DoRIAT: A BAYESIAN FRAMEWORK FOR INTERPRETING AND ANNOTATING DOCKING RUNS.

TECHNICAL REPORT

Christos Maniatis
christos@etcembly.io
Etcembly ltd.
Harwell Campus, Oxfordshire, UK

Zahra Ouaray
zahra@etcembly.io
Etcembly ltd.
Harwell Campus, Oxfordshire, UK

Kai Xiao
kai@etcembly.io
Etcembly ltd.
Harwell Campus, Oxfordshire, UK

Thomas P.E. Dixon
tom@etcembly.io
Etcembly ltd.
Harwell Campus, Oxfordshire, UK

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

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

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

ABSTRACT

The advent of sequence-to-structure deep-learning models have transformed protein engineering landscape by providing an accurate and cost effective way to determine crystal structures. Despite their accuracy, deep-learning predictions tend to give limited insights around protein dynamics. To improve conformation exploration we have developed a machine learning pipeline that combines deep-learning predictions with molecular docking. In this report, we propose **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 binding conformation. We use DoRIAT to score output from docking runs, identify complexes close to the native structure and create ensembles of models with similar binding conformations. Our results demonstrate that the single structural model DoRIAT selects to be the closest representation of the crystal structure lies within the top 10 of docked models, ranked by RMSD, in around 80% of cases.

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 antigens (pHLA), they can constitute a very effective means to protect humans from pathogens or other foreign substances [14]. Characterising the molecular factors governing interactions between T cell receptor (TCR) and pHLA could shed light on T cells' ability to discriminate against antigens [14] 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 [45], can be found in Fig. 8.

Resolving protein crystals has a long and interesting history dating back to the early 1930's [5, 13, 35, 29]. 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, 45]. Technological improvements like crystallisation automates, brighter synchrotron X-ray sources, faster detectors, automated structure solution and refinement pipelines have significantly improved time [17, 45], 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 [65, 18]. However, end-toend protein crystal structure determination remains an expensive and laborious process.

The advent of deep-learning based sequence-to-structure models like AlphaFold [34] 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[33]. 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 [33]. 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 [25], in a pipeline named EMLy™Dock.

More precisely, we model alpha and beta chains of the TCR and the HLA using TCRModel2 [74], an AlphaFold derivative. Then HADDOCK3.0 [20], a physics based docking platform, brings together the TCR, peptide and HLA like similar methods found in literature [15, 26, 9, 64]. The docking process involves two core steps. During the first step different ligand conformations in the active site of protein are sampled [26]. To reduce computational burden, one protein is fixed in space and the second is rotated and translated around the first one [20]. 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. 1. 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 [62].

In this report, we present **Do**cking **R**un **I**ntepretation and **A**nnotation **T**ool (DoRIAT) a method for interpreting and annotating the output of physics based docking programs like [12, 68, 20]. 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 xtal 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.

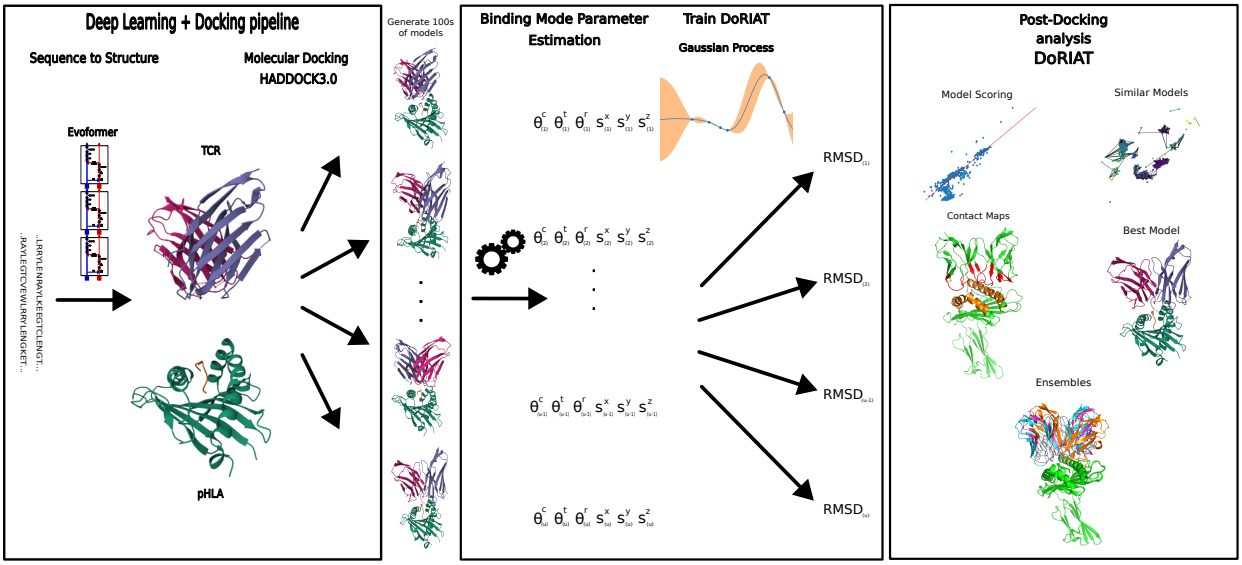


Figure 1: **Schematic representation of EMLy™Dock and DoRIAT.** In house developed pipeline that combines deep-learning based sequence-to-structure model to generate alpha and beta chains, followed by *in-silico* docking with HADDOCK3.0[20] and DoRIAT to interpret docking outcomes and identify the best conformation.

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 [36, 46].

Geometric matching between protein and ligand has repeatedly reaffirmed to have a vital role in determining the geometric complex [75, 39, 26], with early docking algorithms heavily capitalising on it [23, 54, 55, 53, 56]. However, such efforts were quickly abandoned as they struggle to generalise in absence of ground truth introducing many false positives [26].

Force-field-based scoring functions [10, 2, 38] 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 [46]. 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 [31, 50].

In **empirical scoring** functions [28, 32, 70, 24, 8], binding energy decomposes into energy components, such as hydrogen bond, ionic interaction, hydrophobic effect and binding entropy [46]. The contribution of each component is estimated from a set of ligand-protein complexes with measured affinity. Many modern docking engines such as [20, 12] 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 [51, 49, 30, 22, 69] 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 [46]. Similar to empirical-based scoring approaches, knowledge-based ones often fail to generalize outside of their training data.

Consensus scoring [11] combines several different scores to assess the docking conformation. It has been reported [21, 11, 7, 67] that consensus scoring improves prediction of bound conformations and poses [46]. However, binding energies might still be inaccurate and its predictive power depends on the independence of scoring functions it is made of [36, 21, 11, 7, 67].

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 [76], Replica Exchange Free Energy Perturbation [47], Molecular Mechanics Poisson-Boltzmann solvent-accessible surface area (MM-PB/SA) [37]. However, they are computational demanding and can easily become prohibitively expensive for larger molecules[48]. 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. [71] 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 [58, 61, 73] 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.

3 Results

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. 5.1. 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

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.

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. 5.1

we will focus on docking outcomes from EMLy™Dock, Etcembly's 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 don't 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 1.

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 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. 5.1. 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.

	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

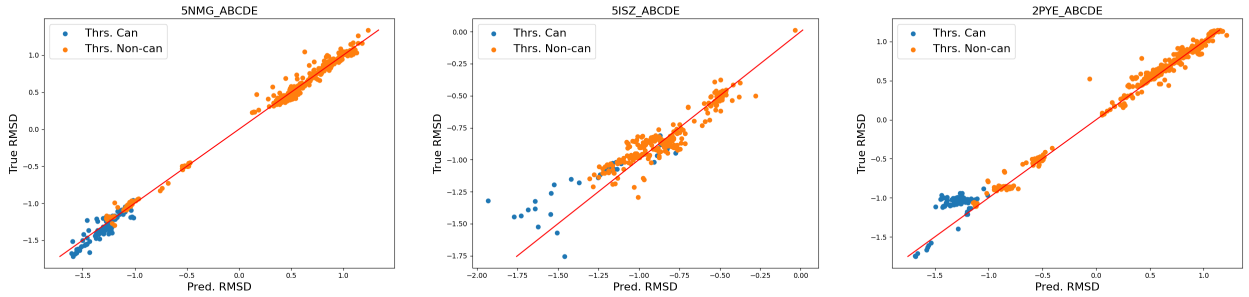
Table 2: **Table summarizing the cut-off thresholds for binding mode parameters.**

3.1 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 [40]. 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 Etcembl'y proprietary data.

Fig. 2 shows a summary of the scoring results for docked conformations of cancer (2PYE), viral (5ISZ) and HIV (5NMG) related systems from the test set. For all examples DoRIAT can accurately predict docked models far from the native structure. All of these models have binding mode parameters that fall outside the canonical range and can be identified with a thresholding approach as indicated by the orange and blue color-coding. For models closer to

the native structure in some complexes, DoRIAT struggles more, as indicated by the larger deviations of predicted means from perfect match in Fig. 2a and 2b. This happens because some of the docked models present binding mode parameters that are on the edge of what our proposed threshold-based approach would deem canonical. A more detailed analysis of this claim is presented later in the subsection. At the same time, there are several examples like 5NMG (Fig. 2c) where the docking engine does not produce these edge cases and therefore DoRIAT performs well.



(a) (b) (c)

Figure 2: **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 2PYE (Fig. 2a), 5ISZ (Fig. 2b) and 5NMG (Fig. 2c). 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.

For TCR-pHLA systems, like 2PYE and 5ISZ, where DoRIAT's prediction is different from ground truth, binding mode parameters of the native complex can be found in the first two lines of Tab 3. Searching through the docking results for 2PYE, we identify some conformations where the threshold based approach would struggle. The most representative example is 2PYE_m1 whose parameters are all within range. For docking simulations of 5ISZ, model 5ISZ_m1 presented in Tab. 3 would be a plausible conformation whose θ^c is borderline but within range. 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. Even though for these models DoRIAT underestimates the models' distance from native structure, it is still able to suggest that all docked models are more than 10 Å away from xtal, minimizing the false positive probability. In contrast, a threshold-based approach would keep these models as they are both within threshold.

