TYPE Original Research
PUBLISHED 12 June 2023
DOI 10.3389/fimmu.2023.1142573

CrossDome: an interactive R package to predict cross-reactivity risk using immunopeptidomics databases

Andre F. Fonseca and Dinler A. Antunes*

Antunes Lab, Center for Nuclear Receptors and Cell Signaling (CNRCS), Department of Biology and Biochemistry, University of Houston, Houston, TX, United States

Check for updates

OPEN ACCESS

EDITED BY
Rene De Waal Malefyt,
Synthekine Inc., United States

REVIEWED BY
Andrew Morin,
Synthekine Inc., United States
Fatemeh Rahbarizadeh,
Tarbiat Modares University, Iran

*CORRESPONDENCE
Dinler A. Antunes
dinler@uh.edu

RECEIVED 11 January 2023
ACCEPTED 23 May 2023
PUBLISHED 12 June 2023

CITATION
Fonseca AF and Antunes DA (2023) CrossDome: an interactive R package to predict cross-reactivity risk using immunopeptidomics databases. *Front. Immunol.* 14:1142573. doi: 10.3389/fimmu.2023.1142573

T-cell-based immunotherapies hold tremendous potential in the fight against cancer, thanks to their capacity to specifically targeting diseased cells. Nevertheless, this potential has been tempered with safety concerns regarding the possible recognition of unknown off-targets displayed by healthy cells. In a notorious example, engineered T-cells specific to MAGEA3 (EVDPIGHLY) also recognized a TITIN-derived peptide (ESDPIVAQY) expressed by cardiac cells, inducing lethal damage in melanoma patients. Such off-target toxicity has been related to T-cell cross-reactivity induced by molecular mimicry. In this context, there is growing interest in developing the means to avoid off-target toxicity, and to provide safer immunotherapy products. To this end, we present CrossDome, a multi-omics suite to predict the off-target toxicity risk of T-cell-based immunotherapies. Our suite provides two alternative protocols, i) a peptide-centered prediction, or ii) a TCR-centered prediction. As proof-of-principle, we evaluate our approach using 16 well-known cross-reactivity cases involving cancer-associated antigens. With CrossDome, the TITIN-derived peptide was predicted at the 99+ percentile rank among 36,000 scored candidates (p-value < 0.001). In addition, off-targets for all the 16 known cases were predicted within the top ranges of relatedness score on a Monte Carlo simulation with over 5 million putative peptide pairs, allowing us to determine a cut-off p-value for off-target toxicity risk. We also implemented a penalty system based on TCR hotspots, named contact map (CM). This TCR-centered approach improved upon the peptide-centered prediction on the MAGEA3-TITIN screening (e.g., from 27th to 6th, out of 36,000 ranked peptides). Next, we used an extended dataset of experimentally-determined cross-reactive peptides to evaluate alternative CrossDome protocols. The level of enrichment of validated cases among top 50 best-scored peptides was 63% for the peptide-centered protocol, and up to 82% for the TCR-centered protocol. Finally, we performed functional characterization of top ranking candidates, by integrating expression data, HLA binding, and immunogenicity predictions. CrossDome was designed as an R package for easy integration with antigen discovery pipelines, and an interactive web interface for users without coding experience. CrossDome is under active development, and it is available at https://github.com/AntunesLab/crossdome.

KEYWORDS

T-cell cross-reactivity prediction, off-target toxicity, antigen prioritization, T-cell therapy, MAGEA3, cancer immunotherapy, computational oncology

1 Introduction

T-cell-based therapies are providing promising approaches for treating several types of cancer. These therapies leverage the cellular immunity mechanisms allowing T-cell lymphocytes to surveil, recognize, and eliminate cells displaying at their surfaces "foreign" peptides bound to Human Leukocyte Antigen (HLA) receptors (1, 2). This class of cancer immunotherapy treatments include i) the use of peptide-based vaccines to trigger the expansion of tumor-specific T-cells (3–6), and ii) the use of adoptive T-cell therapy, which involves collecting, expanding, and transferring tumor-specific T-cells to treat cancer patients (7, 8). In this context, the therapeutic T-cells can be unaltered tumor infiltrating lymphocytes, or genetically modified T-cells engineered to have higher affinity against specific tumor-associated antigens (4, 9, 10). Recently, more effective control of cancer progression was achieved with combined use of adoptive T-cell therapy and immune checkpoint inhibitors (11, 12). There is also growing interest in the development of chimeric antigen receptor T-cells (CAR-T), and tumor-specific antibodies that mimic T-cell receptor (TCR) recognition (13, 14).

However, several limitations are still hindering the broader use of T-cell therapies for cancer treatment (15). For instance, engineering a T-cell receptor is a challenging task that involves potentially conflicting goals, such as enhancing the T-cell response to the tumor-derived peptide, while avoiding side effects caused by T-cell cross-reactivity (16–18). T-cell cross-reactivity is the ability of a single T-cell clonotype to recognize and respond to multiple heterologous peptide-HLA (pHLA) complexes (19–21). From an evolutionary perspective, T-cell cross-reactivity is necessary to maximize the range of unrelated antigens/pathogens that can be recognized by a limited pool of T-cells (i.e., to mediate heterologous immunity between pathogens) (22–24). On the other hand, in T-cell-based therapy, cross-reactivity events have been linked to off-target toxicity risk, i.e., recognition of self-derived peptide-targets leading to autoimmune reactions against healthy tissues (21, 25).

To date, multiple clinical trials have been withdrawn due to T-cell cross-reactivity issues (26, 27). In the most notable example, MAGEA3-specific engineered T-cells were associated with severe off-target toxicity in melanoma patients. It was observed that these therapeutic T-cells were cross-reactive with a TITIN-derived peptide, causing lethal cardiotoxicity in at least five patients (26). Other cross-reactivity events have been reported in studies involving different tumor-associated antigens, such as MART-1, NY-ESO-1, and AFP (28–30). Additionally, off-target toxicity has also been reported with the use of CAR-T therapy (31–33). It is also important to note that while TCR engineering can increase the risk of dangerous T-cell cross-reactivities (34), this risk exists with any T-cell-based therapy, including the use of unaltered TILs, the stimulation of the patient's own T-cells through peptide-based vaccines (35–37), and the use of TCR-mimic antibodies (13, 14). Therefore, the capacity to determine or predict the potential risk for off-target toxicities during the design and development of T-cell-based therapies is a major bottleneck for the broader use of these powerful immunotherapy approaches.

Unfortunately, there are no standard experimental methods that can be routinely applied to determine the risk of T-cell cross-reactivity in immunotherapies. Alanine scans, or X-scans, of the cognate peptide-target are usually performed to provide an initial assessment of T-cell cross-reactivity potential (38, 39). Such experiments do not directly provide information on potential off-targets, but can be used to guide proteomic searches for similar peptide motifs (40). The more recent development of yeast/phage-display and other high-throughput methods is starting to enable the screening of larger libraries of putative peptide-targets, but broader use of these methods is still limited by the cost and time required for library construction (41). In addition, these T-cell-based screenings are not as useful in the case of peptide-based vaccines, since the T-cell clones responding to the immunization will be different for each immunized individual (e.g., private specificity) (21, 42).

On the computational side, early work has been done in the context of heterologous immunity between viruses, mostly focused on the identification of peptide sequence similarities underlying cross-reactivity events (23, 43–45). This led to the development of a few sequence-based methods for cross-reactivity prediction, which can be further divided into methods based on i) peptide sequence identity (e.g., alignment based methods such as Expitope and iCrossR) and ii) peptide biochemical similarity (e.g., "alignment-free" methods such as dGraph and sCRAP). Expitope and iCrossR rely on the combined use of i) a Levenshtein distance to recover proteome-derived cross-reactive candidates with up to 4 amino acid mismatches to the query (i.e., high sequence identity), and ii) the subsequent ranking/filtering of these candidates based on a "epitope presentation score" (e.g., combined score from prediction algorithms for multiple steps of the class I HLA pathway, including proteasomal cleavage, TAP transport, and HLA binding). The output is also annotated with mRNA expression data to indicate tissue distribution and abundance (46).

An alternative approach for ranking peptides based on biochemical similarity was implemented in dGraph (47), which uses physicochemical properties to connect similar peptides into a network graph. More recently, a hybrid approach named sCRAP (48) was proposed. It uses a biochemical similarity matrix for computing a similarity score against the entire human proteome, and filters the output based on maximum tissue expression and HLA-binding affinity (e.g., predictions from NetMHC4 (49), NetMHCpan4.1 (50), and HLAthena (51)). sCRAP also enables customizing the score to increase the contribution of specific peptide positions, which could be used to bias the search based on potential hotspots for TCR recognition. In fact, attention to TCR hotspots on the pHLA surface is supported by both computational and experimental research showing that T-cell cross-reactivity can involve peptides with very diverse sequences, as long as they share a few key structural/biochemical similarities that are driving T-cell recognition (21, 52–55). These observations have supported the development of structure-based methods for T-cell cross-reactivity prediction (56–60), which have been discussed elsewhere (21, 25).

Although each approach provides interesting contributions to T-cell cross-reactivity analysis, all these methods have notable limitations. Expitope, sCRAP and iCrossR rely on sequence-based HLA-binding prediction algorithms to define the "universe" of possible off-targets out of the entire proteome. In doing so, they add a layer of prediction and introduce potential biases to the cross-reactivity screening. For instance, it is well-known that the accuracy of HLA-binding algorithms varies widely across HLA alleles, being less reliable for less prevalent HLAs (61, 62). Furthermore, cases of T-cell cross-reactivity involving peptides with low sequence identity, which have also been reported (21, 52, 63), cannot be predicted by Expitope and iCrossR due to the limited number of allowed mismatches.

Tools such as dGraph do not limit the number of mismatches and produce a relationship network for a set of peptides based on overall biochemical similarity. However, dGraph was developed for antibody cross-reactivity prediction (e.g., with linear epitopes involved in allergies), and was not directly evaluated for T-cell cross-reactivity (47). In addition, it does not include a clear statistical threshold to determine the level of significant similarity, hence hampering the interpretability of predictions. The more recent sCRAP is also not limited to a given number of mismatches, and introduces the use of position-specific weights to try and account for the TCR-specific hotspots. However, these weights were pre-defined for the MAGEA3-specific T-cell, and there was no rationale for customizing those weights for other targets of interest (48). Finally, structure-based methods are still limited by the number of available structures and the computational cost of large-scale modeling, being generally unsuitable for proteome-wide screenings (21, 25).

In this context, we developed CrossDome, a tool that performs peptide screening on multi-omics data from healthy tissues and predicts the risk for off-target toxicity with unrelated self-derived peptides. By relying on experimentally determined data from real pHLA complexes (i.e., immunopeptidomics data from mass-spectrometry studies), we overcome the potential biases of sequence-based HLA-binding prediction algorithms. Our approach leverages a large dataset of amino acid's biochemical properties, allowing us to predict peptides that are biochemically similar to desirable targets, without enforcing any sequence identity cut-off. In addition, we demonstrate how structural data on TCRpHLA interactions can be used to tailor CrossDome predictions in order to account for TCR-specific hotspots. Finally, we compute a p-value for each putative cross-reactive target, therefore providing statistical support to estimate the off-target toxicity risk associated with each prediction. We further improve the significance of our results by integrating functional data to evaluate expression level, HLA-binding, and immunogenicity of putative off-targets, which can help prioritize candidates for experimental validation. CrossDome is released as an R package with support for technical and non-technical users, enabling both lower level control and a user-friendly application for users without coding experience.

