frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 29 July 2022
DOI 10.3389/fimmu.2022.937587

Check for updates

OPEN ACCESS

EDITED BY
Marilia Cascalho,
University of Michigan, United States

REVIEWED BY
Chang Liu,
Washington University in St. Louis,
United States
Antony Nicodemus Antoniou,
Northumbria University,
United Kingdom

*CORRESPONDENCE
Matthias Niemann
matthias.niemann@pirche.com

SPECIALTY SECTION
This article was submitted to
Alloimmunity and Transplantation,
a section of the journal
Frontiers in Immunology

RECEIVED 06 May 2022
ACCEPTED 11 July 2022
PUBLISHED 29 July 2022

CITATION
Niemann M, Matern BM and
Spierings E (2022) Snowflake: A deep
learning-based human leukocyte
antigen matching algorithm
considering allele-specific
surface accessibility.
Front. Immunol. 13:937587.
doi: 10.3389/fimmu.2022.937587

Snowflake: A deep learning-based human leukocyte antigen matching algorithm considering allele-specific surface accessibility

Matthias Niemann[1]*, Benedict M. Matern[2] and Eric Spierings[2,3]

[1]Research and Development, PIRCHE AG, Berlin, Germany, [2]Center for Translational Immunology, University Medical Center, Utrecht, Netherlands, [3]Central Diagnostic Laboratory, University Medical Center, Utrecht, Netherlands

Histocompatibility in solid-organ transplantation has a strong impact on long-term graft survival. Although recent advances in matching of both B-cell epitopes and T-cell epitopes have improved understanding of allorecognition, the immunogenic determinants are still not fully understood. We hypothesized that HLA solvent accessibility is allele-specific, thus supporting refinement of HLA B-cell epitope prediction. We developed a computational pipeline named Snowflake to calculate solvent accessibility of HLA Class I proteins for deposited HLA crystal structures, supplemented by constructed HLA structures through the AlphaFold protein folding predictor and peptide binding predictions of the APE-Gen docking framework. This dataset trained a four-layer long short-term memory bidirectional recurrent neural network, which in turn inferred solvent accessibility of all known HLA Class I proteins. We extracted 676 HLA Class-I experimental structures from the Protein Data Bank and supplemented it by 37 Class-I alleles for which structures were predicted. For each of the predicted structures, 10 known binding peptides as reported by the Immune Epitope DataBase were rendered into the binding groove. Although HLA Class I proteins predominantly are folded similarly, we found higher variation in root mean square difference of solvent accessibility between experimental structures of different HLAs compared to structures with identical amino acid sequence, suggesting HLA's solvent accessible surface is protein specific. Hence, residues may be surface-accessible on e.g. HLA-A*02:01, but not on HLA-A*01:01. Mapping these data to antibody-verified epitopes as defined by the HLA Epitope Registry reveals patterns of (1) consistently accessible residues, (2) only subsets of an epitope's residues being consistently accessible and (3) varying surface accessibility of residues of epitopes. Our data suggest B-cell epitope definitions can be refined by considering allele-specific solvent-accessibility, rather than aggregating HLA protein surface maps by HLA class or locus. To support studies on epitope analyses in organ transplantation, the calculation of donor-allele-specific solvent-accessible amino acid mismatches was implemented as a cloud-based web service.

KEYWORDS

HLA, epitope, epitope matching, antibodies, deep-learning, neural network, 3D-structures, structure prediction

Introduction

Histocompatibility of patients and organ donors has been shown to correlate with long-term graft survival in organ transplantation (1). The formation and/or presence of antibodies towards the mismatched donor HLA plays an important role in that. Antibody recognition of donor HLA is dependent on the patient's immune reactivity towards determinants on these mismatched donor HLA. The determinants that specifically interact with these antibodies have been designated as the antibody epitopes. The amino acid composition and tertiary structure of these donor epitopes may differ from the epitopes formed by recipients' HLA proteins, allowing B-cell receptors to distinguish self epitopes from non-self epitopes, allowing for effective immune responses (2).

Several algorithms have been developed to convert HLA mismatches into B-cell epitope mismatches between two individuals. The HLAMatchmaker algorithm was the first HLA-centric algorithm to define antibody-accessibility based on a limited set of experimental crystal structures. The HLAMatchmaker algorithm groups amino-acid positions and their corresponding configurations for defining the so-called Eplets (3). The EPRegistry database has subsequently been developed as a central repository listing these Eplets (4, 5). By defining residue-specific physicochemical properties, potential explanations for epitope-specific antibody affinity were provided (6). The polymorphism of the HLA-gene region, compounded by pairing of two individuals in a transplantation setting, however, yields such a large number of combinations of HLA epitopes, that verifying each mismatched epitope indistinguishably by antibody reaction patterns is challenging if possible at all (7). Moreover, it has been suggested that basically every amino-acid mismatch is capable of triggering an immune response (8, 9). The HLA-EMMA algorithm exploits amino-acid mismatching, by only considering polymorphic amino-acid residues at solvent-accessible positions based on reported experimental structures of various HLA loci as potential B-cell receptor targets (10).

There are less than 100 distinct HLA Class I and less than 50 distinct HLA Class II crystal structures deposited in the RCSB Protein Data Bank (PDB, www.rcsb.org) (11), while more than 13,600 classical HLA Class I and more than 5,600 HLA Class II protein variants have been reported by the IPD-IMGT/HLA database [version 3.47 (12)]. Given this imbalance and the complex and time-consuming process of X-ray crystallography, it is not feasible to generate structure data for every HLA protein in order to identify surface residues. Thus, neural network predictors have been suggested to predict arbitrary proteins' surfaces (13, 14). Refining and integrating aforementioned concepts, we developed a deep-learning prediction pipeline for solvent accessibility defining allele-specific surface epitopes. Our surface prediction is trained on experimental HLA structures of the PDB, which were supplemented by HLA structures predicted by DeepMind's recent AlphaFold protein folding predictor (15–17). Due to the structural differences between HLA Class I and Class II proteins and the different distribution of available experimental structures, our prediction pipeline's implementation needs specific optimization based upon the respective HLA classes. Herein, we present the results for our HLA Class I-specific pipeline. Based on our surface predictor, we implemented a matching algorithm as a web service allowing for real-time application in transplantation diagnostics. The full stack matching algorithm, named Snowflake, is available online (www.pirche.com), featuring machine-readable batch matching and a human-readable detail browser.

