# APE-Gen2.0: Expanding rapid class I peptide-MHC modeling to post-translational modifications and non-canonical peptide geometries

Romanos Fasoulis,[†] Mauricio M. Rigo,[†] Gregory Lizée,[‡] Dinler A. Antunes,[¶] and Lydia E. Kavraki[*,†]

[†]*Department of Computer Science, Rice University, Houston, 77005, USA*

[‡]*Department of Melanoma Medical Oncology - Research, The University of Texas MD Anderson Cancer Center, Houston, 77054, USA*

[¶]*Department of Biology and Biochemistry, University of Houston, Houston, 77004, USA*

E-mail: *kavraki@rice.edu

### Abstract

The recognition of peptides bound to class I Major Histocompability Complexes (MHC-I) receptors by T-cell Receptors (TCRs) is a determinant of triggering the adaptive immune response. While the exact molecular features that drive the TCR recognition are still unknown, studies have suggested that the geometry of the joint peptide-MHC (pMHC) structure plays an important role. As such, there is a definite need in methods and tools that accurately predict the structure of the peptide bound to the MHC-I receptor. In the last few years, many pMHC structural modeling tools have emerged that provide high-quality modeled structures in the general case. However, there are numerous instances of non-canonical cases in the immunopeptidome that the majority of pMHC modeling tools do not attend to, most notably, peptides that exhibit non-standard amino acids and Post- Translational Modifications (PTMs), or peptides that assume non-canonical geometries in the MHC binding cleft. Such chemical and structural properties have been shown to be present in neoantigens, therefore, accurate structural modeling of these instances can be vital for cancer immunotherapy. To this end, we have developed APE Gen2.0, a tool that improves upon its predecessor and other pMHC modeling tools, both in terms of modeling accuracy and the available modeling range of non-canonical peptide cases. Some of the improvements include: (i) the ability to model peptides that have different types of PTMs such as phosphorylation, nitration and citrullination; (ii) a new and improved anchor identification routine in order to identify and model peptides that exhibit a non-canonical anchor conformation; (iii) a web server that provides a platform for easy and accessible pMHC modeling. We further show that structures predicted by APE-Gen2.0 can be used to assess the effects that PTMs have in binding affinity in a more accurate manner than just using solely the sequence of the peptide. APE-Gen2.0 is freely available at `https://apegen.kavrakilab.org`.

# Introduction

The adaptive immune response is a vital component of the immune system of any organism, seeking to destroy pathogens, viruses or cancer cells.[1] The process in which cytotoxic CD8+ T cells recognize and kill infected cells involves a series of steps; as part of the cells' internal processes, intracellular proteins undergo proteasomal cleavage, resulting in smaller amino acid chain fragments, referred to as peptides. Peptides that are 8-15 amino acids long bind to class I Major Histocompability Complex (MHC-I) proteins, forming a peptide-MHC (pMHC) complex. The pMHC complex is then transported to the surface of the cell, where the receptor of the T cell scans the pMHC complex to assess if the peptide is self or foreign, the latter case resulting in T cell activation.[2] Determining which peptides bind to MHC-I proteins, and which pMHC complexes will elicit an immune response are both longstanding problems in computational biology and immunoinformatics.[3] Accurate identification of good peptide targets has an immediate effect on the efficacy of therapeutics such as peptide vaccination[4] or T cell-based therapies.[5]

Most of the methods that predict the binding affinity of the peptide to the MHC-I, the crucial first step in eliciting an immune response, have long been based on analyzing peptide sequences,[6,7] due to the large amounts of binding affinity and mass-spec data that are publicly available.[8] Methods that determine the immunogenicity of a peptide solely based on its amino acid sequence have also started emerging rapidly.[9,10] In contrast to the availability of sequence data and sequence-based methods, the number of available pMHC crystal structures in public databases is order of magnitudes lower.[11,12] However, there is extensive evidence that structural features stemming from the bound peptide are predictive of properties such as binding affinity,[13] stability[14] and peptide immunogenicity.[15,16] Certain chemical modifications, such as single point mutations[15,17] or post-translational modifications (PTMs) such as phosphorylation[18] can cause severe structural alterations, thus, noticeable effects in T cell recognition, with minimum effect on the peptide sequence.[17] Moreover, there have been studies which employed modeled pMHC structures and subsequently extracted structural features that have shown to be predictive of the aforementioned properties, even exhibiting competitive performance in comparison to peptide sequence-based tools.[13,19,20] It follows that devising algorithms and methodologies that provide accurate geometries of pMHC models is crucial in immune response-related tasks.

There are quite a few examples of pMHC structural modeling tools in the literature that employ a diverse set of algorithms and methodologies to achieve good modeling accuracy.[19] These tools have shown in practice to be successful in providing high-quality structural conformations when compared to ground truth crystal structures. The pDOCK protocol[21] involves two input preparatory steps related to the MHC receptor, as well as the calculation of a docking grid, followed by a single docking (Monte Carlo sampling and scoring) and refinement step (using a Monte Carlo procedure). The refinement protocol of Rosetta FlexPepDock,[22] has been tested on modeling peptide conformations in the MHC binding cleft, reporting near-native predictions ($\leq 2$Å). A crucial step in the Rosetta FlexPepDock is the choice of a proper pMHC template from a database of structures, which is used to produce the new model.[23] Moreover, the ab-initio protocol of Rosetta FlexPepDock[24] has also been recently tested on pMHC modeling.[25] Docktope[26] provides a web-based platform for pMHC docking, employing a combination of molecular docking and an energy minimization protocol that achieves, on average, high quality pMHC structures ($\leq 1$Å). It is limited, however, to only four MHC alleles in total. GradDock[27] uses the highly conserved anchor positions of the peptide, and constructs an ensemble of peptide conformations from half-peptides bound to the anchor positions in the MHC cleft. APE-Gen[28] employs a similar approach, by utilizing an anchor alignment process to define the location of the termini positions. It then constructs an ensemble of peptide conformations using a loop modeling algorithm,[29] without using prior knowledge about the middle portion of the peptide, resembling in this way an ab-initio modeling approach. PANDORA[30] uses homology modeling and a loop optimization approach to provide an ensemble of conformations. Incremental docking methods like DINC2.0[31] have been successfully applied to pMHC modeling, due to the large molecular size of the peptides that bind to MHC-I proteins.[32] Lastly, pMHC modeling using a fine-tuned version AlphaFold[33] has been applied in predicting peptide-binding specificity using structure[34] with comparable results to NetMHCpan4.1, a sequence-based method.[7] While, as previously mentioned, all the pMHC modeling tools in the literature are using a diverse set of methodologies to provide accurate bound peptide conformations, all the approaches (with very few exceptions) can in theory be grouped into two categories: approaches that are using information from a known peptide template, and approaches that follow an ab-initio modeling paradigm and sample peptide backbones.

Another common factor to the pMHC modeling methods and tools mentioned above is that they, with a few exceptions, can only model peptides that exhibit canonical geometries, and to peptides that are comprised of the 20 canonical amino-acids. However, there have been numerous known instances of peptides that do not follow canonical geometries, and/or are composed of one or more chemical modifications. Focusing on peptide geometries, in the canonical case, it is the amino acid in the second position of a peptide sequence that assumes the anchor position in the B pocket of the MHC-I, and the last amino acid in same peptide sequence that assumes the anchor position in the F pocket of the MHC-I. However, many non-canonical cases that do not follow this paradigm have been observed in the literature.[35] For instance, numerous pMHC crystal structures have been observed that show N-terminal extension patterns[36–38] or C-terminal extension patterns.[39,40] The majority of pMHC modeling tools do not identify such cases, and the predicted structures that they provide do not match the non-canonical geometries. For example, while Docktope[26] reports near-native results for the majority of the modeled structures, the authors specifically report that they fail on one case: a peptide variant from the MART-1/Melan-A protein[37] (sequence: LAGIGILTV, PDB code: 2GTW). This peptide adopts a non-canonical, bulged conformation, caused by the leucine in the first position assuming the anchor position in the B pocket. As Docktope lacks the ability to identify such non-canonical cases, it models the peptide as a canonical case, with the Alanine in the second position assuming the anchor position, deviating a lot from the ground-truth as a result. Recently, PANDORA[30] applied NetMHCpan4.1[7] as a proxy, in order to identify such non-canonical cases. The authors show that they provide better structural models for the cases where NetMHCpan4.1 correctly identifies a non-canonical case.

In addition, there is a significant number of peptides from the immunopeptidome that exhibit one or more chemical modifications. Specifically, the topic of peptides presented by MHCs exhibiting PTMs has been extensively discussed.[41,42] In the last few years, a plethora of studies are scanning cell lines in the immunopeptidome, emphasizing the importance that peptides that undergo PTMs hold in the adaptive immune response.[43,44] It is now known that PTMs can have a substantial impact in TCR recognition, and could potentially have an impact in therapeutics, with works emphasizing the fact that many neoantigens exhibit PTMs, absent in normal proteins.[45,46] With the advent of mass-spectrometry and the subsequent increase in pMHC data with PTMs included, sequence-based methods have made breakthroughs in binding affinity/MHC presentation prediction of peptides exhibiting PTMs.[47,48] However, the prediction of the structural effects that these subtle modifications will cause, and how those affect TCR recognition, is still a very challenging problem. Moreover, most of the pMHC structural modeling approaches discussed above are not able to model peptides with PTMs. Recently, the study in[25] extended the Rosetta FlexPepDock ab-initio protocol, in order to support structural modeling of post-translationally modified peptides, making this the only pMHC structural modeling tool that models pMHC complexes including PTMs. While the authors report results averaging below the 2Å threshold, the pMHC modeling running times are reported to be 8-16 hours long, making the method non-applicable in fast pMHC modeling scenarios.

In this work, we present APE-Gen2.0, a fast and accurate pMHC structural modeling tool. APE-Gen2.0 not only improves pMHC modeling performance, but also expands the pMHC modeling repertoire to non-canonical cases, both in terms of peptide geometries and chemical modifications. By employing already established tools and plugins for modeling PTMs,[49] APE-Gen2.0 is able to provide, within minutes, geometrical models for peptides exhibiting PTMs common to the pMHC system, such as phosphorylation, citrullination, nitration and acetylation, among others.[41,42] To prove that APE-Gen2.0 structures are useful in downstream tasks, we experimentally determined binding affinities for a small set of phosphorylated peptides and their non- phosphorylated counterparts. In this dataset, APE-Gen2.0 outperforms sequence-based approaches on the task of correctly identifying positive/negative effects that PTMs cause in pMHC binding affinity. Moreover, by developing a dedicated peptide anchor identification module that correctly identifies non-canonical anchor placements in the majority of the cases, APE-Gen2.0 provides correct structural predictions for non-canonical peptide geometries, as confirmed by the reproduction of crystal structures. Finally, in order to facilitate structural pMHC modeling, APE-Gen2.0 is provided as a web-server and is freely available at `https://apegen.kavrakilab.org`.

# Results and discussion

## APE-Gen2.0 accurately reproduces pMHC structures through a combination of backbone sampling and threading

APE-Gen2.0 is an evolution of the previous version,[28] with modifications and improvements present in multiple parts of the previously established pMHC modeling workflow (Figure 1A). Specifically, the ab-initio modeling process of the previous APE-Gen version is now used in tandem with a peptide backbone threading process, which utilizes geometrical information from the middle portion of the bound peptide. This is depicted in the two branches of Figure 1A. As input, APE-Gen2.0 receives a peptide amino acid sequence, as well as an MHC allotype from any organism. As output, APE-Gen2.0 provides an ensemble of plausible peptide conformations bound to the MHC binding cleft (Figure 1B), as well as a ranking of these conformations, based on protein-ligand scoring functions.[50–52]

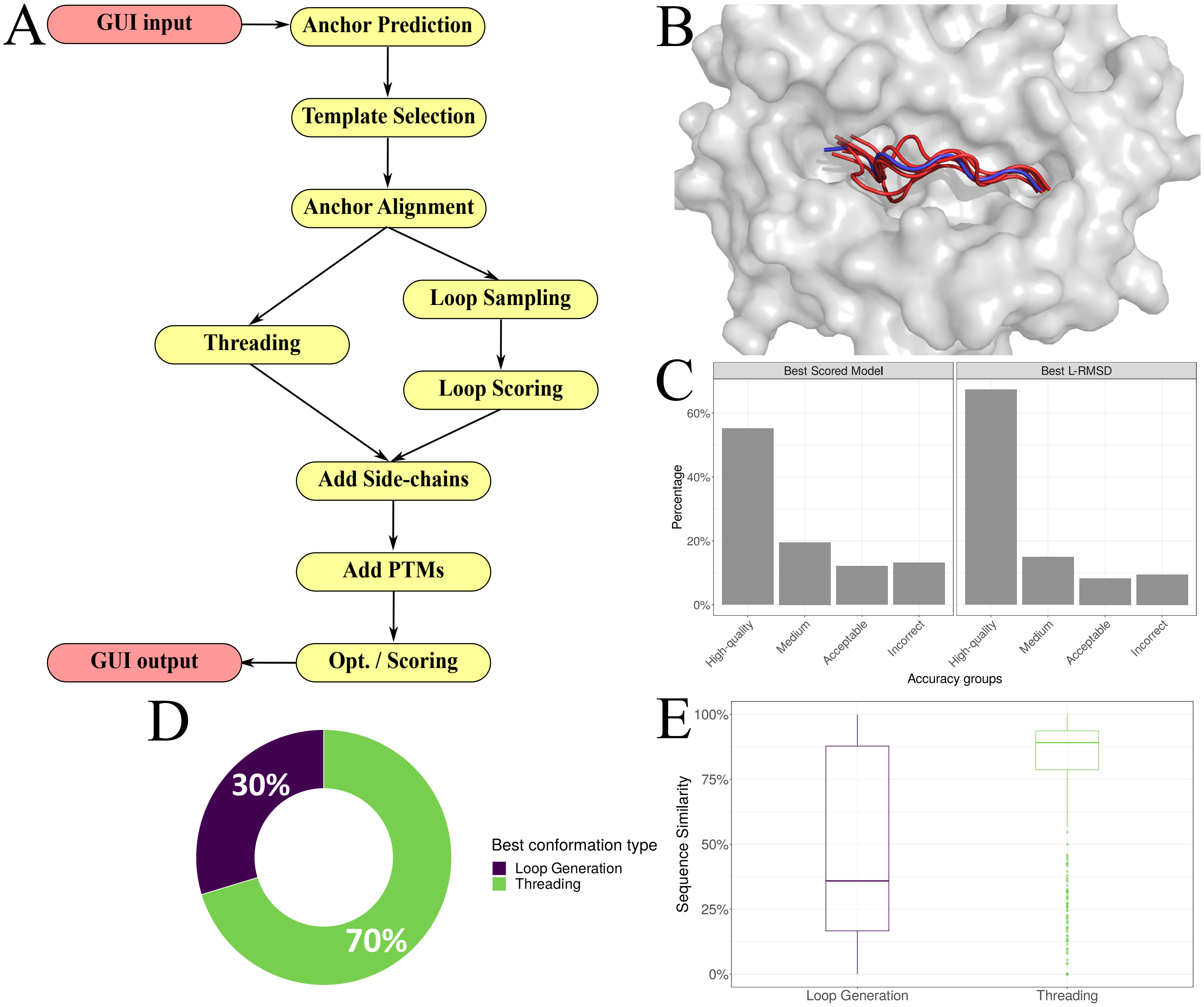


Figure 1: **APE-Gen2.0 produces high-quality ($\leq 1$Å) pMHC models.** **(A)** Overall workflow of APE-Gen2.0. **(B)** Example of the conformational ensemble output of APE-Gen2.0 (depicted in red), given an input peptide sequence and an MHC allotype (example here is PDB code: *1DUZ*, sequence: *LLFGYPVYV*, MHC: *HLA-A*02:01*, structure is depicted in blue). **(C)** APE-Gen2.0 performance on the leave-one-PDB-out cross-validation scenario. Results are shown for both the best scored conformation from the ensemble, and also the conformation with the lowest L-RMSD. **(D)** Around 30% of the best L-RMSD conformations produced by APE-Gen2.0 are a product of the loop sampling and scoring process, and around 70% are a product of peptide threading. **(E)** Comparison of the two backbone reconstruction approaches used by APE-Gen2.0 in regards to the sequence similarity of peptides that are found in pMHC structures in the APE-Gen2.0 database. The Loop Generation box contains the sequence similarity values for the 30% of peptides where Loop Generation performs best in regards to best L-RMSD. The other 70% is contained in the Threading box. The loop sampling and scoring process tends to perform better when the peptide template/templates that are chosen exhibit low sequence similarity to the peptide that is to be modeled.