DoRIAT was also used to analyse docking simulations of unpublished TCR-pHLA complexes that Etcembly possesses. These complexes are much more heterogeneous than public data and are intended to become cancer therapeutics. Their binding mode parameters can be found in rows 3 and 4 of Tab. 3. A direct consequence of the inherent complexity is to have conformations with parameters within range, whose backbone is far from crystal structure. These models cause DoRIAT to struggle with accurate RMSD prediction. The most indicative example is for TCR A, where the native structure has (θ^c, θ^r) angles of $(40.56°, 10.76°)$, but TCR-pHLA complex A_m1 and TCR-pHLA complex A_m2 bind at $(22.75°, -13.67°)$ and $(70.17°, -16.12°)$. There, the discrepancy between DoRIAT's prediction and actual RMSD would be 5 Å and 3 Å respectively. However, these examples would not affect downstream analysis as models TCR-pHLA complex A_m1 and TCR-pHLA complex A_m2 would be flagged for removal due to their high predicted RMSD (19 Å and 7 Å respectively). The native mode of binding for TCR B can be found in line 4 of Tab. 3. For models TCR-pHLA complex B_m1 and TCR-pHLA complex B_m2, DoRIAT tends to under-estimate the predicted RMSD from crystal structure, given their binding mode parameters are far from the native structure, but within range. Fig. 4b confirms that and additionally demonstrates that this phenomenon is very prevalent across this docking run, as most models are far from the line indicating perfect predictions. This is explained by the large number of canonical conformations suggested by the docking platform that are close to TCR-pHLA complex B_m1.

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 following section, we outline an approach that leverages the model to address this challenge.

3.2 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

	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

Table 3: **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).

structurally assessing candidates from our screening 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.

To select a good model we use the binding mode parameters to estimate predicted RMSD (μ_{RMSD}) from Eq. 21 along with its corresponding uncertainty (σ_{RMSD}) from Eq. 22 to solve the constrained minimization problem described in Eq. 25. This objective function seeks the docked model with the lowest RMSD from the crystal structure while resolving ties using a statistical testing inspired constrains. More information can be found in Sec. 3.2.

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. Internal structures are cancer related, while public structures contain a mixture of viral and cancer. All structures are bound to HLA-A*02.

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 [61] to predict RMSD using the code provided in DeepRank2. For DeepRank-GNN-esm [73] and GNN-DOVE, we use the pre-trained models provided in DeepRank-GNN-esm and GNN-DOVE to score models.

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. 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 the optimization problem described in Eq. 25. 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. 3 shows the summary results for 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. Naive GNN and EMLy™Dock scoring functions follow by selecting a model within top 70 about 60% and 40% of times respectively. DeepRank-GNN, DeepRank-GNN-esm and GNN-DOVE select a model within top 70 about 22%, 17% and 17% of times. More results can also be found in Tab. 5. One obvious advantage of DoRIAT is that it has to draw conclusions using combination of binding mode parameters instead of assessing local interactions. Since HADDOCK3.0 [20] can effectively explore the space of conformations and all the examined systems contain an HLA-A*02, for the binding mode parameters within canonical range DoRIAT is essentially interpolating. The DeepRank variations trained on the DoRIAT's training set are trying to infer conformations near the native structure by assessing local interactions which is

challenging. In addition to that, some of the performance degradation can be attributed to the lack of PSSM, which was ignored as it would make unfeasible for assessing large number for TCRs. The DeepRank-GNN-esm and GNN-DOVE have been pre-trained on antibody-antigen docking results, which differ from our TCR-pHLA systems we are interested in. With training on DoRIAT's 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.

To further asses the quality of selections we consider DockQ score [3] which incorporates additional quality measures like F_{nat}, LRMS and iRMS standardized by CAPRI [44]. Tab. 4 and Tab. 6 summarize the DockQ score for model selection in benchmarking across test set and internal TCRs. DoRIAT presents superior performance with all apart from 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 based on the thresholds defined in Tab 2.

In the left out set, there are examples where DoRIAT's selection is out-performed by other methods when DockQ scores are compared. These include TCR-pHLA complex 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 makes the best selection. 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 might sometimes be out-performed by compared methods but it tends to be more consistent in the quality of its selection.

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.

	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

Table 4: **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.

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

Fig. 3 demonstrates that DoRIAT, in most cases, chooses a model close to the crystal structure. However, this task is one of the many possible which can be performed on docking outcomes. An alternative use could be to discriminate between poses, giving clear distinction between completely disconnected conformations from the ones whose global geometry matches. This could be used to create ensembles of similar models around points of interest, expanding the limited view provided by crystals or individual models. In this section, we provide a brief overview of how the DoRIAT's covariance is used to create an ensemble around points of interest using two complexes, TCR-pHLA complex A and TCR-pHLA complex B.

Core to such type of analysis is the identification of similar conformations. Correlation estimated from the GP covariance is the most straightforward way of achieving that goal. The presence of confounding poses, i.e. poses which are in between canonical and non-canonical conformations, will introduce false connections between canonical and non-canonical poses. These misleading connections could be removed by controlling for such examples with partial correlation. For more information the reader is directed to Sec. 5.2.

Fig. 4c and 4d show PCA of binding mode parameters for complexes A and B, with similar conformations being marked with edges. In both examples docking poses form clusters around distinct canonical and non-canonical poses.

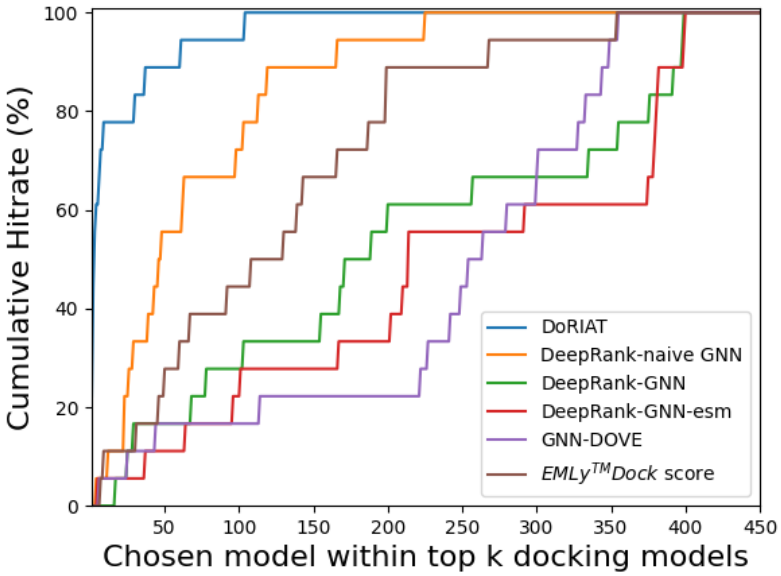


Figure 3: **Cumulative hitrate across compared methods.** Graph summarizing cumulative hitrate for DoRIAT (blue), naive GNN (orange), DeepRank-GNN (green), DeepRank-GNN-esm(red) and EMLy™Dock (purple) across unseen complexes. Each method chooses one docked model that it believes is closest to the crystal 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.

For TCR-pHLA complex A conformations within canonical range have θ^c of about $35\circ$ and $45\circ$, while for TCR-pHLA complex B canonical conformation clusters are formed around θ^c of about $24\circ$ and $65\circ$. In 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 demonstrate a use case where covariance information could aid in a more targeted analysis we are creating two ensembles of models and comparing their average contact maps to the ones observed in the crystal structure. The first ensemble is based on models which are deemed canonical using the thresholds in Tab. 2. For the second ensemble, the optimal model derived from Eq. 25 along with similar models derived by partial correlation coefficient are considered. The quality of each ensemble is assessed by comparing TCR-HLA and TCR-peptide contact maps (see Sec. 5.3) to the equivalent contact maps observed on the crystal using mean structural similarity index (SSIM) [72]. We believe that SSIM is better suited for this task, compared to other distance metrics such as mean squared error, as the positions of contact matter. We are more tolerant if contacts happen in the vicinity of the contacts suggested by the crystal, as proteins are in a constant flux. However, contacts which are far from the crystal's contact have to be penalized stronger as they are more likely to be modelling artifacts.

Fig. 5 and 6 summarize contact maps describing the TCR-HLA and TCR-peptide interactions for the two ensemble methods described above. From Fig. 5a we can see that both the threshold-based ensemble and the ensemble around the best model of Eq. 25 create artificial contacts compared to the crystal structure. These can be found in the region $(\text{TCR}, \text{HLA}) = (25, 52)$. At the same time, both ensembles identify a plethora of correct contacts. Hence, in this case SSIM is 0.979 and 0.975 with threshold-based ensemble being slightly better. For the TCR-epitope contact map Fig. 6a ensemble around optimal model has SSIM 0.815 against 0.805, due to fewer contacts with the incorrect epitope regions. Fig. 5b suggests that for complex B the ensemble around optimal model of Eq. 25 make very similar contacts to the crystal. The threshold-based approach includes several canonical conformations which are far from the native structure, creating erroneous contacts at positions like $(\text{TCR}, \text{HLA}) = (26, 55), (204, 150)$. In this case, SSIM for the TCR-HLA contact map of the threshold-based ensemble is 0.970 and is 0.015 less from the equivalent of optimal model ensemble. For contacts with the epitope we see that each ensemble makes different mistakes, but on average the optimal model ensemble performs slightly better. More precisely, the optimal models ensemble misses the contacts of region $(\text{TCR}, \text{Peptide}) = (200, 5)$ and introduces contacts in the region $(\text{TCR}, \text{Peptide}) = (30, 3)$. The threshold-based ensemble correctly identifies contacts at $(\text{TCR}, \text{Peptide}) = (200, 5)$, but introduces contacts in the

regions $(\text{TCR}, \text{Peptide}) = (30, 3), (48, 3)$. SSIM for the contact map of the optimal model ensemble is 0.850 against 0.843 for the threshold-based ensemble.

In summary, our results suggest the covariance information can help create ensembles with contact maps that are of equal quality, if not better than contact maps created from ensembles derived using threshold-based methods on binding mode parameters. This can aid in the creation of more accurate interface analysis improving downstream analysis.