Materials and methods

Extraction of experimental structures from PDB

HLA protein structures were extracted from the PDB using the full text search term “hla”. The found structures were filtered by structures comprising chains “A”, “B” and “C”, whereas chain B's amino-acid sequence was expected to match beta-2-microglobulin (UniprotKB entry P61769). The supposed alpha chains were aligned to the IPD/IMGT-HLA database (12) (version 3.46.0) using BLAST (18) *via* the Biopython library (19). Structures were only considered further if a BLAST score of 1300 was exceeded. HLA-allele names and identifiers were assigned to the protein structures based on the maximum number of identical amino acid configurations and lowest HLA identifier in case of a tie.

Structural prediction of non-crystallized HLA proteins

Given the large number of known HLA genes and the complexity of generating structural data experimentally, the AlphaFold (15–17) (DeepMind Technologies Limited, London, UK) protein structure inference pipeline (v2.0) was applied to supplement the PDB data. As the public AlphaFold Database does not cover HLA variants, the AlphaFold software was installed on an Amazon Web Services Elastic Cloud Compute

Abbreviations: APE-Gen, Anchored Peptide-MHC Ensemble Generator; BRNN, bidirectional recurrent eural network; IEDB, Immune Epitope DataBase; HLA, human leukocyte antigen; MSE, mean squared error; PDB, Protein Data Bank; ReLU, rectified linear unit; RMSDA, root mean square of dihedral angle differences; RMSSA, root mean square of solvent accessibility differences; SDA, square of dihedral angle difference; SE, squared error; SSA, square of solvent accessibility difference.

server (Amazon Web Services Inc., Seattle, US) to run *de novo* calculations. Generating HLA-protein 3D structures *via* AlphaFold is a heavy computational task. Consequently, only HLA proteins were selected for structure prediction when their serologic and antigenic groups were not already covered by structures from the PDB. For visualization purposes, structures were superimposed as described by Golub et al. (20) *via* the Biopython library (19).

Render peptide in predicted structures' binding groove

The inferred structures comprise the proteins' respective alpha chain and the beta-2-microglobulin, but lack a bound peptide. The presence of such a peptide alters the solvent accessibility for residues involved in the peptide-binding domain. Despite AlphaFolds high prediction accuracy and capability to infer multimeric structures, it is not designed to predict binding of antibodies (17). To supplement the predicted structures with peptides, the Immune Epitope Database (IEDB, www.iedb.org) (21) was screened for nonameric peptides being reported as binders to the respective HLA protein. For predicted alleles without reported binders, the IEDB-reported peptides of the allele with the highest amino acid identity in the extracellular domains were considered as binders. For each allele, a subset of ten peptides was extracted and each peptide's position in the HLA binding groove was predicted by the Anchored Peptide-MHC Ensemble Generator (APE-Gen) (22) to account for peptide-specific changes of accessibility.

Structure comparison

Comparing experimental protein structures by calculating the Euclidean distance of the respective residues' coordinates is sensitive to local differences that propagate through the whole tertiary structure. The relatively small changes in dihedral angles in the backbone may twist the entire molecule, even though large functional units remain identically shaped. Increased Euclidean distances of all subsequent positions will thus be overrepresented. Furthermore, comparing Euclidean distances requires structures to be superimposed. Although axes units are identical, the experimental structures deposited in the PDB have varying coordinate centers and axes orientations, requiring prior computational superimposition of the structures, which may introduce additional bias. As a robust alternative, torsion angles of the amino acids' backbones - characterized as dihedral angles phi (φ) and psi (ψ) - have been considered. The position-specific square of dihedral angle differences (SDA) and the overall root mean SDA (RMSDA) were calculated. Compared to evaluating Euclidean distances of atom pairs of identical amino acid sequence structures, RMSDA is superimposition-independent and local differences do not propagate through the structure (as reviewed by (23), Figure 1). Additionally, the position-specific square of solvent accessible surface area differences (SSA) and the overall root mean SSA (RMSSA) were calculated as a functional distance metric. Predicted structures of the five most frequent proteins of experimental structures reported in the PDB were compared to estimate the prediction performance. However, these predicted structures were not used in the neural network training data set.

$$SDA_i = (\varphi_{a_i} - \varphi_{b_i})^2 + (\psi_{a_i} - \psi_{b_i})^2$$

$$RMSDA = \sqrt{\frac{1}{2n}\sum_{i=1}^{n}(\varphi_{a_i} - \varphi_{b_i})^2 + (\psi_{a_i} - \psi_{b_i})^2}$$

$$SSA_i = (SA_{a_i} - SA_{b_i})^2$$

$$RMSSA = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(SA_{a_i} - SA_{b_i})^2}$$

Calculation of protein-residue-specific surface area

For all experimental and inferred HLA-protein structures, each residues' solvent accessible surface area was calculated considering the Shrake-Rupley algorithm (24), implemented by the Biotite library (25). Essentially, this algorithm is the numerical equivalent of rolling a marble (i.e. probe) over a structure (i.e. HLA protein) to identify which parts of the structure (i.e. residues) get in contact with the marble. The probe radius was set to 1.4 Å corresponding to the size of water molecules. Surface area computation of alpha chain residues considered the proteins' alpha chain, beta-2-microglobulin and bound peptide simultaneously.

