CORAL-DOCK

Multicohort scientific evaluation of pose recovery on Astex and LIT-PCBA

Roberto Ibanez Francisco Arancibia
ISR-BIO

July 28, 2026

Abstract

CORAL-DOCK was evaluated against AutoDock Vina 1.2.7 by redocking two independent structural cohorts: 85 Astex complexes and 127 analyzable protein–ligand pairs from 129 available LIT-PCBA structures (127/129; 98.4%).

CORAL-DOCK recovered a top-1 pose with RMSD ≤ 2 Å in 60/85 Astex complexes (70.6%), compared with 56/85 (65.9%) for Python Vina. In the 127 analyzable LIT-PCBA cases, the corresponding results were 87/127 (68.5%) and 83/127 (65.4%). Paired differences were not statistically conclusive (exact McNemar: $p = 0.219$ and $p = 0.424$). Median top-1 RMSD was lower for CORAL-DOCK in Astex (1.139 vs 1.230 Å) and LIT-PCBA (1.194 vs 1.282 Å), without a significant global paired difference. Top-9 coverage favored Python Vina in Astex (88.2% vs 84.7%) and CORAL-DOCK in LIT-PCBA (88.2% vs 79.5%). Affinity estimates were highly concordant in both cohorts. The results support comparable redocking accuracy and a descriptive top-1 signal in favor of CORAL-DOCK, not universal superiority.

1 Introduction and objectives

The utility of a docking engine depends on its ability to generate conformations close to the crystallographic pose and rank them among the best-scored solutions. Astex provides a diverse structural-validation cohort; LIT-PCBA was designed for virtual screening and supplies an additional cohort of co-crystal structures. The two sources were analyzed separately to address four questions:

1. How frequently does CORAL-DOCK recover a pose close to the crystallographic reference?
2. Does the correct pose rank first or only appear among secondary candidates?
3. Is the signal observed on Astex reproduced in the structural LIT-PCBA cohort?
4. How do results vary across targets and molecular-complexity strata?

2 Materials and methods

2.1 Structural cohorts

Astex contributed 85 PDB protein structures and crystallographic SDF ligands. Both engines produced results for all 85 cases. Ligands ranged from 10 to 41 heavy atoms and from 0 to 16 active torsions.

Within LIT-PCBA, 127 crystallographic pairs across 15 targets were analyzed from 129 available structures. LIT-PCBA active and inactive compound files were not used; this study therefore evaluates redocking rather than virtual-screening enrichment. The explanatory note at the end of the paper describes the two pairs that could not be included in the analysis.

2.2 Preparation and docking protocol

Only receptor `ATOM` records were retained, excluding ligands, cofactors, ions, and waters. Hydrogens were added and PDBQT files were generated with Open Babel 3.1.0. The search center was the geometric center of the crystallographic ligand and the box measured 24 $\times$ 24 $\times$ 24 Å. Seed 181129 was used and up to nine poses were requested. CORAL-DOCK used 8000 search lanes and Python Vina used `exhaustiveness=8`. Receptors remained rigid.

2.3 Metrics and statistical analysis

The primary metric was top-1 success, defined as heavy-atom RMSD ≤ 2 Å from the crystallographic pose. Cumulative top-3, top-5, and top-9 recovery, median and mean RMSD, quartiles, affinity agreement, and complexity-stratified results were also calculated.