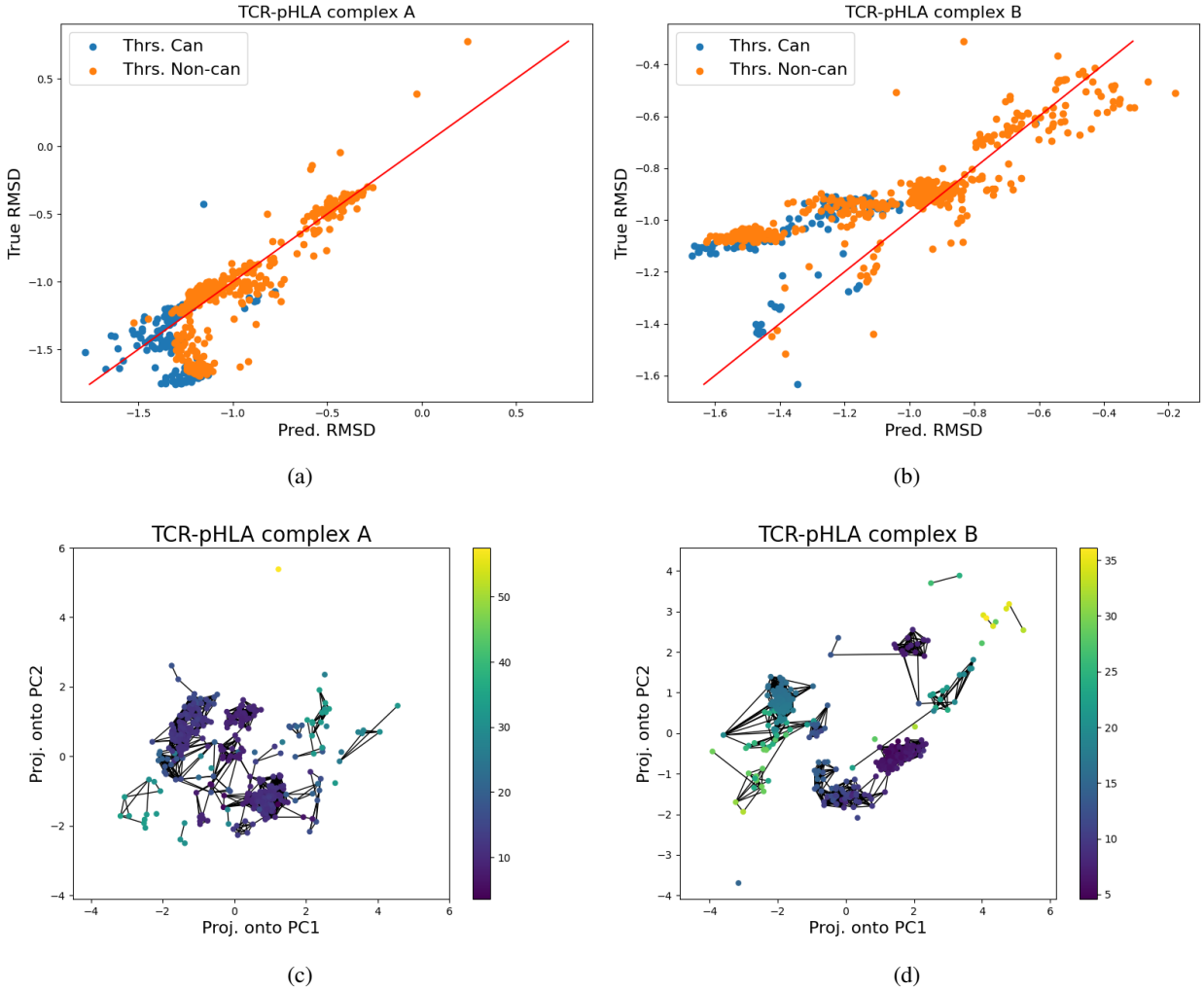
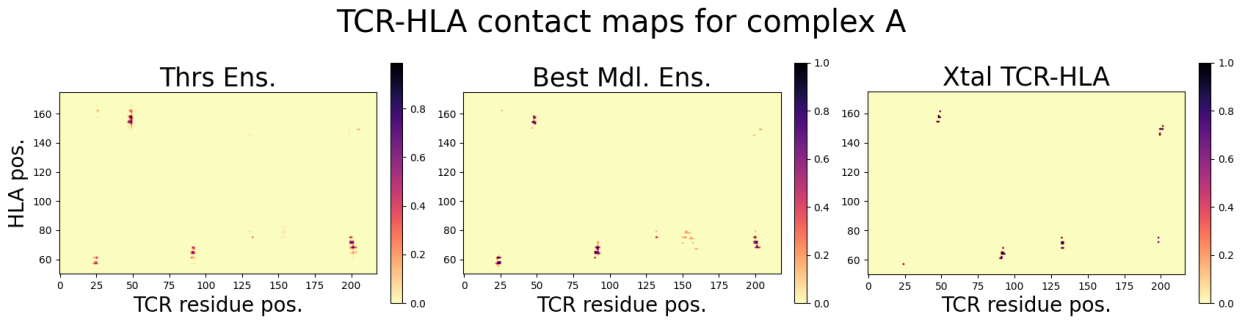


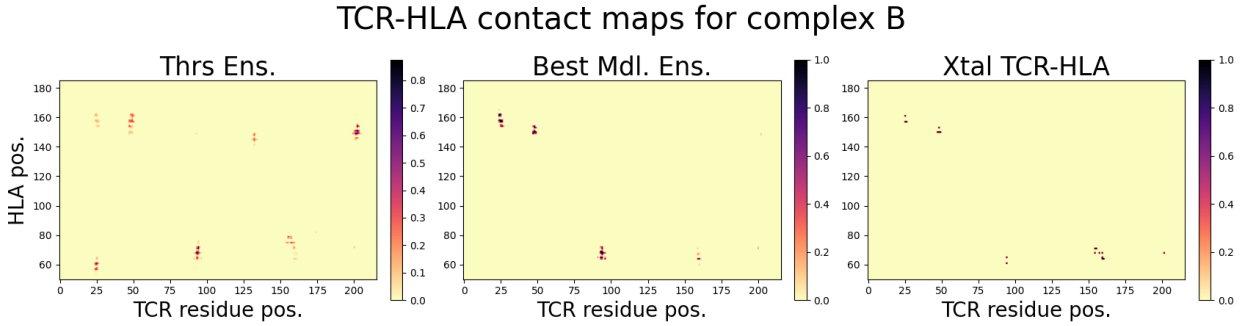
Figure 4: **Interpretation of docking runs for internal data.** Figures summarizing post-docking analysis for TCR A and TCR B. (4a,4b) Scatter plots of measured RMSD as a function of predicted RMSD for TCRs that DoRIAT performs well (TCR A) and struggles (TCR B) respectively. Each dot represents a docked model and is color-coded using cut-off thresholds of Tab. 2. The red line corresponds to perfect match between predicted and measured RMSD. (4c,4d) PCA embedding of binding mode parameters for TCR A and TCR B color-coded by predicted RMSD. Each dot represents a docked model. Edges are placed between conformations with similar binding mode parameters and are computed using 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.



(a)

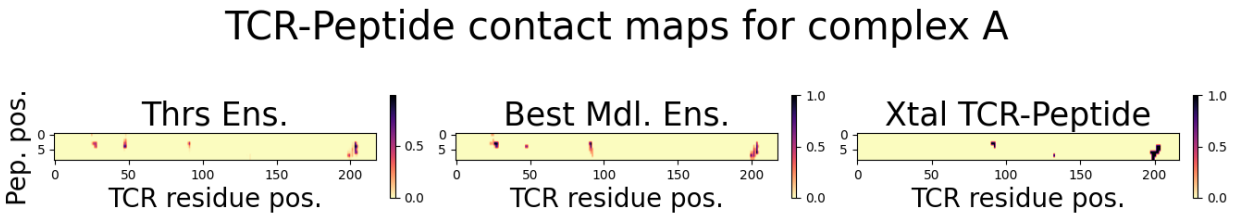


(b)

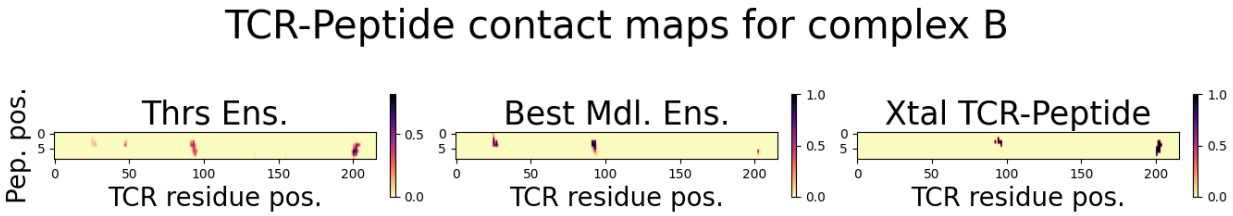
Figure 5: **Contact maps between TCR-HLA** $(8\ \text{Å})$ **for different ensembles for complexes A and B.** Summary of contact maps between the TCR-HLA for complexes A (Fig. 5a) and B (Fig. 5b) for ensembles created using canonical models based on thresholds and similar models around prediction made with Eq. 25. 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 xtal.

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 assisting Etcembly's structural biology team to interpret 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



(a)



(b)

Figure 6: **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. 6a) and B (Fig. 6b) for ensembles created using canonical models based on thresholds and similar models around prediction made with Eq. 25. 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 xtal.

challenges as antibodies lack co-evolutionary constraints, but their large numbers in public repositories could help the algorithm to attain reasonable predictive power.

5 Methods

DoRIAT implements a Gaussian Process regressor to score EMLy™Dock models from six binding mode parameters, three angles and three displacements. A graphical representation for the GP implemented in DoRIAT can be found in Fig. 7

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:

$$\mathbf{x}_{um} = \begin{pmatrix} \theta^c_{um} \\ \theta^t_{um} \\ \theta^r_{um} \\ s^x_{um} \\ s^y_{um} \\ s^z_{um} \end{pmatrix} \tag{1}$$

A detailed description as to how these geometric parameters are obtained can be found in Binding mode parameters. 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

$$\mathscr{K}(x_1, x_2) = \mathbb{E}[f(x_1) f(x_2)] - \mathbb{E}[f(x_1)] \mathbb{E}[f(x_2)] \tag{4}$$

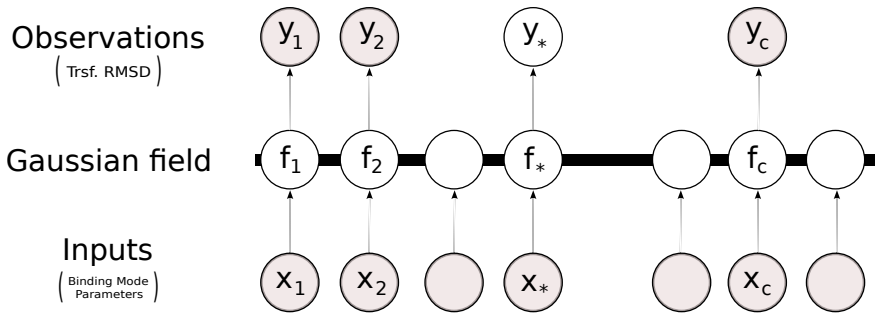


Figure 7: **Graphical model of GP regression.** Each node represents a variable, with shaded nodes corresponding to observed variables. The thick horizontal bar represents a set of fully connected nodes. Figure taken from [57]

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 kernel a generalization of the most popular radial basis function (RBF) kernel designed to address less smooth data. More precisely,

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

where l is the amplitude controlling the marginal variance of the function and ρ is the length-scale controlling spatial variability.

