Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Three properties of a decoy pool decide a protein–protein rescoring comparison: the candidate-set size, its composition, and the relaxation of its near-native poses

Tingguang Sun
Guangxi University of Science and Technology
Qianwei Cheng
Guangxi University of Science and Technology
Lan Liao
lanliao@gxust.edu.cn

Guangxi University of Science and Technology

Research Article

Keywords: protein–protein docking, rescoring, benchmarking, decoy set, knowledge-based potential, candidate-set size

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11412449/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Three properties of a decoy pool decide a protein–protein rescoring comparison: the candidate-set size, its composition, and the relaxation of its near-native poses

Authors. Tingguang Sun[1], Qianwei Cheng[2], Lan Liao[1,*]
Affiliations. (1) Department of Pharmacy, School of Medicine, Guangxi University of Science and Technology, No. 2 Avenue Wenchang, Chengzhong District, Liuzhou, Guangxi 545006, PR China; (2) School of Biological and Chemical Engineering, Guangxi University of Science and Technology, No. 2 Avenue Wenchang, Chengzhong District, Liuzhou, Guangxi 545006, PR China.
Corresponding author. Lan Liao, Department of Pharmacy, School of Medicine, Guangxi University of Science and Technology, No. 2 Avenue Wenchang, Chengzhong District, Liuzhou, Guangxi 545006, PR China; e-mail: lanliao@gxust.edu.cn
ORCID. Tingguang Sun: 0009-0008-1909-404X; Lan Liao: 0009-0000-3988-7129

Abstract

A rescoring benchmark compares scoring functions on a fixed decoy set; neither how many poses are kept nor their character is usually reported. We ask which pool properties decide whether such a comparison is well posed. Three pools were compared under one protocol: a perturbation pool built from the bound pose (all 22 rigid untrained complexes of Benchmark 5), the same perturbation on unbound structures (15 complexes), and the published ZDOCK 3.0.2 decoy set, a real Fourier output (14 complexes). The first two differ only in the structures, the second and third only in the pose generator. Only the bound pool supports the claim that a knowledge-based potential beats a shape-complementarity proxy: there it wins 21 of 22 complexes at $K = 20$ ($p < 0.0001$). The same perturbation on unbound structures removes most of it (8 of 15, $p = 0.79$), and on the search output no advantage remains (all $p \geq 0.09$). The cause is measurable: the potential's score is nearly monotonic in the clash count, and only the bound pool supplies clash-free near-native poses, so elsewhere it penalizes the poses it is asked to find. On the search output, the program's own rescoring algorithm is at chance at $K = 20$ and helps only from $K = 400$; densifying the search fifteenfold leaves the pool unchanged. The comparison is well posed only when the size of the candidate set, its origin and composition, the relaxation of its near-native poses, the measure and the null are stated.

Keywords: protein–protein docking, rescoring, benchmarking, decoy set, knowledge-based potential, candidate-set size

1 Introduction

Protein–protein docking is usually run in two stages. An initial search proposes a large number of candidate orientations, and a scoring function ranks them [1, 2]. The search itself is cheap and geometric — most often a shape-complementarity correlation evaluated by fast Fourier transform [3, 4] — so the poses it returns are only approximately correct, and a second, more expensive function is used to rescore and rerank a subset of them [2]. A knowledge-based statistical potential is a standard choice for that second pass: it scores a pose from the frequencies with which its protein–protein atom-type pairs occur at each separation in a database of known complexes [5, 6, 7, 8].

The claim that such a potential improves rescoring is almost always supported the same way. A decoy set is assembled, every decoy is scored by the existing channel and by the new one, and the two are compared with a summary statistic. The published comparison is made on a fixed pool, and the reader is told neither how many poses the rescorer saw nor what those poses were like.

Both omissions matter, and their relative weight is not obvious. The number of poses is a protocol choice: a transform-based program returns a ranked list whose length is set on the command line, so a rescoring step in a real pipeline sees tens to hundreds of poses, whereas a benchmark is usually evaluated on the complete decoy set, in which nearly every pose is far from native. But the character of the poses is not a free parameter either. If the pool was produced by perturbing the native pose, then its smallest candidate sets are near-native by construction, and a comparison made there is a comparison of the ability of two functions to order nearly correct poses. That is a different question from the one a docking pipeline asks: whether the correct pose can be recovered at all from an output that is mostly wrong. The literature reports that the distinction is real in practice: on the 55 complexes added in Docking Benchmark 5, ZDOCK places a correct model within its top ten predictions for only 38% of cases, and for 50% of the rigid-body subset [9]. For the other 62% of the added cases, and for the other half of the rigid-body cases, the compressed set a pipeline would hand to a rescorer therefore contains no near-native pose at all.

In the protein–protein setting the closest published work to the question asked here is consensus rescoring. CONSRANK ranks a docking ensemble by the conservation of its inter-residue contacts [10], and a systematic evaluation of CONSRANK-like functions on the ZDOCK decoy set found that the size of the subset used to estimate contact and residue frequencies has a major influence on the number of successfully rescored cases [11]. What separates that work from the present study is which quantity is varied. In that work the object being reranked is held fixed at the top 2,000 ZDOCK poses and the frequency-estimation subset changes, whereas here the candidate set handed to the rescorer is itself the variable. A later algorithm, Iter-CONSRANK, filters an ensemble iteratively and uses the shrinking ensemble as a step of the method [12]; that is a different use of the same parameter rather than an analysis of it. The reference programs of the CoDock family rescore a candidate set the way a pipeline does [13, 14], and a recent survey of protein–protein scoring functions documents how hard it is to compare such programs at all [15, 16]. The CAPRI criteria [17, 18] supply the ground-truth definitions used in Section 2.3.

Two further gaps are left open. First, when a benchmark comparison is reported, the conclusion is usually expressed through a single summary number, and the metric that measures overall monotonicity — a rank correlation over the whole pool — is not the metric that measures whether the best pose was found. Truchon and Bayly showed that the choice of metric changes the apparent ranking of virtual-screening methods [19], and Zhao and co-workers described a seesaw effect in which over-weighting early recognition reduces the power of a metric to detect it [20]; a general survey of protein–protein scoring functions reports the same difficulty [15]. Whether the two measures can select different winners in protein–protein rescoring, as a function of the candidate-set size, has not been asked. Second, an improvement in ranking is usually reported without a null: the result is an absolute value, and there is no statement of what the rescorer would have had to beat. Comparing against a deliberately destroyed score — sign-reversed, or shuffled across poses — turns that into a test.

This study treats the candidate-set size and the composition of the decoy pool as variables of the same standing as the choice of scoring function, and asks whether the conclusion survives. Fig. 1 summarizes the design. The poses are held fixed, so every displacement is attributable to the protocol and not to sampling. Two decoy pools are used side by side: a perturbation pool, which is the model the field normally uses, and a published real docking output, which is what a pipeline actually produces. Five scoring channels are compared at a grid of candidate-set sizes on each pool, and two ranking-quality measures are reported together. A formal injection gate is imposed on every channel at every candidate-set size (Section 2.5), and the whole pipeline is recomputed by an independently written implementation (Section 2.6). The analysis is descriptive. The complexes are replicates of one benchmark rather than a random sample of protein–protein complexes, and the conclusions are proof-of-principle observations on these pools, not statistical inference about all docking systems.

2 Methods

2.1 Complexes and the decoy pools

Perturbation pool. Docking Benchmark 5 contains 230 complexes [9]. A complex was retained if its unbound and bound interfaces could be carried through the whole pipeline without a missing chain or a failed atom typing, and if the interface RMSD between the unbound and bound structures was below 1 Å, which restricts the study to rigid-body cases. Of these 230 complexes, 194 have between 1500 and 6000 atoms and 36 do not, and 88 are rigid. The potential was trained on 182 of them, including 65 of the 88 rigid complexes, so 23 rigid complexes are untrained: 12 inside the window and 11 outside it. The split is by complex, so neither structure of a test complex enters training. One of the eleven, 2OOB, has 864 atoms and is the only complex of the benchmark below 1000; it is dropped as the single outlier of size, and no upper limit is imposed, so six of the ten out-of-window test complexes are above 6000 atoms. The test set is therefore the 22 rigid-body complexes the potential was not trained on — 12 inside the window and 10 outside it, four of the ten below 1500 atoms and six above 6000 — and it is every such complex of the benchmark. The set is complete, and no complex was added or dropped after a result was seen: 1AY7, 1BJ1, 1E96, 1EFN, 1EWY, 1EXB, 1JPS, 1KTZ, 1NCA, 1SBB, 1T6B, 1TMQ, 1ZHI, 2AJF, 2GAF, 2JEL, 3BIW, 3LVK, 3PC8, 3VLB, 4H03 and BOYV.

For each complex the receptor is the bound chain A and the ligand is the bound chain B. Decoys are rigid-body perturbations of the ligand about the centroid of its interface residues: the rotation is built from the Rodrigues formula in double precision so that it is exactly orthogonal, and the perturbation scale is drawn log-uniformly over [0.15, 30] Å and converted to a rotation angle through the radius of gyration of the ligand. Four hundred decoys are generated per complex with a fixed seed. Decoy 0 is the identity, which is the native pose and serves as the zero point. It is excluded from every candidate set and from every ranking measure, and it enters only the pool-composition rows of Table 1.

ZDOCK pool. The published ZDOCK 3.0.2 15-degree decoy set for Docking Benchmark 4 [21, 22] is a real docking output: 3600 predictions per complex, the complete rotational search at 15-degree sampling, with the interface Cα RMSD of every prediction against the bound structure distributed with the set. Fifteen of the 22 complexes above are covered by this set. One of them, BOYV, has no near-native pose anywhere in its 3600 predictions (the closest is 5.61 Å at 15°), and is therefore excluded. A comparison at a given K presupposes that a near-native pose exists in the candidate set, and the decoy set itself lists BOYV among the cases that ZDOCK fails on. Fourteen complexes remain: 1AY7, 1BJ1, 1E96, 1EFN, 1EWY, 1JPS, 1KTZ, 1NCA, 1SBB, 1T6B, 1TMQ, 1ZHI, 2AJF and 2JEL.

Two differences from the perturbation pool are intrinsic to using real docking output and are not removed: the docked structures are the unbound rather than the bound structures, and the pose ensemble is a Fourier search rather than a perturbation. Both are properties of the regime the study is about: Section 2.7 builds a third arm that separates them, and Section 4.3 returns to what they do and do not explain.