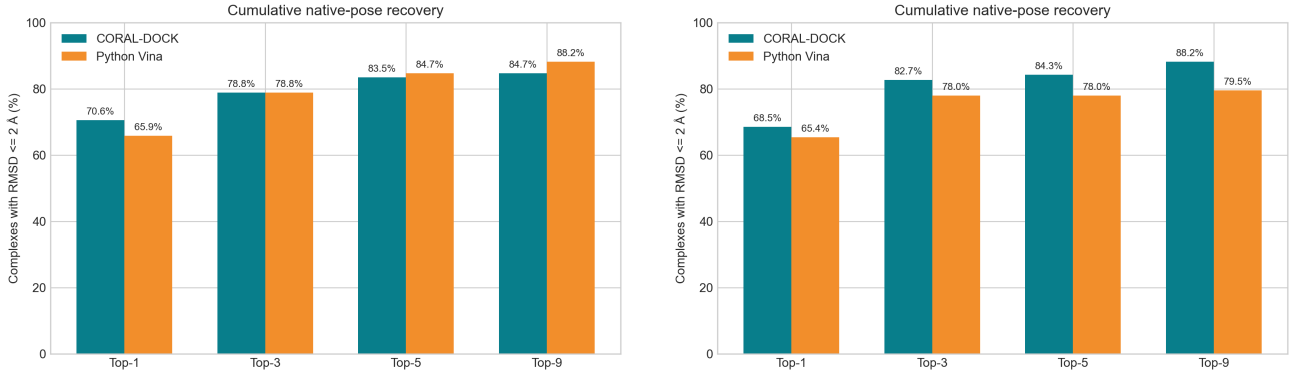
RMSD used direct atom-order matching in the receptor coordinate frame without symmetry correction. Proportions include 95% Wilson intervals. Paired top-1 success was tested with exact McNemar; paired RMSD with Wilcoxon; and the median difference with 20,000 bootstrap resamples. Cohorts were not pooled into a single rate. The LIT-PCBA analysis did not account for clustering of complexes within targets.

3 Results

3.1 Native-pose recovery

Table 1: Cumulative recovery of a pose with RMSD ≤ 2 Å.

Cohort	Level	CORAL-DOCK	Python Vina	Difference
Astex	Top-1	60/85 (70.6%)	56/85 (65.9%)	+4.7 pp
Astex	Top-3	67/85 (78.8%)	67/85 (78.8%)	0.0 pp
Astex	Top-5	71/85 (83.5%)	72/85 (84.7%)	-1.2 pp
Astex	Top-9	72/85 (84.7%)	75/85 (88.2%)	-3.5 pp
LIT-PCBA	Top-1	87/127 (68.5%)	83/127 (65.4%)	+3.1 pp
LIT-PCBA	Top-3	105/127 (82.7%)	99/127 (78.0%)	+4.7 pp
LIT-PCBA	Top-5	107/127 (84.3%)	99/127 (78.0%)	+6.3 pp
LIT-PCBA	Top-9	112/127 (88.2%)	101/127 (79.5%)	+8.7 pp



Astex LIT-PCBA, 127 analyzable cases

Figure 1: Cumulative native-pose recovery by cohort.

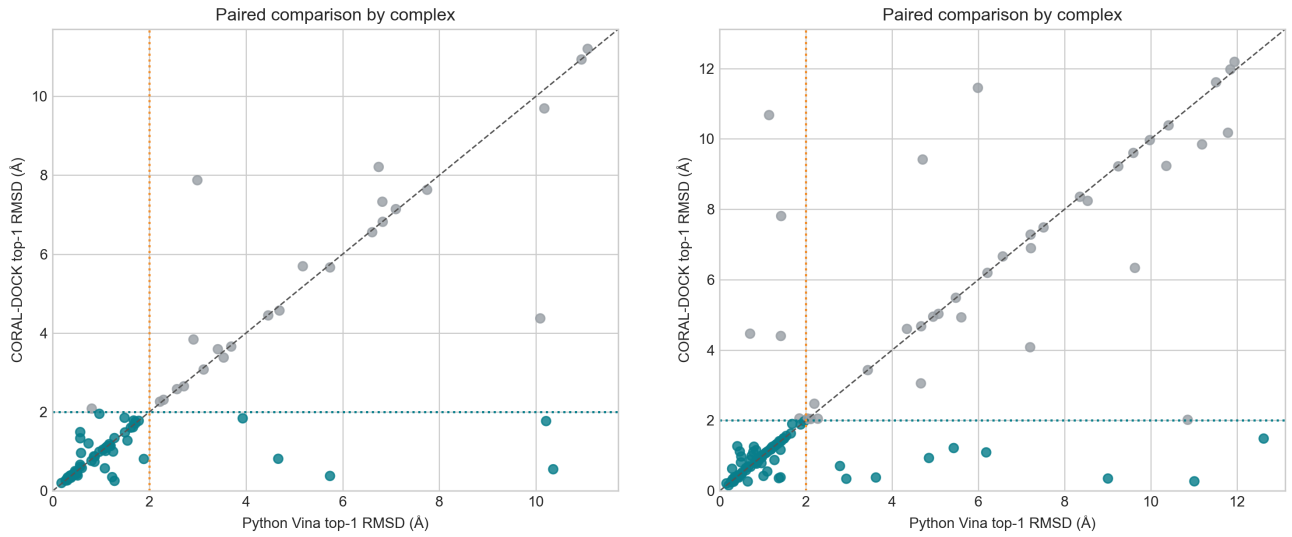
The top-1 direction was consistent and descriptively favored CORAL-DOCK. Top-k coverage was dataset-dependent: Python Vina rescued more secondary poses in Astex, whereas CORAL-DOCK progressively increased its advantage through top-9 in LIT-PCBA.

