Fold

Improving antibody–antigen structure prediction through large-scale distillation of sequence pairs

Mingchen Chen, **Yunlong Cao**, **TorchFold Team**

Changping Laboratory, Beijing, China

Abstract

Antibody–antigen complex prediction remains limited by errors in epitope localization and binding orientation, while experimentally resolved complexes provide sparse training coverage of molecular recognition. Antibody discovery collections contain many additional sequence pairs whose complex structures are unknown. Here we introduce TorchFold, which uses two rounds of confidence-filtered structural distillation to convert these pairs into training examples for antibody–antigen co-folding. The predicted complexes extend epitope coverage and interface contact patterns beyond those represented in SAbDab. Retaining the AlphaFold 3 architecture, TorchFold combines the distilled structures with interface-weighted coordinate supervision and a revised diffusion noise schedule. The complete training programme increases ranked FoldBench interface success from 34.0% to 70.1% at DockQ > 0.23 and achieves competitive performance across five benchmark collections. Among 297 pooled benchmark predictions with ipTM ≥ 0.8, 292 have successful interfaces (98.3%), corresponding to acceptance of 31.1% of the 956 scored interfaces. These results support structural distillation of antibody–antigen sequence pairs as a practical approach to expanding supervision for complex prediction. TorchFold provides training and inference code and an antibody–antigen checkpoint under the MIT license.

Correspondence: Mingchen Chen (mingchenchen@cpl.ac.cn); Yunlong Cao (ylcao@cpl.ac.cn)
Project Page: https://torchx-cpl.github.io
Code: https://github.com/Mingchenchen/TorchX
Web Server: https://torchfold.openi.org.cn/

1 Introduction

AlphaFold 3 (AF3) extended structure prediction to all-atom complexes containing proteins, nucleic acids, small molecules, ions and modified residues [1]. Open co-folding models, including Chai-1, Boltz-1, Protenix, HelixFold3 and OpenFold3, have broadened access to these capabilities [2–6]. Nevertheless, antibody–antigen (Ab–Ag) complexes remain a challenging class [7]. Even when the component structures are well modeled, the antibody may bind the wrong antigen surface or adopt an incorrect orientation. Improving prediction therefore requires learning the geometry of molecular recognition in addition to individual folds.

Experimental complex structures in SAbDab provide direct supervision for this task, but cover only a small fraction of antibody–antigen sequence and interface space [8]. In contrast, antibody discovery campaigns and antigen-specific databases contain many sequence pairs without resolved complex structures. For experimentally supported pairs, the association between an antibody and its target supplies biological information even when the epitope and binding orientation are unknown. Converting these associations into useful structural supervision could expand the examples available for training a complex predictor.

Structural self-distillation offers a route to expanding training supervision beyond experimentally resolved structures. AlphaFold2 demonstrated that confidence-filtered predictions from unlabelled protein sequences can improve structure prediction through further training [9]. Boltz-2 uses structural distillation across several molecular interaction classes, including TCR–pMHC complexes [10]. Antibody-specific adaptation has also improved complex prediction: HelixFold-Multimer combines general structural training with fine-tuning on experimentally resolved antibody–antigen complexes [11]. These advances motivate the systematic use of antibody–antigen sequence collections as an additional source of structural supervision. The central question is whether confidence-filtered complex predictions for pairs lacking resolved structures can expand training coverage of molecular recognition and improve antibody–antigen co-folding on unseen complexes.

Here we introduce TorchFold, which implements this strategy through two rounds of structural distillation of antibody–antigen sequence pairs. To our knowledge, this is the first reported use of large-scale distillation of antibody–antigen pairs without resolved complex structures to expand training supervision and improve Ab–Ag co-folding. The pairs are drawn from antigen-specific databases, patent and private discovery collections, and computational designs. TorchFold retains the AF3 architecture and combines the distilled examples with interface-weighted coordinate supervision and a revised diffusion noise schedule. The complete training programme increases ranked FoldBench success from 34.0% to 70.1%. We evaluate the approach across five benchmark collections to characterize transfer across target sets, the added interface coverage, confidence behavior, inference-time sampling and reduced accuracy on large antigens. Training and inference code and the antibody–antigen checkpoint are provided to support use and further development of the approach.

2 Sequence-pair distillation for antibody–antigen prediction

TorchFold uses an AF3-compatible PyTorch model to generate structural labels for antibody–antigen sequence pairs and then trains on these labels together with experimental complexes (Figure 1a,d). Confidence filtering selects the first round of labels. The resulting checkpoint is used to revisit initially lower-confidence candidates in a second round. An interface-specific coordinate loss and a revised diffusion noise schedule provide additional supervision during fine-tuning (Figure 1b,c). The network architecture is unchanged. Loss definitions and the stage-specific training settings are given in Appendices A.2 and A.5.

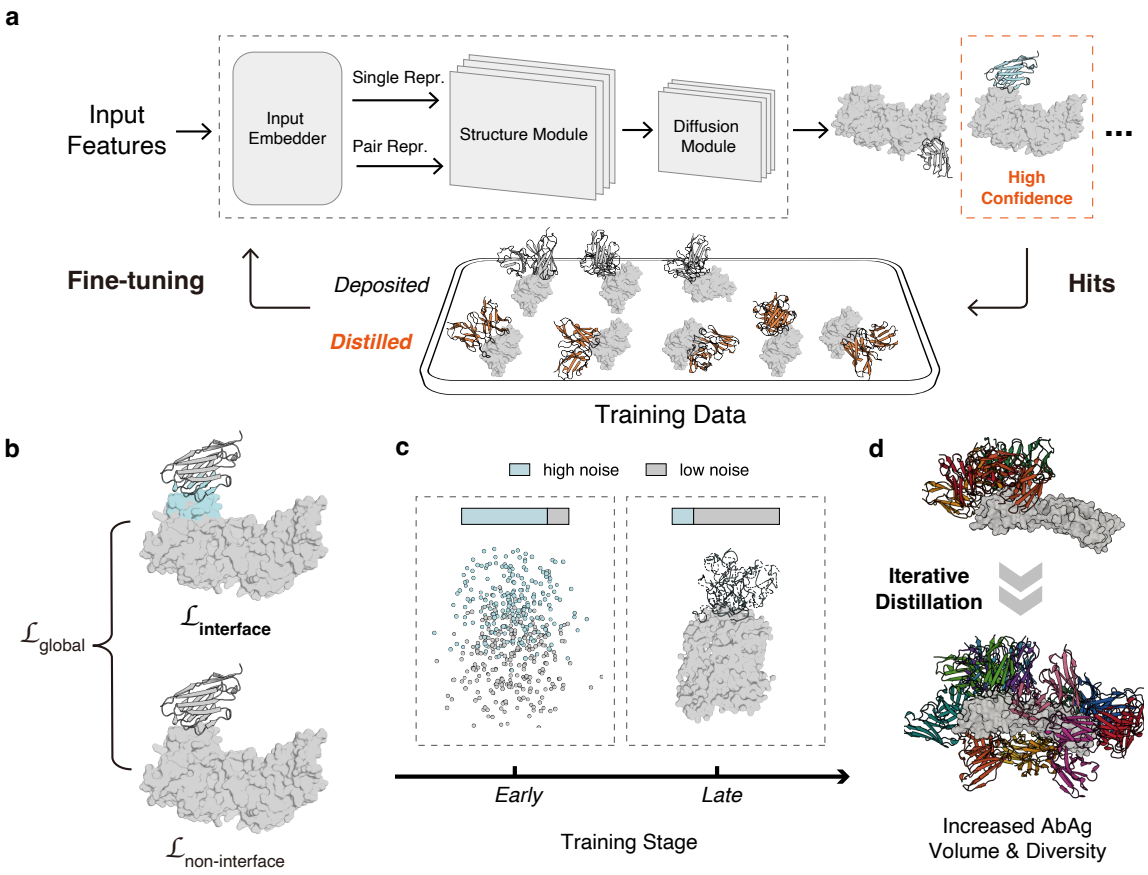

Figure 1 Overview of TorchFold. (a) Model architecture and expansion of training data. (b) Interface-specific loss added to the global coordinate objective. (c) High-noise diffusion schedule from early to late training. (d) Two-round structural distillation of antibody–antigen sequence pairs. Accepted predicted structures provide additional training labels.

2.1 Two-round structural distillation of sequence pairs

TorchFold was trained from scratch on PDB structures with a 30 September 2021 cutoff and protein monomer distillation data from MGnify and Uniclust30. The reported baseline achieves 34.0% ranked FoldBench Ab–Ag success. Subsequent Ab–Ag fine-tuning uses SAbDab with a 1 January 2025 cutoff, after removal of identical sequences, together with predicted complex structures. Experimental training complexes similar to FoldBench-AB or PXMeter-AB targets were removed by structure-based redundancy filtering at the antibody, antigen and complex/interface levels, including CDRs; the remainder were clustered into non-redundant training and validation sets (Appendix A.3). AbAg2526, 2026ARK-AB and CPL-28 have no release overlap with this experimental fine-tuning set.

The from-scratch checkpoint was used to generate complexes for antibody–antigen sequence pairs from PPIFlow designs [12], patent and private discovery collections, the Antigen-Specific Antibody Database (ASD) [13] and CoV-AbDab [14]. ASD curation is described in Appendix A.4. These sources contain different levels of interaction evidence: experimentally supported sequence pairs and computationally proposed pairs should not be interpreted as equivalent biological observations. Complexes with ipTM ≥ 0.8 were retained as first-round structural labels. The acceptance threshold was selected using the observed association between ipTM and DockQ on FoldBench during development.

Fine-tuning on SAbDab and the first-round labels with the interface loss reached 65.9% ranked success (stage 1). The stage 1 checkpoint then re-inferred candidates whose first-round scores lay in $0.5 \leq \text{ipTM} < 0.8$. Predictions passing the acceptance threshold were added for stage 2, which reached 70.1% with the revised

training settings in Table 2. Figure 2c reports approximately 24,500 examples in the final combined Ab–Ag training collection, including experimental examples and distilled labels. Acceptance rates vary across sources (Figure 2b).

