

Supporting Information for ReliaMol-3D: Auditable Post-Generation Reliability Reranking for Pocket-Conditioned 3D Molecular Generation

Shenkuang Wu^{†1}, Jing Zhang^{†2}, Lianming Zhang^{*1,3}, and Pingping Dong^{*1}

¹School of Information Science and Engineering, Hunan Normal University, Changsha, China

²Faculty of Medicine, Hunan Normal University, Changsha, China

³Xiangjiang Laboratory, Changsha, China

[†]These authors contributed equally to this work.

^{*}Corresponding authors: zlm@hunnu.edu.cn; ppdong@hunnu.edu.cn

Supplementary Overview

This Supplementary Material provides additional protocol definitions, figures, tables, and data provenance information supporting the main manuscript. It extends the quantitative evidence for computational receptor-frame reliability-label construction, physical validity, chemical utility, feature-quality robustness, generator transfer, and external stress testing.

Abbreviations used throughout are: quantitative estimate of drug-likeness (QED), synthetic-accessibility proxy (SA), pan-assay interference compound (PAINS), PoseBusters (PB), random forest (RF), histogram-gradient boosting (HGB), multilayer perceptron (MLP), convolutional neural network (CNN), conditional variational autoencoder (CVAE), autoregressive generator (AR), pocket cross-validation (pocket-CV), out-of-distribution (OOD), evidence-reliability weighted fusion (ERWF), reliability-calibrated evidence fusion (RCEF), RCSB Protein Data Bank (RCSB PDB), area under the receiver-operating-characteristic curve (ROC-AUC), and area under the precision-recall curve (PR-AUC).

Protocol and Reproducibility Notes

The main manuscript is centered on auditable post-generation reranking of generated 3D molecules. All supplementary values follow the model definitions, computational endpoint, split protocols, and metric definitions stated in the main Methods section. Tables S2 and S10 summarize the sample sizes, data origins, and evaluation roles used across the reported audits.

The DiffSBDD exact-CrossDocked benchmark uses 10,000 official generated candidates from 100 pockets. The main model comparisons use seeds 11, 22, and 33 with five pocket-held-out folds per seed. The TargetDiff-family evaluation uses official generated outputs from TargetDiff, Pocket2Mol, CVAE, and AR,

totaling 38,073 candidates across 100 pockets. An additional public Database of Useful Decoys: Enhanced (DUD-E) controlled-failure stress test evaluates reliability separation under six targets and deliberately corrupted poses.

The evidence-access convention matches the main manuscript. The primary score-free ReliaMol models use ligand chemistry, internal 3D geometry, and ligand-pocket context evidence at selection time; they exclude Vina/GNINA scores, PB outcomes, direct failure indicators, and generator identity. Docking scores are used only by explicitly labeled score-only or combined-score baselines. PB is used only as an optional strict eligibility gate. Chemistry, docking, PB, diversity, and structural case analyses reported after selection are confirmation or reporting evidence rather than primary-model inputs.

Ungated top-10% results use the fixed pocket-wise budget $K_p(0.1)$. Strict PB-gated results use the effective budget $K_p^{\text{PB}}(0.1)$, apply no fallback or backfilling, and can therefore return fewer candidates than the nominal budget. Reliability and PB pass rates for gated rows are computed among candidates actually selected. Coverage-risk results refer to the ungated fixed-budget policy and are empirical operating-point analyses rather than statistical risk certificates.

In the low-quality feature-quality audit, evidence-reliability weighted fusion (ERWF) is implemented by the diagnostic RCEF network. Neural streams are standardized with training-fold statistics; a missing stream is represented by zeros and masked before softmax gating. Tree-model missing blocks are replaced by training-fold medians. Noisy geometry receives independent Gaussian perturbations with standard deviation 0.75 in standardized units, or the equivalent raw-scale perturbation for tree models. RCEF uses branch dimension 64, hidden dimension 96, 75 epochs, 0.15 evidence dropout, and a gate-entropy weight of 0.005.

Supplementary Figures

Supplementary Figures S1, S2, S3, S4, and S5 extend the threshold, physical-validity, multi-generator, structural case study, empirical coverage-risk, sensitivity, and scalability audits.

Figure S1 shows that changing the contact and clash criteria alters the prevalence of the receptor-reliability label more strongly than it alters the ranking conclusion, indicating that ReliaMol is supported by continuous geometry summaries rather than one exact cutoff. Figure S2 identifies bond-length and bond-angle checks as major contributors to the difference between chemical validity and PB physical validity; the failed-check burden also shows that a subset of candidates violates multiple criteria simultaneously. Figure S3 places this observation across generator families. CVAE has the largest geometry-failure burden and consequently provides the most difficult transfer setting, while the stronger performance on AR, Pocket2Mol, and TargetDiff is consistent with their cleaner raw geometry.