5.1 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 [41, 42, 43]. 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.

$$v_{\text{TCR},j} = \frac{\text{a}_{\text{alpha},J(j)} + \text{a}_{\text{beta},J(j)}}{2} \text{ for } j \text{ in } 0, \ldots \|J\| - 1 \tag{6}$$

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,

$$v_{\text{MHC}_1} = \frac{v_{\text{H}_1} + v_{\text{H}_2}}{\|v_{\text{H}_1} + v_{\text{H}_2}\|_2} \tag{7}$$

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

$$v_{\text{MHC}_2} = \frac{\left(v_{\text{MHC}_1} \times v_{\text{H12}}\right) \times v_{\text{MHC}_1}}{\|\left(v_{\text{MHC}_1} \times v_{\text{H12}}\right) \times v_{\text{MHC}_1}\|} \tag{8}$$

Where $\times$ is the cross product [1]. The centre of HLA mass p_{MHC} is defined as

$$p_{\text{MHC}} = \frac{p_{\text{H1}} + p_{\text{H2}}}{2} \tag{9}$$

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 :

$$n_{\text{MHC}} \cdot k = n_{\text{MHC}} \cdot p_{\text{MHC}} \ , \tag{10}$$

where

$$n_{\text{MHC}} = \frac{v_{\text{MHC}_1} \times v_{\text{MHC}_2}}{\|v_{\text{MHC}_1} \times v_{\text{MHC}_2}\|_2} \tag{11}$$

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 10 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. 8. Let θ^c, θ^t and θ^r be the cross [60], tilt[66] and roll[66] angles. Then

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

$$\theta^c = \begin{cases} \tilde{\theta}_c, \text{ if } \tilde{\theta}_c \geq 90^\circ \\ 360^\circ - \arccos \frac{v_{\text{MHC}_1} \cdot v_{\text{SB}}}{\|v_{\text{MHC}_1}\|_2 \|v_{\text{SB}}\|_2}, \text{ otherwise} \end{cases} \tag{12}$$

$$\theta^t = \arccos \frac{v_{\text{MHC}_1} \cdot v_{rot}}{\|v_{\text{MHC}_1}\|_2 \|v_{rot}\|_2} - 90^\circ \tag{13}$$

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

Let s^z be the z axis displacement of p_{TCR} from the plane 10. Then

$$s^z = \frac{|n_{\text{MHC}} \cdot p_{\text{TCR}} - n_{\text{MHC}} \cdot p_{\text{MHC}}|}{\|n_{\text{MHC}}\|_2} \tag{15}$$

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 10. Then

$$s^x = \frac{(p_{\text{TCR}} - p_{\text{MHC}}) \cdot v_{\text{MHC}_1}}{\|v_{\text{MHC}_1}\|_2} \tag{16}$$

and

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

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

(a) (b)

Figure 8: **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. 8a) and without (Fig. 8b) 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.

[1]An alternative way of obtaining Eq. 8 is by starting with vectors v_{MHC_1} and v_{H12} and creating an orthonormal vector basis with Gram–Schmidt process [16].

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. 6 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. 7 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. 8 estimate v_{MHC_2}.
9: Using Eq. 9 and 11 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. 12, 13 and 14 respectively.
11: x-axis (s^x), y-axis (s^y) and z-axis (s^z) displacements can be estimated from Eq. 16, 17 and 15 respectively.

5.2 Identifying models with similar docking poses using GP covariance.

For a set of docking experiments, an undirected graphical model $\mathscr{G}(V, E)$ can be defined as a collection of vertices (V), representing binding mode parameters of the TCR on the pHLA complex, and edges (E) between vertices representing pairwise relationship. A common representation of E is with a binary adjacency matrix where two linked observations are represented with 1 and 0 otherwise [63]. For the purpose of this paper we will consider the weighted adjacency matrix which also takes into account the importance of connections.

Conditional independence graph is a undirected graphical model summarizing conditional dependencies between variables [63]. Let $\mathbf{X}_{1m_1}$, $\mathbf{X}_{1m_2}$ and $\mathbf{X}_{1m_3}$ denote sets of models with distinct binding mode parameters for TCR with index 1. $\mathbf{X}_{1m_2}$ is said to separate $\mathbf{X}_{1m_1}$ and $\mathbf{X}_{1m_3}$ when there is not path between $\mathbf{X}_{1m_1}$ and $\mathbf{X}_{1m_3}$ that does not pass through $\mathbf{X}_{1m_2}$. In a conditional independence graph, the absence of an edge between two nodes $\mathbf{X}_{1m_1}$ and $\mathbf{X}_{1m_3}$ corresponds to conditional independence of the random vectors $\mathbf{X}_{1m_1}$ and $\mathbf{X}_{1m_3}$ given the separating set $\mathbf{X}_{1m_2}$.

Graphical Gaussian Process Model (GGPM) could be thought as a special case of a conditional independence graph described by Eq. 3. The edge weights of a GGPM correspond to partial correlations between the variables in $f(\mathbf{X}) = \{f_{um} \forall u \in (1, \ldots, U) \text{ and } m \in (1, \ldots, M)\}$. Let $f_{1,m_1} \in f(\mathbf{X}_{1,m_1})$ and $f_{1m_3} \in f(\mathbf{X}_{1,m_3})$, where $f(\mathbf{X}_{1,m_1}) = \{f(\mathbf{x}_{1,m_1}) \forall \mathbf{x}_{1,m_1} \in \mathbf{X}_{1,m_1}\}$ and $f(\mathbf{X}_{1,m_3}) = \{f(\mathbf{x}_{1,m_3}) \forall \mathbf{x}_{1,m_3} \in \mathbf{X}_{1,m_3}\}$. Then the partial correlation between variables f_{1m_1} and f_{1m_3} is a measure of their conditional association, given the remaining elements of $f(\mathbf{X})$. Let $f(\mathbf{X})_{1,-m_1,-m_3}$ denote the set of all variables in $f(\mathbf{X})$ except for f_{1m_1} and f_{1m_3}, the partial correlation is defined as:

$$\rho_{f_{1m_1}, f_{1m_3} | f(\mathbf{X})_{1,-m_1,-m_3}} = \frac{\mathbf{Cov}\left(f_{1m_1}, f_{1m_3} | f(\mathbf{X})_{1,-m_1,-m_3}\right)}{\sqrt{\mathbf{Var}\left(f_{1m_1} | f(\mathbf{X})_{1,-m_1,-m_3}\right)}\sqrt{\mathbf{Var}\left(f_{1m_3} | f(\mathbf{X})_{1,-m_1,-m_3}\right)}} \tag{18}$$

Conditional independence between f_{1m_1} and f_{1m_3} given $f(\mathbf{X})_{1,-m_1,-m_3}$ is equivalent to zero partial correlation between the two variables. For the purpose of this project, we take a more vigilant approach, assuming conditional independence between pairs if their partial correlation is in the range $[-0.05, 0.05]$. For a TCR with index u and models $\tilde{m}_1, \tilde{m}_2$, the edge set of GGPM is defined by the set of all pairs $(f_{u\tilde{m}_1}, f_{u\tilde{m}_2})$ with partial correlation greater than 0.05 in absolute value.

$$\mathbf{E} = \{(\tilde{m}_1, \tilde{m}_2) : |\rho_{f_{u\tilde{m}_1}, f_{u\tilde{m}_2} | f(\mathbf{X})_{u,-\tilde{m}_1,-\tilde{m}_2}}| > 0.05\} \tag{19}$$

The graph defined by binding mode parameters for some TCR with index u and edges of set E (Eq. 19 could be used to group similar docking conformations. At the same time the graph could be exploited to get a systematic way of distinguishing canonical from non-canonical docking poses.

5.3 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. There are several ways of defining interactions, but for the purpose of this paper 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. 5, 6 are created by averaging across individual contact maps of models in each ensemble.

5.4 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. 5 over training data we minimise the negative log marginal likelihood

$$\mathrm{NL}(\mathbf{Y}) = -\ln \left[\int \mathbf{N}\left(f, \sigma^2 \mathbf{I}\right) \mathbf{N}\left(f | 0, \mathscr{K}\left(\mathbf{X}, \mathbf{X}\right)\right) df \right] , \tag{20}$$

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.

6 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)[40] [2] filtered to keep only ones that bind to HLA-A*02. Using pdb-tools [59] we remove all the non-protein atoms such as water, ions, other 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 5.1. To quantify the accuracy of the docked structures, we calculated the backbone distances from xtal 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 [19]. For model training and testing, we split the data based on complex id (TCR-pHLA) labels, maintaining a 3 to 1 ratio respectively.

Acknowledgments

The authors would like to thank Kate Lawrence for proofreading the manuscript.

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.

[2] Until November 2021.

[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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] 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`.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] Michael A. Newton, Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5, 2004. ISSN 14654644. doi:10.1093/biostatistics/5.2.155.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] Katherine H. Shutta, Roberta De Vito, Denise M. Scholtens, and Raji Balasubramanian. Gaussian graphical models with applications to omics analyses. *Statistics in Medicine*, 41, 2022. ISSN 10970258. doi:10.1002/sim.9546.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

7 Appendix

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.

More precisely, we derive analytical solutions for observed RMSD mean, standard deviation, CV and entropy. By analyzing these equations, we justify why using CV in our case yields a reasonable solution.