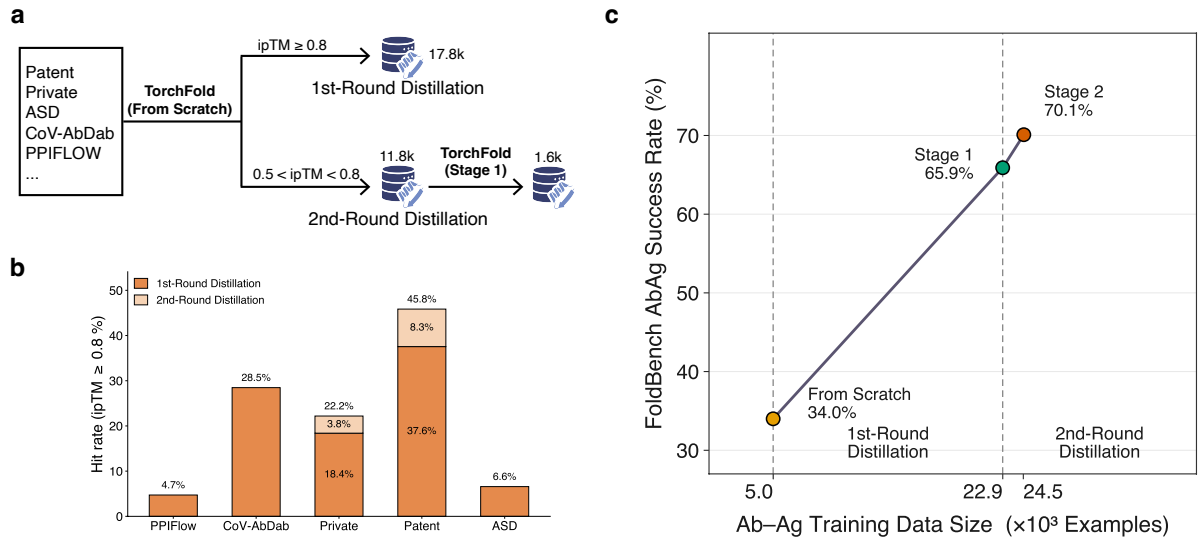

Figure 2 Two-round antibody–antigen distillation. (a) Distillation workflow. (b) Prediction acceptance rate by source. Medium orange, round 1; light orange, round-2 recovery. (c) FoldBench Ab–Ag success versus total Ab–Ag training examples ($\times 10^3$), including experimental structures and predicted labels. Points represent successive training stages with the settings in Table 2.

2.2 Interface-weighted coordinate supervision

The binding surface occupies a small fraction of the atoms in a complex. We add an interface-specific mean-squared error (MSE) to the global coordinate loss to increase its contribution during training. Interface residues or atoms are identified from the training reference structure using a 10 Å interchain distance threshold. For experimental examples the reference is the resolved structure; for distillates it is the accepted model prediction. The reference complex is aligned to the prediction using all valid atoms before the interface term is evaluated. The loss therefore penalizes interface displacement in the global frame, including errors in relative chain orientation, while retaining whole-complex supervision. Stage 1 uses an interface-loss weight of 2.0 and stage 2 a weight of 10.0. Formal definitions are given in Appendix A.2.

2.3 Noise scheduling to reduce residual structural cues

We hypothesized that increasing structural corruption during training would reduce reliance on residual interface coordinates and encourage the model to use sequence and molecular context. The revised schedule, introduced in stage 2, initially emphasizes higher noise and then shifts toward lower noise for structural refinement. The schedule parameter σ_{mean} decreases stepwise from 10 to 2 (Appendix A.2).

3 Distilled complexes expand predicted interface coverage

We compared the epitope coverage and interface contacts of accepted distillates with those of SAbDab complexes. For antigens well represented in the distilled collection, we mapped contacted antigen residues onto the corresponding surfaces (Figure 3a). The distillates include contacted regions absent from the compared SAbDab structures on PD-L1, PD-1, CTLA-4 and CD47. The distillation-only fraction is about three-quarters of the contacted surface on PD-L1. RBD and HEL, which have more extensive experimental coverage, show greater overlap. Structural overlays similarly show predicted binding sites on additional faces of CD47,

CTLA-4, PD-1, PD-L1 and RBD (Figure 3b). These observations describe expanded predicted interface coverage.

We represented interfaces as normalized 20×20 matrices of paratope–epitope amino-acid contacts. Distilled antibodies are enriched in the two-to-three hotspot-count bins and nanobodies in the three-to-four bins (Figure 3c). Size-matched t-SNE visualizations show partial overlap between experimental and distilled contact patterns. Median pairwise cosine distances within the collections range from 0.84 to 0.90 (Figure 3d). Median nearest-neighbour cosine distances within SAbDab are 0.026 for antibodies and 0.027 for nanobodies, whereas distances from distillates to SAbDab have medians of 0.35 and 0.55, respectively. In total, 97% of distilled antibodies and 99.8% of distilled nanobodies exceed the corresponding within-SAbDab median (Figure 3e).

Median interface areas are larger among distillates: 899 versus 807 Å^2 for antibodies and 877 versus 808 Å^2 for nanobodies (Figure 3f). The fraction of CDR residues contacting the antigen is 0.55 versus 0.50 for antibodies, with a smaller shift for nanobodies (Figure 3g). In the matched RBD example, the SAbDab Fab has an interface area of 1236 Å^2 and CDR occupancy of 0.42, compared with 1999 Å^2 and 0.67 for the distilled binder (Figure 3h). The distilled collection thus adds predicted surface coverage and alters the distribution of interface contacts.

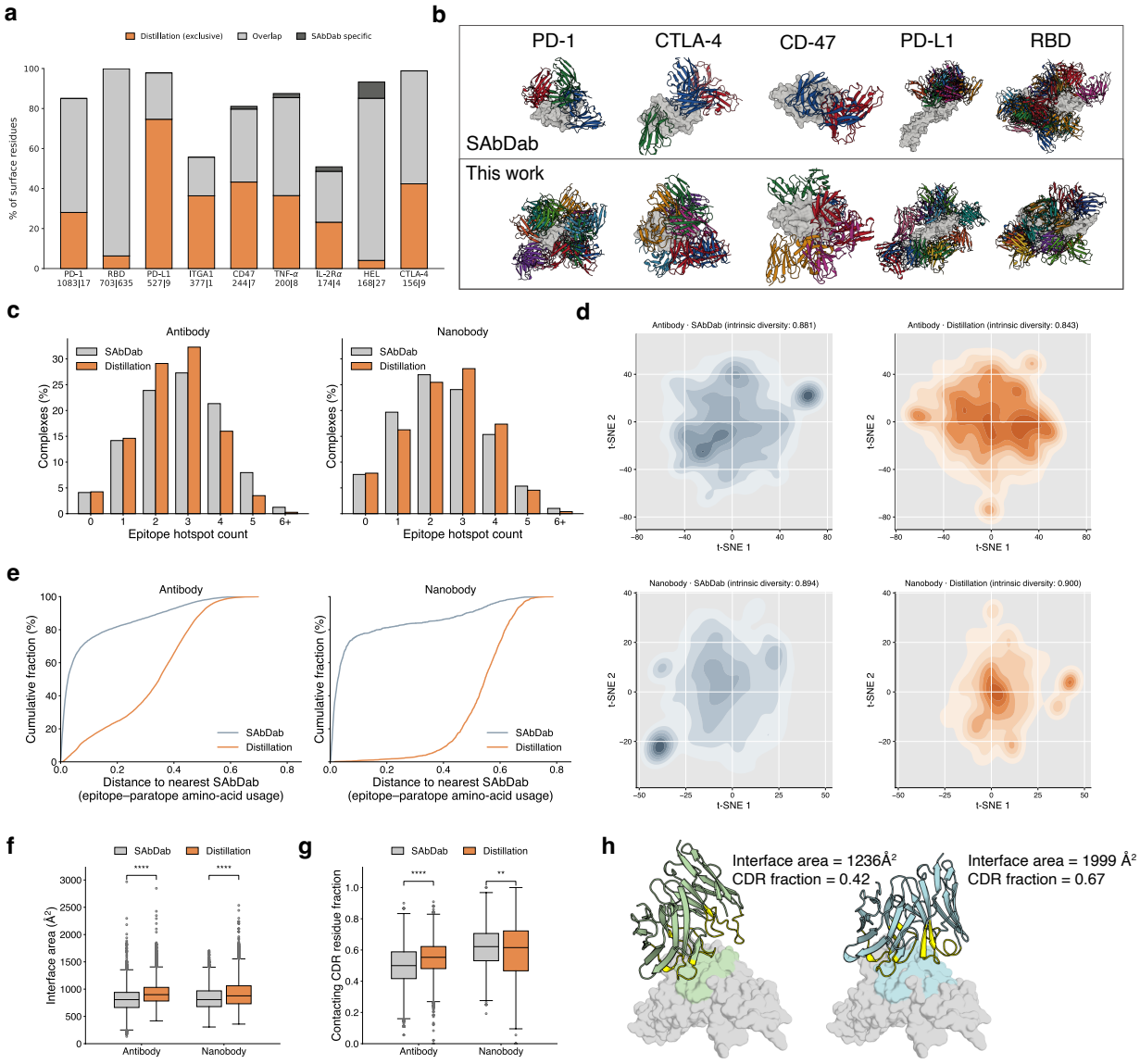

Figure 3 Epitope coverage and interface contacts, distillation versus SAbDab. (a) Antigen-surface residues contacted only in accepted predictions, in both collections, or only in SAbDab. (b) Epitope-cluster overlays on CD47, CTLA-4, PD-1, PD-L1 and RBD (top, SAbDab; bottom, distillation). (c) Epitope hotspot-count distributions. (d) t-SNE of normalized 20×20 paratope–epitope amino-acid contact matrices. (e) Distance to the nearest SAbDab complex (SAbDab: leave-one-out). (f) Interface area, $(\mathrm{SASA_{Ab}} + \mathrm{SASA_{Ag}} - \mathrm{SASA_{complex}})/2$. (g) Fraction of CDR residues contacting the antigen (10 Å Cα). (h) Matched RBD epitope: SAbDab (left) versus distillation (right). Gray surface, antigen; yellow, CDRs. Gray, SAbDab; orange, distillation. Asterisks, Bonferroni-adjusted Mann–Whitney tests.

4 Antibody–antigen structure prediction

4.1 Performance across five benchmark collections

We compare TorchFold with AF3 [1], Protenix-v2 [4], OpenDDE [15] and AF3-TF on FoldBench-AB [7], 2026ARK-AB [15, 16], AbAg2526 (Appendix D), CPL-28 and PXMeter-AB [16] (Figure 4). AF3-TF denotes checkpoint transfer from TorchFold into the AF3 implementation; technical details are provided in Appendix A.1. The reported unit is a scored chain interface, rather than necessarily a unique antibody–antigen

complex. An interface is successful at DockQ > 0.23, with acceptable ($0.23 <$ DockQ ≤ 0.49), medium ($0.49 <$ DockQ ≤ 0.8) and high (DockQ > 0.8) quality. Ranked success uses the model's top selection; oracle success uses the best DockQ within the sampled ensemble and is an upper bound for selection at that sampling budget.