3.2 Paired top-1 comparison

Table 2: Paired contingency for top-1 success.

Cohort	Both	CORAL only	Vina only	Both fail	McNemar
Astex	55	5	1	24	$p = 0.219$
LIT-PCBA	78	9	5	35	$p = 0.424$

Neither comparison reached statistical significance. In LIT-PCBA, the nine CORAL-DOCK-only successes included cases from ESR1, GBA, MAPK1, PKM2, PPARG, and TP53; the five Vina-only successes came from ADRB2, KAT2A, PKM2, PPARG, and TP53. The repeated direction is a descriptive signal rather than sufficient evidence of superiority.



Astex LIT-PCBA

Figure 2: Paired top-1 RMSD comparison by complex.

3.3 RMSD distribution

Table 3: Top-1 RMSD statistics.

Cohort	Statistic	CORAL-DOCK	Python Vina
Astex	Median	1.139 Å	1.230 Å
Astex	Mean	2.243 Å	2.561 Å
Astex	First–third quartile	0.502–2.578 Å	0.569–3.417 Å
Astex	Lower paired RMSD	50/85	35/85
LIT-PCBA	Median	1.194 Å	1.282 Å
LIT-PCBA	Mean	2.680 Å	2.998 Å
LIT-PCBA	First–third quartile	0.664–3.241 Å	0.727–4.690 Å
LIT-PCBA	Lower paired RMSD	69/127	58/127

The paired median CORAL-DOCK minus Python Vina difference was -0.006 Å in Astex (95% bootstrap CI: -0.373 to $+0.203$; Wilcoxon $p = 0.219$) and -0.003 Å in LIT-PCBA (95% CI: -0.264 to $+0.011$; $p = 0.140$). Both intervals included zero.