Figure S4 connects these aggregate patterns to structural inspection in a representative pocket: the selected molecules differ not only in ranking score, but also in the local geometric and receptor-contact evidence used to interpret shortlist reliability. Figure S5 then tests whether the ranking remains useful beyond the most selective operating point. Reliability decreases gradually as coverage expands, remains high under moderate geometry-feature perturbation, and requires little model-stage computation once molecular and receptor features have been constructed. Together, these figures support robustness, interpretability, and practical reranking efficiency as distinct parts of the evidence chain.

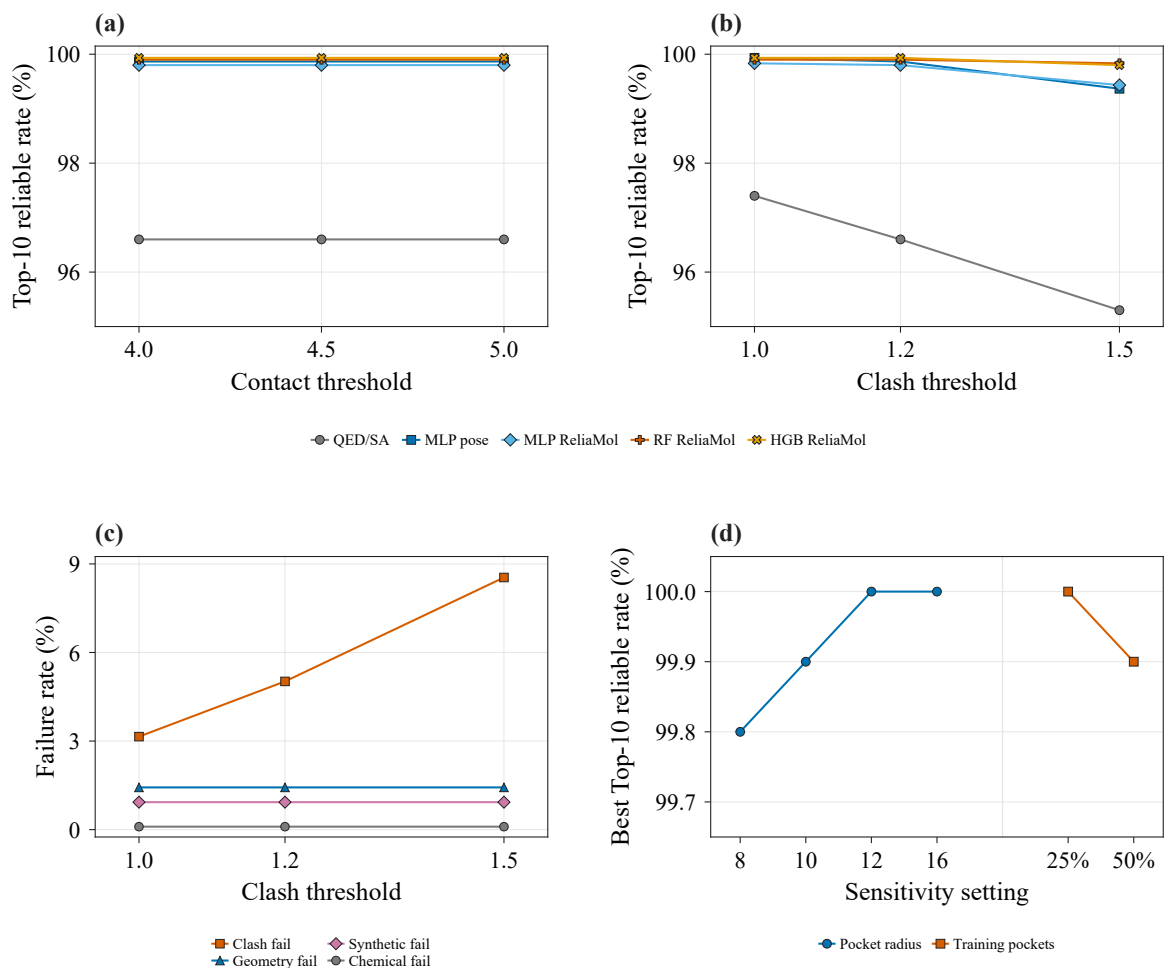


Figure S1: Sensitivity audits for the reliability label and reranking endpoint. Panels (a) and (b) report top-10% reliability under contact- and clash-threshold variation; panel (c) reports failure-component rates across clash thresholds; panel (d) reports pocket-radius and training-pocket-fraction sensitivity. Compact labels such as “RF ReliaMol” denote the corresponding score-free ReliaMol variants.