7.1 Mathematical Derivations

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 [27] we get the following result.

$$\mu_{RMSD} = \int_{-\infty}^{\infty} \epsilon \Phi(x) \mathbf{N}(x, \mu^*, \sigma^*) dx = \epsilon \Phi \left(\frac{\mu^*}{\sqrt{1+\sigma^{*2}}} \right) \tag{21}$$

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 [27] 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:

$$\sigma_{RMSD} = \epsilon \sqrt{\mathbf{F} \left(\vec{m} | \vec{0}, V \right) - \Phi \left(\frac{\mu^*}{\sqrt{1+\sigma^{*2}}} \right)^2} \tag{22}$$

So, CV is estimated by:

$$\text{CV}_{RMSD} = 100 \frac{\sigma_{RMSD}}{\mu_{RMSD}} \tag{23}$$

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 $-\Phi(x) \ln [\Phi(x)]$. 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 $-\Phi(x) \ln [\Phi(x)]$ we see that $\int_{-\infty}^{\infty} -\Phi(x) \ln [\Phi(x)] \, dx = 0.90319728556$. Hence $c\mathbf{N}(x, \mu_L, \sigma_L)$ is a really good approximation of $-\Phi(x) \ln [\Phi(x)]$. Hence:

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

$$E_{RMSD} \approx c\mathbf{N}(\mu^*, \mu_L, \sqrt{\sigma^{*2} + \sigma_L^2}) \tag{24}$$

7.2 Objective function for selecting models.

A key aspect of post-docking analysis is to identify a conformation close to the native structure. The most intuitive way for picking such conformation, would based on μ_{RMSD} (Eq. 21). Our model selection is a variation of the standard approach that accounts for performance degradation observed in some of the most difficult complexes in test set and internal TCRs. Here, we describe our decision making process to identify good models.

The most accurate way for assessing these models would be to simulate their scoring, perform null hypothesis testing on the difference of each pair, control for Type I error with [52] and select the best model. Assuming we can work out the parameters of null distribution for each pair and with many optimizations, that approach requires tens of thousands of comparisons per TCR, creating bottlenecks in tasks like screening. Hence we need a less computationally demanding alternative.

Our starting point is analyze how predicted RMSD (μ_{RMSD}) and its corresponding uncertainty (σ_{RMSD}) behave as a function of (μ^*, σ^*) within some realistic thresholds derived from test set. A common pattern observed Fig. 9a and 9b is that we can link μ^* and σ^* such that we maintain μ_{RMSD} or σ_{RMSD} constant. For negative μ^* to keep μ_{RMSD} or σ_{RMSD} constant, the GP predicted standard deviation σ^* has to be inflated.

The best models is selected by:

$$\begin{aligned} &\min \mu_{RMSD} \\ &\text{sub. to } \mathrm{CV}_{RMSD} \geq \alpha_{0.95} \ , \end{aligned} \tag{25}$$

where $\alpha_{0.95}$ is the 95th percentage quantile of CV_{RMSD} in the docking run. As we are interested for the best model, minimizing the predicted RMSD is the sensible choice. The CV_{RMSD} based constrain is there to simulate a pairwise statistical comparison test. More precisely, it is trying to resolve situations where there are two models $(\mathbf{M}_1, \mathbf{M}_2)$, one with (μ_1^*, σ_1^*) and (μ_1^*, σ_1^*) where $\mu_1^* < \mu_2^*$ and $\sigma_1^* \geq \sigma_2^*$. Prominent examples, include models with same predicted RMSD but different predicted uncertainty and visa-versa. In Fig. 9a this would correspond to points belonging on the same boundary line between regions of different color. Since $\mathbf{M}_1$, $\mathbf{M}_2$ latent scores follow a Normal distribution it is easy to demonstrate that $\mathbf{P}(\mathbf{M}_1 \leq \mathbf{M}_2) > \frac{1}{2}$, $\mathbf{M}_1$ should be preferred. From Fig. 9b we see that following the path between parameters (μ_1^*, σ_1^*) and (μ_2^*, σ_2^*) that keeps predicted RMSD constant also inflates the predicted standard deviation.

This objective tends give good results in practice as it avoids models which are within canonical range but whose backbone is far from native structure. The reason is that it will prefer conformations with higher uncertainty and lower mean.

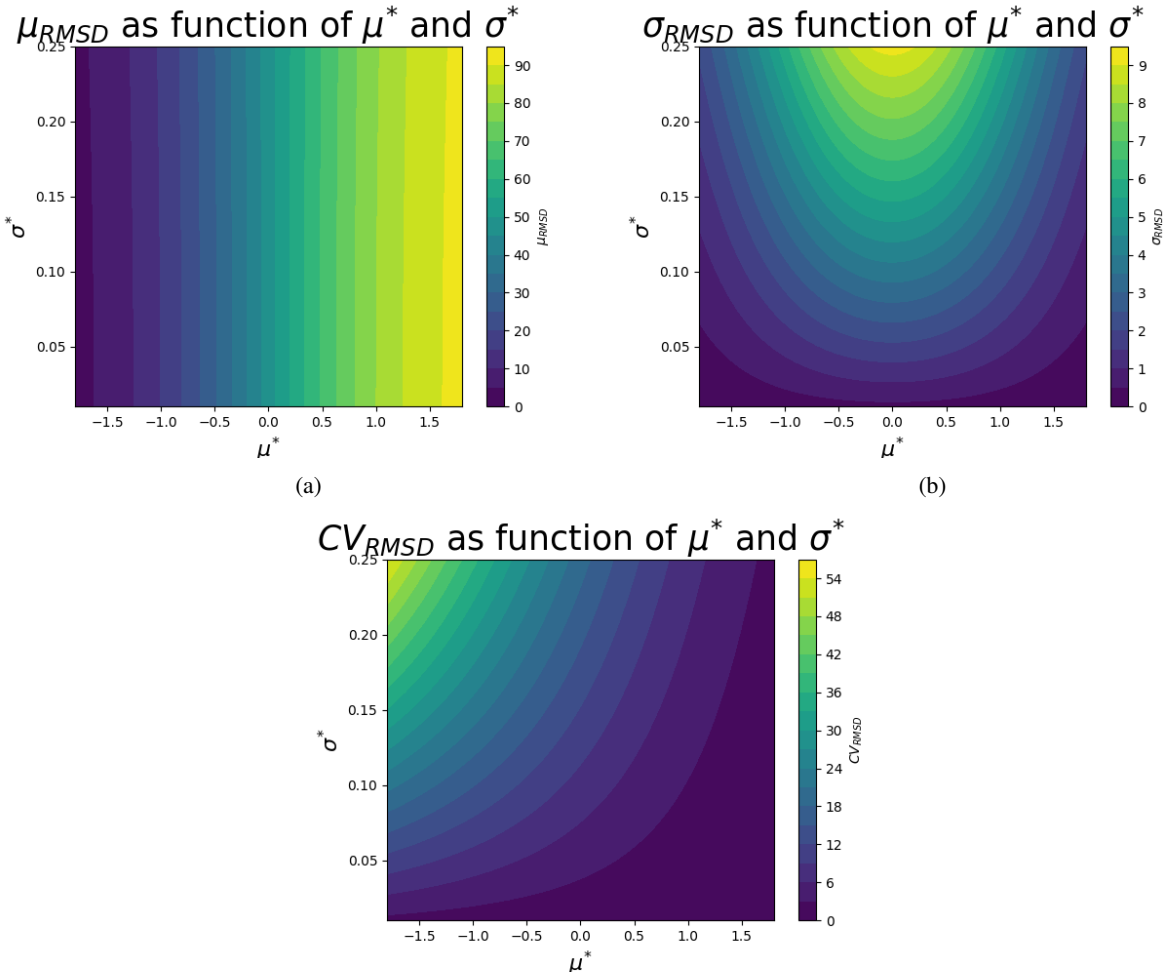
There are examples where this objective can be problematic and it involves two or more models with identical predicted RMSD and a coefficient of variation above minimum threshold. However, we have never encountered an example like that.

7.3 Cumulative hitrate across compared methods

In Sec. 3.2 we select a model from the docking output using different scoring approaches. Here we present a table summarizing the cumulative hitrate across compared methods.

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

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.



(a) (b) (c)

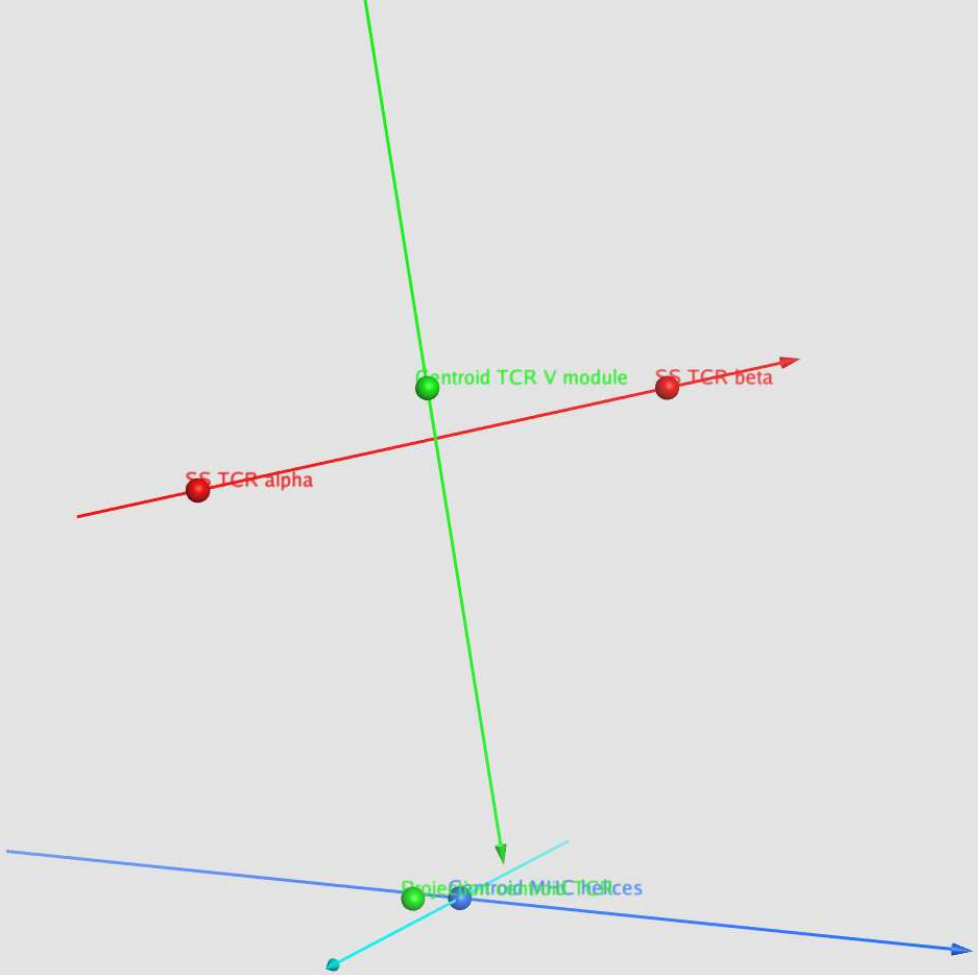
Figure 9: **Contour plots of Eq. 21,22 and 23** Contour plots of predicted mean (μ_{RMSD}, Fig 9a), standard-deviation (σ_{RMSD}, Fig 9b) and coefficient of variation (CV$_{RMSD}$, Fig 9c) as a function of parameters μ^* and σ^*. Thresholds for μ^* and σ^* are estimated from test set observations.