Poses were reconstructed from the ZDOCK output files and the mark_sur-processed input structures distributed with the decoy set. The reconstruction follows the two code paths of create_lig (distributed with ZDOCK [21]), which differ according to the switch flag in the output header. It was validated against the compiled create_lig binary: for eight complexes and four predictions each, the reconstructed ligand coordinates agree with the binary to at most 5.3×10^{-4} Å, which is within the rounding of the three-decimal PDB output and therefore exact. Two further checks confirm the alignment of the pose list with the distributed RMSD list: the output file is ordered by decreasing ZDOCK score, and for 1BJ1 the reconstructed first prediction has the same interface RMSD (1.19 Å) and the same number of predictions within 2.5 Å (45) as the statistics file distributed with the decoy set.

The ground truth on the ZDOCK pool is the interface Cα RMSD distributed with the decoy set. Using the distributed value keeps the replication independent of the coordinate-based recomputation, and the recomputation is unavailable for this pool because the unbound files retain the original PDB chain identifiers while the bound reference in our benchmark copy uses renumbered chains, so an unambiguous atom correspondence cannot be built. The consequence is that only the interface RMSD is available on this pool; the CAPRI quality class, the fraction of native contacts and the ligand RMSD are not, and the analysis of Section 3.5 consequently rests on the interface RMSD alone.

Table 1 quantifies the difference.

2.2 Scoring channels

Five channels are scored on the same pose coordinates, and all five are defined in Table S1.

Shape-complementarity proxy. $-\text{Nclose} + 20 \cdot \text{Nclash}$, where Nclose counts receptor–ligand atom pairs closer than 5.0 Å and Nclash those closer than 2.4 Å. This stands for the pre-existing fast-Fourier-transform score that a rescoring step is meant to improve on. It is a proxy, not the native score of any program. It is a controlled stand-in for shape-complementarity scoring, and it does not reproduce every feature of the production FFT-based docking score functions it represents. Section 4.4 states which aspects it omits.

Knowledge-based potential, four reference states. A distance-dependent atom-pair potential of mean force, $E(r) = -kT \ln[H_{obs}(r)/H_{exp}(r)]$, built from the 182 training complexes with a coarse nine-type protein classification, with a 12 Å cutoff in 0.5 Å bins and a value capped at ±8 kcal/mol. The four reference states differ only in Hexp, so any difference between them is attributable to the reference state alone: spherical uniform density (the classical form), the global shape or mean field, a DFIRE-type power law with exponent 2.5 [23], and the global shape with an added hard-core exclusion below 2.8 Å. The mean reference state is the primary channel; the other three support the ablation of Section 3.9.

Random ranking and oracle. A random ordering, and an ordering by the interface root-mean-square deviation (i-RMSD), provide the null and the ceiling. Neither is a scoring function, and both are reported only in Section 3.10.

No score is converted into a binding affinity or into a ratio anywhere in this study: every reported quantity is a rank, a rank correlation, a count or a standardised difference derived from ranks.

2.3 Candidate sets and the two ranking-quality measures

For every complex and every candidate-set size K, the candidate set is the K poses the pre-existing channel — the shape proxy — ranks first, with ties broken by input order and the native pose excluded. Two measures are then computed inside that set. Every channel is scored on this same set, so a pose redundancy common to all of them acts on each of them equally and cannot by itself produce a difference between two channels.

The *pooled rank correlation* is the Spearman correlation between a channel's score and the i-RMSD over the whole candidate set, averaged over complexes. It measures overall monotonicity and is dominated by the large number of grossly incorrect poses.

The *best-pose rank* is the position a channel assigns to the pose with the smallest i-RMSD, taken as a median over complexes. It measures whether the one pose a screen wants is placed near the top, which is the quantity a docking pipeline consumes.

The two measures are reported together because they ask different questions; Section 3.3 and Section 3.6 show that they select different channels.

2.4 The paired test

For each K the two channels are compared complex by complex on the best-pose rank. A complex is a win for the knowledge-based potential if the potential assigns the nearest-native pose a smaller rank than the shape proxy, a loss if it assigns a larger one, and a tie if the ranks are equal. The win and loss counts are converted into a two-sided exact binomial p value. The test treats the complexes as independent replicates of one benchmark, which is an assumption of the design; Section 4.6 states what follows from it and what does not.

All p values reported here are two-sided exact binomial tests. No correction for multiple comparisons is applied. The test is evaluated at every candidate-set size on every pool, giving sixty-nine cells across the scans of Sections 3.2, 3.5, 3.6 and 3.12 and in Tables S7, S12 and S17–S18. Under a Bonferroni correction at that count, only the perturbation-pool cells at $K = 20$ and $K = 50$ ($p < 0.0001$) would remain significant on their own. The conclusions are therefore not read off single cells. They rest instead on the trend of the p values with the candidate-set size K within each pool — increasing on the bound pool and decreasing on the two unbound pools (Section 3.6); the individual cells between $p = 0.01$ and $p = 0.05$ are treated as descriptive, and no claim rests on one of them alone. The paired test is reported with an effect size as well as a p value, because at this sample size the two answer different questions. The effect size is the standardised mean of the per-complex difference in best-pose rank, dz, signed so that a positive value means the potential ranks the nearest-native pose higher; the proportion of complexes in which the potential ranks that pose higher is reported with it. Because no hypothesis was registered in advance, these p values are descriptive summaries of the win–loss split within each tested set, not inference about a population of protein–protein complexes.

2.5 The injection gate

An improvement in ranking is a positive claim, and a positive claim needs a null. The null used here is an injection gate: a null-control procedure that compares the real scoring result against sign-reversed and shuffled versions of the same scores, and asks whether the channel still carries ranking signal about the nearest-native pose. For every channel and every K the score was destroyed in two ways: its sign was reversed, and its values were shuffled across poses. The shuffle is repeated 200 times with a fixed seed and the resulting best-pose ranks are averaged. A cell passes when the mean rank of the nearest-native pose under the true score is below 70% of the larger of the two mean destroyed ranks; it fails otherwise. The larger of the two is used so that a failure is a conjunction over both nulls: a cell fails only when the true score is worse than 70% of the sign-reversed score and worse than 70% of the shuffled one. The criterion is deliberately a margin criterion rather than a significance test: if the shuffled score does as well as or better than the true one, the channel carries no ranking information about the near-native pose at that K. Reporting its absolute value as a result would then be

unwarranted. The gate is applied to all five channels at all candidate-set sizes on both pools. The 70% threshold is a chosen convention; sensitivity across alternative thresholds is reported in Table S10.

2.6 Independent recomputation, determinism and cross-checks

Three independent checks support the numbers reported here. The candidate-set sweep and the per-complex tables were recomputed independently, in a second environment and on a second machine, and with separately written code for every statistic (Python 3 with NumPy throughout), and agree field for field. The atom-typing classifier and the potential were reimplemented independently and agree over the whole training set. Deliberate faults were injected into both, and the checks detected them. Every run of the pipeline is bit-exact, and input structures are validated before scoring. The fields covered, the counts and the outcomes are listed in Table S2, which also records one input-structure corner case and the coverage limit it implies.

2.7 The control arm: the same protocol on the structures the search was given

The ZDOCK arm of Section 2.1 differs from the perturbation arm in two ways at once, because its structures are unbound rather than bound and its poses come from a Fourier search rather than a perturbation. It therefore cannot say which of the two removes the effect, and a third arm is built to separate them. The perturbation protocol was applied again, unchanged, to the structures the search is given: the unbound receptor and the unbound ligand distributed with the ZDOCK decoy set, in the placement the benchmark supplies them in. Fifteen of the twenty-two complexes are covered by the decoy set, and all fifteen are used.

Three properties of that placement were checked before it was used. The two partners are given in contact — 54 to 125 ligand atoms lie within 5 Å of the receptor, complex by complex — so the input is a docked arrangement and not an arbitrary frame. Two distances agree with the bound arm to better than 1 Å complex by complex: the ligand interface centroid from the coordinate origin, and the median distance of the ligand atoms from it. The rotation convention of the generator therefore means the same thing in both arms. And the atom sets reproduce the bound complexes, complex by complex, to within a few atoms. Two further properties are reported because they turn out to matter, and both are measured rather than assumed. The placement the unbound arm treats as the reference lies a median of 0.70 Å from the crystal complex, and it carries a median of 17 atom pairs closer than 2.4 Å; the reference placement of the bound arm carries none.

The truth for this arm is the displacement from that reference placement, which is the definition the bound arm already uses: there the reference placement is the crystal complex and the identity pose reproduces it exactly, so in both arms the quantity being ranked is the distance from the pose the protocol starts from. The evaluator is a second, independent implementation of the i-RMSD definition rather than a reuse of the one used elsewhere. It computes the interface and the backbone mask separately for the receptor and for the ligand and only then combines them. In several of the unbound files the two partners share a chain letter and a residue numbering — six of fifteen, with up to 230 shared residue keys — so a residue key is not a unique identifier across the two sides, and a chain-name-based evaluator would silently mix the two partners. The two evaluators are compared on the bound arm over its twelve in-window complexes and all 400 poses of each: the largest disagreement is 0.00 Å.

Score columns, candidate-set construction, the K grid, the paired test, the injection gate and the two ranking-quality measures are identical in all three arms. The arms differ only in the input structures and in how the poses were produced.

3 Results

3.1 The pools are not the same object at the same K

Table 1 is the reference point for this study. In PB, 47.4% of all poses are within 2 Å of the reference placement and the twenty poses the shape proxy ranks first have a median i-RMSD of 0.37 Å: the compressed candidate set is essentially the reference pose. In PU, which uses the identical generator on unbound structures, the same two numbers are 46.4% and 0.49 Å, and the compressed set at K = 20 is 91.9% near-native — close enough that composition alone cannot explain what Section 3.6 finds. In SU, 0.12% of all poses are within 2 Å, the twenty poses the shape proxy ranks first have a median i-RMSD of 15.9 Å, and the compressed candidate set at K = 20 contains near-native poses at a rate of 0.4%, against 100% in PB (Fig. 2; Table S8 gives the composition complex by complex).

The pools also differ in size, and that difference compounds the first: the two perturbation pools hold 400 poses, so K = 400 is the entire pool, whereas the ZDOCK pool holds 3600 and K = 400 is its top ninth. A comparison at equal K is therefore not a comparison of equal conditions, and no single K is privileged between them. What the three pools share is the protocol: the same channels, the same candidate-set construction, the same two measures and the same gate.