An improvement of APE-Gen2.0 over its predecessor is the creation of an expanded template structure database, in order for it to be used in the peptide threading step. In particular, we created a database of pMHC crystal structures, collected from public databases (see Methods). The data collection and filtering process resulted in a total of 699 pMHC structures, with no duplicates (see Methods). We subsequently used this template database in order to assess the modeling accuracy of APE-Gen2.0. More specifically, we performed a cross-docking scenario, using the leave-one-PDB-out cross-validation scheme proposed in.[30] We removed structures from the evaluation that contain additional chains, foreign molecules, or any modifications that might alter the structural pose of the peptide (see Methods), resulting in a total of 569 structures for evaluation. L-RMSD results can be seen in Figure 1C (detailed L-RMSD results can be found in the Supporting Information file **Data S1**). More than 50% of the conformations produced by APE-Gen2.0 are high-quality conformations ($\leq$ 1Å). It is interesting that the difference in modeling quality distributions between the best scored model and the best L-RMSD model from the ensemble does not differ by a big margin. That hints to the fact that Vinardo,[50] the default scoring function included in APE-Gen2.0 (see Methods), even though it is designed for smaller ligands, is properly ranking the peptide conformations in terms of L-RMSD closeness to the crystal structure.

It is important to note that, the loop sampling and scoring protocol and the peptide threading protocol do not operate as mutually exclusive, and should be used in tandem. In fact, almost one in three conformations produced by APE-Gen2.0 that are closer to the crystal structure in terms of L-RMSD are produced by the backbone loop sampling process (Figure 1D). We wanted to further investigate the distinct features that this 30% of structures, generated by backbone sampling and optimization and outperforming the peptide threading process, have. Figure 1E shows the peptide sequence identity percentage for when each backbone construction method performs best. Peptide threading performs better than backbone loop sampling and optimization when the sequence of the peptide to be modeled has a large sequence identity with a peptide in the database. On the contrary, when the sequence identity is low, it is more probable that backbone loop sampling yields better results. The above observation necessitates concurrent usage of loop sampling/scoring and peptide backbone threading during modeling.

Lastly, a good choice of backbone sampling to backbone threading ratio ensures that sufficient variability exists in the generated peptide loops, and also ensures the best possible ensemble in terms of accuracy. We found that, if the resulting ensemble generated by APE-Gen2.0 contains 75%-80% conformations stemming from the backbone sampling process, and 20%-25% from the peptide threading protocol, then this results in the best possible L-RMSD to the crystal structure (**Figure S1**). Specifically, the mean best L-RMSD of the workflow that combines both peptide threading and backbone loop sampling is lower than the mean best L-RMSD of the peptide threading only workflow, and much lower from the mean best L-RMSD of the ab-initio modeling. This is true for C$\alpha$, backbone, and full-atom L-RMSD (**Figure S1**). It is worth underlying though, that, by employing backbone sampling, there is a small L-RMSD performance drop when considering only the best scored conformation. This hints that the scoring function used in APE-Gen2.0 is not impairing sampling, but it might sometimes impair the proper ranking of conformations within the predicted ensemble (**Figure S1**). However, the instances of incorrect scoring are quite rare. As such, the backbone sampling to peptide threading ratio value that we employed in the rest of the experiments for this paper was 80%.

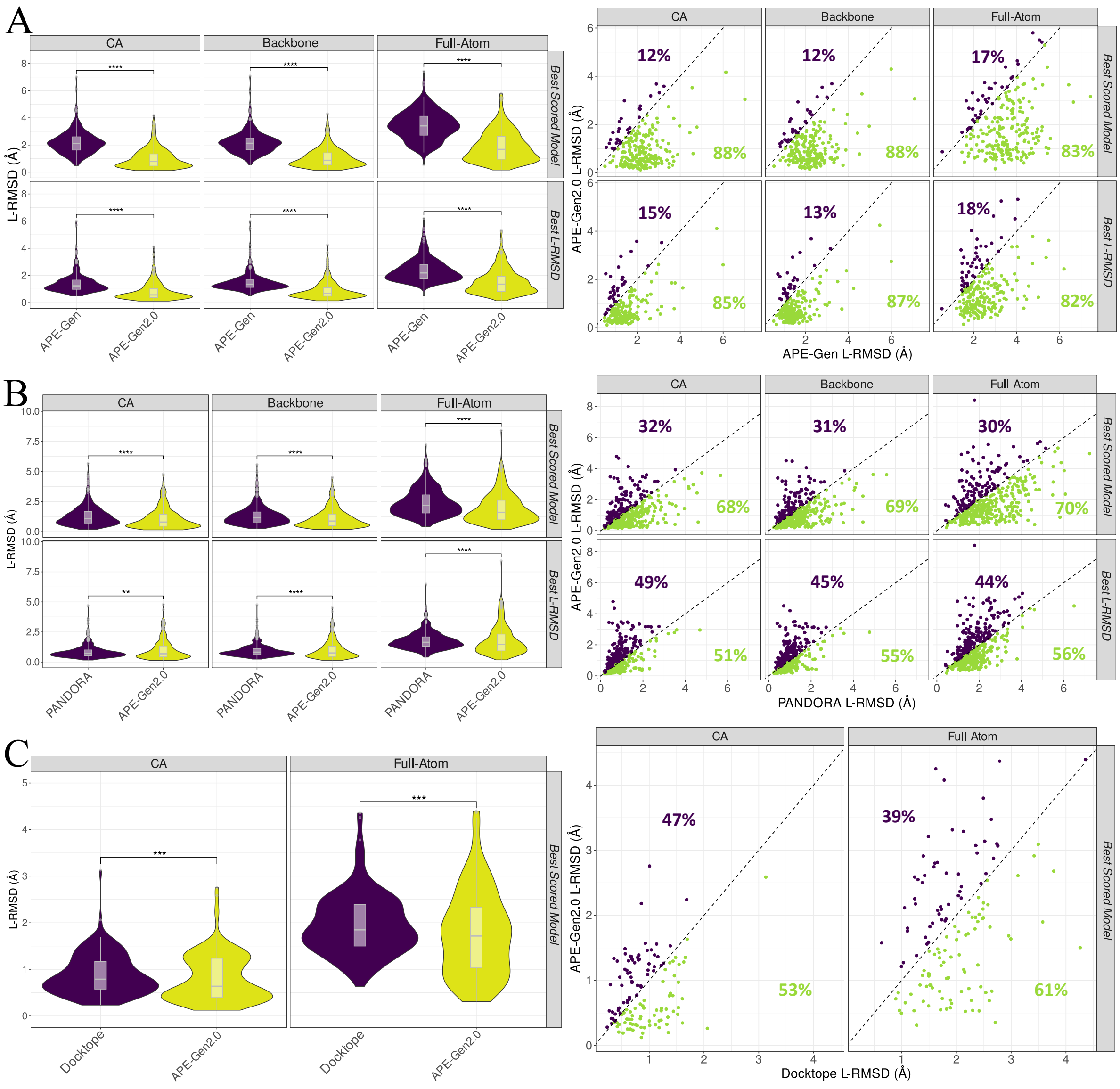


Figure 2: **Leave-one-PDB-out comparison of APE-Gen2.0 to other pMHC modeling tools.** L-RMSD comparison between APE-Gen2.0 and three different pMHC modeling tools from the literature. On the left side, violin plots (with the inserted box plot) depict the distribution of L-RMSD values for each method. On the right side, per-PDB-code L-RMSD comparisons are depicted. Each point represents a unique structure, and its coordinates represent L-RMSD values from APE-Gen2.0 and a different pMHC modeling tool. Percentages in green denote the percentage of structures that APE-Gen2.0 exhibits better L-RMSD results. Percentages in blue denote the percentage of structures that APE-Gen2.0 is outperformed. **(A)** APE-Gen2.0 comparison with its previous version. Both the best scored model and the best model in terms of L-RMSD to the crystal structure are considered. **(B)** APE-Gen2.0 comparison with PANDORA. Both the best scored model and the best model in terms of L-RMSD to the crystal structure are considered. L-RMSD values for PANDORA are taken from.[30] **(C)** APE-Gen2.0 comparison with Docktope. Only the best scored model is being considered in this benchmark (best model results are not provided in the Docktope paper). L-RMSD values for Docktope are taken from.[26]

## APE-Gen2.0 outperforms other pMHC modeling tools

We wanted to assess how APE-Gen2.0 is performing in comparison to other pMHC modeling tools in the literature. We benchmarked APE-Gen2.0 with a selection of pMHC modeling tools that is diverse in regards to the algorithms and methodologies that are employed: APE-Gen, PANDORA and Docktope. The previous version of APE-Gen is using an ab-initio, sampling and scoring approach, without prior knowledge or template guidance for the middle portion of the peptide.[28] PANDORA is a homology modeling-based pMHC modeling tool that is using MODELLER[53] functions and protocols to predict pMHC complexes.[30] Finally, Docktope is a web-based tool that is predicting pMHC structures using a molecular docking/energy minimization protocol.[26] The aforementioned pMHC structural modeling tools were evaluated based on two different methodologies. First, we test APE-Gen2.0 by comparing its performance to the results reported by other tools in the literature, using a leave-one-PDB-out experiment as previously proposed.[30] However, as each tool reports results on different sets of pMHC structures, we also wanted to run all the tools on the same benchmark dataset. For this reason, we constructed a smaller left-out test dataset. We did this by selecting a set of PDB codes that were not found in template databases created by other pMHC modeling tools, in this way, creating an unbiased evaluation (the reader can find more details on the two evaluation schemes in Methods).

Aggregated L-RMSD results for all methods, as well as per-PDB-code L-RMSD comparisons for the leave-one-PDB-out experiment are depicted in Figure 2 (see **Tables S1-S3** in Supporting Information for median and mean L-RMSD values). Emphasizing on the comparison of APE-Gen2.0 to its predecessor (see Figure 2A), it can be seen that APE-Gen2.0 does significantly better, both in terms of the best scored conformation, as well as in terms of the best generated conformation in terms of closeness to the crystal structure. The reason for this is the employment of crystal structures as templates during the modeling process (peptide backbone threading). APE-Gen, it being mostly a loop sampling approach, can potentially produce backbones that are far from the crystal structure, causing many L-RMSD values to increase beyond the acceptable threshold. When compared to PANDORA, APE-Gen2.0 presents a better performance in terms of mean and median L-RMSD. APE-Gen2.0 also outperforms PANDORA when comparing on a per-PDB-code basis in all categories (Figure 2B). While APE-Gen2.0 still outperforms PANDORA in terms of median L-RMSD when considering the best possible conformation to the crystal structure, PANDORA slightly outperforms APE-Gen2.0 in terms of overall mean C$\alpha$ and backbone L-RMSD (Supporting Information, **Table S2**). However, this is not true when considering the best scored conformation. It follows that, while the two tools are mostly comparable when considering the best L-RMSD conformation, PANDORA's scoring function, molpdf,[30,53] is not properly ranking the produced conformations. Lastly, APE-Gen2.0 also outperforms Docktope in all areas (Figure 2C and **Table S3**). When considering C$\alpha$ L-RMSD performance only, Docktope still produces high-quality conformations. However, it is restricted to very few alleles,[26] and it is much slower in terms of performance time. On the contrary, APE-Gen2.0 can provide a prediction for any allele, within minutes.

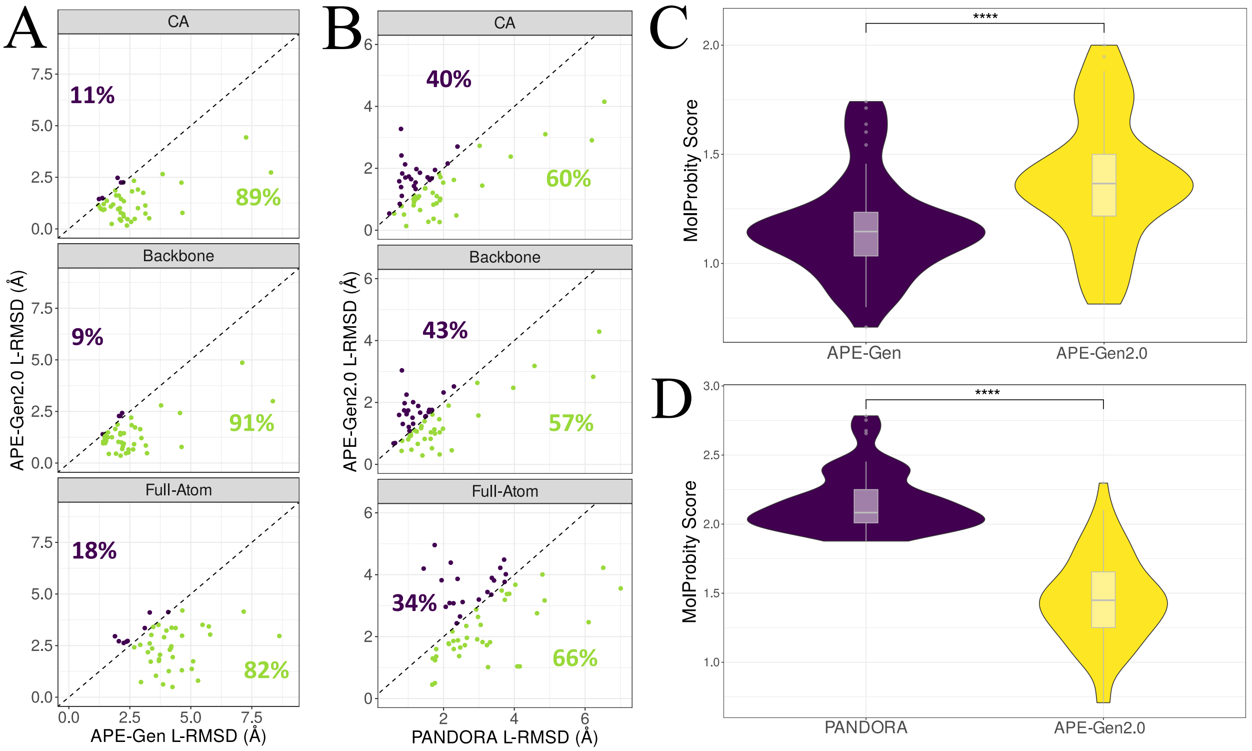


Figure 3: **Comparison of APE-Gen2.0 to other pMHC modeling tools on a left-out test set**. L-RMSD and MolProbity score comparisons between APE-Gen2.0 and two different pMHC modeling tools from the literature. Comparisons are done using top ranking conformation provided by each tool (best scored model). For per-PDB-code L-RMSD comparisons, each point represents a unique structure, and its coordinates represent L-RMSD values from APE-Gen2.0 and a different pMHC modeling tool. Percentages in green denote the percentage of structures that APE-Gen2.0 exhibits better L-RMSD results. Percentages in blue denote the percentage of structures that APE-Gen2.0 is outperformed. Violin plots (with the inserted box plot) depict the distribution of MolProbity score values for each method. **(A)** Per-PDB-code L-RMSD comparison of APE-Gen2.0 to its previous version. **(B)** Per-PDB-code L-RMSD comparison of APE-Gen2.0 to PANDORA. **(C)** MolProbity score comparison of APE-Gen2.0 to its previous version ($p < 0.0001$). **(D)** MolProbity score comparison of APE-Gen2.0 to PANDORA ($p < 0.0001$).