Supplementary Tables

Supplementary Tables are organized into ten evidence blocks: reliability-label rules, experimental scope, mechanism and sensitivity, evidence-weight interpretation, chemistry, physical validity, generator shift and external stress, failure examples, claim traceability, and data provenance. This organization keeps the Supplementary Material compact while preserving the numerical anchors needed for independent assessment. Tables S1, S2, S3, S4, S5, S6, S7, S8, S9, and S10 provide the corresponding evidence blocks.

Supplementary Analysis Summary

The supplementary evidence supports the central claim of the main manuscript. Generated-candidate reliability is a post-generation receptor-frame triage problem; exact receptor provenance should be controlled;

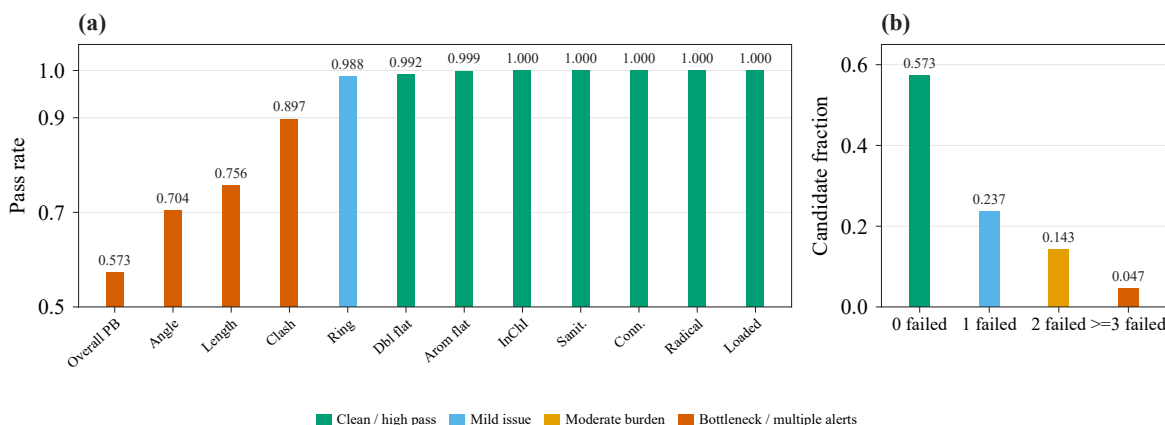


Figure S2: PoseBusters fast-check component pass rates and failed-check burden on DiffSBDD generated candidates.

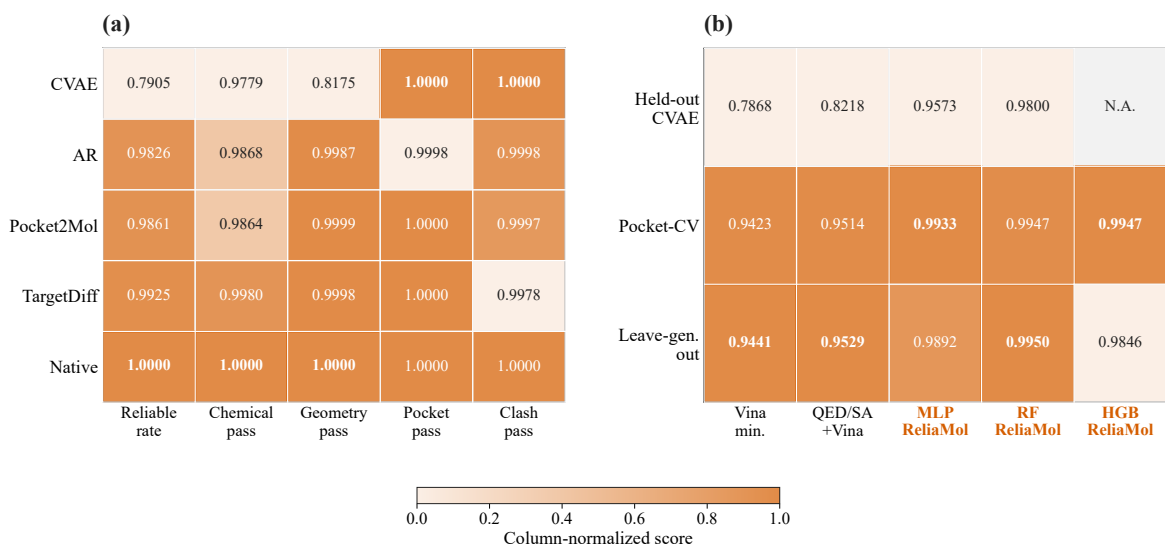


Figure S3: Supplementary scorecard for raw generator reliability, component pass rates, and cross-setting top-10% reliability. Cell values report rates, shading gives the column-normalized score, and orange tick labels mark ReliaMol rerankers. Compact figure labels such as “RF ReliaMol” denote the score-free variants RF-ReliaMol, HGB-ReliaMol, and MLP-ReliaMol used in the text.