Table 1 The three decoy pools.

property	PB, perturbation on bound	PU, perturbation on unbound	SU, ZDOCK search on unbound
complexes	22	15	14
poses per complex	400	400	3600
poses in total	8800	6000	50400
near-native share of the pool, ≤ 2 Å	47.4%	46.4%	0.12%
near-native share of the pool, ≤ 2.5 Å	52.1%	50.8%	0.25%
median of the closest pose per complex	0.00 Å	0.00 Å	1.36 Å
top-20 median, interface RMSD	0.37 Å	0.49 Å	15.90 Å
rank of the nearest-native pose	46	73	869
rank expected from a random ordering	200	200	1800
near-native share at K = 20	100.0%	91.9%	0.4%

PB is built by rigid-body perturbation of the bound pose, PU by the identical perturbation of the unbound structures the search is given, and SU is the published ZDOCK 3.0.2 15-degree decoy set for Docking Benchmark 4. "Near-native share" is the fraction of all poses whose interface RMSD is within 2 Å. "Top-20 median" is computed in two steps: for each complex, the median interface RMSD of the twenty poses the shape proxy ranks first; then the median of that value over the complexes. "Rank of the nearest-native pose" is the median, over the complexes of the pool, of the position the shape proxy assigns to the closest pose; in PB and PU that pose is the identity pose, in SU the closest prediction, and the value expected from a random ordering is given for reference. The composition rows cover the pool as generated, the identity pose of each complex included; the last row is the candidate set of Section 2.3, from which that pose is excluded. PB in this table is the full 22-complex pool; Sections 3.6 to 3.9 use the fifteen-complex subset that the ZDOCK decoy set covers.

The three pools are not the same object at the same K, and the two perturbation arms are not the same object either. In PB the twenty poses the shape proxy ranks first are essentially at the native pose; in PU they are a fraction of an angstrom further away; in SU they are more than 15 Å away, and the compressed candidate set contains almost no near-native pose at all. The comparison at equal K is therefore between different objects, which is the point of this table. One further asymmetry has to be read with the table: the interface RMSD of the two perturbation pools is recomputed from the coordinates as a backbone RMSD over interface residues, whereas the interface RMSD of the ZDOCK pool is the Cα measure distributed with the decoy set (Section 2.1). The columns are therefore not all measured with the same ruler, and only comparisons within a column are read.

3.2 On the perturbation pool the candidate-set size moves the verdict

Against the shape proxy, the knowledge-based potential places the nearest-native pose higher — that is, at a smaller rank — at every K (Table 2). The paired test over the 22 complexes passes at $p < 0.0001$ at K = 20 and K = 50, at $p = 0.0009$ at K = 100 and at $p = 0.0169$ at K = 200. It stops passing at K = 400 ($p = 0.0525$), so the reversal of the conclusion lies between K = 200 and K = 400. The same comparison in effect-size terms changes more than the p value suggests. The standardised mean gain in best-pose rank, dz, is +2.1 at K = 20 and +0.1 at K = 400, while the median per-complex gain is still 20 ranks at K = 400. What is lost at the large candidate sets is therefore not the typical size of the gain but its consistency across complexes.

The mechanism is visible in the same table. The shape proxy places the nearest-native pose at median rank 15.5 at K = 20 and at 45.5 from K = 100 onward, while its pooled correlation rises from -0.304 to a maximum of +0.789 at K = 200 and is +0.708 at K = 400. A pooled correlation rewards the rejection of grossly incorrect poses, which is easy and improves as the pool fills with them; the rank of the nearest-native pose measures the ordering of nearly correct poses, which is hard and does not improve. The candidate-set size therefore determines which of the two jobs a comparison measures, and Fig. 3 shows the three quantities for all three arms side by side.

Where the significance boundary falls is set by the size of the test set as well as by K: the p value follows the count of complexes on which the potential wins, so the boundary sits at a smaller K when fewer complexes are analysed. This is a shift in the boundary and not a change of conclusion, and it carries the same message as the rest of the study — the verdict is a property of the protocol, and the size of the test set is part of the protocol.

Table 2 Candidate-set size on the perturbation pool (22 complexes, 400 poses each).

K	poses	i-RMSD ≤ 2 Å	rho, shape	rho, potential	best rank, shape	best rank, potential	rank gain	paired W/L	p (two-sided)
20	438	100.0%	-0.304	+0.882	15.5	1.0	14.5	21/1	< 0.0001

K	poses	i-RMSD ≤ 2 Å	rho, shape	rho, potential	best rank, shape	best rank, potential	rank gain	paired W/L	p (two-sided)
50	1087	100.0%	-0.199	+0.870	33.5	4.0	29.5	21/1	< 0.0001
100	2178	99.6%	+0.121	+0.767	45.5	11.5	34.0	19/3	0.0009
200	4378	80.4%	+0.789	+0.498	45.5	16.0	29.5	17/5	0.0169
400	8778	47.3%	+0.708	+0.387	45.5	16.0	29.5	16/6	0.0525

Each complex is ranked by the shape proxy and the top K poses are kept; the native pose, where it ranks inside the top K, is removed and not replaced (Section 2.3), so the pose count is below 22 times K. Inside that set the shape proxy and the knowledge-based potential (mean reference state) are compared on the rank each assigns to the nearest-native pose. Larger rank is worse. 'Rank gain' is the difference between the two medians of its row, not the median of the per-complex gains reported in the text (Section 3.2). The paired column counts the complexes on which the potential is better or worse, and the last column is the two-sided exact binomial p value for that split.

3.3 At large K the pooled correlation and the best-pose rank pick different channels

On the perturbation pool the channel with the highest pooled correlation is the knowledge-based potential at K ≤ 100 and the shape proxy at K ≥ 200, while the channel that places the nearest-native pose highest is the potential at every K (Table 3). At K = 200 and K = 400 the two measures therefore select different channels, and the disagreement is not a tie-breaking artifact: the shape proxy's pooled correlation at K = 400 is +0.708 against +0.387 for the potential, while its median best-pose rank is 45.5, against 16.0 for the potential.

The practical reading is that "which channel is better" is not a well-posed question until the measure is named. A benchmark that reports only the pooled correlation will prefer the shape proxy at large K; a benchmark that reports only the best-pose rank will prefer the potential. Both are reporting a real property of the same two functions on the same data.

Table 3 The two ranking-quality measures.

pool	K	highest pooled rho	smallest best-pose rank	conflict
perturbation	20	mean	mean, mhc	no
perturbation	50	mean	mean	no
perturbation	100	mean	mean	no
perturbation	200	shape	mean	yes
perturbation	400	shape	mean	yes
ZDOCK	20	sph	mean, pow25	no
ZDOCK	50	sph	sph	no
ZDOCK	100	sph	sph	no
ZDOCK	200	sph	pow25	no
ZDOCK	400	sph	mean	yes

For each pool and K the table gives the channel with the highest pooled rank correlation, the channel that places the nearest-native pose highest, and whether the two measures select different channels. The last column applies the definition used in the text: the measures conflict when the shape proxy has the higher pooled correlation of the two and the potential has the better best-pose rank of the two. It is a statement about those two channels only, so it cannot be read off the two preceding columns, which name the best of all five channels; the ZDOCK pool at K = 400 is a case in point, where sph has the highest pooled correlation and the conflict is nevertheless recorded.

3.4 The injection gate fails only where the signal is absent

On the perturbation pool three of the twenty-five channel-by-K cells fail, and they fail in a pattern: all three are the shape channel, and all three are at K = 20, 50 and 100 — the compressed candidate sets in which, as Table 2 shows, the shape proxy's ordering is negatively correlated with the truth (Table 4). At K = 200 and K = 400 the same channel passes, and both are K values at which its pooled correlation is clearly positive. If the test were broken, the failures would not be confined to one channel and to the small candidate sets. The concentration of the failures is what establishes that the test has resolving power, and it independently confirms the K dependence found in Section 3.2.

Resolved per complex, the same pattern appears (Table 5; the per-complex detail is in Table S4): the shape channel passes on 4, 6, 13, 22 and 22 of the 22 complexes at K = 20, 50, 100, 200 and 400 (Fig. 4). The per-complex criterion is not the 70 percent rule of the pooled cell test: on a given complex a channel passes when its true rank beats both destroyed ranks, so the per-complex counts and the pooled cells are two separate tests and are not expected to agree.

Table 4 The injection gate on both pools.

pool	K	shape	sph	mean	pow25	mhc	passes
perturbation	20	FAIL	OK	OK	OK	OK	4/5
perturbation	50	FAIL	OK	OK	OK	OK	4/5
perturbation	100	FAIL	OK	OK	OK	OK	4/5

pool	K	shape	sph	mean	pow25	mhc	passes
perturbation	200	OK	OK	OK	OK	OK	5/5
perturbation	400	OK	OK	OK	OK	OK	5/5
ZDOCK	20	FAIL	FAIL	FAIL	FAIL	FAIL	0/5
ZDOCK	50	FAIL	FAIL	FAIL	FAIL	FAIL	0/5
ZDOCK	100	FAIL	FAIL	FAIL	FAIL	FAIL	0/5
ZDOCK	200	FAIL	FAIL	FAIL	FAIL	FAIL	0/5
ZDOCK	400	FAIL	FAIL	FAIL	FAIL	FAIL	0/5
ZDOCK	1000	FAIL	FAIL	FAIL	FAIL	FAIL	0/5
ZDOCK	2000	OK	FAIL	FAIL	FAIL	FAIL	1/5
ZDOCK	3600	OK	FAIL	FAIL	FAIL	FAIL	1/5

For every channel and candidate-set size the score is sign-reversed and, separately, shuffled across poses (200 permutations, fixed seed, ranks averaged); the cell passes when the mean rank of the nearest-native pose under the true score is below 70 percent of the larger of the two mean destroyed ranks. OK = passes, FAIL = does not. The perturbation pool fails three of twenty-five cells, all of them the shape channel at $K \leq 100$; the ZDOCK pool fails every cell up to K = 1000.

Table 5 Pass counts per candidate-set size on both pools.

K	perturbation pool (n=22)	ZDOCK pool (n=14)
20	4	5
50	6	7
100	13	4
200	22	4
400	22	6
1000	-	7
2000	-	10
3600	-	11

Each entry is the number of complexes on which the shape channel passes at that candidate-set size; the denominators are given in the column headers. A dash marks a candidate-set size the pool does not reach.

3.5 On real docking output no candidate-set size gives the potential an advantage

Table 6 is the counterpart of Table 2 on the ZDOCK pool, and the conclusion does not reproduce. No candidate-set size gives the potential a split below p = 0.05: the paired split is 7/6 at K = 20 (p = 1.0000), 10/4 at K = 50 (p = 0.1796), 9/5 at K = 100 (p = 0.4240), 10/4 at K = 200 (p = 0.1796) and 9/5 at K = 400 (p = 0.4240). Extended to the full pool, the potential wins on 6 of 14 complexes at K = 1000, 4 of 14 at K = 2000 and 3 of 14 at K = 3600, with the result that, at the largest candidate set, the only cell close to significance runs in the direction opposite to the claim (p = 0.0923).

