EMBER: PREDICTING TCR–PMHC BINDING WITHOUT A DOCKED COMPLEX

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose **EMBER**, a structure-based framework for TCR–pMHC binding prediction that eliminates the need for a predicted docked complex, and with it the placement error that dominates such predictions. Traditional structure-based predictors utilize the co-folded complex structure of the receptor and the antigen and interpret binding from its contact surface. Consequently, every downstream feature inherits any deviation the structure-prediction tool introduced. Methodologically, EMBER folds the TCR and pMHC independently, encodes them as SE(3)-invariant residue graphs, and compares their learned representations without constructing a shared coordinate frame. A relative pose is therefore not an input of the model, so placement error cannot propagate into the prediction. Computationally, EMBER preprocesses $\mathcal{O}(N+M)$ monomer folds where a complex-based predictor needs $\mathcal{O}(NM)$ operations. Experimentally, under epitope-disjoint evaluation with six retrained baselines, EMBER is first on every metric under all four folding tools, by **0.9 to 14.4%** in micro-AUROC, **18.5 to 25.3%** in micro-AUPRC and **10.6 to 24.1%** in MCC. Its ranking is also up to **85%** more stable across folding tools.

1 INTRODUCTION

We propose **EMBER**, a structure-based predictor of TCR–pMHC binding that **never forms a docked complex**. It folds each molecule once as an unbound monomer, encodes the two independently, and scores the pair by comparing SE(3)-invariant descriptors, so the relative pose of the two molecules never enters the computation. Because TCR–pMHC binding determines adaptive immune specificity (Garcia & Adams, 2005) while the candidate space is far too large for exhaustive experimental screening, EMBER provides a scalable way to prioritize likely binders, including for unseen epitopes.

Existing structure-based TCR–pMHC predictors make this decision through a predicted docked complex, first folding or docking the two molecules into a common assembly and then inferring binding from the predicted interface, using residue contacts, geometric features, or structure-derived graphs (Slone et al., 2025a;b; Li et al., 2025; Abeer et al., 2026; Polster et al., 2026; Siriarchawatana et al., 2026). The design follows the biology, since recognition happens at the interface and sequence alone does not explain specificity (Garcia & Adams, 2005; La Gruta et al., 2018). However, it also makes the downstream predictor dependent on the accuracy of the upstream complex prediction. In particular, these methods typically retain only residues within a fixed distance of the opposing molecule (Slone et al., 2025a; Li et al., 2025; Abeer et al., 2026). The predicted relative pose therefore determines not only the geometric features presented to the model, but also which residues and intermolecular contacts exist in its input graph. An error in rigid-body placement can thus change the input itself, rather than merely perturb its coordinates. Although Polster et al. (2026) attribute their model's limitations to structural prediction error, whether predicted TCR–pMHC complexes are reliable in these task-critical quantities has not been systematically tested.

To address this limitation, we propose **EMBER**, which removes complex assembly rather than correcting its errors. Each molecule is folded as an unbound monomer and represented as a residue graph, with Cα-centered nodes carrying local geometric descriptors and cached protein language-model embeddings. Edges connect nearest neighbors and encode spatial separation and biochemical compatibility. Because all input features are invariant to rigid-body transformations, no absolute coordinates or inter-molecular pose enter the representation. Given these pose-independent graphs, a shared equivariant encoder processes the two molecules separately, restricting message passing to within each molecule and updating neighborhoods from the evolving coordinates at each layer. Cross-molecule interaction occurs only through bidirectional attention between the CDR3 loops and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

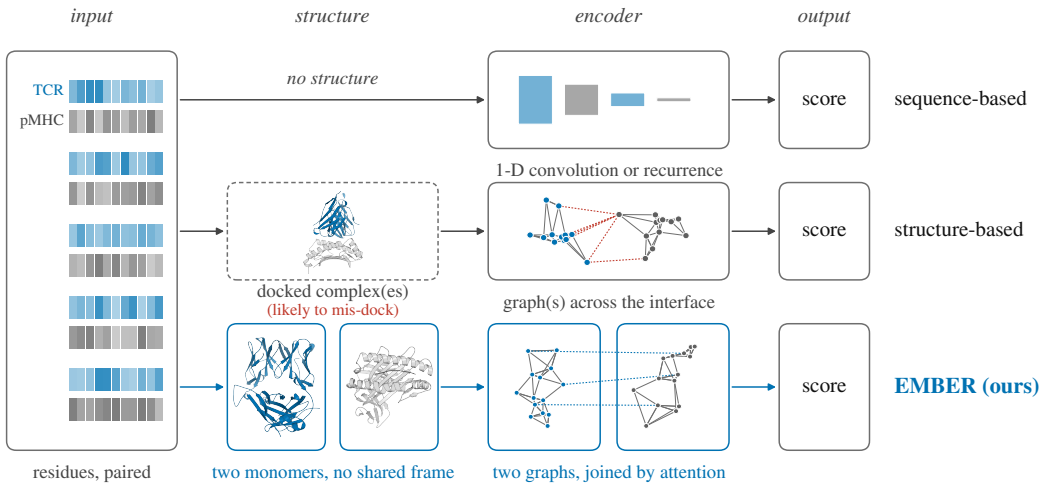


Figure 1: Three routes to binding prediction. All methods score the same molecular pair. Sequence-based methods use no structural information; structure-based methods dock the molecules and construct a joint interface graph with pose-dependent contacts; EMBER instead folds each molecule independently and connects the two monomer graphs only through attention.

peptide, whose residue sets are defined by sequence annotation rather than predicted geometry. Thus, EMBER neither constructs a shared coordinate frame nor selects residues using inter-molecular distance cutoffs. Pose independence is therefore imposed by the model architecture and is preserved throughout training.

Contributions. We present **EMBER** (Figure 1), a monomer-only TCR–pMHC structure-based predictor which is provably independent of the relative pose of the two molecules. Our main contributions are:

- We identify **assembly-induced interface corruption (AIC)**, a previously uncharacterized failure mode in structure-based TCR–pMHC prediction, where pose errors corrupt the residues and contacts exposed to downstream predictors.
- We propose EMBER, a pose-free structure-based predictor that folds the TCR and pMHC independently as unbound monomers. Each molecule is encoded as an SE(3)-invariant residue graph with a shared encoder, and binding is scored by bidirectional attention over the resulting representations.
- EMBER preprocesses $\mathcal{O}(N+M)$ monomer folds, against the $\mathcal{O}(NM)$ complex predictions a complex-based predictor needs.
- Experimentally, EMBER leads six published baselines retrained on identical data and splits, first on every metric under all four folding tools. It is also significantly more stable to the choice of folding tool.

2 Related Work

Existing approaches to TCR–pMHC binding prediction fall broadly into **sequence-based** and **structure-informed** families. Sequence models, including NetTCR (Montemurro et al., 2021; Jensen & Nielsen, 2024) and ERGO-II (Springer et al., 2021), learn directly from amino-acid sequences and benefit from substantially larger sequence datasets than are available for solved TCR–pMHC complexes (Vita et al., 2015; Shugay et al., 2018; Leem et al., 2018). However, recent benchmarking and systematic evaluation show that generalization to unseen epitopes remains limited (Lu et al., 2026; Deng et al., 2026). TEIM predicts the residue contacts themselves from sequence and HeteroTCR builds a heterogeneous graph over sequence embeddings, so neither reads a structure (Peng et al., 2023; Yu et al., 2024). Advances in protein structure prediction enabled a second family of methods that incorporates predicted three-dimensional structure (Jumper et al., 2021; Abramson et al., 2024). Most construct a predicted TCR–pMHC complex and infer binding from its interface using residue contacts, geometric features, and graph representations (Bradley, 2023; Slone et al., 2025a;b; Li et al., 2025; Polster et al., 2026; Siriarchawatana et al., 2026). These methods typically define the interface by a distance cutoff of 8 to 14 Å, so the predicted pose determines not only residue geometry but also which residues and intermolecular contacts enter the model. Other approaches reduce this dependence by using the structure of only one binding partner (Zhao et al., 2023; Ji et al., 2024; Ge et al., 2025;

Shang et al., 2026) or by using structure only during training (Li et al., 2026), but do not exploit the structural information of both molecules. Prior work has already reported unreliable docking confidence (Deleuran & Nielsen, 2025) and that non-interacting complexes adopt conformations geometrically similar to interacting ones (Robben, 2026).

EMBER occupies a distinct point in this design space. STAG, SageTCR and the classifier of Abeer et al. (2026) identify only the residues within 10 to 14 Å of the opposing molecule as the interface (Slone et al., 2025a; Li et al., 2025). STAG-LLM keeps every residue and wires the assembly with a 10 Å radius graph (Slone et al., 2025b), so there the pose decides the contacts rather than the residues. EMBER processes each monomer in full and links the two only through sequence-annotated residue sets, so neither choice is exposed to the pose. TRAP avoids the assembly by dropping the receptor structure and representing the TCR by sequence (Ge et al., 2025), while EMBER keeps the structure of both. caTCR (Ji et al., 2024) also folds its two inputs separately, though it folds only the CDR3β loop and the epitope and leaves the MHC out of the model. It reduces each to a matrix of pairwise Cα distances and reads that matrix with a convolutional network, whose neighborhoods follow the residue order rather than the spatial arrangement. EMBER carries coordinates as an equivariant field and updates them at every layer. Running an E(n)-equivariant network over the joint graph, as Abeer et al. (2026) does, leaves the score invariant to motion of the whole assembly while the relative pose stays in the input. In contrast, EMBER removes the relative pose by Proposition 1.

3 Assembly-induced Interface Corruption

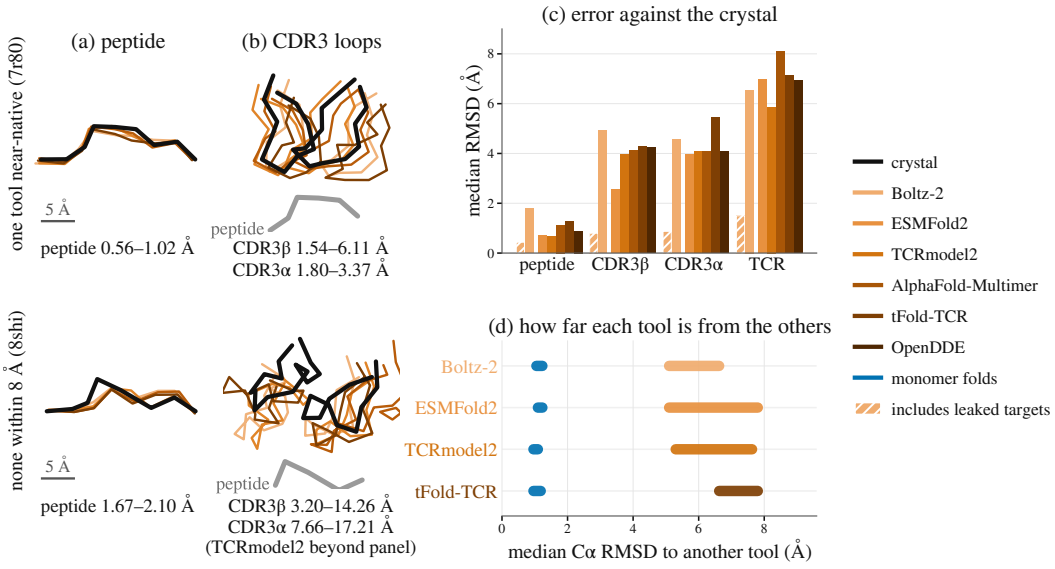


Figure 2: The predicted interface, and where its error lives. (a, b) Two targets in the plane of the peptide, with the crystal drawn in black. The upper one lies inside Boltz-2's training window, the lower one is an example of generalization failure where no tool places the receptor within 8 Å. **(c)** Median Cα RMSD after superposing on the MHC heavy chain. Boltz-2 appears twice, hatched over all 49 targets and solid over the 10 released after its training cutoff. **(d)** Each bar spans one tool's pairings with the others, each tool's own color for receptor placement on the 49 complexes and blue for whole-monomer agreement over the same 250 TCR monomers. AlphaFold-Multimer and OpenDDE have no monomer folds. TCRmodel2 is the TCR-specialized version of AlphaFold-Multimer, and OpenDDE shares its AlphaFold3 architecture with Boltz-2, so each is already represented.

In this section, we test whether predicted TCR–pMHC complexes provide reliable structural inputs for downstream binding prediction. We evaluate six co-folding tools on 49 class-I TCR–pMHC crystal structures from STCRDab (Leem et al., 2018). The tools are Boltz-2 (Passaro et al., 2025), ESMFold2 (Candido et al., 2026), TCRmodel2 (Yin et al., 2023), AlphaFold-Multimer (Evans et al., 2021), OpenDDE (Aureka AI OpenDDE Project, 2026) and tFold-TCR (Wu et al., 2025). We focus on two quantities consumed by structure-based predictors: the intermolecular contacts at the interface and the relative placement of the two molecules.

We term **assembly-induced interface corruption (AIC)** the failure mode in which rigid-body placement error changes the interface presented to the downstream model. Because traditional

structure-based models construct interface graphs using distance cutoffs of 8–14 Å, an error of comparable magnitude can alter which residues and intermolecular contacts enter the graph, producing misleading information. In short, we found that no tool reconstructs the native interface reliably. Setting Boltz-2 aside, the median contact Jaccard at 5 Å runs from 0.188 to 0.459 and the median receptor Cα RMSD from 5.87 to 8.09 Å. Boltz-2 is best in almost every row and reaches a Jaccard of 0.833, but 39 of the 49 targets were seen by it during training. On the 10 targets released after Boltz-2's training cutoff, its Jaccard falls from 0.833 to 0.155. The other five tools move far less on those same targets, ESMFold2 from 0.459 to 0.393 and AlphaFold-Multimer from 0.308 to 0.303. Therefore the lead of Boltz-2 reflects memorization rather than generalization ability. Table 8 and Table 9 list the details.

More specifically, we locate the error in assembly rather than in monomer folding. Peptides remain close to their crystal positions, whereas CDR3 loops and the receptor as a whole are displaced substantially further (Figure 2c). Across tools, receptor placements differ by 5.10 to 7.80 Å on the same complexes, while independently folded TCR monomers agree to within 0.96 to 1.23 Å over the same 250 TCR monomers (Figure 2d). Even the two tools that agree most on placement sit 5.10 Å apart, four times the widest monomer disagreement. The tools largely agree on the individual molecular structures and disagree on how to assemble them.

Most co-folding tools are stochastic, so one natural solution is to fold each complex K times and average the error away. Write the AIC of sample k as a displacement vector v_k and define *directional coherence* $\gamma = \left\| \frac{1}{K} \sum_{k=1}^{K} v_k / \|v_k\| \right\| \in [0, 1]$, which is 1 when every sample misses the same way. When the displacements have similar length, averaging K samples leaves at least a fraction γ of the error, so measuring γ shows what sampling can achieve. Five tools fold each target 20 times and reach γ of 0.84 to 0.992, four times the 0.21 expected from random directions (tFold-TCR is deterministic and offers no distribution to sample). Equal-weight averaging accordingly moves the displacement by -2.5% to $+6.7\%$. Moreover, this bias is not specific to one tool. In a common frame defined by the peptide-binding groove, five independently built tools share one displacement direction, with a cross-tool γ of 0.621 and agreement beyond random on 40 of 48 targets ($p = 3.3 \times 10^{-6}$). Therefore, we claim that assembly-induced interface corruption is a systematic bias.