MSAs for every method were generated with the official AF3 search pipeline [1] (Jackhmmer). OpenDDE's published evaluations instead use LocalColabFold (MMseqs2) against a different sequence database. Differences between our OpenDDE numbers and those reported by OpenDDE [15] may therefore reflect differences in search method and sequence database.

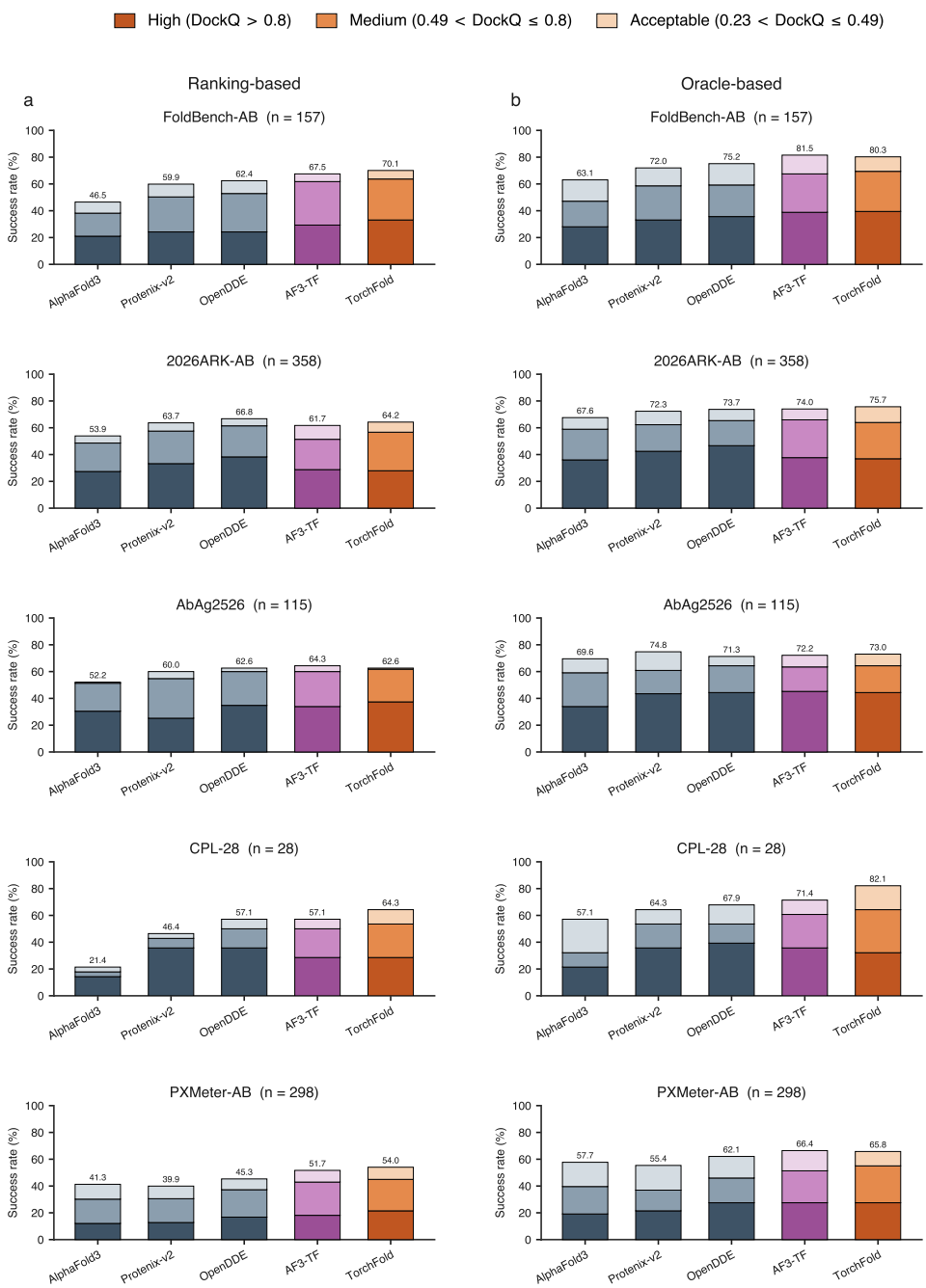

Figure 4 Antibody–antigen DockQ success. (a) Ranking-based and (b) oracle-based success on FoldBench-AB, 2026ARK-AB, AbAg2526, CPL-28 and PXMeter-AB. n is the number of scored interfaces. Stacked segments: high (DockQ > 0.8), medium ($0.49 <$ DockQ ≤ 0.8), acceptable ($0.23 <$ DockQ ≤ 0.49).

FoldBench-AB is the antibody–antigen split of FoldBench, a community all-atom benchmark spanning nine complex classes [7]. TorchFold recovers 70.1% of interfaces by rank, against 62.4% OpenDDE, 59.9% Protenix-v2 and 46.5% AF3. The difference was concentrated in high-quality predictions (DockQ > 0.8: 33.1% versus 24.2% OpenDDE and Protenix-v2, 21.0% AF3).

CPL-28 comprises 12 structures and 28 scored Fab–antigen chain interfaces, all against SARS-CoV-2 Spike RBD. The closely related antibodies make this collection useful for probing epitope discrimination on a familiar

antigen, but its single-antigen composition does not test generalization across antigen families. Ranked success is 64.3% versus 57.1% OpenDDE and 21.4% AF3. Oracle on this panel reaches 82.1%, so many of the misses are ranking errors rather than missing modes.

Whether the FoldBench gain extends to more recent complexes is tested on AbAg2526 and 2026ARK-AB. AbAg2526 is 71 SAbDab complexes released from 1 January 2025 to 31 August 2026 (Appendix D, 115 scored interfaces). TorchFold and OpenDDE both reach 62.6% ranked success. TorchFold places a larger fraction of those successes in the high-quality bin (37.4% versus 34.8%). 2026ARK-AB follows the PXMeter antibody–antigen protocol on newly deposited structures [15, 16]. Here OpenDDE leads rank (66.8% versus 64.2%), and the lead is concentrated at DockQ > 0.8. TorchFold remains slightly ahead on total oracle success (75.7% versus 73.7%), indicating that sample selection contributes to the difference at the acceptable threshold.

PXMeter-AB is the antibody–antigen subset of PXMeter, a large-scale unified evaluation [16]. TorchFold reaches 54.0% ranked success versus 45.3% OpenDDE, 41.3% AF3 and 39.9% Protenix-v2 ($n = 298$). Across all collections oracle success exceeds ranked success, indicating that additional successful interfaces are present in the sampled ensembles but are not selected. Performance varies across collections: TorchFold has higher ranked success on three collections, matches OpenDDE on AbAg2526 and has lower ranked success on 2026ARK-AB.

As an additional checkpoint-transfer experiment, we loaded the TorchFold antibody–antigen checkpoint into an adapted AF3 implementation (AF3-TF; Appendix A.1). Its success rates across the five benchmark collections were comparable to those of TorchFold, supporting transferability of the learned antibody–antigen prediction capability across the two implementations.

4.2 Accuracy across DockQ thresholds

Figure 5 reports the fraction of top-ranked interfaces exceeding each DockQ threshold. This shows whether a model's advantage extends from acceptable to more accurate interfaces. Over the full interval from 0 to 1, the area under this survival curve equals the mean DockQ. It summarizes the same predictions without selecting a single success threshold.

On FoldBench-AB and CPL-28, the improvement persists through the medium and high thresholds (AUC 0.54 and 0.49). On AbAg2526 the TorchFold and OpenDDE traces overlie (AUC 0.51), matching the tied rank rates. On 2026ARK-AB, OpenDDE stays above (AUC 0.54 versus 0.48), again where its high-quality share is largest. On PXMeter-AB, TorchFold remains higher across the sweep (AUC 0.40 versus 0.34 OpenDDE, 0.30 AF3 and 0.28 Protenix-v2). The principal differences therefore extend beyond the acceptable threshold, although they vary by benchmark.

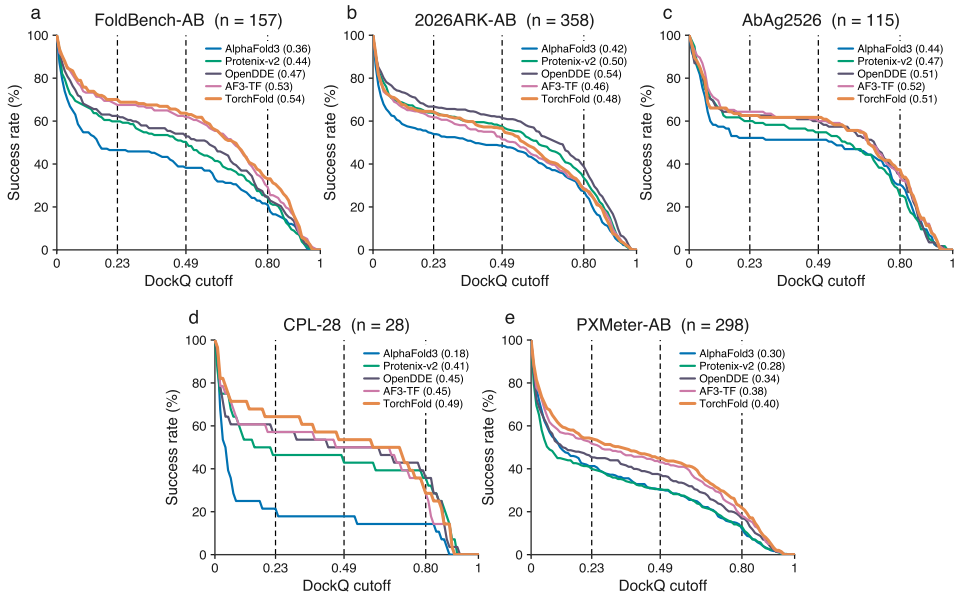

Figure 5 Cumulative DockQ success-rate curves. Fraction of interfaces whose top-ranked prediction reaches DockQ $\geq$ the cutoff on the x-axis. Dashed lines: acceptable, medium and high thresholds. Parentheses: area under the curve.