The design cannot resolve an effect at this sample size, and the null is not treated as a demonstration. The smallest win fraction that a paired test on 14 complexes would take to a two-sided p below 0.05 is 0.86 (Table S11), and the observed fraction is 0.71 at its highest, at K = 50 and K = 200. What this section reports is therefore that no effect of the size seen on the perturbation pool is detectable at n = 14, not that the effect is absent.

Both measures are close to their chance values throughout. The pooled correlations of both channels lie within ±0.13 of zero at every K, against +0.882 for the potential on the perturbation pool at K = 20. The median rank the shape proxy assigns to the nearest-native pose is 11.5 at K = 20 and 869 at K = 3600, against 9.0 and 2059 for the potential; since a random ordering would give 10.5 and 1800, the shape proxy enriches the nearest-native pose by a factor of about two at the full pool but not at K = 20, while the potential enriches it at neither.

The result survives three further checks. Ranking the pool by ZDOCK's own score instead leaves the picture essentially unchanged — no K gives the potential an advantage — and at K = 3600 the test becomes significant in the opposite direction (12/2, p = 0.0129; Table S7). Under that ranking the candidate set is if anything emptier: its near-native share is 0.0% at every K up to 2000 and 0.1% at K = 3600 (Table S7), so the scarcity of near-native poses is a property of the search output and not of the proxy used to compress it. Subsampling the pool to 400 poses per complex, so that the equal-K comparison is between equally sized pools, leaves it unchanged across three seeds (Table S12). The twenty poses the shape proxy ranks first on the ZDOCK pool are not grossly interpenetrating: they carry a median of 13.2 clashes below 2.4 Å against 2.2 on the perturbation pool, but 728 contacts against 544 (Table S9). The proxy therefore selects over-packed rather than overlapping poses, and its failure to enrich near-native poses is not a steric artifact.

The two measures disagree at K = 400 only, and in the direction seen on the perturbation pool at $K \geq 200$: there the shape proxy has the higher pooled correlation while the potential has the better best-pose rank. Everywhere else on the ZDOCK pool neither measure separates the channels. A fourth check, using the rescoring algorithm distributed with the program itself on the same poses, is reported in Section 3.12; a fifth, on the same program's output at a second angular grid, is reported in Section 3.13.

Table 6 Candidate-set size on the ZDOCK pool (14 complexes, 3600 poses each).

K	poses	i-RMSD ≤ 2 Å	rho, shape	rho, potential	best rank, shape	best rank, potential	rank gain	paired W/L	p (two-sided)
20	280	0.4%	-0.015	+0.110	11.5	9.0	2.5	7/6	1.0000
50	700	0.3%	-0.002	+0.108	27.5	19.5	8.0	10/4	0.1796
100	1400	0.4%	+0.009	+0.035	68.5	37.0	31.5	9/5	0.4240
200	2800	0.2%	+0.027	+0.033	116.0	118.0	-2.0	10/4	0.1796
400	5600	0.2%	+0.043	+0.014	207.0	144.0	63.0	9/5	0.4240
1000	14000	0.2%	+0.026	+0.011	534.0	658.0	-124.0	6/8	0.7905
2000	28000	0.1%	+0.018	-0.012	635.5	1356.0	-720.5	4/10	0.1796
3600	50400	0.1%	-0.030	-0.046	869.0	2059.0	-1190.0	3/10	0.0923

Same protocol as Table 2, extended to the full range of K that a 3600-pose pool admits. The perturbation pool holds only 400 poses, so its table stops at K = 400; here K = 400 is the top ninth of the pool and K = 3600 is the whole of it. No K reaches significance, and at the largest K the only cell close to significance runs against the potential.

3.6 Only the pool built from the bound structures supports the effect

The three arms vary two factors, and they are arranged so that each adjacent pair holds one of them fixed: PB and PU share the pose generator, and PU and SU share the structures, so each comparison changes one thing and one thing only (Table 7). Against the shape proxy, the potential places the nearest-native pose higher on 14 of the 15 complexes of the bound arm, at both K = 20 and K = 50 ($p = 0.0010$ in both cases), and on 8 and 9 of the 15 when the same generator is applied to unbound structures ($p = 0.79$ and $p = 0.61$). The pooled correlation separates them further: the potential's mean correlation is +0.872 at K = 20 on the bound arm and -0.055 on the unbound arm, and the median rank it assigns to the nearest-native pose is 1.0 against 11.0.

Replacing the generator with the ZDOCK search while keeping the unbound structures changes much less than the first substitution does. At K = 20 the test is 7/6 on the search arm against 8/6 on the unbound perturbation arm, the mean pooled correlations are +0.110 against -0.055, and the median best-pose ranks are 9.0 against 11.0. Every difference between these two arms is smaller than the resolution of a fifteen-complex comparison, and both measures point the same way in both arms. Table 1 explains why the first substitution costs so much more than the second: PB and PU have almost the same composition, so what separates them is not how many near-native poses the candidate set holds but what those poses are like.

The gate gives the same ordering. Of its 25 cells, 23 pass on the bound arm and 3 on the unbound perturbation arm — counts for the fifteen-complex subsets of Fig. 4, and therefore not the same cells as the twenty-two-complex pool of Table 4 — and none pass at any K up to 1000 on the search arm. The progression runs in the same direction as the two measures and has the same shape: most of the drop comes from the structures, and the search accounts for the remainder (Table S14 gives the same comparison complex by complex at K = 50 and K = 400, and Table S15 gives the gate on all three arms).

The candidate-set size still moves the bound arm's verdict — from $p = 0.0010$ at K = 20 to $p = 0.30$ at K = 400. On the unbound arms it moves the verdict in the opposite sense. At the full pool the test therefore runs against the potential on both of them (2 of 15, $p = 0.0074$, on the unbound perturbation arm; 3 of 14, $p = 0.0923$, on the search arm). The disagreement between the two ranking-quality measures, by contrast, survives in all three arms: it appears at K = 200 and 400 on the bound arm, at K = 50 on the unbound perturbation arm and at K = 400 on the search arm. That part of the result therefore does not depend on the structures or on the search. It is the part of the study most likely to generalize, and it is stated as an observation rather than a demonstrated regularity: four events, each in the same direction, at candidate-set sizes that differ from arm to arm.

Table 7 The controlled experiment.

arm	coords	poses	K	near-native share	best rank, shape	best rank, potential	rho, shape	rho, potential	paired W/L	p (two-sided)
PB	bound	perturbation	20	100.0%	16.0	1.0	-0.285	+0.872	14/1	0.0010
PB	bound	perturbation	50	100.0%	30.0	5.0	-0.151	+0.861	14/1	0.0010
PB	bound	perturbation	100	99.4%	32.0	12.0	+0.226	+0.754	13/2	0.0074
PB	bound	perturbation	200	77.2%	32.0	15.0	+0.819	+0.457	11/4	0.1185
PB	bound	perturbation	400	45.4%	32.0	15.0	+0.672	+0.303	10/5	0.3018
PU	unbound	perturbation	20	91.9%	16.0	11.0	-0.163	-0.055	8/6	0.7905
PU	unbound	perturbation	50	85.6%	39.0	27.0	+0.007	-0.100	9/6	0.6072
PU	unbound	perturbation	100	77.7%	41.0	62.0	+0.269	-0.216	6/9	0.6072
PU	unbound	perturbation	200	59.9%	82.0	152.0	+0.487	-0.470	4/11	0.1185
PU	unbound	perturbation	400	46.3%	82.0	203.0	+0.374	-0.116	2/13	0.0074

arm	coords	poses	K	near-native share	best rank, shape	best rank, potential	rho, shape	rho, potential	paired W/L	p (two-sided)
SU	unbound	Fourier search	20	0.4%	11.5	9.0	-0.015	+0.110	7/6	1.0000
SU	unbound	Fourier search	50	0.3%	27.5	19.5	-0.002	+0.108	10/4	0.1796
SU	unbound	Fourier search	100	0.4%	68.5	37.0	+0.009	+0.035	9/5	0.4240
SU	unbound	Fourier search	200	0.2%	116.0	118.0	+0.027	+0.033	10/4	0.1796
SU	unbound	Fourier search	400	0.2%	207.0	144.0	+0.043	+0.014	9/5	0.4240

PB and PU use the same pose generator and differ only in the structures the ligand is moved through; PU and SU use the same structures and differ only in how the poses are produced. Per arm and candidate-set size: the near-native share of the candidate set, the median rank each of the two channels assigns to the nearest-native pose, the same two channels as mean pooled rank correlations, and the paired test of the best-pose rank over the complexes of that arm (two-sided exact binomial p value). PB is restricted to the fifteen complexes that the ZDOCK decoy set covers so that all three arms refer to the same systems.

Two things follow. Replacing the bound structures with unbound ones, keeping the pose generator fixed, removes most of the advantage the potential has at the small candidate-set sizes. The paired test goes from p = 0.0010 to p = 0.79 at K = 20 and from p = 0.0010 to p = 0.61 at K = 50, and the pooled correlation of the potential falls from +0.872 to -0.055 at K = 20. Replacing the pose generator with a Fourier search, keeping the unbound structures fixed, removes the rest. The effect therefore requires both conditions, and Tables 8 and 9 identify what in the unbound arms is doing the work.

3.7 The unbound arms lose the effect because their near-native poses are strained

The potential's score is very nearly a monotonic function of the clash count in every arm: the median rank correlation between the two is +0.881 on the bound arm, +0.976 on the unbound perturbation arm and +0.841 on the search arm (Tables 9 and S16). This attribution survives a control: partialling out the number of scored pairs leaves these correlations at +0.84, +0.79 and +0.77, while partialling out the clash count removes the positive correlation with the pair count (-0.38, +0.09 and 0.00). The score is a clash counter, not a contact counter. What differs between the arms is which poses carry the clashes.

On the bound arm the reference placement is the crystal complex and has no close contact at all, and the poses nearest to it have fewer clashes than the rest of the pool — a median of 0.00 against 3.00, rank correlation +0.056. The potential's short-range term has nothing to penalize, and its score tracks the number of favorable contacts, which is what makes it work as a ranking function.

On the unbound perturbation arm the reference placement is the two structures the search is given, and it carries a median of 17 close contacts. The poses nearest to it carry more, not fewer, than the rest of the pool — 19.00 against 13.00, rank correlation -0.230 — so the potential, whose score rises with the clash count, penalizes precisely the poses it is being asked to find. The search arm reproduces this from an independent data source. Its nearest-native poses carry a median of 23.5 close contacts, more than the rest of that pool, and the rank correlation between clash count and distance from the reference has the same sign as on the unbound perturbation arm (-0.114 against -0.230). This mechanism is therefore the same in both unbound arms, and it explains why the effect survives only on the bound one.