Finding 1. The dominant structural error in predicted TCR–pMHC complexes lies in rigid-body assembly. It is large enough to change which residues and contacts enter a downstream graph, and its direction is systematic, so multiple sampling and ensembling cannot remove it.

4 METHOD

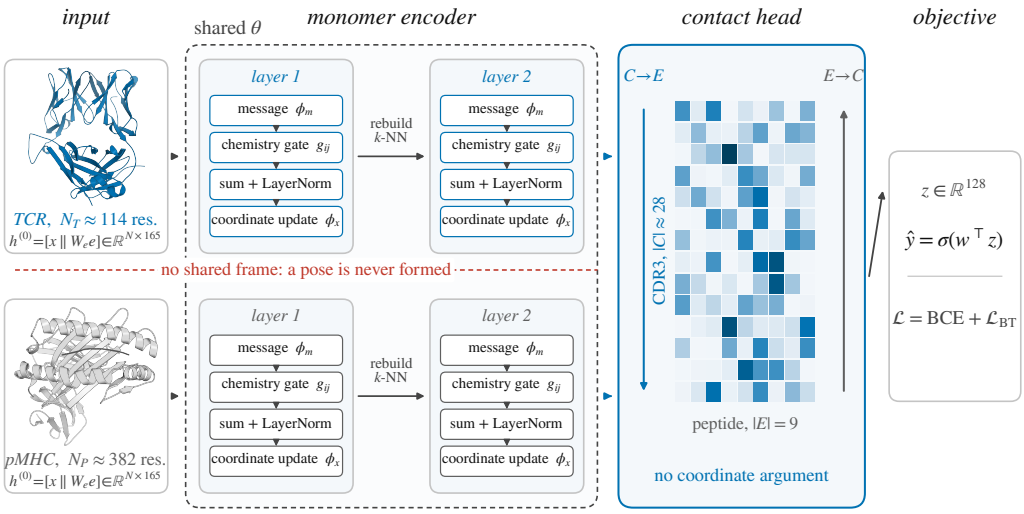


Figure 3: EMBER. The top row is the TCR and the bottom row the pMHC. Tensor shapes are annotated on the flow, and both encoder layers are drawn. In the contact head, C indexes the CDR3β residues and E the peptide residues, so the map carries one row per CDR3β residue and one column per peptide residue. Its values are schematic. The dashed box encloses both towers because they share weights and exchange no message.

Algorithm 1 EMBER forward pass.

Require: monomer graphs $\mathcal{G}^T, \mathcal{G}^P$; encoder θ; depth L; neighbors k
for $\mathcal{G} \in \{\mathcal{G}^T, \mathcal{G}^P\}$ **do**
  $h^{(0)} \leftarrow [\, x \,\|\, W_e e \,]$; $p^{(0)} \leftarrow$ Cα coordinates *# shared θ, no message between graphs*
  for $\ell = 1$ **to** L **do**
    $\mathcal{E}^{(\ell-1)} \leftarrow \mathrm{kNN}(p^{(\ell-1)}, k)$; $h^{(\ell)}, p^{(\ell)} \leftarrow \mathrm{EGCL}_\theta\big(h^{(\ell-1)}, p^{(\ell-1)}, \mathcal{E}^{(\ell-1)}\big)$ *# Eq. 4.1–4.2*
  end for
  $H \leftarrow h^{(L)}$ stacked by residue
end for
$C \leftarrow$ CDR3 rows of H^T; $E \leftarrow$ peptide rows of H^P *# invariant scalars only*
$z_t \leftarrow \mathrm{proj}_1(\mathrm{pool}_C(\mathrm{MHA}_1(H^T_C, H^P_E)))$; $z_p \leftarrow \mathrm{proj}_2(\mathrm{pool}_E(\mathrm{MHA}_2(H^P_E, H^T_C)))$
$z \leftarrow \mathrm{LN}(W_z[\, z_t \,\|\, z_p \,])$
return $\hat{y} = \sigma(u)$, $u = w^\top z$ *# no coordinate enters the contact head*

We propose **EMBER** (Figure 3), a framework for TCR–pMHC binding prediction that folds each molecule independently and represents it as a residue graph with SE(3)-invariant features. A shared-weight encoder processes the two graphs separately. A bidirectional attention head then compares and pools their residue representations, and a linear layer produces the binding score. Algorithm 1 states the forward pass as pseudocode, and Table 5 introduces the symbols in detail.

Problem Definition. EMBER represents the receptor and antigen as residue graphs, $\mathcal{G}^T$ and $\mathcal{G}^P$. The receptor graph carries both TCR chains, and the encoder is given the β chain alone, $N_T \approx 114$ nodes (Appendix D). The antigen graph holds the MHC heavy chain, β_2-microglobulin and the peptide, $N_P \approx 382$ nodes. Each residue i is described by its Cα coordinate $p_i \in \mathbb{R}^3$, an invariant local descriptor $x_i \in \mathbb{R}^{37}$, and a frozen ESM-C embedding $e_i \in \mathbb{R}^{1152}$ (Candido et al., 2026). The descriptor x_i depends only on local geometry and residue identity. It is therefore invariant to rigid motion, and it introduces no absolute coordinate and no preferred axis. The encoder input is $h_i^{(0)} = [x_i \,\|\, W_e e_i] \in \mathbb{R}^{165}$ and $p_i^{(0)} = p_i$, with $W_e \in \mathbb{R}^{128\times1152}$ learned. Graph edges connect the $k = 30$ nearest neighbors in Cα distance $d_{ij} = \|p_i - p_j\|$. Each edge carries $a_{ij} = [\rho(d_{ij}) \,\|\, d_0/d_{ij} \,\|\, b_{ij}] \in \mathbb{R}^{22}$, where ρ is a 16-dimensional radial basis over $[0, 20]$ Å and $d_0 = 4$ Å. The five entries of b_{ij} read residue identity alone: hydrogen bonding, hydrophobicity, van der Waals radius, charge product, and aromaticity.

Monomer Encoder. A single encoder θ runs on the two graphs without any message between them. It first projects $h_i^{(0)}$ to the working width $D = 256$. Every layer then repeats the same four steps in order, a message on each edge, a chemistry gate on that message, a node update and a coordinate update, and closes by rebuilding the neighborhood from the new coordinates. Writing $\mathcal{N}(i)$ for the neighbors of i, σ for the logistic sigmoid, LN for layer normalization and $c_{ij} = [\, h_i^{(\ell-1)} \,\|\, h_j^{(\ell-1)} \,\|\, a_{ij} \,]$ for the edge context, layer ℓ maps $(h^{(\ell-1)}, p^{(\ell-1)}) \mapsto (h^{(\ell)}, p^{(\ell)})$ by

$$m_{ij} = \phi_m(c_{ij}), \qquad g_{ij} = \sigma\big(\phi_g(b_{ij})\big) \in (0,1), \tag{4.1}$$

$$h_i^{(\ell)} = \mathrm{LN}\Big(h_i^{(\ell-1)} + \sum_{j\in\mathcal{N}(i)} g_{ij}\, m_{ij}\Big), \quad p_i^{(\ell)} = p_i^{(\ell-1)} + \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} \big(p_i^{(\ell-1)} - p_j^{(\ell-1)}\big)\, \phi_x(c_{ij}), \tag{4.2}$$

after which the edge set is rebuilt from the updated coordinates as $\mathrm{kNN}(p^{(\ell)}, k)$. This is the E($n$)-equivariant layer of Satorras et al. (2021) with two changes. First, the gate g_{ij} reads residue identity only. Chemistry therefore modulates how much geometry is passed, and equivariance is untouched. Second, the graph is rebuilt at every layer. Coordinates reach h only through d_{ij} inside a_{ij}, so the updated coordinates would never influence h without the rebuild.

Contact Head. The two molecules meet in the bidirectional contact head that takes no coordinate argument. Write $H^T \in \mathbb{R}^{N_T\times D}$ and $H^P \in \mathbb{R}^{N_P\times D}$ for the final states $h^{(L)}$ stacked by residue. Let C index the CDR3β rows of the receptor and E the peptide rows of the antigen. Both index sets come from sequence annotation rather than from the predicted geometry, so they cannot be altered by AIC, unlike a distance cutoff. Both directions use the standard attention of Vaswani et al. (2017) with four heads of width 32. **TCR to peptide**, $\mathrm{MHA}_1(H^T_C, H^P_E)$, lets the CDR3$\beta$ rows query the

peptide. **Peptide to TCR**, $\mathrm{MHA}_2(H_E^P, H_C^T)$, lets the peptide rows query the CDR3β rows. Their per-query outputs are pooled over C and over E and projected, giving $z_t = \mathrm{proj}_1(\mathrm{pool}_C(\cdot))$ and $z_p = \mathrm{proj}_2(\mathrm{pool}_E(\cdot))$, both in $\mathbb{R}^{128}$. The pair is scored from their concatenation, $z = \mathrm{LN}(W_z[\, z_t \| z_p \,]) \in \mathbb{R}^{128}$, logit $u = w^\top z$, and $\hat{y} = \sigma(u)$.

TCR to peptide pools over C rather than over E, so peptide length enters only through the attention over keys and never sets the pooled dimension. It carries little receptor-side signal of its own, and it needs no canonical frame. In contrast, Peptide to TCR needs a canonical frame, because peptide length ranges from 8 to 11 residues in our dataset and pooling over raw positions would give position k a different chemical role. It therefore pools on a canonical frame of $S = 9$ slots. Class-I binding pins the peptide at both ends regardless of length, the free N-terminus in the A pocket, the second residue P2 in the B pocket and the C-terminal residue PΩ in the F pocket (Madden, 1995; Szeto et al., 2021). A monotone map $\pi : \{1, \ldots, |E|\} \to \{1, \ldots, S\}$ sends those three to slot 1, slot 2 and slot S, and spreads the interior residues over the remaining slots, placing surplus residues and gaps centrally, where a longer peptide bulges away from the groove (Tynan et al., 2005). An 8-mer and an 11-mer therefore present the same anchor at the same slot, and length is absorbed where class-I presentation already tolerates it. Writing O for the per-residue output of MHA_2, $\mathrm{pool}_E(O) = \big[\, q_1 \| \cdots \| q_S \,\big], q_s = \big[\, \mathrm{mean}_{\pi(i)=s}\, O_i \;\big\|\; \max_{\pi(i)=s}\, O_i \,\big]$, and $\mathrm{proj}_2 : \mathbb{R}^{2S \cdot 128} \to \mathbb{R}^{128}$.

Objective Function. Let a batch $\mathcal{B} = \mathcal{B}^+ \cup \mathcal{B}^-$ hold binders $\mathcal{B}^+$ and non-binders $\mathcal{B}^-$, and let $y_i \in \{0, 1\}$ be the label of pair i. EMBER is trained on $\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{BT}}$ with

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \big[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \big], \mathcal{L}_{\mathrm{BT}} = -\frac{1}{|\mathcal{B}^+||\mathcal{B}^-|} \sum_{a \in \mathcal{B}^+,\, b \in \mathcal{B}^-} \log \sigma(u_a - u_b).$$

The first term calibrates the per-pair probability. The second ranks candidates within the batch (Bradley & Terry, 1952). See Appendix F.3 for detailed hyperparameters.

4.1 Theoretical Analysis

A rigid motion (R, t) acts on a residue graph by $p_i \mapsto Rp_i + t$, leaving residue identities untouched.

Definition 4.1 (encoder inputs). The encoder input $h^{(0)}$, the edge features a and the k-NN edge set are unchanged by (R, t).

Lemma 4.1 (one layer). If $h^{(\ell-1)}$ is unchanged by (R, t) and $p^{(\ell-1)} \mapsto Rp^{(\ell-1)} + t$, then layer ℓ leaves $h^{(\ell)}$ unchanged and sends $p^{(\ell)} \mapsto Rp^{(\ell)} + t$.

Proof. See Appendix B.1. □

Proposition 1 (Pose independence). Let $s(\mathcal{G}^T, \mathcal{G}^P) = \hat{y}$ be the binding score. For any rotations $R_1, R_2 \in SO(3)$ and translations $t_1, t_2 \in \mathbb{R}^3$ applied *independently* to the two molecules,

$$s\big(R_1\mathcal{G}^T + t_1,\ R_2\mathcal{G}^P + t_2\big) = s\big(\mathcal{G}^T, \mathcal{G}^P\big).$$

Proof. See Appendix B.2. □

EMBER therefore never takes in the relative pose, so AIC cannot mislead the result.

5 Experiments

We evaluate EMBER against six published baselines on a curated set of four public datasets, over five epitope-disjoint folds and three seeds.

Dataset. We utilize four public datasets—McPAS-TCR, VDJdb, IEDB, and 10x Genomics—covering 45,909 TCR–pMHC pairs across 343 epitopes (Vita et al., 2015; Tickotsky et al., 2017; Shugay et al.,

2018; 10x Genomics, 2019). Following Slone et al. (2025a), 7,371 measured binders are augmented with sampled non-binders, yielding a binder prevalence of 0.161.

Baselines. To evaluate EMBER against diverse TCR–pMHC predictors, we compare with six baselines spanning three structural, one hybrid, and two sequence-based methods. Structural baselines include STAG-LLM (Slone et al., 2025b), SageTCR (Li et al., 2025), and the dual-encoder classifier of Abeer et al. (2026); TRAP (Ge et al., 2025) combines pMHC structure with TCR sequence, whereas NetTCR-2.2 (Jensen & Nielsen, 2024) and ERGO-II (Springer et al., 2021) read no structure. We retrain each method using its published hyperparameters across five folds and three random seeds, reporting means over all 15 runs and standard deviations across seeds.

Metrics. To evaluate the performance of EMBER, we measure ranking quality, classification accuracy, and cross-tool stability. Ranking is assessed by micro- and macro-averaged AUROC and AUPRC; classification uses MCC at a fixed threshold of 0.5, reported alongside MCC at the threshold that maximizes it on the test fold, which no deployed system has and which we report only to bound what threshold tuning could buy a baseline. Stability is quantified by pairwise Kendall's τ_b, Kendall's concordance W across all four folding tools, and Overlap@k between top-ranked candidates. Appendix F.2 provides the definitions.