Per-PDB-code L-RMSD comparisons in regards to the left-out dataset are depicted in Figures 3A-B. Comparisons here are done on the basis of the top ranking conformation provided by each tool (best scored model), which is a more realistic scenario when the crystal structure is not known. (best L-RMSD performance on the same test dataset is shown in **Figures S2A-B**). Similar to the leave-one-PDB-out experiment, APE-Gen2.0 outperforms its predecessor (Figure 3A), as well as PANDORA (Figure 3B). When considering the best

L-RMSD conformation, APE-Gen2.0 still outperforms APE-Gen (**Figures S2A**), however, PANDORA performs better (**Figures S2B**). This leads us to the same conclusion as in the leave-one-PDB-out experiment, that is, molpdf, the scoring function that is used internally by PANDORA for pose selection and ranking, is not properly ranking PANDORA's produced conformations, with APE-Gen2.0 performing better in this more realistic scenario (Figure 3B).

Recent studies have suggested that solely looking on L-RMSD values might be misleading, as the aforementioned pMHC structural modeling methods might be introducing physically implausible structures.[54] Therefore, we wanted to quantify such plausibility in APE-Gen2.0 structures, and how they compare to structures generated by other pMHC structural modeling tools. We used Molprobity, a quality assessment tool that validates structures and structural models on a global and on a local scale.[55] More specifically, we used the Molprobity score metric, a single score corresponding to each structural model (see Methods). The Molprobity score (lower is better) is a log-weighted combination of identified clashes, the percentage of Ramahandran outliers, as well as the percentage of side-chain rotamers of bad quality[56] (see also Methods). Interestingly enough, APE-Gen2.0 produces a higher MolProbity score than its predecessor (Figure 3C). However, even though the Molprobity score is not designed to be a threshold-based metric, the Molprobity score threshold of 2.0 has been previously used by the authors of Molprobity for potential loop fragment conformations selection for filling gaps in protein structures.[56] From this perspective, even though higher than its predecessor, APE-Gen2.0 still produces good Molprobity scores, with much better L-RMSD results compared to the previous version (Figure 3A). The same however is not true for PANDORA, as the median Molprobity score for structures generated by PANDORA is greater than 2.0, and much higher than APE-Gen2.0 (Figure 3C). Specifically, manually inspected PANDORA structures exhibit a substantial amount of steric clashes, and a substantial percentage of Ramachandran outliers. The same results and conclusions can be observed when considering the best L-RMSD conformation (**Figures S2C-D**).

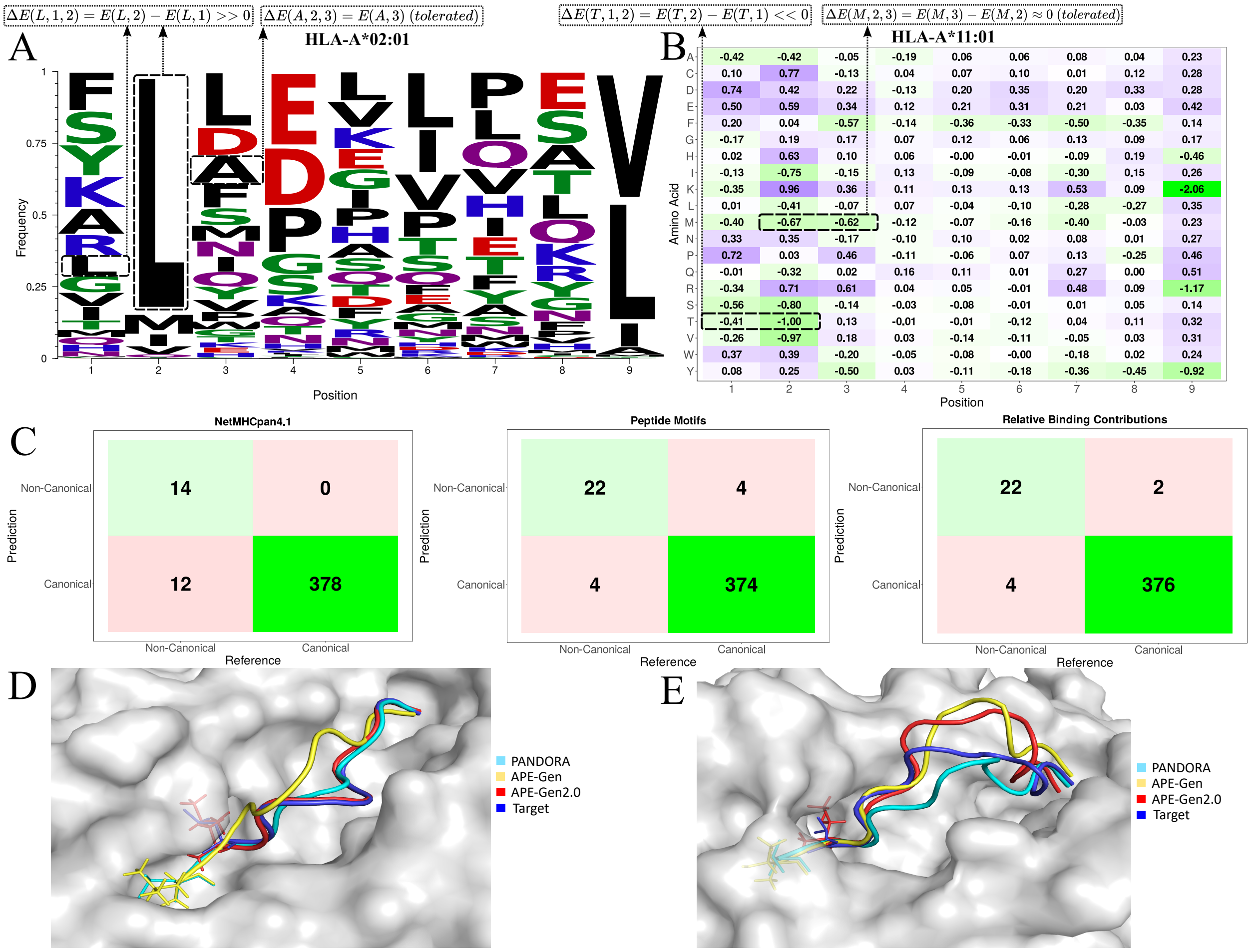


Figure 4: **APE-Gen2.0 correctly identifies and models non-canonical peptide geometries.** **(A)** Simple frequencies from peptide binding motifs can be indicative of non-canonical anchor placements (example here is the MART-1/Melan-A peptide variant *LAGIGILTV* binding to HLA-A*02:01). **(B)** Relative binding affinity contributions taken from[57] can also be indicative of non-canonical anchor placements (example here is Avian Influenza A(H7N9) Virus-derived peptide *TMVMELIRMIK* binding to HLA-A*11:01). **(C)** Confusion matrices of three different methods on non-canonical anchor identification. **(D)** Structure prediction of *LAGIGILTV* bound to HLA-A*02:01 by 3 different pMHC modeling tools (target structure in blue). **(E)** Structure prediction of *TMVMELIRMIK* bound to HLA-A*11:01 by 3 different pMHC modeling tools (target structure in blue).

## An anchor identification module allows detection and modeling of non-canonical peptide geometries

In the majority of the structures deposited at PDB, independently of the peptide length, the N-terminus anchor corresponds to the amino acid in position 2 of the peptide, and the

C-terminus anchor is the amino acid in the last position of the peptide. However, there have been many studies that have reported non-canonical peptide anchor configurations, either in the N-terminus side,[36,37] or the C-terminus side.[39,40] Detecting and correctly modeling these cases can strengthen the accuracy of pMHC modeling tools, and expand the pMHC modeling repertoire. We collected pMHC binding motifs generated by MHCFlurry2.0,[6] as well as relative binding affinity contributions generated by the PMBEC matrix study in[57] (see Methods). As previously proposed by,[35] we also noticed that either position-weight matrices derived by peptide binding motifs, or relative binding affinity contributions, can be predictive of non-canonical anchor conformations. Specifically, in Figure 4A, it is shown that simple differences of amino-acid occurrence frequencies from peptide motifs can identify that, in the case of the MART-1/Melan-A peptide variant LAGIGILTV,[37] it is leucine, the first amino acid in the peptide sequence, that assumes the anchor position. When considering HLA-A*02:01, leucine is prominent in position 2 of the peptide binding motif (corresponding to the B pocket), without being overly frequent in position 1 of the binding motif (corresponding to the A pocket). At the same time, while alanine is not overly frequent in position 3 of the binding motif, it is almost never expressed in the B pocket. As such, we can assume that leucine will overtake the B pocket anchor, resulting in a bulged conformation and a non-canonical geometry. A similar reasoning, from the scope of binding affinity contributions, is followed in Figure 4B with the Avian Influenza A(H7N9) virus-derived peptide TMVMELIRMIK, bound to HLA-A*11:01.[36] From the relative binding affinity contribution matrix, we can see that threonine's absence in position 2 contributes to substantial binding affinity loss. At the same time, methionine can be critical for good binding in position 2, but can also be critical in position 3. As such, we could hypothesize that threonine will assume the B pocket anchor, shifting methionine to the right, resulting in a non-canonical geometry. We subsequently devised a simple algorithm that, by using simple thresholds for relative binding affinity contribution differences, separates canonical/non-canonical cases (see Methods and the Supplementary text section in the Supporting Information material).

These thresholds are kept intentionally simple and linear, in order to avoid overfitting and maintain explainability. A visual interpretation of the relative binding affinity thresholds and the simple boundaries that are formed as a result can be seen in **Figure S3** in Supporting Information.

To quantify the overall improvement of our anchor identification in comparison to NetMHCpan4.1,[7] which is used as a proxy for anchor identification in PANDORA,[30] we benchmarked both approaches in our constructed crystal structure database. As relative binding affinity contributions are not available for all alleles in the database, we kept only MHC structures for which the relative binding affinity contributions are available. This resulted in 404 data points in total, out of which 26 data points exhibit non-canonical anchor conformations. Note that one of these structures (PDB code: *5TRZ*) exhibits non-canonical conformations in both N-terminus and C-terminus, therefore this structure counts as two separate non-canonical data points. Confusion matrices in Figure 4C show that relative binding affinity contribution differences predicts much more non-canonical cases correctly than NetMHCpan4.1. This is reflected on the calculated $F_1$ scores too, as the $F_1$ score for NetMHCpan4.1 is equal to *0.70* compared to the $F_1$ score when using the relative binding affinity contribution matrices, it being equal to *0.88.* We further compared the performance of our simple algorithm based on relative binding affinity contributions compared to the performance of the expert system based on peptide binding motifs (as seen in Figure 4A). Relative binding affinity contributions end up in fewer false positives.

The correct identification of non-canonical anchors in APE-Gen2.0 allowed us to fetch the appropriate peptide template for the anchor alignment step. To clearly show this, we used APE-Gen2.0 to model the two aforementioned non-canonical cases (PDB codes: *2GTW*, *4MJ6*). In Figure 4D we can see that, contrary to APE-Gen, which is not able to properly model non-canonical conformations, APE-Gen2.0 correctly predicts the non-canonical configuration and outputs a correct structural model. It is important to underline here that, because NetMHCpan4.1 cannot identify the correct anchor placement, PANDORA ends up forcing Ala2 in the B pocket. Similarly, in Figure 4E, the homolog that is being fetched by PANDORA, the peptide TIAMELIRMIK, while very similar to the H7N9 virus-derived peptide in terms of sequence, is very different structurally (PDB code: *4MJ5*). In contrast, APE-Gen2.0 correctly identifies the non-canonical anchor and ends up modeling the H7N9 virus-derived peptide with the correct anchor placement.

## APE-Gen2.0 models post-translationally modified peptides in a rapid and accurate manner

By incorporating the PyTMS tool[49] in the APE-Gen2.0 modeling workflow, we are able to model pMHC complexes that include PTMs rapidly and accurately. We collected a small set of peptides bound to MHC complexes from the PDB that exhibit at least one PTM (see Methods). In Figure 5A, performance of APE-Gen2.0 on C$\alpha$, backbone and full-atom L-RMSD on this small set of pMHC complexes is depicted (see **Table S4** in Supporting Information for L-RMSD results per PDB code). We can see that the C$\alpha$ L-RMSD median is below 2Å, indicating that APE-Gen2.0 correctly models pMHC complexes that exhibit PTMs, with only few structures surpassing that threshold. MolProbity scores for all aforementioned structures were also calculated, in order to assess the biological plausibility of the structures. We can observe an obvious separation between phosphorylated peptide structures, and structures exhibiting either citrullination or nitration (see **Figure S4**). We hypothesize that the openMM energy minimization step, which is only supported currently for phosphorylated peptides (see Methods), is crucial in providing structures that are free of steric clashes, Ramahandran outliers and bad quality side-chain rotamers.