This also sets the scope of Section 3.9. The conclusion there — that a hard-core reference state does not fail by over-penalizing the minor steric conflicts that crystal structures carry — holds, and it holds because the bound pool supplies clash-free near-native poses. It is a statement about that pool and not about rescoring in general: on a pool built from unbound structures the clash term is exactly what breaks the ranking, and a potential with a stronger repulsive wall would do worse.

Table 8 Partitioning of the confounded factors.

quantity	K	PU	change vs PB (coordinate source only)	change vs SU (pose source only)
paired p (two-sided)	20	0.7905	+0.7896	-0.2095
paired p (two-sided)	50	0.6072	+0.6063	+0.4277
paired p (two-sided)	100	0.6072	+0.5999	+0.1833
paired p (two-sided)	200	0.1185	+0.0000	-0.0611
paired p (two-sided)	400	0.0074	-0.2944	-0.4166
best rank, potential	20	11.0	+10.0	+2.0
best rank, potential	50	27.0	+22.0	+7.5
best rank, potential	100	62.0	+50.0	+25.0

quantity	K	PU	change vs PB (coordinate source only)	change vs SU (pose source only)
best rank, potential	200	152.0	+137.0	+34.0
best rank, potential	400	203.0	+188.0	+59.0
rho, potential	20	-0.055	-0.927	-0.164
rho, potential	50	-0.100	-0.961	-0.208
rho, potential	100	-0.216	-0.970	-0.251
rho, potential	200	-0.470	-0.928	-0.503
rho, potential	400	-0.116	-0.420	-0.130
gate, cells passed of 25	-	3	-20	+3

The table isolates the coordinate source by comparing PB with PU, and the pose source by comparing PU with SU; PB is the fifteen-complex arm of Table 7. A change column gives the value in the other arm subtracted from the PU value, so a change near zero means that substitution left the quantity where it was.

Table 9 Strain of the reference placement in each arm.

arm	clashes at the reference pose	clashes at the nearest-native pose	clashes, median pose	rank corr, clash vs distance from reference	rank corr, potential score vs clash count
PB	0.00	0.00	3.00	+0.056	+0.881
PU	17.00	19.00	13.00	-0.230	+0.976
SU	n/a	23.50	18.00	-0.114	+0.841

For each arm the table gives the number of atom pairs closer than 2.4 Å at the reference pose, at the nearest-native pose and at the median pose of the pool, together with two rank correlations: clash count against distance from the reference, and potential score against clash count. In every row PB is the fifteen-complex arm of Table 7, so that all three arms refer to the same systems. A potential whose score rises with the clash count, in a pool whose near-native poses carry more clashes than the rest, is being asked to rank against itself.

3.8 The injection gate fails almost everywhere on real docking output

On the ZDOCK pool every one of the twenty-five cells fails at K ≤ 400, and every cell still fails at K = 1000; only the shape channel passes at K = 2000 and K = 3600, giving one pass in five (Table 4). Resolved per complex, the shape channel passes on 5, 7, 4, 4 and 6 of the 14 complexes at K = 20, 50, 100, 200 and 400, against the 14 of 14 expected if the true score were reliably better than a destroyed version of itself.

It is not that the potential loses on real docking output; it is that no channel, including the shape proxy, carries ranking information about the nearest-native pose that survives its own destruction. The gate is reporting, correctly, that the quantity being ranked — the single closest pose in a set that contains almost no near-native poses — is not recoverable from the scores at these candidate-set sizes.

The conclusion is robust to the gate's threshold at the low end and weaker at the high end. On the perturbation pool 22 of 25 cells pass at a 60 percent margin, 22 at 70 and 23 at 80. On the ZDOCK pool 1 of 40 cells passes at 60 percent, 2 at 70 and 9 at 80 (Table S10). The 70 percent margin used here is therefore a stated convention and not a derived constant, and an 80 percent margin would be too permissive to be relied on.

The reading is supported by the pool composition. The near-native share inside the candidate set on the ZDOCK pool is 0.4% at K = 20 and 0.2% at K = 400 (Table 1), so a compressed candidate set typically contains no near-native pose at all. When the object to be ranked is absent, the rank assigned to it is noise, and a gate that compares the true score against a shuffled version of itself returns the correct answer: nothing.

3.9 A potential can satisfy the shape criteria and still rank poorly

The four reference states of Table 10 and Fig. 5 were built from identical pair counts, so any difference between them is attributable to the reference state alone, and the ablation is recomputed here on both pools.

The hard-core reference state satisfies all three shape criteria: 45 of 45 type pairs are repulsive at 2.8 Å, 39 of 45 have a negative minimum in the contact range, and 39 of 45 rise again by 11 Å. It is nevertheless the worst of the four at ranking, with a pooled correlation of +0.053 at K = 400 and a best-pose rank of 131.0 on the perturbation pool. It behaves almost as a count of favorable contacts rather than as a discriminator: the rank correlation between its score and the number of scored atom pairs has a median of +0.795 across the twenty-two complexes (Table S6). The uniform-sphere reference state, the classical form, fails two of the three criteria badly — only 6 of 45 type pairs have a minimum in the contact range and only 2 of 45 rise again — yet it has the highest pooled correlation of the four (+0.687), and the shape proxy has the highest of any channel (+0.708).

The mechanism is not an over-penalization of small clashes. If the hard-core wall were punishing the nearest-native pose for the minor steric conflicts that crystal structures carry, that pose would have more clashes than the rest of the set; it

has fewer, a median of 0.00 against 3.00 below 2.4 Å (Table S6). The failure is that the potential has become a contact counter, and a contact counter rewards over-packed poses.

The same ordering holds on the ZDOCK pool, where the hard-core state is again the worst of the four, but every absolute value there lies within 0.13 of zero, so the ablation is interpretable only where the measures have range, which is the perturbation pool (Section 3.7 explains why the unbound pool behaves as it does).

Table 10 Reference-state ablation.

Reference state	Definition	A: 2.8 Å	B: 3.5-7 Å	C: 11 Å	rho, perturbation	rho, ZDOCK	best rank, perturbation
mean	mean field	27/45	39/45	39/45	+0.387	+0.014	16.0
mhc	mean field + hard core	45/45	39/45	39/45	+0.053	-0.029	131.0
pow25	power law, alpha = 2.5	43/45	28/45	22/45	+0.546	+0.076	62.5
sph	uniform sphere, r^2	44/45	6/45	2/45	+0.687	+0.127	48.5
shape proxy (reference)	-	-	-	-	+0.708	+0.043	45.5

The training data are identical across the four reference states, so a difference can only come from the reference state. Columns A, B and C count, out of the 45 atom-type-pair classes, the pairs that pass each of the three shape criteria: A repulsive at 2.8 Å, B a negative minimum between 3.5 and 7 Å, C rising again by 11 Å. Those three columns are properties of the potential and do not depend on the test set. The last two columns are the pooled rank correlation and the median best-pose rank at K = 400, recomputed on the 22-complex perturbation pool and on the 14-complex ZDOCK pool separately.

On the ZDOCK pool every value of the ZDOCK column of this table lies within 0.13 of zero, so the ablation is read on the perturbation pool. The scope of this section is set by Section 3.7: the conclusion that the hard-core state does not fail by over-penalizing small clashes holds for the bound pool, which supplies clash-free near-native poses, and not for a pool built from unbound structures.

3.10 Inside the K = 100 candidate set the binary success measures are saturated

Table 11 reports the full channel comparison inside the K = 100 candidate set of the twenty-two-complex perturbation pool, including three binary success measures, defined in the note. The random-ranking control already reaches 99.6% on acc@1 and 96.5% on high@1, so the three measures have no discriminating power at this candidate-set size and no conclusion is drawn from them. They are reported only to document the saturation, which is the reason the rest of this study uses ranks and rank correlations. This saturation follows directly from the pool composition of Section 3.1. When 99.6% of the candidate set at K = 100 is within 2 Å of native, a binary success measure cannot separate any two channels, and a benchmark that reports only such measures cannot see a difference that exists.

Table 11 Full channel comparison inside the K = 100 candidate set of the twenty-two-complex perturbation pool.

channel	best rank, median	best rank, mean	ρ(score, i-RMSD)	acc@1	acc@10	high@1
shape	45.5	42.8	+0.121	100.0%	100.0%	100.0%
sph	19.5	23.5	+0.630	100.0%	100.0%	100.0%
mean	11.5	14.2	+0.767	100.0%	100.0%	100.0%
pow25	22.5	25.0	+0.622	100.0%	100.0%	100.0%
mhc	28.5	26.0	+0.632	100.0%	100.0%	100.0%
random ranking (negative control)	—	—	—	99.6%	100.0%	96.5%
ranked by i-RMSD (oracle ceiling)	1.0	—	+1.000	100.0%	100.0%	100.0%

Best-pose rank (median and mean over the twenty-two complexes), the mean of the per-complex rank correlation between score and interface RMSD, and three binary measures. Of these, acc@1 is the fraction of complexes whose top-ranked pose is within 2 Å of native; acc@10 the fraction with at least one of the ten top-ranked poses within 2 Å; high@1 the fraction whose top-ranked pose is of high CAPRI accuracy. The random control is the analytical expectation under a random ranking, averaged over complexes. The binary measures are saturated in this candidate set, as the random control shows, and are reported only to document that saturation.

3.11 Leave-one-out and ground-truth checks support the result

Two supporting analyses are in the supplementary material. The K = 20 and K = 50 conclusions survive the removal of any single complex (Table S3); at K = 400, where the full set is not decisive, one complex can move p across 0.05, which the low power at that size already implies. Recomputing the best-pose rank against the CAPRI quality class instead of the

i-RMSD gives the same median rank and the same paired counts at every K (Table S5), so the choice of ground truth does not change the result; per-complex diagnostics are in Table S6.

3.12 The rescoring algorithm distributed with the program is also at chance on a compressed candidate set

ZDOCK decoy sets include a version of the ZDOCK 3.0.2 output re-ranked by IRAD, the scoring algorithm distributed with the program and the successor of ZRANK [6, 21]. The pool used here is the 6° grid of Section 3.13, with 54,000 predictions per complex. IRAD is a composite of shape, statistical and desolvation terms rather than a bare statistical potential, and it is a control that the three pools of Section 3.6 cannot provide. IRAD and the raw ZDOCK score rank the same 54,000 poses per complex, so a comparison between them isolates the rescoring step with no confounding from pool, pose set or sampling density, and it can be read without any of the machinery of Section 2.