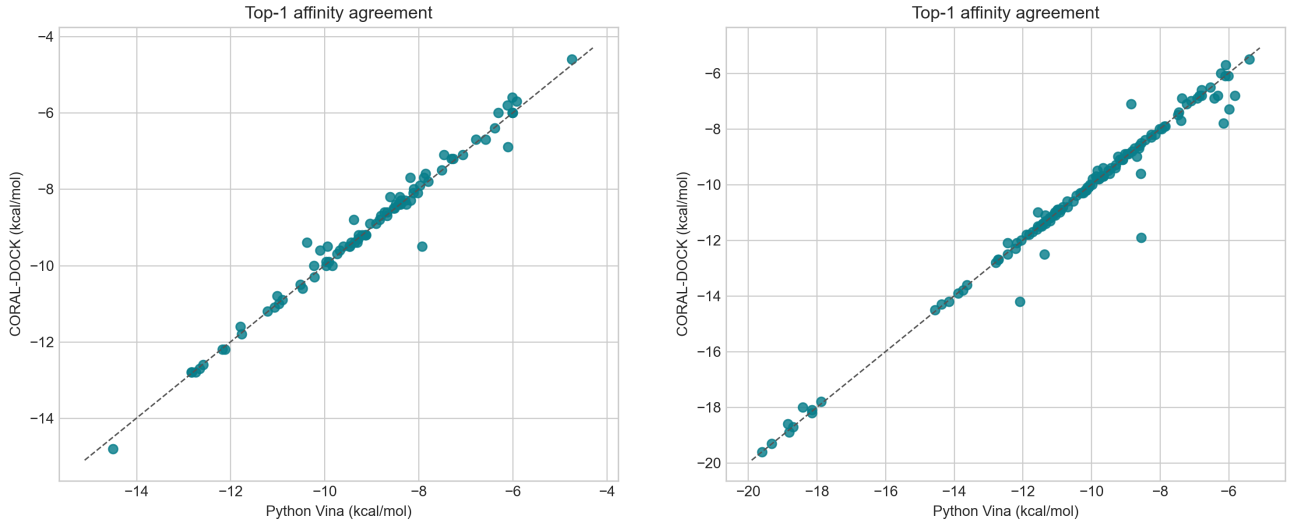
3.4 Conformational coverage and ranking

In Astex, CORAL-DOCK rescued 12 additional complexes between top-1 and top-9, compared with 19 for Python Vina. Seven cases failed with both engines even at top-9. In LIT-PCBA, CORAL-DOCK rescued 25 additional complexes and Vina rescued 18; eleven cases failed with both. These shared failures are priority candidates for reviewing protonation, tautomerism, functional waters, cofactors, receptor flexibility, and search sufficiency.

3.5 Affinity agreement

Table 4: Top-1 pose affinity agreement.

Metric	Astex	LIT-PCBA
Pearson r	0.990	0.988
Spearman ρ	0.978	0.977
Mean CORAL-DOCK – Vina delta	+0.044	-0.059
Mean absolute error (kcal/mol)	0.147	0.186
RMSE (kcal/mol)	0.276	0.478
Absolute difference ≤ 0.5 kcal/mol	81/85	118/127



Astex LIT-PCBA

Figure 3: Affinity agreement between CORAL-DOCK and Python Vina.

Energy agreement was high in both cohorts. Because the systems derive from the Vina scoring-function family, this demonstrates numerical consistency between implementations rather than thermodynamic accuracy against experimental affinities.

3.6 Target-level heterogeneity in LIT-PCBA

Table 5: Top-1 and top-9 success by LIT-PCBA target.

Target	n	C1	V1	C9	V9	Target	n	C1	V1	C9	V9
ADRB2	8	75.0	87.5	100.0	100.0	MAPK1	14	71.4	64.3	100.0	78.6
ALDH1	8	50.0	50.0	50.0	75.0	MTORC1	11	100.0	100.0	100.0	100.0
ESR1_ago	15	86.7	86.7	100.0	100.0	OPRK1	1	100.0	100.0	100.0	100.0
ESR1_ant	15	80.0	73.3	86.7	73.3	PKM2	9	55.6	44.4	100.0	88.9
FEN1	1	0.0	0.0	100.0	0.0	PPARG	15	53.3	46.7	66.7	60.0
GBA	6	66.7	33.3	100.0	66.7	TP53	6	50.0	50.0	100.0	50.0
IDH1	13	53.8	53.8	76.9	76.9	VDR	2	100.0	100.0	100.0	100.0
KAT2A	3	33.3	66.7	66.7	66.7						

C1 and V1 denote CORAL-DOCK and Vina top-1 success; C9 and V9 denote top-9. Between-target heterogeneity was larger than the aggregate between-engine difference. Percentages for targets with small n are unstable and should not be generalized.

3.7 Effect of molecular complexity

Table 6: Top-1 success stratified by ligand complexity.