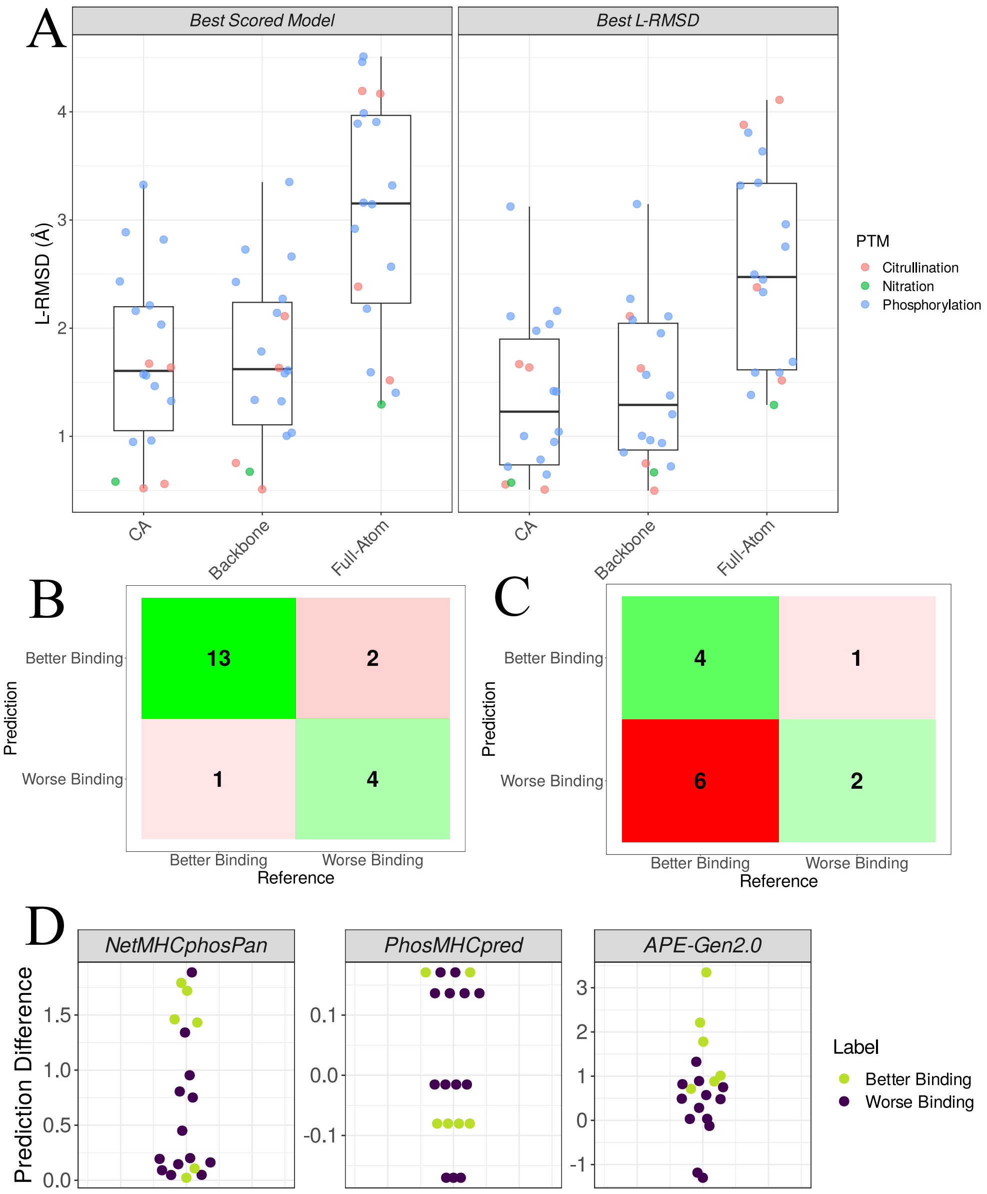


Figure 5: **APE-Gen2.0 modeling of post-translationally modified peptides. (A)** L-RMSD performance on a set of post-translationally modified peptides bound to MHCs. Different colors correspond to different PTM categories. Both the best scored model and the best model in terms of L-RMSD to the crystal structure are considered. **(B)** Confusion matrix denoting the performance of APE-Gen2.0 on the small IEDB dataset of phosphorylated/non-phosphorylated peptides, in the task of identifying positive/negative binding effects in presence/absence of phosphorylation. **(C)** Confusion matrix denoting the performance of APE-Gen2.0 on the small IEDB dataset of citrullinated/non-citrullinated peptides, in the task of identifying positive/negative binding effects in presence/absence of citrullination. **(D)** APE-Gen2.0 is compared to sequence-based methods NetMHCphosPan1.0[47] and PhosMHCpred[48] on the in-house dataset of 19 phosphorylated/non-phosphorylated peptide pairs. The y-axis in the beeswarm plots denotes the difference in predictions for a phosphorylated/non-phosphorylated peptide pair. Ideally, peptide pairs where PTM results in better binding (light green)/ worse binding (dark blue) should be separated.

We also compared APE-Gen2.0 modeling with the only other method that models pMHC complexes with PTMs, Rosetta FlexPepDock,[25] on a subset of phosphorylated pMHC complexes. Backbone L-RMSD results can be seen in **Table S5** for four different pMHC complexes. In general, all methods are competing with each other, with no clear winner. However, it is worth mentioning that the time of modeling with the Rosetta FlexpepDock protocol is reported to be 10–16 hours long,[25] while APE-Gen2.0 can provide a model within minutes.

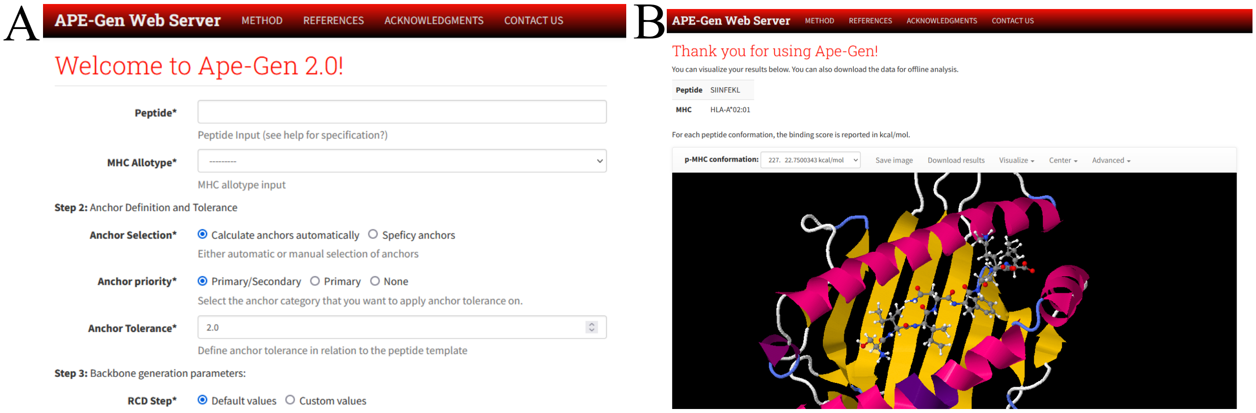
Additionally, we wanted to check whether modeled APE-Gen2.0 structural models hint at downstream effects that the PTM might have on the pMHC complex, particularly on binding affinity. We collected a small set of phosphorylated peptides from the IEDB that also come with their non-phosphorylated counterpart, comprising two alleles HLA-A*02:01 and HLA-B*40:02, for which the effects that phosphorylation has in binding affinity are known.[42,58,59] The aforementioned phosphorylated/non-phosphorylated pairs were modeled using APE-Gen2.0, using a 5-experiment protocol, where the modeling is repeated 5 different times to enhance robustness (see Methods). After modeling, for each phosphorylated/non-phosphorylated pair, we compare the two values resulting from the aforementioned protocol. If the score is better for the phosphorylated peptide in comparison to its non-phosphorylated counterpart, it is predicted that binding affinity is to be enhanced as a result of the PTM, and vice versa. The confusion matrix resulting from this classification can be seen in Figure 5B. APE-Gen2.0, except one case of a False Negative, predicts correctly whether a phosphorylation will result in a better binding affinity. While APE-Gen2.0 incorrectly classifies as positives two of the negative instances, the Area under the ROC Curve (AUC) performance is equal to *0.798*, a value bigger than random prediction (AUC = *0.5*). It is important here to note that critical factors for this performance include both considering the whole ensemble produced by APE-Gen2.0, as well as the openMM optimization step (see Methods). Omitting one of these steps results in a close to, or even below 0.5 average AUC (**Figure S5A**). Additionally, these factors do not just contribute to the better AUC, but to the overall stability of the scoring itself. Specifically, looking at the left part of (**Figure S5A**), it is evident that scores calculated by optimizing the structures through openMM and considering the whole ensemble when scoring are not just the best scores in terms of performance, but also, the most stable scores, resulting in $> 0.5$ AUCs for all 5 experiments.

The same experiment was performed with small set of 13 HLA-A*02:01 citrullinated peptides from the IEDB (see Methods). It is important to note that the force fields parameters that are used for the openMM energy minimization step do not support the citrullinated arginine (see Methods). As such, for the case of citrullination, including any PTM that is no phosphorylation, the optional openMM energy minimization step cannot be performed. This shows in the confusion matrix results in Figure 5C. Even though the calculated AUC given the ensemble of generated APE-Gen2.0 conformation is *0.7*, a value better than random prediction, the results are much more unstable. We hypothesize that the lack of an optimization/energy minimization step in the case of citrullination reduces accuracy. Moreover, AUC values for the citrullinated peptides fluctuate between experiments, with some experiments producing AUC values equal or below 0.5 (**Figure S5A**). This means that the lack of an optimization/energy minimization step not only reduces accuracy, but also stability. Future work will emphasize using force field parameters that support a larger number of PTMs,[60] as relaxed structures show to be more useful for downstream analysis. The full list of citrullinated/non-citrullinated peptide pairs from IEDB, as well as the Vinardo scores for different APE-Gen2.0 runs can be found in **Data S4** in Supporting Information.

To further confirm the potential of using the energy-minimized ensemble of APE-Gen2.0 conformations for downstream tasks, we further employed a small in-house dataset of 19 phosphorylated peptides from 5 different alleles, also including their non-phosphorylated counterpart (see Methods). Results of this comparison can be seen in Figure 5C (see **Data S5** in Supporting Information for the full list of peptides). Similar to the set of IEDB-deposited phosphorylated peptides, scoring the openMM optimized generated ensemble can distinguish between an increase/non-increase in binding affinity in the existence/absence of a PTM. As before, considering the optimized ensemble not only yields the best positive/negative instance separation results (**Figure S5B**), but the most stable ones throughout different experiments in regards to performance (**Figure S5C**). Moreover, we wanted to see how already existing sequence-based binding affinity prediction methods expanded to phosphorylated peptides, specifically, NetMHCphosPan1.0[47] and PhosMHCpred[48] can detect changes in binding affinity due to the existence of a PTM, and how they compare to our scoring protocol. These methods provide an ideal comparison as, contrary to the IEDB dataset, the methods were not exposed to the in-house peptides that we are testing, making this an unbiased comparison. Interestingly enough, sequence-based methods are not able to rank the positive and negative instances as good as APE-Gen2.0. APE-Gen2.0 can almost clearly separate the positive and negative instances, with the positive instances rising mostly to the top of the beeswarm plot (Figure 5D).

## A web server to facilitate pMHC modeling

To further make the tool accessible and facilitate structural pMHC modeling, APE-Gen2.0 is offered as a freely accessible web server at `https://apegen.kavrakilab.org`. The user interface is comprised of two different tabs: the job submission tab and the results tab. In the job submission page (Figure 6A), users can define the peptide sequence and the MHC allotype of their choice. Additionally, users can define specific parameters, such as the preferred scoring function to be used during the molecular docking step, as well as the total number of conformations that they want to generate, among others. All these options are provided in a clean and user-friendly way, to accommodate for both basic and advanced users of the tool. In the results tab, the user can visualize the results generated from the APE-Gen2.0 workflow (Figure 6B). Individual peptide conformations bound to the MHC of choice can be visualized, along with the scoring function results. The whole pMHC structural ensemble can then be downloaded and be utilized in further downstream analysis.



Welcome to Ape-Gen 2.0!

Peptide*

Peptide Input (see help for specification?)

MHC Allotype* --------

MHC allotype input

**Step 2:** Anchor Definition and Tolerance

Anchor Selection* ● Calculate anchors automatically ○ Speficy anchors

Either automatic or manual selection of anchors

Anchor priority* ● Primary/Secondary ○ Primary ○ None

Select the anchor category that you want to apply anchor tolerance on.

Anchor Tolerance* 2.0

Define anchor tolerance in relation to the peptide template

**Step 3:** Backbone generation parameters:

RCD Step* ● Default values ○ Custom values

Thank you for using Ape-Gen!

You can visualize your results below. You can also download the data for offline analysis.

| | |
|---|---|
| **Peptide** | SIINFEKL |
| **MHC** | HLA-A*02:01 |

For each peptide conformation, the binding score is reported in kcal/mol.

p-MHC conformation: 227. 22.7500343 kcal/mol

Figure 6: **The APE-Gen2.0 web server. (A)** The job submission page. Users can type-in the peptide sequence/MHC allotype pair, along with the chosen APE-Gen2.0 modeling parameters. **(B)** The APE-Gen2.0 modeling results page. Users can visualize different conformations from the produced ensemble, as well as download the ensemble for downstream analysis.

# Conclusions

The field of structural modeling of pMHC complexes, ever since its beginnings,[21] has evolved dramatically, with multiple methods and tools being published at an increasing rate.[61] However, there are many pMHC pairs that cannot be modeled accurately by current pMHC modeling tools, most notably, peptides that exhibit PTMs, or peptides that assume non-canonical geometries in the MHC-I cleft. In this work, we have developed APE-Gen2.0, an update from the original version of the tool.[28] APE-Gen2.0 combines and extends the best of the methodologies of previously published pMHC tools to further increase pMHC modeling accuracy (Figure 1C). It also innovates in expanding the pMHC modeling repertoire to non-canonical cases in terms of peptide geometries and chemical modifications.

APE-Gen2.0 provides a conformational ensemble that stems from both peptide backbone threading (resulting backbones are closer to the chosen template) and peptide backbone sampling (resulting backbones diverge from the chosen template). The combination of the backbone sampling and peptide threading processes ensures that, no matter the sequence or structural homology level of the peptide that is to be modeled with other peptide structures in the database, enough conformational space is adequately explored. This is done with no additional computational cost, as peptide threading modifies in-place the peptide amino acids with no added computational burden, and the algorithm that is being used to sample backbones is very fast (see Methods for more details). The presence of peptide conformations stemming from both backbone construction processes provides the best overall accuracy in terms of best L-RMSD (**Figure S1**). Moreover, looking into the best conformations produced by APE-Gen2.0 in the leave-one-PDB-out cross-validation experiment, we see that, although the majority of those are stemming from the peptide threading process, a significant portion of those also comes from backbone sampling (Figure 1D). Thus, the concurrent usage of both backbone construction processes is crucial for the best results in terms of accuracy and diversity of conformations.

It is important though to note that, while the combination of backbone construction processes gives the best L-RMSD results, in some cases, the APE-Gen2.0 scoring function fails to properly rank the conformations produced by the ensemble. This leads to the combination of backbone construction processes underperforming the simple peptide threading process when considering only the best conformation in terms of score (**Figure S1**). As such, future work will involve the creation of more accurate pMHC scoring functions that are accustomed to the intricacies of the pMHC system. The authors of GradDock,[27] as well as the study in[62] have already produced ideas leading to scoring functions that are pMHC specific. While the validation of those previous works is limited to specific alleles or scenarios, this is still a promising research avenue to pursue.

APE-Gen2.0 also excels in identifying cases of peptides that assume non-canonical geometries when bound to MHC-I. We prove that, by imposing simple thresholds in either amino acid occurrence frequencies found in peptide motifs, or in relative binding affinity contributions of amino acids in each peptide position (**Figure S3**) that the majority of cases of irregular geometries can be identified (Figure 4C). In regards to previous approaches attempting to identify those cases, PANDORA[30] employs NetMHCpan4.1 to identify non-canonical anchor configurations. Briefly, given a peptide sequence and an MHC allotype, NetMHCpan4.1[7] determines the 9-mer binding core of the peptide that results in the best binding affinity out of all binding affinities predicted by all possible binding cores. While this can be a good proxy for non-canonical anchor identification, it is not specific to anchor identification. The previously discussed example, the infamous 9-mer MART-1/Melan-A peptide variant bound to HLA-A*02:01[37] (Figure 4A), exhibits a non-canonical configuration in the N-terminus part, where, it's the leusine in position 1 that acts as the anchor in the B-pocket. This, in principle, cannot be identified by NetMHCpan4.1 as a non-canonical configuration, as the 9-mer binding core of the peptide with the biggest binding potential is the peptide itself. For similar reasons, the Avian Influenza A(H7N9) Virus-derived peptide TMVMELIRMIK, bound to HLA-A*11:01[36] (Figure 4B), although it exhibits the same non-canonical configuration as the melanoma peptide, it cannot be identified by NetMHCpan4.1 as non-canonical. Our anchor identification module identifies both of these cases as non-canonical, and APE-Gen2.0 creates modeled ensembles that follow the predicted non-canonical geometry of the crystal structure (Figure 4D and Figure 4E). Moreover, inspired by sequence-based consensus methods that combine many peptide binding predictors and have shown better results in peptide binding prediction and target identification scenarios,[63–65] future work will include combining different sources of peptide motif frequencies and matrices (stemming from different binding affinity predictors) and relative binding affinity contributions, in order to construct an even more robust anchor prediction module. It is also important to note that our proposed anchor identification module is specialized in identifying N-terminus or C-terminus anchor positions, however, it cannot explicitly identify secondary anchors found in the middle portion of the peptide. These secondary anchors have shown to arise in certain mouse alleles, as well as certain human alleles such as the HLA-B*08:01 allele.[66,67] Even though finding the appropriate pMHC template during modeling results in APE-Gen2.0 correctly modeling secondary anchors in the majority of cases, we plan to expand our anchor identification module to explicitly identify secondary anchors. It has been previously shown that peptide binding motifs exhibit conservation in secondary anchor positions,[66] so it is highly probable that peptide binding motif information or relative binding affinity contribution information can also be exploited for secondary anchor identification.