On a compressed candidate set the rescoring step does not help. With the candidate set taken to be the top K of each channel's own score, the two are indistinguishable at K = 20. Both put the nearest-native pose at a median rank of 11.0 out of 20, against 10.5 for a random ordering, and a near-native pose is present in the set at all on 1 of 15 complexes for IRAD and 2 of 15 for the ZDOCK score. (One complex, BOYV, has no near-native pose anywhere in this pool — its closest prediction is 5.11 Å on the 6° grid, against 5.61 Å on the 15° grid of Section 2.1 — and is retained; its ranks are noise, identically for both channels.) The same holds for the candidate set a pipeline actually holds, the top K by the ZDOCK score. Rescoring that set with IRAD does not move the nearest-native pose significantly at $K \leq 200$ (Table S18).

Rescoring does help once the candidate set is large, and the transition occurs over the same range of K as the reversal on the perturbation pool (Section 3.2). On the set a pipeline holds, IRAD places the nearest-native pose higher than the ZDOCK score does on 12 of 15 complexes at K = 400 ($p = 0.035$) and on 13 of 15 at K = 3600 ($p = 0.0074$), while no cell at $K \leq 200$ reaches significance (Table S19 gives the two ranks complex by complex). A real docking pool is therefore not unrankable: a tuned rescoring function of the program's own family does recover ordering information from it, but only at candidate-set sizes at which the pose to be ranked is present often enough to be found.

The knowledge-based potential studied here is not failing because such a pool is uninformative. It is failing because at the candidate-set sizes a pipeline uses, the pose it is asked to find is usually not in the set, and no channel recovers it there — not the shape score, not our potential, and not IRAD.

3.13 The sampling density of the search does not change the pool

Section 3.5 rests on one published decoy set. The benchmark supplies the output of the same program at two angular grids, 15° (3,600 predictions per complex) and 6° (54,000), and both files carry the score of the program itself, so one channel can be compared on the same 14 complexes without rebuilding a single pose. BOYV is absent from both, as in Table 6. The grids agree closely on every quantity Section 3.5 uses: the median near-native share of a complex's own pool is 0.06% at 15° and 0.09% at 6° (median over complexes, not the pooled share of Table 1), the nearest-native pose sits at a median depth of 10.6% and 10.1% of the pool, and the whole-pool rank correlation of the program's own score against closeness to native is +0.176 and +0.175 (Tables S20 and S21). A near-native pose is present in the top 20 for 3 of 14 complexes at 15° and 2 of 14 at 6°, and in the top 400 for 6 of 14 at both. Densifying the search therefore adds poses without adding the ones that matter, and the conclusion of Section 3.5 does not depend on the angular grid it was measured on.

4 Discussion

4.1 What decides a rescoring comparison

Four protocol quantities decide the outcome of the comparison studied here — three of them properties of the pool and the fourth a property of the measure — and they are not independent. The candidate-set size decides whether the comparison measures the rejection of grossly incorrect poses or the ordering of nearly correct ones. The measure decides which of those two jobs is being reported, and at large K the two measures select different channels. The composition of the decoy pool decides whether the second job is available at all: on a perturbation pool the compressed candidate sets are near-native by construction, so the ordering of nearly correct poses is what is being measured and a knowledge-based potential wins; on real docking output the compressed candidate sets contain almost no near-native pose, so there is nothing to order and the gate returns nothing. And the structures the pool is built from decide whether the potential's own short-range term agrees with the truth. Bound coordinates supply near-native poses that are relaxed and therefore clash-free; unbound ones do not; and because the potential's score is nearly monotonic in the clash count, that difference alone decides the verdict (Section 3.7). The effect is a property of this family of atom-pair mean-force potentials, in which the short-range term dominates the score. It need not extend to machine-learning scoring functions, whose terms are learned rather than built from an explicit pair count.

The pool can outweigh the other three. A benchmark can report its K, its measure and its gate and still be uninterpretable, because the pools of Table 7 support opposite readings at the same K. The composition of the pool is therefore not a caveat on a result; it is a parameter of the experiment that has to be reported with it.

4.2 How to make a rescoring benchmark well posed

Six requirements follow, one for each defect that this study exhibits on a real pool. Each is stated as a check that can be run before a benchmark is reported, and Table 12 sets them out together with the evidence for each.

Table 12 A checklist for a well-posed rescoring benchmark.

Check	What this study shows	How to run it
Report the composition of the candidate set, not only its size	The bound pool is 100% near-native at K = 20; the ZDOCK pool is 0.4%.	Give the near-native share at the reported K, the rank of the closest pose and its clash count.
Gate the comparison against a destroyed score	Three of 25 cells fail on the bound arm, all in one channel; all 25 fail on the search output at K ≤ 1000.	Reverse the sign, and separately shuffle across poses; require the true score to beat 70% of both (200 shuffles).
Name the measure, and report both when they can disagree	They select different channels on the bound arm at K = 200 and 400, on the unbound perturbation arm at K = 50 and on the search arm at K = 400.	Report the pooled rank correlation and the best-pose rank together.
Choose K to match the pipeline being discussed	The bound arm's verdict reverses between K = 200 and K = 400; at n = 14 the detectable win fraction is 0.86.	Use the K a pipeline holds, and report the detectable fraction alongside the observed one.
Do not generalise from a perturbation pool	The same potential wins 21 of 22 complexes on the bound pool and 8 of 15 when the identical generator is applied to unbound structures.	Use a published decoy set where one exists, or state that the pool was built by perturbation.
Do not compare channels on a saturated binary measure	A random ordering already reaches 99.6% on acc@1 at K = 100.	Report ranks, not only binary success counts.

For a reader who runs docking rather than builds benchmarks, the six requirements have one practical consequence: state which pose the claim is about, how many poses were offered to the rescorer, and what those poses were like. The first is the rank of the nearest-native pose rather than a correlation over the whole set, the second is K, and the third is the near-native share together with the contact strain of the nearest-native pose. Each is a single number that a pipeline can log, and without them a rescoring comparison cannot be reproduced or interpreted, however carefully the scoring functions themselves are implemented.

4.3 Why the perturbation pool was a reasonable but insufficient model

The perturbation pool was built to isolate the ability of a scoring channel to order poses, holding the pose ensemble's statistical character fixed and varying only the protocol. It does that, and the perturbation is deliberately centered on the interface so that small perturbations stay at the interface and a quality gradient is formed. The consequence is that half of the perturbations are smaller than about 2 Å, and the smallest candidate sets are therefore near-native by construction — a property no Fourier search has.

Two intrinsic differences distinguish real docking outputs from perturbation-generated ensembles, and neither can be removed by design: the docked structures are unbound, so the best achievable i-RMSD is a fraction of an ångström and the near-native basin is shallower, and the pose ensemble is a rotational search rather than a perturbation about the native orientation. Both reduce the near-native share of the compressed candidate set; Section 3.6 partitions the loss and finds that the structures carry most of it. The observed difference — 100% against 0.4% at K = 20 — is far too large to be a second-order effect of either.

4.4 What the shape proxy omits

The shape proxy is −Nclose + 20·Nclash with a hard 5 Å contact cut and a linear clash penalty. A production fast-Fourier-transform shape score differs in three ways that matter for the comparison made here. It is evaluated on a grid and includes a pairwise statistical term, rather than being a bare atom-pair count [5], so it rewards a properly packed interface rather than a merely contact-rich one. It includes a soft core that penalizes overlap continuously rather than at a single distance threshold, and it is defined on surface atoms, so that buried and exposed atoms are not counted alike. All three are reasons why a real shape score can be expected to do better than the proxy on a real pool, and the sweep reported in Table S7

shows that ZDOCK's own score, which has all three properties, does not change the conclusion. The proxy is used throughout because it is the same object on both pools, which is what a controlled comparison requires.

4.5 What the failure of the gate does and does not mean

The gate's failure on the ZDOCK pool (Section 3.8) is a statement about the pair (score, candidate set) rather than about the potentials themselves: on these candidate sets the score does not rank the nearest-native pose better than a destroyed version of itself. A potential can be a good ranker of nearly correct poses — which Section 3.2 shows on the perturbation pool — and be indistinguishable from a shuffled score on a set that contains almost no nearly correct poses. Both statements are true of the same potential on the same complexes, and the second is the one that applies to a real pipeline that keeps the top 100 poses of a Fourier search. Throughout, the scoring function reorders a candidate set that a search has already produced; when the same function guides the search itself, it helps to choose the poses it is later asked to rank, and the argument above does not carry over unchanged. That null result is a statement about detectability at n = 14, not a demonstration of absence: the ZDOCK subset has limited power to detect an effect of the size seen on the perturbation pool.

4.6 Limitations

Six limits are recorded rather than argued away.

First, the ZDOCK pool covers 14 complexes, and the paired test at that size has little power: an effect of the size observed on the perturbation pool would not reach significance. The null result of Section 3.5 is therefore a statement that no effect of that size is detectable at n = 14, not a demonstration that the effect is absent.

Second, the ground truth on the ZDOCK pool is the interface Cα RMSD distributed with the decoy set rather than a recomputation from coordinates. This makes the replication independent of our implementation, but it also means the two pools are compared through slightly different definitions of the same quantity, and the CAPRI quality class is unavailable on the ZDOCK pool.

Third, the analysis is descriptive: the paired test treats the complexes as independent and no hypothesis was registered in advance, so the p values describe the split within the tested sets.

Fourth, both pools are rigid-body. The perturbation model does not simulate conformational change, and the complexes were restricted to those whose unbound and bound interfaces agree to within 1 Å. The position of that subset in the benchmark is given in Table S13: the studied complexes have a median unbound-to-bound interface RMSD of 0.68 Å, against 1.21 Å over all 230 triaged complexes, and the 142 complexes above 1 Å, 51 of them above 2.2 Å, are excluded. Conformational flexibility is ubiquitous in real interactions and would be expected to reduce the near-native share of a candidate set further, that is, to move a real pool even further from the perturbation pool. Our analysis is therefore restricted to rigid-body cases with small unbound-to-bound interface rearrangement; systems that undergo large conformational change on binding are outside its scope. The real search output examined here also comes from a single program, whose sampling is a rigid-body Fourier correlation. A search that relaxes its poses as it samples would be expected to deliver near-native poses with fewer close contacts, and the mechanism of Section 3.7 then predicts that the potential's advantage is restored rather than lost. That prediction is testable; testing it is left for future work.

Fifth, one input structure, 1NCA, carries a corrupted element column in the copy of the benchmark used here and required a repair (Table S2). The repair was verified to leave the existing per-complex values unchanged, and the gate that now detects such corruption is reported with its positive control.