2 Material and methods

2.1 Collecting reference experimental datasets

Immunopeptidomics data on naturally occurring self-derived peptides was retrieved from several sources, including HLA Ligand Atlas (64), HLAthena (51), SysteMHC (65), IEDB (66), and two other published datasets (67, 68). The data was filtered considering only 9-mer peptides presented by Class I HLA alleles (Supplementary Figure S1A). Next, we combined data sources into a non-redundant local database, organized by HLA allele restriction. In turn, each allele produces a single background database for the CrossDome screening algorithm. In order to evaluate the false-positive ratio, we also collected published data on 16 peptides previously identified as T-cell cross-reactivity off-targets for one of four well-known tumor-associated antigens: MAGEA3 (27, 41, 69, 70), NY-ESO-1 (69), TMEM161A (71) and AFP (72) (Supplementary Table 1). Finally, we also obtained a dataset of 60 synthetic peptides experimentally-determined to be recognized by the A3A TCR, being therefore cross-reactive with the cognate peptide from MAGEA3 (Supplementary Figure S1B). This data was produced by a yeast-display screening experiment previously reported by Gee et al. (2018) (73), and includes peptide sequences with up to 6 mismatches in relation to the cognate peptide (i.e., only 33% sequence identity). We also complemented this dataset with known off-targets of MAGEA3 (e.g., TITIN and MAGEA6). The full set of MAGEA3 off-targets is available in the CrossDome repository on Github.

2.2 Implementing a new similarity model based on biochemical profiles

The biochemical properties of amino acids have been used in previous work to estimate the similarity between peptides or protein binding motifs (44, 45, 74, 75). Here, we used a library with over 500 biochemical properties from AAIndex (76), a gold standard database of amino acids properties. This database has been used as a source of amino acid features for machine learning (77), and was used here to implement a new model to measure peptide similarity. First, the AAIndex data were summarized by using principal component analysis (PCA). The biochemical properties were then summarized into 12 principal components, holding 95% variance in the dataset (Supplementary Figure S2). The resulting eigenvectors were converted into a matrix of biochemical properties, spanning the 20 natural amino acids.

Each peptide was represented using this matrix, hereafter referred to as a biochemical profile (BP). In turn, the biochemical profiles of two different peptides can be used to compute a distance between these peptides, named as relatedness score (RdS). In order to compute this relatedness score, we implemented a weighted Euclidean distance. The weighted vector can be derived from TCR hotspots in the peptide sequence, i.e., position-specific weights related to known bonds/interactions between TCR and peptide molecules. This implementation penalizes biochemical profiles that deviate in hotspot positions. The relatedness score was normalized by peptide length, where low values represent highly similar peptide pairs (i.e., stronger candidates for cross-reactivity). Note that in its current implementation, our algorithm is limited to the analysis of 9-mers, which account for most of the peptides displayed by Class I HLA alleles (51). Finally, we compared the performance of the biochemical-profile-based approach with evolution-based substitution matrices. This revealed the extent to which our BP-based approach can capture peptide similarities beyond what would be found with a standard substitution matrix (e.g., BLOSUM (78)), keeping all other parameters equal (e.g., same query, Page 3positive control, and universe of peptides). The comparison was performed using Biostrings, an R Bioconductor package (79).

2.3 Monte Carlo simulation of peptide pairs and statistical validation

Statistical thresholds are essential to determine confidence levels in computational analysis, but no reference thresholds have been provided in previous methods for T-cell cross-reactivity prediction. To determine confidence levels in CrossDome predictions, we conducted a Monte Carlo simulation using peptide pairs derived from our immunopeptidomics database. The analysis was designed to produce 5 million simulated pairs, covering a wide range of class I HLA alleles. Next, an individual relatedness score was calculated for each peptide pair, using the aforementioned methods. The resulting RdS distribution was tested using the Shapiro-Wilk test (80). Then, we utilized the "*p*-norm" function to derive probability values for each peptide pair. Finally, the statistical threshold was defined based on the highest p-value associated with any of the experimentally-validated peptide pairs (Supplementary Table 1). This procedure allowed us to determine the relatedness score sensitivity to identify "real" cases within the background noise.

2.4 Uncovering TCR-peptide interactions from structural data

X-ray crystallography data for the A3A/MAGEA3/HLA-A*01:01 TCRpHLA complex was retrieved from the Protein Data Bank (81) to determine the molecular interactions between the engineered TCR and the cognate MAGEA3-derived peptide (PDB ID: 5BRZ) (70). The crystal structure was processed and cleaned using the PDBFixer tool from OpenMM suite (82). Hydrogen atoms were included assuming neutral pH (pH = 7.0), using the CHARMM36 force field parameterization protocol (83). The GetContacts package (84) was used to derive TCR-peptide interactions from the 3D structure, which we used to create TCR contact maps (CM). All interactions supported in this package were included. The hydrogen bond threshold was changed to 4.0 A, following the parameters used for curated TCR contacts on IEDB (85).

Two different CMs were produced for the same TCR in this analysis: CM-crystal and CM-custom. On the one hand, CM-crystal was obtained by performing a per-peptide-position cumulative sum of contacts derived from the reference crystal structure with GetContacts. These per-peptide-position contacts were then converted into a frequency vector, and pseudo counts (penalty = 0.5) were included in positions without TCR interactions. The resulting penalty vector for CM-crystal had the following values: w = (3.0, 0.5, 0.5, 4.0, 2.0, 0.5, 1.0, 1.0, 0.5). CM-custom was designed with knowledge-based weights informed by both data derived from the molecular dynamics simulation, and data regarding HLA binding requirements. We performed a 100 ns molecular dynamics simulation using the Gromacs 2021.2 package (86). The resulting data was divided into distinct time points for analysis, starting with the input structure (i.e., static data from crystal) and extending in 10 ns increments from the simulation (i.e., dynamic data). A stride step equal to 50 was adopted to recover frames from the simulation, which were used as input for GetContacts in order to obtain both the type and frequency of TCR-peptide interactions. Since the molecular dynamics simulation also enables energy calculations that are not available in GetContacts, we manually accounted for the occurrence of short-range Coulomb interactions with TCR residues (Supplementary Figure S3). Finally, we also accounted for the importance of peptide positions 3 and 9 for binding to HLA-A*01 alleles, as described in the SYFPHEITHI database (87). The resulting empirical penalty vector for CM-custom had the following values: w = (3.0, 0.5, 2.0, 4.0, 2.0, 0.5, 1.0, 1.0, 2.0).

2.5 Integrating functional data and third-party predictions

The clinical relevance of each candidate predicted by CrossDome might depend on additional functional properties, such as off-target expression and immunogenicity (i.e., capacity to trigger T-cell response) 8 (46, 48, 69). Therefore, to provide additional support for target prioritization, we incorporated into CrossDome results data from i) gene expression, ii) HLA binding affinity, iii) and peptide immunogenicity.

Transcriptomics data was retrieved from Human Protein Atlas (88), and highlights two essential aspects for the source antigen: i) abundance, which indirectly affects the number of HLA-displayed peptides at the cell surface, and ii) localization, which allows characterizing a candidate profile as tissue-specific or ubiquitous. We collected 37 healthy tissues spanning mRNA expression for 25,000 coding genes. To mitigate discrepancies from transcript to protein expression, the expression levels were normalized to transcripts per million protein-coding genes (pTPM) for healthy tissue in the database. Next, we classified the gene expression using the TissueEnrich package (89). The genes were separated into five groups: tissue-enriched, tissue-enhanced, group-enriched, low-tissue-specificity, and not-expressed profiles.

Finally, all eluted peptides included in our reference database were annotated with predicted immunogenicity (e.g., immunogenicity score) and predicted HLA binding (e.g., predicted IC50 binding affinity), for each included HLA Class I allele. HLA binding affinities were predicted with MHFlury 2.0 (90), which also accounts for intracellular peptide processing. Immunogenicity values were predicted with a recently published method called DeepImmuno (77). Taken together, these data can provide additional "clues" regarding T-cell toxicity and tolerability levels (69, 91). It is worth highlighting that our approach to off-target prediction differs from previous methods in that we do not rely on machine learning algorithms to identify "real" HLA-binders. Instead, we use these methods as additional criteria for prioritizing high-risk candidates.

2.6 Software development and web application

An overview of the CrossDome algorithm is provided in Supplementary Figure S4. In summary, for a peptide-centered prediction CrossDome requires as input only a peptide sequence, and its HLA allele restriction. Currently, CrossDome does not search across different HLA alleles, so the universe of self-derived peptides is restricted to the HLA allele of interest. For a TCR-centered prediction, the user would also provide the TCR contact map.

After ranking the peptides based on the relatedness score, CrossDome annotates each predicted off-target with a calculated p-value, and generates multiple plots with additional information on gene-expression, HLA binding and peptide immunogenicity. For software development, we adopted the R language (version 4.4) and S4 object-oriented pattern. To ensure best practices, we leverage R development guidelines, including packages for software design, testing, and documentation, such as devtools (92), usethis (93), testthat (94), and roxygen2 (95), respectively. Next, we create a user-friendly application using the Shiny framework (96). The web application was built upon an interactive interface to produce data visualization and searchable data tables.

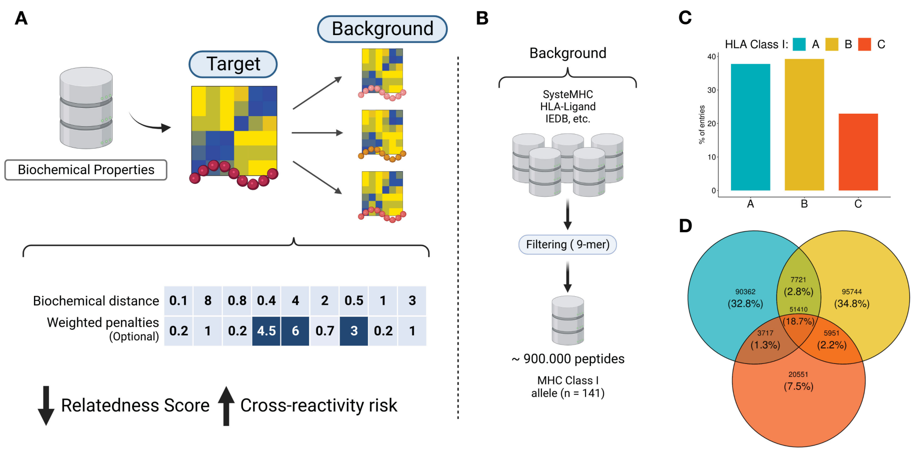
3 Results

3.1 Immunopeptidomics data can be leveraged for off-target toxicity prediction

Currently, the assessment of T-cell off-target toxicity risk in clinical and research settings is limited and heavily dependent on the accuracy of sequence-based HLA binding prediction tools. This dependence on additional layers of prediction can significantly increase the risk of false positives (97, 98). CrossDome, on the other hand, addresses this issue by screening for potentially cross-reactive peptides within “real” peptides, experimentally-determined by immunopeptidomics assays. This approach reduces the number of spurious candidates, and minimizes potential biases of HLA binding prediction tools. First, the methodology described in section 4.2 allowed us to convert peptide sequences into biochemical profiles (Figure 1A), which were in turn used to calculate distances between peptides (e.g., RdS). Using the RdS, we can perform pairwise comparisons between a tumor-associated query peptide, and a large reference dataset of self-derived peptides from immunopeptidomics databases. We created a reference dataset by retrieving over 900,000 eluted peptides from five different databases, covering 141 HLA Class I alleles (Figure 1B and Supplementary Figure S1A). In this combined reference database, HLA-B-restricted peptides are the most prevalent (40%), followed by HLA-A (38%) and HLA-C (22%) (Figure 1C). Although some peptides are shared among loci (Figure 1D), the large majority of the displayed peptides are HLA-exclusive.

3.2 CrossDome’s relatedness score outperforms alignment-based methods

As a proof of principle for CrossDome, we used the tumor-associated peptide from MAGEA3 (EVDPIGHLY) as a query, and evaluated the capacity of our method to recover the known cross-reactive peptide derived from TITIN (ESDPIVAQY) among the top-ranked putative off-targets. Using the aforementioned relatedness score, the TITIN off-target was predicted at the 99+ percentile rank, at position 27 out of 36,000 peptides displayed by HLA-A*01:01 (Figure 2A). The list of best-scored peptides reported



Biochemical distance	0.1	8	0.8	0.4	4	2	0.5	1	3
Weighted penalties (Optional)	0.2	1	0.2	4.5	6	0.7	3	0.2	1

Relatedness Score Cross-reactivity risk

FIGURE 1
The rationale behind the biochemical approach implementation. **(A)** Biochemical properties from AAIndex are summarized using dimensionality reduction techniques. The eigenvectors are utilized to convert peptide sequences into biochemical profiles. Next, similar peptides are screened in the immunopeptidomics database. Each pairwise comparison produces a distance-based metric called relatedness score (RdS), which is used for peptide-centered predictions. The relatedness calculation can optionally utilize a position-specific weighted vector, for TCR-centered predictions. Higher values are related with strong penalties (e.g., positions in dark blue). **(B)** A comprehensive immunopeptidomics database was created by leveraging several public datasets. **(C, D)** Quantitative description of HLA alleles in the CrossDome database. Image created with BioRender.com.

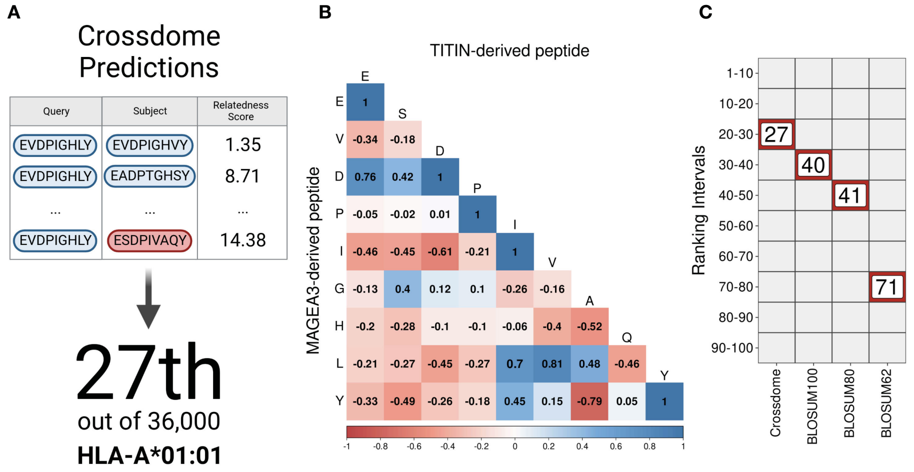


FIGURE 2
CrossDome performance against traditional approaches. **(A)** Evaluating CrossDome using MAGEA3-TITIN. Titin-derived peptide was recovered at position 27th out of 36,000 eluted peptides. **(B)** Residue-level comparison using Pearson correlation test. **(C)** Ranking-related heatmap displaying MAGEA3-TITIN positioning across distinct methods.

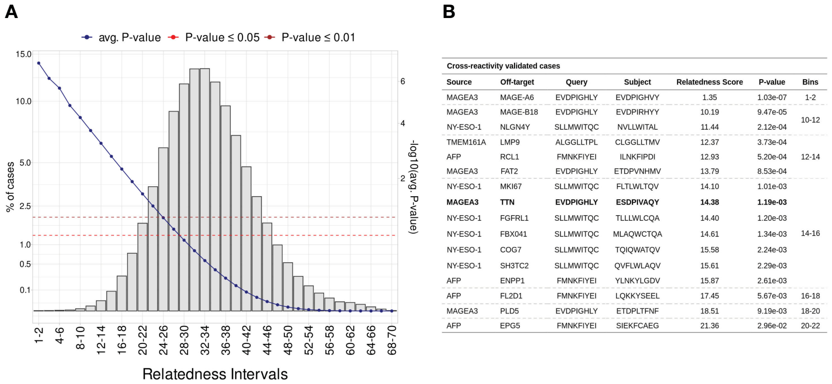
by CrossDomealso included MAGEA3 paralogs, other validated off-targets, and a few other highly similar peptides (Supplementary Table S3). In addition to the relatedness score, we measured residue-level correlation among the two peptides. It allowed us to evaluate the peptide composition and correlation between biochemically similar residues. In general, the MAGEA3-TITIN pair shows a strong correlation (Pearson p-value $\leq 2.2e^{-16}$). Figure 2B diagonal displays the residue-based correlation among MAGEA3 and TITIN-derived peptides. Note that highly correlated residues recapitulate expected amino acid biochemical similarities. For instance, we observe high correlation between polar residues with the same charge (e.g., glutamic acid and aspartic acid), and between non-polar residues (e.g., leucine and valine), while we see low correlation between polar residues with opposite charge (e.g., aspartic acid and histidine).

However, our implementation of the relatedness score goes beyond these obvious associations when determining peptide similarity. To demonstrate this point, we compared our ranking based on the relatedness score with alternative rankings based on traditional alignment-based metrics (e.g., substitution matrices). Keeping all other factors equal (e.g., same query and same reference universe of peptides), the highest BLOSUM matrix (i.e., BLOSUM100) placed the TITIN off-target at the 40th position. Higher numbers on BLOSUM matrices are more accurate for highly similar sequences, therefore not reasonable to predict cross-reactivity between unrelated peptides (99). On the other hand, BLOSUM62 should provide higher sensitivity for low-similarity peptides, but it displayed an even worse performance in our experiment, ranking TITIN at the 71st position. Figure 2C shows a ranking comparison between searches using the relatedness score, or substitution matrices. In fact, the relatedness score implemented in CrossDome outperformed alignment-based metrics regarding both sensitivity, as observed by the ranking of the TITIN-derived peptide, and computational performance, as the average run-time is 15 times faster than sequence alignment. Moreover, unlike sequence alignment methods, our approach does not depend on parameters such as gap and mismatch penalties.

3.3 A statistical threshold can be used to estimate off-target toxicity risk

Considering that most documented cases of T-cell off-target toxicity have been associated with molecular mimicry (20, 21, 52, 70), we expected that validated target/off-target pairs should in general present low relatedness scores. In this context, the relatedness distribution obtained in the Monte Carlo simulation described in section 4.3 can both be used to understand the dispersion of validated cases, and to identify confidence boundaries in our predictions (i.e., statistical threshold). As a result, we obtained a RdS distribution that largely resembles a Gaussian (u = 32.77, sd = 6.05, Shapiro-Wilk test ≤ 0.05). Next, we divided the distribution into intervals (i.e., "bins") encompassing worse to best-scored peptide pairs (Figure 3A). As expected, the bins with low relatedness score values (i.e., best-scored cases, Bin ≤ 16), were highly populated with experimentally validated cases (Figure 3B;. Supplemantary Table S1).

For instance, the peptide pair related to the MAGEA3-MAGEA6 cross-reactivity was ranked among the most meaningful values (RdS = 1.35, p-value $\leq 1.03e^{-02}$). MAGEA6 belongs to the melanoma-associated antigens, a paralog group with high sequence similarity (100). The MAGEA6-specific peptide deviates from MAGEA3 by a single conserved residue substitution at position eight (leucine to valine). Bin 14-16 holds the largest number of validated cases (n = 7), including the MAGEA3-TITIN pair. Finally, the AFP-EPG5 pair represents the last detectable cross-reactivity event (RdS = 21.36, p-value $\leq 2.96e^{-02}$). The standard threshold for CrossDomepredictions was defined as p-value ≤ 0.01, based on this RdS distribution.



Cross-reactivity validated cases						
Source	**Off-target**	**Query**	**Subject**	**Relatedness Score**	**P-value**	**Bins**
MAGEA3	MAGE-A6	EVDPIGHLY	EVDPIGHVY	1.35	1.03e-07	1-2
MAGEA3	MAGE-B18	EVDPIGHLY	EVDPIRHYY	10.19	9.47e-05	10-12
NY-ESO-1	NLGN4Y	SLLMWITQC	NVLLWITAL	11.44	2.12e-04	
TMEM161A	LMP9	ALGGLLTPL	CLGGLLTMV	12.37	3.73e-04	12-14
AFP	RCL1	FMNKFIYEI	ILNKFIPDI	12.93	5.20e-04	
MAGEA3	FAT2	EVDPIGHLY	ETDPVNHMV	13.79	8.53e-04	
NY-ESO-1	MKI67	SLLMWITQC	FLTLWLTQV	14.10	1.01e-03	14-16
MAGEA3	**TTN**	**EVDPIGHLY**	**ESDPIVAQY**	**14.38**	**1.19e-03**	
NY-ESO-1	FGFRL1	SLLMWITQC	TLLLWLCQA	14.40	1.20e-03	
NY-ESO-1	FBX041	SLLMWITQC	MLAQWCTQA	14.61	1.34e-03	
NY-ESO-1	COG7	SLLMWITQC	TQIQWATQV	15.58	2.24e-03	
NY-ESO-1	SH3TC2	SLLMWITQC	QVFLWLAQV	15.61	2.29e-03	
AFP	ENPP1	FMNKFIYEI	YLNKYLGDV	15.87	2.61e-03	
AFP	FL2D1	FMNKFIYEI	LQKKYSEEL	17.45	5.67e-03	16-18
MAGEA3	PLD5	EVDPIGHLY	ETDPLTFNF	18.51	9.19e-03	18-20
AFP	EPG5	FMNKFIYEI	SIEKFCAEG	21.36	2.96e-02	20-22

FIGURE 3
Establishing a statistical threshold for relatedness distribution. **(A)** Benchmarking putative cross-reactivity pairs in allele agnostic simulation. In total, 5 million peptide pairs were simulated and categorized based on relatedness intervals. For each pairwise comparison was calculated an empirical p-value. The left-sided y-axis shows the percentage of cases per bin (i.e., intervals). The blue line represents the average p-value in each interval (second y-axis). Additionally, red and brown are associated with standard p-value thresholds, 0.05 and 0.01, respectively. **(B)** Table covering experimentally validated CR cases. MAGEA3-TITIN was retrieved at 14-16 bin, which holds less than 0.1% of putative cases.

3.4 Contact maps can be inferred from structure and used to refine predictions

Although peptide similarity can provide us with a baseline probability for observing T-cell cross-reactivity, different cross-reactivity patterns can be observed for different T-cell clones (21). In other words, cross-reactivity is ultimately determined by the particular T-cell clone that is tested, and by the TCR-specific interactions with the cognate pHLA complex (52, 53). Here, we investigated the possibility of refining the CrossDomesearch in order to provide a TCR-specific off-target toxicity prediction. For that, we derived the molecular interactions between an engineered TCR (A3A) and its cognate MAGEA3-derived peptide-target from the available crystal structure of the TCRpMHC complex (Figure 4A). The contact map revealed a high preference for peptide positions 1, 4, 5, 7, and 8 (Figure 4B). Positions 4 and 5 showed higher interaction type diversity, including van der Waals, hydrogen bonds, and hydrophobic interactions.

Our A3A-derived contact map was consistent with curated data from IEDB-3D (85), and produced promising results when used to tailor CrossDome predictions for MAGEA3. Therefore, it supported our idea that the weights for CrossDome predictions can be derived from a reference TCRpHLA structure (i.e., with the cognate pHLA). However, we reasoned that significant differences could appear between contact maps derived from static (e.g., a single crystal structure) and dynamic sources (e.g., data derived from NMR experiments or molecular dynamics simulations). To investigate that, we calculated contacts for distinct time points in a 100 ns long molecular dynamics simulation (Figure 4B). In the overall profile obtained with the GetContacts package, position-

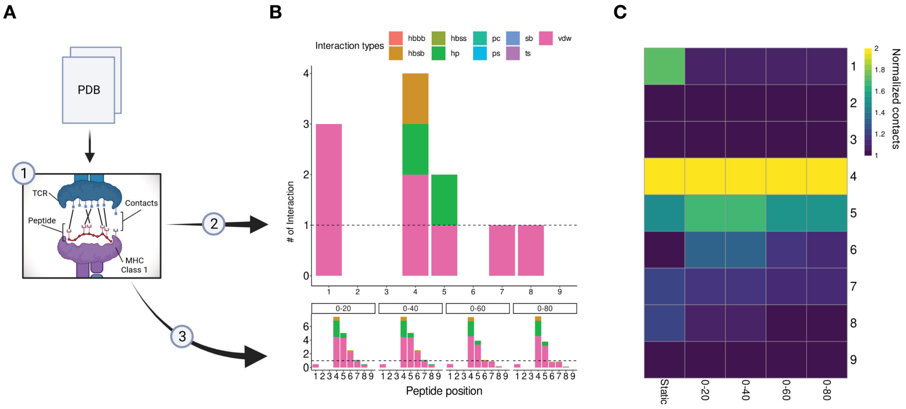


FIGURE 4
A3A-TCR contact maps construction. **(A)** A flowchart representing the contacts/interactions analysis. TCRpMHC interactions are depicted as black lines. The A3A crystal was submitted to two analyses (2), static, and (3) dynamic (molecular dynamics). **(B)** Each analysis produces a contact map summarized by peptide positions. The interaction types and frequency were retrieved using getcontacts. In total, nine interaction types can be detected, including hydrogen bond-related (hbbb, hbsb, hbss); salt bridge (sb); pi-cation (pc); pi-stacking (ps); t-stacking (ts); hydrophobic (hp); van der Waals (vdw). The dynamic contact map was summarized into four distinct time points: 0-20, 0-40, 0-60, and 0-80 nanoseconds. The overall profile showed similar hotspots compared to the static map. **(C)** Normalized contacts are displayed on the heatmap. Image created with BioRender.com.

specific preferences displayed by contact maps from molecular dynamics were highly similar to that of the crystal, including interaction types. However, a few interactions were lost, related to positions 1, 7, and 8 (Figure 4C). The reduction of TCR interactions over the simulated time might indicate these peptide positions were more involved with HLA interactions. In the case of position 1, we were still able to detect short-range Coulomb interactions with TCR residues in the simulation (Supplementary Figure S3). This type of interaction is not supported by GetContacts, but it can be computed with the gmx energy tool from Gromacs.

3.5 CrossDome predictions are consistent across protocols

CrossDome can predict T-cell cross-reactivity using two distinct approaches: i) based only on the biochemical profile (BP) of the peptides, or ii) using a combination of peptide's biochemical profile and TCR contact map (BP + CM). Note that in the first case we have a peptide-centered prediction, regardless of TCR information. While, on the second case, we have a TCR-centered prediction. To evaluate the overall performance of CrossDome using these different protocols, we leveraged a dataset of over 60 unrelated peptides known to be cross-reactive with MAGEA3, as described in section 2.1. As expected, the CM-based predictions increased the overall number of experimentally validated cases under statistical confidence (Figure 5A). Moreover, we could observe an incremental increase in sensitivity among protocols. The percentage of experimentally-validated cases in the top 50 ranking was equal to 63%, 71%, and 82% for BP only, CM-crystal, and CM-custom, respectively.

Subsequently, we analyzed qualitative changes in each protocol regarding the top 50 off-target candidates (Figure 5B), which were broadly shared between protocols (n = 33). The high level of agreement between protocols demonstrates CrossDome's accuracy even without the CM imputation. Further, at the sequence level, we noticed a reduction in the substitution rate at peptide positions 1, 3, and 9. Specifically, top ranking peptides on the CM-based screenings showed an even greater conservation of glutamic acid, aspartic acid, and tyrosine at these positions (Figure 5C), reflecting differences in the weighted vector values in each analysis (see section 2.4). The glutamic acid (position 1) displayed the largest increase across protocols, ranging from 74% up to 92% in residue conservation. The ranking of the known TITIN-derived cross-reactive peptide also improved between protocols, from the 27th position on BP to 8th and 6th position on CM-crystal and CM-custom, respectively. Together, these findings validate the contact maps as a reliable resource to tailor CrossDome predictions for a specific TCR.

3.6 CrossDome outperforms sCRAP when predicting known off-target toxicity cases

We used the aforementioned dataset of 16 validated cross-reactive peptide pairs (Figure 3B) to compare the top-ranking predictions by CrossDome with those of the recently published sCRAP tool (48). The same 4 cancer-associated antigens were used as queries for each tool (Supplementary Table S1), with default protocols. On CrossDome, we used the BP-based prediction (i.e., no TCR-based position-specific weights). On sCRAP, we used recommended settings (i.e., including default position-specific

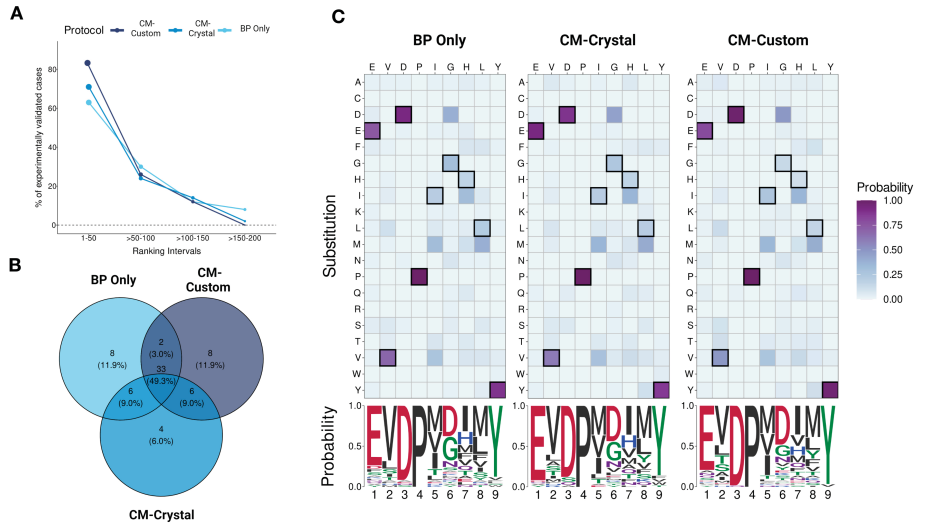


FIGURE 5
Evaluating protocols using yeast-displayed peptides. **(A)** Overall performance across CrossDome protocols. The y-axis corresponds to experimentally validated off-targets. The data was filtered considering best-scored candidates (p-value <0.01). The ranking was divided into four intervals. In total, three screenings were carried out, i) Biochemical properties (BP) only, ii) A3A-derived static contact map, and iii) manually curated contact map. **(B)** Venn diagram showing TOP50 best-scored peptides overlap among protocols. **(C)** TOP50 substitution heatmaps and seqLogos. The heatmaps present the substitution probability for each position across candidates. The seqlogos summarizes the most prevalent residue per position. The amino acid colors are related to biochemical classifications.

weights). In this context, CrossDome was able to predict 10 out of 16 off-targets within the top 50 sequences. For the other 6 cases, while not in the top 50, CrossDome still predicted the off-target at the 99+ percentile rank (Figure 3). sCRAP, on the other hand, was only able to predict 5 out of 16 within the top 50, and we could not determine the percentile rank for the other predictions since only the top 100 entries are provided.

Note that a fair direct comparison between peptide rankings obtained with these tools might not be possible, since they differ in multiple aspects (e.g., different search algorithms, different reference universe of self-derived peptides, different use of third-party methods, etc.). Instead, we decided to evaluate if there was any overlap between the top-ranking predictions by CrossDome and sCRAP in these experiments. To determine this overlap, we used only the top 50 predicted peptides by each tool, for each query (Supplementary Table S2). Any agreement between tools with such different implementations highlights their potential to identify dangerous off-targets. However, it is important to note that the entire list of candidates predicted by CrossDome is comprised of real immunopeptidomics-derived peptide targets. In addition, with the exception of the MAGEA3 prediction, multiple of these off-targets are also predicted to be both immunogenic and strong HLA-binders. In spite of that, only a fraction of these off-targets is also predicted by sCRAP. This difference is probably driven by the fact that sCRAP predictions are highly populated by predicted peptides, for which there is no available immunopeptidomics validation. Interestingly, in the case o MAGEA3, only the TITIN-derived peptide was kept as a candidate target after considering HLA binding and immunogenicity predictions. On the other hand, 12 of the top 50 predictions by CrossDome were also in the top 50 by sCRAP, with no further support to exclude them. The full list of top 50 predictions for both tools can be found in Supplementary Table S3.

3.7 MAGEA3-specific predictions can be refined based on mRNA expression and immunogenicity

In order to demonstrate the impact of incorporating mRNA expression for off-target localization and tolerability assessment, we applied this additional analysis to the best-scored peptides derived from the MAGEA3 screening using BP-based protocol. Note that a few peptides were dropped due to lack of similarity with RefSeq Protein database (101). Figure 6A shows the expression profile summary across all best-scored peptides (RdS P-value ≤ 0.01). The MAGEA3 screening presented higher percentages associated with low-specificity followed by tissue-enriched, tissue-enhanced, and not-expressed groups. In the tissue-enriched group, several candidates are strongly associated with the Heart/Skeletal Muscle (TITIN, TIMM50, PSMA3, etc.), Testis (MAGEA6, MAGEA11, etc.), Liver (LCAT, ABCC2, etc.), and Cerebellum (FAT2). Next, we displayed the expression levels for the top 50 peptides (Figure 6B). The experimental relevant MAGEA3 off-targets, TITIN and FAT2, were correctly assigned to respective tissues (41, 102, 103). Curiously, other liver-biased candidates identified here, such as LCAT, were not associated with off-target toxicity in previous clinical trials (26).

Next, we hypothesized that genes with ubiquitous (low-specificity) expression could be considered dangerous candidates due to a putative broad autoimmune response. For instance, NOP53, a ribosome biogenesis factor, showed a high expression level across all tissues (avg. PPM expression = 317.32). Furthermore, in terms of biochemical similarity, the NOP53-related peptide showed a relatedness score equal to 12.95. However, the NOP53 peptide (EVAPAGASY) has no evidence of T-cell positive assays on IEDB (66). The lack of T-cell assays reporting NOP53 can be indicative of either a low immunogenicity profile (e.g., low binding HLA affinity or lack of immunogenic features), or potential immune tolerance mechanisms.

To further investigate the immunogenicity potential of top-ranked putative off-targets, we conducted computational predictions of both HLA binding and peptide immunogenicity (Figure 6C). MAGEA3-specific predictions for HLA-binding (BA = 33.50) and immunogenicity (IS = 0.68) were used as references (dashed red line). In total, 23 off-target candidates were predicted as putative strong binders, i.e., IC50 ≤ 50 nM. Additionally, three candidates were reported with a meaningful immunogenicity level (e.g., > 0.65). Only two of the candidates were predicted to be both strong HLA-binders and highly immunogenic. As expected, TITIN has a similar binding affinity to the cognate MAGEA3 peptide (< 50 nM) with a superior immunogenic score (> 0.90), and would have been predicted as a dangerous off-target using our package. MAGEA6, another experimentally validated cross-reactive peptide, also displayed a similar profile. FAT2 and MAGEA18 had worse scores than MAGEA3 and other validated cases. The combination of high affinity (i.e., low predicted binding value) and high immunogenicity score can indicate the most dangerous candidates on CrossDome predictions.

3.8 Increased usability promoted through an R package and user-friendly interface

R is a well-established language in the bioinformatics community. To improve CrossDome's usability, we developed an R package containing several functions for predicting, analyzing, and visualizing cross-reactivity risk. Currently, the package allows the screening of putative off-targets by selecting a peptide-target (query) and our immunopeptidome database (subject). This database can also be combined with, or replaced by, a customized database. The CrossDome immunopeptidome database includes peptide immunogenicity and binding affinity predictions across several HLA Class I alleles. On average, the CrossDome run takes less than 1 minute per allele in a workstation machine (e.g., Intel Core i7 Processor, 32GB RAM), therefore allowing for batch analysis using several peptide-targets across distinct HLA alleles.

To foster reproducibility of results, we provided a tutorial/vignette series from basic usage to MAGEA3 analysis (Supplementary Materials). The MAGEA3 tutorial details contact map usage and calculation, including the comparison between BP and BP + CM predictions. In addition, the package was designed to store data reporting essential parameters and outputs in each execution.

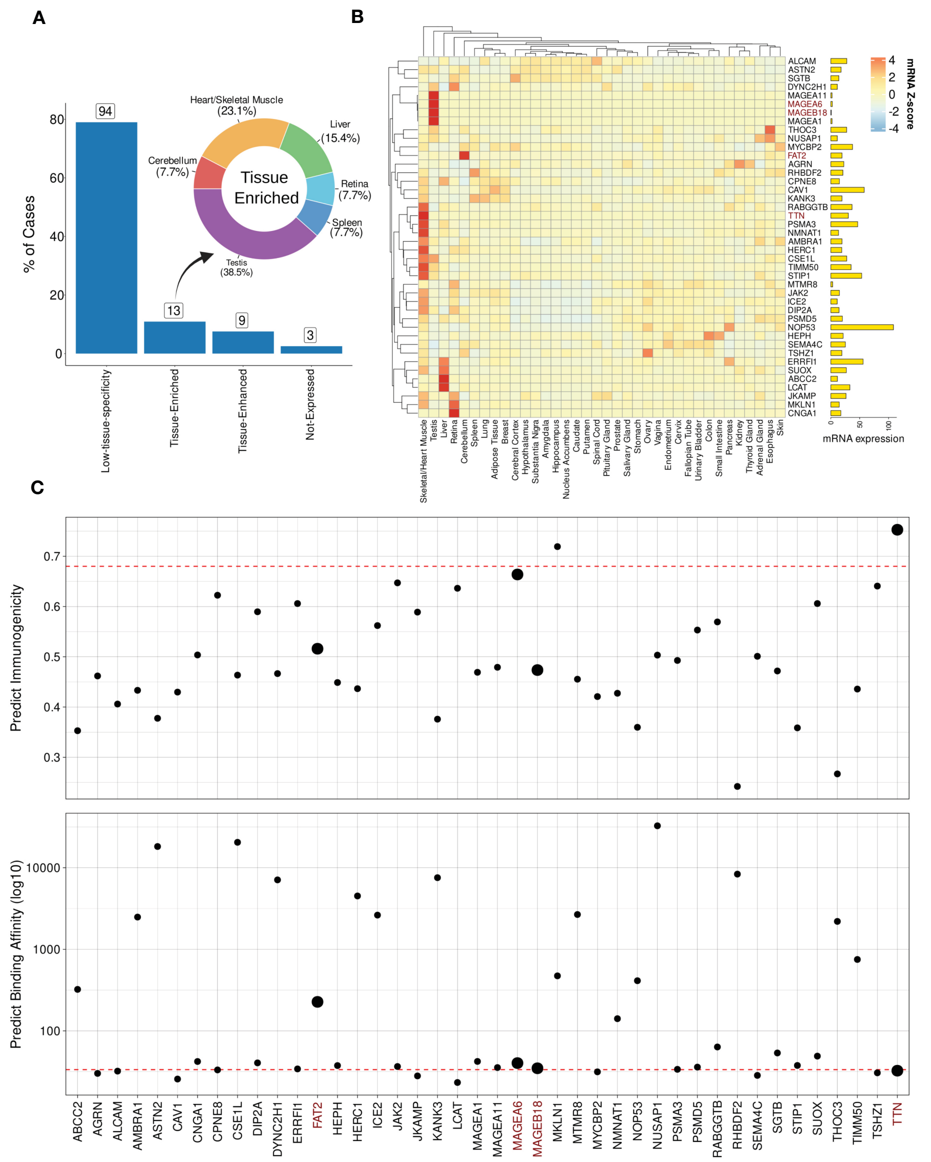


FIGURE 6
Expression and immunogenic profiles related to the cross-reactive candidates. **(A)** Tissue specificity groups across all best-scored candidates (p-value < 0.01). The donut plot displays tissues/organs related to "Tissue-enriched" candidates, i.e., genes with tissue-biases expression. **(B)** Heatmap with TOP50 best-scored peptides showing Z-score mRNA expression (pTPM). High (red) to low (blue) expression values on color key. Finally, total mRNA expression is represented on bar plot. **(C)** Peptide immunogenicity and MHC-antigen binding affinity predictions. The red dashed line represents the MAGEA3-derived peptide (EVDPIGHLY). The data points size reflects experimentally validated candidates. TTN corresponds to TITIN gene aliases.

CrossDome results can be manipulated using dplyr, a well-known R package for data science (104), therefore promoting greater versatility for bioinformaticians and computational biologists. Finally, we developed a web application that allows CrossDome basic usage. Currently, the app generates an interactive table supporting filtering, ranking, and downloading (Figure 7).

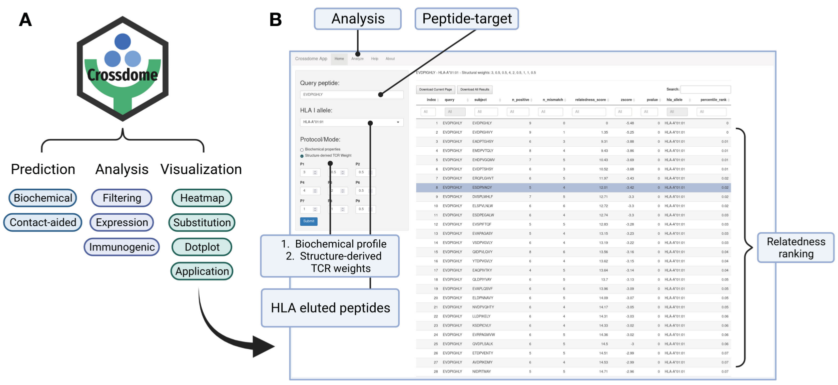
4 Discussion

Here we describe CrossDome, a software suite for predicting off-target toxicity risk for T-cell-based cancer immunotherapies. Using experimentally-validated cases as positive controls, we demonstrate that for a given tumor-associated target of interest for immunotherapy, this implementation can identify self-derived peptides that represent a potential risk for off-target toxicity mediated by T-cell cross-reactivity. Therefore, our method reduces the screening space from many thousand peptides (e.g., entire host proteome) to dozens of high-risk candidates, also providing information about the immunogenic profile and tissue-specificity of these putative off-targets. More importantly, CrossDome goes beyond previously proposed methods by providing a p-value associated with each off-target prediction. Based on the computed distribution of known cross-reactivities involving cancer-associated antigens, the p-values can be used to define a significance cut-off for the off-target toxicity risk.

It is also important to note that most of the previously published methods for T-cell cross-reactivity prediction are based on a "target-centered" perspective. By considering properties of the peptide or the pHLA complex as the key to find potentially similar off-targets, these methods can provide a baseline prediction of cross-reactivity (46, 48, 57, 59, 105). However, different T-cell clones will express unique TCRs, which can have different specificities towards pHLA complexes (24, 53). In turn, this can be reflected in different cross-reactivity patterns/profiles among T-cell clones recognizing the same cognate peptide (21, 39, 63). In order to account for that, predictions with CrossDome can be performed using two alternative approaches: one based on the biochemical profile of the query peptide, and the other based on a combination of peptide's biochemical profile and a TCR contact map. In the first case, we have the standard peptide-centered prediction, regardless of TCR information. This option is useful when the T-cell information is not available, or different T-cells can be triggered depending on the subject (e.g., peptide based vaccine design). Additionally, by changing the reference database, this approach can be used to identify relatedness between tumor associated antigens and microbial-derived peptides, therefore extending its applicability towards distance-to-self calculations on vaccine development studies (106–109). On the other hand, the second case provides a TCR-centered prediction (i.e., clone-specific off-target toxicity prediction). This option is preferred for users interested in the cross-reactivity profile of particular T-cell clones (e.g., in TCR-based immunotherapy). Tailoring CrossDome predictions with TCR information helps filtering out spurious candidates while recovering even more diverse peptide sequences, and mitigates the need for exhaustive search through experimental approaches (41, 52, 110). Consequently, CrossDome can reduce the time and costs associated with prioritizing antigens for T-cell-based immunotherapy, potentially accelerating their transition to clinical trials.

Our implementation choices on CrossDome are supported by extensive research on T-cell cross-reactivity, previously performed by us and by others. For instance, the role of pHLA structural similarity in T-cell recognition has been previously discussed, and even leveraged for cross-reactivity prediction (21, 58, 59, 111). It is also well known that T-cell recognition is driven by a few hotspots in the pHLA surface, and that T-cell cross-reactivity can be observed between peptides with very different sequences, as long as they share the same hotspots for TCR interaction (52, 63). Our work is also informed by previous implementations, which leveraged peptide sequence similarity, HLA binding prediction, and tissue expression patterns (46, 48, 105). Is worth noting that these tools rely on underlying AI-based methods for HLA-binding prediction, which have been a standard in the field. In addition, AI-based methods have enabled many other recent breakthroughs in biosciences (112). However, these methods were not yet successfully applied to the problem of T-cell cross-reactivity prediction, mostly due to the lack of large enough labeled training datasets. This landscape should change in the future, as data from high-throughput experimental methods for T-cell activation becomes more broadly available. In fact, available cognate TCR/peptide sequences from databases such as VDJdb (113) are already been leveraged to train AI-based models for the prediction of TCR specificity (69, 114–116). CrossDome can be used in combination with these methods to further accelerate the identification of peptide-targets and TCRs requiring experimental validation.

Different from previous methods, our tool does not rely on HLA-binding prediction to define the universe of self-derived peptides used in the search for off-targets. Instead, it relies on a local database of "real" peptides from immunopeptidomics studies. On one hand, this is a major advantage since it reduces false positive predictions (i.e., predicted off-targets that cannot be displayed by HLAs). On the other hand, this implementation restricts the universe to available experimental data, which might still be incomplete. Fortunately, immunopeptidomics has become a standard in the field, and we will continue updating our reference database as new datasets become available (65, 117). Note that CrossDome is also currently limited to the analysis of 9-mers, which account for most of the peptides displayed by class I HLA alleles. Future work will enable the expansion towards longer peptides, including those restricted to class II HLA alleles, therefore enabling cross-reactivity prediction for CD4+ T-cells. Although