4.3 Test-time scaling

We evaluate whether additional inference-time sampling improves antibody–antigen prediction (Figure 6). For each FoldBench-AB target we generate 100 random seeds with 5 samples per seed. The ranked prediction is the sample with the highest model ranking score; the oracle prediction is the sample with the highest ground-truth DockQ. We report ranked success (DockQ > 0.23), ranked high-quality success (DockQ > 0.8), and oracle success (DockQ > 0.23). The oracle setting measures the performance available in the sampled set and separates sampling capacity from ranking quality.

As shown in Figure 6, TorchFold lies above AlphaFold3 at every seed count on FoldBench-AB. Ranked success rises from 62.4% at one seed to 72.4% at 100 seeds, against 41.9% and 52.6% for AlphaFold3. The ranked high-quality rate rises from 27.9% to 32.7%, against 18.9% and 21.2%. This compares sampling budgets, not measured runtime or total training cost.

The oracle curves show a larger scaling effect. TorchFold oracle success rises from 68.4% at one seed to 93.6% at 100 seeds; AlphaFold3 rises from 48.6% to 87.8%. The growing gap between ranked and oracle performance indicates that both models generate native-like interfaces that are not selected as the top-ranked prediction. The gap quantifies the potential benefit of improved selection within these ensembles.

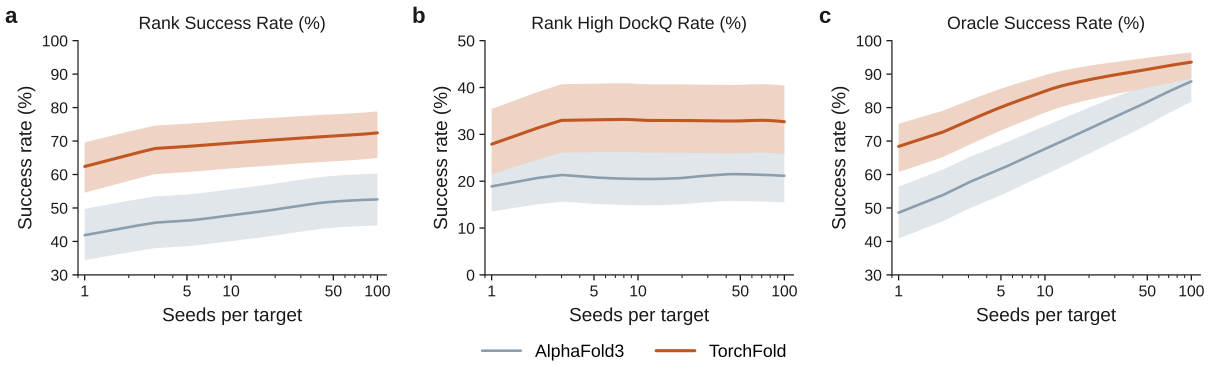

Figure 6 Test-time scaling on FoldBench-AB. 100 seeds × 5 samples per target. (a) Ranking-based selection (DockQ > 0.23). (b) Ranking-based selection of DockQ > 0.8. (c) Oracle-based selection (DockQ > 0.23).

4.4 Confidence as a filter for structural accuracy

We examined the relationship between ipTM and DockQ among top-ranked predictions pooled across the five collections (956 scored interfaces; Figure 7). For TorchFold, 292 of 297 predictions with ipTM ≥ 0.8 have DockQ > 0.23 (98.3%), and 288 of 297 have DockQ > 0.49 (97.0%). The mean DockQ in the accepted subset is 0.80 and the median is 0.84. This threshold accepts 297/956 interfaces (31.1% of the pooled evaluation). At ipTM ≥ 0.7, precision at DockQ > 0.23 is 416/441 (94.3%), with acceptance of 46.1% of interfaces. The nominal interface-level 95% Wilson interval for 292/297 is 96.1–99.3%. Excluding FoldBench-AB, which was used to choose the threshold during development, leaves 236/241 (97.9%) at 241/799 coverage (30.2%; Table 7).

OpenDDE assigns ipTM ≥ 0.8 to more predictions (464/956) but includes 44 unsuccessful interfaces (420/464, 90.5%). AF3 is more selective (159/956) and reaches 155/159 (97.5%). These values use the same numerical cutoff, not the same accepted fraction; a precision–coverage comparison is in Figure 12. Per-benchmark scatters and complex-level counts are in Appendix E.

High precision also coexists with missed opportunities for confident selection: 118 of the 523 TorchFold top-ranked interfaces with DockQ > 0.49 have ipTM < 0.7. Together with the rank–oracle gaps, this motivates evaluation of improved ranking and confidence coverage.

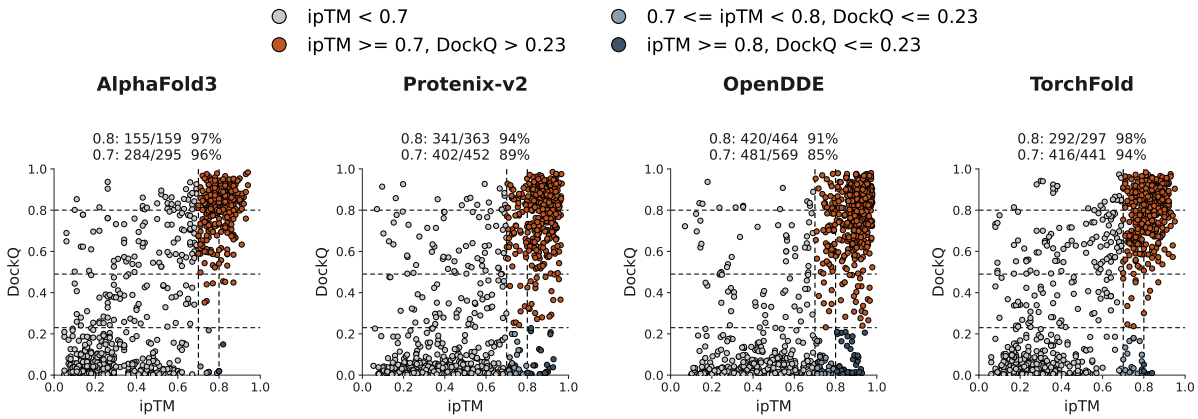

Figure 7 Ranked ipTM versus DockQ, pooled across benchmarks. Top-ranked antibody–antigen pair ipTM against DockQ for the evaluated models. The pooled TorchFold evaluation contains 956 scored interfaces. Vertical guides at ipTM = 0.7 and 0.8. Horizontal guides at DockQ = 0.23/0.49/0.80. Panel titles: successes (DockQ > 0.23) over models above each ipTM cutoff. Per-benchmark scatters in Figure 11.

4.5 Prediction accuracy decreases with antigen length

In the size-stratified analysis (956 scored interfaces), TorchFold recovers 74% of interfaces for antigens of ≤ 50 residues ($n = 68$), 65% at 51–150 ($n = 228$), 64% at 151–300 ($n = 354$), 57% at 301–600 ($n = 264$) and 38% beyond 600 residues ($n = 42$) (Figure 8). It leads or ties in the shorter bins and matches Protenix-v2 in the longest bin. Accuracy decreases for all evaluated models, although the longest-antigen group is small.

Length is associated with other differences in target composition and training coverage. In particular, ASD curation excludes antigens longer than 500 residues (Appendix A.4). Accuracy on large antigens remains a limitation of the current pipeline.

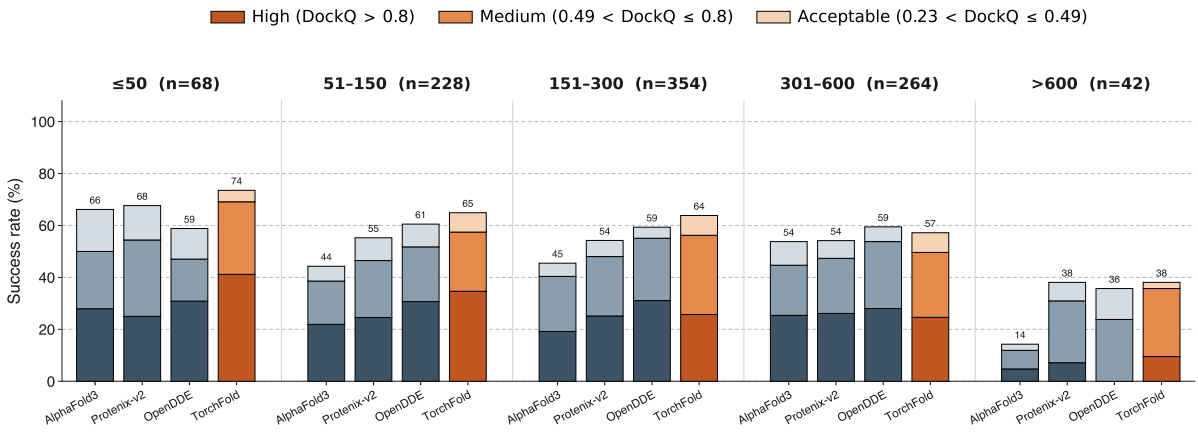

Figure 8 DockQ success versus antigen length. Interfaces pooled across FoldBench-AB, 2026ARK-AB, AbAg2526, CPL-28 and PXMeter-AB, binned by antigen residue count. Stacked segments: high (DockQ > 0.8), medium (0.49 < DockQ $\leq$ 0.8), acceptable (0.23 < DockQ $\leq$ 0.49). n is the number of interfaces in the bin.

4.6 Recovery of experimentally observed binding modes

Superposition on three crystallographic complexes shows where the predictions succeed or fail (Figure 9). All three examples are FoldBench-AB targets. Experimental coordinates are gray, TorchFold cyan and OpenDDE green. Fab constant domains are omitted. In 7TXW the TorchFold-predicted 1G2 Fv contacts both EGF-like domains of the malaria antigen Pfs25, as observed in the crystal (DockQ 0.90 and 0.85 for the two interfaces). In 8AV2, TorchFold-predicted VHH-4.80 occupies the crystallographic site on the FnIII domain of leptin receptor (DockQ 0.88). In 9JBQ the TorchFold-predicted h1F3 variable region recapitulates the crystallographic binding mode on PcrV of the *Pseudomonas* type III secretion system (DockQ 0.96).