Sixth, our control arm cannot fully separate two co-occurring properties of unbound starting structures: a small positional offset from the crystal complex (median 0.70 Å), and interface steric strain (median 17 close clashes). Both contribute to the observed outcome, yet the present setup cannot quantify their individual contributions. A follow-up experiment that relaxes the unbound interface before the perturbation would separate them, and it is not part of this study. What it could change is the attribution of the mechanism in Section 3.7, not the transferability result of Section 3.5, which rests on the difference between bound and unbound structures taken as a whole.

5 Conclusions

We measured, at fixed poses and over three decoy pools, how the size of the candidate set handed to a rescoring step changes the comparison between a knowledge-based potential and a shape-complementarity proxy, and under which conditions the comparison is well posed at all.

On a rigid-body perturbation pool covering all 22 untrained rigid complexes of Docking Benchmark 5, the potential places the nearest-native pose at a median rank of 1.0 at K = 20 and 16.0 at K = 400, against 15.5 and 45.5 for the shape proxy. The paired test moves from $p < 0.0001$ at K = 20 and K = 50 through $p = 0.0169$ at K = 200 to $p = 0.0525$ at K = 400, so the reversal lies between K = 200 and K = 400. The two ranking-quality measures select different channels at K ≥

200, and a formal injection gate fails in three of the twenty-five cells, all of them the shape channel at $K \leq 100$ — a pattern a broken test would not produce.

On the published ZDOCK 3.0.2 decoy set, a real Fourier search output covering 14 of those complexes with 3600 poses each, the same protocol gives a different answer, and the reason is visible before any statistic is computed. At $K = 20$ the perturbation candidate set is 100% near-native while the ZDOCK candidate set is 0.4%; the twenty poses the shape proxy ranks first are 0.37 Å from native on the first pool and 15.9 Å away on the second. Consequently no candidate-set size shows a significant advantage for the potential (all $p \geq 0.09$, and the only cell close to significance, at the full pool, runs the other way), and the injection gate fails every cell up to $K = 1000$.

Applying the identical perturbation protocol to the unbound structures that the search is given, on the same fifteen complexes, shows that the search is not what the effect depended on. The effect is gone for the most part as soon as the structures change: the paired test goes from $p = 0.0010$ to $p = 0.79$ at $K = 20$ and the potential's pooled correlation from +0.872 to -0.055, although the composition of the two pools is almost the same, 47.4% against 46.4% near-native. What differs is the strain of the near-native poses. The potential's score is nearly a monotonic function of the clash count, the bound pool's reference placement has no clash while the unbound one has a median of 17, and only in the first case does the potential's short-range term agree with the truth. The search arm shows the same signature on poses it produced itself.

The candidate-set composition, including whether the near-native poses it contains are structurally relaxed, is an integral part of the benchmark measurement. A claim that a rescoring function improves protein–protein docking is well posed only when the size of the candidate set, its near-native composition, the relaxation of those near-native poses, the measure used and the null that was beaten are all stated with it. A benchmark whose candidate sets are near-native and relaxed by construction measures the ordering of nearly correct poses. That is a legitimate question and it has a clear answer here, but it is not the question a docking pipeline asks, and its conclusions do not transfer.

References

1. Vakser IA (2014) Protein-protein docking: from interaction to interactome. Biophys J 107(8):1785–1793. https://doi.org/10.1016/j.bpj.2014.08.033
2. Vajda S, Hall DR, Kozakov D (2013) Sampling and scoring: a marriage made in heaven. Proteins 81(11):1874–1884. https://doi.org/10.1002/prot.24343
3. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 89(6):2195–2199. https://doi.org/10.1073/pnas.89.6.2195
4. Lawrence MC, Colman PM (1993) Shape complementarity at protein/protein interfaces. J Mol Biol 234(4):946–950. https://doi.org/10.1006/jmbi.1993.1648
5. Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z (2007) Integrating statistical pair potentials into protein complex prediction. Proteins 69(3):511–520. https://doi.org/10.1002/prot.21502
6. Pierce B, Weng Z (2007) ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 67(4):1078–1086. https://doi.org/10.1002/prot.21373
7. Huang SY, Zou X (2008) An iterative knowledge-based scoring function for protein-protein recognition. Proteins 72(2):557–579. https://doi.org/10.1002/prot.21949
8. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213(4):859–883. https://doi.org/10.1016/S0022-2836(05)80269-4
9. Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jiménez-García B, Bates PA, Fernandez-Recio J, Bonvin AMJJ, Weng Z (2015) Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. J Mol Biol 427(19):3031–3041. https://doi.org/10.1016/j.jmb.2015.07.016
10. Oliva R, Vangone A, Cavallo L (2013) Ranking multiple docking solutions based on the conservation of inter-residue contacts. Proteins 81(9):1571–1584. https://doi.org/10.1002/prot.24314
11. Launay G, Ohue M, Prieto Santero J, Matsuzaki Y, Hilpert C, Uchikoga N, Hayashi T, Martin J (2020) Evaluation of CONSRANK-like scoring functions for rescoring ensembles of protein-protein docking poses. Front Mol Biosci 7:559005. https://doi.org/10.3389/fmolb.2020.559005
12. Ricciardelli T, Barradas-Bautista D, Cao Z, Chawla M, Cavallo L, Oliva R (2025) Increasing the fraction of correct solutions in ensembles of protein-protein docking models by an iterative consensus algorithm. Protein Sci 34(11):e70314. https://doi.org/10.1002/pro.70314
13. Kong R, Wang F, Zhang J, Wang F, Chang S (2019) CoDockPP: a multistage approach for global and site-specific protein-protein docking. J Chem Inf Model 59(8):3556–3564. https://doi.org/10.1021/acs.jcim.9b00445

14. Pang M, He W, Lu X, She Y, Xie L, Kong R, Chang S (2023) CoDock-Ligand: combined template-based docking and CNN-based scoring in ligand binding prediction. BMC Bioinformatics 24(1):444. https://doi.org/10.1186/s12859-023-05571-y
15. Shirali A, Stebliankin V, Karki U, Shi J, Chapagain P, Narasimhan G (2025) A comprehensive survey of scoring functions for protein docking models. BMC Bioinformatics 26(1):25. https://doi.org/10.1186/s12859-024-05991-4
16. Huang SY (2014) Search strategies and evaluation in protein-protein docking: principles, advances and challenges. Drug Discov Today 19(8):1081–1096. https://doi.org/10.1016/j.drudis.2014.02.005
17. Janin J, Henrick K, Moult J, Ten Eyck L, Sternberg MJE, Vajda S, Vakser I, Wodak SJ (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52(1):2–9. https://doi.org/10.1002/prot.10381
18. Lensink MF, Méndez R, Wodak SJ (2007) Docking and scoring protein complexes: CAPRI 3rd edition. Proteins 69(4):704–718. https://doi.org/10.1002/prot.21804
19. Truchon JF, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model 47(2):488–508. https://doi.org/10.1021/ci600426e
20. Zhao W, Hevener KE, White SW, Lee RE, Boyett JM (2009) A statistical framework to evaluate virtual screening. BMC Bioinformatics 10:225. https://doi.org/10.1186/1471-2105-10-225
21. Pierce BG, Hourai Y, Weng Z (2011) Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS One 6(9):e24657. https://doi.org/10.1371/journal.pone.0024657
22. Hwang H, Vreven T, Janin J, Weng Z (2010) Protein-protein docking benchmark version 4.0. Proteins 78(15):3111–3114. https://doi.org/10.1002/prot.22830
23. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11(11):2714–2726. https://doi.org/10.1110/ps.0217002

Statements and Declarations

Funding

This work was supported by the National Natural Science Foundation of China (grant no. 32360634), the Guangxi Natural Science Foundation Project (grant no. 2025GXNSFAA069786), and the Doctoral Start-up Fund, Guangxi University of Science and Technology (grant no. 17z09).

Competing interests

The authors declare no competing interests.

Ethics approval

Not applicable. This study is a computational analysis of publicly available protein structures and published decoy sets. It involved no human participants, human material, or animals, so no institutional review board approval was required.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The Docking Benchmark 5 structures are available from the docking benchmark [9], and the ZDOCK 3.0.2 decoy sets for Docking Benchmark 4, both the raw output and the version re-ranked by IRAD, are available from the Weng laboratory (https://zlab.wenglab.org/zdock/decoys.shtml). The per-pose score tables for both pools, the candidate sweep results and the injection-gate results are provided as supplementary material.

Code availability

The analysis scripts that generate every number and figure in this article are provided as supplementary material. Their entry points are four scripts: one generates the rigid-body decoys, one reconstructs the ZDOCK poses and scores them, one builds the candidate sets, and one runs the paired tests and the injection gate.

Authors' contributions

Tingguang Sun Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft.

Qianwei Cheng Conceptualization, Methodology, Validation, Supervision, Project administration, Funding acquisition, Writing – review and editing.

Lan Liao Conceptualization, Supervision, Funding acquisition, Writing – review and editing.

Figures

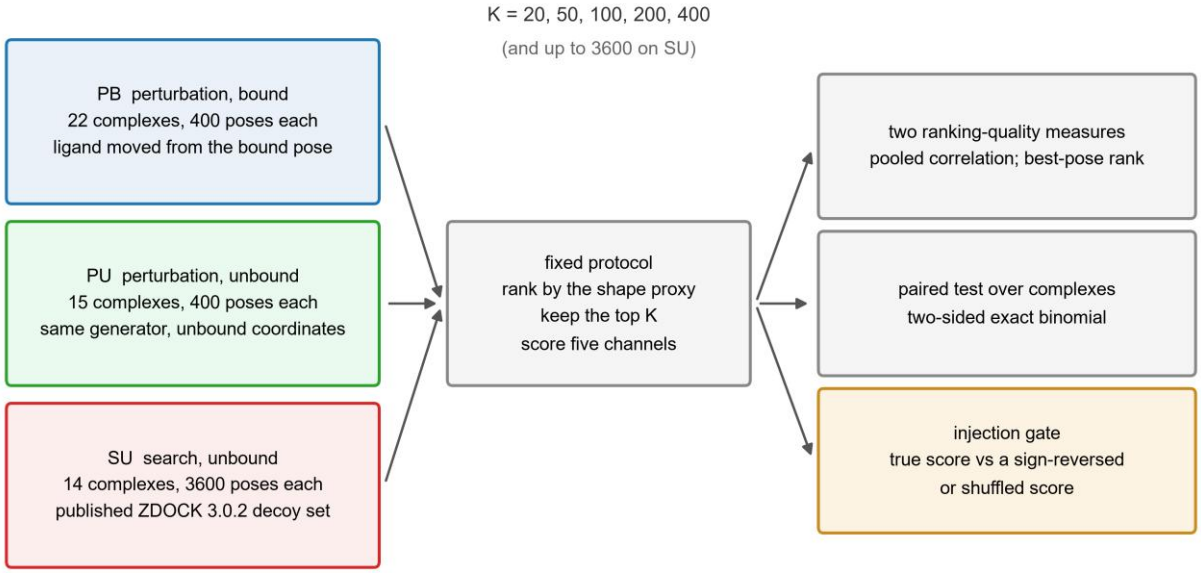


Fig. 1 Design of the study. Three decoy pools — PB, a perturbation of the bound structures; PU, the same perturbation applied to the unbound structures; SU, the published ZDOCK search output — are passed through one fixed protocol, which ranks each pool by the shape proxy, keeps the top K poses and scores five channels on them. The injection gate of Section 2.5 is imposed on every channel-by-K cell, and two ranking-quality measures are reported for each cell.