Table 1: Binding prediction under epitope-disjoint five-fold cross-validation, means over five folds and three seeds, with the standard deviation over seeds. *structures* names the tool that produced the input, and within each group the best value per column is **bold** and the second best is underlined. The sequence baselines read no structure and are not ranked. EMBER is first in every cell, ahead of the second by **6.1 to 21.7%** in macro-AUROC, **13.6 to 22.0%** in macro-AUPRC, **0.9 to 14.4%** in micro-AUROC, **18.5 to 25.3%** in micro-AUPRC, **51.2 to 74.5%** in MCC@0.5 and **10.6 to 24.1%** in MCC*. MCC* is MCC at the threshold that maximizes it on the test fold. Appendix F.4 gives the per-fold values.

family	structures	method	macro-AUROC	macro-AUPRC	micro-AUROC	micro-AUPRC	MCC@0.5	MCC*
structure-based	tFold-TCR	**EMBER**	**0.758** ± 0.010	**0.585** ± 0.008	**0.666** ± 0.003	**0.366** ± 0.014	**0.251** ± 0.059	**0.292** ± 0.017
		STAG-LLM	0.620 ± 0.018	0.473 ± 0.020	0.582 ± 0.008	0.292 ± 0.006	0.166 ± 0.024	0.264 ± 0.018
		SageTCR	0.622 ± 0.009	0.443 ± 0.004	0.547 ± 0.011	0.240 ± 0.007	0.122 ± 0.027	0.182 ± 0.036
		Abeer	0.591 ± 0.008	0.434 ± 0.006	0.550 ± 0.011	0.225 ± 0.014	0.071 ± 0.007	0.157 ± 0.024
		TRAP	0.651 ± 0.015	0.513 ± 0.014	0.499 ± 0.010	0.182 ± 0.010	0.034 ± 0.000	0.095 ± 0.024
	ESMFold2	**EMBER**	**0.759** ± 0.004	**0.584** ± 0.003	**0.670** ± 0.005	**0.372** ± 0.013	**0.273** ± 0.029	**0.304** ± 0.011
		STAG-LLM	0.714 ± 0.013	0.507 ± 0.016	0.659 ± 0.003	0.312 ± 0.007	0.172 ± 0.001	0.245 ± 0.004
		SageTCR	0.618 ± 0.002	0.437 ± 0.014	0.545 ± 0.015	0.227 ± 0.003	0.115 ± 0.016	0.174 ± 0.030
		Abeer	0.590 ± 0.003	0.431 ± 0.002	0.579 ± 0.008	0.257 ± 0.028	0.108 ± 0.049	0.188 ± 0.025
		TRAP	0.649 ± 0.006	0.514 ± 0.012	0.497 ± 0.025	0.180 ± 0.003	0.041 ± 0.017	0.096 ± 0.008
	TCRmodel2	**EMBER**	**0.769** ± 0.007	**0.593** ± 0.004	**0.668** ± 0.003	**0.367** ± 0.002	**0.267** ± 0.009	**0.301** ± 0.003
		STAG-LLM	0.632 ± 0.009	0.477 ± 0.005	0.586 ± 0.008	0.295 ± 0.018	0.153 ± 0.029	0.264 ± 0.019
		SageTCR	0.596 ± 0.028	0.418 ± 0.011	0.537 ± 0.012	0.235 ± 0.004	0.112 ± 0.018	0.174 ± 0.011
		Abeer	0.583 ± 0.023	0.434 ± 0.017	0.597 ± 0.006	0.291 ± 0.021	0.145 ± 0.024	0.236 ± 0.019
		TRAP	0.630 ± 0.008	0.486 ± 0.011	0.487 ± 0.009	0.172 ± 0.004	0.019 ± 0.006	0.077 ± 0.006
	Boltz-2	**EMBER**	**0.753** ± 0.004	**0.579** ± 0.007	**0.661** ± 0.006	**0.372** ± 0.003	**0.278** ± 0.013	**0.311** ± 0.012
		STAG-LLM	0.710 ± 0.002	0.505 ± 0.003	0.655 ± 0.008	0.314 ± 0.026	0.169 ± 0.012	0.254 ± 0.020
		SageTCR	0.597 ± 0.001	0.422 ± 0.013	0.527 ± 0.012	0.224 ± 0.009	0.115 ± 0.003	0.146 ± 0.004
		Abeer	0.600 ± 0.003	0.437 ± 0.009	0.580 ± 0.014	0.255 ± 0.032	0.054 ± 0.026	0.188 ± 0.023
		TRAP	0.642 ± 0.015	0.502 ± 0.006	0.491 ± 0.011	0.175 ± 0.002	0.016 ± 0.013	0.082 ± 0.009
sequence-based	-	NetTCR-2.2	0.688 ± 0.007	0.536 ± 0.012	0.604 ± 0.010	0.319 ± 0.009	0.207 ± 0.026	0.257 ± 0.006
	-	ERGO-II	0.668 ± 0.016	0.502 ± 0.016	0.593 ± 0.009	0.284 ± 0.010	0.061 ± 0.026	0.224 ± 0.012

5.1 Binding Prediction and Stability against Retrained Baselines

We conduct experiments on five epitope-disjoint folds, ensuring that all test peptides are unseen during training. We run the comparison on structures generated by four different tools, retraining every method on that tool's structures (Table 1 and Figure 4). A structure-based predictor inherits its upstream tool, and the folding tools disagree, so we also measure how far that choice moves the answer. For each epitope and fold we rank that epitope's candidate TCRs with the same method trained on each tool's structures, then compare the rankings (Table 2).

Results. EMBER is first on every metric under every folding tool, and the size of the lead depends on which part of the ranking a metric reads. AUROC weighs the whole candidate list, and there the strongest baseline comes close. STAG-LLM trails EMBER by 0.011 in micro-AUROC under ESMFold2 and by 0.006 under Boltz-2, so the whole-list metric barely separates the two. Metrics that weigh the top of the list give a different reading. On micro-AUPRC the same baseline trails by 0.058 to 0.074, and on MCC at a fixed threshold of 0.5 it trails by 51 to 75%. Calibration accounts for part of that last figure, since tuning the threshold on the test fold narrows the lead to 10 to 24%.

What remains after the tuning is discrimination, and a deployed screen meets both terms at once, because it reads the top of the list at a threshold fixed in advance. Read one fold at a time, the lead holds on all five folds in 16 of the 24 cells, and the eight that fall short are two micro-AUROC cells, two micro-AUPRC cells and all four MCC* cells (Table 15).

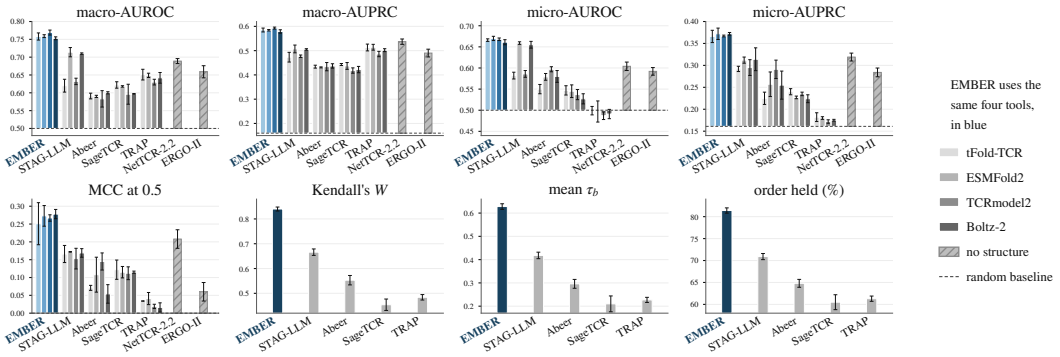


Figure 4: Means over five folds and three seeds, with every bar starting at chance. Structural methods carry one bar per folding tool, light to dark in the order tFold-TCR, ESMFold2, TCRmodel2, Boltz-2. A method whose four bars are uneven depends on the upstream tool. The last three panels read the same rankings three ways, over the four tools of Table 2. Kendall's W carries bootstrapped intervals.

Table 2: Cross-tool agreement of methods trained on structures from four folding tools. Agreement is evaluated over 142 epitopes with at least ten candidates each, comprising 44.6k candidates in total.

method	W	mean τ_b	order held	Overlap@10
EMBER	**0.840** $\pm$ 0.008	**0.628** $\pm$ 0.012	**81.4%** $\pm$ 0.6	**0.729** $\pm$ 0.013
STAG-LLM	0.666 $\pm$ 0.013	0.418 $\pm$ 0.014	70.9% $\pm$ 0.7	0.675 $\pm$ 0.008
Abeer	0.553 $\pm$ 0.019	0.296 $\pm$ 0.019	64.8% $\pm$ 0.9	0.599 $\pm$ 0.001
TRAP	0.484 $\pm$ 0.011	0.227 $\pm$ 0.011	61.3% $\pm$ 0.6	0.515 $\pm$ 0.010
SageTCR	0.454 $\pm$ 0.023	0.210 $\pm$ 0.034	60.5% $\pm$ 1.7	0.573 $\pm$ 0.046

EMBER is the most stable method by a wide margin. Its concordance is $W = 0.840$, against 0.454 to 0.666 for the four structure-consuming baselines, and that is **26% to 85% higher agreement** than any of them. Read pairwise, 81.4% of candidate pairs keep their relative order when the folding tool is swapped, against 60.5% to 70.9% for the baselines. The spread across seeds says what the comparison can resolve. EMBER's W varies by 0.008 across seeds and the closest baseline by 0.013, against a separation of 0.174. The least stable baseline varies most, up to 0.023 in W and 0.034 in τ_b, so an experiment expecting an effect that size needs more seeds rather than more folds.

Finding 2. EMBER leads every metric under every folding tool, most clearly where every epitope counts equally and at the threshold a deployed system runs at. Its ranking is also the least disturbed by a change of folding tool, by a margin far outside the spread across training seeds.

5.2 Influence of Structure

We conduct three analyses to assess how structural information influences prediction. First, we directly use each folding tool's interface confidence as a training-free binding score. Second, we stratify co-folded pairs into confidence percentile bands and evaluate each method at its own operating threshold (Figure 5); because confidence is available only for co-folded complexes, only the three complex-based methods see the structure a band scores. For each band, we report the normalized positive likelihood ratio, $\mathrm{LR}^+ = \mathrm{TPR}/\mathrm{FPR}$, relative to the method's lowest-confidence band, thereby removing differences in operating point. Finally, we ablate EMBER to quantify the contribution of structural information.

Results. The complex carries evidence that two monomers do not. Within one pMHC, the interface confidence separates a true binder from a decoy at Cohen's d of 0.20 to 0.60. The independently folded receptor reaches only 0.05 to 0.12 on the same pairs. Read untrained as a score, the same confidence beats all six retrained baselines under TCRmodel2. No method that reads the complex turns this evidence into discrimination. Every method gains from the least to the most confident band. The two sequence baselines see no structure, so their gain measures what the pair alone is worth. The three complex consumers do not exceed that level, averaging 16% against 17%. The metric itself is

not saturated, since EMBER reaches 56% without reading the complex. Nor are the flat curves an artefact of weak classifiers, since STAG-LLM has the second highest micro-AUROC of the seven methods and the second lowest gain.

Removing the structure entirely from EMBER costs 11.1% of that margin in micro-AUPRC. Replacing the two monomers with the predicted complex costs 12.3%, and it costs more still at a fixed threshold. A docked assembly is therefore worth no more to EMBER than no structure at all. Removing the canonical peptide frame causes MCC@0.5 to fall by 23.4%, and the fall is 8.3% once the threshold is tuned on the test fold. So, the canonical frame improves calibration and per-epitope precision.

Table 3: Main ablation, on the ESMFold2 structures. *remove structure* keeps the residue identities and the original ESM-C output. *docked complex* feeds EMBER the co-folded complex. *no peptide frame* removes the canonical peptide frame. The arms run on ESMFold2 because it gives the most favourable input, leading every interface and placement metric on the post-cutoff targets of Table 9. Appendix F.5 gives two further ablations.

arm	macro-AUROC	macro-AUPRC	micro-AUROC	micro-AUPRC	MCC@0.5	MCC*
EMBER (control)	0.759 ± 0.004	0.584 ± 0.003	0.670 ± 0.005	0.372 ± 0.013	0.273 ± 0.029	0.304 ± 0.011
remove structure	0.753 ± 0.008	0.570 ± 0.006	0.649 ± 0.003	0.349 ± 0.001	0.236 ± 0.018	0.283 ± 0.010
docked complex	0.745 ± 0.009	0.578 ± 0.013	0.654 ± 0.009	0.346 ± 0.033	0.232 ± 0.063	0.274 ± 0.027
no peptide frame	0.740 ± 0.008	0.558 ± 0.004	0.664 ± 0.007	0.360 ± 0.004	0.209 ± 0.017	0.279 ± 0.009

Finding 3. Co-folding encodes binding evidence in its own confidence, yet no complex consumer converts it into discrimination, and supplying the complex to EMBER costs accuracy. The structure contributes a geometric scaffold, so two unbound monomers give up nothing measurable.

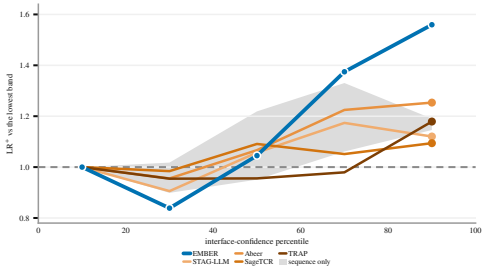


Figure 5: How LR^+ varies with the structure confidence level. The gray band spans the two sequence baselines. The first two bands hold the same share of binders, 0.112 each, so the step between them carries no signal.

Table 4: Preprocessing cost on a single A100-SXM4-80GB GPU. All tools were benchmarked on a fixed pool of 50 TCR–pMHC pairs, 50 TCRs and 20 pMHCs. TCRmodel2 uses 416-residue complexes rather than 610-residue inputs, making its reported savings not directly comparable. Times include inference only. Appendix F.6 gives the per-tool caveats.

	one structure (s)			this evaluation set (GPU-h)		
tool	TCR 228 aa	pMHC 382 aa	complex 610 aa	EMBER 21,830 monomers	complex 45,909 pairs	saving
tFold-TCR	**0.672**	**1.618**	**6.329**	**4.2**	80.7	19.4×
ESMFold2	5.190	10.522	25.629	32.0	326.8	10.2×
Boltz-2	10.92	17.50	23.784	66.9	303.3	4.5×
TCRmodel2	102.8	111.5	125.6	624.1	1,602.3	2.6×

Preprocessing Cost. EMBER has a preprocessing cost of $\mathcal{O}(N+M)$. All four tools mentioned ran on one A100-SXM4-80GB over one fixed set of samples, and Table 4 gives the result. The evaluation set contains 21,477 TCR, 353 pMHC, and 45,909 candidate pairs. EMBER reduces preprocessing cost by reusing each monomer structure across the pairs it appears in. Folding every monomer with tFold-TCR costs 4.2 GPU-h. The same set cost 80.7, 303.3, 326.8 and 1,602.3 GPU-h as complexes under tFold-TCR, Boltz-2, ESMFold2 and TCRmodel2, a saving of up to 19.4× against complexes.

6 Conclusion

We propose **EMBER**, a pose-independent framework for TCR–pMHC binding prediction that avoids **assembly-induced interface corruption (AIC)**, a failure mode in which complex-placement errors alter the residues and contacts presented to downstream predictors. Testing across 49 crystal structures on six folding tools, we show that such placement errors are substantial. By encoding each molecule independently, EMBER removes relative pose from the binding score and is first on every metric under all four folding tools, improving MCC at the fixed threshold of 0.5 by **51 to 75%** over the strongest baseline and ranking stability across folding tools by **26 to 85%** over structural baselines. It also cuts preprocessing by up to **19.4**× on the same evaluation set. These results suggest a different way of using predicted structure for the TCR–pMHC prediction task. EMBER uses only the trustworthy part of TCR and pMHC structures, and it is both more accurate and more stable for it.

Ethics Statement

This work uses only publicly released data. The binding pairs come from McPAS-TCR, VDJdb, IEDB and the 10x Genomics public datasets. The records are molecular measurements rather than identifiable donor information, and no new human or animal data was collected. The AIC benchmark uses deposited PDB entries only.

Two limitations bear on how the predictions should be used. First, the public repertoires are dominated by a small number of HLA alleles and a long tail of epitopes, so accuracy on an allele or a peptide far from that distribution is not established by our results. Reporting macro alongside micro metrics makes the imbalance visible rather than removing it. Second, EMBER is a screening model that narrows a candidate list. A predicted binder is a hypothesis for experimental testing, and none of our evaluations speaks to safety, cross-reactivity or clinical suitability.