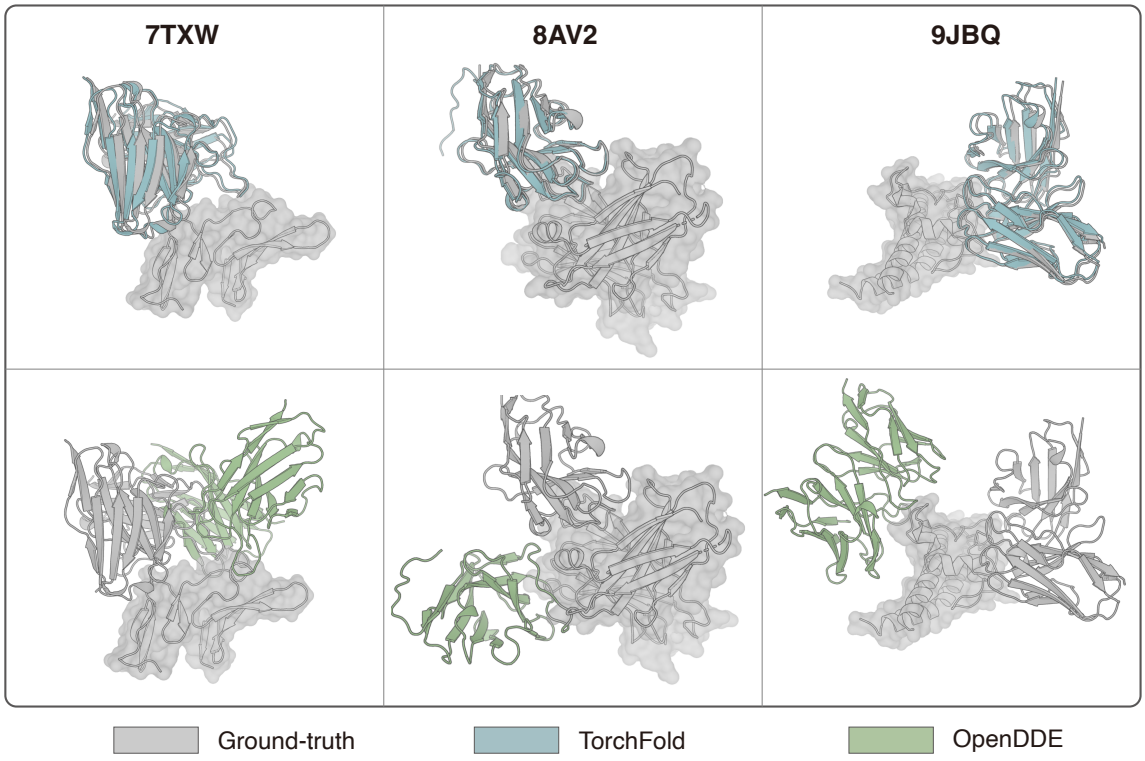

Figure 9 Recognition of three antibody–antigen epitopes. Top, TorchFold (cyan); bottom, OpenDDE (green), each aligned to the experimental antigen and overlaid on the crystallographic antibody (gray). Antigens are shown as cartoon plus a transparent surface. Fab CH1/CL domains are omitted. All three complexes are FoldBench-AB targets. Left to right: 1G2–Pfs25 (7TXW), VHH-4.80–LEP-R FnIII (8AV2), h1F3–PcrV (9JBQ).

Top-ranked models from AF3, Protenix-v2 and OpenDDE do not form these interfaces (DockQ 0.01–0.06). Across 25 samples per target (5 seeds $\times$ 5 diffusion samples), none of the three methods reaches DockQ ≥ 0.23 on 7TXW or 9JBQ, and neither Protenix-v2 nor OpenDDE does so on 8AV2. One AF3 sample of 8AV2 reached DockQ 0.64 but was not ranked first (DockQ 0.01). In the incorrect complexes, global lDDT remains 0.88–0.90: antibody and antigen are folded, yet the binder contacts a non-native surface. These examples illustrate errors in epitope assignment and relative orientation despite high global structural similarity.

4.7 Evaluation of supplied complex structures

The same checkpoint can evaluate supplied coordinates through the confidence pathway without diffusion-based structure generation. This score-only mode, TorchScore, was assessed on AF3-generated ensembles for 55 experimental complexes. The complete analysis and timings are provided in Appendix C.

5 Discussion

TorchFold develops structural distillation of antibody–antigen sequence pairs as a route to expanded supervision for complex prediction. Antigen-specific sequence collections encode interaction associations that are incompletely represented in experimental structure databases. By assigning confidence-filtered structural labels to these pairs and incorporating them into training, the method connects antibody discovery data to the learning of binding geometries. The complete training programme produces substantial gains over the starting checkpoint and competitive performance across multiple Ab–Ag benchmarks within an AF3-class architecture.

The value of self-training depends on the reliability and coverage of the accepted labels. A teacher can provide useful structural supervision on a subset of inputs even when its overall prediction accuracy is limited. Learning across those examples may improve prediction on other pairs, while a subsequent checkpoint can revisit candidates that were initially below the acceptance threshold. Matched teacher–student evaluation on complexes absent from both experimental and distilled training sources would further quantify this transfer.

The complete training programme combines additional structural labels, interface-weighted supervision and a revised noise schedule. The interface analyses show expanded predicted interface coverage and different contact distributions relative to SAbDab. The unchanged architecture demonstrates that additional capability can be obtained through training. Whether the same recipe further benefits from model scaling, and which of these changes transfer most readily, remain open questions.

Confidence filtering both enables and limits the approach. It can enrich useful labels, but may preserve teacher errors and preferentially retain familiar or readily scored interfaces. Experimental interaction evidence and computational design confidence should therefore be tracked separately when combining sequence sources. High ipTM precision on known binding pairs supports structural triage of predicted complexes. The observed gap between oracle and ranked accuracy leaves room for better selection. Direct reranking on the same ensembles is a natural next step.

Large antigens remain challenging. Their lower accuracy is associated with a broader search problem and a training distribution that includes length-based exclusions. Additional domain context, glycans and conformational variation are plausible contributors. Evaluations on unfamiliar antigen families and prospective experimental complexes would further define the method's range of application.

The broader opportunity is to use sequence-level interaction knowledge as a source of structural supervision wherever resolved complexes are scarce. Whether the same recipe transfers to other co-folding backbones or recognition systems is an empirical question. The released TorchFold implementation and checkpoint provide a basis for testing that possibility and for improving antibody–antigen structure prediction.

6 Data availability

Experimental structures are obtained from the PDB and SAbDab [8]. Public sequence-pair sources include ASD [13] and CoV-AbDab [14]; the training collection also includes patent, private discovery and PPIFlow design data. The experimental and predicted sources and their curation are described in Appendix A.3.

7 Code availability

GPU and NPU inference code is available at https://github.com/Mingchenchen/TorchX and https://gitcode.com/AI4Science/Mingchenchen, respectively. The training pipeline, including data-processing scripts and loss implementations, and the antibody–antigen model weights are released under the MIT license together with the inference code. The web server is available at https://torchfold.openi.org.cn/.

8 Acknowledgments

This work was funded by Changping Laboratory (CPNL project 2026D-03-01).

9 Author contributions

Mingchen Chen[1,†], Yunlong Cao[1,†], and TorchFold Team[1]

[1]Changping Laboratory, Beijing, China

[†]Corresponding author

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. 630(8016):493–500.

[2] Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life.

[3] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1 Democratizing Biomolecular Interaction Modeling.

[4] ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction.

[5] PaddleHelix Team. Technical report of HelixFold3 for biomolecular structure prediction.

[6] OpenFold3-preview-2 Technical Report.

[7] Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng, and Siqi Sun. Benchmarking all-atom biomolecular structure prediction with FoldBench. 17(1):442.

[8] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: The structural antibody database. 42(D1):D1140–D1146.

[9] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. 596(7873):583–589.

[10] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Towards Accurate and Efficient Binding Affinity Prediction.

[11] Jie Gao, Jing Hu, Lihang Liu, Yang Xue, Kunrui Zhu, Xiaonan Zhang, and Xiaomin Fang. PRECISE ANTIGEN-ANTIBODY STRUCTURE PREDICTIONS ENHANCE ANTIBODY DEVELOPMENT WITH HELIXFOLD-MULTIMER.

[12] Qilin Yu, Liangyue Guo, Xiayan Qin, Xikun Huang, Baihui Tian, Hongzhun Wang, Yu Liu, Yunzhi Lang, Di Wang, Zhouhanyu Shen, Jie Lin, and Mingchen Chen. High-Affinity Protein Binder Design via Flow Matching and In Silico Maturation.

[13] Arkadiusz Czerwiński, Paweł Dudzic, Konrad Wójtowicz, Igor Jaszczyszyn, Weronika Bielska, Sonia Wrobel, Samuel Demharter, Roberto Spreafico, Victor Greiff, and Konrad Krawczyk. ASD: Antigen-specific antibody database. 18(1):2623330.

[14] Matthew I J Raybould, Aleksandr Kovaltsuk, Claire Marks, and Charlotte M Deane. CoV-AbDab: The coronavirus antibody database. 37(5):734–735.

[15] Aureka AI OpenDDE project. Folding, Reasoning, and Scaling with Open-source Drug Discovery Engine.

[16] Wenzhi Ma, Zhenyu Liu, Jincai Yang, Chan Lu, Hanyu Zhang, and Wenzhi Xiao. From Dataset Curation to Unified Evaluation: Revisiting Structure Prediction Benchmarks with PXMeter.

[17] Yu Liu, Di Wang, Zhengyi Li, Shuxian Gao, Hongzhun Wang, Zhouhanyu Shen, Zhiqi Ma, Yucheng Zhang, and Mingchen Chen. TorchFold: A PyTorch and NPU-compatible reimplementation of AlphaFold3.

[18] Claudio Mirabello and Björn Wallner. DockQ v2: Improved automatic quality measure for protein multimers, nucleic acids, and small molecules. 40(10):btae586.

Appendix

Contents

A Training details

A.1 Architecture

TorchFold is a complete PyTorch reimplementation of AlphaFold 3, migrating all core components from JAX/Haiku to native `nn.Module` [17]. The architecture is strictly aligned with the official AF3 specification [1]—Input Embedder, MSA module and Pairformer trunk, Diffusion head, Confidence head, and Distogram head—so official AF3 checkpoints load without conversion. We introduce no architectural modifications. In Figure 1a, the block labelled Structure Module, placed immediately before diffusion, is the Pairformer trunk.