ReliaMol improves receptor-reliability selection through receptor-aware 3D evidence; PoseBusters is used as an independent physical-validity gate; selected molecules retain chemical diversity and basic drug-likeness-rule compatibility; empirical risk remains low across the evaluated fixed-budget coverage range; coordinate-noise sensitivity remains limited after true feature recomputation; score-free reranking transfers to official TargetDiff-family generated outputs; and the public DUD-E controlled-failure stress test supports external reliability separation.

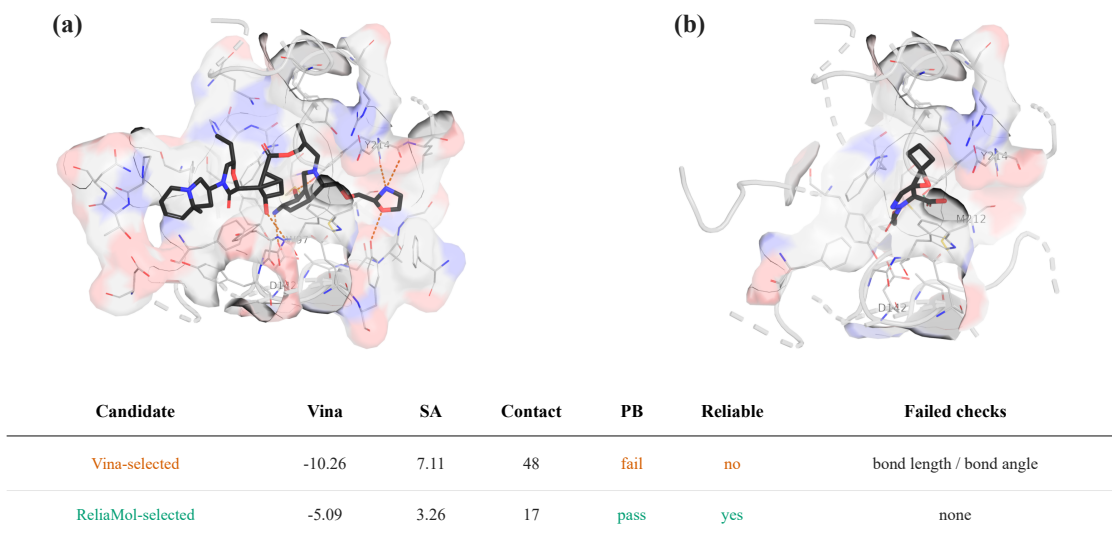


Figure S4: Extended qualitative structural case study for the representative 4z2g pocket, including two selected poses and a compact table of computational audit endpoints.

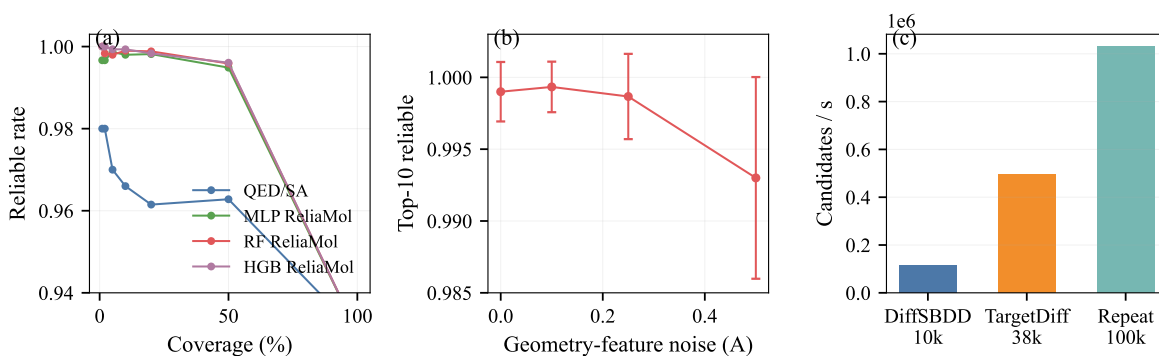


Figure S5: Empirical coverage-risk, sensitivity, and model-stage scalability audits. Panel (a) reports reliability for the ungated fixed-budget policy as a function of selected coverage; panel (b) perturbs held-out geometry distance-summary features at test time; panel (c) reports RF inference throughput from precomputed molecular and receptor features. Runtime in panel (c) isolates model inference from molecular and receptor feature construction.