TCR–pMHC prediction supports immunotherapy and vaccine design, and the same capability could in principle inform the design of epitopes that evade recognition. The barrier to that use is experimental rather than computational, the model is trained on public data with published baselines, and we judge the benefit to open immunological research to outweigh this risk.

Reproducibility Statement

Every component needed to reproduce the results is specified in the paper. The method section defines the architecture, Algorithm 1 states the forward pass, and Table 5 lists every symbol with the value it takes in the reported configuration. Appendix C gives graph construction and the node and edge features, and Appendix F.3 gives the hyperparameters and the optimiser schedule.

Appendix F.1 specifies the evaluation protocol in full, covering the composition of the curated set, the five epitope-disjoint folds and their sizes, the train, validation and test rotation, the checkpoint-selection rule, the construction of the non-binders and the row coverage of each method. Appendix F.2 defines each metric. Every number in the paper is a mean over five folds and three training seeds, with the standard deviation over seeds reported alongside, and Appendix F.4 gives the per-fold values behind those means. Baselines are retrained under this protocol with their own published hyperparameters. Appendix F.6 reports the hardware and the measured preprocessing cost for each folding tool. Appendix B proves Proposition 1, and Appendix D states which parts of the design the released configuration exercises.

Source code, configuration files and the scripts are released at `https://anonymous.4open.science/r/ember-6BCA/`.

AI Use Statement

In our paper and project, we use large language models (LLMs) to improve the conciseness and precision of the writing. We also use LLMs to assist with intermediate steps in theoretical derivations and to generate code for figure visualization. All technical claims, proofs, experiments, and final results are independently checked and validated by the authors.

References

10x Genomics. A new way of exploring immunity: linking highly multiplexed antigen recognition to immune repertoire and phenotype. Technical report, 10x Genomics, 2019.

ANM Nafiz Abeer, Raj S Roy, Xiaoning Qian, and Byung-Jun Yoon. Structure-based tcr-pmhc binding prediction and generalization to unseen peptides. *npj Drug Discovery*, 3(1):31, 2026.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630:493–500, 2024.

Aureka AI OpenDDE Project. Folding, reasoning, and scaling with an open-source drug discovery engine. *arXiv preprint arXiv:2607.03787*, 2026. OpenDDE.

Philip Bradley. Structure-based prediction of T cell receptor:peptide-MHC interactions. *eLife*, 12: e82813, 2023.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Daniel S. Liu, Tom Sercu, Alexander Rives, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. ESMFold2 and ESM-C.

Sebastian N. Deleuran and Morten Nielsen. NetTCR-struc, a structure driven approach for prediction of TCR-pMHC interactions. *Frontiers in Immunology*, 16:1616328, 2025.

Yang Deng, Jinhao Que, Guangfu Xue, Yideng Cai, Wenyi Yang, Yilin Wang, Yi Hui, Zuxiang Wang, Yi Lin, Wenyang Zhou, Zhaochun Xu, Qinghua Jiang, and Haoxiu Sun. AI-driven computational methods and benchmarking for T-cell antigen identification. *Briefings in Bioinformatics*, 27(2): bbag123, 2026.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.

K. Christopher Garcia and Erin J. Adams. How the T cell receptor sees antigen—a structural view. *Cell*, 122(3):333–336, 2005.

Jingxuan Ge, Jike Wang, Qing Ye, Liqiang Pan, Yu Kang, Chao Shen, Yafeng Deng, Chang-Yu Hsieh, and Tingjun Hou. TRAP: a contrastive learning-enhanced framework for robust TCR–pMHC binding prediction with improved generalizability. *Chemical Science*, 16:9881–9894, 2025.

Mathias Fynbo Jensen and Morten Nielsen. Enhancing TCR specificity predictions by combined pan- and peptide-specific training, loss-scaling, and sequence similarity integration. *eLife*, 12:RP93934, 2024.

Hongchen Ji, Xiang-Xu Wang, Qiong Zhang, Chengkai Zhang, and Hong-Mei Zhang. Predicting TCR sequences for unseen antigen epitopes using structural and sequence features. *Briefings in Bioinformatics*, 25(3):bbae210, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

M. G. Kendall and B. Babington Smith. The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3):275–287, 1939.

Nicole L. La Gruta, Stephanie Gras, Stephen R. Daley, Paul G. Thomas, and Jamie Rossjohn. Understanding the drivers of MHC restriction of T cell receptors. *Nature Reviews Immunology*, 18 (7):467–478, 2018.

Jinwoo Leem, Saulo H. P. de Oliveira, Konrad Krawczyk, and Charlotte M. Deane. STCRDab: the structural T-cell receptor database. *Nucleic Acids Research*, 46(D1):D406–D412, 2018.

Jiarui Li, Zixiang Yin, Yunbei Zhang, Janet Wang, Samuel J. Landry, Zhengming Ding, and Ramgopal R. Mettu. Structure-regularized interpretable TCR-epitope prediction. *arXiv preprint arXiv:2606.30902*, 2026.

Xiangyi Li, Chuance Sun, Weiran Huang, Yanjing Wang, and Buyong Ma. SageTCR: a structure-based model integrating residue- and atom-level representations for enhanced TCR-pMHC binding prediction. *Briefings in Bioinformatics*, 26(5):bbaf496, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yanping Lu, Yuyan Wang, Meng Xu, Bingbing Xie, Yumeng Yang, Haodong Xu, and Shengbao Suo. Assessment of computational methods in predicting TCR–epitope binding recognition. *Nature Methods*, 23:248–259, 2026.

Dean R. Madden. The three-dimensional structure of peptide–MHC complexes. *Annual Review of Immunology*, 13:587–622, 1995.

Raúl Méndez, Raphaël Leplae, Marc F. Lensink, and Shoshana J. Wodak. Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. *Proteins: Structure, Function, and Bioinformatics*, 60(2):150–169, 2005.

Alessandro Montemurro, Viktoria Schuster, Helle Rus Povlsen, Amalie Kai Bentzen, Vanessa Jurtz, William D. Chronister, Austin Crinklaw, Sine R. Hadrup, Ole Winther, Bjoern Peters, Leon Eyrich Jessen, and Morten Nielsen. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCR α and β sequence data. *Communications Biology*, 4:1060, 2021.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, et al. Boltz-2: towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025.

Xingang Peng, Yipin Lei, Peiyuan Feng, Lemei Jia, Jianzhu Ma, Dan Zhao, and Jianyang Zeng. Characterizing the interaction conformation between T-cell receptors and epitopes with deep learning. *Nature Machine Intelligence*, 5:395–407, 2023.

Mark Polster, Josua Stadelmaier, Elias Ball, Jonas Scheid, Jens Bauer, Annika Nelde, Manfred Claassen, Marissa Dubbelaar, Juliane S. Walz, and Sven Nahnsen. t2pmhc: A structure-informed graph neural network to predict TCR-pMHC binding. *bioRxiv*, 2026.

Michael Robben. AI predicted TCR-pMHC structures differentiate immune interactions. *bioRxiv*, 2026.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 9323–9332, 2021.

Chun Shang, Kevin C. Chan, and Ruhong Zhou. Decoding TCR recognition via geometric deep learning of immunological fingerprints. *Briefings in Bioinformatics*, 27(2):bbag048, 2026.

Mikhail Shugay, Dmitriy V. Bagaev, Ivan V. Zvyagin, Renske M. Vroomans, Jeremy Chase Crawford, Garry Dolton, Ekaterina A. Komech, Anastasiya L. Sycheva, Anna E. Koneva, Evgeniy S. Egorov, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, 46(D1):D419–D427, 2018.

Paopit Siriarchawatana, Supawadee Ingsriswang, Challika Kaewborisuth, and Anan Jongkaewwattana. TCRLens: structure-aware equivariant graph learning for TCR-pMHC-I recognition and immunogenic epitope discovery. *Bioinformatics Advances*, 6:vbag066, 2026.

Jared K. Slone, Anja Conev, Mauricio M. Rigo, Alexandre Reuben, and Lydia E. Kavraki. TCR-pMHC binding specificity prediction from structure using graph neural networks. *IEEE Transactions on Computational Biology and Bioinformatics*, 22(1):171–178, 2025a.

Jared K. Slone, Minying Zhang, Peixin Jiang, Amanda Montoya, Emily Bontekoe, Barbara Nassif Rausseo, Alexandre Reuben, and Lydia E. Kavraki. STAG-LLM: Predicting TCR-pHLA binding with protein language models and computationally generated 3D structures. *Computational and Structural Biotechnology Journal*, 27:3885–3896, 2025b.

Ido Springer, Nili Tickotsky, and Yoram Louzoun. Contribution of T cell receptor alpha and beta CDR3, MHC typing, V and J genes to peptide binding prediction. *Frontiers in Immunology*, 12: 664514, 2021.

Christopher Szeto, Christian A. Lobos, Andrea T. Nguyen, and Stephanie Gras. TCR recognition of peptide–MHC-I: Rule makers and breakers. *International Journal of Molecular Sciences*, 22(1): 68, 2021.

Nili Tickotsky, Tal Sagiv, Jaime Prilusky, Eric Shifrut, and Nir Friedman. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*, 33(18): 2924–2929, 2017.

Fleur E. Tynan, Scott R. Burrows, Ashley M. Buckle, Craig S. Clements, Natalie A. Borg, John J. Miles, Travis Beddoe, James C. Whisstock, Matthew C. Wilce, Sharon L. Silins, Jacqueline M. Burrows, Lars Kjer-Nielsen, Lyudmila Kostenko, Anthony W. Purcell, James McCluskey, and Jamie Rossjohn. T cell receptor recognition of a 'super-bulged' major histocompatibility complex class I-bound peptide. *Nature Immunology*, 6(11):1114–1122, 2005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Randi Vita, James A. Overton, Jason A. Greenbaum, Julia Ponomarenko, Jason D. Clark, Jason R. Cantrell, Daniel K. Wheeler, Jeffrey L. Gabbard, Deborah Hix, Alessandro Sette, and Bjoern Peters. The immune epitope database (IEDB) 3.0. *Nucleic Acids Research*, 43(D1):D405–D412, 2015.

Fandi Wu, Yu Zhao, Yang Xiao, Chenchen Qin, Fang Wang, Zihan Wu, Long-Kai Huang, Xiao Liu, Jiangning Song, Bing He, Jamie Rossjohn, and Jianhua Yao. Fast and accurate modeling of TCR-peptide-MHC complexes using tFold-TCR. *bioRxiv*, 2025.

Rui Yin, Helder V. Ribeiro-Filho, Valerie Lin, Ragul Gowthaman, Melyssa Cheung, and Brian G. Pierce. TCRmodel2: high-resolution modeling of T cell receptor recognition using deep learning. *Nucleic Acids Research*, 51(W1):W569–W576, 2023.

Zilan Yu, Mengnan Jiang, and Xun Lan. HeteroTCR: A heterogeneous graph neural network-based method for predicting peptide-TCR interaction. *Communications Biology*, 7:684, 2024.

Yu Zhao, Bing He, Fan Xu, Chen Li, Zhimeng Xu, Xiaona Su, Haohuai He, Yueshan Huang, Jamie Rossjohn, Jiangning Song, and Jianhua Yao. DeepAIR: A deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis. *Science Advances*, 9(32):eabo5128, 2023.

Supplementary Material

A Notation **14**

B Proof of Main Text **16**
B.1 Lemma 4.1 16
B.2 Proposition 1 16

C Graph Construction and Features **16**

D Scope of the Released Model **18**

E AIC Observation Details **18**

F Experiment Results **19**
F.1 Evaluation Protocol 19
F.2 Metric Definitions 21
F.3 Hyperparameter 22
F.4 Binding Prediction, Fold by Fold 23
F.5 Architecture Ablations 28
F.6 Preprocessing Benchmark 28

A Notation

Every symbol used by EMBER, with the value it takes in the reported configuration. Indices i, j always run over residues of a single monomer. No symbol in the table ranges over both molecules at once, which is the notational form of Proposition 1. Calligraphic letters denote sets and graphs, uppercase Latin letters denote matrices and whole-molecule quantities, lowercase Latin letters denote per-residue vectors and scalars, and Greek letters denote learned networks and fixed nonlinearities. Symbols follow that order within each block.

Table 5: Notation for EMBER. Shapes are per residue unless stated otherwise.

symbol	meaning	shape or value
graphs and inputs		
$\mathcal{G}^T, \mathcal{G}^P$	residue graphs of the receptor and of the antigen, built from separately folded monomers and never merged	—
$\mathcal{N}(i)$	the k nearest neighbors of i at layer ℓ	$\lvert\mathcal{N}(i)\rvert = k$
N_T, N_P	residues in $\mathcal{G}^T$ (β chain) and in $\mathcal{G}^P$	$\approx 114, \approx 382$
W_e	learned bottleneck that compresses e_i	$\mathbb{R}^{128\times1152}$
i, j	residue indices within one graph	—
p_i	Cα coordinate of residue i, in that monomer's own frame	$\mathbb{R}^3$
x_i	invariant node descriptor: curvature ratios, dihedrals, solvent-accessible area, neighbor count, pLDDT, residue one-hot, secondary-structure one-hot	$\mathbb{R}^{37}$
e_i	frozen per-residue ESM-C embedding, cached once	$\mathbb{R}^{1152}$
$h_i^{(0)}$	encoder input, $[\, x_i \,\|\, W_e e_i \,]$	$\mathbb{R}^{165}$
k	neighbors per node	30
d_{ij}	Cα distance $\lVert p_i - p_j \rVert$	Å
b_{ij}	biochemical compatibility from residue identity alone: hydrogen bonding, hydrophobic product, van der Waals ratio, charge product, aromatic flag	$\mathbb{R}^5$
a_{ij}	edge feature, $[\, \rho(d_{ij}) \,\|\, d_0/d_{ij} \,\|\, b_{ij} \,]$, $d_0 = 4$ Å	$\mathbb{R}^{22}$
ρ	radial basis expansion of d_{ij} over $[0, 20]$ Å	$\mathbb{R}^{16}$
monomer encoder		
L	encoder layers	2