Neural network design

The experimental and inferred HLA structures comprise only a fraction of known HLA Class I proteins. To create a database of surface accessible residues for all HLA Class I proteins, a long short-term memory bidirectional recurrent neural network (BRNN) (26) was implemented, that chained the HLA alpha chains' amino-acid configurations as one-hot-encoded inputs and the residues' surface areas as output. The network configuration stacked a bidirectional layer with rectified linear unit (ReLU) activation of 100 neurons, a second bidirectional layer with 64 ReLU neurons, a third bidirectional layer with 32 ReLU neurons and the output layer

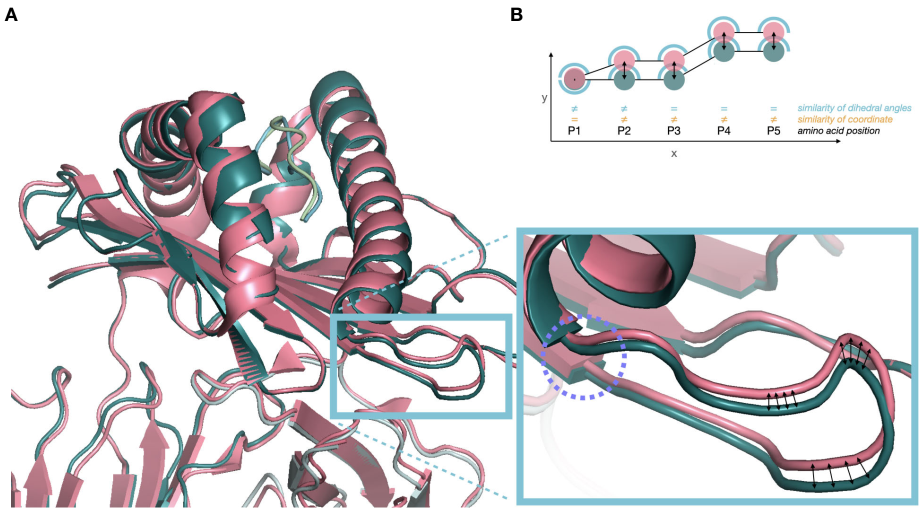


FIGURE 1
(A) Cartoon plot of two independent crystal structures of HLA-A*02:614 in overlay. In pink structure 3MRE (PDB DOI: 10.2210/pdb3MRE/pdb) and in green structure 3GSR (PDB DOI: 10.2210/pdb3GSR/pdb). Superimposition was based on atoms of identical amino acid residues. The blue box indicates small changes in dihedral angles (purple circle) propagating into large Euclidean distances in the following loop structure. **(B)** Euclidean distance between atom positions (orange) may overestimate the structural difference (P3-P5) due to distance-propagation of angular difference at P1/P2 (blue) or suboptimal global superimposition. However, dihedral angles at P3-P5 are identical, only considering P1/P2 as structurally different.

considering a linear activation function. Position-specific surface area was limited to 100, corresponding to the 85th percentile, in order to account for protruding residues with disproportionate surface area. Surface area was divided by 100 resulting in values ranging from 0 to 1. Network training used Adam optimization for efficient gradient descent (27). The model-generation and inference was implemented in the Python programming language (Python Software Foundation, version 3.5.1) using Keras (https://keras.io)/Tensorflow 2 (https://www.tensorflow.org).

Comparison of calculated surface area with other methods

The performance of the applied surface prediction pipeline was evaluated by comparing the predicted surface accessibility of 72 IMGT-IPD/HLA-reported HLA Class I reference alleles (based on IMGT 3.46.0 (28),) with the predictions provided by NetSurfP 2.0 (13) and PaleAle 4.0 (14). Relative solvent accessibility predicted by PaleAle was divided by 100 to match the other predictors' outputs. The produced relative solvent accessibility scores were mean-centered. Position-specific squared error (SE) and mean squared errors (MSE) were calculated pairwise between the predictors and reference alleles. Amino acid sequence positions with large SE were visualized in PyMol 2.3.0. Amino acid positions with solvent accessibility scores exceeding the respective predictor's mean solvent accessibility were considered accessible and inaccessible otherwise.

Epitope matching based on surface-accessible amino acid residues

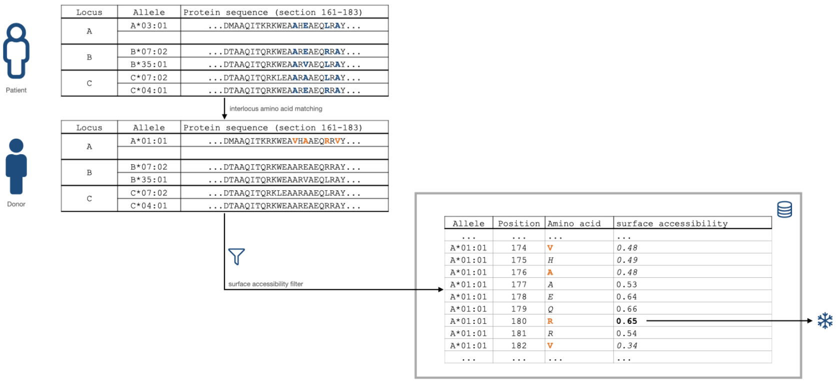
The proposed method considers donor amino-acid mismatches at allele-specific positions exceeding a surface accessibility threshold (i.e. are predicted to be surface-accessible) between two individuals as epitope mismatches (Figure 2). The sum of such amino acid mismatches is considered the Snowflake epitope mismatch score. The algorithm is parameterized in the surface accessibility threshold and the mode of reference loci: (1) intralocus matching considers only the same recipient locus, whereas (2) interlocus matching references all recipient's HLA Class I proteins.

Statistical analysis

All calculations were executed in R software (R 3.6.1, R Foundation for Statistical Computing, Vienna, Austria).

Results

The PDB search yielded 1593 structures that were filtered down to 676 HLA Class-I experimental structures (Figure 3). The majority of experimental HLA Class I structures correspond to the A*02:01 (n=272, ~40%) (Figure 4). Based on the



Locus	Allele	Protein sequence (section 161-183)
A	A*03:01	...DMAAQITKRKWEAAHEAEQLRAY...
B	B*07:02	...DTAAQITQRKWEAAREAEQRRAY...
	B*35:01	...DTAAQITQRKWEAARVAEQLRAY...
C	C*07:02	...DTAAQITQRKLEAARAAEQLRAY...
	C*04:01	...DTAAQITQRKWEAAREAEQRRAY...