Query peptide: EVDPIGHLY

HLA I allele: HLA-A*01:01

Protocol/Mode:
○ Biochemical properties
● Structure-derived TCR Weight

P1 3 | 0.5 | P2 0.5
P4 4 | 2 | P6 0.5
P7 1 | 1 | P9 0.5

EVDPIGHLY - HLA-A*01:01 - Structural weights: 3, 0.5, 0.5, 4, 2, 0.5, 1, 1, 0.5

index	query	subject	n_positive	n_mismatch	relatedness_score	zscore	pvalue	hla_allele	percentile_rank
All	All	All	All	All	All	All	All	All	All
1	EVDPIGHLY	EVDPIGHLY	9	0	0	-5.48	0	HLA-A*01:01	0
2	EVDPIGHLY	EVDPIGHVY	9	1	1.35	-5.25	0	HLA-A*01:01	0
3	EVDPIGHLY	EADPTGHSY	6	3	9.31	-3.88	0	HLA-A*01:01	0.01
4	EVDPIGHLY	EMDPVTQLY	8	4	9.43	-3.86	0	HLA-A*01:01	0.01
5	EVDPIGHLY	EHDPVGQMV	7	5	10.43	-3.69	0	HLA-A*01:01	0.01
6	EVDPIGHLY	EVDPTSHSY	6	3	10.52	-3.68	0	HLA-A*01:01	0.01
7	EVDPIGHLY	ERQPLGHVT	6	5	11.97	-3.43	0	HLA-A*01:01	0.02
8	EVDPIGHLY	ESDPIVAQY	5	4	12.01	-3.42	0	HLA-A*01:01	0.02
9	EVDPIGHLY	DVSPLMHLF	7	5	12.71	-3.3	0	HLA-A*01:01	0.02
10	EVDPIGHLY	ELSPVLNLW	6	6	12.72	-3.3	0	HLA-A*01:01	0.02
11	EVDPIGHLY	ESDPEGALW	6	4	12.74	-3.3	0	HLA-A*01:01	0.03
12	EVDPIGHLY	EVSPIFTQF	5	5	12.83	-3.28	0	HLA-A*01:01	0.03
13	EVDPIGHLY	EVAPAGASY	5	4	13.15	-3.23	0	HLA-A*01:01	0.03
14	EVDPIGHLY	VSDPVGVLY	6	4	13.19	-3.22	0	HLA-A*01:01	0.03
15	EVDPIGHLY	QIDPVLQVY	8	6	13.56	-3.16	0	HLA-A*01:01	0.04
16	EVDPIGHLY	YTDPVGVLY	6	4	13.62	-3.15	0	HLA-A*01:01	0.04
17	EVDPIGHLY	EAQPIVTKY	4	5	13.64	-3.14	0	HLA-A*01:01	0.04
18	EVDPIGHLY	QLDPIYVAY	6	5	13.7	-3.13	0	HLA-A*01:01	0.05
19	EVDPIGHLY	EVAPLQSVF	6	6	13.96	-3.09	0	HLA-A*01:01	0.05
20	EVDPIGHLY	ELDPNNAVY	6	5	14.09	-3.07	0	HLA-A*01:01	0.05
21	EVDPIGHLY	NVDPVQHTY	6	4	14.17	-3.05	0	HLA-A*01:01	0.05
22	EVDPIGHLY	LLDPIKELY	6	4	14.31	-3.03	0	HLA-A*01:01	0.06
23	EVDPIGHLY	KSDPICVLY	6	4	14.33	-3.02	0	HLA-A*01:01	0.06
24	EVDPIGHLY	EVRPAGMVW	6	5	14.36	-3.02	0	HLA-A*01:01	0.06
25	EVDPIGHLY	QVDPLSALK	6	5	14.5	-3	0	HLA-A*01:01	0.06
26	EVDPIGHLY	ETDPVENTY	5	5	14.51	-2.99	0	HLA-A*01:01	0.07
27	EVDPIGHLY	AVDPIKDMY	6	4	14.53	-2.99	0	HLA-A*01:01	0.07
28	EVDPIGHLY	NIDPITMAY	5	5	14.71	-2.96	0	HLA-A*01:01	0.07

FIGURE 7
CrossDome R package. **(A)** Main features include expression heatmap, immunogenic plot, best-scored substitutions plot, and web application. **(B)** CrossDome user-interface built with R's Shiny framework.

not associated with off-target toxicity, cross-reactivity involving these cells is a promising future direction due to newly discovered cytotoxic effects, and their role in mediating the production of autoreactive antibodies following vaccination (118, 119).

Another original component of our study relates to the demonstration of how structural information from a TCRpHLA complex can be used to derive a TCR contact map. Such contact map can be used by CrossDome as a per-peptide-position weighting system, enabling the aforementioned TCR-centered prediction of off-target toxicity. Note that automated extraction of the contact map from a TCRpHLA complex is not yet available on CrossDome, but it is a future implementation already being developed by our team. The best predictions in our TCR-centered experiments were obtained with a customized set of weights (CM-custom), considering dynamic contact maps, Coulomb interactions, and HLA binding motif. However, it is important to note that even the contact map derived from the TCR-peptide bonds detected on a single structure (CM-crystal) was already enough to recover over 71% of the validated cases in our dataset. Interested users can derive such static contact map from a growing number of crystal structures of TCRpHLA complexes being made available at PDB (81) and IEDB (66). Alternatively, we are also investigating if these contact maps can be accurately derived from 3D models, as new TCRpHLA modeling methods become available (120–123). If successful, this effort could enable future automation of structure-based contact map extraction from TCR sequences, such as those produced by single-cell TCR sequencing.

Finally, CrossDome can be easily incorporated into existing antigen discovery pipelines, therefore aiding the selection of better and safer peptide-targets and T-cell clonotypes for immunotherapy applications. The tool is under active development, and the beta version is available at https://github.com/AntunesLab/crossdome.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

DA conceived the original idea behind this work. DA and AF designed the methods and experiments. AF selected the datasets, implemented the methods, and executed the experiments. DA and AF analyzed and interpreted the results. AF wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

AF and DA were supported in part by funds from the University of Houston. Preliminary work on this project was partially funded by the Cancer Prevention and Research Institute of Texas (CPRIT) through a fellowship from the Gulf Cost Consortia on the Computational Cancer Biology Training Program (Grant No. RP170593).

Acknowledgments

We thank Jayvee Abella, Ph.D., and Alexandra Bishop, B.S., for the work in the implementation and proof-of-concept validation of a preliminary version of the code for computing biochemical similarities between peptides. We thank Jonatas Cesar, Ph.D., for assistance with mathematical modeling and experimental design that greatly improved this project. We thank Mavin H. Gee and K. Christopher Garcia for sharing the dataset of 60 synthetic peptides also recognized by a MAGEA3-specific T-cell clone, as previously reported in (73). Finally, we thank Finn Beruldsen for his helpful comments on the final manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1142573/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Background and validation datasets. (A) Schematic representation of sources used to build the reference self-derived background dataset for CrossDome. (B) Schematic representation of sources used for validation experiments.

SUPPLEMENTARY FIGURE 2
Dimensionality reduction of biochemical properties. (A) Scree plot presenting the explained variance per principal component (dimensions of the Principal Component Analysis). Twelve dimensions are accumulating 95% explained variance among AAIndex databases. (B) Two-dimensional plot displaying amino acid distribution according to the first and second principal components (dimensions) from the dimensionality reduction. Similar residues are grouped at closest distances, e.g., Isoleucine, Valine, and Leucine. Proline and Arginine are outliers.

SUPPLEMENTARY FIGURE 3
Histograms of Coulomb interactions. Short-range Coulomb interactions were calculated using the software gmx energy from Gromacs, and divided into intervals (x-axis). The energy reflects the interaction between the A3A

MAGEA3-specific TCR molecule and each amino acid from the MAGEA3-derived peptide. The y-axis shows the fraction frame reporting that energy range (molecular dynamics). Low values are related to proximity, i.e., putative strong interaction with the TCR molecule. Positive values are related to repulsion from TCR molecular, potentially favoring MHC interaction.

SUPPLEMENTARY FIGURE 4

CrossDome workflow diagram. The required and optional inputs are displayed on top. Peptide and HLA allele parameters only support 9-mer sequences and Class I alleles, respectively. TCR contacts and known off-targets (e.g., positive controls) can be introduced as optional inputs. The "CrossDome screening" process is performed over a reference "universe" database, using the relatedness score to find biochemically similar peptides. Next, a statistical threshold (p-value < 0.01) and data integration are applied to the potential off-target candidates. The "Gene-donor expression" step maps mRNA expression data from genes to each candidate. The expression profile provides measurements across several healthy tissues. The "Tissue specificity plot" summarizes the expression profile and tissue specificity. Additionally, the "Substitution plot" provides insights into sequence substitution among the best-scored candidates. Finally, the "immune-related predictions" combine immunogenicity and binding affinity predictions for each candidate. The package functions are named below each step in the workflow.

SUPPLEMENTARY TABLE 1

Additional information on the 16 validated off target toxicity cases, involving 4 well-known tumor-associated antigens.

SUPPLEMENTARY TABLE 2

Overlap between sCRAP and the top 50 ranking by CrossDome, for 4 different queries.

SUPPLEMENTARY TABLE 3

Full list of peptides among the top 50 ranking for both sCRAP and CrossDome, for 4 different queries.

References

1. Vandiedonck C, Knight JC. The human major histocompatibility complex as a paradigm in genomics research. *Brief Funct Genomic Proteomic* (2009) 8:379–94. doi: 10.1093/bfgp/elp010

2. Lizee G, Overwijk WW, Radvanyi L, Gao J, Sharma P, Hwu P. Harnessing the power of the immune system to target cancer. *Annu Rev Med* (2013) 64:71–90. doi: 10.1146/annurev-med-112311-083918

3. June CH. Adoptive t cell therapy for cancer in the clinic. *J Clin Invest* (2007) 117:1466–76. doi: 10.1172/jci32446

4. Rosenberg SA, Restifo NP, Yang JC, Morgan RA, Dudley ME. Adoptive cell transfer: a clinical path to effective cancer immunotherapy. *Nat Rev Cancer* (2008) 8:299–308. doi: 10.1038/nrc2355

5. Rosenberg SA, Yang JC, Sherry RM, Kammula US, Hughes MS, Phan GQ, et al. Durable complete responses in heavily pretreated patients with metastatic melanoma using t-cell transfer immunotherapy. *Clin Cancer Res* (2011) 17:4550–7. doi: 10.1158/1078-0432.ccr-11-0116

6. Robbins PF, Morgan RA, Feldman SA, Yang JC, Sherry RM, Dudley ME, et al. Tumor regression in patients with metastatic synovial cell sarcoma and melanoma using genetically engineered lymphocytes reactive with NY-ESO-1. *J Clin Oncol* (2011) 29:917–24. doi: 10.1200/jco.2010.32.2537

7. Perica K, Varela JC, Oelke M, Schneck J. Adoptive T cell immunotherapy for cancer. *Rambam Maimonides Med J* (2015) 6:e0004. doi: 10.5041/RMMJ.10179

8. Yee C, Lizee GA. Personalized therapy: tumor antigen discovery for adoptive cellular therapy. *Cancer J* (2017) 23:144–8. doi: 10.1097/PPO.0000000000000255

9. Arnaud M, Bobisse S, Chiffelle J, Harari A. The promise of personalized TCR-based cellular immunotherapy for cancer patients. *Front Immunol* (2021) 12:562701636. doi: 10.3389/fimmu.2021.562701636

10. Chrusciel E, Urban-Wojciuk Z, Arcimowicz´ Ł, Kurkowiak M, Kowalski J, Gliwinski M, et al. Adoptive cell therapy–harnessing antigen-specific t cells to target solid tumours. *Cancers* (2020) 12:683. doi: 10.3390/cancers12030683

11. Houot R, Schultz LM, Marabelle A. Kohrt h. t-cell–based immunotherapy: adoptive cell transfer and checkpoint inhibition. *Cancer Immunol Res* (2015) 3:1115–22. doi: 10.1158/2326-6066.568cir-15-0190

12. Chapuis AG, Roberts IM, Thompson JA, Margolin KA, Bhatia S, Lee SM, et al. T-Cell therapy using interleukin-21-Primed cytotoxic T-cell lymphocytes combined with cytotoxic T-cell lymphocyte antigen-4 blockade results in long-term cell persistence and durable tumor regression. *J Clin Oncol* (2016) 34:3787–95. doi: 10.1200/JCO.2015.65.5142

13. Duan Z. Ho m. T-cell receptor mimic antibodies for cancer immunotherapy. *Mol Cancer Ther* (2021) 20:1533–41. doi: 10.1158/1535-7163.MCT-21-0115

14. Yang X, Nishimiya D, Lochte S, Jude KM, Borowska M, Savvides CS, et al. Facile repurposing of peptide-MHC-restricted antibodies for cancer immunotherapy. *Nat Biotechnol* (2023). doi: 10.1038/s41587-022-01567-w

15. Özcan M, Jensen KM, Chamberlain CA, Donia M, Svane IM. Principles of adoptive t cell therapy in cancer. *Semin Immunopathol* (2019) 41:49–58. doi: 10.1007/s00281-018-0703-z