A.1.0.1 AF3-TF checkpoint transfer. AF3 generates fixed Fourier noise embeddings at runtime, so these quantities are absent from its checkpoint. TorchFold instead represents them as trainable parameters that are updated and stored during training. Consequently, direct conversion of a TorchFold checkpoint to the AF3 parameter tree would omit the learned embeddings. For AF3-TF, we minimally adapted the AF3 inference code to load and use TorchFold's learned Fourier weights and biases together with the converted checkpoint; no model layers were changed.

A.2 Hyperparameters and loss

The AF3 loss decomposition is retained, $\mathcal{L}_{\text{total}} = \alpha_{\text{diff}}\mathcal{L}_{\text{diff}} + \alpha_{\text{dist}}\mathcal{L}_{\text{dist}} + \alpha_{\text{conf}}\mathcal{L}_{\text{conf}}$, with $\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{smooth-LDDT}} + \mathcal{L}_{\text{bond}}$. Antibody–antigen fine-tuning adds an interface-specific MSE and a high-noise dynamic training schedule, defined below.

A.2.0.1 Interface-specific MSE. Interface residues are identified from the training reference structure (experimental coordinates or accepted distilled coordinates). In the residue-level setting, residue i is an interface residue if its representative atom is within d_{int} of a representative atom from a different chain:

$$\mathcal{I}_{\text{res}} = \left\{ i \,\middle|\, \exists j,\ a_i \neq a_j,\ \left\|\mathbf{x}_i^* - \mathbf{x}_j^*\right\|_2 < d_{\text{int}} \right\}, \tag{1}$$

where a_i is the chain identifier of residue i, $\mathbf{x}_i^*$ is its reference representative-atom coordinate, and $d_{\text{int}} = 10\,\text{Å}$ by default. All valid atoms of an interface residue inherit the interface mask. Alternatively, an atom is an interface atom if it is within d_{int} of any resolved atom from another chain.

Let $\mathcal{I}$ be the set of interface atoms. Before the coordinate error is evaluated, the reference structure is globally aligned to the prediction by a weighted Kabsch alignment over all valid atoms of the complex, not over the interface alone. The interface loss therefore measures interface accuracy in the global frame rather than after an independent interface-only rigid-body alignment.

For each interface atom $i \in \mathcal{I}$, the squared error after global alignment is

$$e_i = \|\mathbf{x}_i - \tilde{\mathbf{x}}_i^*\|_2^2, \tag{2}$$

where $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i^*$ are the predicted and globally aligned reference coordinates. Following AF3, each atom has a modality-dependent weight w_i. The interface-specific MSE is

$$L_{\text{interface}} = \frac{\sum_{i\in\mathcal{I}} w_i e_i}{|\mathcal{I}| + \epsilon}, \tag{3}$$

where $|\mathcal{I}|$ is the number of interface atoms and ϵ avoids division by zero. The coordinate loss is

$$L_{\text{coord}} = L_{\text{global}} + \lambda_{\text{int}} L_{\text{interface}}, \tag{4}$$

with $\lambda_{\text{int}} = 2$ in stage 1 and 10 in stage 2. Interface atoms therefore contribute to both the global aligned MSE and the extra term, increasing their relative weight while preserving whole-complex supervision.

A.2.0.2 High-noise schedule. The revised stage 2 schedule varies the noise parameter σ_{mean} with training step. It is initialized at $\sigma_{\text{mean}} = 10$ and decreased stepwise to 2, reduced by a fixed amount every K steps. With s the current step,

$$\sigma_{\text{mean}}(s) = \max\left(2,\, 10 - \Delta\sigma \left\lfloor \frac{s}{K} \right\rfloor\right), \tag{5}$$

where $\Delta\sigma$ is the decrement applied every K steps.

A.3 Training data

Stage-wise data mixtures are reported together with the schedule in Table 2.

From-scratch. Following the AF3 data-preparation protocol [1], we curate training data from the Protein Data Bank with a cutoff date of September 30, 2021 (Table 2). The Weighted PDB set follows the AF3 scheme that upweights recent and high-resolution structures. Distillation data (AF2/OpenFold predictions on MGnify, plus an OpenProteinSet subset of AF2 predictions on Uniclust30) provide single-chain structural diversity. Evaluation targets are excluded from this from-scratch collection.

TorchFold uses a pickle-based feature format in which each sample is a pre-featurized dictionary containing sequence features, atom coordinates, MSA profiles, template information, and CCD-derived reference conformers. The pipeline supports multi-source concatenation via `ConcatDataset`; contiguous, spatial, and spatial-interface cropping with configurable sampling weights and ligand/non-standard-residue preservation; length-grouped distributed sampling to reduce padding; optional epitope-mask sampling from ground-truth contacts; and per-data-source noise schedules (p_{mean}, p_{std}).

Fine-tuning. Antibody–antigen fine-tuning uses experimental complexes and confidence-filtered pseudo-structures as separate sources (Table 2 and Figure 2a). Distillation acceptance is ipTM ≥ 0.8. Per-source acceptance rates are in Figure 2b. SAbDab with a 1 January 2025 cutoff, after removal of identical sequences, provides experimentally resolved structural supervision in every round. The from-scratch checkpoint is used to generate first-round distillates: PPIFlow, patent and private discovery collections, ASD DB (Appendix A.4) and CoV-AbDab. Stage 1 trains on SAbDab plus this first-round mix. Second-round distillation re-infers first-round candidates with $0.5 \leq$ ipTM < 0.8 using the stage 1 checkpoint.

AbAg2526 and 2026ARK-AB are restricted to complexes released after this cutoff, and CPL-28 consists of structures that have not been released, so none of these three evaluations overlaps the SAbDab fine-tuning set by release date. FoldBench-AB and PXMeter-AB include earlier structures. To prevent leakage into those collections, we performed structure-based redundancy filtering of experimental training complexes rather than relying solely on deposition dates. Candidate training complexes were compared against each test target at the antibody, antigen and complex/interface levels using sequence and structural similarity, including antibody CDR regions and complex/interface structures. The similarity thresholds followed the criteria used in FoldBench [7]. Training structures exceeding these thresholds to any FoldBench-AB or PXMeter-AB target were removed, and the remaining data were clustered to define non-redundant training and validation sets.

A.4 ASD DB curation

ASD [13] compiles antibody–antigen sequence pairs and does not require a deposited complex. Before folding we removed pairs that already have experimental structures and reduced the remainder by antigen- and CDR-level clustering (Table 1).

Antigen sequences were clustered with MMseqs2 at 95% identity, then cluster representatives were re-clustered at 70%, giving 5,059 antigen clusters. Within each antigen cluster, antibody CDR sequences were clustered and only the centroid was retained. Clustering started at 95% CDR identity; if more than 500 representatives remained, the identity cutoff was lowered through 90, 80, 70, 60, 50 and 40% until the count fell below 500,

and that cutoff was used. Antigens longer than 500 residues were then discarded. The remaining pairs were folded, and complexes with ipTM ≥ 0.8 were accepted as first-round labels.

Table 1 ASD DB curation funnel. Counts are antibody–antigen pairs unless noted.

Stage	Count	Criterion
0	∼1,000,000	ASD pairs
1	530,000	Exclude pairs with experimental structures
2	5,059 Ag clusters	MMseqs2 antigen clustering at 95% identity, then re-cluster representatives at 70%
3	125,000	Within each antigen cluster, cluster antibody CDRs and keep the centroid. Start at 95% CDR identity; if > 500 representatives remain, lower the cutoff through 90, 80, 70, 60, 50, 40% until the count is < 500
4	79,061	Discard antigens longer than 500 residues

A.5 Training schedule

TorchFold is trained in three from-scratch stages on Huawei Ascend NPUs, followed by two antibody–antigen distillation rounds initialized from Scratch-3 (Table 2). Stage 1 is trained on experimental Ab–Ag structures and first-round labels. Stage 2 adds accepted second-round labels and changes the training settings listed below. The from-scratch protocol follows AF3 and Protenix [1, 4]. Scratch-2 and Scratch-3 are the AF3-style from-scratch fine-tuning stages (larger crop / confidence-head calibration), not the later Ab–Ag rounds.

From-scratch. Scratch-1 (initial training; 75K steps) uses crop size 384, 48 diffusion noise samples, and no bond loss, establishing core folding on PDB plus distillation. Scratch-2 (15K steps) raises the crop to 640 to capture longer-range interactions, turns on bond loss, and keeps the diffusion batch at 48 on the larger crop. Scratch-3 (10K steps) freezes the structure and distogram pathway and trains only the confidence head (PAE, pLDDT, PDE) on PDB-only data, with no distillation, to calibrate confidence estimates. Learning rates are 1.8×10^{-3} then 5×10^{-4}.

Fine-tuning. Stage 1 (FoldBench Ab–Ag 65.9%) loads the from-scratch checkpoint at crop 600 on SAbDab plus the first-round mix—PPIFlow, patent, private, ASD DB and CoV-AbDab—upweights interface cropping, and uses interface-loss weight 2.0. Stage 2 (70.1%) raises the crop to 768 and the diffusion batch to 48, switches on a revised noise schedule, sets the interface-loss weight to 10.0, trains the structure, distogram and confidence pathways with a binned confidence loss, and adds the second-round distillates obtained by re-inferring round-1 candidates with $0.5 \leq$ ipTM < 0.8 using the stage 1 checkpoint. This round also switches on a Protenix data adapter and fused kernels.

Table 2 Training stages for TorchFold. From-scratch stages (S-1–S-3) were trained on Huawei Ascend NPUs. From-scratch mix: Weighted PDB (ground-truth structures, weighted sampling), protein monomer distillation (AF2/OpenFold predictions on MGnify, uniform), and an OpenProteinSet subset (AF2 predictions on Uniclust30, uniform). Fine-tuning source rows mark which collections are used at each stage (✓). Distillation acceptance is ipTM ≥ 0.8. The 2nd-round source is first-round candidates with $0.5 \leq$ ipTM < 0.8, re-inferred with the stage 1 checkpoint. Per-source acceptance rates are in Figure 2b. The baseline success rate is reported separately from the two Ab–Ag fine-tuning stages.