Locus	Allele	Protein sequence (section 161-183)
A	A*01:01	...DMAAQITKRKWEAVHAAEQRRVY...
B	B*07:02	...DTAAQITQRKWEAAREAEQRRAY...
	B*35:01	...DTAAQITQRKWEAARVAEQLRAY...
C	C*07:02	...DTAAQITQRKLEAARAAEQLRAY...
	C*04:01	...DTAAQITQRKWEAAREAEQRRAY...

Allele	Position	Amino acid	surface accessibility
...	...	...	...
A*01:01	174	**V**	*0.48*
A*01:01	175	*H*	*0.49*
A*01:01	176	**A**	*0.48*
A*01:01	177	*A*	0.53
A*01:01	178	*E*	0.64
A*01:01	179	*Q*	0.66
A*01:01	180	**R**	**0.65**
A*01:01	181	*R*	0.54
A*01:01	182	**V**	*0.34*
...	...	...	...

FIGURE 2
Snowflake HLA matching algorithm schematic. Donor amino-acid mismatches with a predicted threshold-exceeding allele- and position-specific solvent accessibility (0.00 to 1.00, 0.00 being inaccessible, 1.00 being exposed) increase the Snowflake score by one.

experimental structures' distinct HLA proteins, an additional 37 Class-I alleles (14 HLA-A, 16 HLA-B, 7 HLA-C) were selected for structure prediction, covering all HLA Class I antigenic groups. Rendering epitopes into the binding grooves of 37 predicted HLA structures failed in 41 of 370 cases (median = 0, mean = 1.11) due to internal errors in APE-Gen where energy values exceed numeric limits. Distributions of position-specific solvent accessible surface area are provided in Supplementary Figure 1. Variance in surface area can be observed both for positions in interaction with the bound peptide (29), as well as residues outside the binding groove.

The distribution of pairwise RMSDA for the most common structures is similar compared to the overall RMSDA distribution considering structural variance across all HLA Class I proteins. Predicted structures' RMSDA with their corresponding experimental structures is similar to the RMSDA deviation between experimental structures of the same allele (Figure 5). Certain regions of the HLA have higher SDA compared to rigid regions with low SDA (Supplementary Figure 2). Noteworthy, SDA does not correspond to levels of solvent accessibility (Supplementary Figure 3, Spearman's rho = 0.135, p< 0.001).

There is considerable variance in RMSSA observed within the various reported experimental structures of the same allele. The RMSSA across all HLA Class I experimental structures is however significantly higher compared to the allele-specific RMSSAs, indicating a higher variance between individual alleles (Figure 6). Also for the position-wise variance of surface

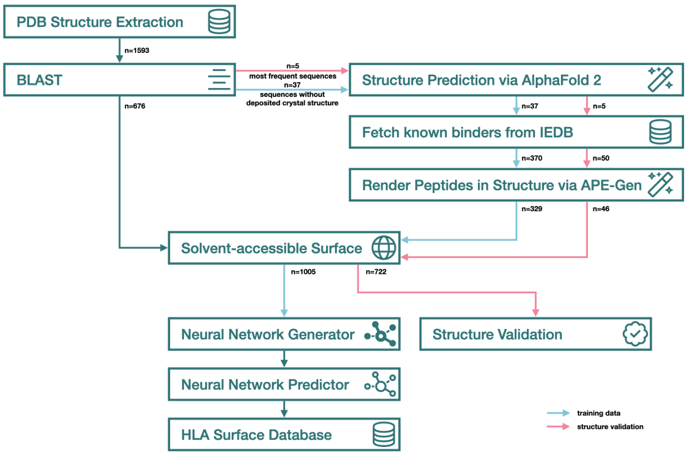


FIGURE 3
Snowflake surface accessibility prediction pipeline. Blue arrows represent training data, pink arrows represent validation data, green arrows represent data flow identical in both steps.

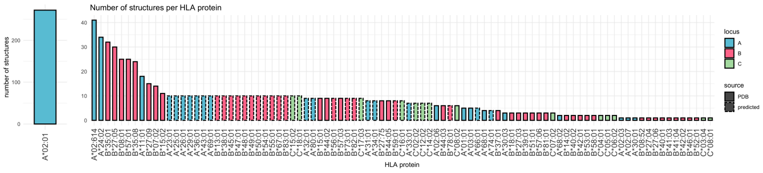


FIGURE 4
Bar chart indicating the distribution of alleles in experimental structures. The number of structures is encoded on the y axis. Color indicates locus, dashed border indicates structure prediction as outlined in the Materials and Methods section.

area, certain regions of the HLA are subject to higher variance in SSA than others (Supplementary Figure 4).

Compare surface prediction with reference tools

After training the Snowflake surface predictor over 400 epochs, the loss function converged to ~0.02, so further training did not alter neural network weights. The average predicted relative solvent accessibility ranged lower for NetSurfP and PaleAle (i.e. reference predictors) compared to Snowflake (Figure 7A). After centering, the reference predictors produced similar predictions with low MSE across reference alleles (Figure 7B). Snowflake had considerably higher MSEs considering the reference predictors (Figure 7B).

The low MSE between NetSurfP and PaleAle was also observed in the low position-specific SE (Supplementary Figure 5) at the majority of amino acid positions. Position-specific SE for Snowflake was substantially higher in certain regions (Supplementary Figure 5). Classifying solvent accessibility by exceeding the respective predictor's mean score yielded a considerable number of allele-dependent (e.g P3, P110, P190) and allele-independent (e.g. P31, P167 or P219) deviations between the Snowflake and the reference predictors (Figure 8). As an example, the solvent accessibility prediction and SE of A*02:01 have been provided in Supplementary Figure 6.

Generating visual representations of some of these characteristic positions indicates allele-specific differences in surface prediction (Figure 9). Amino acid position 167 of A*01:01 is not predicted to be solvent-accessible by Snowflake, PaleAle and NetSurfP. Despite PaleAle and NetSurfP predicting the same position also not being accessible in an A*02:01,

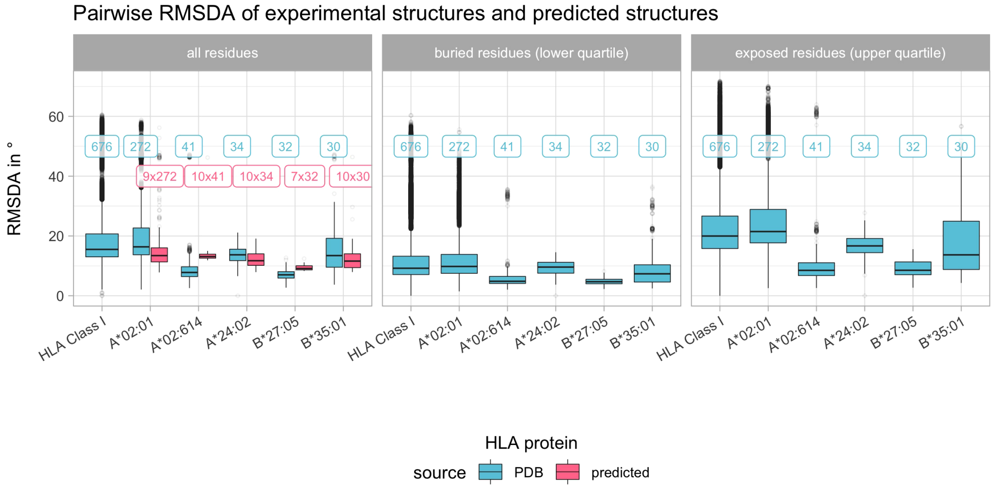


FIGURE 5
Boxplot indicating the pairwise root mean square dihedral angle differences (RMSDA) in degree across all experimental HLA Class I structures and pairwise across the most common HLA Class I structures. Color depicts the source of the structure data (blue indicates experimental structures from the PDB, pink indicates predicted structures as described in Materials and Methods). Horizontal panels consider separate distributions of all, only buried or only exposed residues. Boxplots depict the median (horizontal line), first to third quartile (box); the highest and lowest values within 1.5x IQR (whiskers) and outliers (circles), respectively. Labels refer to the number of structures being compared, with blue labels referring to experimental structures and pink labels referring to predicted structures being compared with experimental structures.

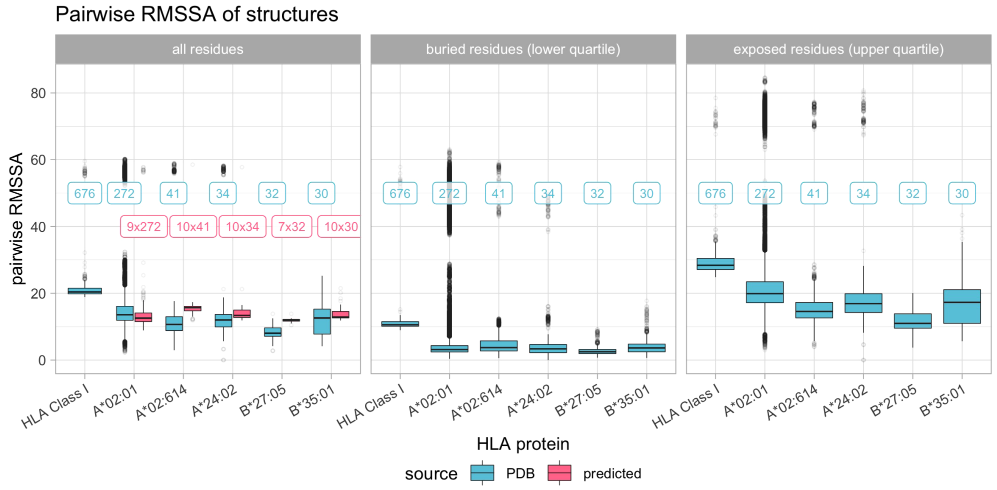


FIGURE 6
Boxplot indicating the pairwise root mean square surface area differences (RMSSA) across all experimental HLA Class I structures and pairwise across the most common HLA Class I structures. Color depicts the source of the structure data (blue indicates experimental structures from the PDB, pink indicates predicted structures as described in Materials and Methods). Horizontal panels consider separate distributions of all, only buried or only exposed residues. Boxplots depict the median (horizontal line), first to third quartile (box); the highest and lowest values within 1.5× IQR (whiskers) and outliers (circles), respectively. Labels refer to the number of structures being compared, with blue labels referring to experimental structures and pink labels referring to predicted structures being compared with experimental structures.

Snowflake estimates P167 being accessible (Figure 9B). In the visual representation accessibility appears plausible. Notably, positions 166 and 167 of A*01:01 have been described by the Eplet 166DG. Although the corresponding amino acid positions of A*02:01 are both surface accessible, no corresponding Eplet 166EW is defined.

For position 110 of A*02:01 the predictors disagree on accessibility; PaleAle and NetSurfP negate solvent accessibility, whilst Snowflake affirm accessibility. By visual representation, accessibility of this position however appears plausible (Figure 9C).

The positions 119 and 120 of A*02:01 appear to be masked by the beta-2-microglobulin structure (Figure 9D). Opposed to Snowflake and PaleAle both agreeing on the inaccessibility of these positions, NetSurfP predicts accessibility. Coloring the residues depending on their predicted solvent accessibility, shows the masking effect of the bound beta-2-microglobulin, rendering hidden amino acid positions, despite being potentially polymorphic, antibody-inaccessible (Figure 10).

HLA Class I antibody-verified eplets are defined as one to three amino acids. At each of respective positions, distinct

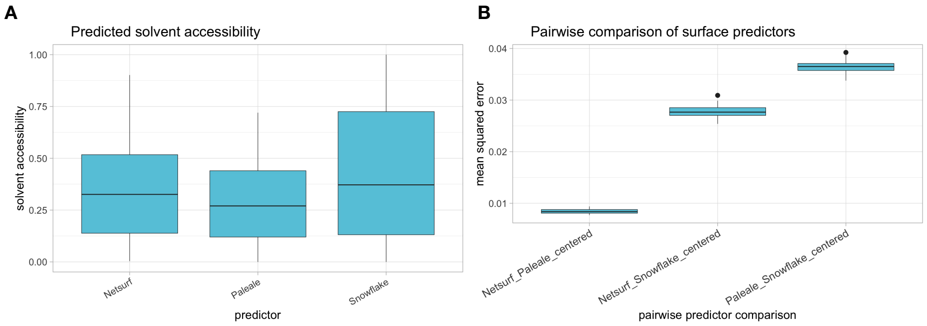


FIGURE 7
(A) Solvent accessibility distribution of reference alleles as predicted by NetSurfP (i.e. “rsa”), PaleAle (i.e. Relative Solvent Accessibility) and Snowflake. **(B)** Pairwise mean squared distance between centered predictions of NetSurfP and PaleAle, NetSurfP and Snowflake, and PaleAle and Snowflake. Boxplots depict the median (horizontal line), first to third quartile (box); the highest and lowest values within 1.5× IQR (whiskers) and outliers (circles), respectively.

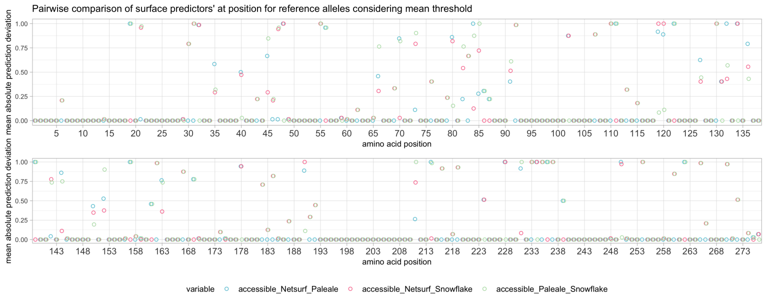


FIGURE 8
Pairwise comparison of predictions considering reference alleles with respective predictor's mean as cutoff for surface accessibility.

distributions of solvent accessibility were observed (Figure 11), leading to different patterns of solvent accessibility: (i) Eplets with consistently high predicted solvent accessibility (e.g. 127K, 193PV), (ii) Eplets with a consistent pattern of solvent accessibility, where at least one amino acid position of the Eplet has high predicted solvent accessibility (e.g. 145KHA, 44RT), and (iii) Eplets with high variance of predicted solvent accessibility at one or more positions (e.g. 144K, 163R, 65RNA). Similar patterns were observed for non-antibody-verified Eplets (Supplementary Figure 7). Compared to antibody-verified Eplets, positions of non-antibody-verified Eplets were more often predicted to be inaccessible.

Snowflake matching was integrated in the previously described PIRCHE Risk and Acceptable Mismatch Profile epitope matching suite (30). The PIRCHE web service allows semi-automated HLA typing data import (GL String, HML, CSV) and provides interfaces to major HLA typing kit vendors' software suites to load HLA typing data quickly and error-free. Interactive bar charts provide interlocus and intralocus Snowflake scores along with PIRCHE scores, allele frequencies,

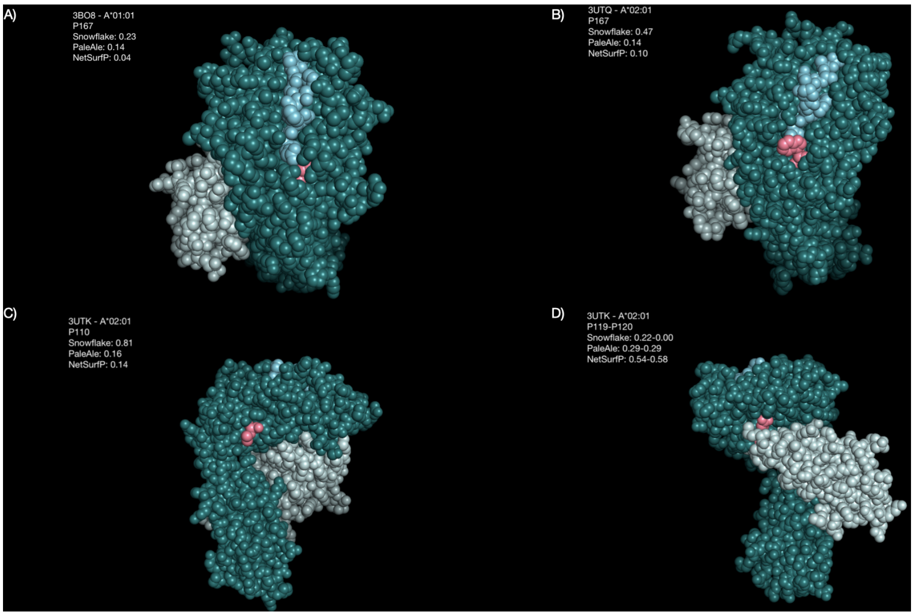


FIGURE 9
Highlighted amino acid positions of selected structures. The HLA alpha chain is colored in dark green, beta-2-microglobulin is colored in light green, the bound peptide is colored in blue, and the highlighted amino acid positions are colored pink. **(A)** Position P167 is predicted by Snowflake and the reference predictors to be buried in A*01:01. **(B)** The same position P167 in A*02:01 is predicted by Snowflake to be accessible, by the reference predictors again predicted to be buried. **(C)** The reference predictors estimate position P110 in A*02:01 to be buried, whilst Snowflake considers it accessible. **(D)** For A*02:01, positions P119 and P120 are contact points with the beta-2-microglobulin. Whilst Snowflake considers these positions not being accessible, NetSurfP considers them as accessible.