16. Lim WA, June CH. The principles of engineering immune cells to treat cancer. *Cell* (2017) 168:724–40. doi: 10.1016/j.cell.2017.01.016

17. Rath JA, Arber C. Engineering strategies to enhance TCR-based adoptive t cell therapy. *Cells* (2020) 9:1485. doi: 10.3390/cells9061485

18. D'Ippolito E, Schober K, Nauerth M, Busch DH. T Cell engineering for adoptive t cell therapy: 585 safety and receptor avidity. *Cancer Immunology Immunother* (2019) 68:1701–12. doi: 10.1007/s00262-019-02395-9

19. Sewell AK. Why must t cells be cross-reactive? *Nat Rev Immunol* (2012) 12:669–77. doi: 10.1038/nri3279

20. Mandl JN, Germain RN. Focusing in on t cell cross-reactivity. *Cell* (2014) 157:1006–8. doi: 10.1016/j.cell.2014.05.002

21. Antunes DA, Rigo MM, Freitas MV, Mendes MFA, Sinigaglia M, Lizee G, et al. Interpreting T-cell cross-reactivity through structure: implications for TCR-based cancer immunotherapy. *Front Immunol* (2017) 8:121. doi: 10.3389/fimmu.2017.0121

22. Welsh RM, Selin LK. No one is naive: the significance of heterologous T-cell immunity. *Nat Rev Immunol* (2002) 2:417–26. doi: 10.1038/nri820

23. Welsh RM, Che JW, Brehm MA, Selin LK. Heterologous immunity between viruses. *Immunol Rev* (2010) 235:244–66. doi: 10.1111/j.0105-2896.2010.00897.x

24. Degauque N, Brouard S, Soulillou JP. Cross-reactivity of TCR repertoire: current concepts, challenges, and implication for allotransplantation. *Front Immunol* (2016) 7:89. doi: 10.3389/fimmu.2016.00089

25. Lee CH, Salio M, Napolitani G, Ogg G, Simmons A, Koohy H. Predicting cross-reactivity and antigen specificity of T cell receptors. *Front Immunol* (2020) 11:565096. doi: 10.3389/fimmu.2020.565096

26. Linette GP, Stadtmauer EA, Maus MV, Rapoport AP, Levine BL, Emery L, et al. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* (2013) 122:863–71. doi: 10.1182/blood-2013-03-490565

27. Morgan RA, Chinnasamy N, Abate-Daga D, Gros A, Robbins PF, Zheng Z, et al. Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J Immunother* (2013) 36:133–51. doi: 10.1097/CJI.0b013e3182829903

28. Zhao Y, Bennett AD, Zheng Z, Wang QJ, Robbins PF, Yu LYL, et al. High-affinity TCRs generated by phage display provide CD4+ t cells with the ability to recognize and kill tumor cell lines. *J Immunol* (2007) 179:5845–54. doi: 10.4049/jimmunol.179.9.5845

29. Cai L, Galva LDC, Peng Y, Luo X, Zhu W, Yao Y, et al. Preclinical studies of the off-target reactivity of AFP158-specific TCR engineered t cells. *Front Immunol* (2020) 11:607. doi: 10.3389/fimmu.2020.00607

30. Hellman LM, Foley KC, Singh NK, Alonso JA, Riley TP, Devlin JR, et al. Improving t cell receptor on-target specificity *via* structure-guided design. *Mol Ther* (2019) 27:300–13. doi: 10.1016/j.ymthe.2018.12.010

31. Brudno JN, Kochenderfer JN. Toxicities of chimeric antigen receptor t cells: recognition and management. *Blood* (2016) 127:3321–30. doi: 10.1182/blood-2016-04-703751

32. Bedoya F, Frigault MJ, Maus MV. The flipside of the power of engineered t cells: observed and potential toxicities of genetically modified t cells as therapy. *Mol Ther* (2017) 25:314–20. doi: 10.1016/j.ymthe.2016.11.011

33. Hammill JA, Kwiecien JM, Dvorkin-Gheva A, Lau VW, Baker C, Wu Y, et al. A cross-reactive small protein binding domain provides a model to study off-tumor CAR-t cell toxicity. *Mol Ther - Oncolytics* (2020) 17:278–92. doi: 10.1016/j.omto.2020.04.001

34. Stone JD, Harris DT, Kranz DM. TCR affinity for p/MHC formed by tumor antigens that are self proteins: impact on efficacy and toxicity. *Curr Opin Immunol* (2015) 33:16–22. doi: 10.1016/j.coi.2015.01.003

35. van den Berg JH, Gomez-Eerland R, van de Wiel B, Hulshoff L, van den Broek D, Bins A, et al. Case report of a fatal serious adverse event upon administration of T cells transduced with a MART-1- specific T-cell receptor. *Mol Ther* (2015) 23:1541–50. doi: 10.1038/mt.2015.60

36. Dudley ME, Wunderlich JR, Robbins PF, Yang JC, Hwu P, Schwartzentruber DJ, et al. Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. *Science* (2002) 298:850–4. doi: 10.1126/science.1076514

37. Weber JS, Yang JC, Atkins MB, Disis ML. Toxicities of immunotherapy for the practitioner. *J Clin Oncol* (2015) 33:2092–9. doi: 10.1200/JCO.2014.60.0379

38. Zoete V, Irving M, Ferber M, Cuendet MA, Michielin O. Structure-based, rational design of T cell receptors. *Front Immunol* (2013) 4:268. doi: 10.3389/fimmu.2013.00268

39. Border EC, Sanderson JP, Weissensteiner T, Gerry AB, Pumphrey NJ. Affinity-enhanced T-cell receptors for adoptive T-cell therapy targeting MAGE-A10: strategy for selection of an optimal candidate. *Oncoimmunology* (2019) 8:e1532759. doi: 10.1080/2162402X.2018.1532759

40. Arber C, Feng X, Abhyankar H, Romero E, Wu MF, Heslop HE, et al. Survivin-specific T cell receptor targets tumor but not T cells. *J Clin Invest* (2015) 125:157–68. doi: 10.1172/JCI75876

41. Kula T, Dezfulian MH, Wang CI, Abdelfattah NS, Hartman ZC, Wucherpfennig KW, et al. T-Scan: a genome-wide method for the systematic discovery of T cell epitopes. *Cell* (2019) 178:1016–28. doi: 10.1016/j.cell.2019.07.009

42. Cornberg M, Wedemeyer H. Hepatitis c virus infection from the perspective of heterologous immunity. *Curr Opin Virol* (2016) 16:41–8. doi: 10.1016/j.coviro.2016.01.005

43. Wedemeyer H, Mizukoshi E, Davis AR, Bennink JR, Rehermann B. Cross-reactivity between hepatitis c virus and influenza a virus determinant-specific cytotoxic T cells. *J Virol* (2001) 75:11392–400. doi: 10.1128/JVI.75.23.11392-11400.2001

44. Vieira GF, Chies JA. Immunodominant viral peptides as determinants of cross-reactivity in the immune system–can we develop wide spectrum viral vaccines? *Med Hypotheses* (2005) 65:873–9. doi: 10.1016/j.mehy.2005.05.041

45. Frankild S, de Boer RJ, Lund O, Nielsen M, Kesmir C. Amino acid similarity accounts for T cell cross-reactivity and for "holes" in the T cell repertoire. *PloS One* (2008) 3:e1831. doi: 10.1371/journal.pone.0001831

46. Jaravine V, Raffegerst S, Schendel DJ, Frishman D. Assessment of cancer and virus antigens for cross-reactivity in human tissues. *Bioinformatics* (2017) 33:104–11. doi: 10.1093/bioinformatics/btw567

47. Braun BA, Schein CH, Braun W. DGraph clusters flaviviruses and β-coronaviruses according to their hosts, disease type, and human cell receptors. *Bioinf Biol Insights* (2021) 15:117793222110203. doi: 10.1177/11779322211020316

48. Yarmarkovich M, Marshall QF, Warrington JM, Premaratne R, Farrel A, Groff D, et al. Cross-HLA targeting of intracellular oncoproteins with peptide-centric CARs. *Nature* (2021) 599:477–84. doi: 10.1038/s41586-021-04061-6

49. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* (2016) 32:511–7. doi: 10.1093/bioinformatics/btv639

50. 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:W449–54. doi: 10.1093/nar/gkaa379

51. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat Biotechnol* (2019). doi: 10.1038/s41587-019-0322-9

52. Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, et al. Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* (2014) 157:1073–87. doi: 10.1016/j.cell.2014.03.047

53. Adams JJ, Narayanan S, Birnbaum ME, Sidhu SS, Blevins SJ, Gee MH, et al. Structural interplay between germline interactions and adaptive recognition determines the bandwidth of TCR-peptide MHC cross-reactivity. *Nat Immunol* (2016) 17:87–94. doi: 10.1038/ni.3310

54. Bentzen AK, Such L, Jensen KK, Marquard AM, Jessen LE, Miller NJ, et al. T Cell receptor fingerprinting enables in-depth characterization of the interactions governing recognition of peptide–MHC complexes. *Nat Biotechnol* (2018) 36:1191–6. doi: 10.1038/nbt.4303

55. Cole DK, Bulek AM, Dolton G, Schauenberg AJ, Szomolay B, Rittase W, et al. Hotspot autoimmune t cell receptor binding underlies pathogen and insulin peptide cross-reactivity. *J Clin Invest* (2016) 126:2191–204. doi: 10.1172/jci85679

56. Antunes DA, Rigo MM, Silva JP, Cibulski SP, Sinigaglia M, Chies JA, et al. Structural in silico analysis of cross-genotype-reactivity among naturally occurring HCV NS3-1073-variants in the context of HLA-A*02:01 allele. *Mol Immunol* (2011) 48:1461–7. doi: 10.1016/j.molimm.2011.03.019

57. Mendes MF, Antunes DA, Rigo MM, Sinigaglia M, Vieira GF. Improved structural method for T-cell cross-reactivity prediction. *Mol Immunol* (2015) 67:303–10. doi: 10.1016/j.molimm.2015.06.017

58. Dhanik A, Kirshner JR, MacDonald D, Thurston G, Lin HC, Murphy AJ, et al. In-silico discovery of cancer-specific peptide-HLA complexes for targeted therapy. *BMC Bioinf* (2016) 17. doi: 10.1186/s12859-016-1150-2

59. Mendes MFA, de Souza Bragatte M, Vianna P, de Freitas MV, Pohner I, Richter S, et al. MatchTope: a tool to predict the cross reactivity of peptides complexed with major histocompatibility complex I. *Front Immunol* (2022) 13:930590. doi: 10.3389/fimmu.2022.930590

60. Lin X, George JT, Schafer NP, Chau KN, Birnbaum ME, Clementi C, et al. Rapid assessment of T-cell receptor specificity of the immune repertoire. *Nat Comput Sci* (2021) 1:362–73. doi: 10.1038/s43588-021-00076-1

61. Pearngam P, Sriswasdi S, Pisitkun T, Jones AR. MHCVision: estimation of global and local false discovery rate for MHC class i peptide binding prediction. *Bioinformatics* (2021) 37:3830–8. doi: 10.1093/bioinformatics/btab479

62. Zhang L, Udaka K, Mamitsuka H, Zhu S. Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. *Briefings Bioinf* (2012) 13:350–64. doi: 10.1093/bib/bbr060

63. Zhang S, Bakshi RK, Suneetha PV, Fytili P, Antunes DA, Vieira GF, et al. Frequency, private specificity, and cross-reactivity of preexisting hepatitis c virus (HCV)-specific CD8+ T cells in HCV-seronegative individuals: implications for vaccine responses. *J Virol* (2015) 89:8304–17. doi: 10.1128/JVI.00539-15

64. Marcu A, Bichmann L, Kuchenbecker L, Kowalewski DJ, Freudenmann LK, Backert L, et al. HLA ligand atlas: a benign reference of HLA-presented peptides to improve t-cell-based cancer immunotherapy. *J ImmunoTher Cancer* (2021) 9:e002071. doi: 10.1136/jitc-2020-002071