Table 5 continued

symbol	meaning	shape or value
D	encoder width	256
H^T, H^P	final states $h^{(L)}$ stacked by residue	$\mathbb{R}^{N_T \times D}, \mathbb{R}^{N_P \times D}$
$h_i^{(\ell)}$	invariant node state after layer ℓ; $h_i^{(0)}$ is projected to width D before layer 1	$\mathbb{R}^D$
$p_i^{(\ell)}$	coordinate after layer ℓ; $p_i^{(0)} = p_i$	$\mathbb{R}^3$
c_{ij}	edge context at layer ℓ, $[\, h_i^{(\ell-1)} \,\|\| \, h_j^{(\ell-1)} \,\|\, a_{ij} \,]$	$\mathbb{R}^{2D+22}$
m_{ij}	message carried on edge (i, j)	$\mathbb{R}^D$
g_{ij}	chemistry gate $\sigma(\phi_g(b_{ij}))$ scaling that message	$(0, 1)$
θ	encoder parameters, shared by both graphs	—
ℓ	layer index, $\ell = 1, \ldots, L$	—
ϕ_m	message network, two hidden layers with SiLU	$\mathbb{R}^{2D+22} \to \mathbb{R}^D$
ϕ_g	gate network on b_{ij}, two hidden layers with SiLU	$\mathbb{R}^5 \to \mathbb{R}$
ϕ_x	coordinate-weight network, $\tanh$-bounded output	$\mathbb{R}^{2D+22} \to (-1, 1)$
σ	logistic sigmoid	—
LN	layer normalization over the feature axis	—
contact head and objective		
$\mathcal{B}$	a training batch of pairs, $\mathcal{B} = \mathcal{B}^+ \cup \mathcal{B}^-$	—
$\mathcal{B}^+, \mathcal{B}^-$	the binders and the non-binders of $\mathcal{B}$	$\lvert\mathcal{B}^-\rvert = 3\lvert\mathcal{B}^+\rvert$
$\mathcal{L}$	objective, $\mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{BT}}$	—
C	rows of H^T holding CDR3β residues, fixed by sequence annotation	—
E	rows of H^P holding peptide residues, fixed by sequence annotation	—
W_z	projection of the concatenated pair	$\mathbb{R}^{128 \times 256}$
S	slots of the canonical peptide frame	9
d	width of one attention head, four heads in all	32
z_t	direction 1 output: CDR3 rows query the peptide	$\mathbb{R}^{128}$
z_p	direction 2 output: peptide rows query the CDR3 loops	$\mathbb{R}^{128}$
z	pair embedding, $\mathrm{LN}(W_z[\, z_t \| z_p \,])$	$\mathbb{R}^{128}$
w	classifier weight	$\mathbb{R}^{128}$
u	pair logit, $w^\top z$	$\mathbb{R}$
$\hat{y}$	predicted binding probability, $\sigma(u)$	$(0, 1)$
y	binary label, 1 for a binder and 0 otherwise	$\{0, 1\}$
π	monotone map from peptide residues onto the canonical slots	$\{1..\lvert E\rvert\} \to \{1..S\}$
$\mathrm{MHA}(X, Y)$	multi-head attention, queries from X, keys and values from Y	—
pool_C	mask-aware mean and max over the rows of C, concatenated	$\mathbb{R}^{256}$
pool_E	per-slot mean and max under π, concatenated over the S slots	$\mathbb{R}^{2S \cdot 128}$
proj_1	linear map onto the contact dimension	$\mathbb{R}^{256} \to \mathbb{R}^{128}$
proj_2	linear map onto the contact dimension	$\mathbb{R}^{2304} \to \mathbb{R}^{128}$
pose independence		
R_1, R_2	rotations applied independently to the two molecules	$SO(3)$
t_1, t_2	translations applied independently to the two molecules	$\mathbb{R}^3$
$g = (R, t)$	one rigid motion; g_1 acts on the receptor and g_2 on the antigen	—
s	the score as a function of the two graphs, $s(\mathcal{G}^T, \mathcal{G}^P) = \hat{y}$	$(0, 1)$
N, M	distinct receptors and targets in a screen, not residue counts	—

B Proof of Main Text

Write (R_1, t_1) for the motion applied to the receptor and (R_2, t_2) for the one applied to the antigen. The two are unrelated, which is what separates Proposition 1 from invariance to a single motion of a joint structure.

B.1 Lemma 4.1

Proof of Lemma 4.1. In Equation (4.1) ϕ_m receives only the edge context c_{ij} and ϕ_g only b_{ij}. The hypothesis leaves $h^{(\ell-1)}$ unchanged, and a_{ij}, rebuilt from $p^{(\ell-1)}$, is unchanged by the argument of Definition 4.1 applied to $p^{(\ell-1)}$, so m_{ij} and the gate g_{ij} are unchanged. Their gated sum followed by LN gives $h^{(\ell)}$ unchanged. Writing $q_i = p_i^{(\ell-1)}$ and substituting $q \mapsto Rq + t$ into the coordinate update of Equation (4.2),

$$(Rq_i+t)+\frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)}\big(Rq_i-Rq_j\big)\,\phi_x(\cdot) = R\Big(q_i+\frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)}(q_i-q_j)\,\phi_x(\cdot)\Big)+t = Rp_i^{(\ell)}+t,$$

where ϕ_x is unchanged because its arguments are, and the translation cancels inside each difference. The edge set rebuilt from $p^{(\ell)}$ is the same by the argument of Definition 4.1.

B.2 Proposition 1

Proof of Proposition 1. Apply Definition 4.1 to each graph, then Lemma 4.1 by induction over the L layers. So H^T is unchanged by (R_1, t_1) and H^P by (R_2, t_2). The two towers share weights but exchange no message, so neither output depends on the other graph's frame and the two statements hold simultaneously. The contact head is a function of H^T, H^P and the row masks selecting C and E. It admits no position argument, and the masks are index sets fixed by sequence, so the score depends on p only through H. Hence $s\big(R_1\mathcal{G}^T + t_1,\ R_2\mathcal{G}^P + t_2\big) = s\big(\mathcal{G}^T, \mathcal{G}^P\big)$.

Why the independence matters. The guarantee comes from the head's signature rather than from its weights, so it holds at initialization and cannot be traded away by training. The model has no relative-pose input, so there is nothing for training to ignore. An equivariant network applied to a docked complex, as in Siriarchawatana et al. (2026), gives only the weaker statement that the score is unchanged under a single motion of the whole assembly. That quotients out the choice of laboratory frame but keeps the relative pose, which is exactly the quantity we measure to be unreliable. Proposition 1 quotients out the relative pose itself, because no operation in the model could read it.

C Graph Construction and Features

What is cached. One monomer is one graph, and a node is a residue represented by its Cα. The cache holds the node scalars $x \in \mathbb{R}^{N\times 37}$ of Table 6 and the coordinates $p \in \mathbb{R}^{N\times 3}$, and nothing else. Edges are not cached, because the encoder rebuilds the k-NN graph and every edge feature from the updated coordinates at each layer. The per-residue language-model embedding lives in a separate store and its bottleneck is a learned module rather than a preprocessing step, which is why the cached node vector is 37 wide rather than the 165 the encoder sees.

What makes the representation invariant. Every entry of Table 6 and every biochemical term of Table 7 is a function of residue identity or of relative geometry, so a rigid motion leaves all of them fixed by construction. Appendix B turns this into the statement proved for the binding score.

Table 6: The 37 node scalars. Every entry reads residue identity or relative geometry alone, so none of them introduces an absolute coordinate or a preferred axis. Dimensions 0–13 and 34–36 are the 17 structure-derived ones that the no-structure ablation zeroes, which leaves the residue one-hot standing.

dims	name	definition
0–1	curvature	two ratios of the eigenvalues $\lambda_0 \geq \lambda_1 \geq \lambda_2$ of the centered covariance of the 10 nearest Cα, namely $\lambda_2/(\lambda_0+\lambda_1+\lambda_2)$ and $\lambda_1\lambda_2/\lambda_0^2$
2	electrostatic	$\sum_j q_j/(1+d_j)$ over heavy atoms of other residues within 10 Å of the Cα, on a per-atom partial-charge table
3	SASA	Shrake–Rupley, summed over the atoms of the residue
4	coordination	heavy atoms of other residues within 8 Å
5–7	backbone dihedral	ϕ, ψ, ω in radians, undefined values set to 0
8–11	side-chain dihedral	χ_1 to χ_4 in radians, undefined values set to 0
12	density	heavy atoms of other residues within 5 Å
13	pLDDT	B-factor rescaled to $[0, 1]$, with the scale read from the file rather than assumed per tool
14–33	residue identity	20-way one-hot
34–36	secondary structure	3-way one-hot from P-SEA, helix, strand and coil

Table 7: The 22 invariant edge scalars and the five biochemical terms. Core sits before secondary on purpose, so the ablation that drops the secondary chemistry is a zeroing of the last two columns. The order is fixed by the cached graphs and cannot be permuted.

cols	name	definition
0–15	radial basis	16 Gaussians with centers spaced evenly over $[0, 20]$ Å and $\gamma \approx 5.63$
16	radial scalar	inverse distance d_0/d_{ij} with $d_0 = 4$ Å
17–19	core chemistry	hydrogen bonding, hydrophobic compatibility, van der Waals ratio
20–21	secondary chemistry	charge product, aromatic pairing
the five biochemical terms, each a function of the two residue identities alone		
hydrogen bonding		$(\mathrm{don}_i \wedge \mathrm{acc}_j) \vee (\mathrm{acc}_i \wedge \mathrm{don}_j)$
hydrophobic compatibility		$(\mathrm{kd}_i/4.5)\,(\mathrm{kd}_j/4.5)$, a product rather than a difference, so two hydrophobic ends give a positive value and a mismatch a negative one
van der Waals ratio		$\min(r_i, r_j)/\max(r_i, r_j)$ with $r = (3V/4\pi)^{1/3}$ on residue volumes
charge product		$q_i q_j$ on net residue charge at pH 7.4
aromatic pairing		$\mathrm{arom}_i \wedge \mathrm{arom}_j$ over PHE, TYR, TRP and HIS

Why the edge carries an inverse distance. The edge already holds a charge product $q_i q_j$ but no radial decay, and the Coulomb kernel $q_i q_j/d$ cannot be written as any one of the other terms. Left to itself the network has to synthesize $1/d$ out of 16 local Gaussian bumps, which are not monotone in d and represent it poorly. A single monotone radial term multiplied by the charge column is the Coulomb form exactly. Node dimension 2 of Table 6 already uses the same inverse-distance proxy, and this term is its edge-level counterpart. It is a function of $\|p_i - p_j\|$ alone, so it leaves the invariance argument untouched.

The hydrogen-bonding sets. Donors are ARG, ASN, GLN, HIS, LYS, SER, THR, TRP and TYR. Acceptors are ASN, ASP, GLN, GLU, HIS, SER, THR and TYR.

The invariance gate. A unit test applies a random rotation and translation to a structure and requires that the node scalars and the invariant edge scalars are unchanged elementwise and that the equivariant edge field rotates with the input. It runs before any bulk build and before any training run. One source of nondeterminism is known and tolerated. When the 30th and the 31st neighbor of a residue sit within 10^{-3} Å of each other, a rotation can swap them and change the edge set, which we measure on 0.58% of nodes for Boltz-2 structures and 0.86% for tFold-TCR and ESMFold2, so the gate is allowed to retry.

D SCOPE OF THE RELEASED MODEL

Receptor chains. The cached receptor graph carries both TCR chains. The encoder is given the β chain alone. Dropping the α chain removes its node block and rebuilds the k-NN edges, so nothing of that chain reaches the encoder. Appendix F.5 reports the arm that keeps it.

Paired α chains. A source database records an α chain together with its β chain for part of the set only. Matching each $(\mathrm{CDR3}\alpha, \mathrm{CDR3}\beta)$ pair against the paired records of VDJdb, McPAS-TCR and IEDB confirms $6{,}759$ of the $21{,}477$ receptors and $19{,}462$ of the $45{,}909$ rows. The rate is 70.7% on binders and 37.0% on non-binders. It is a lower bound, because our copy of the 10x Genomics block carries no per-receptor provenance.

E AIC OBSERVATION DETAILS

Targets. The 49 class-I complexes carry 35 distinct peptides and 31 distinct MHC heavy chains, so no single system dominates the set. Every tool folds the same five chains taken from the crystal entry, which makes residue correspondence with the native structure hold by construction and lets contacts be compared without any alignment step.

Leakage control. Templates are capped per target at the day before that target's deposition, so no tool can retrieve the answer it is being scored against. As a check that the cap is sufficient, splitting the 49 by deposition date leaves the possibly-seen half the *worse* half on docking ($p = 0.010$), the opposite of what memorization would produce.

Table 8: Interface accuracy of six co-folding tools on the 49 crystal structures, best per row in bold. Contacts are computed on raw coordinates without superposition, so they are what a downstream graph reads. Every RMSD is measured after superposing the prediction on the MHC heavy chain, except interface RMSD, which follows the CAPRI definition (Méndez et al., 2005) and re-superposes the interface locally. A target counts as well-docked when the receptor lands within 8 Å of its crystal position. Boltz-2 leads almost every row, and Table 9 shows why that is not a capability result.

	Boltz-2	ESMFold2	TCRmodel2	AF-Multimer	OpenDDE	tFold-TCR
interface contacts, median [mean]						
contact recall @4 Å ↑	**0.87 [0.73]**	0.52 [0.49]	0.41 [0.40]	0.38 [0.40]	0.39 [0.38]	0.26 [0.27]
contact recall @5 Å ↑	**0.92 [0.78]**	0.62 [0.58]	0.52 [0.51]	0.53 [0.48]	0.49 [0.44]	0.30 [0.33]
contact recall @8 Å ↑	**0.93 [0.84]**	0.71 [0.68]	0.68 [0.62]	0.64 [0.60]	0.65 [0.51]	0.50 [0.49]
contact precision @5 Å ↑	**0.89 [0.77]**	0.62 [0.59]	0.55 [0.49]	0.44 [0.45]	0.40 [0.43]	0.34 [0.36]
contact Jaccard @5 Å ↑	**0.83 [0.69]**	0.46 [0.44]	0.34 [0.36]	0.31 [0.33]	0.26 [0.33]	0.19 [0.22]
geometry (Å), median [mean]						
interface RMSD ↓	**0.43 [1.07]**	1.80 [2.31]	2.31 [3.68]	2.78 [4.02]	2.65 [5.89]	3.28 [4.21]
TCR docking RMSD ↓	**1.50 [2.80]**	6.96 [8.12]	5.87 [8.70]	8.09 [12.89]	6.93 [16.21]	7.14 [9.80]
peptide RMSD ↓	**0.43 [0.69]**	0.72 [1.01]	0.66 [0.99]	1.13 [1.70]	0.88 [1.11]	1.26 [1.84]
CDR3β loop RMSD ↓	**0.80 [1.68]**	2.56 [3.50]	3.95 [6.17]	4.13 [5.58]	4.24 [11.27]	4.27 [5.86]
CDR3α loop RMSD ↓	**0.87 [1.92]**	3.96 [4.45]	4.08 [5.60]	4.07 [5.71]	4.09 [11.55]	5.44 [6.29]
receptor placement, fraction of the 49 targets						
docking $<$ 2 Å (near-native) ↑	**59%**	6%	8%	0%	10%	4%
docking $<$ 4 Å ↑	**78%**	18%	37%	14%	37%	12%
docking $<$ 5 Å ↑	**80%**	29%	41%	16%	39%	18%
docking $<$ 8 Å ↑	**94%**	59%	78%	49%	55%	57%
docking $>$ 8 Å (mis-docked) ↓	**6%**	41%	22%	51%	45%	43%
interface RMSD $<$ 2 Å (CAPRI medium) ↑	**80%**	59%	43%	33%	43%	20%
interface RMSD $<$ 4 Å (CAPRI acceptable) ↑	**96%**	90%	84%	76%	67%	65%
contact recall @5 Å split by placement, median						
well-docked targets ↑	**0.92**	0.72	0.63	0.64	0.69	0.42
mis-docked targets ↑	0.20	**0.44**	0.14	0.36	0.05	0.18

Table 9: The same table on the 10 targets released after Boltz-2 training cutoff of 2023-06-01, the only ones it cannot have memorized. Every tool is scored on the same 10, so the comparison is matched. Boltz-2 goes from first on every contact metric to last on Jaccard and precision. The other five move far less, which is what separates memorization from a harder batch of targets.

	Boltz-2	ESMFold2	TCRmodel2	AF-Multimer	OpenDDE	tFold-TCR
interface contacts, median [mean]						
contact recall @4 Å ↑	0.19 [0.25]	**0.47 [0.49]**	0.27 [0.31]	0.39 [0.42]	0.23 [0.21]	0.17 [0.20]
contact recall @5 Å ↑	0.28 [0.31]	**0.57 [0.56]**	0.43 [0.43]	0.49 [0.50]	0.30 [0.29]	0.26 [0.26]
contact recall @8 Å ↑	0.45 [0.51]	**0.70 [0.69]**	0.64 [0.60]	0.66 [0.65]	0.47 [0.41]	0.41 [0.41]
contact precision @5 Å ↑	0.26 [0.30]	**0.56 [0.61]**	0.40 [0.40]	0.45 [0.46]	0.36 [0.28]	0.33 [0.33]
contact Jaccard @5 Å ↑	0.15 [0.19]	**0.39 [0.43]**	0.25 [0.27]	0.30 [0.34]	0.20 [0.17]	0.17 [0.17]
geometry (Å), median [mean]						
interface RMSD ↓	3.41 [3.24]	**1.62 [1.89]**	2.97 [3.41]	2.29 [3.52]	3.57 [5.97]	3.89 [4.28]
TCR docking RMSD ↓	6.55 [7.08]	**6.02 [5.74]**	6.78 [7.99]	7.34 [9.83]	8.35 [16.32]	9.64 [9.92]
peptide RMSD ↓	1.79 [1.56]	**0.89 [1.07]**	1.75 [1.54]	1.98 [1.69]	1.97 [1.69]	2.13 [2.07]
CDR3β loop RMSD ↓	4.91 [4.41]	**2.44 [2.32]**	3.96 [5.04]	3.15 [4.15]	4.78 [11.72]	4.23 [4.92]
CDR3α loop RMSD ↓	4.55 [5.18]	**3.10 [3.92]**	4.99 [5.74]	3.34 [4.81]	6.37 [12.99]	6.87 [6.84]
receptor placement, fraction of the 10 targets						
docking <2 Å (near-native) ↑	0%	**20%**	0%	0%	0%	0%
docking <4 Å ↑	10%	**40%**	20%	20%	10%	20%
docking <5 Å ↑	10%	**40%**	20%	20%	20%	20%
docking <8 Å ↑	**70%**	**70%**	**70%**	**70%**	40%	40%
docking >8 Å (mis-docked) ↓	**30%**	**30%**	**30%**	**30%**	60%	60%
interface RMSD <2 Å (CAPRI medium) ↑	10%	**60%**	20%	50%	0%	0%
interface RMSD <4 Å (CAPRI acceptable) ↑	80%	**100%**	90%	90%	70%	50%
contact recall @5 Å split by placement, median						
well-docked targets ↑	0.40	**0.72**	0.47	0.63	0.38	0.30
mis-docked targets ↑	0.20	0.34	**0.42**	0.40	0.27	0.21

Sampling. Table 10 gives the per-tool figures behind the coherence argument. Two of its rows need a caveat. Boltz-2's 0.754 Å cannot be read as accuracy, because 39 of the 49 targets were released before its training cutoff. Split on that cutoff, the median is 0.68 Å inside the training window against 1.72 Å outside it, Mann–Whitney $p = 7.6 \times 10^{-5}$, and no other tool shows the split. AF-Multimer needs no such split, since its training cutoff precedes every target in the set, and restricting it to the ten clean targets leaves 1.589 Å and $\gamma = 0.969$. Its confidence column is also not on the same footing as the others, because AlphaFold publishes no per-chain-pair value and the global score is used in its place, so that one row cannot be compared against the others on calibration.

Table 10: Twenty co-folds per target do not average AIC away. Each stochastic tool folds all 49 targets 20 times, which is 980 structures per tool. AIC is the median over targets of the per-target median. γ is the directional coherence of the 20 samples, against 0.21 for uniformly random directions. The change column is the median over targets of $1 -$ equal/single, positive meaning better, with a two-sided Wilcoxon signed-rank p. The oracle column picks the best of the 20 per target and is not available to any deployed system.

tool	AIC (Å)	γ	equal-weight change	Wilcoxon p	oracle
Boltz-2	**0.754**	0.949	$+3.6\%$	0.194	$+44.4\%$
ESMFold2	2.227	0.839	$+6.7\%$	0.328	$+52.0\%$
AF-Multimer	2.965	**0.992**	$+1.3\%$	0.851	$+18.6\%$
TCRmodel2	3.062	0.985	-2.5%	0.078	$+27.3\%$
OpenDDE	3.961	0.946	$+0.6\%$	0.465	$+34.6\%$
tFold-TCR	3.52	deterministic, no distribution to sample			

F Experiment Results

F.1 Evaluation Protocol

Composition. The curated evaluation set contains 7,371 measured binding pairs over 343 epitopes, 7,142 distinct TCRs and 353 class-I pMHCs. Sampled non-binders extend it to 45,909 rows at a prevalence of 0.161, which is the chance level for every precision metric in the paper. The 343 epitopes are presented on 353 pMHC, because ten of them appear with two MHC alleles. The

distribution is long tailed, and the ten most frequent epitopes account for most of the entries. That is a property of the field's data, not of our protocol, and it is the reason macro and micro aggregations are reported side by side.

Negative samples. The non-binders come with the STAG release and are used without resampling (Slone et al., 2025a). STAG builds them two ways. It takes non-binding pairs from the 10x Genomics dextramer panel (10x Genomics, 2019), and it swaps TCRs, holding the pMHC of a recorded binder fixed and drawing a TCR from the repertoire to pair with it. The released file carries no provenance column, so the two blocks cannot be separated. Table 11 gives the composition.

Table 11: Composition of the 38,538 non-binders. Quartiles run over the 353 pMHC of the set.

quantity	value
non-binders per binder within one pMHC, median	5.5
quartiles	4.9, 6.0
maximum	32
pMHC carrying at least one binder	353
pMHC carrying at least one non-binder	348
distinct TCRs	21,477
in a binding pair	7,142
in a non-binding pair	19,174
in non-binding pairs only	14,335
non-binding rows carried by those TCRs	22,314
times a TCR is reused across non-binders, median	2
90th percentile	4
maximum	26
pairs labeled both binding and non-binding	0

Fold sizes. Every epitope forms its own cluster and each cluster is assigned to exactly one fold. No epitope appears in more than one fold.

Table 12: The five epitope-disjoint folds.

fold	rows	positives	negatives	pos : neg	epitopes
0	12,327	1,876	10,451	1 : 5.6	65
1	9,342	1,593	7,749	1 : 4.9	67
2	9,337	1,531	7,806	1 : 5.1	69
3	8,543	1,329	7,214	1 : 5.4	70
4	6,360	1,042	5,318	1 : 5.1	72
total	45,909	7,371	38,538	1 : 5.2	343

Rotation. For test fold F the validation fold is $F+1 \bmod 5$ and the remaining three train. Checkpoint selection is on validation micro-AUROC, a three-epoch rolling mean with no checkpoint before epoch 10. The test fold is consulted neither in training nor in selection.

Per fold, then averaged. Every metric in the paper is computed within a fold and the five values are averaged. None is computed on the five folds pooled. Each fold is a separately trained model on its own logit scale, so pooling lets between-model score offsets drive the AUROC and silently weights by fold size. An earlier version of this analysis pooled, and reported one configuration at micro-AUROC 0.676 against its true 0.642.

Row coverage. Each method is scored on every row it covers. Coverage differs between methods by under 0.5% of the fold, because a few structures are missing under one tool and the sequence baselines drop rows they cannot parse. EMBER emits a logit and the baselines emit probabilities. Only the EMBER side is passed through a sigmoid, which is monotone, so AUROC and AUPRC are untouched and MCC@0.5 on the probability equals MCC@0 on the logit.

Degenerate folds. Every fold of every seed is kept in Table 1, including the ones on which a baseline fails. SageTCR collapses to a constant logit on fold 0 under all four structure sources, with a score standard deviation of zero and as few as one distinct value across 12,318 candidate pairs, so its

AUROC there is 0.5 and its MCC at a fixed threshold is 0. TRAP does not go fully constant but its scores span a range too narrow to cross the 0.5 threshold on folds 0 and 4, which also drives MCC@0.5 to 0. Dropping those folds would report each baseline on the folds it happens to survive, so we keep them and report the mean over all five.

F.2 Metric Definitions

Ranking. For a scored set S, $\mathrm{AUROC}(S) = \Pr(\hat{y}^+ > \hat{y}^-)$ over random positive–negative pairs and $\mathrm{AUPRC}(S) = \int_0^1 \mathrm{Prec}(r)\,\mathrm{d}r$. Writing $\mathcal{E}$ for the epitopes of a fold and S_e for the candidates of epitope e,

$$\text{micro-}M = M\Big(\bigcup\nolimits_{e\in\mathcal{E}} S_e\Big), \qquad \text{macro-}M = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} M(S_e), \qquad M \in \{\mathrm{AUROC}, \mathrm{AUPRC}\},$$

so micro is dominated by the frequent epitopes and macro counts every epitope once. A precision metric starts at the class prevalence rather than at zero, which is 0.161 here, so a loss is read against the margin a method holds above that level.

Classification. At a threshold t a pair counts as predicted positive when $\hat{y} \geq t$. Writing TP, TN, FP and FN for the four counts this produces,

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \qquad \mathrm{MCC@0.5} = \mathrm{MCC}\big|_{t=0.5}.$$

MCC runs over $[-1, 1]$ and is 0 for a constant or random prediction, so it is not inflated by the class imbalance the way accuracy is. We fix $t = 0.5$ for every method, which is the cut a deployed system runs at. We also report $\mathrm{MCC}^* = \max_t \mathrm{MCC}(t)$ taken on the test fold itself. No deployed system has that threshold, so MCC^* is an oracle value and we read it only as a bound on what threshold tuning could buy a baseline. We also report $\mathrm{MCC@}t$, where t is fitted on the four folds that are not being scored, so it is the cut a deployed system can set. Table 13 gives the three side by side, with calibration and the operating point each one produces.

Stability. Stability is measured on rankings rather than scores. For two rankings of the same n candidates,

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}, \qquad \text{order held} = \frac{1 + \tau_b}{2}$$

with C and D the concordant and discordant pairs and T_x, T_y those tied in one ranking only (Kendall, 1938; 1945). Across all $m = 4$ rankings at once we use Kendall's concordance $W = 12S/\big(m^2(n^3 - n)\big)$ (Kendall & Babington Smith, 1939), where S sums the squared deviations of each candidate's rank sum from the mean. $W = 1$ is perfect agreement and $W = 0$ is none. Overlap@k is the intersection of two top-k sets divided by k. Intervals are bootstrapped over epitope-folds.

Table 13: Three ways to set the decision threshold, with the calibration and the operating point each one produces. MCC@0.5 uses the fixed cut. MCC@t uses a threshold fitted on the four folds that are not being scored, which is the cut a deployed system can set. MCC* uses the threshold that maximizes MCC on the scored fold, which is an oracle. The first and third columns repeat Table 1. Precision, recall and predicted-positive rate are read at t. Brier and ECE are threshold free, and ECE uses ten equal-mass bins. t is the median of the four per-fold thresholds. Means over five folds and three seeds. Best MCC@t per group in bold.

structures	method	MCC@0.5	MCC@t	MCC*	precision	recall	pred. pos.	Brier	ECE	t
tFold-TCR	**EMBER**	0.251	**0.233**	0.292	0.517	0.206	0.068	0.126	0.070	0.452
	STAG-LLM	0.166	0.170	0.264	0.442	0.132	0.049	0.188	0.216	0.720
	SageTCR	0.122	0.137	0.182	0.399	0.101	0.039	0.133	0.040	0.318
	Abeer	0.071	0.108	0.157	0.293	0.120	0.058	0.237	0.303	0.745
	TRAP	0.034	0.046	0.095	0.409	0.015	0.006	0.200	0.226	0.639
ESMFold2	**EMBER**	0.273	**0.257**	0.304	0.505	0.253	0.086	0.126	0.077	0.435
	STAG-LLM	0.172	0.224	0.245	0.465	0.224	0.079	0.192	0.210	0.748
	SageTCR	0.115	0.117	0.174	0.336	0.105	0.049	0.135	0.041	0.298
	Abeer	0.108	0.110	0.188	0.297	0.115	0.054	0.231	0.300	0.706
	TRAP	0.041	0.061	0.096	0.402	0.019	0.007	0.196	0.224	0.621
TCRmodel2	**EMBER**	0.267	**0.209**	0.301	0.535	0.180	0.060	0.127	0.082	0.487
	STAG-LLM	0.153	0.169	0.264	0.422	0.164	0.066	0.200	0.237	0.736
	SageTCR	0.112	0.128	0.174	0.338	0.119	0.052	0.134	0.043	0.248
	Abeer	0.145	0.123	0.236	0.329	0.134	0.066	0.224	0.289	0.701
	TRAP	0.019	0.035	0.077	0.253	0.015	0.007	0.201	0.224	0.641
Boltz-2	**EMBER**	0.278	**0.263**	0.311	0.623	0.192	0.048	0.127	0.079	0.559
	STAG-LLM	0.169	0.183	0.254	0.454	0.160	0.058	0.198	0.218	0.800
	SageTCR	0.115	0.114	0.146	0.310	0.122	0.060	0.134	0.046	0.257
	Abeer	0.054	0.090	0.188	0.251	0.145	0.089	0.251	0.319	0.683
	TRAP	0.016	0.042	0.082	0.270	0.014	0.006	0.192	0.206	0.661
-	NetTCR-2.2	0.207	0.241	0.257	0.493	0.212	0.065	0.128	0.060	0.360
	ERGO-II	0.061	0.199	0.224	0.474	0.161	0.056	0.282	0.373	0.799

F.3 HYPERPARAMETER

The objective sums a cross-entropy term and a Bradley–Terry ranking term, both defined in the method. Negatives are redrawn every epoch rather than fixed once, and checkpoints are selected on a rolling three-epoch mean of validation micro-AUROC, because taking the single best epoch selects on the validation fold.

Table 14: Every hyperparameter of the reported EMBER runs. Values are shared by all arms. The arms differ only in the flag named in Appendix F.5.

group	setting	value
graph	node descriptor x	37 invariant scalars
	language model	ESM-C 600M, frozen, 1152-d
	learned bottleneck W_e	$1152 \to 128$
	encoder input $h^{(0)}$	165-d
	neighbors k	30, rebuilt every layer
	edge feature a_{ij}	16 RBF over $[0, 20]$ Å $+ d_0/d_{ij}$ ($d_0 = 4$ Å) $+ 5$ biochemical
encoder	layers L	2
	width	256
	activation	SiLU
	coordinate update	Tanh-bounded coefficient
head	attention heads	4×32
	contact dimension	128
	pooling	mask-aware mean + max
optimiser	algorithm	AdamW (Loshchilov & Hutter, 2019)
	learning rate	3×10^{-5}
	weight decay	0.01
	schedule	5 warm-up epochs, then cosine
	epochs	100
sampling	negatives per positive	3, redrawn every epoch
	positive sampling	uniform at natural epitope frequency
selection	criterion	rolling mean (window 3) of validation micro-AUROC
	floor	epoch 10

F.4 Binding Prediction, Fold by Fold

Every mean in Table 1 is an average of five per-fold values, and the tables below give all of them. The spread across folds is wide enough to be worth stating plainly: fold 0 is the easiest fold on the micro metrics for every method that separates the classes at all, and the three weakest methods peak elsewhere only because they barely rank on fold 0 either. On the macro metrics, which average over epitopes and so discount that one large epitope, no method peaks on fold 0.

The last column of each table is MCC at the threshold that maximizes it on that fold, the oracle value of Table 13, so the cost of the fixed cut can be read fold by fold.

Fold by fold against the strongest baseline. The means of Table 1 say which method is ahead on average. Table 15 asks the same question one fold at a time, pairing on the fold because that is the only unit two methods share. Seeds are not a shared unit, since each method draws its own, so seeds are averaged within a fold before pairing. EMBER wins all five folds in 16 of the 24 cells. The exceptions are micro-AUROC under ESMFold2 and Boltz-2, where the median difference is $+0.005$ and $+0.009$ and three folds go its way, micro-AUPRC under tFold-TCR and ESMFold2, where four folds do, and MCC^*, where no tool reaches five. Both exceptions sit where the main text already reads the lead as small, on the whole-list metric and after the threshold has been tuned on the test fold.

Table 15: Fold-by-fold comparison against the strongest baseline in each cell. Seeds are averaged within a fold first, so each cell is five paired observations, one per epitope-disjoint fold. The entry gives the folds EMBER wins and the median paired difference. Five folds out of five is a one-sided Wilcoxon p of 0.031, which is the smallest value five paired observations can produce, so no cell can report a stronger test. The baseline is STAG-LLM except for macro-AUPRC under tFold-TCR, ESMFold2 and TCRmodel2 and macro-AUROC under tFold-TCR, which are against TRAP, and micro-AUROC under TCRmodel2, which is against Abeer.

metric	tFold-TCR	ESMFold2	TCRmodel2	Boltz-2
macro-AUROC	**5/5** (+0.100)	**5/5** (+0.045)	**5/5** (+0.130)	**5/5** (+0.036)
macro-AUPRC	**5/5** (+0.067)	**5/5** (+0.060)	**5/5** (+0.131)	**5/5** (+0.064)
micro-AUROC	**5/5** (+0.062)	3/5 (+0.005)	**5/5** (+0.075)	3/5 (+0.009)
micro-AUPRC	4/5 (+0.060)	4/5 (+0.031)	**5/5** (+0.058)	**5/5** (+0.022)
MCC@0.5	**5/5** (+0.082)	**5/5** (+0.048)	**5/5** (+0.075)	**5/5** (+0.068)
MCC*	3/5 (+0.017)	4/5 (+0.026)	4/5 (+0.017)	4/5 (+0.025)

Table 16: Per-fold results, structures from tFold-TCR. Each cell is the mean over three training seeds, ± the spread across those seeds. Every *mean* row averages the five folds within a seed first, so it matches Table 1. No metric is computed on the five folds pooled. Candidate counts differ between methods by under 0.5%, so they are not repeated here.

method / fold	macro-AUROC	macro-AUPRC	micro-AUROC	micro-AUPRC	MCC@0.5	MCC*
EMBER						
0	0.713 ± 0.014	0.530 ± 0.023	0.855 ± 0.010	0.656 ± 0.049	0.512 ± 0.279	0.648 ± 0.045
1	0.776 ± 0.024	0.611 ± 0.025	0.628 ± 0.005	0.278 ± 0.008	0.144 ± 0.010	0.154 ± 0.004
2	0.788 ± 0.010	0.615 ± 0.002	0.582 ± 0.012	0.283 ± 0.015	0.172 ± 0.020	0.208 ± 0.023
3	0.777 ± 0.014	0.613 ± 0.018	0.625 ± 0.013	0.273 ± 0.024	0.152 ± 0.033	0.169 ± 0.031
4	0.736 ± 0.008	0.555 ± 0.004	0.639 ± 0.002	0.343 ± 0.004	0.275 ± 0.016	0.279 ± 0.014
mean	0.758 ± 0.010	0.585 ± 0.008	0.666 ± 0.003	0.366 ± 0.014	0.251 ± 0.059	0.292 ± 0.017
STAG-LLM						
0	0.620 ± 0.021	0.480 ± 0.036	0.792 ± 0.028	0.455 ± 0.061	0.445 ± 0.107	0.474 ± 0.116
1	0.642 ± 0.058	0.475 ± 0.067	0.490 ± 0.015	0.242 ± 0.026	0.036 ± 0.037	0.192 ± 0.033
2	0.649 ± 0.027	0.496 ± 0.029	0.532 ± 0.005	0.223 ± 0.008	0.090 ± 0.033	0.171 ± 0.016
3	0.642 ± 0.015	0.500 ± 0.014	0.574 ± 0.002	0.275 ± 0.020	0.127 ± 0.031	0.221 ± 0.015
4	0.546 ± 0.031	0.411 ± 0.022	0.522 ± 0.021	0.264 ± 0.002	0.134 ± 0.114	0.263 ± 0.003
mean	0.620 ± 0.018	0.473 ± 0.020	0.582 ± 0.008	0.292 ± 0.006	0.166 ± 0.024	0.264 ± 0.018
SageTCR						
0	0.512 ± 0.016	0.243 ± 0.079	0.519 ± 0.028	0.156 ± 0.006	0.000 ± 0.000	0.120 ± 0.113
1	0.643 ± 0.027	0.474 ± 0.020	0.510 ± 0.017	0.251 ± 0.019	0.176 ± 0.047	0.199 ± 0.023
2	0.668 ± 0.008	0.511 ± 0.030	0.554 ± 0.022	0.253 ± 0.021	0.128 ± 0.062	0.189 ± 0.025
3	0.671 ± 0.018	0.525 ± 0.027	0.568 ± 0.016	0.260 ± 0.029	0.154 ± 0.069	0.180 ± 0.041
4	0.615 ± 0.045	0.459 ± 0.020	0.584 ± 0.018	0.282 ± 0.012	0.149 ± 0.004	0.223 ± 0.005
mean	0.622 ± 0.009	0.443 ± 0.004	0.547 ± 0.011	0.240 ± 0.007	0.122 ± 0.027	0.182 ± 0.036
Abeer						
0	0.492 ± 0.049	0.341 ± 0.036	0.601 ± 0.085	0.202 ± 0.065	0.087 ± 0.096	0.130 ± 0.094
1	0.618 ± 0.065	0.455 ± 0.063	0.494 ± 0.030	0.225 ± 0.021	0.032 ± 0.060	0.158 ± 0.039
2	0.672 ± 0.016	0.506 ± 0.021	0.546 ± 0.009	0.221 ± 0.008	0.072 ± 0.016	0.143 ± 0.016
3	0.613 ± 0.028	0.465 ± 0.031	0.558 ± 0.036	0.205 ± 0.027	0.065 ± 0.017	0.124 ± 0.056
4	0.557 ± 0.007	0.402 ± 0.008	0.550 ± 0.043	0.274 ± 0.018	0.100 ± 0.050	0.230 ± 0.020
mean	0.591 ± 0.008	0.434 ± 0.006	0.550 ± 0.011	0.225 ± 0.014	0.071 ± 0.007	0.157 ± 0.024
TRAP						
0	0.573 ± 0.043	0.419 ± 0.058	0.552 ± 0.087	0.187 ± 0.058	0.000 ± 0.000	0.110 ± 0.094
1	0.682 ± 0.010	0.554 ± 0.026	0.423 ± 0.010	0.170 ± 0.005	0.058 ± 0.018	0.101 ± 0.010
2	0.703 ± 0.030	0.577 ± 0.020	0.511 ± 0.022	0.191 ± 0.010	0.038 ± 0.009	0.094 ± 0.006
3	0.662 ± 0.029	0.525 ± 0.051	0.515 ± 0.014	0.174 ± 0.012	0.043 ± 0.027	0.084 ± 0.029
4	0.635 ± 0.045	0.488 ± 0.055	0.494 ± 0.006	0.188 ± 0.019	0.029 ± 0.050	0.088 ± 0.051
mean	0.651 ± 0.015	0.513 ± 0.014	0.499 ± 0.010	0.182 ± 0.010	0.034 ± 0.000	0.095 ± 0.024

Table 17: Per-fold results, structures from ESMFold2. Conventions as in Table 16.

method / fold	macro-AUROC	macro-AUPRC	micro-AUROC	micro-AUPRC	MCC@0.5	MCC*
EMBER						
0	0.737 ± 0.015	0.563 ± 0.020	0.855 ± 0.013	0.677 ± 0.045	0.675 ± 0.051	0.690 ± 0.036
1	0.774 ± 0.014	0.605 ± 0.008	0.660 ± 0.003	0.324 ± 0.017	0.161 ± 0.041	0.206 ± 0.014
2	0.773 ± 0.006	0.587 ± 0.011	0.577 ± 0.005	0.263 ± 0.013	0.165 ± 0.003	0.184 ± 0.012
3	0.781 ± 0.016	0.617 ± 0.013	0.621 ± 0.012	0.262 ± 0.017	0.137 ± 0.041	0.174 ± 0.011
4	0.729 ± 0.004	0.545 ± 0.006	0.639 ± 0.005	0.334 ± 0.007	0.228 ± 0.022	0.267 ± 0.015
mean	0.759 ± 0.004	0.584 ± 0.003	0.670 ± 0.005	0.372 ± 0.013	0.273 ± 0.029	0.304 ± 0.011
STAG-LLM						
0	0.705 ± 0.010	0.489 ± 0.010	0.819 ± 0.017	0.463 ± 0.052	0.360 ± 0.031	0.472 ± 0.049
1	0.758 ± 0.013	0.550 ± 0.025	0.608 ± 0.024	0.293 ± 0.016	0.131 ± 0.003	0.190 ± 0.025
2	0.728 ± 0.006	0.543 ± 0.009	0.603 ± 0.012	0.237 ± 0.009	0.117 ± 0.020	0.126 ± 0.019
3	0.714 ± 0.024	0.523 ± 0.019	0.633 ± 0.003	0.271 ± 0.030	0.131 ± 0.021	0.194 ± 0.010
4	0.664 ± 0.036	0.432 ± 0.056	0.635 ± 0.003	0.297 ± 0.036	0.123 ± 0.025	0.242 ± 0.012
mean	0.714 ± 0.013	0.507 ± 0.016	0.659 ± 0.003	0.312 ± 0.007	0.172 ± 0.001	0.245 ± 0.004
SageTCR						
0	0.486 ± 0.013	0.260 ± 0.035	0.501 ± 0.056	0.150 ± 0.020	0.000 ± 0.000	0.182 ± 0.132
1	0.634 ± 0.014	0.458 ± 0.031	0.512 ± 0.031	0.243 ± 0.007	0.163 ± 0.020	0.175 ± 0.010
2	0.651 ± 0.033	0.487 ± 0.021	0.541 ± 0.011	0.237 ± 0.016	0.143 ± 0.038	0.165 ± 0.027
3	0.680 ± 0.035	0.519 ± 0.037	0.568 ± 0.010	0.231 ± 0.010	0.129 ± 0.021	0.148 ± 0.004
4	0.639 ± 0.023	0.460 ± 0.011	0.604 ± 0.061	0.274 ± 0.028	0.141 ± 0.028	0.202 ± 0.006
mean	0.618 ± 0.002	0.437 ± 0.014	0.545 ± 0.015	0.227 ± 0.003	0.115 ± 0.016	0.174 ± 0.030
Abeer						
0	0.484 ± 0.024	0.329 ± 0.013	0.723 ± 0.043	0.388 ± 0.152	0.303 ± 0.196	0.344 ± 0.150
1	0.614 ± 0.014	0.460 ± 0.032	0.514 ± 0.025	0.220 ± 0.010	0.009 ± 0.008	0.144 ± 0.006
2	0.609 ± 0.026	0.438 ± 0.029	0.544 ± 0.001	0.215 ± 0.005	0.073 ± 0.024	0.132 ± 0.008
3	0.636 ± 0.025	0.486 ± 0.022	0.546 ± 0.005	0.200 ± 0.006	0.049 ± 0.012	0.125 ± 0.014
4	0.609 ± 0.052	0.442 ± 0.047	0.567 ± 0.002	0.260 ± 0.027	0.106 ± 0.014	0.196 ± 0.033
mean	0.590 ± 0.003	0.431 ± 0.002	0.579 ± 0.008	0.257 ± 0.028	0.108 ± 0.049	0.188 ± 0.025
TRAP						
0	0.592 ± 0.008	0.438 ± 0.017	0.500 ± 0.124	0.155 ± 0.032	0.000 ± 0.000	0.102 ± 0.090
1	0.657 ± 0.031	0.515 ± 0.042	0.439 ± 0.009	0.174 ± 0.006	0.064 ± 0.008	0.096 ± 0.010
2	0.680 ± 0.009	0.551 ± 0.015	0.536 ± 0.007	0.194 ± 0.005	0.030 ± 0.004	0.081 ± 0.023
3	0.678 ± 0.025	0.557 ± 0.019	0.528 ± 0.011	0.187 ± 0.005	0.045 ± 0.020	0.098 ± 0.012
4	0.639 ± 0.041	0.506 ± 0.054	0.484 ± 0.010	0.190 ± 0.021	0.064 ± 0.056	0.102 ± 0.062
mean	0.649 ± 0.006	0.514 ± 0.012	0.497 ± 0.025	0.180 ± 0.003	0.041 ± 0.017	0.096 ± 0.008

Table 18: Per-fold results, structures from TCRmodel2. Conventions as in Table 16.