Lastly, to our knowledge, by using PyTMS,[49] APE-Gen2.0 is the first method to offer a rapid modeling protocol of post-translationally modified peptides bound to MHC-I. We showed that APE-Gen2.0 can provide near-native ($\leq 2$Å) conformations of phosphorylated, citrullinated and nitrated peptides within minutes (Figure 5A). However, as previously mentioned, while providing the structures is in itself important, proving that these structures can be of use in downstream analysis and tasks is equally important. To this end, we collected two datasets of phosphorylated peptides and their non-phosphorylated counterparts, a dataset from IEDB and a smaller, in-house dataset. In both datasets, APE-Gen2.0 provides correct predictions in regards to the effects of the phosphorylation to the binding affinity (Figure 5B and Figure 5D). Surprisingly enough, on the smaller in-house dataset and on the same task, APE-Gen2.0 even outperforms sequence-based tools that have been explicitly trained on the task of binding prediction of phosphorylated peptides[47,48] (Figure 5D). This shows, even on a small scale, without any explicit training or fine-tuning as done by sequence-based methods, that structural information obtained from pMHC models can be of invaluable help in downstream analysis. It is important to note that a huge factor in obtaining these results was the use of the APE-Gen2.0 ensemble output, combined with the openMM energy optimization step (**Figure S5**). As it stands, the energy minimization step can only be performed on peptides with canonical amino acids or phosphorylated peptides (see Methods). This partially explains the worse performance on the same task in the citrullinated peptides scenario (Figure 5C). As such, future work will emphasize on using additional force field parameters,[60] in order to expand the openMM energy minimization step to other PTMs. Additionally, as there have been already examples in the literature that use pMHC modeled structures to learn binding affinity[13] or immunogenicity labels,[20] future work will emphasize modeling a larger dataset of phosphorylated peptides and use it in downstream tasks. Given that the scoring function alone could discern effects of the existence/absence of phosphorylation to the binding affinity (Figure 5B and Figure 5D), we hypothesize that further fine-tuning scoring functions on specific binding affinity labels of phosphorylated/non-phosphorylated peptides can further improve performance. Future work will also include expanding the PTM repertoire of APE-Gen2.0. Currently, APE-Gen2.0 uses PyTMS, a fast and accurate tool that has however a finite selection of PTMs.[49] As previously done in,[25] we plan to expand APE-Gen2.0 to more PTMs. Lastly, future work will also include the expansion of APE-Gen2.0 to class-II pMHCs. Specifically, we are interested in modeling post-translationally modified peptides bound to class-II MHCs, as PTMs are quite prominent in the class-II MHC.[68,69] The field of studying post-translationally modified peptides bound to MHCs and their clinical relevance has started to flourish,[43,44] and we hope that structural modeling of these peptides in a fast and accurate way will take center stage, and further advance the field.

# Methods

## Template collection and curation

APE-Gen2.0 relies heavily on a meticulously curated and labeled database of pMHC structures. The following section describes the collection, filtering, and labeling of these structures that are used as templates in the pMHC modeling process.

### pMHC Structure Collection

A collection of pMHC class-I structures was acquired from the IMGT/3D-structureDB database.[11] Namely, the IMGT receptor description that was chosen was MH1, resulting in 1084 entries (tested on February 2nd 2023). Dubious crystal structure files that result in parsing errors are manually inspected, and subsequently removed if they are deemed to not be adequate for further processing. Crystal structures with missing peptides or missing peptide residues were also removed. For each remaining file, we follow a modified pipeline to the one already proposed by:[30] i) Duplicate chains stemming from multiple copies of the biological assembly are removed, ii) files are renumbered in terms of atom and residue indexes using pdb-tools,[70] and iii) the $\beta$2-microglobulin is removed, as it does not contribute to the proposed pMHC modeling process.

Moreover, we identify the following categories of crystal structures where a factor other than the MHC molecule or the peptide itself contributes to the conformation of the peptide: i) the peptide residues contain one or more post-translational modifications (PTMs) or altered/non-canonical amino-acids that can lead to an altered peptide pose in comparison with a non-altered version of the peptide, ii) there is an additional small molecule in the pMHC binding cleft in close proximity to the peptide that might affect the peptide's structural pose, and iii) other chains that can be present in the crystal structure, e.g. T-cell receptors (TCRs), Killer-cell immunoglobulin-like receptors (KIRs), antigen processing (TAP)1/2, tapasin, calreticulin, ERp57, among others, that have been shown to affect the peptide's pose.[71] We opted in keeping all of the aforementioned structures in the APE-Gen2.0 crystal structure database, as it was shown that they were helpful as templates during the template selection step (see **Figure S6** in Supporting Information). For structures belonging to categories ii) and iii), we manually removed the small molecule/other chains. For any peptide exhibiting a PTM/altered/non-canonical residue, we reverted its residue to a canonical form based on the *Parent* residue entry in the Protein Data Bank (PDB).[12]

### MHC allotype and peptide identification

To identify the MHC allotype that is present in the PDB file, unlike the study in,[30] we did not use the IMGT/3D-structureDB nomenclature, as there were valid PDB files that had missing G-ALPHA1 and G-ALPHA2 entries (corresponding to the two $\alpha$-helices). Instead, we extracted the MHC $\alpha$-chain sequence from the PDB and performed a pairwise sequence alignment to all MHCs with known sequence. The MHC allotype with a sequence resulting in the greatest similarity to the sequence found in the PDB file was chosen as the MHC allotype label for this file. The peptide sequence, as well as its length, was also extracted from the PDB file. As previously described, PTM/altered/non-canonical residues in the peptide sequence were converted to canonical ones based on their parent form. Finally, if there are more than one structures with identical peptide sequence and MHC allotypes, only one is kept, namely, the one that has the better resolution. As, in most cases, such structures are almost identical when super-imposed, this was done to keep the database of crystal structures diverse, but more importantly, to avoid data leakage during the leave-one-PDB-out cross validation evaluation. The aforementioned data collection, filtering and labeling process resulted overall in 699 distinct pMHC structures.

### Anchor Identification and Labeling

A major decision factor for the selection of the peptide template is the anchor placement of the peptide residues in the cleft. As such, there was a need to develop a protocol for identifying and labeling, given a crystal structure in the database, the peptide residues that assume the anchor positions in the MHC binding cleft. We identified the following features that are descriptive of a peptide anchor:

- *Relative accessible surface area*: For each peptide residue, the Relative accessible Surface Area (RSA) is defined as:

  $$RSA_i = \frac{SASA_i}{Max_{SASA}(i)}$$

  where $i$ is a given peptide residue and $SASA_i$ is the Solvent Accessible Surface Area for this residue, denoting the surface area of the residue that is accessible to a solvent. $Max_{SASA}(i)$ denotes the maximum value that SASA can receive for a given residue $i$. This normalization results in RSA values being comparable among different residues that might have different side chain volumes that could skew the SASA value. Applied directly to peptide residues, a higher RSA would imply that a peptide residue is more exposed, while a low RSA value would imply that the peptide residue is found deep within the cleft, and is likely to be an anchor. RSA is computed using the NACCESS 2.1.1[72] tool, with choosing the default parameters and utilizing a standard 1.4 Å radius probe. The $Max_{SASA}$ values are taken from the default parameters of NACCESS. When calculating the RSA for each peptide residue $i$, the rest of the residues were removed from consideration, as neighboring peptide residues to the residue $i$ are sure to affect the SASA surface.

- *Distance to the $\beta$-sheet calculation*: Given that the *beta*-sheet floor formed by the two polypeptide $\alpha$-chains is roughly planar,[73] and the bound peptide is positioned roughly in parallel to the $\beta$-sheet, we can assume that peptide residues that are closer to the $\beta$-sheet are more probable to be anchors. Specifically, we used the *z-dist* formulation by[74] to calculate, for each peptide residue, its distance to the $\beta$-sheet floor.

For each pMHC structure, two major peptide anchors are assumed. The first anchor position is located in the N-terminus side of the peptide, and it is always placed in the B region of the MHC binding cleft.[75] The other anchor is located the C-terminus side of the peptide, and it is always placed in the F region of the MHC binding cleft.[75] Scanning through the APE-Gen2.0 crystal structure database, it can be inferred that, for the N-anchor positions, it is always the case that it is one of the first three residues of the peptide that take the anchor position in the B region of the MHC. Similarly, for the C-terminus anchors, it is always residues from position 7 of the peptide onward that compete for the anchor position in the F region, independently of the peptide length.

Since the crystal structure database is too large for manually inspecting and defining the anchors, the following protocol was devised for anchor identification:

- The C$\alpha$ and all atom z-dist was calculated for both the N-terminus side (first three residues) and the C-terminus side of the peptide (position 7 of the peptide onward). The two residues exhibiting the minimum C$\alpha$ and all atom z-dist, one for each residue group (N, C) are anchor candidates. The very few times C$\alpha$ and all atom z-dist end up in different candidates, manual inspection on these crystal structures is performed to determine the closest residue to the $\beta$-sheet floor.

- RSA was calculated for the same residue groups (N, C). The residue with the minimum RSA is considered an anchor candidate.
- A residue is considered an anchor if it is a candidate both in terms of z-dist and RSA. Manual inspection to determine the major anchors is only necessary when the z-dist and RSA consensus results in two different candidates.

Out of 699 structures in our crystal structure database, the above protocol results in 41 non-canonical cases, which we manually inspect to confirm that they are actual non-canonical cases.

## APE-Gen2.0 workflow

The workflow of APE-Gen2.0 can be seen in (Figure 1A). It is composed of many individual parts, which all contribute to the final ensemble of conformations that are produced as an output of APE-Gen2.0. In the following subsections, we will examine the individual parts of the workflow in more detail.

### Anchor prediction module

We collected relative binding energy contribution matrices from[57] for all supported alleles. The relative energy contribution matrices, of size $20 \times N$ (20 being the twenty canonical amino acids and $N$ being the peptide length), denote the binding affinity contribution of a specific amino acid *aa* in a specific position *pos*, and are calculated as specified in.[57] Similarly, peptide motif frequencies where collected from,[6] with the $20 \times N$ matrix denoting the frequency of an amino acid *aa* in a specific position *pos*.

We observed that mere relative binding energy contributions or binding affinity motif frequencies correlate with anchor placements. We subsequently developed a formal strategy to extract features from these matrices that are predictive of anchor placements. More specifically, we defined the energy difference feature $\Delta E(aa, pos, pos')$:

$$\Delta E(aa, pos, pos') = E(aa, pos') - E(aa, pos) \tag{1}$$

Values of the energy function $E$ are taken from the relative binding contribution matrices (or frequencies extracted from peptide motifs).

For the anchor prediction module, we have designed an expert system for identification of possible non-canonical anchor configurations, based on the energy difference $\Delta E$. For each non-canonical candidate position (positions 1 and 3 for the N-terminus side and positions 7 up to position (length of peptide - 1) for the C-terminus), we set simple and interpretable $\Delta E$ thresholds that, when satisfied, result in a non-canonical configuration. In Supporting Information, the reader can find the expert system using the relative binding affinity contributions from the PMBEC work in[57] (similar thresholds where defined from peptide motifs as comparison, and are not shown in the manuscript).

### Template selection

Similar to the previous version,[28] APE-Gen2.0 just needs the amino acid sequence of the peptide and an MHC allotype (or sequence) in order to predict the bound pMHC structure. To achieve this, it needs one (or more) peptide template (e.g. an already experimentally defined crystal structure) as a prior for the prediction of the 3D conformation of the peptide in question, as well as an MHC template for the receptor. In this section, we describe in more detail the protocols used for selecting a peptide template and an MHC template that are to be used for predicting the final ensemble of pMHC conformations.

- *Peptide template*: The choice of peptide templates is performed through a pipeline of different filtering and scoring mechanisms:

1. **Anchor filtering**: The anchor configuration of a peptide in the MHC binding cleft is a vital component that contributes majorly in the final peptide conformation. If the anchor configuration is known, the peptide template that is to be used to guide the pMHC modeling should exhibit the same anchor configuration. Using the anchor prediction module, given the peptide sequence and the MHC allotype, we predict the major anchor placements that the peptide will have in the MHC binding cleft. We then define *the major anchor difference*, calculated as the positional, index difference between the anchor in the C-terminus part of the peptide and the anchor in the N-terminus part of the peptide. As an example, assuming a 9-mer with canonical anchor configuration, its major anchor difference would be $9 - 2 = 7$. We subsequently filter and only keep peptide templates that exhibit the same anchor difference calculated from the anchors output of the anchor prediction module. In cases when the anchor prediction module fails to predict the anchor placement (e.g. for alleles that peptide motifs or relative binding affinity contributions are not provided), no crystal structures are filtered out, and all are considered for the next steps in the peptide template selection.

2. **Filter by organism**: Given distinct geometrical differences between alleles of different organisms (for example, human vs. mice alleles), no templates corresponding to different organisms than the MHC allotype given as input are considered. Crystal structures from different organisms are considered if and only if there are no crystal structures in the template database that correspond to the organism that the MHC allotype input belongs to.

3. **Template similarity**: For the remaining peptide templates that passed through the anchor filter and the organism filter discussed above, the best candidate needs to be selected. The best candidate is based on two distinct similarity measures: A) the similarity of the MHC allotype in question to other MHCs in the crystal

structure database and B) the similarity of the peptide sequence to-be-modeled with other peptide sequences in the crystal structure database.

In regards to *Allele Similarity (AS)*, there are two main similarity measures when comparing two alleles; similarity in terms of sequence, and similarity in terms of binding preferences. Although there is obvious overlap between the two, there are also distinct differences.[76] In this work, we hypothesize that, for two MHCs that have similar binding preferences (bind to similar peptides), it is highly probable that the MHC binding cleft, including the possible conformation of the peptides, also exhibit similarities in terms of geometry. As such, for a given MHC allotype, MHCs from the crystal structure database that exhibit similar binding preferences are given priority for the peptide template selection. To quantify the similarity based on binding preferences, we download the peptide binding motifs from MHCFlurry2.0.[6] We define MHC similarity as the similarity between two MHC binding motifs. Specifically, we interpret a peptide binding motif as a $20 \times N$ normalized frequency matrix ($N$ being the peptide length). For two different matrices $P$ and $Q$ corresponding to two different alleles, we define their allele similarity $AS(P, Q)$ as:

$$AS(P, Q) = 1 - \frac{1}{\sqrt{2}} \|\sqrt{P} - \sqrt{Q}\|_2$$

The second part of the equation is the Hellinger distance[77] between matrix $P$ and $Q$. $AS(P, Q)$ is valued from [0-1], higher values denoting bigger similarity between motifs. As such, motif similarities between the MHC allotype to be modeled and the database of candidate templates are calculated in a pairwise manner. Templates having scores closer to 1 are given the priority. If the MHC allotype in question is identical to one of the MHCs in the template database, that template takes the most priority, since the similarity value is the max value of 1.

As there are peptide binding motifs corresponding to different peptide lengths,[6] in practice, allele similarity depends also on the peptide length $N$ in question, $AS(P, Q, N)$.