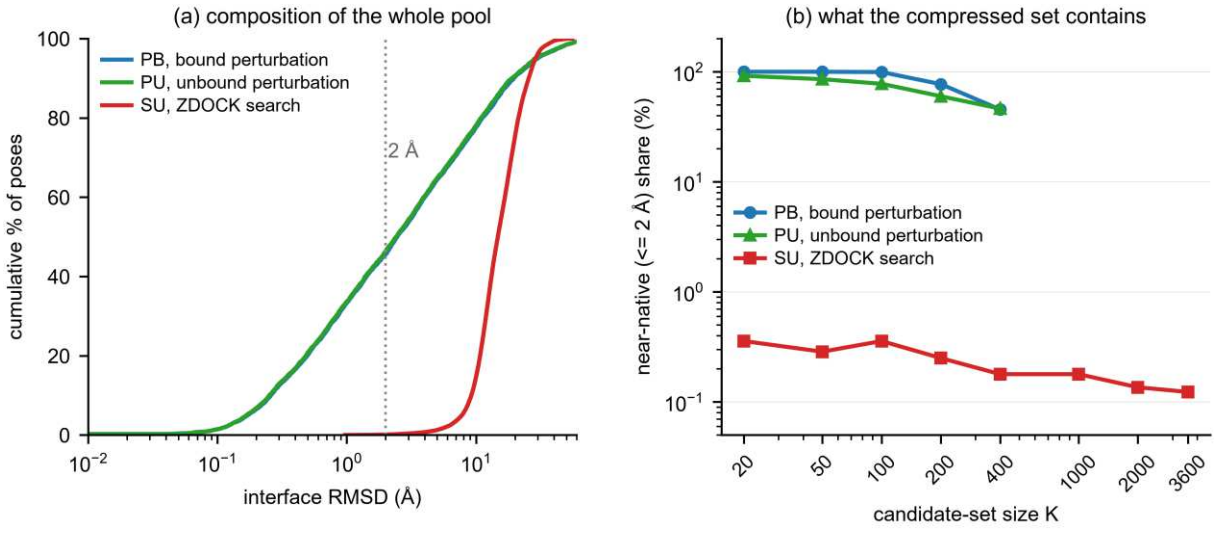


Fig. 2 Composition of the pools. (a) Distribution of the pose interface root-mean-square deviation (i-RMSD) over all poses of each pool; the x-axis is logarithmic and the dotted line marks the 2 Å near-native threshold. (b) The near-native share (i-RMSD ≤ 2 Å) inside the candidate set as a function of the candidate-set size K; both axes of panel (b) are logarithmic. PB is the perturbation pool built from the bound structures, PU the identical perturbation applied to the unbound structures, and SU the published ZDOCK search output. The bound arm is restricted to the fifteen complexes the ZDOCK decoy set covers, so that the three arms refer to the same systems; panel (b) reproduces the corresponding column of Table 7.

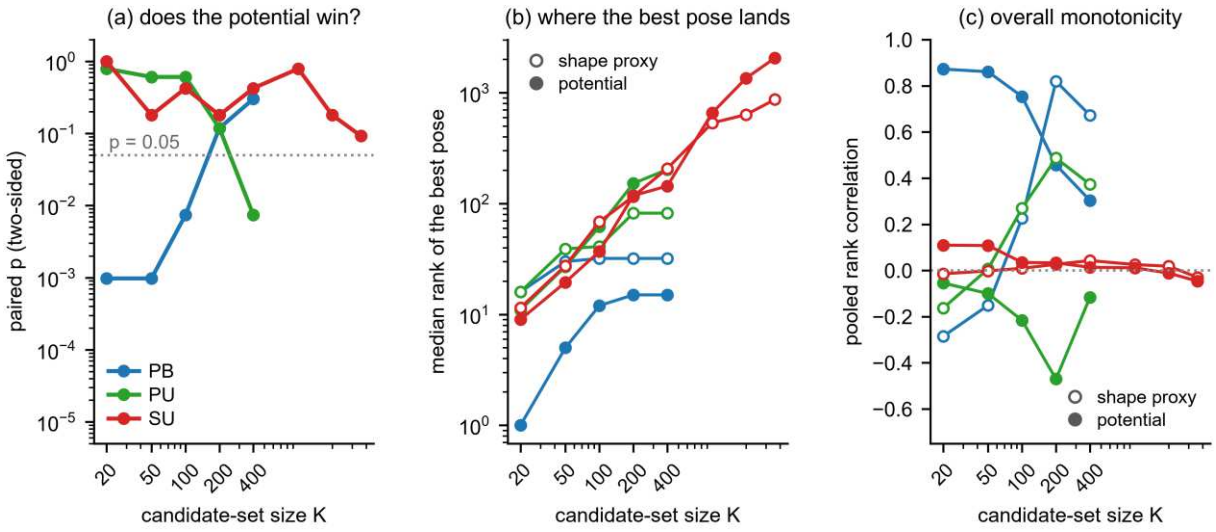


Fig. 3 The K response of the pools, against the candidate-set size K. (a) The paired-test p value (two-sided exact binomial), on logarithmic axes, with the p = 0.05 level dotted. (b) The median rank the channel assigns to the nearest-native pose. (c) The pooled rank correlation. Colours denote the arm (PB, PU, SU); filled markers are the knowledge-based potential and open markers the shape proxy.

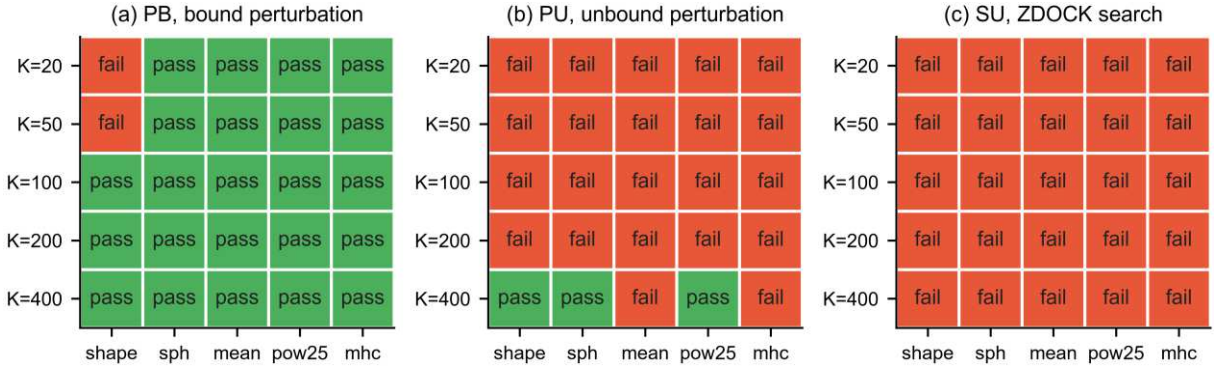


Fig. 4 The injection gate, applied to each of the three arms (panels a, b and c: PB, PU and SU). Rows are the candidate-set size K and columns the five scoring channels. A cell is marked pass when the true score places the nearest-native pose below 70% of the larger of the mean ranks obtained from a sign-reversed and from a shuffled score, and fail otherwise; the gate is defined in Section 2.5.

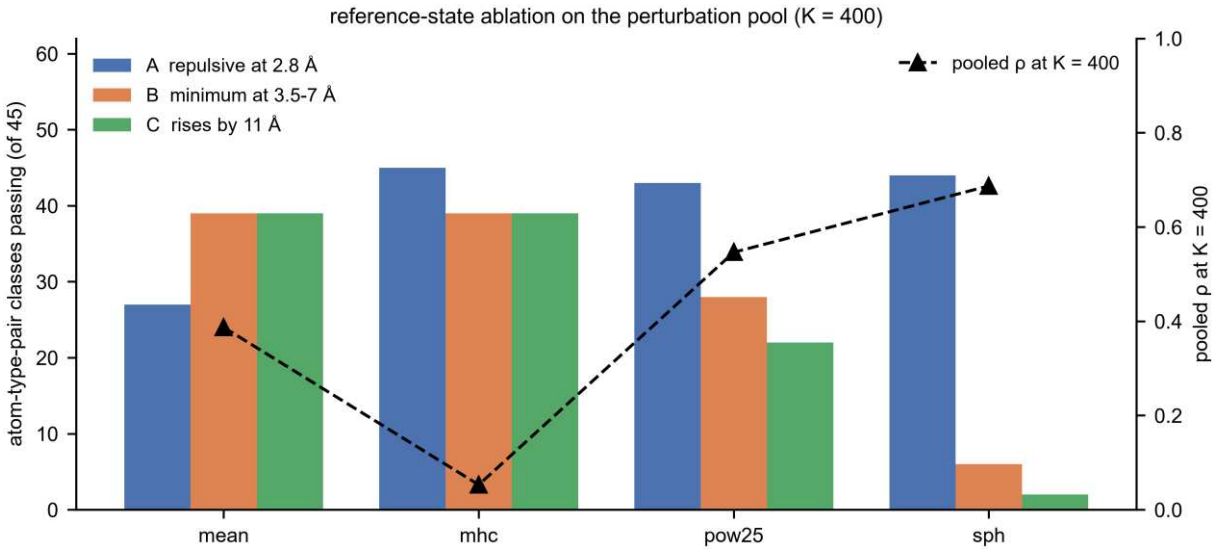


Fig. 5 Reference-state ablation on the perturbation pool at K = 400. Bars give, for each reference state (mean, mean field; mhc, mean field plus hard core; pow25, power law with α = 2.5; sph, uniform sphere), the number of the 45 atom-type-pair classes that satisfy each of the three shape criteria — A, repulsive at 2.8 Å; B, a negative minimum between 3.5 and 7 Å; C, rising again by 11 Å. The dashed line with triangles, read on the right-hand axis, is the pooled rank correlation ρ of the same four reference states at K = 400.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx