DoRIAT: A Bayesian framework for interpreting and annotating TCR-pHLA docking runs.

Christos Maniatis
Etcembly LTD
Harwell Campus, Oxfordshire, UK
christos@etcembly.com

Zahra Ouaray
Etcembly LTD
Harwell Campus, Oxfordshire, UK
zahra@etcembly.com

Chengkai Xiao
Etcembly LTD
Harwell Campus, Oxfordshire, UK
kai@etcembly.com

Thomas P.E. Dixon
Etcembly LTD
Harwell Campus, Oxfordshire, UK
tom@etcembly.com

Charlie Naylor
Etcembly LTD
Harwell Campus, Oxfordshire, UK
charlie@etcembly.io

James Snowden
Etcembly LTD
Harwell Campus, Oxfordshire, UK
james@etcembly.io

Michelle Teng
Etcembly LTD
Harwell Campus, Oxfordshire, UK
michelle@etcembly.io

Jacob Hurst
Etcembly LTD
Harwell Campus, Oxfordshire, UK
jacobhurst@etcembly.io

Abstract

The advent of sequence-to-structure deep-learning models has transformed the protein engineering landscape by providing an accurate and cost-effective way to determine crystal structures. Despite their accuracy, deep-learning predictions tend to offer limited insights into protein dynamics. To improve conformational exploration in the context of T cell receptor (TCR) docking, we have developed a machine learning pipeline that combines deep-learning predictions with molecular docking. In this report, we introduce **Do**cking **R**un **I**ntepretation and **A**nnotation **T**ool (DoRIAT). In contrast to frameworks that score models based on interface interactions, DoRIAT uses a set of parameters that summarize the binding conformation. We use DoRIAT to annotate TCR-pHLA docking runs, score the resulting complexes, identify models close to the native crystal structure, and create ensembles of models with similar binding conformations. Our results demonstrate that the single structural model selected by DoRIAT as the closest representation of the crystal structure lies within the top 10 docked models, ranked by Root Mean Squared Distance (RMSD), in approximately 80% of HLA-A*02 complexes considered.

1 Introduction

T cell-mediated immunity is a crucial aspect of our immune system, playing a vital role in defending against pathogens and cancerous cells. As T cells have the ability to scrutinise the entire proteome within a cell, by interrogating short peptides presented on the surface of cell by human leukocyte

European Conference on Neural Information Processing Systems (EurIPS) Workshop: SimBioChem 2025.

antigens (pHLA), they can constitute a very effective means to protect humans from pathogens or other foreign substances [15]. Characterising the molecular factors governing interactions between T cell receptor (TCR) and pHLA could shed light on T cells' ability to discriminate against antigens [15] and pave the way for developing novel T cell therapies targeting cancer and autoimmune diseases. An example of a complete TCR-pHLA complex from the public domain, determined by X-ray crystallography [48], can be found in Fig. 4.

Resolving protein crystals has a long and interesting history dating back to the early 1930's [5, 14, 38, 31]. However, the idea that the knowledge of a protein structure could aid in the design of specific ligands, appeared a few years after the launch of the Protein Data Bank, in 1971 [4, 6, 48]. Technological improvements like crystallisation automates, brighter synchrotron X-ray sources, faster detectors, automated structure solution and refinement pipelines have significantly improved time [18, 48], making structural resolution an integral part of pre-clinical target-based drug design. Structure-based design has now delivered drugs for a number of important diseases, including cancer, HIV, glaucoma and hypertension [66, 19]. However, end-to-end protein crystal structure determination remains an expensive and laborious process.

The advent of deep-learning based sequence-to-structure models like AlphaFold [37] have transformed the landscape for *in silico* protein engineering. The initial versions of AlphaFold did not accurately model multi-chain complexes, however this has been addressed by a later release [36]. Although modern sequence-to-structure deep-learning models can generate highly accurate structural predictions, they frequently display a significant degree of structural homogeneity. Therefore, a crucial knowledge gap persists with regards to the incorporation of molecular dynamics and flexibility, as highlighted in [36]. A single structural model gives a static picture of a system which in reality is in a constant state of flux. An ensemble-based approach can overcome this problem and more comprehensively describe the preserved and transient interactions. Our approach has been to model the various TCR-pHLA complex components with sequence-to-structure models and bring them together with traditional docking algorithms, similar to [27], in a pipeline named EMLy™Dock.

More precisely, we model alpha and beta chains of the TCR and the HLA using TCRModel2 [77], an AlphaFold derivative. Then HADDOCK3.0 [22], a physics based docking platform, brings together the TCR, peptide and HLA like similar methods found in literature [16, 28, 10, 65]. The docking process involves two core steps. During the first step different ligand conformations in the active site of protein are sampled [28]. To reduce computational burden, one protein is fixed in space and the second is rotated and translated around the first one [22]. Subsequently, correct docking poses are delineated from incorrect ones with a force-field based scoring function. A summary of the described process is shown in Fig. 3. Each docking run generates hundreds of possible protein conformations which are analyzed to suggest mutations for discovery and engineering of TCRs to create cancer therapeutics [64].

In this report, we present **Do**cking **R**un **I**nterpretation and **A**nnotation **T**ool (DoRIAT) a method for interpreting and annotating the output of physics based docking programs like [13, 69, 22]. DoRIAT is a Gaussian Process (GP) based regression model which associates root mean squared distance (RMSD) between TCR-pHLA docked model and the crystal structure (xtal) using geometric parameters derived from the TCR-pHLA complex. We use DoRIAT to give a favourable scores to conformations which are more likely to initiate immune response (i.e. canonical binders), identify complexes close to the resolved crystal structure and identify similar docking poses to create an ensemble of models with similar binding conformations. This enables a number of downstream use cases that are of deep interest for *in silico* guided protein engineering.

2 Methods

DoRIAT implements a Gaussian Process regressor to score EMLy™Dock models from six binding mode parameters, three angles and three displacements. [1]

Consider U TCRs for which EMLy™Dock can give us M models. For each TCR $u \in (1, \ldots, U)$ and model $m \in (1, \ldots, M)$, $\mathbf{x}_{um}$ describing the binding mode parameters and is given by:

$$\mathbf{x}_{um} = \left(\theta^c_{um}, \quad \theta^t_{um}, \quad \theta^r_{um}, \quad s^x_{um}, \quad s^y_{um}, \quad s^z_{um}\right) \tag{1}$$

[1] A Python implementation of DoRIAT along with data preprocessing and docking data are available at `https://zenodo.org/records/14763708`

A detailed description as to how these geometric parameters are obtained can be found in Sec. A.4. The associated response $\mathbf{y}_{um}$ describing the docked model's distance from native structure is modelled as a noisy observation of the function evaluated at $\mathbf{x}_{um}$, $f_{um} = f(\mathbf{x}_{um})$ through likelihood:

$$\mathbf{y}_{um}|f_{um} \sim \mathbf{N}\left(f_{um}, \sigma^2\mathbf{I}\right) \tag{2}$$

$\mathbb{P}(\mathbf{y}_{um}|f_{um})$ models the data distribution at $\mathbf{x}_{um}$ and the function f

$$f \sim \mathcal{GP}(0, \mathscr{K}) \tag{3}$$

where $\mathscr{K}$ is a positive semi-definite function that determines the covariance of f at locations x_1 and x_2 and reflects our beliefs regarding the behaviour of f. In this work we chose the Matérn $\frac{5}{2}$ kernel described by the following formula.

$$\mathscr{K}_{l,\rho}(x_1, x_2) = l^2\left(1 + \frac{\sqrt{5}\|x_1 - x_2\|_2}{\rho} + \frac{5\|x_1 - x_2\|_2^2}{3\rho^2}\right)\exp\left(-\frac{\sqrt{5}\|x_1 - x_2\|_2}{\rho}\right) \tag{4}$$

where l is the amplitude controlling the marginal variance of the function and ρ is the length-scale controlling spatial variability. We chose the Matérn $\frac{5}{2}$ kernel to model association between rmsd and the six binding mode parameters because it offers a good balance between generalization and smoothness. For more information please refer to the ablation study of Sec. A.17.

2.1 Objective function for selecting models.

Our model selection strategy is a variation of μ_{RMSD} minimisation which at the same time drastically reduces search space. More precisely, we select the best model is by:

$$\begin{aligned} &\min \mu_{RMSD} \\ &\text{sub. to } \mathrm{CV}_{RMSD} \geq \alpha_{0.95}\,, \end{aligned} \tag{5}$$

where $\alpha_{0.95}$ is the 95^{th} percentage quantile of CV_{RMSD} in the docking run. Minimizing the μ_{RMSD} is there to give the best model and CV_{RMSD} is a constrain used to reduce the search space while resolving situations where multiple models present very similar μ_{RMSD}. More information can be found in Sec. A.8.

This objective can be problematic when multiple models with identical μ_{RMSD} present a coefficient of variation above threshold. In practice, we have never encountered an example like that.

2.2 Identifying models with similar docking poses using GP covariance.

The covariance function of Eq. 3 not only encodes assumptions about the function f, but also dictates the similarity [59] between docking poses. We use that property to build a graph $\mathscr{G}(V, E)$ of vertices (V) and edges (E) representing binding mode parameters and connected poses respectively. The connection between poses is dictated by partial correlation, which measures dependence while accounting for effects of confounding conformations.

Let $f(\mathbf{X}) = \{f_{um} \forall u \in (1, \ldots, U) \text{ and } m \in (1, \ldots, M)\}$ be the set of RMSD predicted for U TCRs and M models, where f described in Eq. 3, $\mathbf{X} = \{x_{um} \forall u \in (1, \ldots, U) \text{ and } m \in (1, \ldots, M)\}$. Without loosing generality we chose TCR with index 1 and two conformations m_1 and m_3. Their predicted RMSDs are given by f_{1,m_1} and f_{1,m_3} respectively. The partial correlation between variables f_{1m_1} and f_{1m_3} is a measure of their conditional association, given the remaining elements $f(\mathbf{X})_{1,-m_1,-m_3}$ of $f(\mathbf{X})$ and it is defined as:

$$\rho_{f_{1m_1}, f_{1m_3}|f(\mathbf{X})_{1,-m_1,-m_3}} = \frac{\mathbf{Cov}\left(f_{1m_1}, f_{1m_3}|f(\mathbf{X})_{1,-m_1,-m_3}\right)}{\sqrt{\mathbf{Var}\left(f_{1m_1}|f(\mathbf{X})_{1,-m_1,-m_3}\right)}\sqrt{\mathbf{Var}\left(f_{1m_3}|f(\mathbf{X})_{1,-m_1,-m_3}\right)}} \tag{6}$$

For the purpose of this project, we assume that any two models with binding mode parameters whose partial correlation is higher than 0.05 are related. This lays the foundation for choosing similar models and building ensembles around conformations of interest.

2.3 Contact Maps

A contact map is a two-dimensional representation used to visualize amino acid interactions between proteins of interest. Each residue in the protein complex is assigned a position on the map and colours are used to indicate interactions. We consider two residues to interact if their C_A atoms are at most 8 Å apart. We use a binary representation to indicate amino acid interactions, with residues meeting our definition being labelled with 1 and 0 otherwise.

In Sec. 3 contact maps are used to asses different ensembles. We separately consider the TCR-HLA and TCR-Peptide contacts as they can give us some idea about interactions that drive affinity and specificity respectively. The contact maps of Fig. 7, 8 are created by averaging across individual contact maps of models in each ensemble.

3 Results

Table 1: **Nomenclature.** Table briefly describing parameters used to summarize TCR-pHLA complexes. Binding mode parameters are bolded. A detailed description of how these parameters are computed can be found in Sec. A.4

v_{rot}	Alpha and beta chain rotation symmetry vector pointing towards the HLA.
v_{SB}	Vector between the disulphide bonds' centroids in alpha and beta chain pointing towards beta chain.
p_{TCR}	TCR's centre of mass.
p_{MHC}	MHC's centre of mass.
v_{MHC_1}, v_{MHC_2}	Vectors defining the TCR-pHLA interface plane.
n_{MHC}	Vector perpendicular to v_{MHC_1} and v_{MHC_2} pointing towards TCR.
$\boldsymbol{\theta^c}$	Angle between v_{MHC_1} and v_{SB}.
$\boldsymbol{\theta^t}$	Angle between v_{MHC_1} and v_{rot}.
$\boldsymbol{\theta^r}$	Angle between v_{MHC_2} and v_{rot}.
$\boldsymbol{s^x}$	v_{MHC_1} direction displacement of p_{TCR} from p_{MHC} in the interface plane.
$\boldsymbol{s^y}$	v_{MHC_1} direction displacement of p_{TCR} from p_{MHC} in the interface plane.
$\boldsymbol{s^z}$	Distance of p_{TCR} projected TCR-pHLA interface plane.

DoRIAT is a tool for the exploratory analysis of TCR-pHLA structures produced by docking simulations. It aims to establish a robust docking model scoring system from six binding mode parameters characterising the relative position of the TCR to the pHLA, briefly described in Tab. 1 and more detailed in Sec. A.4. In turn, this is used for a set of downstream tasks such as identify and rank near-native conformations or detect similar conformations. In this paper we will focus on docking outcomes from EMLy™Dock, our in-house built docking pipeline, however similar approach could be used to understand simulated conformation from any docking engine.

TCRs can bind in multiple conformations to pHLA. This is also highlighted by the hundreds of models that rank well according to HADDOCK's scoring function, which serves as a proxy for affinity. When TCR-pHLA binding exceeds geometric tolerances, important cell surface interactions do not take place and signaling is disrupted [1]. We refer to docking geometries that fail to induce activation as non-canonical and we have actively tried to remove them in post-docking analysis. Our first approach for identifying and eliminating non-canonical poses was to derive thresholds from limited publicly available X-ray crystallography structures which can be found in Tab. 2. As the structures used to estimate these ranges presented significant levels of similarity and were different from the systems we were internally interested, their use created challenges in our engineering effort. To overcome the problem of setting geometric limitations, we redefine the problem as a prediction of distance to the crystal structure, where signaling is more likely. We developed DoRIAT, a GP regressor based scoring function that takes into account combinations of binding mode parameters to identify models close to the native structure, schematically described in Fig 3.

Briefly, for each TCR u, DoRIAT associates RMSD distance $\mathbf{y}_{um}$ between C_A atoms of docked model m with its corresponding crystal structure with TCR-pHLA binding mode parameters $\mathbf{x}_{um}$ using a

Table 2: **Table summarizing the cut-off thresholds for binding mode parameters.** To compute these thresholds, we consider public and proprietary TCRs the majority of which are bound to HLA-A*02.

	θ^c	θ^t	θ^r	s^x	s^y	s^z
min	19	-25	-30	-17	-7	20
max	75	25	25	12	10	30

GP f. Our kernel of choice for the GP is the Matérn with $\nu = \frac{5}{2}$ with amplitude l and length-scale ρ. As the RMSD data are initially scaled between $[0, 1]$ and subsequently inverse-transformed by the cumulative cdf of a standard normal, observation noise model is a normal distribution with mean the output of the GP f and variance σ^2. More details can be found in Sec. A.4. DoRIAT then constructs a constrained optimization problem involving the predicted RMSD and its corresponding coefficient of variation (CV) to select a model close to the native structure. DoRIAT's covariance matrix is used to identify similar docking conformations allowing for the creation of ensembles around points of interest.

DoRIAT was trained on docking outputs from 43 public TCRs and tested on 15 public and 3 proprietary TCRs, each with 600 docked conformations of TCR–pHLA complexes involving viral or cancer peptides. The model effectively distinguishes non-native binding modes that deviate strongly from canonical parameters, outperforming traditional threshold-based methods in identifying implausible complexes. However, DoRIAT struggles with cases closer to the native structure, particularly when binding parameters lie near threshold boundaries. Tests on both public and proprietary cancer-related TCR–pHLA complexes revealed that while DoRIAT sometimes underestimates RMSD, it still provides reliable indicators for filtering inaccurate models. Overall, DoRIAT offers finer-grained, more reliable post-docking analysis than threshold-based approaches. More information can be found in Sec. A.9.

3.1 DoRIAT outperforms state of the art in identifying models close to the native structure.

By benchmarking EMLy™Dock on public and internal TCR-pHLA complexes, we have identified that the median RMSD between the best simulated pose and the crystal structure is 3.1 Å. This can become particularly useful when structurally assessing candidates from our screening for therapeutics pipeline where access to a crystal structure is not available. Having a method that can suggest a docked models close to the native structure allows to cost-effectively evaluate if a candidate has engineering potential.

The comparison between different model selection approaches is performed using 18 TCR-pHLA complexes, 15 from DoRIAT's test set and 3 internal which include TCR-pHLA complex A and TCR-pHLA complex B. All these structures are bound to HLA-A*02. Additional experiments between Chai-1's aggregate score [21] and DoRIAT on internal structures are included in Sec. A.15. Internal structures are cancer related, while public structures contain a mixture of viral and cancer. Complementary results covering additional HLA alleles (A*01, B*07 and C*02) can be found in Sec. A.14.

To evaluate the effectiveness of different methods in identifying models close to the native structure, we consider, alongside DoRIAT, EMLy™Dock score (HADDOCK score in water refinement), three DeepRank variations and GNN-DOVE. HADDOCK score is a weighted sum of electrostatic, van der Waals, desolvation and distance restraints energy functions to discriminate native looking structures from the rest. Hence, the lower the score the closer to the native structure should be. For DeepRank we train a naive Graph Neural Network (naive-GNN) and DeepRank-GNN [63] to predict RMSD using the code provided in DeepRank2. For DeepRank-GNN-esm [76] and GNN-DOVE, we use the pre-trained models provided in DeepRank-GNN-esm and GNN-DOVE to score models.

All models are set to their default hyper-parameters and are trained on the same data split as DoRIAT using Adam optimizer with 5 step patience early stopping policy for a maximum of 20 steps. For the naive GNN and DeepRank-GNN the computationally intensive PSSM step is skipped.

For each of the compared methods, one docked conformation is selected based on each method's scoring approach. For DoRIAT the model is selected such that it solves Eq. 5. Model selection on DeepRank variations naive-GNN and DeepRank-GNN is performed based on predicted RMSD.

For DeepRank-GNN-esm and GNN-DOVE selection is based on fraction of native contacts and the probability that the docking decoy has a CAPRI acceptable quality respectively. For EMLy™Dock selection is based on lowest HADDOCK score. These predictions are then ranked based on their RMSD from crystal structure. Subsequently, we define the hitrate as a binary variable specific to each method which is 1 for the selected position and 0 otherwise.

Fig. 1 summarizes the cumulative hitrate as a function of ranking. The structural model that DoRIAT selects to be the closest representation of the crystal structure, out-performs compared methods by being within top 10 of docked models, ranked by RMSD, 78% of times (14 out of 18) with 4 times identifying the best model. In comparison, Naive GNN and EMLy™Dock identify models within the top 70 in 60% and 40% of cases respectively, while DeepRank-GNN, DeepRank-GNN-esm, and GNN-DOVE achieve only 22%, 17%, and 17%. More results can be found in Tab. 5. Since the docking engine effectively explores the space of conformations and creates plausible interfaces, predicting canonical models from local interactions becomes challenging. Hence, using binding mode parameters gives DoRIAT significant benefit over alternatives. DeepRank's performance further degrades due to the lack of PSSM, which was ignored as it would make unfeasible the assessment of large TCR batchess. The DeepRank-GNN-esm and GNN-DOVE have been pre-trained on antibody-antigen docking results, which differ from our TCR-pHLA systems, therefore performance is not optimal. However, if we were to retrain or fine-tune them on DoRIAT's training data we would expect them to face some of the challenges encountered by the other DeepRank variations. EMLy™Dock's scoring function is the third highest performer with the exception of TCR-pHLA complex A where it performs better even compared to DoRIAT. For the additional alleles, DoRIAT selects a model within top 20 two out of three times while the other methods struggle to get a rank less than 100.

To further assess the quality of selections, we consider DockQ score [3] which incorporates additional measures like F_{nat}, LRMS and iRMS standardized by CAPRI [47]. Tab. 3 and Tab. 6 summarize the DockQ scores for model selection in benchmarking across test set and internal TCRs. DoRIAT presents superior performance with all but one selection having at least acceptable quality and an average DockQ score of 0.42. The second best performer is the naive-GNN DeepRank variation with 5 out of 18 selected models having acceptable quality and an average DockQ score of 0.24. The rest DeepRank based models, EMLy™Dock and GNN-DOVE have 7 or fewer acceptable models and an average score of less than 0.20. Furthermore, some of the selections made by DeepRank variations, GNN-DOVE and EMLy™Dock are not guaranteed to be within canonical range.

In the left-out set, DoRIAT is sometimes outperformed by other methods based on DockQ scores, such as in TCR-pHLA complexes A, TCR-pHLA complex C, 3QFJ and 4MNQ. For the first two, EMLy™Dock's selections are slightly better. However, TCR-pHLA complex C has x-axis and y-axis displacements (s^x, s^y) of $(13.1\text{ Å}, -9.4\text{ Å})$, which is outside of canonical range. For 3QFJ and 4MNQ naive-GNN method performs best. In all these examples, the main reason other methods get better DockQ scores is attributed to better interface quality. This is expected as DoRIAT focuses on the global geometric properties of the complex. Even in these cases, DoRIAT's selections fall into the same CAPRI quality class with top performers. Overall, DoRIAT exhibits greater consistency and robustness in the quality of selections which come at an additional computing cost compared to EMLy™Dock.

In presence of larger training dataset, we would expect the disparity between DoRIAT, naive-GNN and DeepRank-GNN performance on post-docking analysis to close. However, DoRIAT alternatives would struggle as modern docking engines are effective at creating plausible interfaces even for non-canonical conformations. Hence, in this setting, binding mode parameters provide a robust alternative for building machine learning models without the need for massive amounts of data.

3.2 DoRIAT covariance can help resolve the post-docking landscape.

Fig. 1 demonstrates that DoRIAT often selects models close to the crystal structure. Beyond that, DoRIAT can also discriminate between poses, giving clear distinction between disconnected conformations and those with similar global geometry, enabling the creation of model ensembles around points of interest. This expands the limited perspective offered by crystal structures or individual models. In this section, we briefly overview how DoRIAT's covariance is used to create ensembles around points of interest using two complexes, TCR-pHLA complex A and TCR-pHLA complex B.

Table 3: **Assessment of model selections using DockQ score for test set and internal TCRs. (Part1)** Best perfroming selection is marked with bold. Cells are color-coded to indicate whether selection is canonical (green) or not (red) according to the thresholds of Tab. 2. DockQ scores are in the $[0, 1]$ range.

	DoRIAT	DeepRank (naive-GNN)	DeepRank-GNN	DeepRank-GNN-esm	GNN-DOVE	EMLy™Dock
TCR-pHLA complex A	0.34	0.16	0.1	0.20	0.12	**0.46**
TCR-pHLA complex B	**0.24**	0.05	0.06	0.04	0.07	0.11
2PYE	**0.54**	0.11	0.03	0.02	0.02	0.02
5ISZ	**0.49**	0.13	0.04	0.28	0.29	0.22
5NMG	**0.41**	0.25	0.24	0.03	0.02	0.25

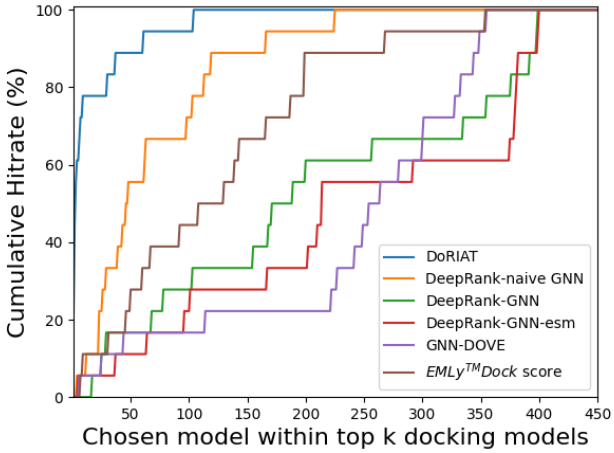


Figure 1: **Cumulative hitrate across compared methods.** Graph summarizing cumulative hitrate for DoRIAT (blue), naive GNN (orange), DeepRank-GNN (green), DeepRank-GNN-esm(red) and EMLy™Dock (purple) across unseen complexes. Each method chooses one docked model that it believes is closest to the crystal structure. These picks are then ranked using the measured RMSD from crystal structure. We set the lowest ranking to 450 as all methods reach 100% cumulative rate before this position.

Fig. 2c and 2d show PCA of binding mode parameters for complexes A and B, with edges connecting similar conformations. In both examples docking poses cluster around distinct canonical and non-canonical poses. For complex A conformations within canonical range have θ^c of about $35°$ and $45°$, while for complex B canonical conformation cluster around θ^c of about $24°$ and $65°$. In the absence of a crystal, an ensemble could be created either by considering all canonical models or by picking a model of interest and building the ensemble around it.

To illustrate how covariance information could aid in a more targeted analysis, we create two ensembles of models and compare their average contact maps to those of the crystal structure. The first ensemble includes models which are deemed canonical based on Tab. 2 thresholds. The second uses the optimal model from Eq. 5 along with similar models identified via partial correlation coefficient. Ensemble quality is assessed by comparing TCR-HLA and TCR-peptide contact maps to crystal contact maps using mean structural similarity index (SSIM) [73]. SSIM is better suited for this task, compared to other distance metrics such as mean squared error, as it accounts for contact position. Proximity to crystal contacts is more tolerable, reflecting protein flexibility, while distant deviations are penalized more heavily as they likely indicate modeling artifacts.

Fig. 7 and 8 summarize TCR-HLA and TCR-peptide contact maps for the two ensemble methods. In Fig. 7a, both the threshold and optimal model threshold introduce artificial contacts compared to the crystal structure, notably at $(\text{TCR}, \text{HLA}) = (25, 52)$. At the same time, both ensembles identify many correct contacts, resulting in SSIM scores of 0.979 and 0.975, with threshold-based ensemble being slightly better. For the TCR-epitope contact map (Fig. 8a), the optimal model ensemble

achieves an SSIM 0.815 against 0.805 for the threshold-based method, due to fewer contacts in incorrect epitope regions. Fig. 7b suggests that for complex B the ensemble around optimal model matches the crystal contacts, while the threshold-based approach presents erroneous contacts, such as $(\text{TCR}, \text{HLA}) = (26, 55), (204, 150)$. In this case, SSIM for the TCR-HLA contact map of the threshold-based ensemble is 0.970, 0.015 lower than that of the optimal model ensemble. The optimal model ensemble, for TCR-peptide contacts, misses contacts at $(\text{TCR}, \text{Peptide}) = (200, 5)$ and adds incorrect ones at $(\text{TCR}, \text{Peptide}) = (30, 3)$ (Fig. 8b). The threshold-based ensemble correctly identifies contacts at $(\text{TCR}, \text{Peptide}) = (200, 5)$, but introduces contacts in the regions $(\text{TCR}, \text{Peptide}) = (30, 3), (48, 3)$ (Fig. 8b). SSIM for the the optimal model ensemble contact map is 0.850 against 0.843 for the threshold-based ensemble.

In summary, our results suggest the covariance information can create ensembles with contact maps of equal quality or better quality than those derived from threshold-based methods on binding mode parameters. This enables more accurate interface analysis, improving downstream analysis.

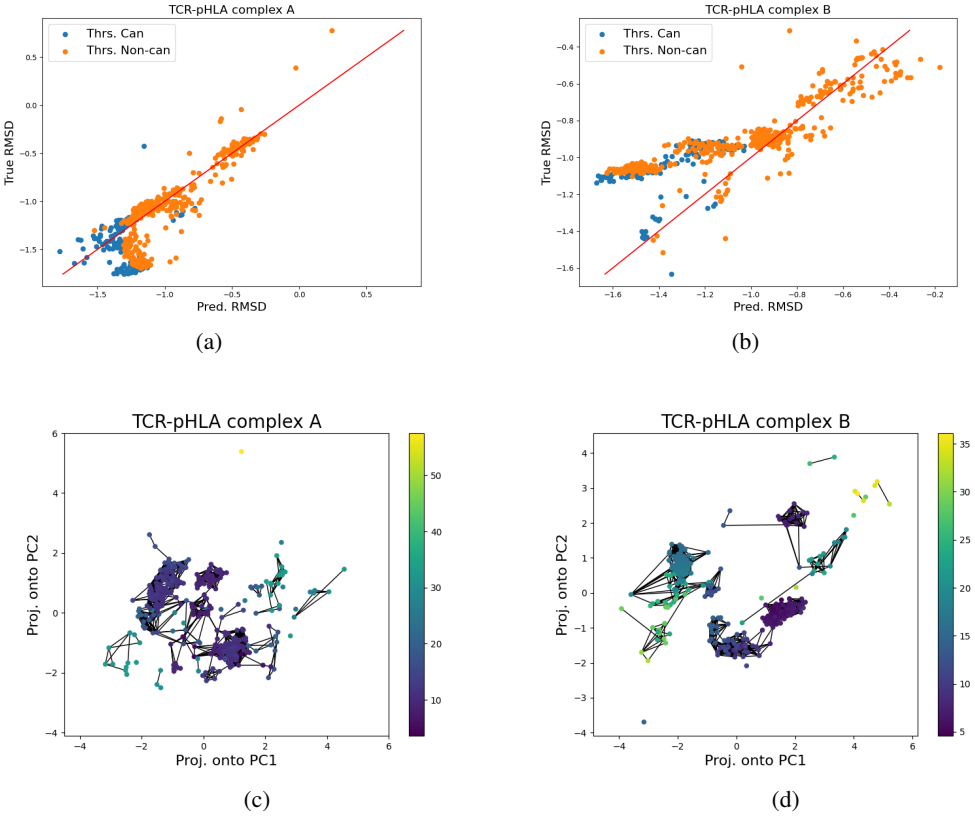


Figure 2: **Interpretation of docking runs for internal data.** Figures summarizing post-docking analysis for TCR A and TCR B. (2a,2b) Scatter plots of measured RMSD as a function of predicted RMSD for TCRs that DoRIAT performs well (TCR A) and struggles (TCR B) respectively. Each dot represents a docked model and is color-coded using cut-off thresholds of Tab. 2. The red line represents a perfect match between predicted and measured RMSD. RMSD values are scaled to 1 and inverse-transformed using normal cumulative density function. (2c,2d) PCA embedding of binding mode parameters for TCR A and TCR B, color-coded by predicted RMSD. Each dot represents a docked model. Each dot corresponds to a docked model, with edges connecting conformations that share similar binding mode parameters, as determined by partial correlation from GP covariance.

4 Conclusion

The advent of sequence-to-structure models and their combination with more traditional physics based docking platforms has opened up new possibilities for *in silico* protein engineering. However,

traditional docking engines introduce, among native conformations, a series of docking poses which are either impossible to activate the T cell or are far from the native structure observed in X-ray crystallography. We introduce DoRIAT, a Gaussian process regressor which uses binding mode parameters to interpret docking output. By considering combinations of binding mode parameters instead of a threshold-based approach, DoRIAT can reliably generalize on TCR-pHLA complexes of interest.

DoRIAT has been benchmarked in a set of tasks vital for *in silico* protein engineering. One of DoRIAT's standout features is its ability to identify models close to the crystal structure. In contrast to compared methods which heavily rely on interface contacts, DoRIAT's strategy of assessing docked conformations using binding mode parameters allows for accurate and scalable identification of models near to the crystal structure. Additionally, by leveraging DoRIAT's covariance, conformations that exhibit similar binding geometry can be selected creating an ensemble of models around points of interest. These ensembles could be used for a more thorough interface analysis of TCR-pHLA systems of interest. Overall, the described capabilities offer a cost effective solution for evaluating engineering candidates.

DoRIAT has been developed as tool for interpreting EMLy™Dock's output. Over time it has offered important insights in the effort to engineer picomolar affinity TCRs. We believe that insights provided in this report will offer an alternative way of analyzing docking results which extends beyond TCRs. A similar method to DoRIAT could be developed to analyze antibody-antigen complexes. This is going to come with its own challenges as antibodies lack co-evolutionary constraints, but their large numbers in public repositories could help the algorithm to attain reasonable predictive power.

Acknolwedgments

The authors gratefully acknowledge Kate Lawrence for proofreading the manuscript and Marc Argent for maintaining the infrastructure that supported the execution of the modeling pipeline and DoRIAT.

Lay Summary

T cell-mediated immunity is a cornerstone of the body's defense system, protecting against infections and cancer. Understanding the molecular interactions between T cell receptors (TCRs) and pHLA molecules is key to advancing therapies for cancer and autoimmune diseases. To address this challenge, we have developed a cutting-edge platform that combines deep-learning-based sequence-to-structure models with traditional docking algorithms to analyze TCR-pHLA interactions. At the heart of this effort is DoRIAT, a tool that enables the generation and interpretation of hundreds of potential protein conformations, identifying the most promising candidates for novel T cell therapies.

DoRIAT is a tool for interpreting docking results. It scores and prioritizes conformations most likely to trigger immune responses, identifies models closely resembling experimental crystal structures, and creates ensembles of binding models for more detailed analyses. This unique approach bridges the gap between static protein modeling and the dynamic reality of molecular interactions.

By leveraging this innovative combination of AI and molecular docking, we are positioned to redefine insilico protein engineering. Our solutions offer a scalable, cost-effective way to accelerate the discovery and optimization of T cell therapies, driving us toward a future of transformative treatments for cancer and autoimmune diseases.

References

[1] Jarrett J. Adams, Samanthi Narayanan, Baoyu Liu, Michael E. Birnbaum, Andrew C. Kruse, Natalie A. Bowerman, Wei Chen, Aron M. Levin, Janet M. Connolly, Cheng Zhu, David M. Kranz, and K. Christopher Garcia. T cell receptor signaling is limited by docking geometry to peptide-major histocompatibility complex. *Immunity*, 35, 2011. ISSN 10747613. doi: 10.1016/j.immuni.2011.09.013.

[2] Johan Aqvist, Victor B. Luzhkov, and Bjørn O. Brandsdal. Ligand binding affinities from md simulations. *Accounts of Chemical Research*, 35, 2002. ISSN 00014842. doi: 10.1021/ar010014p.

[3] Sankar Basu and Björn Wallner. Dockq: A quality measure for protein-protein docking models. *PLoS ONE*, 11, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0161879.

[4] C. R. Beddel, P. J. Goodford, F. E. Norrington, S. Wilkinson, and R. Wootton. Compounds designed to fit a site of known structure in human haemoglobin. *British Journal of Pharmacology*, 57, 1976. ISSN 14765381. doi: 10.1111/j.1476-5381.1976.tb07468.x.

[5] J. D. Bernal and D. Crowfoot. X-ray photographs of crystalline pepsin [3]. *Nature*, 133, 1934. ISSN 00280836. doi: 10.1038/133794b0.

[6] Frances C. Bernstein, Thomas F. Koetzle, Graheme J.B. Williams, Edgar F. Meyer, Michael D. Brice, John R. Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *Archives of Biochemistry and Biophysics*, 185, 1978. ISSN 10960384. doi: 10.1016/0003-9861(78)90204-7.

[7] C. Bissantz, G. Folkers, and D. Rognan. Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry*, 43, 2000. ISSN 00222623. doi: 10.1021/jm0010441.

[8] Hans Joachim Böhm. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3d database search programs. *Journal of Computer-Aided Molecular Design*, 12, 1998. ISSN 0920654X. doi: 10.1023/a:1007999920146.

[9] Ewen Callaway. Alphafold is running out of data — so drug firms are building their own version. *Nature*, 2025. doi: 10.1038/d41586-025-00868-9.

[10] Carlos J. Camacho and Sandor Vajda. Protein-protein association kinetics and protein docking. *Current Opinion in Structural Biology*, 12, 2002. ISSN 0959440X. doi: 10.1016/S0959-440X(02)00286-5.

[11] Heather A. Carlson and William L. Jorgensen. An extended linear response method for determining free energies of hydration. *Journal of Physical Chemistry*, 99, 1995. ISSN 00223654. doi: 10.1021/j100026a034.

[12] Paul S. Charifson, Joseph J. Corkery, Mark A. Murcko, and W. Patrick Walters. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry*, 42, 1999. ISSN 00222623. doi: 10.1021/jm990352k.

[13] Rong Chen and Zhiping Weng. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins: Structure, Function and Genetics*, 47, 2002. ISSN 08873585. doi: 10.1002/prot.10092.

[14] G. L. Clark and K. E. Corrigan. The crystal structure of insulin [8]. *Physical Review*, 40, 1932. ISSN 0031899X. doi: 10.1103/PhysRev.40.639.

[15] Charlotte H. Coles, Catriona McMurran, Angharad Lloyd, Miriam Hock, Linda Hibbert, Marine C.C. Raman, Conor Hayes, Patrick Lupardus, David K. Cole, and Stephen Harper. T cell receptor interactions with human leukocyte antigen govern indirect peptide selectivity for the cancer testis antigen mage-a4. *Journal of Biological Chemistry*, 295, 2020. ISSN 1083351X. doi: 10.1074/jbc.RA120.014016.

[16] Jason B. Cross, David C. Thompson, Brajesh K. Rai, J. Christian Baber, Kristi Yi Fan, Yongbo Hu, and Christine Humblet. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *Journal of Chemical Information and Modeling*, 49, 2009. ISSN 1549960X. doi: 10.1021/ci900056c.

[17] Biswa Nath Datta. Numerical linear algebra and applications. *Numerical Linear Algebra and Applications*, 2010. doi: 10.1137/1.9780898717655.

[18] Zbigniew Dauter and Alexander Wlodawer. Progress in protein crystallography. *Protein & Peptide Letters*, 23, 2016. ISSN 09298665. doi: 10.2174/0929866523666160106153524.

[19] Andrew M. Davis, Stephen A. St-Gallay, and Gerard J. Kleywegt. Limitations and lessons in the use of x-ray structural information in drug design. *Drug Discovery Today*, 13, 2008. ISSN 13596446. doi: 10.1016/j.drudis.2008.06.006.

[20] Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matthew D. Hoffman, and Rif A. Saurous. Tensorflow distributions. *CoRR*, abs/1711.10604, 2017. URL `http://arxiv.org/abs/1711.10604`.

[21] Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *Cold Spring Harbor Laboratory*, 2024. doi: 10.1101/2024.10.10.615955.

[22] Cyril Dominguez, Rolf Boelens, and Alexandre M.J.J. Bonvin. Haddock: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125, 2003. ISSN 00027863. doi: 10.1021/ja026939x.

[23] Miklos Feher. Consensus scoring for protein-ligand interactions. *Drug Discovery Today*, 11, 2006. ISSN 13596446. doi: 10.1016/j.drudis.2006.03.009.

[24] Miklos Feher, Eugen Deretey, and Samir Roy. Bhb: A simple knowledge-based scoring function to improve the efficiency of database screening. *Journal of Chemical Information and Computer Sciences*, 43, 2003. ISSN 00952338. doi: 10.1021/ci030006i.

[25] Daniel Fischer, Shuo Liang Lin, Haim L. Wolfson, and Ruth Nussinov. A geometry-based suite of moleculardocking processes. *Journal of Molecular Biology*, 248, 1995. ISSN 00222836. doi: 10.1016/S0022-2836(95)80063-8.

[26] Daniel K. Gehlhaar, Gennady M. Verkhivker, Paul A. Rejto, Christopher J. Sherman, David R. Fogel, Lawrence J. Fogel, and Stephan T. Freer. Molecular recognition of the inhibitor ag-1343 by hiv-1 protease: conformationally flexible docking by evolutionary programming. *Chemistry and Biology*, 2, 1995. ISSN 10745521. doi: 10.1016/1074-5521(95)90050-0.

[27] Marco Giulini, Constantin Schneider, Daniel Cutting, Nikita Desai, Charlotte M. Deane, and Alexandre M.J.J. Bonvin. Towards the accurate modelling of antibody-antigen complexes from sequence using machine learning and information-driven docking. *bioRxiv*, 2023.

[28] Inbal Halperin, Buyong Ma, Haim Wolfson, and Ruth Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function and Genetics*, 47, 2002. ISSN 08873585. doi: 10.1002/prot.10115.

[29] Marcelo Hartmann. Extending owen's integral table and a new multivariate bernoulli distribution. *arXiv preprint arXiv:1704.04736*, 2017.

[30] Richard D. Head, Mark L. Smythe, Tudor I. Oprea, Chris L. Waller, Stuart M. Green, and Garland R. Marshall. Validate: A new method for the receptor-based prediction of binding affinities of novel ligands. *Journal of the American Chemical Society*, 118, 1996. ISSN 00027863. doi: 10.1021/ja9539002.

[31] D. C. HODGKIN. The x-ray analysis of the structure of penicillin. *Advancement of science*, 6, 1949. ISSN 0001866X.

[32] Rodrigo V. Honorato, Panagiotis I. Koukos, Brian Jiménez-García, Andrei Tsaregorodtsev, Marco Verlato, Andrea Giachetti, Antonio Rosato, and Alexandre M.J.J. Bonvin. Structural biology in the clouds: The wenmr-eosc ecosystem. *Frontiers in Molecular Biosciences*, 8, 2021. ISSN 2296889X. doi: 10.3389/fmolb.2021.729513.

[33] Alexey V. Ishchenko and Eugene I. Shakhnovich. Small molecule growth 2001 (smog2001): An improved knowledge-based scoring function for protein-ligand interactions. *Journal of Medicinal Chemistry*, 45, 2002. ISSN 00222623. doi: 10.1021/jm0105833.

[34] Richard M. Jackson. Q-fit: A probabilistic method for docking molecular fragments by sampling low energy conformational space. *Journal of Computer-Aided Molecular Design*, 16, 2002. ISSN 0920654X. doi: 10.1023/A:1016307520660.

[35] Ajay N. Jain. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *Journal of Computer-Aided Molecular Design*, 10, 1996. ISSN 0920654X. doi: 10.1007/BF00124474.

[36] Abramson Josh, Adler Jonas, Dunger Jack, Evans Richard, Green Tim, Pritzel Alexander, Ronneberger Olaf, Willmore Lindsay, Ballard Andrew J, Bambrick Joshua, Bodenstein Sebastian W., David A. Evans, Hung Chia-Chun, O'Neill Michael, Reiman David, Tunyasuvunakool Kathryn, Wu Zachary, Žemgulytė Akvilė, Arvaniti Eirini, Beattie Charles, Bertolli Ottavia, Bridgland Alex, Cherepanov Alexey, Congreve Miles, Cowen-Rivers Alexander I., Cowie Andrew, Figurnov Michael, Fuchs Fabian B., Gladman Hannah, Jain Rishub, Khan Yousuf A., Low Caroline M. R., Perlin Kuba, Potapenko Anna, Savy Pascal, Singh Sukhdeep, Stecula Adrian, Thillaisundaram Ashok, Tong Catherine, Yakneen Sergei, Zhong Ellen D., Zielinski Michal, Žídek Augustin, Bapst Victor, Kohli Pushmeet, Jaderberg Max, Hassabis Demis, and Jumper John M. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630, 2024. ISSN 14764687. doi: 10.1038/s41586-024-07487-w.

[37] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596, 2021. ISSN 14764687. doi: 10.1038/s41586-021-03819-2.

[38] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, and V. C. Shore. Structure of myoglobin: A three-dimensional fourier synthesis at 2 . resolution. *Nature*, 185, 1960. ISSN 00280836. doi: 10.1038/185422a0.

[39] Douglas B. Kitchen, Hélène Decornez, John R. Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery*, 3, 2004. ISSN 14741776. doi: 10.1038/nrd1549.

[40] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research*, 33, 2000. ISSN 00014842. doi: 10.1021/ar000033j.

[41] Peter Kollman. Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews*, 93, 1993. ISSN 15206890. doi: 10.1021/cr00023a004.

[42] Irwin D. Kuntz, Jeffrey M. Blaney, Stuart J. Oatley, Robert Langridge, and Thomas E. Ferrin. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161, 1982. ISSN 00222836. doi: 10.1016/0022-2836(82)90153-X.

[43] Jinwoo Leem, Saulo H.P. De Oliveira, Konrad Krawczyk, and Charlotte M. Deane. Stcrdab: The structural t-cell receptor database. *Nucleic Acids Research*, 46, 2018. ISSN 13624962. doi: 10.1093/nar/gkx971.

[44] Marie Paule Lefranc. Unique database numbering system for immunogenetic analysis; current literature. *Immunology Today*, 18, 1997. ISSN 01675699. doi: 10.1016/S0167-5699(97)01163-8.

[45] Marie Paule Lefranc. Imgt locus on focus. a new section of experimental and clinical immunogenetics. *Experimental and Clinical Immunogenetics*, 15, 1998. ISSN 02549670. doi: 10.1159/000019049.

[46] Marie Paule Lefranc. The imgt unique numbering for immunoglobulins, t-cell receptors, and ig-like domains. *Immunologist*, 7, 1999. ISSN 11925612.

[47] Marc F. Lensink and Shoshana J. Wodak. Docking, scoring, and affinity prediction in capri. *Proteins: Structure, Function and Bioinformatics*, 81, 2013. ISSN 08873585. doi: 10.1002/prot.24428.

[48] Laurent Maveyraud and Lionel Mourey. Protein x-ray crystallography and drug discovery. *Molecules*, 25, 2020. ISSN 14203049. doi: 10.3390/molecules25051030.

[49] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. Molecular docking: A powerful approach for structure-based drug discovery. *Current Computer Aided-Drug Design*, 7, 2012. ISSN 15734099. doi: 10.2174/157340911795677602.

[50] Yilin Meng, Danial Sabri Dashti, and Adrian E. Roitberg. Computing alchemical free energy differences with hamiltonian replica exchange molecular dynamics (h-remd) simulations. *Journal of Chemical Theory and Computation*, 7, 2011. ISSN 15499618. doi: 10.1021/ct200153u.

[51] Bill R. Miller, T. Dwight McGee, Jason M. Swails, Nadine Homeyer, Holger Gohlke, and Adrian E. Roitberg. Mmpbsa.py: An efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation*, 8, 2012. ISSN 15499626. doi: 10.1021/ct300418h.

[52] John B.O. Mitchell, Roman A. Laskowski, Alexander Alex, and Janet M. Thornton. Bleep - potential of mean force describing protein-ligand interactions: I. generating potential. *Journal of Computational Chemistry*, 20, 1999. ISSN 01928651. doi: 10.1002/(SICI)1096-987X(199908)20:11<1165::AID-JCC7>3.0.CO;2-A.

[53] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil. Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *British Journal of Pharmacology*, 153, 2008. ISSN 00071188. doi: 10.1038/sj.bjp.0707515.

[54] Ingo Muegge and Yvonne C. Martin. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *Journal of Medicinal Chemistry*, 42, 1999. ISSN 00222623. doi: 10.1021/jm980536j.

[55] R. Norel, H. J. Wolfson, and R. Nussinov. Small molecule recognition: Solid angles surface representation and molecular shape complementarity. *Combinatorial Chemistry & High Throughput Screening*, 2, 1999. ISSN 13862073. doi: 10.2174/1386207302666220204193837.

[56] Raquel Norel, Daniel Fischer, Haim J. Wolfson, and Ruth Nussinov. Molecular surface recognition by a computer vision-based technique. *Protein Engineering, Design and Selection*, 7, 1994. ISSN 17410126. doi: 10.1093/protein/7.1.39.

[57] Raquel Norel, Shuo L. Lin, Haim J. Wolfson, and Ruth Nussinov. Shape complementarity at protein–protein interfaces. *Biopolymers*, 34, 1994. ISSN 10970282. doi: 10.1002/bip.360340711.

[58] Raquel Norel, Donald Petrey, Haim J. Wolfson, and Ruth Nussinov. Examination of shape complementarity in docking of unbound proteins. *Proteins: Structure, Function and Genetics*, 36, 1999. ISSN 08873585. doi: 10.1002/(SICI)1097-0134(19990815)36:3<307::AID-PROT5>3.0.CO;2-R.

[59] Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian processes for machine learning. *Gaussian Processes for Machine Learning*, 2018. doi: 10.7551/mitpress/3206.001.0001.

[60] Nicolas Renaud, Cunliang Geng, Sonja Georgievska, Francesco Ambrosetti, Lars Ridder, Dario F. Marzella, Manon F. Réau, Alexandre M.J.J. Bonvin, and Li C. Xue. Deeprank: a deep learning framework for data mining 3d protein-protein interfaces. *Nature Communications*, 12, 2021. ISSN 20411723. doi: 10.1038/s41467-021-27396-0.

[61] João P.G.L.M. Rodrigues, João M.C. Teixeira, Mikaël Trellet, and Alexandre M.J.J. Bonvin. Pdb-tools: A swiss army knife for molecular structures. *F1000Research*, 7, 2018. ISSN 20461402. doi: 10.12688/f1000research.17456.1.

[62] Markus G. Rudolph, Robyn L. Stanfield, and Ian A. Wilson. How tcrs bind mhcs, peptides, and coreceptors. *Annual Review of Immunology*, 24, 2006. ISSN 07320582. doi: 10.1146/annurev.immunol.23.021704.115658.

[63] Manon Réau, Nicolas Renaud, Li C. Xue, and Alexandre M.J.J. Bonvin. Deeprank-gnn: a graph neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*, 39, 2023. ISSN 13674811. doi: 10.1093/bioinformatics/btac759.

[64] Paul Shafer, Lauren M. Kelly, and Valentina Hoyos. Cancer therapy with tcr-engineered t cells: Current strategies, challenges, and prospects. *Frontiers in Immunology*, 13, 2022. ISSN 16643224. doi: 10.3389/fimmu.2022.835762.

[65] Graham R. Smith and Michael J.E. Sternberg. Prediction of protein-protein interactions by docking methods. *Current Opinion in Structural Biology*, 12, 2002. ISSN 0959440X. doi: 10.1016/S0959-440X(02)00285-3.

[66] Andrew J. Souers, Joel D. Leverson, Erwin R. Boghaert, Scott L. Ackler, Nathaniel D. Catron, Jun Chen, Brian D. Dayton, Hong Ding, Sari H. Enschede, Wayne J. Fairbrother, David C.S. Huang, Sarah G. Hymowitz, Sha Jin, Seong Lin Khaw, Peter J. Kovar, Lloyd T. Lam, Jackie Lee, Heather L. Maecker, Kennan C. Marsh, Kylie D. Mason, Michael J. Mitten, Paul M. Nimmer, Anatol Oleksijew, Chang H. Park, Cheol Min Park, Darren C. Phillips, Andrew W. Roberts, Deepak Sampath, John F. Seymour, Morey L. Smith, Gerard M. Sullivan, Stephen K. Tahir, Chris Tse, Michael D. Wendt, Yu Xiao, John C. Xue, Haichao Zhang, Rod A. Humerickhouse, Saul H. Rosenberg, and Steven W. Elmore. Abt-199, a potent and selective bcl-2 inhibitor, achieves antitumor activity while sparing platelets. *Nature Medicine*, 19, 2013. ISSN 10788956. doi: 10.1038/nm.3048.

[67] M. K. Teng, A. Smolyar, A. G.D. Tse, J. H. Liu, J. Liu, R. E. Hussey, S. G. Nathenson, H. C. Chang, E. L. Reinherz, and J. H. Wang. Identification of a common docking topology with substantial variation among different tcr-peptide-mhc complexes. *Current Biology*, 8, 1998. ISSN 09609822. doi: 10.1016/s0960-9822(98)70160-5.

[68] G. E. Terp, B. N. Johansen, I. T. Christensen, and F. S. Jørgensen. A new concept for multidimensional selection of ligand conformations (multiselect) and multidimensional scoring (multiscore) of protein-ligand binding affinities. *Journal of Medicinal Chemistry*, 44, 2001. ISSN 00222623. doi: 10.1021/jm001090l.

[69] Oleg Trott and Arthur J. Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31, 2010. ISSN 0192-8651. doi: 10.1002/jcc.21334.

[70] G. Verkhivker, K. Appelt, S. T. Freer, and J. E. Villafranca. Empirical free energy calculations of ligand-protein crystallographic complexes. i. knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Engineering, Design and Selection*, 8, 1995. ISSN 17410126. doi: 10.1093/protein/8.7.677.

[71] G. M. Verkhivker, D. Bouzida, D. K. Gehlhaar, P. A. Rejto, S. Arthurs, A. B. Colson, S. T. Freer, V. Larson, B. A. Lutty, T. Marrone, and P. W. Rose. Deciphering common failures in molecular docking of ligand-protein complexes. *Journal of Computer-Aided Molecular Design*, 14, 2000. ISSN 0920654X. doi: 10.1023/A:1008158231558.

[72] Xiao Wang, Sean T. Flannery, and Daisuke Kihara. Protein docking model evaluation by graph neural networks. *Frontiers in Molecular Biosciences*, 8, 2021. ISSN 2296889X. doi: 10.3389/fmolb.2021.647915.

[73] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13, 2004. ISSN 10577149. doi: 10.1109/TIP.2003.819861.

[74] Andrew Gordon Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). *Cold Spring Harbor Laboratory*, 2015. doi: 10.48550/arXiv.1503.01057.

[75] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1 democratizing biomolecular interaction modeling. *Cold Spring Harbor Laboratory*, 2025. doi: 10.1101/2024.11.19.624167.

[76] Xiaotong Xu and Alexandre M.J.J. Bonvin. Deeprank-gnn-esm: A graph neural network for scoring protein-protein models using protein language model. *Bioinformatics Advances*, 4, 2024. ISSN 26350041. doi: 10.1093/bioadv/vbad191.

[77] 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, 2023. ISSN 13624962. doi: 10.1093/nar/gkad356.

[78] Piotr Zielenkiewicz and Andrzej Rabczenko. Protein-protein recognition: Method of finding complementary surfaces of interacting proteins. *Journal of Theoretical Biology*, 111, 1984. ISSN 10958541. doi: 10.1016/S0022-5193(84)80193-9.

[79] G. C.P. Van Zundert, J. P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P. L. Kastritis, E. Karaca, A. S.J. Melquiond, M. Van Dijk, S. J. De Vries, and A. M.J.J. Bonvin. The haddock2.2 web server: User-friendly integrative modeling of biomolecular complexes. *Journal of Molecular Biology*, 428, 2016. ISSN 10898638. doi: 10.1016/j.jmb.2015.09.014.

[80] Robert W. Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *Journal of Chemical Physics*, 22, 1954. ISSN 10897690. doi: 10.1063/1.1740409.

A Appendix

A.1 Schematic representation of the pipeline

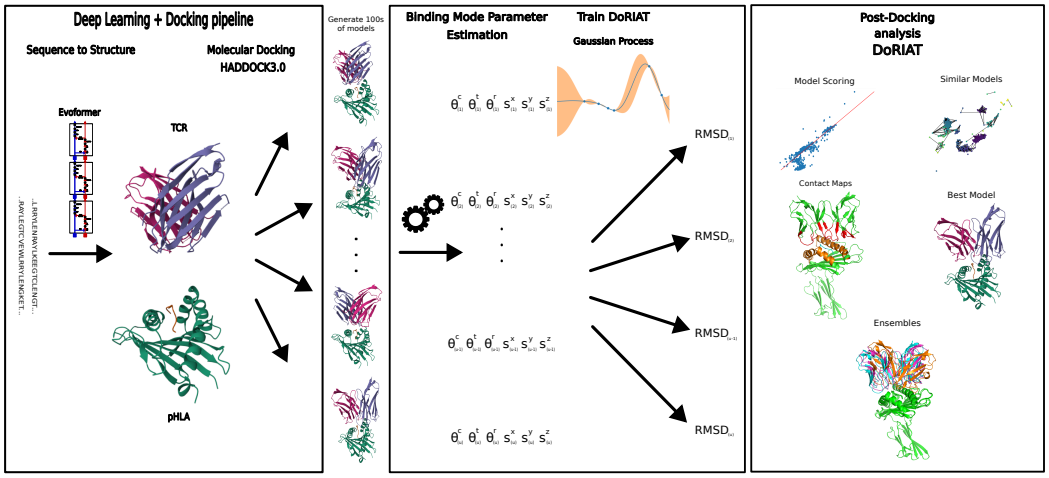


Figure 3: **Schematic representation of EMLy™Dock and DoRIAT.** In-house developed pipeline that combines deep-learning based sequence-to-structure model to generate alpha and beta chains, followed by *in silico* docking with HADDOCK3.0 [22] and DoRIAT to interpret docking outcomes and identify the best conformation.

A.2 Related Work

The binding of proteins and other molecules is fundamental to numerous biological processes. However, predicting binding affinity from structural information alone, within reasonable time, is a challenging task, particularly when experimental data is limited or unavailable. Field practitioners have developed a plethora of approaches which are either used as part of the docking process to score conformations, or post-docking to remove some of the spurious conformations that are less likely to

initiate immune response. In this section we review these two approaches explaining some of their underlying assumptions.

Docking scoring functions were developed as a proxy of protein-ligand binding affinity and can be divided into geometric matching, force-field-based, empirical and knowledge-based [39, 49].

Geometric matching between protein and ligand has repeatedly reaffirmed to have a vital role in determining the geometric complex [78, 42, 28], with early docking algorithms heavily capitalising on it [25, 56, 57, 55, 58]. However, such efforts were quickly abandoned as they struggle to generalise in absence of ground truth introducing many false positives [28].

Force-field-based scoring functions [11, 2, 41] assess the binding energy by calculating the sum of electrostatics, calculated by a Coulombic formulation and van der Waals interactions described by the Lennard-Jones potential function [49]. To address the slow computational speed of force-field-based scoring functions, distant amino acid interactions are ignored. To improve upon standard force-field functions a series of extensions involving hydrogen bonds and solvation energy are included [34, 53].

In **empirical scoring** functions [30, 35, 71, 26, 8], binding energy decomposes into energy components, such as hydrogen bond, ionic interaction, hydrophobic effect and binding entropy [49]. The contribution of each component is estimated from a set of ligand-protein complexes with measured affinity. Many modern docking engines such as [22, 13] are equipped with empirical scoring functions, accounting for a plethora of described phenomena that contribute to affinity. The main limitation of empirical scoring functions is that they might fail to generalize beyond the set of training complexes.

Knowledge-based scoring functions [54, 52, 33, 24, 70] are derived by looking into the inter-atomic contact frequencies and protein-ligand distances. Their key assumption is that favorable interactions will be more frequently observed [49]. Similar to empirical-based scoring approaches, knowledge-based ones often fail to generalize outside of their training data.

Consensus scoring [12] combines several different scores to assess the docking conformation. It has been reported [23, 12, 7, 68] that consensus scoring improves prediction of bound conformations and poses [49]. However, binding energies might still be inaccurate and its predictive power depends on the independence of scoring functions it is made of [39, 23, 12, 7, 68].

A limitation shared across the scoring functions described above is their limited treatment of solvation energy. More rigorous approaches for estimating binding affinity would include Free Energy Perturbation [80], Replica Exchange Free Energy Perturbation [50], Molecular Mechanics Poisson-Boltzmann solvent-accessible surface area (MM-PB/SA) [40]. However, they are computational demanding and can easily become prohibitively expensive for larger molecules[51]. A work-around would be to use them for scoring post-docking.

The introduction of powerful hardware accelerators like GPUs and TPUs along with parallel file system technologies has given rise to machine learning approaches. Machine learning methods are extremely flexible and in numerous cases have proved their ability to extract patterns from millions of data points. [72] have introduced Graph Neural Network–based DOcking decoy eValuation scorE (GNN-DOVE), a method that scores docking models by looking at atom chemical properties and inter-atom distances in the TCR-pHLA interface. Similar to GNN-DOVE, DeepRank and its variations [60, 63, 76] use graph neural networks (GNNs) to learn residue level features and score biomolecular complexes. Subsequently they could be used for binding affinity prediction but also to disentangle crystal artefacts from protein interactions of potential biological interest.

A.3 Information driven docking

Information-driven docking integrates experimental or computational data to improve the prediction of biomolecular interactions, enhancing accuracy and efficiency [22]. HADDOCK uses Ambiguous Interaction Restraints (AIRs) to handle incomplete or ambiguous data, enabling the modeling of flexible interactions [79]. Unlike unambiguous restraints, AIRs allow for multiple possible interactions between residue groups, making them ideal for uncertain data. This flexibility enables HADDOCK to explore a broader range of conformations while still being guided by the available data [32].

To systematically explore the interaction space, we devised three distinct docking protocols, each defining active and passive residues differently. Protocol A a more permissive approach, with only a minimal number of active residues, allowing for a less constrained exploration. In contrast, protocols

B and C were designed to be more targeted, by employing a larger set of active residues in the binding interface. This tiered approach enables a balanced exploration of the conformation space, combining the flexibility of Protocol A with the precision of Protocols B and C, thereby capturing both broad and specific aspects of the interaction landscape.

To train DoRIAT, we used all three protocols. For best model selection we choose the best model from subset of models docked with Protocols B and C, but the ranking was done on all three data. The ensembles were created from Protocol B and C models.

A.4 Binding mode parameters

Binding mode parameters are essential in assessing T cell activation and subsequent immune response. DoRIAT depends on three angles and three displacements. To estimate these parameters, 2 vectors and a centre of mass for the TCR and pHLA have to be characterized respectively. Here we describe the procedure to obtain DoRIAT features and summarize the process into a pseudo-code.

Let v_{rot} be the vector parallel to the alpha and beta chain rotation symmetry axis pointing towards the HLA, p_{TCR} be the TCR's centre of mass, and v_{SB} the vector in the direction defined by alpha and beta chain disulphide-bridges pointing towards beta chain. Let a_{alpha} and a_{beta} be the set of amino acids' carbon alpha (C_A) atomic coordinates in alpha and beta chains respectively aligned based on the IMGT numbering [44–46]. Let J be the indices of subsets of a_{alpha} and a_{beta} with common IMGT numbering and $\text{a}_{\text{alpha},J}$ and $\text{a}_{\text{beta},J}$ the relevant subsets. We consider only points with common IMGT numbering in order not to tip the centre of mass and symmetry axis towards the beta chains which is commonly longer. Let v_{TCR} be the set of midpoints between alpha and beta chain atomic C_A coordinates with the same IMGT numbering.

$$v_{\text{TCR},j} = \frac{\text{a}_{\text{alpha},J(j)} + \text{a}_{\text{beta},J(j)}}{2} \text{ for } j \text{ in } 0, \dots \|J\| - 1 \tag{7}$$

Then v_{rot} is the normalized vector pointing towards the HLA parallel to the best fit line of v_{TCR} coordinates and p_{TCR} is the mean of v_{TCR} coordinates. v_{SB} is the normalized vector parallel to the line connecting disulphide bonds in alpha and beta chains, with the vector pointing towards beta chain.

Let v_{H_1} and v_{H_2} be the normalized vectors parallel to the best fit lines running through C_A atoms' coordinates in HLA helix 1 and 2 respectively. Helices 1 and 2 are defined to be the parts of HLA with IMGT numbering ranges $52-88$ and $141-176$ respectively. To make results consistent when estimating v_{H_1} and v_{H_2} we make sure the vectors point towards the last residue in the amino acid sequence. v_{MHC_1} is the normalized bisecting vector of the parallelogram spanned by v_{H_1} and v_{H_2}. Hence,

$$v_{\text{MHC}_1} = \frac{v_{\text{H}_1} + v_{\text{H}_2}}{\|v_{\text{H}_1} + v_{\text{H}_2}\|_2} \tag{8}$$

Let p_{H1} and p_{H2} be the center of mass estimated from C_A atom coordinates of helices 1 and 2 respectively and v_{H12} the vector parallel to the line defined by points $p_{\text{H1}}, p_{\text{H2}}$ pointing towards helix 2. Then

$$v_{\text{MHC}_2} = \frac{(v_{\text{MHC}_1} \times v_{\text{H12}}) \times v_{\text{MHC}_1}}{\| (v_{\text{MHC}_1} \times v_{\text{H12}}) \times v_{\text{MHC}_1} \|} \tag{9}$$

Where $\times$ is the cross product [2]. The centre of HLA mass p_{MHC} is defined as

$$p_{\text{MHC}} = \frac{p_{\text{H1}} + p_{\text{H2}}}{2} \tag{10}$$

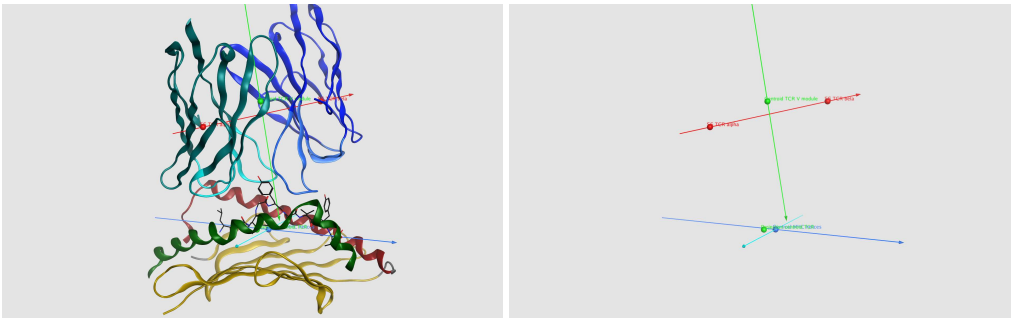
Using vectors $v_{\text{MHC}_1}, v_{\text{MHC}_2}$ and the HLA centre of mass p_{MHC} we define the HLA plane. The equation of the HLA plane for any given point k on the plane is given by :

$$n_{\text{MHC}} \cdot k = n_{\text{MHC}} \cdot p_{\text{MHC}} \,, \tag{11}$$

where

$$n_{\text{MHC}} = \frac{v_{\text{MHC}_1} \times v_{\text{MHC}_2}}{\|v_{\text{MHC}_1} \times v_{\text{MHC}_2}\|_2} \tag{12}$$

[2] An alternative way of obtaining Eq. 9 is by starting with vectors v_{MHC_1} and v_{H12} and creating an orthonormal vector basis with Gram–Schmidt process [17].



(a) (b)

Figure 4: **Vectors and points required to obtain binding mode parameters for 4FTV structure.** Visual depiction of v_{rot}(vector in light green), v_{SG}(vector in red), v_{MHC_1}(vector in blue), v_{MHC_2}(vector in cyan), p_{TCR} (top light green sphere), p_{MHC} (blue sphere) with (Fig. 4a) and without (Fig. 4b) ribbon representation of the complex overlayed. The bottom light green sphere on the HLA plane is the projection of p_{TCR} used to estimate the displacements s^x,s^y and s^z.

Given v_{rot},v_{SB}, p_{TCR} for the TCR, along with v_{MHC_1}, v_{MHC_2}, p_{MHC} for the HLA and the HLA plane described by Eq 11 we can estimate the DoRIAT parameters. Visualization of v_{rot},v_{SB}, p_{TCR}, v_{MHC_1}, v_{MHC_2} and p_{MHC} for the case of the 4FTV crystal structure can be found in Fig. 4. Let θ^c, θ^t and θ^r be the cross [62], tilt[67] and roll[67] angles. Then

$$\tilde{\theta_c} = \arccos \frac{v_{\text{MHC}_1} \cdot v_{\text{SB}}}{\|v_{\text{MHC}_1}\|_2 \|v_{\text{SB}}\|_2}$$

$$\theta^c = \begin{cases} \tilde{\theta_c}, \text{ if } \tilde{\theta_c} \geq 90^\circ \\ 360^\circ - \arccos \frac{v_{\text{MHC}_1} \cdot v_{\text{SB}}}{\|v_{\text{MHC}_1}\|_2 \|v_{\text{SB}}\|_2}, \text{ otherwise} \end{cases} \tag{13}$$

$$\theta^t = \arccos \frac{v_{\text{MHC}_1} \cdot v_{rot}}{\|v_{\text{MHC}_1}\|_2 \|v_{rot}\|_2} - 90^\circ \tag{14}$$

$$\theta^r = \arccos \frac{v_{\text{MHC}_2} \cdot v_{rot}}{\|v_{\text{MHC}_2}\|_2 \|v_{rot}\|_2} - 90^\circ \tag{15}$$

Let s^z be the z axis displacement of p_{TCR} from the plane 11. Then

$$s^z = \frac{|n_{\text{MHC}} \cdot p_{\text{TCR}} - n_{\text{MHC}} \cdot p_{\text{MHC}}|}{\|n_{\text{MHC}}\|_2} \tag{16}$$

Let s^x and s^y be the v_{MHC_1} and v_{MHC_2} axis displacements of p_{TCR} from p_{MHC} on the plane 11. Then

$$s^x = \frac{(p_{\text{TCR}} - p_{\text{MHC}}) \cdot v_{\text{MHC}_1}}{\|v_{\text{MHC}_1}\|_2} \tag{17}$$

and

$$s^y = \frac{(p_{\text{TCR}} - p_{\text{MHC}}) \cdot v_{\text{MHC}_2}}{\|v_{\text{MHC}_2}\|_2} \tag{18}$$

The procedure to estimate the binding mode parameters is summarized in Alg. 1/

A.5 Data Preprocessing

The development of DoRIAT, involved a dataset of 58 crystal structures of human TCRs obtained from the Structural T cell Receptor Database (STCRDab)[43] [3] filtered to keep only ones that bind to HLA-A*02. Using pdb-tools [61] we remove all the non-protein atoms such as water, ions, other

[3]Until November 2021.

Algorithm 1 Pseudocode for estimating binding mode parameters

Require: A PDB file with the TCR-pHLA complex atoms and their corresponding 3D coordinates.

1: Align complex amino acids based on IMGT numbering.
2: Obtain C_A atoms coordinates from alpha and beta chain with common IMGT numbering.
3: Estimate v_{TCR} from Eq. 7 and obtain v_{rot} and p_{TCR}.
4: Obtain C_A atoms in the HLA and define vectors v_{H_1}, v_{H_2} along with centre of masses p_{H_1} and p_{H_2} for helices 1 and 2.
5: Define vector v_{SB} to be parallel to the line connecting alpha and beta chain disulphide bonds with a direction towards beta chain.
6: Using Eq. 8 estimate v_{MHC_1}.
7: Define vector $v_{\text{H}_{12}}$ as the vector parallel to the line connecting p_{H_1} and p_{H_2} with direction towards helix 2.
8: Using Eq. 9 estimate v_{MHC_2}.
9: Using Eq. 10 and 12 estimate the HLA center of mass and the normal to the HLA plane defined by v_{MHC_1} and v_{MHC_2}.
10: Cross (θ^c), tilt (θ^t) and roll (θ^r) angles can be estimated from Eq. 13, 14 and 15 respectively.
11: x-axis (s^x), y-axis (s^y) and z-axis (s^z) displacements can be estimated from Eq. 17, 18 and 16 respectively.

ligands for the TCR-pHLA complex. Then alpha and beta chains were extracted into fasta format. Subsequently TCR alpha and beta chains are modeled using sequence-to-structure deep-learning based models.

Using EMLy™Dock we obtain 600 possible conformation of how the TCR-pHLA complex looks. From these docked structures, we calculate six parameters summarizing the global mode of TCR binding to the pHLA: cross, tilt, roll, shiftx, shifty, and shiftz. More information about these parameters can be found in Section A.4. To quantify the accuracy of the docked structures, we calculated the backbone distances from crystal using only C_A atoms. Crystal structures with missing residues were ignored. Subsequently, backbone distances were normalized to a range between 0 and 1 by dividing with the largest observed distance plus 5 between the docked models and corresponding crystal structures. The normalized distance ranges were then converted to $(-\infty, \infty)$ range to enable the seamless use of GP regressor from Tensorflow probability [20]. For model training and testing, we split the data based on complex id (TCR-pHLA) labels, maintaining a 3 to 1 ratio respectively.

As TCR-pHLA complexes contain four amino acid chains and the deep-learning methods examined are designed to score complexes containing only two chains, pdb files are preprocessed accordingly by merging alpha with beta chains and HLA with peptide in the described order.

A.6 Implementation

For a training set of independent variables $\mathbf{X} = \{\mathbf{x}_{um} \forall u \in (1, \ldots, U), (1, \ldots, M)\}$ and dependent variables $\mathbf{Y} = \{\mathbf{y}_{um} \forall u \in (1, \ldots, U), (1, \ldots, M)\}$. To tune the GP kernel parameters l and ρ of Eq. 4 over training data we minimise the negative log marginal likelihood

$$\text{NL}(\mathbf{Y}) = -\ln\left[\int \mathbf{N}\left(f, \sigma^2\mathbf{I}\right)\mathbf{N}\left(f|0, \mathscr{K}(\mathbf{X}, \mathbf{X})\right)df\right], \tag{19}$$

using Adam with learning rate 0.02. To minimise resource use, we have an early stopping policy applied if the negative log-likelihood drops by less than 1 unit in 10 epochs. To prevent over-fitting we tried cross validation but final parameters did not defer much.

A.7 Mathematical Derivations

DoRIAT is trained and tested with response variable lying in the $(-\infty, \infty)$ range. For predictions and model delineation dependent values have to be converted back into they original Å scale. In this section, we provide the detailed mathematical calculations required to avoid costly simulation along with the justification for using a non-standard approach to select the best model.

Suppose that DoRIAT for some input variable x^* has predicted μ^* and σ^* as latent mean and standard deviation.

Let μ_{RMSD} denote the RMSD average in observed space. To estimate μ_{RMSD} we work as follows:

$$\mu_{RMSD} = \int_{-\infty}^{\infty} \epsilon \Phi(x) \mathbf{N}(x, \mu^*, \sigma^*) dx$$

where Φ is the commulative cdf of standard normal and ϵ is the maximum observed RMSD we could observe. Using [29] we get the following result.

$$\mu_{RMSD} = \int_{-\infty}^{\infty} \epsilon \Phi(x) \mathbf{N}(x, \mu^*, \sigma^*) dx = \epsilon \Phi \left(\frac{\mu^*}{\sqrt{1 + \sigma^{*2}}} \right) \tag{20}$$

Let σ_{RMSD} denote the RMSD standard deviation in observed space. To estimate σ_{RMSD} we work as follows:

$$\sigma_{RMSD} = \sqrt{\int_{-\infty}^{\infty} \epsilon^2 \Phi(x)^2 \mathbf{N}(x, \mu^*, \sigma^*) dx - \mu_{RMSD}^2}$$

Again using [29] we get:

$$\int_{-\infty}^{\infty} \epsilon^2 \Phi(x)^2 \mathbf{N}(x, \mu^*, \sigma^*) dx = \epsilon^2 \mathbf{F} \left(\vec{m} | \vec{0}, V \right)$$

where $\mathbf{F}$ is the bivariate normal cdf with mean $\vec{0}$ and variance V and

$$V = \begin{bmatrix} \sigma^2 + 1 & \sigma^2 \\ \sigma^2 & \sigma^2 + 1 \end{bmatrix}, \vec{m} = \begin{bmatrix} \mu^* \\ \mu^* \end{bmatrix}, \vec{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Hence:

$$\sigma_{RMSD} = \epsilon \sqrt{\mathbf{F} \left(\vec{m} | \vec{0}, V \right) - \Phi \left(\frac{\mu^*}{\sqrt{1 + \sigma^{*2}}} \right)^2} \tag{21}$$

So, CV is estimated by:

$$\text{CV}_{RMSD} = 100 \frac{\sigma_{RMSD}}{\mu_{RMSD}} \tag{22}$$

Predicted entropy E_{RMSD} is given by the following formula:

$$E_{RMSD} = -\int_{-\infty}^{\infty} \Phi(x) \ln \left[\Phi(x) \right] \mathbf{N}(x, \mu^*, \sigma^*) dx$$

Since there is not closed form solution to the integral we first need to approximate $-\Phi(x) \ln \left[\Phi(x) \right]$. A very good approximation can be obtained by Laplace approximation. By doing so we find that the mean is $\mu_L = \Phi^{-1} \left(\frac{1}{e} \right)$ and standard deviation $\sigma_L = e\mathbf{N} \left(\Phi^{-1} \left(\frac{1}{e} \right), 0, 1 \right)$. Using numerical integration on $-\Phi(x) \ln \left[\Phi(x) \right]$ we see that $\int_{-\infty}^{\infty} -\Phi(x) \ln \left[\Phi(x) \right] dx = 0.90319728556$. Hence $c\mathbf{N}(x, \mu_L, \sigma_L)$ is a really good approximation of $-\Phi(x) \ln \left[\Phi(x) \right]$. Hence:

$$E_{RMSD} \approx \int_{-\infty}^{\infty} c\mathbf{N}(x, \mu_L, \sigma_L) \mathbf{N}(x, \mu^*, \sigma^*) dx$$

$$E_{RMSD} \approx c\mathbf{N}(\mu^*, \mu_L, \sqrt{\sigma^{*2} + \sigma_L^2}) \tag{23}$$

A.8 Objective function justification

In this section we attempt to provide some additional insights as to why we chose Eq. 5 as our objective function.

To construct an objective function that selects good models we start by examining μ_{RMSD}. Eq. 20 and Fig.5a indicate that if were to chose models based on μ^* and σ^*, we would have a hard time as different combinations will yield the same results. Hence, it is important for our objective function to break such ties.

To resolve ties in model selection, we prioritize models with a high CV_{RMSD}. In the face of uncertainty, we are making the optimistic assumption that we are overestimating μ_{RMSD}. In practice we have observed that this assumption works well, as models close to the crystal structure present

higher coefficient of variation compared to the remaining ones. By applying a constraint that retains the top 5% of models in terms of CV_{RMSD}, we effectively filter out less favourable conformations. Then by selecting the model with the lowest μ_{RMSD} among models within that subset we increase the chance of getting a desirable results.

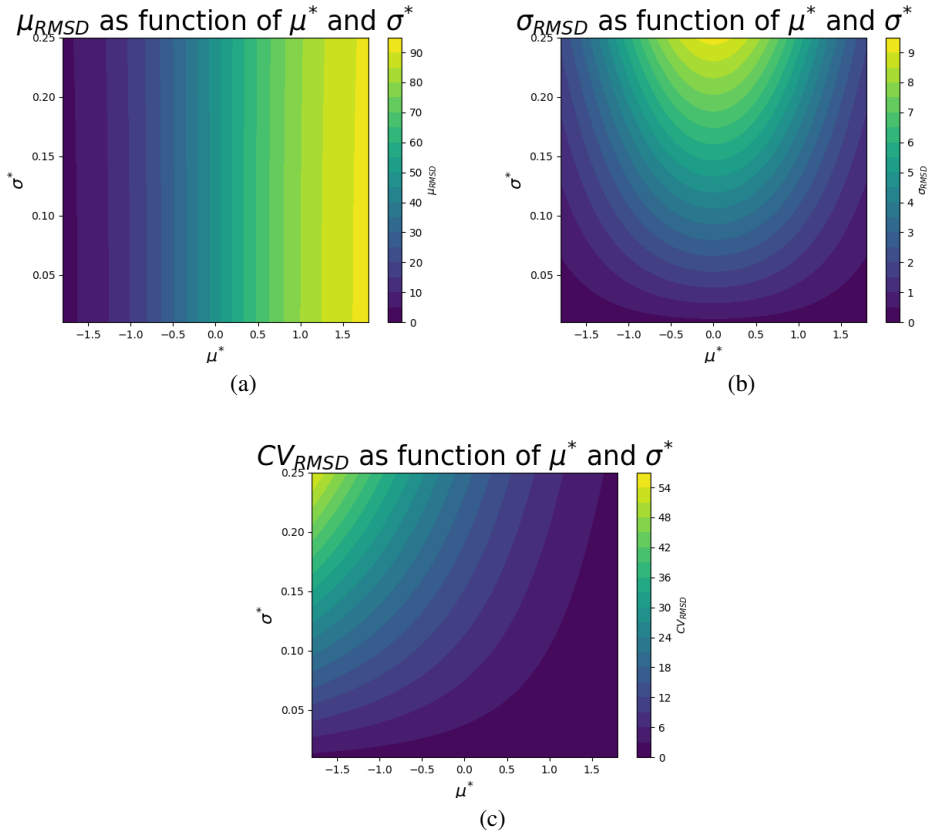


Figure 5: **Contour plots of Eq. 20,21 and 22** Contour plots of predicted mean (μ_{RMSD}, Fig 5a), standard-deviation (σ_{RMSD}, Fig 5b) and coefficient of variation (CV_{RMSD}, Fig 5c) as a function of parameters μ^* and σ^*. Thresholds for μ^* and σ^* are estimated from test set observations.

A.9 DoRIAT accurately scores output models of docking run.

DoRIAT is trained on docking output of 43 TCRs and subsequently tested on 15 TCRs sourced from STCRDab [43]. For each TCR the docking platform generates 600 possible conformations of the TCR-pHLA complex. Both training and testing data comprise of a mixture of complexes of a TCR bound to viral or cancer peptides presented by HLA-A*02. For testing we also include proprietary data.

Fig. 6 summarizes the scoring results for docked conformations of HIV (5NMG), viral (5ISZ) and cancer (2PYE) related systems from the test set.

DoRIAT accurately predicts docked models far from the native structure, where binding mode parameters range outside the canonical range and can be identified with a thresholding approach as indicated by the orange and blue color-coding. DoRIAT struggles with models closer to the native structure, as indicated by larger deviations in Fig. 6c, 6b, 2a and 2b. The common trait across conformations where DoRIAT struggles is that they present binding mode parameters on the edge of what threshold-based approach would deem canonical.

For TCR-pHLA systems like 2PYE and 5ISZ, where DoRIAT's predictions differ from the ground truth, the native complex parameters are listed in the first two lines of Tab 4. For 2PYE and 5ISZ, models like 2PYE_m1 and 5ISZ_m1 have parameters θ^t and θ^c close to thresholds presented in Tab

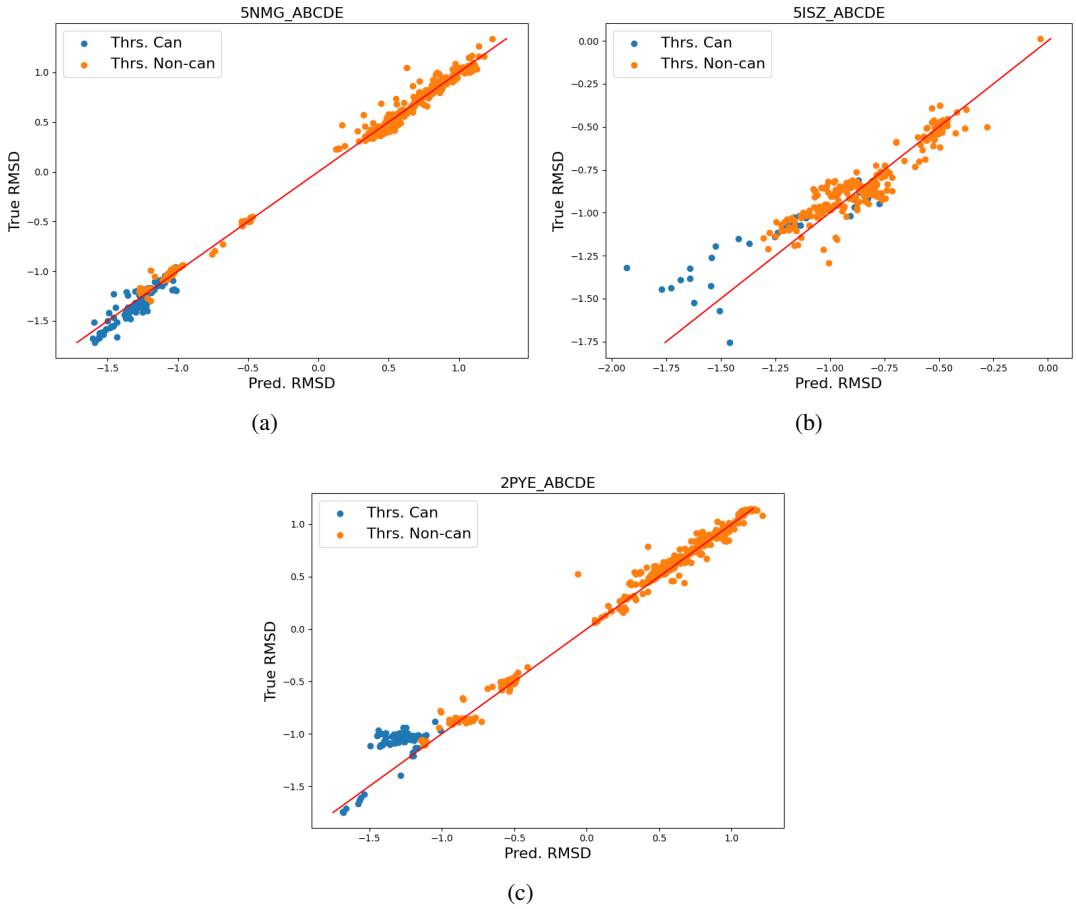


Figure 6: **Measured RMSD as a function of predicted RMSD for three TCRs in the test set.** Measured RMSD as a function of predicted RMSD for crystal structures 5NMG (Fig. 6a), 5ISZ (Fig. 6b) and 2PYE (Fig. 6c). Each dot represents a docked model and is color-coded using cut-off thresholds of Tab. 2. The red line corresponds to perfect match between predicted and measured RMSD. RMSD values are scaled to 1 and inverse-transformed using normal cumulative density function.

2 and highlight challenges for threshold-based method. In absence of ground truth the presented models could be considered plausible TCR-pHLA complexes, even though their backbone distance from their corresponding native structures is greater than 11.5 Å.While DoRIAT underestimates the models' distance from native structure, it is still able to suggest that all considered models are more than 10 Å away from crystal structure, unlike the threshold based approach which would incorrectly retain them.

DoRIAT was also applied to docking simulations of unpublished TCR-pHLA complexes, which are more heterogeneous than public data and intended for cancer therapeutics. Their binding mode parameters can be found in rows 3 and 4 of Tab. 4. A direct consequence of the inherent complexity is to have conformations with parameters within range, whose backbone is far from crystal structure causing DoRIAT to struggle with accurate RMSD prediction. TCR A's native structure has (θ^c, θ^r) angles of $(40.56°, 10.76°)$, while TCR-pHLA complex A_m1 and TCR-pHLA complex A_m2 bind at $(22.75°, -13.67°)$ and $(70.17°, -16.12°)$. For these examples, the discrepancy between DoRIAT's prediction and actual RMSD would be 5 Å and 3 Å respectively. However, these models would not affect downstream analysis as they would easily be flagged for removal due to their high predicted RMSD (19 Å and 7 Å respectively). TCR B's listed in Tab. 4 (row 4) is one of the most difficult TCRs to make predictions. TCR-pHLB complex B_m1 and TCR-pHLA complex B_m2 are

Table 4: **Binding mode parameters for section's examined examples.** Summary of the binding poses for native complexes (first four lines) and docking outcomes with canonical parameters that represent structures distant from the ground truth (remaining six lines).

	θ^c	θ^t	θ^r	s^x	s^y	s^z
2PYE	66.73	8.21	-17.03	4.39	-4.96	24.74
5ISZ	56.33	2.66	-7.34	3.34	2.71	24.53
TCR A	40.56	6.40	10.76	0.72	-5.18	23.06
TCR B	65.67	-3.74	-8.53	2.87	-0.29	25.29
2PYE_m1	41.31	23.97	8.88	4.66	1.12	22.71
5ISZ_m1	21.75	19.46	5.81	11.48	-1.04	22.14
TCR-pHLA complex A_m1	22.75	22.49	-13.67	9.75	-6.61	23.08
TCR-pHLA complex A_m2	70.17	-4.87	-16.12	6.02	-6.93	23.59
TCR-pHLA complex B_m1	21.55	7.06	9.32	6.69	-0.53	24.08
TCR-pHLA complex B_m2	45.56	-18.38	-6.85	-2.42	-6.46	24.85

part of a bigger trend presented in Fig. 2b, where DoRIAT underestimates the RMSD from crystal structure as binding mode parameters are far from native structure, but within range.

By jointly evaluating binding mode parameters, DoRIAT is more effective than the threshold-based method, offering higher control over the granularity of the post-docking analysis. However, DoRIAT's training on limited data creates unique challenges in subsequent tasks like finding a model close to the native structure. In Sec. 3.1 , we outline an approach that leverages the model to address this challenge.

A.10 Cumulative hitrate across compared methods

In Sec. 3.1 we select a model from the docking output using different scoring approaches. Here we present a table summarizing the cumulative hitrate across compared methods.

Table 5: **Table summarizing cumulative hitrate across compared methods.** Table summarizing cumulative hitrate for DoRIAT, naive GNN, DeepRank-GNN, DeepRank-GNN-esm and EMLy™Dock across unseen complexes. Each method chooses one docked model that believes it is closest to the crystal structure. These picks are then ranked using the measured RMSD from crystal structure.

Hitrate / Top	1	10	20	40	70	120	170	250
DoRIAT	22.22	77.78	77.78	88.89	94.44	100.00	100.00	100.00
DeepRank (naive-GNN)	0.00	5.56	11.11	38.89	66.67	88.89	94.44	100.00
DeepRank-GNN	0.00	0.00	5.56	16.67	22.22	33.33	44.44	61.11
DeepRank-GNN-esm	0.00	5.56	5.56	11.11	16.67	27.78	33.33	55.56
GNN-DOVE	0.00	5.56	5.56	11.11	16.67	22.22	22.22	44.44
EMLy™Dock	0.00	11.11	11.11	16.67	38.89	50.00	72.22	88.89

A.11 Benchmarking model selections with DockQ

In Sec. 3.1 we select a model from the docking output using different scoring approaches. Apart from ranking predictions based on RMSD, we also compare them using DockQ score. Since DockQ takes into account interface properties apart from ligand RMSD, it provides a more holistic assessment of the docking pose. Here we provide DockQ scores from the remaining test set and internal TCRs.