Cohort	Stratum	n	CORAL-DOCK	Python Vina
Astex	0–5 torsions	33	69.7%	63.6%
Astex	6–10 torsions	35	80.0%	77.1%
Astex	> 10 torsions	17	52.9%	47.1%
LIT-PCBA	0–5 torsions	41	61.0%	56.1%
LIT-PCBA	6–10 torsions	51	76.5%	68.6%
LIT-PCBA	> 10 torsions	35	65.7%	71.4%
Astex	≤ 20 heavy atoms	34	61.8%	52.9%
Astex	21–35 heavy atoms	48	75.0%	75.0%
Astex	> 35 heavy atoms	3	100.0%	66.7%
LIT-PCBA	≤ 20 heavy atoms	19	42.1%	31.6%
LIT-PCBA	21–35 heavy atoms	86	73.3%	69.8%
LIT-PCBA	> 35 heavy atoms	22	72.7%	77.3%

Astex showed lower recovery above 10 torsions. LIT-PCBA showed no monotonic degradation with flexibility or size. These strata combine different targets and do not support a causal attribution to any single molecular property.

4 Discussion

CORAL-DOCK showed a top-1 signal that was favorable and reproducible in direction: +4.7 percentage points in Astex and +3.1 in LIT-PCBA. It also produced a lower median RMSD in both cohorts. However, paired tests were inconclusive and confidence intervals for the median difference included zero. The supported interpretation is comparability under this protocol, not universal superiority.

Top-1 represents the utility of automatically accepting the first solution; top-5/top-9 represents the ability to rescue a conformation through inspection or re-ranking. The reversal of the top-9 pattern between Astex and LIT-PCBA demonstrates that ranking and coverage should be reported separately and across more than one cohort.

LIT-PCBA was designed as a virtual-screening benchmark. This study used only crystallographic complexes for redocking: it did not use active and inactive compounds and did not calculate AUROC, PR-AUC, BEDROC, or early enrichment. Active–inactive discrimination cannot be inferred from these results.

Protocol limitations include a single seed, RMSD without symmetry correction, rigid receptors, Open Babel preparation, exclusion of waters/cofactors, and non-equivalent search-effort parameters. Target-level clustering in LIT-PCBA was not modeled.

5 Conclusions

1. Top-1 success descriptively favored CORAL-DOCK in Astex (70.6% vs 65.9%) and LIT-PCBA (68.5% vs 65.4%).
2. Top-1 differences were not statistically conclusive in either cohort.
3. Top-9 coverage favored Vina in Astex and CORAL-DOCK in LIT-PCBA.
4. Affinity estimates were numerically consistent between implementations but were not validated against experimental data.
5. The data support comparable redocking accuracy and a favorable top-1 signal for CORAL-DOCK under this protocol.

References

1. Hartshorn MJ, Verdonk ML, Chessari G, et al. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem*. 2007;50:726–741. doi:10.1021/jm061277y.
2. Tran-Nguyen VK, Jacquemard C, Rognan D. LIT-PCBA: An unbiased data set for machine learning and virtual screening. *J Chem Inf Model*. 2020;60:4263–4273. doi:10.1021/acs.jcim.0c00155.
3. 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;31:455–461. doi:10.1002/jcc.21334.
4. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model*. 2021;61:3891–3898. doi:10.1021/acs.jcim.1c00203.
5. Buttenschoen M, Morris GM, Deane CM. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem Sci*. 2024. doi:10.1039/D3SC04185A.

Explanatory note on the LIT-PCBA cohort

Of the 129 available LIT-PCBA structures, 127 could be included in the paired analysis. Pairs `IDH1--4xs3` and `MAPK1--5ax3` could not be analyzed because of a technical incompatibility while reading their PDBQT input files. Consequently, a complete between-engine comparison was not generated for those two pairs. They are not docking or pose-recovery failures and were not included in metric denominators; all LIT-PCBA results in this article refer to the 127 analyzable cases.