Table S1: Reliability-label construction. A candidate is reliable only if all listed failure indicators are false. The thresholds are fixed before reranking experiments and are used consistently unless a threshold-sensitivity experiment explicitly perturbs them.

Failure type	Proxy rule	Interpretation
Chemical failure	RDKit sanitization fails, heavy atoms < 6 , rare element count > 0 , or absolute formal charge > 2	Removes chemically invalid or unusual generated molecules
Internal geometry failure	Missing 3D coordinates, any internal nearest-neighbor distance < 0.65 Å, pose radius < 0.8 Å, or pose radius > 8.0 Å	Removes collapsed, over-expanded, or internally clashing poses
Pocket-contact failure	Number of ligand atoms within 4.5 Å of receptor atoms is zero	Requires at least one direct ligand-pocket contact
Receptor-clash failure	Number of ligand atoms within 1.2 Å of receptor atoms is greater than zero	Removes severe ligand-receptor overlaps
Synthetic proxy failure	SA proxy > 6.8 , heavy atoms > 70 , or rotatable bonds > 22	Removes extreme molecules under a lightweight synthesis proxy

Table S2: Experimental scope and evidence tiers. The supplementary tables are organized as evidence blocks rather than one table per experiment.

Evidence block	Scale	Repeats/splits	Main numerical anchor	Claim role
DiffSBDD	10,000 candidates, 100 pockets	3 seeds $\times$ 5 folds	QED/SA 0.9660 vs ReliaMol	Primary benchmark
exact-CrossDocked	10,000 candidates, 100 pockets	Exact vs public receptor frame	0.9980–0.9993 top-10% Public RCSB PDB reliability rate 0.3816; exact-CrossDocked 0.9304	Provenance control
Feature mechanism	10,000 candidates, 100 pockets	RF diagnostic grouping	Reliability importance: ligand-pocket 0.6224, internal 0.2158, chemistry 0.1619	Why fusion works
Physical validity and docking	9,901–10,000 candidates	Strict PB gate and score-only ranking	Strict PB gate + RF-ReliaMol gives 1.0000 selected reliability and 1.0000 selected PB pass; Vina gives 0.9750 reliability	Complementary evidence
Chemistry and diversity	9,901 scored / 10,000 full candidates	3 seeds $\times$ 5 folds	RF keeps 9.88 scaffolds/pocket, Tanimoto 0.1030, PAINS-free 0.9797	Utility preservation
Sensitivity and empirical coverage–risk	10,000 candidates	Thresholds, fixed-budget coverage, noise, bootstrap	RF reliability 0.9990 at 10% coverage and 0.9959 at 50% coverage	Sensitivity audit
TargetDiff-family transfer	38,073 candidates, 100 pockets	Pocket-CV and leave-generator-out	RF 0.9947 pocket-CV, 0.9950 leave-generator-out, 0.9800 held-out CVAE	Generator shift
External DUD-E stress test	2,154 public controlled-failure poses, 6 targets	Three independent seeds, leave-one-target	MLP-ReliaMol with label-definition-independent inputs: ROC-AUC 0.9981 and top-10% reliability 0.9846	External OOD stress

Table S3: Mechanism, empirical coverage–risk, and sensitivity audit. Values are grouped by evidence stream and evaluation stressor. Coverage rows use the ungated fixed-budget policy.

Audit	Baseline / reference	ReliaMol or diagnostic result	Interpretation
<i>A. Evidence-stream diagnostic importance</i>			
Reliability label	Positive rate 0.9304	Chemistry 0.1619; internal 3D 0.2158; ligand-pocket 0.6224	Pocket-aware geometry dominates reliability selection
Geometry failure	Positive rate 0.0143	Internal 3D importance 0.7258	Internal geometry catches pose collapse and local clashes
Receptor-clash failure	Positive rate 0.0502	Ligand-pocket importance 0.8823	Pocket distances catch receptor overlaps
Synthetic proxy failure	Positive rate 0.0093	Chemistry importance 0.7865	Ligand descriptors carry synthesis-proxy evidence
<i>B. Empirical coverage–risk, calibration, and perturbation</i>			
Fixed-budget coverage 10%	QED/SA reliability 0.9660; empirical risk 0.0340	RF 0.9990; HGB 0.9993	Most selective evaluated operating point
Fixed-budget coverage 50%	QED/SA reliability 0.9628; empirical risk 0.0372	RF 0.9959; HGB 0.9961	Reliability persists beyond the top-10% subset
Calibration/bootstrap	QED/SA reference	HGB Brier 0.0218; bootstrap delta +0.0334, 95% CI [0.0177, 0.0517]	Scores are stable across pockets
Threshold sensitivity	Clash threshold 1.0–1.5 Å	Raw reliability range 0.8978–0.9479	Ranking conclusion remains stable under label-prevalence changes
True coordinate perturbation	QED/SA top-10% 0.9660	RF top-10% 0.9983, 0.9957, and 0.9900 at 0.10, 0.25, and 0.50 Å	Stable after feature recomputation
Model-stage scalability	10,000 candidates with precomputed features	RF training 0.559 s; inference 0.087 s	Reranker fitting and inference are fast