method / fold	macro-AUROC	macro-AUPRC	micro-AUROC	micro-AUPRC	MCC@0.5	MCC*
EMBER						
0	0.729 ± 0.010	0.540 ± 0.006	0.856 ± 0.009	0.665 ± 0.025	0.623 ± 0.034	0.681 ± 0.021
1	0.799 ± 0.018	0.639 ± 0.027	0.647 ± 0.003	0.283 ± 0.014	0.134 ± 0.027	0.175 ± 0.002
2	0.794 ± 0.006	0.605 ± 0.018	0.568 ± 0.007	0.261 ± 0.017	0.172 ± 0.014	0.185 ± 0.020
3	0.775 ± 0.022	0.626 ± 0.021	0.620 ± 0.002	0.275 ± 0.023	0.149 ± 0.045	0.179 ± 0.015
4	0.747 ± 0.009	0.554 ± 0.018	0.648 ± 0.015	0.351 ± 0.017	0.256 ± 0.047	0.284 ± 0.009
mean	0.769 ± 0.007	0.593 ± 0.004	0.668 ± 0.003	0.367 ± 0.002	0.267 ± 0.009	0.301 ± 0.003
STAG-LLM						
0	0.599 ± 0.042	0.457 ± 0.028	0.794 ± 0.020	0.481 ± 0.068	0.372 ± 0.110	0.492 ± 0.080
1	0.640 ± 0.051	0.469 ± 0.035	0.478 ± 0.014	0.226 ± 0.026	0.060 ± 0.007	0.173 ± 0.026
2	0.673 ± 0.016	0.518 ± 0.004	0.535 ± 0.005	0.224 ± 0.004	0.060 ± 0.035	0.170 ± 0.010
3	0.669 ± 0.021	0.512 ± 0.022	0.569 ± 0.010	0.271 ± 0.018	0.078 ± 0.024	0.222 ± 0.020
4	0.577 ± 0.030	0.427 ± 0.019	0.553 ± 0.013	0.272 ± 0.023	0.196 ± 0.018	0.262 ± 0.017
mean	0.632 ± 0.009	0.477 ± 0.005	0.586 ± 0.008	0.295 ± 0.018	0.153 ± 0.029	0.264 ± 0.019
SageTCR						
0	0.470 ± 0.035	0.187 ± 0.026	0.465 ± 0.048	0.145 ± 0.009	0.000 ± 0.000	0.112 ± 0.087
1	0.624 ± 0.036	0.473 ± 0.023	0.522 ± 0.013	0.249 ± 0.013	0.105 ± 0.045	0.176 ± 0.023
2	0.617 ± 0.048	0.465 ± 0.049	0.548 ± 0.007	0.241 ± 0.027	0.127 ± 0.053	0.167 ± 0.034
3	0.677 ± 0.017	0.524 ± 0.019	0.572 ± 0.007	0.249 ± 0.010	0.138 ± 0.009	0.174 ± 0.007
4	0.593 ± 0.035	0.439 ± 0.016	0.578 ± 0.001	0.292 ± 0.010	0.190 ± 0.020	0.238 ± 0.026
mean	0.596 ± 0.028	0.418 ± 0.011	0.537 ± 0.012	0.235 ± 0.004	0.112 ± 0.018	0.174 ± 0.011
Abeer						
0	0.486 ± 0.065	0.348 ± 0.056	0.780 ± 0.011	0.519 ± 0.098	0.372 ± 0.177	0.517 ± 0.004
1	0.600 ± 0.029	0.443 ± 0.028	0.531 ± 0.021	0.226 ± 0.011	0.088 ± 0.076	0.144 ± 0.020
2	0.633 ± 0.025	0.476 ± 0.016	0.542 ± 0.008	0.214 ± 0.005	0.072 ± 0.028	0.137 ± 0.022
3	0.618 ± 0.026	0.485 ± 0.025	0.563 ± 0.015	0.215 ± 0.017	0.062 ± 0.034	0.142 ± 0.051
4	0.580 ± 0.012	0.421 ± 0.023	0.570 ± 0.007	0.280 ± 0.024	0.132 ± 0.019	0.240 ± 0.040
mean	0.583 ± 0.023	0.434 ± 0.017	0.597 ± 0.006	0.291 ± 0.021	0.145 ± 0.024	0.236 ± 0.019
TRAP						
0	0.578 ± 0.028	0.409 ± 0.032	0.460 ± 0.069	0.147 ± 0.017	0.000 ± 0.000	0.058 ± 0.031
1	0.635 ± 0.068	0.498 ± 0.099	0.430 ± 0.056	0.159 ± 0.021	0.003 ± 0.013	0.068 ± 0.039
2	0.692 ± 0.022	0.573 ± 0.012	0.521 ± 0.011	0.197 ± 0.006	0.045 ± 0.007	0.104 ± 0.013
3	0.671 ± 0.018	0.549 ± 0.016	0.526 ± 0.002	0.191 ± 0.003	0.046 ± 0.018	0.111 ± 0.010
4	0.572 ± 0.019	0.401 ± 0.024	0.496 ± 0.025	0.167 ± 0.016	0.000 ± 0.000	0.043 ± 0.026
mean	0.630 ± 0.008	0.486 ± 0.011	0.487 ± 0.009	0.172 ± 0.004	0.019 ± 0.006	0.077 ± 0.006

Table 19: Per-fold results, structures from Boltz-2. Conventions as in Table 16.

method / fold	macro-AUROC	macro-AUPRC	micro-AUROC	micro-AUPRC	MCC@0.5	MCC*
EMBER						
0	0.748 ± 0.014	0.572 ± 0.021	0.861 ± 0.006	0.671 ± 0.021	0.627 ± 0.022	0.666 ± 0.034
1	0.770 ± 0.013	0.598 ± 0.022	0.628 ± 0.039	0.316 ± 0.013	0.173 ± 0.033	0.225 ± 0.001
2	0.770 ± 0.010	0.586 ± 0.010	0.591 ± 0.009	0.285 ± 0.012	0.182 ± 0.040	0.219 ± 0.021
3	0.751 ± 0.015	0.598 ± 0.012	0.603 ± 0.008	0.275 ± 0.021	0.173 ± 0.028	0.189 ± 0.027
4	0.727 ± 0.011	0.538 ± 0.006	0.623 ± 0.006	0.311 ± 0.006	0.236 ± 0.028	0.256 ± 0.036
mean	0.753 ± 0.004	0.579 ± 0.007	0.661 ± 0.006	0.372 ± 0.003	0.278 ± 0.013	0.311 ± 0.012
STAG-LLM						
0	0.696 ± 0.007	0.483 ± 0.005	0.811 ± 0.028	0.460 ± 0.097	0.342 ± 0.022	0.488 ± 0.093
1	0.741 ± 0.011	0.544 ± 0.012	0.613 ± 0.011	0.304 ± 0.016	0.135 ± 0.038	0.206 ± 0.014
2	0.734 ± 0.007	0.525 ± 0.022	0.607 ± 0.007	0.243 ± 0.002	0.114 ± 0.012	0.141 ± 0.015
3	0.718 ± 0.014	0.534 ± 0.005	0.633 ± 0.003	0.274 ± 0.020	0.121 ± 0.012	0.202 ± 0.005
4	0.661 ± 0.023	0.441 ± 0.027	0.614 ± 0.017	0.290 ± 0.013	0.131 ± 0.014	0.231 ± 0.017
mean	0.710 ± 0.002	0.505 ± 0.003	0.655 ± 0.008	0.314 ± 0.026	0.169 ± 0.012	0.254 ± 0.020
SageTCR						
0	0.482 ± 0.016	0.226 ± 0.064	0.452 ± 0.043	0.140 ± 0.015	0.000 ± 0.000	0.033 ± 0.017
1	0.611 ± 0.010	0.457 ± 0.028	0.475 ± 0.029	0.229 ± 0.030	0.155 ± 0.054	0.172 ± 0.034
2	0.659 ± 0.029	0.503 ± 0.024	0.554 ± 0.003	0.241 ± 0.004	0.134 ± 0.038	0.171 ± 0.011
3	0.644 ± 0.035	0.508 ± 0.024	0.578 ± 0.009	0.237 ± 0.009	0.109 ± 0.007	0.148 ± 0.014
4	0.591 ± 0.019	0.416 ± 0.017	0.575 ± 0.023	0.272 ± 0.014	0.177 ± 0.008	0.205 ± 0.013
mean	0.597 ± 0.001	0.422 ± 0.013	0.527 ± 0.012	0.224 ± 0.009	0.115 ± 0.003	0.146 ± 0.004
Abeer						
0	0.516 ± 0.026	0.357 ± 0.021	0.753 ± 0.033	0.409 ± 0.148	0.086 ± 0.153	0.387 ± 0.123
1	0.621 ± 0.017	0.460 ± 0.015	0.488 ± 0.017	0.217 ± 0.014	0.010 ± 0.079	0.150 ± 0.027
2	0.648 ± 0.014	0.472 ± 0.012	0.538 ± 0.018	0.202 ± 0.014	0.049 ± 0.045	0.101 ± 0.027
3	0.624 ± 0.054	0.465 ± 0.048	0.562 ± 0.035	0.208 ± 0.029	0.065 ± 0.049	0.129 ± 0.040
4	0.594 ± 0.017	0.432 ± 0.018	0.558 ± 0.013	0.238 ± 0.006	0.062 ± 0.053	0.172 ± 0.019
mean	0.600 ± 0.003	0.437 ± 0.009	0.580 ± 0.014	0.255 ± 0.032	0.054 ± 0.026	0.188 ± 0.023
TRAP						
0	0.552 ± 0.038	0.395 ± 0.039	0.461 ± 0.071	0.146 ± 0.019	0.000 ± 0.000	0.054 ± 0.034
1	0.686 ± 0.029	0.558 ± 0.027	0.472 ± 0.031	0.185 ± 0.008	0.009 ± 0.028	0.109 ± 0.009
2	0.681 ± 0.008	0.540 ± 0.019	0.521 ± 0.002	0.188 ± 0.001	0.019 ± 0.008	0.090 ± 0.006
3	0.672 ± 0.075	0.544 ± 0.083	0.501 ± 0.031	0.169 ± 0.022	0.024 ± 0.022	0.075 ± 0.054
4	0.622 ± 0.053	0.473 ± 0.088	0.502 ± 0.002	0.186 ± 0.015	0.030 ± 0.053	0.084 ± 0.042
mean	0.642 ± 0.015	0.502 ± 0.006	0.491 ± 0.011	0.175 ± 0.002	0.016 ± 0.013	0.082 ± 0.009

Table 20: Per-fold results, sequence baselines. These methods read no structure, so they are listed once. Each cell is the mean over three training seeds, ± the spread across those seeds.

method / fold	macro-AUROC	macro-AUPRC	micro-AUROC	micro-AUPRC	MCC@0.5	MCC*
NetTCR-2.2						
0	0.676 ± 0.011	0.530 ± 0.006	0.806 ± 0.021	0.541 ± 0.063	0.453 ± 0.128	0.519 ± 0.083
1	0.699 ± 0.024	0.519 ± 0.029	0.478 ± 0.009	0.230 ± 0.008	0.093 ± 0.005	0.161 ± 0.008
2	0.702 ± 0.003	0.550 ± 0.007	0.541 ± 0.003	0.231 ± 0.008	0.120 ± 0.027	0.162 ± 0.018
3	0.689 ± 0.024	0.549 ± 0.036	0.565 ± 0.006	0.222 ± 0.009	0.093 ± 0.018	0.145 ± 0.038
4	0.676 ± 0.021	0.530 ± 0.013	0.631 ± 0.047	0.370 ± 0.016	0.276 ± 0.008	0.299 ± 0.014
mean	0.688 ± 0.007	0.536 ± 0.012	0.604 ± 0.010	0.319 ± 0.009	0.207 ± 0.026	0.257 ± 0.006
ERGO-II						
0	0.658 ± 0.034	0.489 ± 0.027	0.755 ± 0.041	0.428 ± 0.082	0.219 ± 0.085	0.383 ± 0.105
1	0.702 ± 0.007	0.540 ± 0.013	0.482 ± 0.005	0.234 ± 0.011	-0.070 ± 0.006	0.175 ± 0.021
2	0.706 ± 0.022	0.541 ± 0.033	0.555 ± 0.005	0.246 ± 0.015	0.039 ± 0.007	0.173 ± 0.015
3	0.645 ± 0.020	0.480 ± 0.020	0.572 ± 0.003	0.229 ± 0.014	0.060 ± 0.003	0.165 ± 0.012
4	0.631 ± 0.030	0.462 ± 0.027	0.599 ± 0.043	0.285 ± 0.045	0.054 ± 0.043	0.222 ± 0.061
mean	0.668 ± 0.016	0.502 ± 0.016	0.593 ± 0.009	0.284 ± 0.010	0.061 ± 0.026	0.224 ± 0.012

F.5 Architecture Ablations

Two more ablations were run on the same protocol and are reported here. Removing the graph but keeping the structure-derived node descriptors costs 13.1% of the micro-AUPRC margin the control holds above chance, which is close to what removing the structure entirely costs, so most of what the structure contributes arrives through the neighborhood rather than through the per-residue descriptors. Restoring the TCR α chain costs 8.6% of that margin and 10.0% of MCC@0.5, which is the measurement behind the scope decision of Appendix D.

The docked-complex arm. *docked complex* replaces the two monomer structures with the ESMFold2 co-folded complex of the same pair. The two halves already share one frame. The arm concatenates them into one node set and runs a single encoder pass. The control runs one pass per molecule. Cross-molecular edges appear on their own, because the k-NN graph is rebuilt on the union at every layer. No cross-edge flag and no new feature is added. The five biochemical edge terms read residue identity alone, so they were already defined across the boundary. The contact head is unchanged. The complex cache keeps the node order of the monomer graph, so the arm reuses the same language-model cache and the same CDR3β and peptide row sets. Capacity is unchanged as well. Both arms share one encoder between the two molecules, and the docked arm runs that same encoder over the union. The two arms hold 3,027,721 parameters under identical names and shapes. Node features are recomputed from the complex instead of copied from the monomer graphs, because a co-folded half differs from its separately folded monomer. Cross-molecular edges carry distance and residue identity only, so the score stays invariant to a rigid motion of the pair. It is not invariant to independent motions of the two halves. That second invariance is what Proposition 1 states and what this arm gives up.

Table 21: The full ablation ladder, on the ESMFold2 structures. The first four rows are Table 3, repeated so the two remaining arms can be read against them. *no graph* replaces the encoder with a simple multilayer perceptron, which removes the neighborhood and the coordinate update while keeping the structure-derived node descriptors. *alpha chain restored* puts the TCR α chain back into the encoder input (Appendix D). Means over five folds and three seeds, $\pm$ the spread across seeds.

arm	macro-AUROC	macro-AUPRC	micro-AUROC	micro-AUPRC	MCC@0.5	MCC*
EMBER (control)	0.759 $\pm$ 0.004	0.584 $\pm$ 0.003	0.670 $\pm$ 0.005	0.372 $\pm$ 0.013	0.273 $\pm$ 0.029	0.304 $\pm$ 0.011
remove structure	0.753 $\pm$ 0.008	0.570 $\pm$ 0.006	0.649 $\pm$ 0.003	0.349 $\pm$ 0.001	0.236 $\pm$ 0.018	0.283 $\pm$ 0.010
docked complex	0.745 $\pm$ 0.009	0.578 $\pm$ 0.013	0.654 $\pm$ 0.009	0.346 $\pm$ 0.033	0.232 $\pm$ 0.063	0.274 $\pm$ 0.027
no peptide frame	0.740 $\pm$ 0.008	0.558 $\pm$ 0.004	0.664 $\pm$ 0.007	0.360 $\pm$ 0.004	0.209 $\pm$ 0.017	0.279 $\pm$ 0.009
no graph	0.750 $\pm$ 0.002	0.561 $\pm$ 0.002	0.656 $\pm$ 0.006	0.344 $\pm$ 0.007	0.249 $\pm$ 0.012	0.275 $\pm$ 0.020
alpha chain restored	0.739 $\pm$ 0.020	0.568 $\pm$ 0.026	0.643 $\pm$ 0.004	0.354 $\pm$ 0.014	0.246 $\pm$ 0.025	0.280 $\pm$ 0.012

F.6 Preprocessing Benchmark

What is timed. The sample behind Table 4 is fixed once and shared by all four tools, because folding time scales with residue count. Residue medians are 228 for a receptor, 382 for a target and 610 for a complex. Each cell is GPU time for the inference call alone. Model load is excluded and timed separately. The first structure of each run is excluded as well, since it pays one-time kernel autotuning.

Two rows carry a cost the table does not show. TCRmodel2 spends 45 to 48 percent of its wall clock outside inference, on container start, imports, parameter loading, template search and featurization, so its 2.6$\times$ is a ratio between two numbers that both carry that overhead. Boltz-2 cannot fold without a multiple sequence alignment, and that alignment is not in its row. Every chain carries an explicit alignment path and the 9-mer peptide carries none, which matches the production configuration rather than the cheaper one. Timing it against a public alignment server would have compared a different quantity to tFold-TCR and ESMFold2, which need no alignment at all.

Limits. Folding a complex is not expensive in itself, and on a fixed candidate set the advantage is a constant factor rather than a change of regime. Each cell is one measurement rather than a mean over repeats, so no error bar is claimed, though the within-cell spread is tight at p_{90} over median of 1.03 to 1.08.