Table 6: **Assessment of model selections using DockQ score for test set and internal TCRs. (Part2)** Best perfroming selection is marked with bold. Cells are color-coded to indicate whether selection is canonical (green) or not (red) according to the thresholds of Tab. 2. DockQ scores are in the $[0, 1]$ range.

	DoRIAT	DeepRank (naive-GNN)	DeepRank-GNN	DeepRank-GNN-esm	GNN-DOVE	EMLy™Dock
TCR-pHLA complex C	0.24	0.16	0.02	0.03	0.04	**0.29**
1QRN	**0.54**	0.51	0.02	0.02	0.03	0.42
2BNQ	**0.20**	0.09	0.01	0.02	0.04	0.06
2VLJ	**0.52**	0.21	0.05	0.02	0.20	0.06
3QFJ	0.51	**0.60**	0.10	0.58	0.56	0.59
4MNQ	0.54	**0.57**	0.38	0.03	0.03	0.03
5C07	**0.37**	0.26	0.03	0.03	0.02	0.03
5COA	**0.32**	0.1	0.11	0.02	0.02	0.05
5C0B	**0.41**	0.19	0.03	0.02	0.03	0.04
5C0C	**0.25**	0.09	0.03	0.03	0.03	0.03
5EU6	**0.41**	0.17	0.19	0.1	0.09	0.03
5MEN	**0.64**	0.10	0.43	0.02	0.02	0.28
5YXU	**0.64**	0.62	0.06	0.58	0.5	0.54

A.12 Contact maps of DoRIAT's ensembles.

To complement the analysis presented in the main text, we include here the detailed contact map comparisons between model ensembles and the corresponding crystal structures.

The first set of figures shows average contact maps for the canonical ensemble, defined by the thresholds in Tab. 2. The second set corresponds to the covariance-based ensemble, which includes the optimal model from Eq. 5 and structurally similar models identified via partial correlation analysis.

For each ensemble, TCR–HLA and TCR–peptide contact maps are compared to those of the crystal structure. The mean Structural Similarity Index (SSIM) values reported in the main text summarize the overall agreement, while the images below visually illustrate the spatial distribution of contact deviations. Warmer regions indicate greater correspondence with experimental contacts, whereas cooler regions highlight structural discrepancies or flexible regions within the modeled complexes.

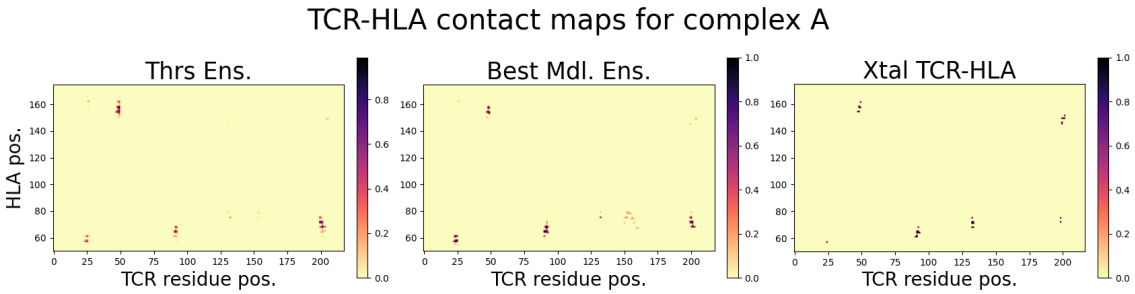
A.13 Empirical estimation of DoRIAT's computational complexity

To assess the empirical scaling behavior of DoRIAT, we conducted an experiment using six synthetic datasets of increasing size. Each dataset consisted of binding mode parameters of docked models derived from 15, 29, 43, 58, 67, and 70 TCR-pHLA complexes respectively. For datasets with more than 70 complexes, we observed that training could not be completed due to memory constraints on our hardware setup. This provides a practical reference point for the upper bound of tractable model size under our current compute availability, since exact GP inference and storage scales cubically and quadratically with the number of training examples respectively.

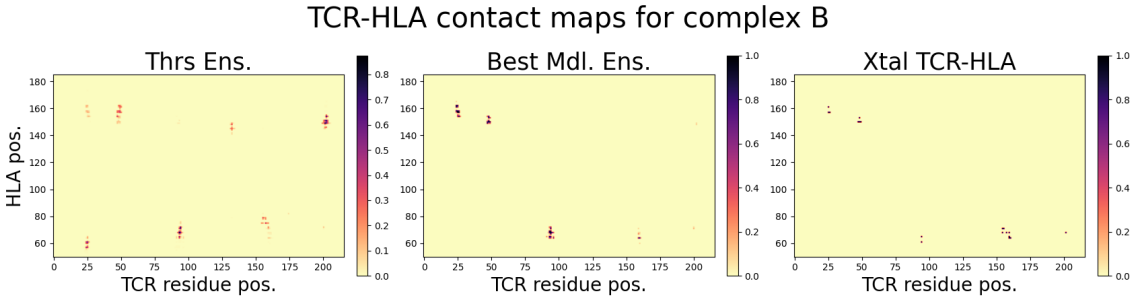
Each complex contributed 600 docked models, resulting in total dataset sizes ranging from $9,000$ to $42,000$ samples. For inference we used an NVIDIA A100 with 80Gb of vram and CUDA version 12.2. All datasets were split into training and testing subsets in a 3 to 1 ratio. DoRIAT was trained on the training portion of each dataset, and the wall-clock time required to fit the model was recorded.

Our empirical observations, summarized in Fig. 9 indicate that the training time scales approximately as $O(n^{2.8})$, where n is the number of training complexes. This scaling is consistent with the expected cubic complexity of exact GP inference, albeit with a slightly sub-cubic exponent, potentially due to implementation-level optimizations of Tensorflow.

A principled strategy to extend the scalability of the model is to employ Structured Kernel Interpolation (SKI) [74], which replaces the full kernel computation with an interpolation scheme defined over a grid of inducing points. In our setting, this approach is particularly well-suited to the structure of the input space. Three of the six geometric descriptors correspond to angular parameters, each confined to a biologically meaningful and relatively narrow domain. The remaining translational

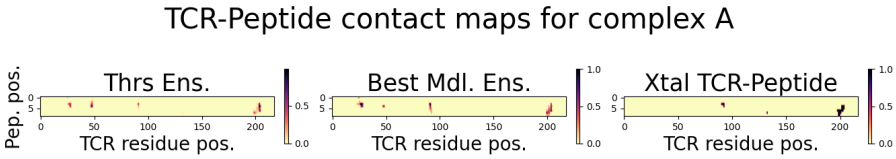


(a)

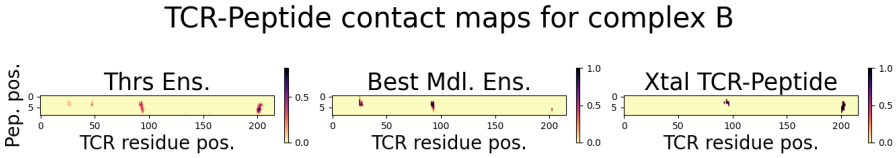


(b)

Figure 7: **Contact maps between TCR-HLA $(8\ \text{Å})$ for different ensembles for complexes A and B.** Summary of contact maps between the TCR-HLA for complexes A (Fig. 7a) and B (Fig. 7b) for ensembles created using canonical models based on thresholds and similar models around prediction made with Eq. 5. Here a contact is defined if two amino acids have a distance of less than 8 Å. Each contact map is created by averaging contact maps of models in each ensemble. Positions on the contact map are colored based on the average contacts between TCR and HLA. For comparison we include the contact maps observed in the crystal structure.



(a)



(b)

Figure 8: **Contact maps between TCR-peptide $(8\ \text{Å})$ for different ensembles for complexes A and B.** Summary of contact maps between the TCR-peptide for complexes A (Fig. 8a) and B (Fig. 8b) for ensembles created using canonical models based on thresholds and similar models around prediction made with Eq. 5. Here a contact is defined if two amino acids have a distance of less than 8 Å. Each contact map is created by averaging contact maps of models in each ensemble. Positions on the contact map are colored based on the average contacts between TCR and peptide. For comparison we include the contact maps observed in the crystal structure.

parameters, while theoretically unbounded, are in practice restricted by the finite physical dimensions of the pHLA surface. Together, these constraints imply that the joint input space occupies a compact and well-structured region of $\mathbb{R}^6$, making it amenable to grid-based interpolation.

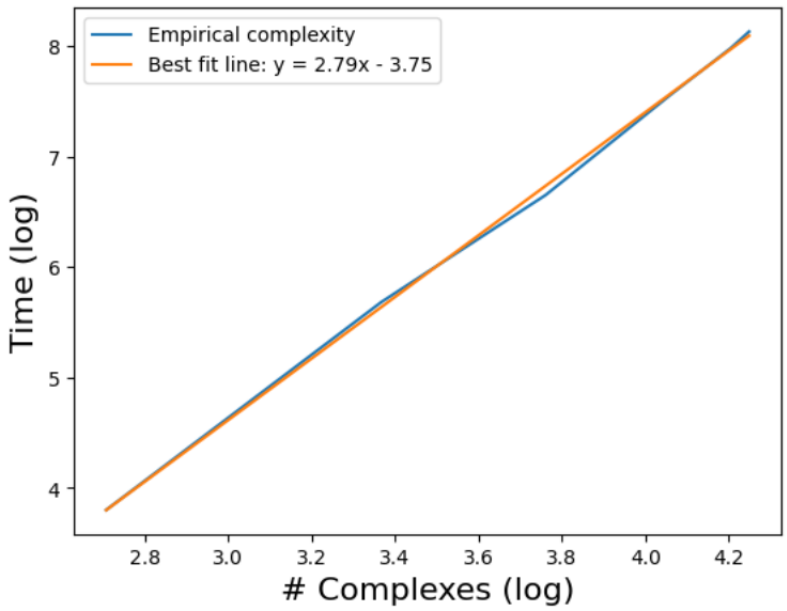


Figure 9: **DoRIAT's inference complexity from empirical experimentation**

A.14 Assessing DoRIAT's Generalization Across HLA Alleles

All the benchmarking presented in the main section heavily focused on TCR–pHLA complexes restricted to the HLA-A*02 allele, which constitutes the majority of publicly available X-ray crys-

tallography structures. However, to evaluate whether DoRIAT's performance is allele-specific or generalizable across distinct HLA backgrounds, we extended our analysis to include representative examples from the three major HLA classes. Specifically, we selected crystallographically resolved TCR–pHLA complexes from STCRDab with identifiers 5BRZ (HLA-A*01), 6AVG (HLA-B*07), and 8SHI (HLA-C*06). These structures represent allelic variants with distinct binding grooves, anchor residue preferences, and TCR recognition topologies, making them a suitable test bed for assessing DoRIAT's robustness beyond the HLA-A*02 setting.

For each complex, we generated 600 docked models using the EMLy™Dock pipeline, adhering to the docking protocols described in Sec. A.3. Due to conference deadline, the full computational pipeline, particularly the modules designed for more effective exploration were not used, limiting the breadth of models examined. This is reflected on the DockQ scores which tend to be lower across all complexes compared to the ones presented in the main section.

Each docking run was subsequently analyzed with DoRIAT, which assigned scores to the docked models based on combinations of binding mode parameters and selected conformations predicted to be closest to the native crystal structure according to the optimization criterion defined in Eq. 5. This evaluation mirrors the procedure used for HLA-A*02 complexes, ensuring comparability of performance metrics across different alleles. To contextualize DoRIAT's relative to alternative scoring methods, we performed a systematic comparison against four alternative post-docking scoring frameworks; naive-GNN and DeepRank-GNN trained to predict RMSD as well as pretrained DeepRank-GNN-esm and GNN-DOVE. For each method, we assessed the rank of the selected model and compared DockQ scores, following the analysis described in Sec 3.1.

Tab. 7 evaluates each method's ability to select a docked structure that is closest to the true crystal structure, using RMSD as the metric with lower ranks corresponding to more accurate selections. Across all complexes shown (HLA-A*01, HLA-B*07, HLA-C*06), DoRIAT obtains markedly better ranks than all alternative methods, including naïve-GNN, DeepRank-GNN, GNN-DOVE, and EMLy™Dock. The differences are especially pronounced in complexes like HLA-A*01 and HLA-B*07, where competing methods select models far from the crystal structure, while DoRIAT selects a candidate within top 20 for two out of three structures considered. Even for the more challenging HLA-C*06 complex, where all methods struggle more, DoRIAT's rank (34) is substantially better than the others which pick models at ranks $116\tilde{}561$. These results demonstrate that DoRIAT is able to select better docked conformations across a range of HLA molecules in addition to HLA-A*02.

A complementary evaluation, measured via DockQ focusing on the quality of the docked structure that each method selects is shown in Tab. 8. For all complexes analyzed, the structures selected by DoRIAT achieve the highest DockQ scores among all methods (0.22, 0.21, 0.15 for HLA-A*01, HLA-B*07 and HLA-C*06 respectively), and crucially, every DoRIAT-selected structure lies within the canonical quality range. Alternative methods produce much lower DockQ scores (in most cases between $0.01\tilde{}0.04$), reflecting selections that may show some interface interaction but are far from correct geometry. For HLA-A*01 EMLy™Dock and DeepRank-GNN pick comparatively better structures (DockQ 0.10) but still trail DoRIAT while on the remaining structures selections tend to be poor.

Taken together, the RMSD and DockQ results highlight DoRIAT's ability to select structures that are simultaneously close to the true crystal geometry and exhibit reasonable interface topology. While alternative methods occasionally perform adequately on one metric for a given complex, none achieve comparable performance across both evaluations. DoRIAT is the only method that consistently identifies near-native and canonical-quality structures across the test set, underscoring the robustness of its scoring mechanism for all HLA alleles in addition to HLA-A*02.

Table 7: **Ranking of selected model for different complexes and methods.** Table summarizing ranking of selected models for DoRIAT, naive GNN, DeepRank-GNN, DeepRank-GNN-esm and EMLy™Dock across unseen complexes. Each method chooses one docked model that believes it is closest to the crystal structure. These picks are then ranked using the measured RMSD from crystal structure.

Rank	DoRIAT	DeepRank (naive-GNN)	DeepRank-GNN	DeepRank-GNN-esm	GNN-DOVE	EMLy™Dock
5BRZ	17	599	178	417	309	116
6AVG	14	570	570	560	564	235
8SHI	34	285	429	233	481	561

Table 8: **Assessment of model selections using DockQ score for selected alleles outside HLA-A*02.** Best performing selection is marked with bold. Cells are color-coded to indicate whether selection is canonical (green) or not (red) according to the thresholds of Tab. 2. DockQ scores are in the $[0, 1]$ range.

	DoRIAT	DeepRank (naive-GNN)	DeepRank-GNN	DeepRank-GNN-esm	GNN-DOVE	EMLy™Dock
5BRZ	0.22	0.03	0.10	0.03	0.04	0.10
6AVG	0.21	0.03	0.03	0.01	0.03	0.03
8SHI	0.15	0.07	0.04	0.04	0.03	0.01

A.15 Comparison between DoRIAT and Chai-1's aggregate score on unseen TCR-pHLA Complexes

Deep learning-based methods have transformed the field of protein complex modeling by learning patterns directly from large datasets containing crystal structures. Notably, seqeunce-to-structure models have demonstrated strong performance on benchmark datasets [36, 21, 75]. At this stage however, their ability to generalize to novel protein complexes, particularly those of pharmaceutical interest remains uncertain [9]. In such settings, confidence metrics like Chai-1's aggregate score[4], which is learned from static structural data, may not be sufficient to guide the selection of near-native docking poses. In contrast, post-docking, task-specific machine learning methods like DoRIAT are explicitly trained to incorporate structural and biophysical constraints, potentially offering more accurate identification of physically realistic models.

To evaluate the effectiveness of DoRIAT in selecting good binding poses, we consider TCR-pHLA docking runs of three TCRs intended for therapeutic use and for each TCR, we choose the best 4 docked complexes according to Eq. 5. Subsequently, we select 4 complexes produced by Chai-1, a state of the art sequence-to-structure deep learning model, with the highest aggregate score. The comparison is performed on average DockQ, RMSD and iRMSD values.

As shown in Tab. 9, DoRIAT selections consistently yielded higher DockQ scores and lower RMSD and iRMSD values, compared to Chai-1's aggregate score, across all three complexes. For instance, TCR-pHLA complex A achieved an average DockQ of 0.37, RMSD of 5.75 Å, and iRMSD of 2.99 Å. In comparison, as presented in Tab. 10, the models selected by Chai-1 showed lower average DockQ scores and higher RMSD and iRMSD values. For TCR-pHLA complex A, Chai-1 produced a DockQ of 0.34, RMSD of 7.77 Å, and iRMSD of 3.47 Å. The differences were even more pronounced for TCR-pHLA complexes B and C, where DoRIAT-selected models had significantly better structural quality metrics than those from Chai-1. These results highlight the superior performance of EMLy™Dock and DoRIAT in selecting accurate docking models, particularly in capturing near-native TCR–pHLA conformations.

Our comparison highlights the limitations of relying solely on intrinsic confidence metrics like Chai-1's aggregate score for selecting accurate protein complex models, especially in cases involving

[4]Chai-1 offers a high-level number representing the overall quality of the prediction. For more information the reader is directed to `https://neurosnap.ai/blog/post/interpreting-chai-1-alphafold3-metrics-and-visualizations-on-neurosnap/67943bec1d1c9fa86479c694`

structurally diverse or unseen sequences such as TCR-pHLA interfaces. While aggregate score offers a useful internal measure of structural uncertainty, it is not specifically optimized for the docking selection task. In contrast, DoRIAT leverages geometric parameters tailored to the TCR-pHLA docking runs, demonstrating superior performance in identifying near-native conformations across multiple complexes.

Table 9: **Average DockQ, RMSD and iRMSD values of the top 4 models produced by EMLy™Dock and selected by DoRIAT.**

EMLy™Dock + DoRIAT	DockQ	RMSD	iRMSD
TCR-pHLA complex A	0.37	5.75	2.99
TCR-pHLA complex B	0.20	7.53	3.32
TCR-pHLA complex C	0.22	10.10	2.73

Table 10: **Average DockQ, RMSD and iRMSD values of the top 4 models produced by Chai-1.**

Chai-1	DockQ	RMSD	iRMSD
TCR-pHLA complex A	0.34	7.77	3.47
TCR-pHLA complex B	0.09	18.19	6.17
TCR-pHLA complex C	0.10	16.91	6.13

A.16 DoRIAT's Robustness Across Data Splits

To assess the robustness and generalization of DoRIAT, we performed 5 train–test splits on the dataset described in Sec. 3. Each split consists of 43 TCR-pHLA complexes docking sets for training and 15 for testing, maintaining a balanced representation of viral and cancer-derived pHLA complexes.

To rigorously evaluate the robustness and generalization capacity of DoRIAT, we conducted 5 independent train–test splits on the full dataset described in Sec. 3. Similar to the split described in the results section the split was done using a 3 to 1 ratio. This procedure ensures that model performance is not biased by a particular data partition and provides a more comprehensive view of DoRIAT's behavior under different training conditions.

For each split, DoRIAT was trained from scratch and evaluated on the corresponding held-out test data. Model behavior was monitored through the evolution of the negative log-likelihood (NLL) during training, which consistently demonstrated stable convergence. To assess predictive accuracy, the predicted mean RMSD values were compared against the true RMSD between the docked and experimental TCR–pHLA structures, revealing a strong linear correlation across all splits. Furthermore, analysis of the Gaussian Process (GP) predictive uncertainty as a function of the predicted mean showed that the model reliably assigned higher uncertainty to less confident predictions, reflecting appropriate uncertainty calibration.

Across the five independent train–test splits, DoRIAT produced consistent NLL trajectories, comparable correlation patterns between predicted and true RMSD values, and similar uncertainty distributions. This is also confirmed by Tab. 11 where kernel's amplitude and length scale parameters are inferred for each split. These results indicate that the model's performance is stable and reproducible across different random partitions of the dataset, with no meaningful variation attributable to the specific choice of training and test sets.

Table 11: **Kernel parameters inferred for each train/test split.**

Split	Amplitude (l)	Length (ρ)
1	0.45	25.87
2	0.52	29.80
3	0.45	25.87
4	0.47	26.88
5	0.50	29.11

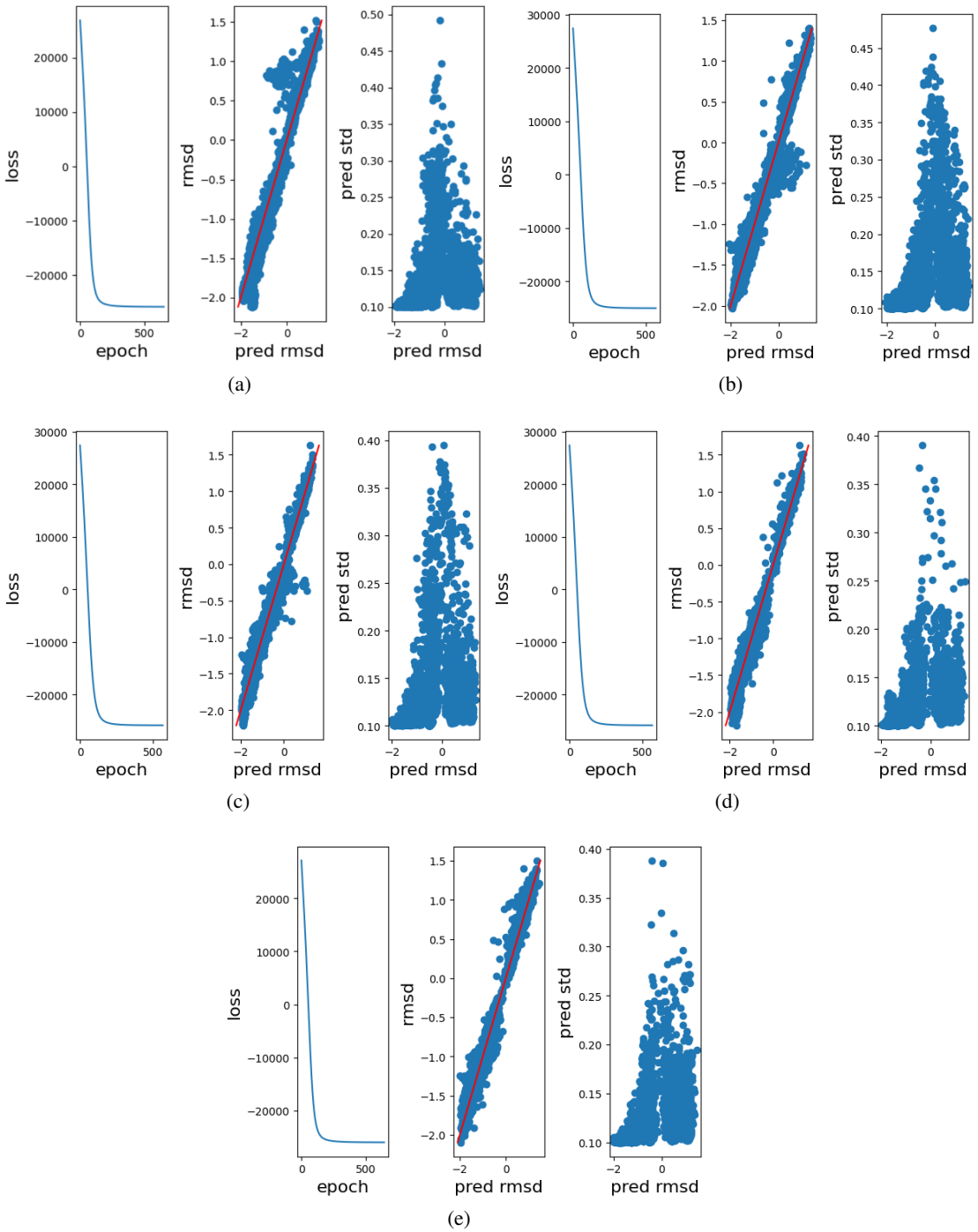


Figure 10: Model performance across 5 independent train–test splits. Each panel corresponds to one split and shows: Negative log-likelihood across training epochs, correlation between predicted and true RMSD values, and predicted standard deviation as a function of the predicted RMSD.

A.17 Ablation study

To assess the sensitivity of DoRIAT to kernel selection and input feature composition, we performed a series of ablation experiments. We first examined the impact of the covariance function by varying the Gaussian Process kernel while keeping all six structural parameters (cross, tilt, roll, shiftx, shifty, shiftz) intact. Specifically, we compared the baseline Matérn $\frac{5}{2}$ kernel against alternative formulations, including Matérn $\frac{3}{2}$, Matérn $\frac{1}{2}$ and Gaussian kernels. This analysis allowed us to determine how kernel smoothness and prior assumptions about function variability influence model performance and generalization.

In the second set of experiments, we investigated the relative contribution of each binding pose descriptor by systematically removing one parameter at a time while keeping the kernel fixed. This step isolates the individual effect of each geometric feature on DoRIAT's ability to interpret and annotate TCR–pHLA docking runs. For all experimental variants, we recorded the log-likelihood on both training and test sets, as well as the root-mean-square deviation (RMSD) between predicted and ground-truth values on the test set. Together, these analyses provide a detailed characterization of the factors most critical to the robustness and interpretability of the DoRIAT framework.

The kernel ablation results can be found in Tab. 12 and indicate that DoRIAT performs consistently well across the tested covariance functions. Best results are found for Matérn $\frac{3}{2}$ and $\frac{5}{2}$ kernels yielding nearly identical performance in terms of both negative log-likelihood (NLL) and RMSD on the test set (RMSD ≈ 0.092). [5] In contrast, the Matérn $\frac{1}{2}$ and Gaussian kernels resulted in slightly higher test RMSD values (0.095 and 0.104, respectively) and moderately higher test NLLs, suggesting that overly rough or excessively smooth priors limit the model's ability to capture fine-grained structural variation. Given the similar generalization and the smoother covariance structure of the Matérn $\frac{5}{2}$ kernel, we adopt it as the default kernel for subsequent experiments.

Table 12: Ablation study on DoRIAT kernels

	NLL (train)	NLL (test)	rmsd (pred. vs. ground truth in test)
Gussian	-24276.02	-8674.94	0.10
Matérn $\frac{1}{2}$	-22865.89	-8433.77	0.10
Matérn $\frac{3}{2}$	-25122.28	-9094.49	0.09
Matérn $\frac{5}{2}$	-25044.46	-8978.55	0.09

In the second part of the ablation study presented in Tab. 13 , where individual binding mode parameters were omitted, all six geometric descriptors contributed positively to model performance. This confirms that each parameter captures a complementary aspect of the TCR–pHLA docking landscape. Among them, *cross* and *shiftz* emerged as the most critical; removing either led to substantial degradations in predictive accuracy, with RMSD increasing to 0.30 and 0.22 respectively, and a pronounced increase in NLL. Excluding the remaining parameters (tilt, roll, shiftx, shifty) also produced measurable, though less severe, performance drops. These findings highlight the interdependence of the six geometric descriptors and underscore that accurate interpretation of docking configurations requires the full set of binding mode parameters to preserve structural fidelity and predictive robustness.

[5] Note: This RMSD is not the same as the one predicted by DoRIAT. DoRIAT predicts RMSD between docked model and xtal structure, subsequently these predicted values along with ground truth rmsd between docked models and xtal are used to compute the RMSD values presented in Tab 12

Table 13: Ablation study on DoRIAT parameters

	NLL (train)	NLL (test)	rmsd (pred. vs. ground truth in test)
no cross	-5910.15	-3795.83	0.30
no tilt	-23817.27	-8720.36	0.13
no roll	-22599.71	-8170.09	0.12
no shiftx	-22547.85	-8395.61	0.11
no shifty	-22501.91	-8423.96	0.13
no shiftz	-16749.94	-5752.90	0.22

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the main claims presented in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: DoRIAT was developed to analyze TCR–pHLA docking simulations based on six binding mode parameters, as described in the abstract and introduction. Following the same framework, DoRIAT can be extended to antibody–antigen complexes, which share similar binding mode characteristics; an effort that is currently underway. However, this methodology is not directly applicable to the broader class of protein–protein complexes.

   Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provides the full set of assumptions and proofs for each theoretical result, as detailed in Sec. 2 and Appendix Sec. A.4, A.6, A.7, A.8 and A.13.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: A Python implementation of DoRIAT, along with data preprocessing scripts and docking dataset is publicly available at `https://zenodo.org/records/14763708`. Sec. 1 and A.3 provide a high level overview of the modeling pipeline used to generate the TCR–pHLA docking runs. There certain details regarding the modeling approach have not been disclosed as it part of proprietary code. Certain implementation details of the modeling approach are not disclosed, as they involve proprietary code. However, since DoRIAT can

be applied to any docking output generated by the pipeline, the availability of the docking data ensures that the main claims of the paper are reproducible.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A Python implementation of DoRIAT, along with data preprocessing scripts and docking dataset is publicly available at `https://zenodo.org/records/14763708`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides relevant information in Sec. 2 and Appendix Sec.A.6, A.7, A.8 and A.13. A Python implementation of DoRIAT, along with data preprocessing scripts and docking dataset is publicly available at `https://zenodo.org/records/14763708`.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Sec. A.16 provides DoRIAT's performance across 5 train/test splits. For each of the splits a scatter plot summarizing the GP predicted standard deviation as a function of predicted RMSD.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Relevant results can be found under Sec. A.13.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research presented in the paper fully conforms to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Societal impact of DoRIAT is discussed under Sec. **??**.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No safeguards were necessary for this work.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All creators and original owners of assets used in the paper are properly credited. Specifically, HADDOCK3.0 is used under a legally purchased license, ensuring full compliance with its terms of use. Additionally, wherever work from other researchers has been incorporated, appropriate references are provided to acknowledge their contributions.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets are introduced in this dataset.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No data from crowdsourcing initiatives or human subject experiments were used in this study.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: DoRIAT is developed using data from a public database that has obtained all necessary approvals.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM was used for editing and formating purpose.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.