Table S4: RCEF evidence-weight audit for ERWF under low-quality evidence stress. Values are mean gate weights $\pm$ standard deviation across three seeds and five pocket-held-out folds. The weights are learned allocations over chemistry, internal 3D, and ligand-pocket context evidence; they summarize sample-level evidence use rather than probabilities of correctness.

Stress condition	Chemistry weight	Internal 3D weight	Ligand-pocket weight	Interpretation
Clean	0.317 ± 0.014	0.300 ± 0.010	0.383 ± 0.012	Pocket context receives the largest clean-evidence allocation.
Missing chemistry	< 0.001	0.437 ± 0.013	0.563 ± 0.013	Missing chemistry is suppressed and the remaining weight shifts to geometry.
Missing internal 3D	0.451 ± 0.017	< 0.001	0.549 ± 0.017	The absent internal-geometry stream is masked, while pocket context remains dominant.
Missing pocket context	0.514 ± 0.017	0.486 ± 0.017	< 0.001	The model falls back to ligand chemistry and internal geometry when pocket context is unavailable.
Noisy internal 3D	0.310 ± 0.015	0.315 ± 0.018	0.375 ± 0.015	Moderate internal-geometry noise preserves a clean-like allocation pattern.
Noisy pocket context	0.305 ± 0.018	0.288 ± 0.015	0.407 ± 0.026	The model still gives substantial weight to pocket context under moderate pocket-evidence noise.
Noisy all geometry	0.300 ± 0.019	0.303 ± 0.020	0.397 ± 0.026	Geometry noise broadens uncertainty but does not collapse the evidence allocation.

Table S5: Chemistry, diversity, medicinal-chemistry, and coordinate-perturbation audit. Panel A uses the 9,901-candidate subset with complete Vina/GNINA scores; panel B uses all 10,000 candidates unless a perturbation setting is stated. Structural-alert filters are downstream medicinal-chemistry checks with a separate optimization objective from receptor reliability. The strict PB-gated row uses the effective no-backfill budget.

Audit / method	Setting	Reliability	PB pass	Median QED	Median SA	Scaffolds / pocket	Pairwise Tanimoto	Rule/audit note
<i>A. Utility/diversity on the 9,901-candidate Vina/GNINA-scored subset</i>								
QED/SA	top-10%	0.9660	0.7640	0.7792	3.2019	9.90	0.1036	Baseline
Vina	top-10%	0.9750	0.5950	0.4848	4.1932	9.90	0.1102	Affinity-like score only
RF-ReliaMol	top-10%	0.9990	0.7380	0.5700	3.6413	9.88	0.1030	Reliability gain without diversity collapse
HGB-ReliaMol	top-10%	0.9993	0.7660	0.5744	3.4941	9.82	0.1007	Highest reliability in this subset
Strict PB gate + RF-ReliaMol	Nominal top-10%; effective budget	1.0000	1.0000	0.5841	3.5184	6.11	0.1009	No fallback or backfilling
<i>B. Medicinal-chemistry rules and true coordinate perturbation</i>								
Raw pool	100%	0.9304	–	–	–	93.32	–	Lipinski ≤ 1 : 0.9218; Veber: 0.7577; PAINS-free: 0.9740
QED/SA	top-10%	0.9660	–	–	–	9.89	–	Lipinski ≤ 1 : 1.0000; Veber: 0.9990; PAINS-free: 0.9840
RF-ReliaMol	top-10%	0.9990	–	–	–	9.79	–	Lipinski ≤ 1 : 0.9893; Veber: 0.9027; PAINS-free: 0.9797
RF-ReliaMol	0.25 Å perturb.	0.9957	–	–	–	–	–	ROC-AUC 0.8956; PR-AUC 0.9894
RF-ReliaMol	0.50 Å perturb.	0.9900	–	–	–	–	–	ROC-AUC 0.8455; PR-AUC 0.9825

Table S6: Physical-validity, docking-score, and baseline audit. PB checks and non-docking rerank rows use the full 10,000-candidate DiffSBDD pool; Vina/GNINA score-only rows use the 9,901 candidates with complete docking scores. Ungated reranking uses a fixed 10% budget, whereas the strict PB-gated row uses the effective no-backfill budget. The RF-ReliaMol PB pass value therefore differs slightly from Table S5, whose panel A is restricted to the complete-score subset.