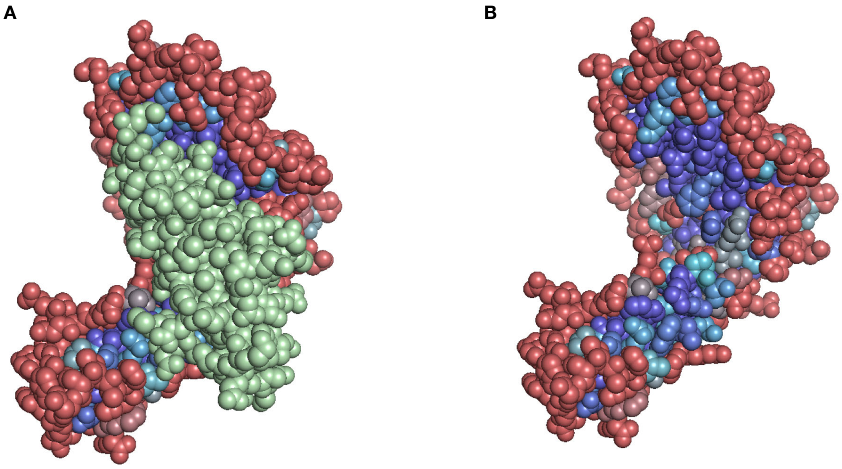


FIGURE 10
(A) Solvent accessibility-colored residues of A*02:01. Beta-2-microglobulin is colored in green. **(B)** Residues masked by the beta-2-microglobulin structure are predicted to be inaccessible. Color (gradient from red to purple to blue) based on Snowflake score, red tones indicating higher solvent accessibility, blue tones indicate lower solvent accessibility.

predicted shared T cell epitopes and mean fluorescence intensities of single antigen bead assays (Figure 12).

Discussion

A wide range of universal software tools characterizing potential B-cell epitopes has been described in the literature (e.g. ElliPro (31), NetSurfP (13), PaleAle (14), DiscoTope (32), BepiPred (33)). Specifically for transplant histocompatibility matching, these tools supported approaches of HLA B-cell epitope definition and matching. HLAMatchmaker (3) (www.epitopes.net) and EPRegistry (4, 5) (www.epregistry.com.br) define a manually curated database of antibody-accessible amino acid configurations (i.e. eplets) based on a limited number of experimental HLA structures. Verifying each of these eplets by antibodies and estimating their immunogenicity however remains challenging (7, 34). The EMS-3D model considers a multidimensional score of electrostatic dissimilarity between donor and recipient HLA, which may be an additional metric to characterize the HLA surface (6).

In the present study, we describe the Snowflake algorithm, a newly developed deep learning-based HLA-matching algorithm that considers allele-specific surface accessibility. Snowflake applies a strategy comparable to the one used by the pHLA3D database (phla3d.com.br) (35). Both approaches complement the experimental structures of the PDB by *in silico* predicted structures of pHLAs using a homology modeling pipeline approach in order to spatially locate amino acid positions in HLA proteins without available structures. Snowflake applies AlphaFold for structure prediction as it appears to outperform the “satisfaction of spatial restraints”-approach of MODELLER, which was used to construct the pHLA3D database (15). As Snowflake considers multiple binding peptides during training, peptide-specific bias in solvent accessibility of binding groove

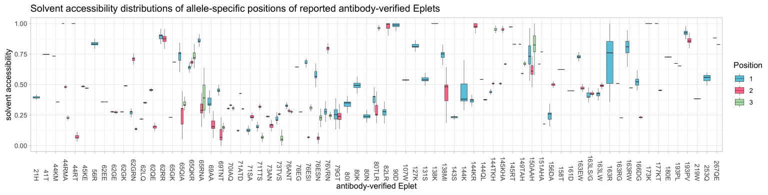
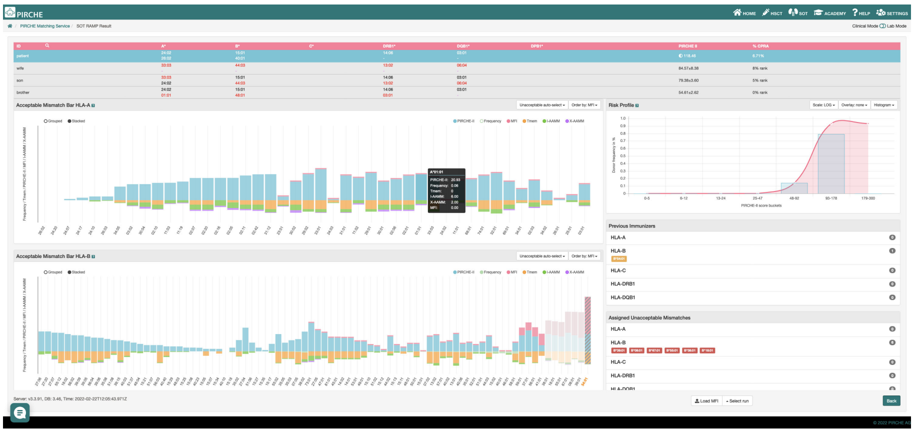


FIGURE 11
Distributions of solvent accessibility of respective antibody-verified Eplets positions. Color corresponds to the amino acid position within the Eplet. Boxplots depict the median (horizontal line), first to third quartile (box); the highest and lowest values within 1.5× IQR (whiskers), respectively.