In regards to *Peptide Similarity (PS)*, it has been shown that, given an MHC allotype, similar peptides in terms of sequence are also similar in structure.[16,78] As such, the peptide template to be selected must also have as high peptide sequence similarity as possible to the peptide that is to be modeled. As previous work has suggested,[30] the peptide sequence to be modeled is aligned in a pairwise manner with all the peptide sequences in the crystal structure database. The BLOSUM62 matrix[79] is used to score the pairwise alignment of the peptide sequences. However, it is important to underline that the alignment has to be structurally aware, meaning that the anchors of two peptides sequences need to be correctly aligned. As such, we perform a pairwise sequence alignment with anchor constraints. This is done to avoid giving high scores to peptide templates that are very similar in terms of sequence alignment but different structurally. To employ the anchor constraint pairwise sequence alignment protocol, we use the anchors given by the anchor prediction module. When the anchor prediction module is not available for predictions (for example, no relative affinity contribution matrix is available for a rare MHC allotype), then a simpler version of the pairwise sequence alignment is performed, but with appropriate gap penalties to avoid structurally incorrect alignments.

Finally, the $AS$ score and the $PS$ score are averaged to create a template similarity score $TS$:

$$TS = \frac{AS + PS}{2}$$

Crystal structures available in the database are ranked in decreasing order, and the template with the highest $TS$ score is chosen to be the peptide template. In case of ties, one of the top scoring peptide templates is randomly chosen.

- *MHC template*: The choice of MHC template will depend on whether the MHC allotype input exists in our crystal structure database, as well as the peptide sequence input:

    1. **MHC filtering/modeling**: If the given MHC allotype exists in the crystal structure database, we simply filter out from consideration all the MHC templates that host a different MHC. Otherwise, similarly to the previous version of the tool,[28] the $\alpha$-chain amino-acid sequence of the MHC allotype is retrieved and matched with all of the $\alpha$-chain sequences of the MHCs in the crystal structure database. The crystal structure that exhibits the greatest similarity to the allotype in terms of sequence is used as a template for MODELLER[53] to model the structure of the MHC allotype in question. As previously mentioned,[28] many rounds of MODELLER are being run, and the conformation with the best DOPE score is retrieved.

    2. **Peptide similarity**: In the scenario where, multiple crystal structures in the database exist with the same MHC, priority is given to the ones that have bound peptides that are closer in sequence to the peptide to be modelled. Priority is given by scoring, which, in turn is done by aligning, in a pairwise manner, the sequences of the peptide in question with the peptides bound to the MHC. The BLOSUM62 matrix[79] is employed in order to score the sequence alignment.

### Peptide alignment

Given a peptide template that contains the peptide to be used as a guide, and an MHC template that contains the MHC of interest, they are subsequently superimposed and aligned using PyMOL (`http://www.pymol.org/`). After the alignment, the MHC from the peptide template and the peptide from the MHC template are removed.

### Peptide backbone threading

Given the pMHC pair resulting from the peptide alignment phase, as well as the result of the pairwise sequence alignment with anchor constraints during the template similarity computation step, we alter the amino acids of the template peptide with the amino acids of the peptide to be modelled, by:

1. *Deleting* residues from the peptide template that are not to be used for the new peptide to be modelled. This can happen only in the N-termini and C-termini ends of the peptide, for instance, in scenarios where position 1 is predicted to be used as the N-terminal anchor,[37] while the peptide template exhibits a canonical anchor placement, and position 1 in the peptide template needs to be deleted as a result.

2. *Mutating* residues from the peptide template to their new amino acid identities taken from the peptide to be modelled. When, for a given position, there exists the same amino acid in both the peptide template and the peptide sequence to be modelled, the mutation process is skipped for this particular position. The mutation of the residues is performed by PDBFixer.[80]

3. *Inserting* residues that are not in the peptide template but exist in the sequence of the peptide to be modelled. This again can happen only in the N-termini and C-termini ends of the peptide, for instance, in cases where there is an extended configuration either in the N-terminus,[38] or in the C-terminus.[40] The insertion of the residues is also performed by PDBFixer.[80]

We emphasize that throughout the backbone threading process, both the peptide and the receptor are considered, resulting in avoidance of any steric clashes that could arise in the absence of the receptor after amino acid mutation/insertion.

It is important to also note here that the coined term *peptide backbone threading* is a portmanteau of protein threading. The scoring method that matches a peptide template to the peptide to-be-modelled is not just using sequence information, but also structural information in terms of anchor constraints. In practice, in the majority of the cases, peptide threading ends up in similar results to homology modeling, in that the resulting conformation will be very close to the peptide template. However, the mutation approach proposed here is much faster computationally than a homology modeling software like MODELLER,[53] as there is no need for a sequence alignment or a refinement step, especially since all refinement steps are applied later in the APE-Gen2.0 workflow (Figure 1A).

### Backbone Sampling and Scoring

To avoid ending in potentially high modeling error in case the chosen peptide template is not appropriate for peptide threading, we keep a modified version of the loop sampling approach developed in the previous version of the tool.[28] Namely, a big number of peptide backbone conformations are generated using the Random Coordinate Descent algorithm (RCD).[29] Contrary to the previous APE-Gen version,[28] we are generating a much larger set of backbone conformations (the defaults value is 5000 in comparison to the previous value of 100). This is done with very little computational cost, as RCD generates potential conformations really fast, outperforming Cyclic Coordinate Descent, and other loop methods.[81] However, not all peptide conformations are used downstream, as docking/optimizing/scoring all the generated loops would be costly in terms of computation time. Instead, the backbone conformations produced by RCD are ranked by score that reflects the goodness of the loop. Only the top conformations are used downstream. Different scoring functions are being employed for this step and are part of APE-Gen2.0, namely, statistical potentials such as ICOSA[82] and KORP[83] that operate only on backbone atoms. Moreover, RMSD to the template structure is also used. While this option falls under the paradigm of the resulting conformations being closer to the template, still, enough backbone diversity is generated. Finally, as previously proposed,[28] the resulting top backbone conformations obtained go through a final sidechain addition step using PDBFixer.[80]

### Post-Translational Modifications (PTMs)

After pMHC complexes are obtained from the peptide threading and RCD backbone sampling steps, PTMs are also added to the peptide when applicable. PTMs are being added through the PyMOL plugin PyTMS.[49] The PTMs that are currently supported in APE-Gen2.0 are acetylation, carbamylation, citrullination, cystein oxidation/di- oxidation/hydroxidation, di/tri-methylation, methionine oxidation, nitration, nitrosylation, phosphorylation and proline hydroxylation.

### Energy minimization and scoring

The final step to the APE-Gen workflow, as the previous version,[28] involves the optimization of the peptide conformation in the MHC cleft using one of the scoring functions provided by SMINA.[52] Vinardo[50] is being used by default, but Vina[51] is also available in APE-Gen2.0. As before, SMINA is kept intact in the new workflow, as it exhibits a very fast local search protocol, and because of its ability to consider the flexibility of the MHC residues during docking. It is important to note that, as previously reported, some of the favorable, low energy output conformation produced by SMINA might deviate a lot from the proper, anchor restrained pMHC conformation, for example, peptides floating away from the MHC binding cleft. Therefore, when, for a particular peptide conformation, a RMSD difference bigger than 2 Å is detected in the anchor amino acids (N, C) when compared to the chosen peptide template, this conformation is filtered out.

In addition to the local search protocol by SMINA, we also employ an optional energy minimization protocol using OpenMM.[80] This is done to further optimize the conformation of the peptide side chains. During the energy minimization, we apply an external force to the backbone atoms of the peptide in order to keep the backbone intact. The employed force field for the energy minimization is the Amber ff14SB forcefield, with the addition of phosaa14SB parameters in case of presence of phosphorylated residues in the peptide[84] (other PTMs are not yet supported in the optional OpenMM step). The energy tolerance to which the system should be minimized is set to 10× kilojoules/mole.

## Comparisons with other pMHC modeling tools

Throughout the literature, each pMHC modeling tool performs different evaluation experiments. Some tools perform cross-docking leave-one-PDB-out experiments,[26,30] while some tools perform re-docking experiments.[21,28] Here, we chose to evaluate performance based on two different experiments: (A) A leave-one-PDB-out experiment, where APE-Gen2.0 is directly compared to L-RMSD results reported by other tools in the literature and (B) a left-out test set evaluation, where a separate left-out test set is created, and we run and evaluate all the tools on this test dataset. In both evaluation schemes, re-docking is not considered, and we only test the methods on cross-docking. In the following subsections, we will describe in more details the evaluation protocol that we have developed for each method.

### Comparison with PANDORA

Leave-one-PDB-out experiment: Similar to APE-Gen2.0, PANDORA, a homology modeling approach, uses a curated database of pMHC crystal structures to be used as homologs during pMHC modeling.[30] However, the crystal structure database of PANDORA contains duplicate pMHC structures in terms of peptide-MHC pairs (although the PDB codes are different). Additionally, it does not contain crystal structures that include PTMs, or structures that contain additional molecules inside the pMHC binding cleft. As such, any performance gains of APE-Gen2.0 could just be attributed to the different crystal structure database content. To ensure a proper comparison between APE-Gen2.0 and PANDORA, we used the crystal structure database from PANDORA as our crystal structure database of reference instead. This certifies that performance differences between the two methods on this experiment will stem purely from the algorithm and the methodology used in each method. We subsequently used this database to compare APE-Gen2.0 and PANDORA in a leave-one-PDB-out cross-docking scenario as previously proposed.[30] Moreover, in aiming for a proper comparison, similar to PANDORA, we set the maximum number of conformations generated by APE-Gen2.0 to 20 (the default is 100). Finally, as the crystal structure database of PANDORA contains duplicate pMHC structures, we opted in removing those from the database, as this would introduce data leakage. Finally, the evaluation is being done using 427 different structures in total. As PANDORA followed the same leave-one-PDB-out evaluation protocol, the L-RMSD results from PANDORA were taken from the original publication.[30]

Left-out test set experiment: We identified, from our template database, all pMHC pairs that do not appear in PANDORA's template database. We subsequently removed those crystal structures from our database. This acts as the left-out test dataset, which neither APE-Gen2.0 nor PANDORA have access to during the template selection step. From this test set, during evaluation, we filtered out structures that PANDORA could not model (mostly due to MHC allele name support). This resulted in 58 different crystal structures. As before, during modeling those structures with both PANDORA and APE-Gen, the maximum number of conformations generated was set to 20 (PANDORA's default) for a more fair comparison. The list of the PDB codes, along with L-RMSD and Molprobity score results for each one, can be found **Data S2** in Supporting Information.

### Comparison with APE-Gen

Leave-one-PDB-out experiment: To compare APE-Gen2.0 to its predecessor, we used our template database of 699 structures (see the **Template collection and curation** Section for more details). Structures containing additional chains, foreign molecules, or any modifications that might alter the structural pose of the peptide were removed, leaving 569 structures for evaluation in total. For this set of structures, we tested both APE-Gen and APE-Gen2.0 on a cross-docking leave-one-PDB-out experiment as previously proposed.[30] We did not consider the cases where APE-Gen failed to produce conformations during evaluation. This resulted in 229 different structures that we evaluated the performance of APE-Gen and APE-Gen2.0 on.

Left-out test set experiment: We used the same pMHC pairs that we identified when comparing with PANDORA's template database. Similarly, as before, we removed any crystal structure in our template that corresponds to these pMHC pairs. We only consider the cases where APE-Gen successfully produced conformations. PDB codes, L-RMSDs and Molprobity scores for this comparison can be found **Data S3** in Supporting Information.

### Comparison with Docktope

Leave-one-PDB-out experiment: Per the original publication, Docktope was tested on 135 non-redundant pMHC structures.[26] For each one of these structures, by using a molecular docking/energy optimization approach, 1000 conformations were generated. We used these 135 structures to also test APE-Gen2.0. For each of these structures, we applied a cross-docking leave-one-PDB-out protocol, by removing the crystal structure from the APE-Gen2.0 database if it exists. Additionally, we generated 1000 conformation instead of 100 (the default value of APE-Gen2.0) for a fairer comparison with Docktope. Docktope L-RMSD results were taken from the original publication.[26] It is worth underlining that this increased the APE-Gen2.0 execution time from under a minute to 7-8 minutes per complex on average, but it is still well below Docktope's reported execution time of 6 hours maximum.

Left-out test set experiment: As Docktope's web server interface was not functional at the time of assessment (assessed January 30th 2024), we could not model pMHC complexes using Docktope. However, as Docktope is restricted to very few alleles, the test set that could be used for comparison purposes would have been too small to confidently extrapolate. Therefore, all things considered, we opted on not using Docktope for the left-out test set experiment.

## Details on experiments involving post-translationally modified peptides

### Crystal structures involving PTMs

Crystal structures that exhibit PTMs were downloaded from PDB.[12] To enforce non-redundancy and mitigate bias, duplicate structures were removed (example: *3BGM*, *4NNX*). To further mitigate redundancy, non-phosphorylated peptide counterparts that exist in the APE-Gen2.0 crystal structure database were removed during modeling. In total, 13 structures with phosphorylated peptides, 4 structures with citrullinated peptides and 1 structure with a nitrated peptide were used for accessing the accuracy of APE-Gen2.0 in modeling post-translationally modified peptides (see **Table S4** in Supporting Information).

### Post-translationally modified peptides from IEDB

We searched IEDB[8] for peptide entries exhibiting one or more PTMs with a corresponding IC50 value. Specifically, for each PTM that can be modeled by APE-Gen2.0, we search for peptide entries that contain the IC50 value of the peptide, as well as entries that contain the IC50 of the non-PTM variant. In regards to phosphorylation, we found 20 datapoints deposited in the IEDB that contain IC50 values of both the phosphorylated and the non-phosphorylated version of the peptide. 14 of those peptides bind to HLA-A*02:01, and 6 of the peptides bind to HLA-B*40:02. Both the phosphorylated and non-phosphorylated peptides are characterized by a binding affinity value (measured in nM). For the majority of peptides binding to HLA-A*02:01, phosphorylation is seen at position 4, creating a negative charge which improves binding, as previously discussed.[58] A notable exception is the $\beta$-Catenin peptide (*YLDSGIHSGA*, PDB codes: *3FQN*, *3FQR*),[42] where the phosphorylation does not contribute to better binding as expected. The majority of phosphorylated peptides bound to HLA-B*40:02 exhibit the opposite effect, that is, phosphorylation in position 4 mainly decreases binding affinity, as it has been previously observed.[59] In regards to citrullination, 14 datapoints deposited in the IEDB were found, all binding to HLA-A*02:01. As with the phosphorylated peptides, all 14 datapoints contain IC50 values of both the citrullinated and the non-citrullinated version of the peptide. For both the phosphorylated and citrullinated peptides, if IC50 binding affinity values are better than their non-phosphorylated/non-citrullinated counterparts, then we consider the PTM to have positive effects on the binding (labeled as *Better Binding*), else, we consider the PTM to have negative/neutral effects on the binding (labeled as *Worse Binding*). The list of all the IEDB curated peptides can be found in **Data S4** in Supporting Information.