Block	Method/check	Reliability	or component pass	PB pass	Clash-free	ROC-AUC
<i>A. PoseBusters check-level summary</i>						
PB check	Overall PoseBusters pass	0.5732	–	–	–	–
PB check	Bond lengths pass	0.7565	–	–	–	–
PB check	Bond angles pass	0.7040	–	–	–	–
PB check	Internal steric-clash pass	0.8972	–	–	–	–
<i>B. Score-only baselines and strict PB-gated reranking</i>						
Raw rerank	QED/SA	0.9660	0.7640	0.9730	0.6455	
Raw rerank	RF-ReliaMol	0.9990	0.7333	0.9990	0.9819	
Strict PB gate + rerank	RF-ReliaMol	1.0000	1.0000	–	–	
Score-only	Vina	0.9750	0.5950	0.9810	0.8559	
Score-only	GNINA empirical affinity	0.9760	0.5100	0.9870	0.8518	
Score-only	GNINA CNNscore	0.9460	0.5820	0.9720	0.6354	
Score-only	GNINA CNNaffinity	0.8790	0.4400	0.9160	0.5245	
<i>C. Additional simple baselines</i>						
Rule baseline	QED/SA + explicit clash/contact rule	0.9930	–	1.0000	0.9216	
Linear baseline	Logistic regression, ReliaMol features	0.9970	–	0.9990	0.9691	
Unsupervised baseline	Isolation Forest, ReliaMol features	0.9950	–	0.9953	0.8976	
Post-hoc upper bound	Ground-truth reliability-label ordering (non-deployable)	1.0000	–	1.0000	1.0000	

Table S7: Generator-shift and external OOD stress-test audit. The DUD-E block is a public controlled-failure setting for external reliability separation. Its controlled physical-validity proxy is specific to the injected corruption protocol and is not a PoseBusters output.

Block	Method / source	Candidates	Raw reliability / ROC-AUC	top-10% reliability	Geometry / physical-validity / note
<i>A. TargetDiff-family raw generator audit</i>					
Raw generator	CVAE	9,911	0.7905	–	Geometry pass 0.8175
Raw generator	AR	9,295	0.9826	–	Geometry failure 0.0013
Raw generator	Pocket2Mol	9,831	0.9861	–	Geometry failure 0.0001
Raw generator	TargetDiff	9,036	0.9925	–	Clash failure 0.0022
<i>B. Hard held-out CVAE generator split</i>					
Held-out CVAE	Vina-minimize	–	–	0.7868	Geometry pass 0.8028
Held-out CVAE	QED+SA+Vina	–	–	0.8218	Geometry pass 0.8418
Held-out CVAE	MLP-ReliaMol	–	–	0.9573	Geometry pass 0.9713
Held-out CVAE	RF-ReliaMol	–	–	0.9800	Geometry pass 0.9830
<i>C. Public DUD-E controlled-failure external stress test</i>					
DUD-E external	Vina ranking	2,154	ROC-AUC 0.5966	0.0093	Controlled physical-validity proxy pass 0.0370
DUD-E external	MLP pose shape	2,154	ROC-AUC 0.8217	0.3287	Controlled physical-validity proxy pass 1.0000
DUD-E external	MLP-ReliaMol, label-definition-independent inputs	2,154	ROC-AUC 0.9981	0.9846	Controlled physical-validity proxy pass 1.0000
DUD-E external	HGB-ReliaMol, label-definition-independent inputs	2,154	ROC-AUC 0.9983	–	PR-AUC 0.9920

Table S8: Representative exact-CrossDocked failure cases from the 10,000 DiffSBDD candidates. The examples link each reliability criterion to diagnostic values.