65. Shao W, Pedrioli PGA, Wolski W, Scurtescu C, Schmid E, Vizcaino JA, et al. The SysteMHC atlas project. *Nucleic Acids Res* (2018) 46:D1237–47. doi: 10.1093/nar/gkx664

66. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res* (2019) 47:D339–43. doi: 10.1093/nar/gky1006

67. Trolle T, McMurtrey CP, Sidney J, Bardet W, Osborn SC, Kaever T, et al. The length distribution of class i–restricted t cell epitopes is determined by both peptide supply and MHC allele–specific binding preference. *J Immunol* (2016) 196:1480–7. doi: 10.4049/jimmunol.1501721

68. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* (2017) 46:315–26. doi: 10.1016/j.immuni.2017.02.007

69. Karapetyan AR, Chaipan C, Winkelbach K, Wimberger S, Jeong JS, Joshi B, et al. TCR fingerprinting and off-target peptide identification. *Front Immunol* (2019) 10:2501. doi: 10.3389/fimmu.2019.02501

70. Raman MC, Rizkallah PJ, Simmons R, Donnellan Z, Dukes J, Bossi G, et al. Direct molecular mimicry enables off-target cardiovascular toxicity by an enhanced affinity TCR designed for cancer immunotherapy. *Sci Rep* (2016) 6:18851. doi: 10.1038/srep18851

71. Chiou SH, Tseng D, Reuben A, Mallajosyula V, Molina IS, Conley S, et al. Cell specificities in human non-small cell lung cancer enables HLA inference and antigen discovery. *Immunity* (2021) 54:586–602. doi: 10.1016/j.immuni.2021.02.014

72. Luo X, Cui H, Cai L, Zhu W, Yang WC, Patrick M, et al. Selection of a clinical lead TCR targeting alpha-Fetoprotein-Positive liver cancer based on a balance of risk and benefit. *Front Immunol* (2020) 11:623. doi: 10.3389/fimmu.2020.00623

73. Gee MH, Yang X, Garcia KC. Facile method for screening clinical t cell receptors for off-target peptide-HLA reactivity. *Preprint BioRxiv* (2018). doi: 10.1101/472480

74. Ivanciuc O, Midoro-Horiuti T, Schein CH, Xie L, Hillman GR, Goldblum RM, et al. The property distance index PD predicts peptides that cross-react with IgE antibodies. *Mol Immunol* (2009) 46:873–83. doi: 10.1016/j.molimm.2008.09.004

75. Bryant DH, Moll M, Chen BY, Fofanov VY, Kavraki LE. Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction. *BMC Bioinf* (2010) 11:242. doi: 10.1186/1471-2105-11-242

76. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* (2007) 36:D202–5. doi: 10.1093/nar/gkm998

77. Li G, Iyer B, Prasath VBS, Ni Y, Salomonis N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for t-cell immunity. *Briefings Bioinf* (2021) 22. doi: 10.1093/bib/bbab160

78. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci* (1992) 89:10915–9. doi: 10.1073/pnas.89.22.10915

79. Pages H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: efficient manipulation of biological strings. (2020).

80. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika* (1965) 52:591–611. doi: 10.1093/biomet/52.3-4.591

81. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic Acids Res* (2000) 28:235–42. doi: 10.1093/nar/28.1.235

82. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PloS Comput Biol* (2017) 13:e1005659. doi: 10.1371/journal.pcbi.1005659

83. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, et al. CHARMM36m: an improved 765 force field for folded and intrinsically disordered proteins. *Nat Methods* (2017) 14:71–3. doi: 10.1038/nmeth.4067

84. Venkatakrishnan AJ, Fonseca R, Ma AK, Hollingsworth SA, Chemparathy A, Hilger D, et al. Uncovering patterns of atomic interactions in static and dynamic structures of proteins. *Preprint BioRxiv* (2019). doi: 10.1101/840694

85. Ponomarenko J, Papangelopoulos N, Zajonc DM, Peters B, Sette A, Bourne PE. IEDB-3d: structural data within the immune epitope database. *Nucleic Acids Res* (2010) 39:D1164–70. doi: 10.1093/nar/gkq888

86. Pall S, Zhmurov A, Bauer P, Abraham M, Lundborg M, Gray A, et al. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. *J Chem Phys* (2020) 153:134110. doi: 10.1063/5.0018516

87. Rammensee HG, Bachmann J, Emmerich NPN, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* (1999) 50:213–9. doi: 10.1007/s002510050595

88. Digre A, Lindskog C. The human protein atlas–spatial localization of the human proteome in health and disease. *Protein Sci* (2020) 30:218–33. doi: 10.1002/pro.3987

89. Jain A, Tuteja G. TissueEnrich: tissue-specific gene enrichment analysis. *Bioinformatics* (2018) 35:1966–7. doi: 10.1093/bioinformatics/bty890

90. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst* (2020) 11:42–8. doi: 10.1016/j.cels.2020.06.010

91. Cameron BJ, Gerry AB, Dukes J, Harper JV, Kannan V, Bianchi FC, et al. Identification of a titin derived HLA-A1-presented peptide as a cross-reactive target for engineered MAGE A3-directed T cells. *Sci Transl Med* (2013) 5:197ra103. doi: 10.1126/scitranslmed.3006034

92. Wickham H, Hester J, Chang W, Bryan J. *Devtools: tools to make developing r packages easier* (2022). Available at: https://devtools.r-lib.org/https://github.com/r-lib/devtools.

93. Wickham H, Bryan J, Barrett M. *Usethis: automate package and project setup* (2022). Available at: https://usethis.rlib.orghttps://github.com/r-lib/usethis.

94. Wickham H. Testthat: get started with testing. *R J* (2011) 3:5. doi: 10.32614/rj-2011-002

95. Wickham H, Danenberg P, Csardi G, Eugster M. *roxygen2: in-line documentation for r* (2022). Available at: https://roxygen2.r-lib.org/https://github.com/r-lib/roxygen2.

96. Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, et al. *Shiny: web application framework for r*. (2022).

97. Schaap-Johansen AL, Vujovic M, Borch A, Hadrup SR, Marcatili P. T Cell epitope prediction and its application to immunotherapy. *Front Immunol* (2021) 12:712488. doi: 10.3389/fimmu.2021.712488

98. Paul S, Croft NP, Purcell AW, Tscharke DC, Sette A, Nielsen M, et al. Benchmarking predictions 798 of MHC class i restricted t cell epitopes in a comprehensively studied model system. *PloS Comput Biol* (2020) 16:e1007757. doi: 10.1371/journal.pcbi.1007757

99. Pearson WR. Selecting the right similarity-scoring matrix. *Curr Protoc Bioinf* (2013) 43. doi: 10.1002/0471250953.bi0305s43

100. Zajac P, Schultz-Thater E, Tornillo L, Sadowski C, Trella E, Mengus C, et al. MAGE-a antigens and cancer immunotherapy. *Front Med* (2017) 4:18. doi: 10.3389/fmed.2017.00018

101. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* (2007) 35:D61–5. doi: 10.1093/nar/gkl842

102. Vazquez-Lombardi R, Jung JS, Schlatter FS, Mei A, Mantuano NR, Bieberich F, et al. High-throughput t cell receptor engineering by functional screening identifies candidates with enhanced potency and specificity. *Immunity* (2022) 55:1953–1966.e10. doi: 10.1016/j.immuni.2022.09.004

103. Zhao X, Kolawole EM, Chan W, Feng Y, Yang X, Gee MH, et al. Tuning t cell receptor sensitivity through catch bond engineering. *Science* (2022) 376. doi: 10.1126/science.abl5282

104. Wickham H, François R, Henry L, Muller K. *A grammar of data manipulation*. (2020).

105. Haase K, Raffegerst S, Schendel DJ, Frishman D. Expitope: a web server for epitope expression. *Bioinformatics* (2015) 31:1854–6. doi: 10.1093/bioinformatics/btv068

106. Zitvogel L, Kroemer G. Cross-reactivity between cancer and microbial antigens. *OncoImmunology* (2021) 10. doi: 10.1080/2162402x.2021.1877416

107. Redelman-Sidi G, Glickman MS, Bochner BH. The mechanism of action of BCG therapy for bladder cancer–a current perspective. *Nat Rev Urol* (2014) 11:153–62. doi: 10.1038/nrurol.2014

108. Schubert B, de la Garza L, Mohr C, Walzer M, Kohlbacher O. ImmunoNodes – graphical development of complex immunoinformatics workflows. *BMC Bioinf* (2017) 18. doi: 10.1186/s12859-017-1667-z

109. Sioud M. T-Cell cross-reactivity may explain the large variation in how cancer patients respond to checkpoint inhibitors. *Scandinavian J Immunol* (2018) 87:e12643. doi: 10.1111/sji.12643

110. Gee MH, Han A, Lofgren SM, Beausang JF, Mendoza JL, Birnbaum ME, et al. Antigen identification for orphan t cell receptors expressed on tumor-infiltrating lymphocytes. *Cell* (2018) 172:549–563.e16. doi: 10.1016/j.cell.2017.11.043

111. Tarabini RF, Rigo MM, Fonseca AF, Rubin F, Belle R, Kavraki LE, et al. Large-Scale structure- based screening of potential t cell cross-reactivities involving peptide-targets from BCG vaccine and SARS-CoV-2. *Front Immunol* (2022) 12:812176. doi: 10.3389/fimmu.2021.812176

112. Sapoval N, Aghazadeh A, Nute MG, Antunes DA, Balaji A, Baraniuk R, et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun* (2022) 13:1728. doi: 10.1038/s41467-022-29268-7

113. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* (2018) 46:D419–27. doi: 10.1093/nar/gkx760

114. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* (2017) 547:94–8. doi: 10.1038/nature22976

115. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* (2017). doi: 10.1038/nature22383

116. Jokinen E, Dumitrescu A, Huuhtanen J, Gligorijević V, Mustjoki S, Bonneau R, et al. TCRconv: predicting recognition between T cell receptors and epitopes using contextualized motifs. *Bioinformatics* (2023) 39. doi: 10.1093/bioinformatics/btac788

117. Martini S, Nielsen M, Peters B, Sette A. The immune epitope database and analysis resource program 2003-2018: reflections and outlook. *Immunogenetics* (2020) 72:57–76. doi: 10.1007/s00251-019-01137-6

118. Loyal L, Braun J, Henze L, Kruse B, Dingeldey M, Reimer U, et al. Cross-reactive CD4+ T cells enhance SARS-CoV-2 immune responses upon infection and vaccination. *Science* (2021) 374. doi: 10.1126/science.abh1823

119. Oh DY, Fong L. Cytotoxic CD4+ T-cells in cancer: expanding the immune effector toolbox. *Immunity* (2021) 54:2701–11. doi: 10.1016/j.immuni.2021.11.015

120. Gowthaman R, Pierce BG. TCRmodel: high resolution modeling of T cell receptors from sequence. *Nucleic Acids Res* (2018) 46:W396–401. doi: 10.1093/nar/gky432

121. Jensen KK, Rantos V, Jappe EC, Olsen TH, Jespersen MC, Jurtz V, et al. TCRpMHCmodels: structural modelling of TCR-pMHC class I complexes. *Sci Rep* (2019) 9:14530. doi: 10.1007/978-1-4939-9728-2_17

122. Li S, Wilamowski J, Teraguchi S, van Eerden FJ, Rozewicki J, Davila A, et al. Structural modeling of lymphocyte receptors and their antigens. *Methods Mol Biol* (2019) 2048:207–29. doi: 10.3389/fimmu.2021.686127

123. Peacock T, Chain B. Information-driven docking for TCR-pMHC complex prediction. *Front Immunol* (2021) 12:686127. doi: 10.3389/fimmu.2021.812176