus the rule was +0.020; graph versus descriptor LR was -0.027. The ensemble had the highest descriptive mean AUROC (0.881), but its superiority was not a registered paired hypothesis.

Within the 244-record distant subset, descriptor LR ranked above this graph model: graph-minus-descriptor AUROC was -0.095 (adjusted interval -0.184 to -0.016). Its descriptor AUROC of 0.862 coexisted with specificity of only 0.200 at the fixed 0.5 threshold. Good ranking does not establish useful classification of negative molecules.

3. Predict less, with measured limits



Illustrative descriptor-LR confidence policy on S-data, averaged over five seeds. Thresholds were set on separate original-data validation records. Actual retention need not equal nominal validation coverage. The 75% setting is discussed for interpretability, not selected for deployment.

At nominal 75% validation coverage, actual external coverage was 74.95%. Error among accepted records was 5.55%, compared with 12.34% at full coverage. The paired absolute reduction was 6.79 percentage points (pointwise 95% interval 4.60-9.20). Mean error among rejected records was 32.87%. Thresholds ranged from 0.751 to 0.777 across seeds.

Average retention was 80.23% of positive records and 50.13% of negative records. Every seed retained at least 34 negatives. Accepted specificity was 52.90% and sensitivity 99.93%. Thus, even this favorable example misclassified almost half the accepted negative records; high overall accuracy must not be sold as uniformly reliable BBB classification.

A post hoc diagnostic standardized accepted error to the original external class balance: 8.31%, still below 12.34%. Of the observed 6.79-point reduction, 3.61 points came from changed class mixture and 3.18 from lower within-class error under the retained mixture. This descriptive decomposition is separate from the frozen criterion and has no new inferential claim.

What the complete policy grid shows

Population	Passing / nonrandom settings	Result
S-data, 428 records	22 / 72	Exploratory criterion supported for several fixed policies
Distant S-data, 244 records	0 / 72	No policy met all coverage, support and improvement requirements
Controlled broad / moderate / strict	7 / 72; 15 / 72; 11 / 72	Benefits depend on condition and model
Both B3DB cohorts and distant subsets	0 / 72 in each	Minority-class support is inadequate

The criterion required at least five percentage points less error, mean achieved coverage $\geq 50\%$, each-class coverage $\geq 30\%$ in every seed, at least 20 accepted records per class in every seed, a positive paired 95% lower bound and at least 950 valid bootstrap replicates. All four scores and all three nominal coverages are reported for each model; deterministic random retention provides a baseline.

For descriptor LR at nominal 75%, random retention kept 74.35% of external records and had 12.70% error, versus 5.55% under confidence retention. Similarity-only retention at the same nominal setting transported poorly: it retained 38.18% and had 12.60% error. These comparisons have different actual coverage and are descriptive, not matched-coverage causal effects.

The useful research output is a reproducible abstention audit: a fixed policy can concentrate errors in rejected records, but class retention, negative-class performance and chemical distance must accompany the headline accuracy. No external winner was chosen for a production system. Screening 72 nonrandom settings per population creates multiplicity; policy intervals are exploratory and do not control the familywise error across that grid.

No uncertainty method, clinical application or first discovery is claimed. Tong and colleagues already studied external BBB evaluation and uncertainty-based rejection in 2022. The contribution here is the specific controlled comparison, conservative representation audit and transparent transport test with class-support requirements. Priority or publication novelty remains unestablished.

Verification and reproducibility

All 100 registered model runs completed. The frozen protocol and downloaded source hashes match. All split, group, sample-count, class-balance, distance-threshold and fixed-rule invariants passed. The metric implementation matched 2,430 independent scikit-learn calculations, and the vectorized weighted bootstrap kernel passed 40 additional independent checks.

The seed-21 external descriptor and graph fits were rerun from the frozen memberships. All 1,070 calibrated outputs in each representative run reproduced exactly. The graph training history also matched exactly. Raw graph predictions match in their native float32 type; decimal CSV conversion produces a maximum float64 representation difference of $2.98\text{e-}8$. This is serialization precision, not a changed fitted model.

All ten primary analysis outputs reproduced byte-for-byte in a separate directory, including per-record probabilities, metrics, paired comparisons, frozen thresholds and selective-risk tables. These are execution and consistency checks performed within this project, not independent scientific peer review. The original study and any earlier public edition remain unchanged.

Location in the private follow-up folder	Contents
protocol/	Frozen questions, criteria, timestamp and hash
source_manifest.json; raw/	Pinned source revisions and private original records
code/; results/runs/	Pinned implementations, saved fits and predictions
results/analysis/	All metrics, policy settings, contrasts and diagnostics
verification/	Independent checks, representative refits and analysis replay
publication/	This report, editable figure sources and lossless masters

Remaining limits: heterogeneous source labels; unresolved assay-level lineage; conservative stereo-free exclusions; five seeds; small validation sets in controlled experiments; a single untuned graph architecture; intervals conditional on fitted models; and no prospective validation. A further computational study should freeze one candidate policy before using a new assay-resolved holdout, or explicitly report if adequate public evidence cannot be obtained.

Sources and exact study inputs

Tong et al. (2022). Blood-brain barrier penetration prediction enhanced by uncertainty estimation. *Journal of Cheminformatics*. DOI 10.1186/s13321-022-00619-2.

Tong authors repository; S-data source pinned at commit 5486ab6ab1c14cf5948f2807d3ef7c6529ea60f2.

Meng et al. (2021). A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Scientific Data*. DOI 10.1038/s41597-021-01069-5.

B3DB authors repository; source files pinned at commit 75dab1cc607bb7a03f3de5c576cffd223e767844.

Ferri and Garcia-Gomez (2026). Overrepresentation bias leads to performance overestimation in blood-brain barrier permeability prediction models: characterization and mitigation. DOI 10.1021/acs.jcim.5c02891.

Authors code for overrepresentation-bias analysis.

Literature positioning is a targeted comparison, not an exhaustive novelty certification. Published S-data performance is not directly comparable to this report: overlap filters, model definitions, calibration, splitting and policy rules differ. External source records are kept private; no blanket downstream redistribution permission is assumed.

Protocol SHA-256: 2d6e2edcb29b3ef75eb9011fd6c011ae6cf28b1a3bbe71a500d1fa7f598d973e

Original cohort-audit SHA-256:

0b1ff7b47f36370118ad150ed2dae74526c657de0283323ea1387c61493bf833

Original similarity-matrix SHA-256:

61529e3118adfb40b1633cb68bd202cc108e9fb7b37f669666575ffe4528450

The machine-readable manifest records individual source-file hashes. Full per-seed metrics, all 810 policy/population summaries and 4,050 per-seed selective evaluations accompany this report. Undefined bootstrap AUROCs in sparse secondary cohorts are explicitly retained as missing with valid-replicate counts.