Failure class	Rate	Example	Diagnostic values	Takeaway
Chemical	0.0010	2gns-A, candidate 53	Heavy atoms = 3; rare elements = 0; formal charge = 0; RDKit sanitization passed	Too small
Internal geometry	0.0143	2rhy-A, candidate 22	Minimum internal distance = 0.168 Å; close atoms = 2; pose radius = 3.422 Å	Collapsed pose
Pocket contact	0.0000	2gns-A, candidate 20	Contact atoms within 4.5 Å = 3; minimum receptor distance = 3.636 Å	Exact receptor fixes apparent failures
Receptor clash	0.0502	5mgl-A, candidate 88	Clash atoms within 1.2 Å = 9; severe clashes within 0.8 Å = 8; minimum receptor distance = 0.103 Å	Clash-aware audit needed
Synthetic proxy	0.0093	4qlk-A, candidate 70	SA proxy = 10.645; heavy atoms = 78; rotatable bonds = 36; QED = 0.0038	Too large/flexible

Table S9: Claim-to-evidence traceability map.

Claim	Evidence anchor	Manuscript anchor
Post-generation fusion layer	Method formulation, Fig. 1, and Algorithm 1 define reranking after generation	Introduction and Method
Exact receptor provenance matters	RCSB PDB approximation vs exact-CrossDocked: 0.3816 vs 0.9304 reliability rate	Main provenance table; Fig. 2
Top- K reliability improves on DiffSBDD	QED/SA 0.9660 vs ReliaMol 0.9980–0.9993 top-10% reliability	Main DiffSBDD table; Fig. 3
Gain comes from receptor-aware 3D evidence	Internal + pocket geometry reaches 0.9990 top-10%; diagnostic importance localizes failure evidence	Main feature-ablation table; Table S3
Evidence degradation is auditable	RCEF implements ERWF and keeps 0.994 ± 0.002 top-10% reliability when internal 3D evidence is missing; missing-stream gate weights are < 0.001 for the absent stream	Main low-quality evidence table; Table S4
Selection keeps chemical diversity	RF-ReliaMol keeps 9.88 scaffolds/pocket and pairwise Tanimoto 0.1030	Table S5
Reliability remains high across evaluated coverage	Ungated fixed-budget reliability is 0.9990 at 10% and 0.9959 at 50% coverage	Method empirical coverage–risk definition; Table S3
Physical validity uses gate plus rerank	Raw PB pass is 0.5732; strict PB gate + RF-ReliaMol reaches 1.0000 PB pass and 1.0000 reliability among selected candidates with no backfilling	Main physical-validity table; Table S6
Transfer extends beyond DiffSBDD	TargetDiff-family pocket-CV and leave-generator-out; hard CVAE split improves 0.7868 to 0.9800	Main multi-generator table; Table S7
External stress separation is supported	Public DUD-E controlled-failure stress test: MLP-ReliaMol with label-definition-independent inputs reaches ROC-AUC 0.9981 and top-10% reliability 0.9846	Table S7

Table S10: Data provenance and evaluation-protocol record. The table reports scientific data origins, benchmark scale, and analytical use without implementation-specific storage details.

Evidence source	Origin and scale	Analytical role	Protocol record
DiffSBDD generated pool	Official DiffSBDD samples; 10,000 candidates from 100 pockets	Primary reliability-learning and reranking benchmark	Seeds 11, 22, and 33; five pocket-held-out folds per seed; pocket-wise top- K selection
Exact-CrossDocked receptors	Receptor coordinates paired with the 100 DiffSBDD test pockets	Primary coordinate frame for labels and receptor-aware features	Coordinate provenance is fixed before feature construction and model fitting
Public receptor structures	RCSB PDB structures corresponding to benchmark targets	Coordinate-provenance sensitivity comparison	Used only as an approximate receptor-frame comparison
TargetDiff-family pools	Official TargetDiff, Pocket2Mol, CVAE, and AR generated outputs; 38,073 candidates from 100 pockets	Generator-transfer and held-out-generator evaluation	Pocket cross-validation and leave-generator-out protocols
DUD-E stress benchmark	Six public targets; 2,154 native and controlled-failure poses	External reliability-separation stress test	Three seeds with leave-one-target evaluation
Physical-validity evidence	PoseBusters molecular and pose checks on DiffSBDD candidates	Independent physical-validity gate and component audit	Reported separately from the learned receptor-reliability label
Docking evidence	AutoDock Vina and GNINA scores on the scored DiffSBDD subset	Score-only baselines and complementarity analysis	Docking scores are excluded from the primary score-free ReliaMol feature set
Chemical-utility evidence	RDKit descriptors, medicinal-chemistry filters, Murcko scaffolds, and molecular fingerprints	Drug-likeness, alert, and diversity assessment	Evaluated on the same pocket-wise selected subsets used for reliability analysis
Statistical evaluation	Pocket-level resampling, calibration analysis, threshold sweeps, and coordinate perturbations	Uncertainty, robustness, and sensitivity assessment	Identical metric definitions are retained across comparison methods