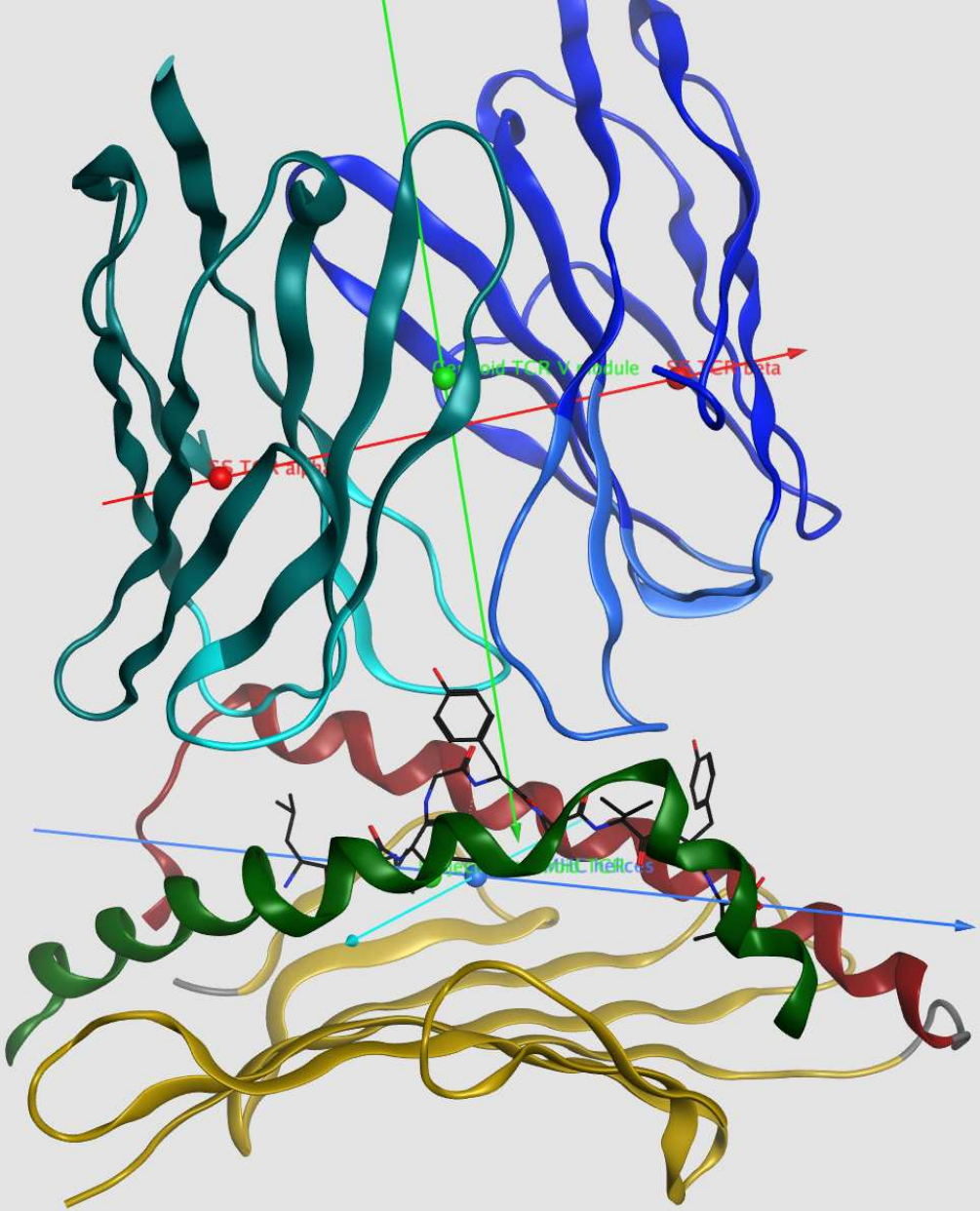
7.4 Benchmarking model selections with DockQ

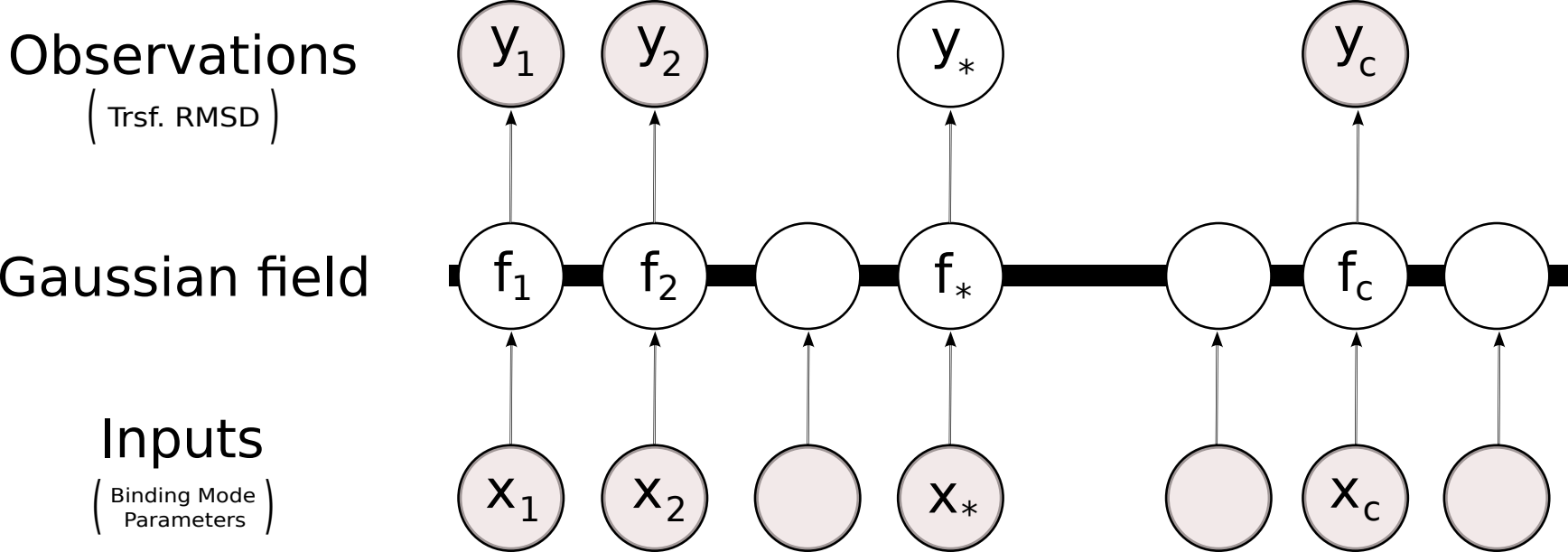
In Sec. 3.2 we select a model from the docking output using different scoring approaches. Apart from ranking predictions based on RMSD, we also compare them using DockQ score. Since DockQ takes into account interface properties apart from ligand RMSD, it provides a more holistic assessment of the docking pose. Here we provide DockQ scores from the remaining test set and internal TCRs.

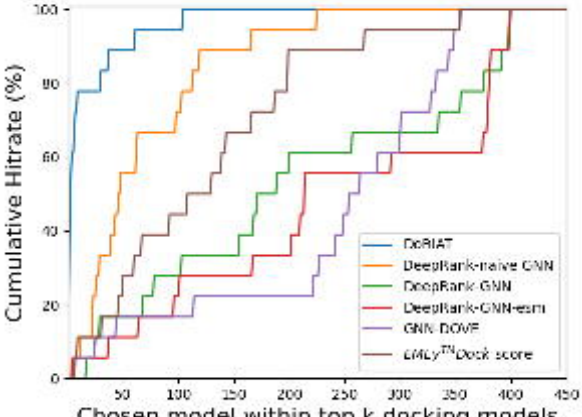
	DoRIAT	DeepRank (naive-GNN)	DeepRank-GNN	DeepRank-GNN-esm	GNN-DOVE	EMLy™Dock
TCR-pHLA complex C	0.24	0.16	0.02	0.03	0.04	**0.29**
1QRN	**0.54**	0.51	0.02	0.02	0.03	0.42
2BNQ	**0.20**	0.09	0.01	0.02	0.04	0.06
2VLJ	**0.52**	0.21	0.05	0.02	0.20	0.06
3QFJ	0.51	**0.60**	0.10	0.58	0.56	0.59
4MNQ	0.54	**0.57**	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

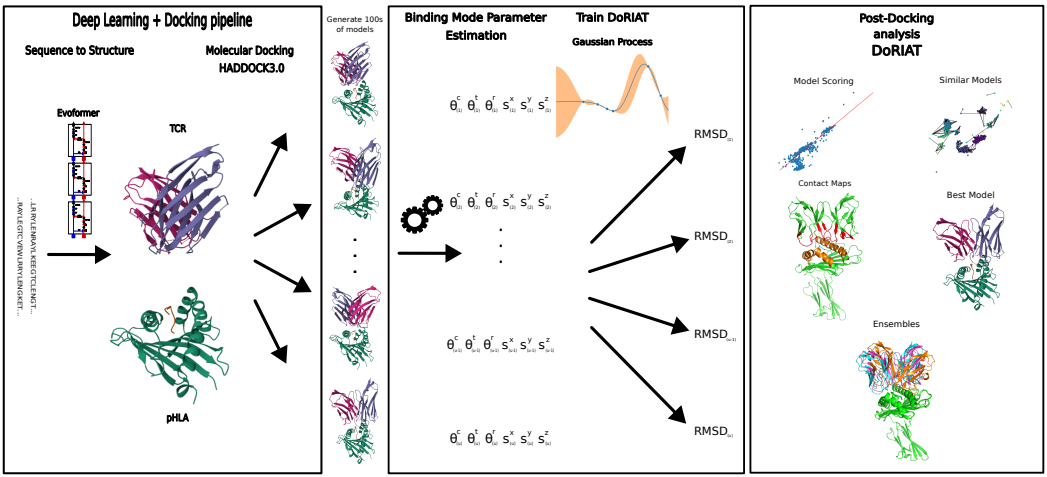
Table 6: **Assessment of model selections using DockQ score for test set and internal TCRs. (Part2)** Best perfroming selection is marked with bold. Cells are color-coded to indicate whether selection is canonical (green) or not (red) according to the thresholds of Tab. 2. DockQ scores are in the $[0,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