	From-scratch			Fine-tuning	
				1st-round	2nd-round
Parameter	**S-1**	**S-2**	**S-3**	**1**	**2**
FoldBench Ab–Ag success (%)	34.0 (baseline)			65.9	70.1
Steps	75K	15K	10K	32K	45K
Crop tokens	384	640	640	600	768
Diffusion batch size	48	48	–	32	48
Learning rate	1.8×10^{-3}	5×10^{-4}	5×10^{-4}	1×10^{-4}	1×10^{-4}
Train structure / distogram	✓	✓	✗	✓	✓
Train confidence head	✗	✗	✓	✓	✓
Revised noise schedule	–	–	–	✗	✓
Interface loss weight	–	–	–	2.0	10.0
From-scratch mix (%)					
Weighted PDB (%)	60	60	100	–	–
Monomer distillation (%)	35	35	0	–	–
OpenProteinSet (%)	5	5	0	–	–
Fine-tuning sources					
SAbDab	–	–	–	✓	✓
PPIFlow	–	–	–	✓	✓
Patent	–	–	–	✓	✓
Private	–	–	–	✓	✓
ASD DB	–	–	–	✓	✓
CoV-AbDab	–	–	–	✓	✓
2nd-round ($0.5 \leq$ ipTM < 0.8)	–	–	–	–	✓

B Engineering optimizations

Training an AF3-scale model (∼386M parameters) with large diffusion batch sizes and long sequences requires careful memory and compute management. TorchFold incorporates the following optimizations. These are execution and numerical safeguards, not architectural deviations from AF3.

B.1 Corrections to the original AF3 description

During reproduction we identified and corrected several issues in the AF3 supplementary algorithms, consistent with those reported by Protenix [4] (Table 3).

Table 3 Corrections applied relative to the AF3 supplementary algorithms.

Location	Correction	Rationale
Alg. 1, Line 10	$\{z_{ij}\}$ = MsaModule (not $+=$)	MSA module contains internal residual connections; an additional residual is redundant.
Alg. 3, Line 8	Use $b_{ij}^{\text{same\_entity}}$	Symmetry-ID comparison is only meaningful within the same entity.
Alg. 18, Line 9	$\vec{\delta}_l = (\vec{x}_l^{\text{noisy}} - \vec{x}_l^{\text{denoised}})/\hat{t}$	Ensures consistency with the Euler ODE step in Line 11.
Alg. 23, Line 2	$\{b_i\}$ += AttentionPairBias	A missing residual causes convergence failure in the diffusion transformer.
Alg. 28, Line 11	$\vec{x}_l^{\text{align}} = R\vec{x}_l + \vec{\mu}_{\text{GT}}$	Typographical error; the alignment target must be the GT mean.
Alg. 31, Line 3	Distance bins: $[33/8, 45/8, \ldots]$ Å	Corrected to yield 15 equal-width bins of 1.25 Å.
Supp. §3.7.1, Eq. (6)	$(\hat{t}^2 + \sigma_{\text{data}}^2)/(\hat{t} \times \sigma_{\text{data}})^2$	Consistent with the EDM per-sample weighting formulation.

B.2 BF16 mixed-precision training

All forward passes through the Evoformer trunk, diffusion head, and confidence head are wrapped in `torch.cuda.amp.autocast(dtype=torch.bfloat16)`. BF16 reduces storage for tensors cast from FP32 while retaining a broad exponent range. Autocast alone does not imply that persistent parameters or optimizer states are stored in BF16. On hardware without BF16, TorchFold falls back to FP16 with `GradScaler`. TF32 is enabled for matrix multiplications and convolutions:

```
torch.backends.cuda.matmul.allow_tf32 = True
torch.backends.cudnn.allow_tf32 = True
torch.set_float32_matmul_precision("high")
```

B.3 Grouped gradient checkpointing

Gradient checkpointing is applied at multiple granularities to trade compute for memory:

- **Pairformer.** The 48-layer Pairformer stack is segmented into groups of configurable size (`pairformer_group_size`, default 2–8). Each group is wrapped in a single `checkpoint` call, so only one group's activations reside in memory during backward.
- **MSA module.** MSA Evoformer blocks are grouped similarly (default group size 2).
- **Template embedding.** The template stack is optionally checkpointed as a single unit.
- **Diffusion head.** Diffusion training samples are chunked along the sample dimension (`diffusion_checkpoint_group_size`), with each chunk independently checkpointed, preventing $O(N_{\text{sample}} \times N_{\text{atom}}^2)$ activation memory from accumulating.
- **Confidence head.** Confidence Pairformer layers are individually checkpointed during training. At inference the head is run one sample at a time.

All checkpoint wrappers use `use_reentrant=False` with a custom context that disables the AMP autocast cache, preventing stale cached values during recomputation:

```
@contextmanager
def _disable_amp_cache():
    torch.amp.autocast_mode._exit_all_autocast()
    yield
```

B.4 Chunked diffusion training and loss

During training the diffusion head processes N_{sample} noise samples (typically 32–48) per step. TorchFold splits them into chunks of configurable size and processes each chunk sequentially. Combined with per-chunk gradient checkpointing, this reduces peak activation memory in proportion to the number of chunks, at a given sequence length, while preserving gradient correctness.

Smooth LDDT and bond loss involve $O(N_{\text{sample}} \times N_{\text{atom}}^2)$ pairwise operations. When $N_{\text{sample}} \geq 8$, TorchFold chunks the loss along the sample dimension (chunk size 4 by default, overridable via `XFOLD_DIFFUSION_SAMPLE_CHUNK`). Each chunk is optionally checkpointed.

B.5 Numerical stability for multi-backend support

TorchFold targets both NVIDIA GPUs and Huawei Ascend NPUs. Several safeguards keep numerics stable across backends:

- **CPU-side SVD.** Kabsch alignment (`WeightedRigidAlign`) performs SVD in FP32 on CPU, then moves results back to the accelerator, avoiding GPU/NPU kernel instabilities in low-rank SVD.
- **AMP-disabled alignment.** Rigid alignment is wrapped in `torch.amp.autocast('cuda', enabled=False)` to prevent BF16 precision loss in the rotation matrix.
- **NaN-safe masking.** Atom cross-attention applies safe masking to avoid NaN propagation on masked positions, which is particularly important on NPU backends.
- **Gather indices on CPU.** Atom-layout gather indices (`GatherInfo.input_shape`) are kept on CPU to avoid device-specific integer indexing issues.

B.6 Distributed training

TorchFold supports multi-node, multi-GPU training via PyTorch DDP with the NCCL backend:

- **Device-aware process group.** `dist.init_process_group` is called with explicit `device_id` and an extended timeout (default 1800 s) for robustness under checkpoint I/O.
- **Length-grouped sampling.** A custom `LengthGroupedDistributedSampler` groups sequences by length across ranks to reduce padding overhead relative to random assignment.
- **Multi-source ratio sampling.** When mixing datasets, per-source sampling ratios can be specified per epoch.
- **Distributed inference (optional).** Diffusion steps can be distributed across ranks via `dist.broadcast` for multi-sample generation.

B.7 Memory-efficient inference and Triton kernels

For long-sequence inference, TorchFold processes confidence-head samples one at a time, calls `torch.cuda.empty_cache()` at strategic points (every 50 training steps, and before the confidence head), and detaches diffusion sample tensors to prevent graph retention during debug/logging.

The `xfold/fastnn` module provides optional Triton fused kernels for layer normalization, dot-product attention, and gated linear units. The default backend is pure PyTorch for portability.

C TorchScore

C.1 Structural scoring evaluation

TorchScore is a score-only loading of TorchFold: given input coordinates, it bypasses the diffusion-based structure module and evaluates the complex through the confidence pathway. The same antibody–antigen checkpoint used for structure prediction is therefore the scorer; no separate network is trained. GPU and NPU latencies are in Appendix C.2.

Antibody–antigen evaluation uses AbAg PDB55 (construction in Appendix C.3). For each of the 55 complexes, AF3 was run with 100 seeds and 5 samples per seed (500 candidates), from which 200 structures were drawn at random. DockQ against the native complex is used to assess interface quality [18]. Coverage and top-1 success are reported cumulatively at acceptable, medium and high thresholds, rather than as disjoint quality bins.

Because the ensemble is AF3-sampled, sampling coverage is the ranking oracle: at least one structure meeting each acceptable, medium and high threshold is present in 43/55, 30/55, and 20/55 cases. Ranking the same 200-sample ensembles, TorchScore reaches 30/55, 26/55, and 14/55 (Table 4). High-quality top-1 therefore remains below the sampling ceiling (14/55 versus 20/55). At score cutoff 0.7, the pooled candidate-level recall for high-DockQ structures is 96.05% at 59.65% precision.

Table 4 Coverage and ranking on the PDB55 200-sample subset. Coverage: at least one sampled structure at the DockQ threshold. Ranking: top-1 by TorchScore. Recall and precision are candidate-level statistics at score cutoff 0.7 and high-DockQ cutoff 0.8.

Method	Acceptable or better	Medium or better	High	High-DockQ recall at cutoff 0.7	High-score precision for High DockQ
AF3 sampling coverage	43/55	30/55	20/55	–	–
TorchScore	30/55	26/55	14/55	2555/2660 (0.9605)	2555/4283 (0.5965)

Across the pooled candidate structures, TorchScore ipTM correlates with DockQ at Pearson $r = 0.869$ (RMSE 0.187; Figure 10).

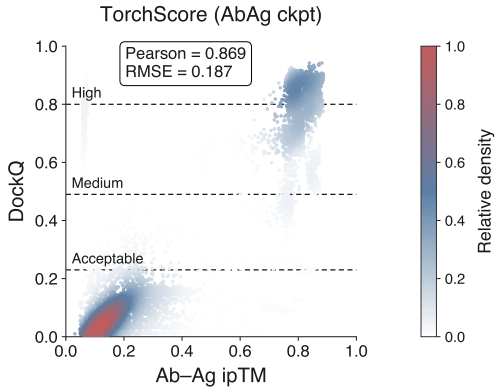

Figure 10 ipTM versus DockQ on AbAg PDB55. TorchScore loaded with the TorchFold Ab–Ag checkpoint. Dashed lines: Acceptable, Medium, and High DockQ thresholds.

C.2 Inference performance on GPU and NPU

TorchScore was run 3 times per input on protein sequences of different lengths. Average end-to-end latency is reported. TorchScore used 11 trunk recycles. Hardware platforms were NVIDIA A100 for GPU runs and Atlas 800T A2 for NPU runs.

Table 5 TorchScore end-to-end latency (seconds) versus sequence length.

Method	37	107	301	436	583	740	1024	2002
TorchScore (NPU, A2)	14.89	16.52	27.16	39.23	62.42	93.24	150.70	934.00
TorchScore (GPU, A100)	7.41	8.43	21.20	45.21	90.00	159.50	360.02	OOM

C.3 AbAg PDB55 construction

AbAg PDB55 was constructed from SAbDab [8] by retaining high-resolution structures (< 4 Å) with valid CDR–antigen contacts, clustering antigen sequences at 95% and 70% identity, keeping complexes released from 1 May 2025 to 31 August 2025 (48 clusters, 98 complexes), and excluding those longer than 512 residues, yielding 55 targets.