### In-house dataset

A total of 19 selected peptides (**Data S5**) across 5 alleles (HLA-A*01:01, HLA-A*02:01, HLA-B*07:02, HLA-B*40:01, HLA-C*07:02) were obtained from Immunotrack company at purity of ¿80% with quality control by reverse-phase HPLC and mass spectrometry (SC1208). All HLA molecules were made and re-folded as described elsewhere.[85] For affinity measurements, peptides were titrated (8 concentrations: 10000 to 0,01 nM) and incubated in the presence of each HLA followed by analysis with conformation-dependent W6/32 antibodies to determine the affinity of the peptides. The affinities were determined by using sigmoidal curve fitting. For the stability assays, peptides were incubated with each HLA to fold complexes. After overnight incubation, the folded complexes were transferred to 384 plates and subjected to stress at increasing urea concentrations at 0, 1, 2, 3, 4, 5, 6, and 7M, followed by analysis with W6/32. Measurements were carried out as duplicates, and reference peptides were included to ensure the performance of the affinity and stability assay.

Similar to the post-translationally modified peptides from IEDB, we want to differentiate between positive and negative/neutral effects of the phosphorylation on the binding affinity. As before, we assign a positive effect if the phosphorylation results in a better binding affinity (labeled as *Better Binding*), and a negative effect if the phosphorylation results in a worse/similar binding affinity (labeled as *Worse Binding*). Neutrality is assigned when a negligible change in binding affinity and a less than 20% change in stability is observed.

### 5-Experiment protocol specifications

The experiment protocol is as follows: post-translationally modified peptides and their non-modified counterparts from both IEDB and Immunotrack were modeled using APE-Gen2.0. Later, the Vinardo[50] scoring function was used to score both the post-translationally modified peptides and their non-modified counterparts. The difference in Vinardo scores was used as a determinant of positive/negative effects that the PTM can have on peptide binding affinity. As the output from APE-Gen2.0 is an ensemble of conformations, to assess the contribution of the ensemble, we used both the conformation that gives the best Vinardo score, as well as the average Vinardo score from the whole ensemble. However, due to the small number of peptides collected from either IEDB or Immunotrack, and the non-deterministic nature of APE-Gen2.0, scoring function results vary between different APE-Gen2.0 runs. As such, we devised a 5-experiment protocol, where the above process is repeated 5 times, in order to avoid large variations in the results. For each experiment, we get a Vinardo score for each peptide pMHC complex. Therefore, after 5 experiments, the 5 Vinardo scores were averaged in one final score for each pMHC structure. For each post-translationally modified peptide and its non-modified counterpart, we compare the two scores resulting from the above 5-experiment protocol. If the Vinardo score is better for the post-translationally modified peptide in comparison to the vanilla peptide, its binding affinity is then predicted to be better (effectively a labeling threshold of 0). Finally, for all pMHC structures modeled, we opted in not applying any constraints on the peptide backbone during the openMM energy minimization steps, as it has been shown that PTMs can lead to severe structural alterations on the peptide backbone.[18]

### Comparison with Rosetta FlexpepDock and Refinement protocols

To our knowledge, the only other effort in modeling pMHC complexes that include PTMs is the work by.[25] Specifically, the authors modified the Rosetta FlexpepDock[24] and Refinement[22] protocols in order to be able to model peptides bound to MHCs that exhibit PTMs. The authors were able to expand the Rosetta protocols to three different PTMs. We wanted to compare APE-Gen2.0 to the modified Rosetta protocols. We collected the 4 phosphorylated peptide-MHC structures in the PDB[12] that are also used in comparisons in.[25] We used APE-Gen2.0 for modeling, setting the number of generated conformations to 1000 instead of 100 (the default value of APE-Gen2.0), as the Rosetta Refinement protocol also generates 1000 conformations by default. The L-RMSD values reported by[25] for Rosetta FlexpepDock protocol however assume 50000 conformations. This will practically cause the L-RMSD values from Rosetta FlexpepDock to be better than if 1000 generated conformations were used instead.

## Evaluation metrics

### Ligand Root Mean Square Deviation (L-RMSD)

To evaluate the quality of a conformation produced by a pMHC modeling tool in comparison to a ground truth crystal structure, we used the Ligand Root Mean Square Deviation (L-RMSD), a standard metric used extensively in the literature:[26,28,30]

$$\text{L-RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i}$$

where $N$ is the total number of atoms found in the peptide, while $d_i$ is the Euclidean distance between a pair of two corresponding atoms $i$ from the two different structures (model and ground truth). To calculate the L-RMSD, we used ProFit,[86] as previously used by.[30] Three different types of L-RMSD were considered: A) C$\alpha$ L-RMSD, calculated by considering only the C$\alpha$ atoms of the peptide, one per position, B) Backbone L-RMSD, considering only the [C$\alpha$, N, O, C] atoms, and C) Full-Atom L-RMSD, taking all the atoms of the peptide into account.

We also define a variant of the CAPRI criteria[87] to categorize L-RMSD values to different categories: A) High-quality conformations (L-RMSD $\leq$ 1Å), B) Medium, (L-RMSD $\leq$ 1.5Å), C) Acceptable (L-RMSD $\leq$ 2Å) and D) Incorrect (L-RMSD $>$ 2Å). The reason for not following the already established CAPRI criteria here is because pMHC modeling tools have long succeeded in producing near-native ($\leq$ 2Å) conformations of most pMHC complexes. As such, we wanted to have a more fine-grained categorization in the 1-2 Å frame.

### $F_1$ Score

To assess the quality of the anchor identification module, we used the $F_1$ score, defined as the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

where TP is the number of True Positives, FP the number of False Positives and FN the number of False Negatives. For the purpose of our anchor identification task, a positive label represents a non-canonical anchor case, while a negative label represents a canonical anchor. the $F_1$ score receives values scaling from 0-1 (closer to 1 indicates better classification). We make usage of the $F_1$ score in this task because of the large class imbalance between canonical and non-canonical anchor cases. Specifically, the number of non-canonical anchor cases is much lower than the canonical case. As such, we do not wish to focus on the number of True Negatives (not present in the $F_1$ score), as identifying a canonical anchor case is an easy task.

### MolProbity Score

We used MolProbity,[55] more specifically, the MolProbity score,[56] in order to assess the validity of our pMHC modeled structures, as well as to compare APE-Gen2.0 MolProbity scores to MolProbity scores taken from other pMHC structural modeling tools in the literature.[56] The MolProbity score is a single log-weighted value, that combines the calculated clashscore (number of serious clashes per 1000 atoms), the percentage of Ramachandran outliers and the percentage of bad side-chain rotamers. A lower MolProbity score value corresponds to a more protein-like model.

## Acknowledgement

The authors thank Anja Conev for recommendations concerning the structure and content of the manuscript, Alexandra Shewchuk for invaluable suggestions and fixes in certain parts of the APE-Gen2.0 code, as well as other colleagues from Kavraki Lab for the insightful propositions in the APE-Gen2.0 workflow.

## Supporting Information Available

Supporting Information contains:

1. Anchor identification module algorithm discussed in the modeling of non-canonical peptide geometries section and the visual representation of the anchor identification module algorithms' threshold boundaries (**Figure S3**), benchmarks of APE-Gen2.0 for different sampling ratios (**Figure S1**) or template selections (**Figure S6**), benchmarks of different APE-Gen2.0 modeling configurations on sets of post-translationally modified peptides, (**Figure S5**), MolProbity scores of APE-Gen2.0 models of peptides exhibiting post-translational modifications (**Figure S4**), L-RMSD values of comparisons of APE-Gen2.0 to other pMHC structural modeling tools in the literature (**Tables**

**S1, S2, S3, S5** and **Figure S2**), and L-RMSD values plus MolProbity scores on a small set of phosphorylated peptides (**Table S4**). (**PDF** file).

2. Full L-RMSD results (reported in Å) on the leave-one-pdb-out cross-validation experiment. (**Data S1**)

3. Comparison of APE-Gen2.0 to PANDORA (full L-RMSD results (reported in Å) plus MolProbity scores) on a left-out test set. (**Data S2**)

4. Comparison of APE-Gen2.0 to its predecessor (full L-RMSD results (reported in Å) plus MolProbity scores) on a left-out test set. (**Data S3**)

5. The collected list of post-translationally modified peptides collected from the IEDB.[8] (**Data S4**)

6. The list of the in-house dataset of phosphorylated/non-phosphorylated peptides pairs, along with a Kd (nM) value (peptides with Kd value equal to 5000 are designated as non-binder peptides), as well as a Stability percentage value (calculated using a control peptide for each allele) for both phosphorylated/non-phosphorylated peptides. (**Data S5**)

# Data and Software Availability

APE-Gen2.0 is freely available online at `https://apegen.kavrakilab.org`. All data that comprise this study are available in the main text, Supporting Information, or in the github repository provided below. Modeled structures that were used in benchmarks can be provided upon request to the authors. APE-Gen2.0 code, building instructions, as well as additional material, can be found in `https://github.com/anon528/cautious-funicular`.

# Funding

R.F. and L.E.K are supported in part by NIH U01CA258512 and by Rice University funds. M.M.R is supported by a Computational Cancer Biology Training Program fellowship (CPRIT Grant No. RP170593). D.A.A is supported in part by University of Houston funds.

# References

(1) Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J. *Molecular Biology of the Cell*, 4th ed.; Garland Pub.: New York, 2002.

(2) Wieczorek, M.; Abualrous, E. T.; Sticht, J.; Álvaro-Benito, M.; Stolzenberg, S.; Noé, F.; Freund, C. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Front. Immunol.* **2017**, *8*.

(3) Peters, B.; Nielsen, M.; Sette, A. T Cell Epitope Predictions. *Annu. Rev. Immunol.* **2020**, *38*, 123–145.

(4) Hajissa, K.; Zakaria, R.; Suppian, R.; Mohamed, Z. Epitope-based vaccine as a universal vaccination strategy against Toxoplasma gondii infection: A mini-review. *Journal of Advanced Veterinary and Animal Research* **2019**, *6*, 174.

(5) Wang, M.; Yin, B.; Wang, H. Y.; Wang, R.-F. Current advances in T-cell-based cancer immunotherapy. *Immunotherapy* **2014**, *6*, 1265–1278.

(6) O'Donnell, T. J.; Rubinsteyn, A.; Laserson, U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst.* **2020**, *11*, 42–48.e7.

(7) Reynisson, B.; Alvarez, B.; Paul, S.; Peters, B.; Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent

motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **2020**, *48*, W449–W454.

(8) Vita, R.; Overton, J. A.; Greenbaum, J. A.; Ponomarenko, J.; Clark, J. D.; Cantrell, J. R.; Wheeler, D. K.; Gabbard, J. L.; Hix, D.; Sette, A.; Peters, B. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **2014**, *43*, D405–D412.

(9) Gfeller, D.; Schmidt, J.; Croce, G.; Guillaume, P.; Bobisse, S.; Genolet, R.; Queiroz, L.; Cesbron, J.; Racle, J.; Harari, A. Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8+ T-cell epitopes. *Cell Syst.* **2023**, *14*, 72–83.e5.

(10) Li, G.; Iyer, B.; Prasath, V. B. S.; Ni, Y.; Salomonis, N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. *Briefings Bioinf.* **2021**, *22*.

(11) Kaas, Q. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res.* **2004**, *32*, 208D–210.

(12) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(13) Conev, A.; Devaurs, D.; Rigo, M. M.; Antunes, D. A.; Kavraki, L. E. 3pHLA-score improves structure-based peptide-HLA binding affinity prediction. *Sci. Rep.* **2022**, *12*.

(14) Borbulevych, O. Y.; Baxter, T. K.; Yu, Z.; Restifo, N. P.; Baker, B. M. Increased Immunogenicity of an Anchor-Modified Tumor-Associated Antigen Is Due to the Enhanced Stability of the Peptide/MHC Complex: Implications for Vaccine Design. *J. Immunol.* **2005**, *174*, 4812–4820.

(15) Devlin, J. R.; Alonso, J. A.; Ayres, C. M.; Keller, G. L. J.; Bobisse, S.; Kooi, C. W. V.;

Coukos, G.; Gfeller, D.; Harari, A.; Baker, B. M. Structural dissimilarity from self drives neoepitope escape from immune tolerance. *Nat. Chem. Biol.* **2020**, *16*, 1269–1276.

(16) Smith, A. R.; Alonso, J. A.; Ayres, C. M.; Singh, N. K.; Hellman, L. M.; Baker, B. M. Structurally silent peptide anchor modifications allosterically modulate T cell recognition in a receptor-dependent manner. *Proc. Natl. Acad. Sci.* **2021**, *118*.

(17) Cole, D. K.; van den Berg, H. A.; Lloyd, A.; Crowther, M. D.; Beck, K.; Ekeruche-Makinde, J.; Miles, J. J.; Bulek, A. M.; Dolton, G.; Schauenburg, A. J.; Wall, A.; Fuller, A.; Clement, M.; Laugel, B.; Rizkallah, P. J.; Wooldridge, L.; Sewell, A. K. Structural Mechanism Underpinning Cross-reactivity of a CD8+ T-cell Clone That Recognizes a Peptide Derived from Human Telomerase Reverse Transcriptase. *J. Biol. Chem.* **2017**, *292*, 802–813.

(18) Zhao, Y.; Sun, M.; Zhang, N.; Liu, X.; Yue, C.; Feng, L.; Ji, S.; Liu, X.; Qi, J.; Wong, C. C.; Gao, G. F.; Liu, W. J. Phosphosite-dependent presentation of dual phosphorylated peptides by MHC class I molecules. *iScience* **2022**, *25*, 104013.

(19) Antunes, D. A.; Abella, J. R.; Devaurs, D.; Rigo, M. M.; Kavraki, L. E. Structure-based Methods for Binding Mode and Binding Affinity Prediction for Peptide-MHC Complexes. *Curr. Top. Med. Chem.* **2019**, *18*, 2239–2255.

(20) Riley, T. P.; Keller, G. L. J.; Smith, A. R.; Davancaze, L. M.; Arbuiso, A. G.; Devlin, J. R.; Baker, B. M. Structure Based Prediction of Neoantigen Immunogenicity. *Front. Immunol.* **2019**, *10*.

(21) Khan, J.; Ranganathan, S. pDOCK: a new technique for rapid and accurate docking of peptide ligands to Major Histocompatibility Complexes. *Immunome Res.* **2010**, *6*, S2.

(22) Raveh, B.; London, N.; Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2029–2040.

(23) Liu, T.; Pan, X.; Chao, L.; Tan, W.; Qu, S.; Yang, L.; Wang, B.; Mei, H. Subangstrom Accuracy in pHLA-I Modeling by Rosetta FlexPepDock Refinement Protocol. *J. Chem. Inf. Model.* **2014**, *54*, 2233–2242.

(24) Raveh, B.; London, N.; Zimmerman, L.; Schueler-Furman, O. Rosetta FlexPepDock ab-initio: Simultaneous Folding, Docking and Refinement of Peptides onto Their Receptors. *PLoS One* **2011**, *6*, e18934.

(25) Bloodworth, N.; Barbaro, N. R.; Moretti, R.; Harrison, D. G.; Meiler, J. Rosetta FlexPepDock to predict peptide-MHC binding: An approach for non-canonical amino acids. *PLoS One* **2022**, *17*, e0275759.

(26) Rigo, M. M.; Antunes, D. A.; de Freitas, M. V.; de Almeida Mendes, M. F.; Meira, L.; Sinigaglia, M.; Vieira, G. F. DockTope: a Web-based tool for automated pMHC-I modelling. *Sci. Rep.* **2015**, *5*.

(27) Kyeong, H.-H.; Choi, Y.; Kim, H.-S. GradDock: rapid simulation and tailored ranking functions for peptide-MHC Class I docking. *Bioinformatics* **2017**, *34*, 469–476.

(28) Abella, J.; Antunes, D.; Clementi, C.; Kavraki, L. APE-Gen: A Fast Method for Generating Ensembles of Bound Peptide-MHC Conformations. *Molecules* **2019**, *24*, 881.