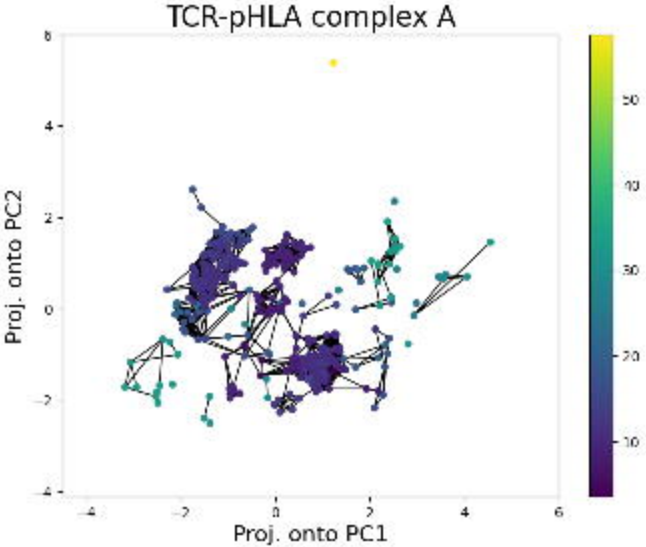

Centroid TCR V module
SS TCR beta
SS TCR alpha
Proje
Centroid MHC helices

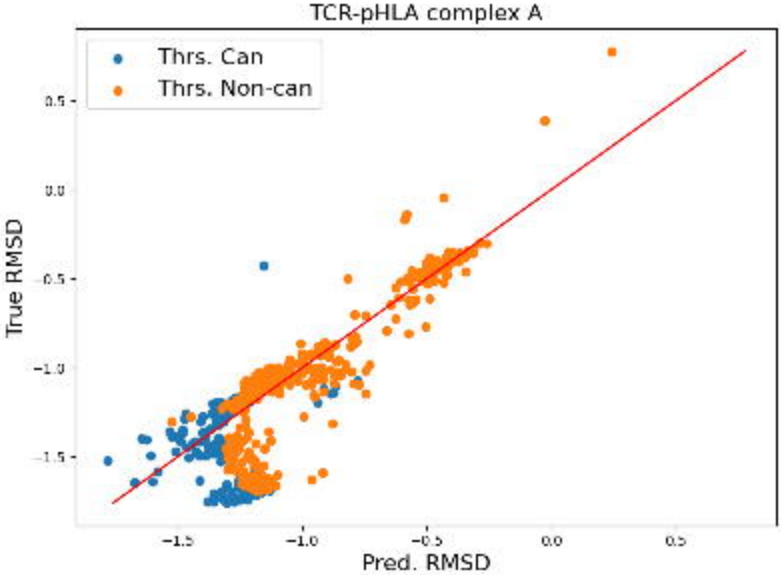

Centroid TCR V module
SS TCR beta
SS TCR alpha
Mhc helices

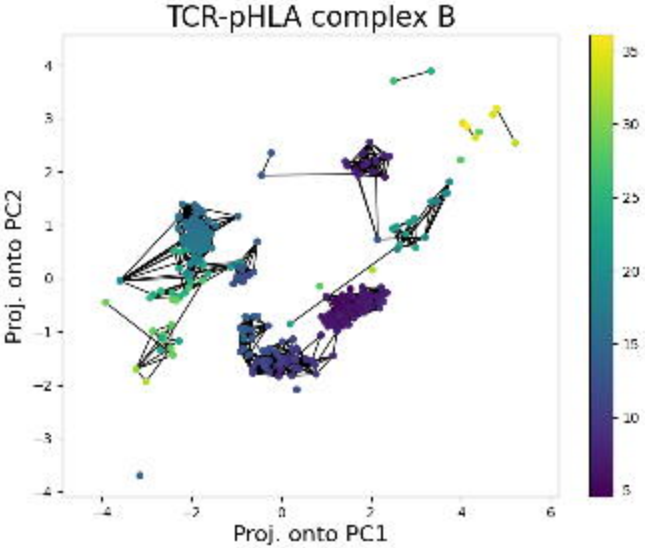

Observations
Trsf. RMSD
Gaussian field
Inputs
Binding Mode Parameters
y_1
y_2
y_*
y_c
f_1
f_2
f_*
f_c
x_1
x_2
x_*
x_c

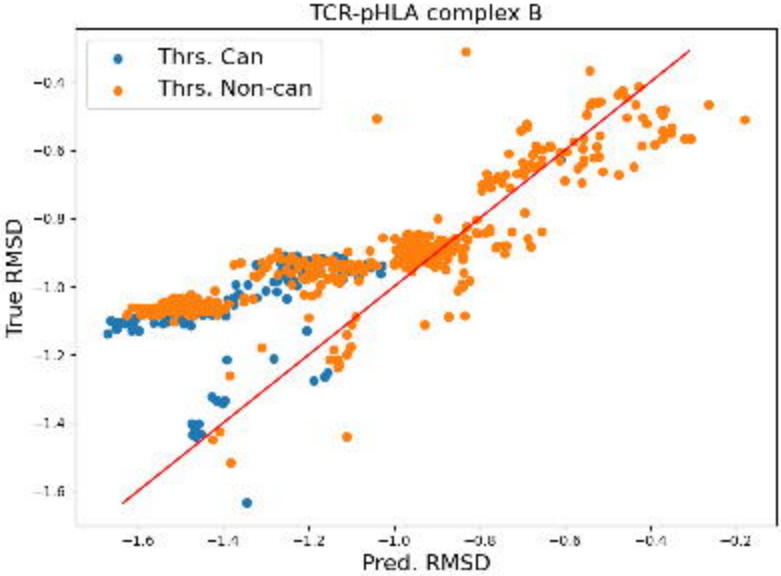

Cumulative Hitrate (%)
Chosen model within top k docking models
0
20
40
60
80
100
50
100
150
200
250
300
350
400
450
DeRIAT
DeepRank-naive GNN
DeepRank-GNN
DeepRank-GNN-esm
GNN DOVF
GMLy^{TN}Dock score

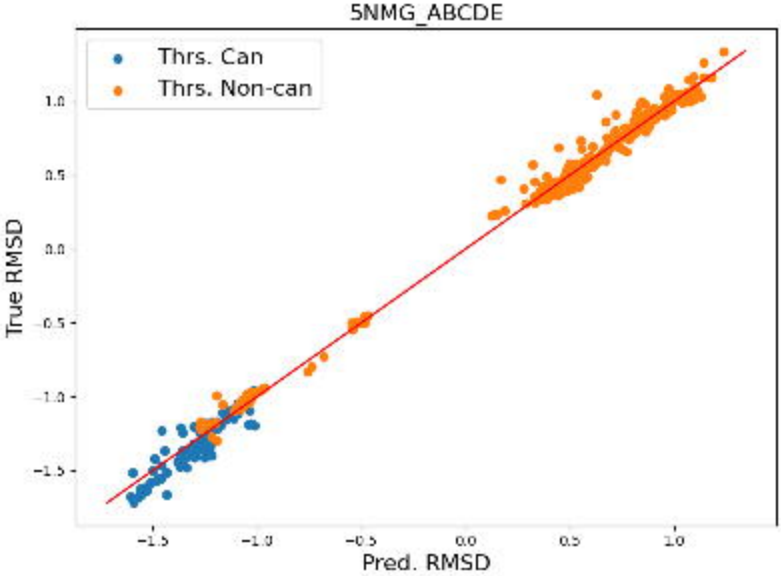

Deep Learning + Docking pipeline
Sequence to Structure
Molecular Docking
HADDOCK3.0
Evoformer
TCR
pHLA
Generate 100s of models
Binding Mode Parameter Estimation
Train DoRIAT
Gaussian Process
RMSD
Post-Docking analysis
DoRIAT
Model Scoring
Similar Models
Contact Maps
Best Model
Ensembles

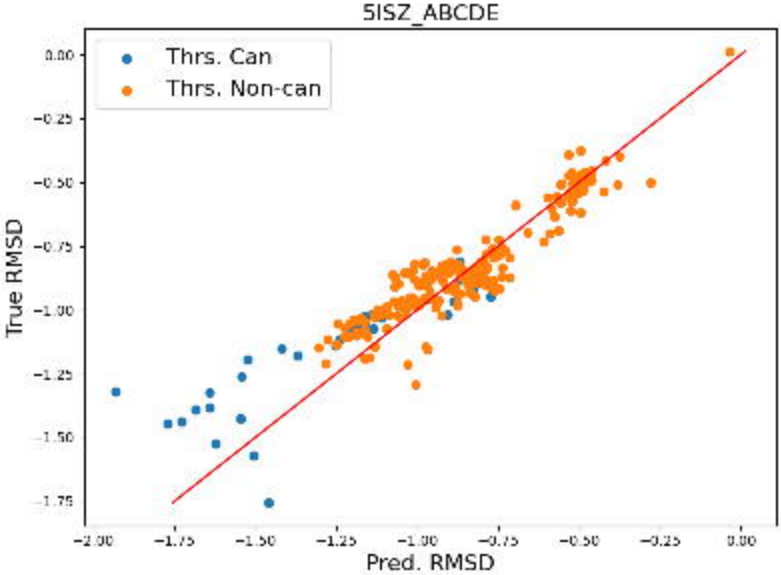

TCR-pHLA complex A
Proj. onto PC2
Proj. onto PC1
6
4
2
0
−2
−4
−4
−2
0
2
4
6
50
40
30
20
10

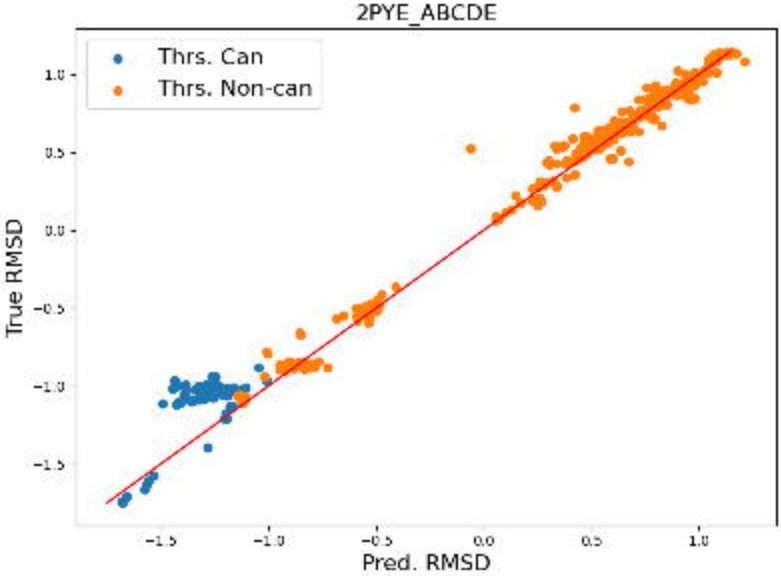

TCR-pHLA complex A
Thrs. Can
Thrs. Non-can
True RMSD
Pred. RMSD
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5