D AbAg2526 construction

AbAg2526 is an antibody–antigen evaluation set with temporal and paired-sequence filtering relative to the specified experimental training set. Structures are taken from SAbDab [8], restricted to entries released from 1 January 2025 to 31 August 2026.

Table 6 AbAg2526 construction funnel.

Stage	Criterion
1	SAbDab annotation table
2	Release in window; resolution ≤ 2.5 Å; single-chain protein antigen; type $\in$ {FV, FAB, SD-H}; one ASU copy per PDB + SAbDab ID
3	$0 <$ polymer chains < 300
4	No severe clash; $10 < \sum$ residues(Ab+Ag) < 2560; no heterologous antigen oligomer; discard if an additional non-antibody polymer chain (other than the designated antigen and any antibody chain in the PDB) makes ≥ 10 heavy-atom contacts (5 Å) to the antigen
5	Antibody–antigen heavy-atom interface at 5 Å; one ASU copy per SAbDab ID + PDB
6	Homology gate versus pre-2025 training PDBs (SAbDab release before 1 January 2025): discard only if CDR identity ≥ 0.85 and antigen identity ≥ 0.95 hit the same training PDB. CDR-only or antigen-only similarity is retained
7	Internal CDR-cluster (0.85) $\times$ antigen-cluster (0.95); keep the highest-resolution member of each cluster

The funnel yields AbAg2526: 71 complexes (115 scored interfaces in Figures 4 and 5). The homology gate excludes pairs close to the same pre-2025 experimental complex.

E ipTM versus DockQ by benchmark

Figure 7 pools top-ranked predictions. The same interface set, split by collection, is shown in Figure 11. Table 7 reports precision together with the fraction of each collection accepted at ipTM ≥ 0.8. Precision is high across the collections, while acceptance coverage ranges from 20.1% to 57.1%. Removing FoldBench-AB does not reduce precision below 97.9%. The ipTM ≥ 0.8 cutoff is the threshold used for distillation acceptance, not a screening rule applied to the benchmarks. Because that cutoff accepts 16.6% of AF3 predictions, 31.1% of TorchFold predictions and 48.5% of OpenDDE predictions, Figure 12 compares precision at matched acceptance rates. At TorchFold's 31.1% coverage, precision is 96.3% (AF3), 95.3% (Protenix-v2) and 96.0% (OpenDDE). Table 8 repeats the TorchFold ipTM ≥ 0.8 summary after collapsing scored interfaces to unique PDB complexes (172/177, 97.2%).

Table 7 TorchFold structural precision and acceptance coverage at ipTM ≥ 0.8. Successful interfaces have DockQ > 0.23. Coverage is accepted/scored; precision is successful/accepted. The pooled row includes all reported benchmark interfaces. The excluding-FoldBench-AB row omits the collection used to choose the threshold.

Collection	Scored	Accepted	Successful	Coverage (%)	Precision (%)
FoldBench-AB	157	56	56	35.7	100.0
2026ARK-AB	358	112	110	31.3	98.2
AbAg2526	115	53	51	46.1	96.2
CPL-28	28	16	16	57.1	100.0
PXMeter-AB	298	60	59	20.1	98.3
Pooled	956	297	292	31.1	98.3
Excluding FoldBench-AB	799	241	236	30.2	97.9

Table 8 TorchFold precision and coverage at ipTM ≥ 0.8, unique PDB complexes. A complex is accepted if it has at least one interface with ipTM ≥ 0.8, and successful if every accepted interface has DockQ > 0.23. The pooled row counts each PDB identifier once (16 accepted PDBs occur in more than one collection). CPL-28 identifiers are unique per scored chain interface. One accepted complex (9pso) has mixed interfaces and is counted as unsuccessful. Nominal 95% Wilson interval for the pooled row: 93.6–98.8%.

Collection	Scored	Accepted	Successful	Coverage (%)	Precision (%)
FoldBench-AB	103	42	42	40.8	100.0
2026ARK-AB	147	55	53	37.4	96.4
AbAg2526	69	34	32	49.3	94.1
CPL-28	28	16	16	57.1	100.0
PXMeter-AB	213	46	45	21.6	97.8
Pooled (unique PDB)	520	177	172	34.0	97.2

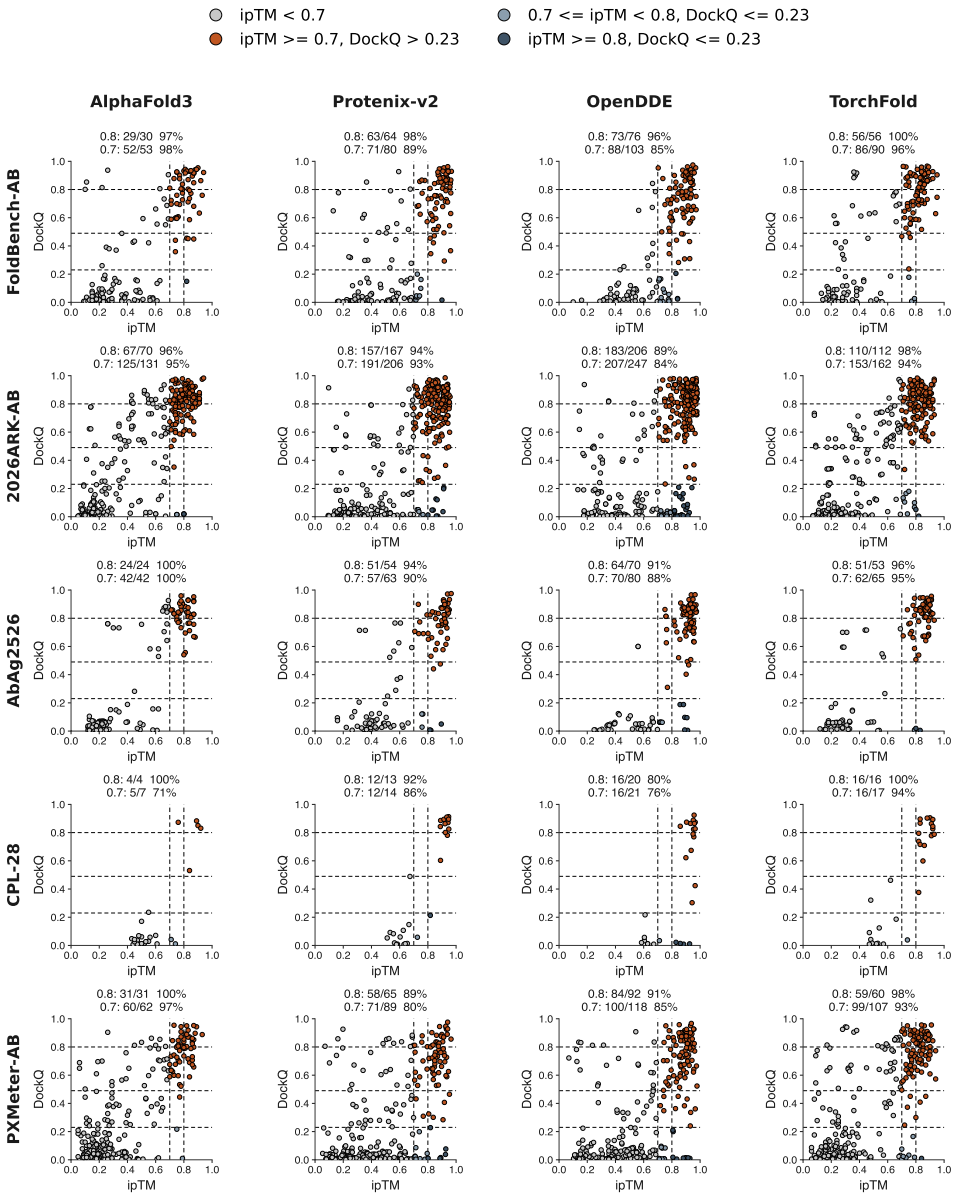

Figure 11 Ranked ipTM versus DockQ by benchmark. Available per-benchmark predictions, using the coloring of Figure 7. Panels show FoldBench-AB, 2026ARK-AB, AbAg2526, CPL-28 and PXMeter-AB. Panel titles: successes (DockQ > 0.23) over models above ipTM 0.8 and 0.7.

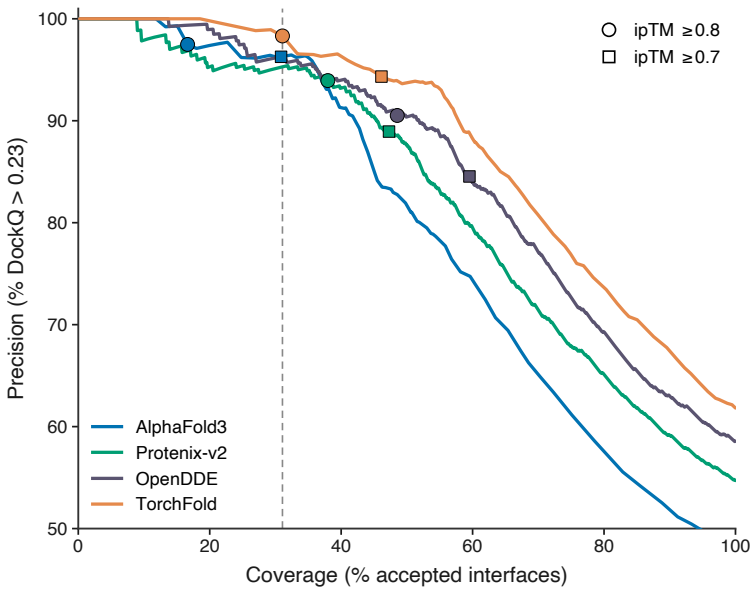

Figure 12 Precision versus coverage while sweeping the ipTM cutoff. Pooled top-ranked predictions, 956 scored interfaces. Coverage is the fraction of interfaces with ipTM at or above the cutoff; precision is the fraction of those interfaces with DockQ > 0.23. Circles, ipTM ≥ 0.8; squares, ipTM ≥ 0.7. The vertical guide marks TorchFold coverage at ipTM ≥ 0.8 (31.1%). At that coverage, precision is 98.3% (TorchFold), 96.3% (AF3), 95.3% (Protenix-v2) and 96.0% (OpenDDE).