(29) López-Blanco, J. R.; Canosa-Valls, A. J.; Li, Y.; Chacón, P. RCD+: Fast loop modeling server. *Nucleic Acids Res.* **2016**, *44*, W395–W400.

(30) Marzella, D. F.; Parizi, F. M.; van Tilborg, D.; Renaud, N.; Sybrandi, D.; Buzatu, R.; Rademaker, D. T.; ‘t Hoen, P. A. C.; Xue, L. C. PANDORA: A Fast, Anchor-Restrained Modelling Protocol for Peptide: MHC Complexes. *Front. Immunol.* **2022**, *13*.

(31) Antunes, D. A.; Moll, M.; Devaurs, D.; Jackson, K. R.; Lizée, G.; Kavraki, L. E.

DINC 2.0: A New Protein–Peptide Docking Webserver Using an Incremental Approach. *Cancer Res.* **2017**, *77*, e55–e57.

(32) Antunes, D. A.; Devaurs, D.; Moll, M.; Lizée, G.; Kavraki, L. E. General Prediction of Peptide-MHC Binding Modes Using Incremental Docking: A Proof of Concept. *Sci. Rep.* **2018**, *8*.

(33) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(34) Motmaen, A.; Dauparas, J.; Baek, M.; Abedi, M. H.; Baker, D.; Bradley, P. Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. *Proc. Natl. Acad. Sci.* **2023**, *120*.

(35) Guillaume, P.; Picaud, S.; Baumgaertner, P.; Montandon, N.; Schmidt, J.; Speiser, D. E.; Coukos, G.; Bassani-Sternberg, M.; Filippakopoulos, P.; Gfeller, D. The C-terminal extension landscape of naturally presented HLA-I ligands. *Proc. Natl. Acad. Sci.* **2018**, *115*, 5083–5088.

(36) Liu, W. J.; Tan, S.; Zhao, M.; Quan, C.; Bi, Y.; Wu, Y.; Zhang, S.; Zhang, H.; Xiao, H.; Qi, J.; Yan, J.; Liu, W.; Yu, H.; Shu, Y.; Wu, G.; Gao, G. F. Cross-immunity Against Avian Influenza A(H7N9) Virus in the Healthy Population Is Affected by Antigenicity-Dependent Substitutions. *J. Infect. Dis.* **2016**, *214*, 1937–1946.

(37) Borbulevych, O. Y.; Insaidoo, F. K.; Baxter, T. K.; Powell, D. J.; Johnson, L. A.; Restifo, N. P.; Baker, B. M. Structures of MART-126/27–35 Peptide/HLA-A2 Complexes

Reveal a Remarkable Disconnect between Antigen Structural Homology and T Cell Recognition. *J. Mol. Biol.* **2007**, *372*, 1123–1136.

(38) Pymm, P.; Illing, P. T.; Ramarathinam, S. H.; O'Connor, G. M.; Hughes, V. A.; Hitchen, C.; Price, D. A.; Ho, B. K.; McVicar, D. W.; Brooks, A. G.; Purcell, A. W.; Rossjohn, J.; Vivian, J. P. MHC-I peptides get out of the groove and enable a novel mechanism of HIV-1 escape. *Nat. Struct. Mol. Biol.* **2017**, *24*, 387–394.

(39) Collins, E. J.; Garboczi, D. N.; Wiley, D. C. Three-dimensional structure of a peptide extending from one end of a class I MHC binding site. *Nature* **1994**, *371*, 626–629.

(40) McMurtrey, C.; Trolle, T.; Sansom, T.; Remesh, S. G.; Kaever, T.; Bardet, W.; Jackson, K.; McLeod, R.; Sette, A.; Nielsen, M.; Zajonc, D. M.; Blader, I. J.; Peters, B.; Hildebrand, W. Toxoplasma gondii peptide ligands open the gate of the HLA class I binding groove. *eLife* **2016**, *5*.

(41) Engelhard, V. H.; Altrich-Vanlith, M.; Ostankovitch, M.; Zarling, A. L. Post-translational modifications of naturally processed MHC-binding epitopes. *Curr. Opin. Immunol.* **2006**, *18*, 92–97.

(42) Petersen, J.; Purcell, A. W.; Rossjohn, J. Post-translationally modified T cell epitopes: immune recognition and immunotherapy. *J. Mol. Med.* **2009**, *87*.

(43) Mangalaparthi, K. K.; Madugundu, A. K.; Ryan, Z. C.; Garapati, K.; Peterson, J. A.; Dey, G.; Prakash, A.; Pandey, A. Digging deeper into the immunopeptidome: characterization of post-translationally modified peptides presented by MHC I. *J. Proteins Proteomics* **2021**, *12*, 151–160.

(44) Kacen, A.; Javitt, A.; Kramer, M. P.; Morgenstern, D.; Tsaban, T.; Shmueli, M. D.; Teo, G. C.; da Veiga Leprevost, F.; Barnea, E.; Yu, F.; Admon, A.; Eisenbach, L.; Samuels, Y.; Schueler-Furman, O.; Levin, Y.; Nesvizhskii, A. I.; Merbl, Y. Post-

translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors. *Nat. Biotechnol.* **2022**, *41*, 239–251.

(45) Malaker, S. A.; Penny, S. A.; Steadman, L. G.; Myers, P. T.; Loke, J. C.; Raghavan, M.; Bai, D. L.; Shabanowitz, J.; Hunt, D. F.; Cobbold, M. Identification of Glycopeptides as Posttranslationally Modified Neoantigens in Leukemia. *Cancer Immunol. Res.* **2017**, *5*, 376–384.

(46) Srivastava, A. K.; Guadagnin, G.; Cappello, P.; Novelli, F. Post-Translational Modifications in Tumor-Associated Antigens as a Platform for Novel Immuno-Oncology Therapies. *Cancers* **2022**, *15*, 138.

(47) Refsgaard, C. T.; Barra, C.; Peng, X.; Ternette, N.; Nielsen, M. NetMHCphosPan - Pan-specific prediction of MHC class I antigen presentation of phosphorylated ligands. *ImmunoInformatics* **2021**, *1-2*, 100005.

(48) Solleder, M.; Guillaume, P.; Racle, J.; Michaux, J.; Pak, H.-S.; Müller, M.; Coukos, G.; Bassani-Sternberg, M.; Gfeller, D. Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands. *Mol. Cell. Proteomics* **2020**, *19*, 390–404.

(49) Warnecke, A.; Sandalova, T.; Achour, A.; Harris, R. A. PyTMs: a useful PyMOL plugin for modeling common post-translational modifications. *BMC Bioinf.* **2014**, *15*.

(50) Quiroga, R.; Villarreal, M. A. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PLoS One* **2016**, *11*, e0155183.

(51) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **2021**, *61*, 3891–3898.

(52) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893–1904.

(53) Webb, B.; Sali, A. Protein Structure Modeling with MODELLER. **2014**, 1–15.

(54) Buttenschoen, M.; Morris, G. M.; Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **2024**,

(55) Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B.; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **2007**, *35*, W375–W383.

(56) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *66*, 12–21.

(57) Kim, Y.; Sidney, J.; Pinilla, C.; Sette, A.; Peters, B. Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior. *BMC Bioinf.* **2009**, *10*.

(58) Jackson, K. R.; Antunes, D. A.; Talukder, A. H.; Maleki, A. R.; Amagai, K.; Salmon, A.; Katailiha, A. S.; Chiu, Y.; Fasoulis, R.; Rigo, M. M.; Abella, J. R.; Melendez, B. D.; Li, F.; Sun, Y.; Sonnemann, H. M.; Belousov, V.; Frenkel, F.; Justesen, S.; Makaju, A.; Liu, Y.; Horn, D.; Lopez-Ferrer, D.; Huhmer, A. F.; Hwu, P.; Roszik, J.; Hawke, D.; Kavraki, L. E.; Lizée, G. Charge-based interactions through peptide position 4 drive diversity of antigen presentation by human leukocyte antigen class I molecules. *PNAS Nexus* **2022**, *1*.

(59) Alpízar, A.; Marino, F.; Ramos-Fernández, A.; Lombardía, M.; Jeko, A.; Pazos, F.; Paradela, A.; Santiago, C.; Heck, A. J.; Marcilla, M. A Molecular Basis for the Presentation of Phosphorylated Peptides by HLA-B Antigens. *Mol. Cell. Proteomics* **2017**, *16*, 181–193.

(60) Croitoru, A.; Park, S.-J.; Kumar, A.; Lee, J.; Im, W.; MacKerell, A. D.; Aleksandrov, A. Additive CHARMM36 Force Field for Nonstandard Amino Acids. *J. Chem. Theory Comput.* **2021**, *17*, 3554–3570.

(61) Perez, M. A. S.; Cuendet, M. A.; Röhrig, U. F.; Michielin, O.; Zoete, V. *Methods Mol. Biol.*; Springer US, 2022; pp 245–282.

(62) Keller, G. L. J.; Weiss, L. I.; Baker, B. M. Physicochemical Heuristics for Identifying High Fidelity, Near-Native Structural Models of Peptide/MHC Complexes. *Front. Immunol.* **2022**, *13*.

(63) Karosiene, E.; Lundegaard, C.; Lund, O.; Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **2011**, *64*, 177–186.

(64) Hundal, J.; Kiwala, S.; McMichael, J.; Miller, C. A.; Xia, H.; Wollam, A. T.; Liu, C. J.; Zhao, S.; Feng, Y.-Y.; Graubert, A. P.; Wollam, A. Z.; Neichin, J.; Neveau, M.; Walker, J.; Gillanders, W. E.; Mardis, E. R.; Griffith, O. L.; Griffith, M. pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. *Cancer Immunol. Res.* **2020**, *8*, 409–420.

(65) Wilson, E. A.; Hirneise, G.; Singharoy, A.; Anderson, K. S. Total predicted MHC-I epitope load is inversely associated with population mortality from SARS-CoV-2. *Cell Rep. Med.* **2021**, *2*, 100221.

(66) Sidney, J.; Assarsson, E.; Moore, C.; Ngo, S.; Pinilla, C.; Sette, A.; Peters, B. Quanti-

tative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res.* **2008**, *4*, 2.

(67) Nguyen, A. T.; Szeto, C.; Gras, S. The pockets guide to HLA class I molecules. *Biochem. Soc. Trans.* **2021**, *49*, 2319–2331.

(68) Collado, J. A.; Guitart, C.; Ciudad, M. T.; Alvarez, I.; Jaraquemada, D. The Repertoires of Peptides Presented by MHC-II in the Thymus and in Peripheral Tissue: A Clue for Autoimmunity? *Front. Immunol.* **2013**, *4*.

(69) Sandalova, T.; Sala, B. M.; Achour, A. Structural aspects of chemical modifications in the MHC-restricted immunopeptidome; Implications for immune recognition. *Front. Chem.* **2022**, *10*.

(70) Rodrigues, J.; Teixeira, J.; Trellet, M.; Bonvin, A. pdb-tools: a swiss army knife for molecular structures [version 1; peer review: 2 approved]. *F1000Research* **2018**, *7*.

(71) Ayres, C. M.; Corcelli, S. A.; Baker, B. M. Peptide and Peptide-Dependent Motions in MHC Proteins: Immunological Implications and Biophysical Underpinnings. *Front. Immunol.* **2017**, *8*.

(72) Hubbard, S.; Thornton, J. NACCESS. Computer Program, Department of Biochemistry and Molecular Biology, University College London, 1993.

(73) Antunes, D. A.; Rigo, M. M.; Freitas, M. V.; Mendes, M. F. A.; Sinigaglia, M.; Lizée, G.; Kavraki, L. E.; Selin, L. K.; Cornberg, M.; Vieira, G. F. Interpreting T-Cell Cross-reactivity through Structure: Implications for TCR-Based Cancer Immunotherapy. *Front. Immunol.* **2017**, *8*.

(74) Abella, J. R.; Antunes, D.; Jackson, K.; Lizée, G.; Clementi, C.; Kavraki, L. E. Markov state modeling reveals alternative unbinding pathways for peptide–MHC complexes. *Proc. Natl. Acad. Sci.* **2020**,

(75) Nguyen, A. T.; Szeto, C.; Gras, S. The pockets guide to HLA class I molecules. *Biochem. Soc. Trans.* **2021**, *49*, 2319–2331.

(76) Sarkizova, S.; Klaeger, S.; Le, P. M.; Li, L. W.; Oliveira, G.; Keshishian, H.; Hartigan, C. R.; Zhang, W.; Braun, D. A.; Ligon, K. L.; Bachireddy, P.; Zervantonakis, I. K.; Rosenbluth, J. M.; Ouspenskaia, T.; Law, T.; Justesen, S.; Stevens, J.; Lane, W. J.; Eisenhaure, T.; Zhang, G. L.; Clauser, K. R.; Hacohen, N.; Carr, S. A.; Wu, C. J.; Keskin, D. B. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **2019**, *38*, 199–209.

(77) Nikulin, M. Hellinger distance. Encyclopedia of Mathematics, 1994.

(78) Ding, Y.-H.; Baker, B. M.; Garboczi, D. N.; Biddison, W. E.; Wiley, D. C. Four A6-TCR/Peptide/HLA-A2 Structures that Generate Very Different T Cell Signals Are Nearly Identical. *Immunity* **1999**, *11*, 45–56.

(79) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **1992**, *89*, 10915–10919.

(80) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659.

(81) Chys, P.; Chacón, P. Random Coordinate Descent with Spinor-matrices and Geometric Filters for Efficient Loop Closure. *J. Chem. Theory Comput.* **2013**, *9*, 1821–1829.

(82) Elhefnawy, W.; Chen, L.; Han, Y.; Li, Y. ICOSA: A Distance-Dependent, Orientation-Specific Coarse-Grained Contact Potential for Protein Structure Modeling. *J. Mol. Biol.* **2015**, *427*, 2562–2576.

(83) López-Blanco, J. R.; Chacón, P. KORP: knowledge-based 6D potential for fast protein and loop modeling. *Bioinformatics* **2019**, *35*, 3013–3019.

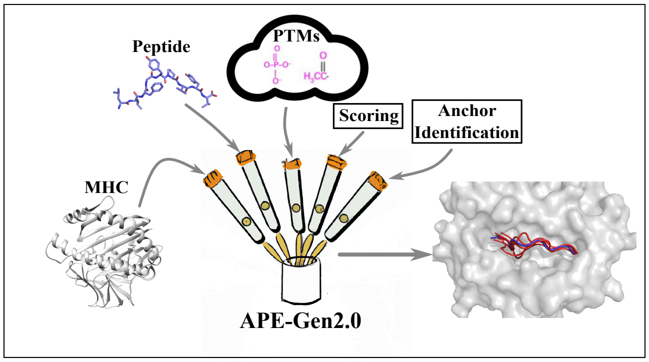
(84) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(85) Harndahl, M.; Justesen, S.; Lamberth, K.; Røder, G.; Nielsen, M.; Buus, S. Peptide Binding to HLA Class I Molecules: Homogenous, High-Throughput Screening, and Affinity Assays. *SLAS Discovery* **2009**, *14*, 173–180.

(86) McLachlan, A. D. Rapid comparison of protein structures. *Acta Crystallogr., Sect. A: Found. Adv.* **1982**, *38*, 871–873.

(87) Lensink, M. F.; Nadzirin, N.; Velankar, S.; Wodak, S. J. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins: Struct., Funct., Bioinf.* **2020**, *88*, 916–938.

# TOC Graphic



**APE-Gen2.0: Expanding rapid class I peptide-MHC modeling to post-translational modifications and noncanonical peptide geometries**

Romanos Fasoulis, Mauricio M. Rigo, Gregory Lizée, Dinler A. Antunes and Lydia E. Kavraki*