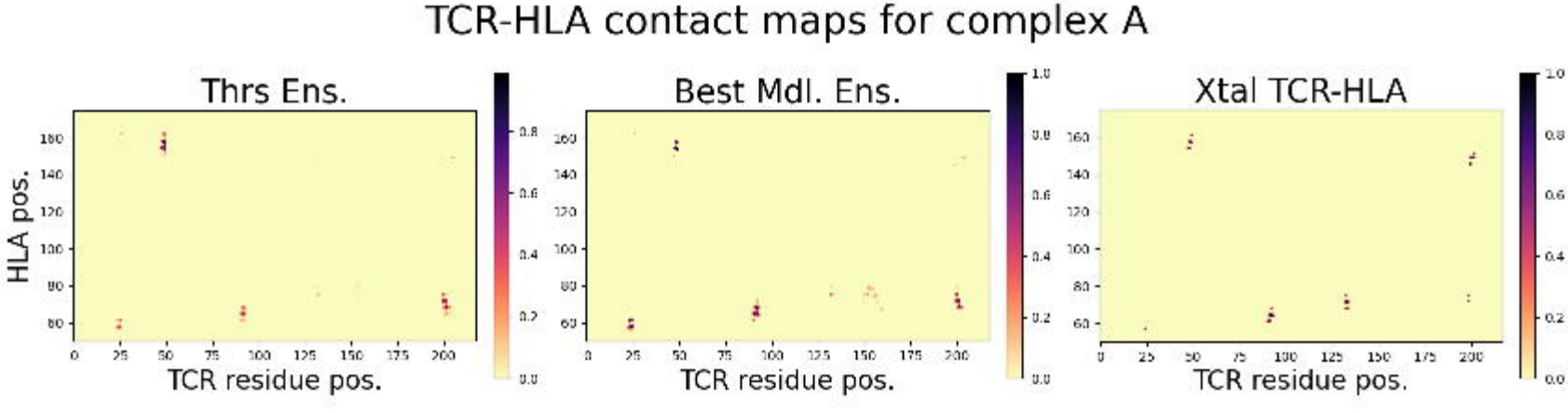

TCR-pHLA complex B
Proj. onto PC2
Proj. onto PC1
4
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
6
35
30
25
20
15
10
5


TCR-pHLA complex B
Thrs. Can
Thrs. Non-can
True RMSD
Pred. RMSD
−0.4
−0.6
−0.8
−1.0
−1.2
−1.4
−1.6
−1.6
−1.4
−1.2
−1.0
−0.8
−0.6
−0.4
−0.2

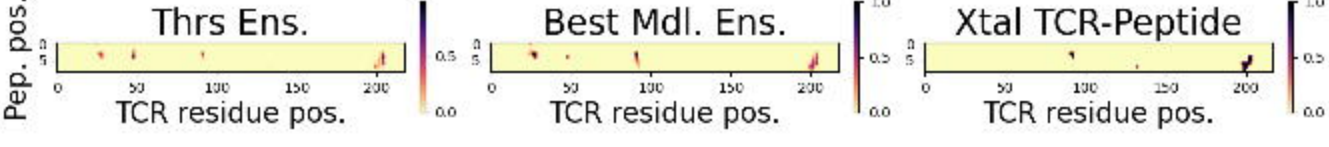

5NMG_ABCDE
Thrs. Can
Thrs. Non-can
True RMSD
Pred. RMSD
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0

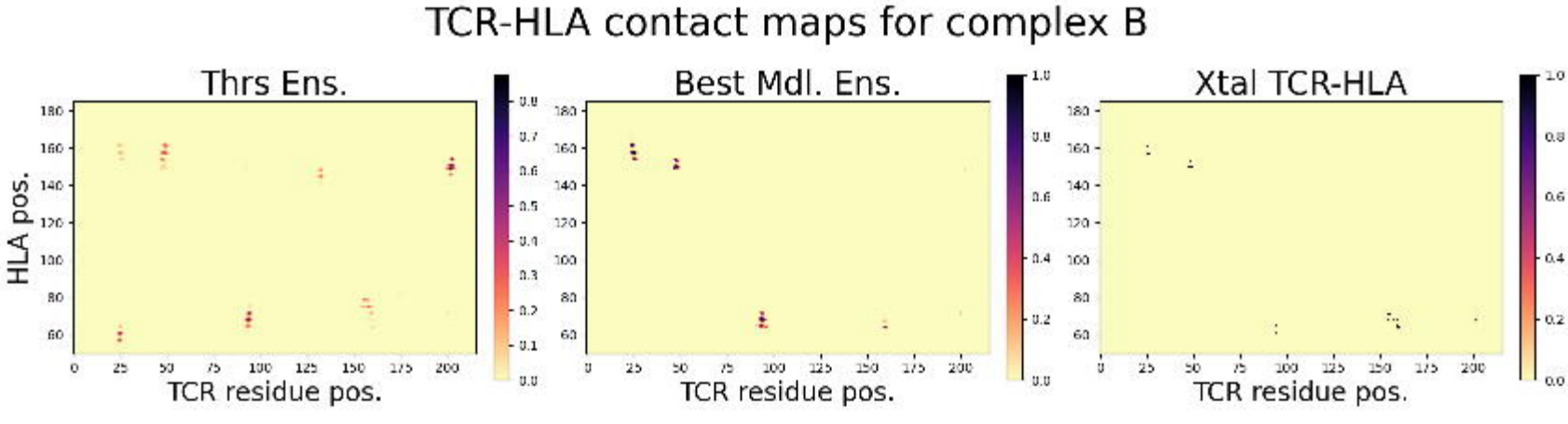

5ISZ_ABCDE
Thrs. Can
Thrs. Non-can
True RMSD
Pred. RMSD
0.00
−0.25
−0.50
−0.75
−1.00
−1.25
−1.50
−1.75
−2.00


2PYE_ABCDE
Thrs. Can
Thrs. Non-can
True RMSD
Pred. RMSD
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0

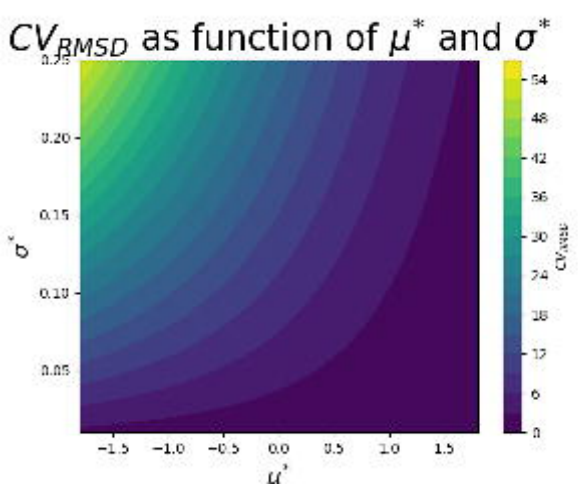

TCR-HLA contact maps for complex A
Thrs Ens.
Best Mdl. Ens.
Xtal TCR-HLA
HLA pos.
TCR residue pos.
TCR residue pos.
TCR residue pos.

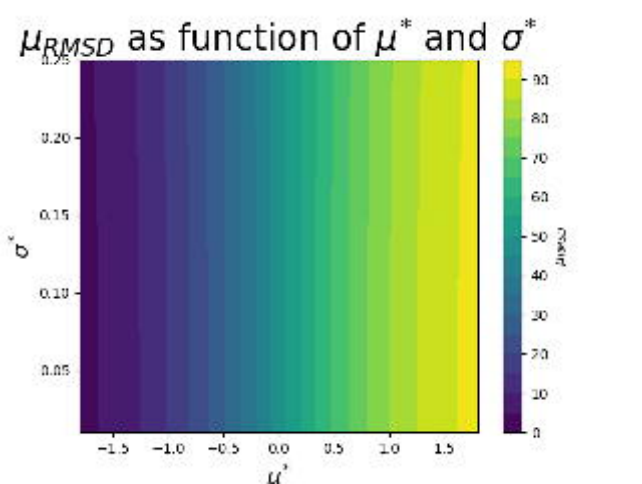
TCR-Peptide contact maps for complex A

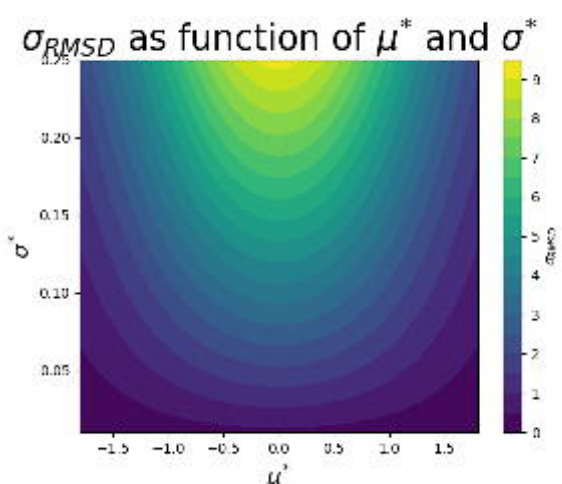



TCR-HLA contact maps for complex B
Thrs Ens.
Best Mdl. Ens.
Xtal TCR-HLA
HLA pos.
TCR residue pos.
TCR residue pos.
TCR residue pos.

TCR-Peptide contact maps for complex B


CV_{RMSD} as function of μ^* and σ^*
0.25
0.20
0.15
0.10
0.05
σ^*
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
μ^*
54
48
42
36
30
24
18
12
6
0
CV_{RMSD}


μ_{RMSD} as function of μ^* and σ^*
0.25
0.20
0.15
0.10
0.05
σ^*
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
μ^*
90
80
70
60
50
40
30
20
10
0
μ_{RMSD}


σ_{RMSD} as function of μ^* and σ^*
0.25
0.20
0.15
0.10
0.05
σ^*
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
μ^*
9
8
7
6
5
4
3
2
1
0
σ_{RMSD}