PIRCHE
HOME HSCT SOT ACADEMY HELP SETTINGS
PIRCHE Matching Service / SOT RAMP Result
Clinical Mode Lab Mode

ID	A*	B*	C*	DRB1*	DQB1*	DPB1*	PIRCHE II	% CPRA
patient	24:02 26:02	15:01 40:01		14:06 -	03:01 -		118.46	6.71%
wife	33:03 -	44:03 -		13:02 -	06:04 -		84.57±8.38	8% rank
son	33:03 24:02	15:01 44:03		14:06 13:02	03:01 06:04		79.38±3.60	5% rank
brother	24:02 01:01	15:01 48:01		14:06 03:01	03:01 -		54.61±2.62	0% rank

Previous Immunizers
HLA-A 0
HLA-B 1
B*54:01
HLA-C 0
HLA-DRB1 0
HLA-DQB1 0

Assigned Unacceptable Mismatches
HLA-A 0
HLA-B 6
B*39:01 B*08:01 B*67:01 B*55:01 B*38:01 B*18:01
HLA-C 0
HLA-DRB1 0
HLA-DQB1 0

Server: v3.2.81, DB: 3.46, Time: 2022-02-22T12:05:43.971Z

FIGURE 12
PIRCHE Risk and Acceptable Mismatch Profile including PIRCHE-II epitope scores (blue bars), mean fluorescence intensities of imported single antigen bead assays (pink bars), shared T-cell epitopes with previous immunizers (orange bars), allele frequencies (pale green bars), interlocus Snowflake scores (purple bars) and intralocus Snowflake scores (dark green bars).

residues is taken into account. Comparing the Snowflake-predicted surface accessibility with the calculated solvent-accessible surface area of pHLA3D database structures confirms a concordance between the approaches for the majority of amino acid positions (Supplementary Figure 8). However, each structure has a number of amino acid positions with large deviation between pHLA3D and Snowflake (i.e. outliers) indicating differences in parts of the predicted structures. Considering protein flexibility and different optimization strategies, these differences may however be technical artifacts of the respective tool chain's characteristics.

To incorporate simulated HLA structures into surface prediction, Snowflake implemented its own HLA surface predictor comparable to both NetSurfP 2.0 and PaleAle 4.0. Similar to Snowflake, these predictors implement a deep learning neural network architecture using BRNN to predict (amongst other characteristics) relative surface accessibility. Both networks have been trained with the experimental structures reported in the PDB (13, 14). Limiting on solvent accessibility as predicted output, Snowflake's deep learning model has been trained exclusively on HLA Class I structures, tailoring its prediction capabilities specifically to HLA Class I proteins. The observed lower numerical distance between predictions of NetSurfP and PaleAle compared to Snowflake and these predictors (Figure 7) may be caused by a different strategy in preprocessing calculated surface area. Considering predicted structures as additional training data is an advantage over NetSurfP and PaleAle, as it incorporates domain-specific knowledge on protein folding into solvent accessibility prediction. It must be considered that learning-based surface prediction is obsolete if the predicted structures have been generated for all known HLA alleles. Proving the predictors' reliability and accuracy requires more experimental structures of HLA proteins that have not been analyzed so far. Given the complexity of X-ray crystallographic experiments, this remains challenging.

Although HLA Class I experimental structures showed little overall structural variance in dihedral angles between alleles (Figure 5), solvent accessibility varied across different alleles, indicating solvent accessibility to be allele-specific (Figure 6). Visual representations of characteristic positions support this hypothesis (Figure 9). Consequently, amino acid matching approaches may be improved by considering solvent accessibility of the respective amino acid of the specific allele, providing a consistent explanation for epitopes' directionality.

Despite strong numeric correlation between Eplets and Snowflake, some difference in epitope definition may arise. The example in Figure 9 shows strong accessibility for positions 166 (glutamic acid) and 167 (tryptophan) of A*02:01, yet there is no Eplet 166EW defined. As described by (3), the interlocus comparison with HLA-B and -C typically shows matched monomorphic configurations at these positions. However, in the meantime there are HLA-B and -C surface-expressible alleles described, which do not share the EW configuration at these positions (data not shown). Consequently, there may be scenarios where a common hypothetical 166EW is not found in the recipient's self-Eplets, potentially provoking an immune response. Similar examples are 253Q without a corresponding common 253E or 131S without a corresponding common 131R. Although these situations are likely rare, this indicates potential gaps in Eplet definitions. Furthermore, verifying Eplets with low population frequency by antibody reaction patterns remains challenging. The herein presented definition of epitope mismatches provides a consistent allele-specific process.

The position-specific solvent accessibility at amino acid positions encoding for antibody-verified Eplets (Figure 11, Supplementary Figure 7) varies for some of the reported Eplets. This variation suggests that the three-dimensional structure of such Eplets is dependent on the complete structure of the allele they are occurring in. This dependency raises the question, if antibody reactions against these Eplets are equally pronounced for alleles with lower accessibility compared to alleles with higher accessibility of the corresponding amino acid positions. Accessibility of non-antibody-verified Eplets generally appeared to be comparatively lower. The lower accessibility may be the cause for lack of antibody verification in the first place, and supports the hypothesis that non-antibody-verifiable internal Eplets act as proxies by impacting the protein surface from within.

HLA epitope-matching algorithms based upon solvent-accessible amino-acid mismatches are valuable tools in antibody-based HLA-mismatch risk classifications. The recently described HLA-EMMA (hla-emma.com) algorithm considers these solvent-accessible amino acid mismatches as potential B-cell receptor targets by defining solvent-accessible amino acid positions per HLA locus. The HLA-EMMA algorithm considers an amino acid position as solvent-accessible, if PaleAle- or NetSurfP-solvent accessibility scores exceed a certain threshold in at least one of the considered alleles of a locus (10). This approach may, however, assign mismatched amino acids of two alleles as epitope mismatch even though the amino acids within the respective alleles aren't surface-accessible (Figures 8, 9A, B). Such misclassifications can result from the fact that accessibility of the involved position was determined by a structural alignment to a different allele. Consequently, by using an allele-specific approach, Snowflake's allele-specific solvent accessibility of mismatched amino-acids may increase specificity.

To confirm the impact of allele-specific solvent accessibility experimentally, binding analyses of HLA epitope-specific monoclonal antibodies to HLA as applied by Bezstarosti et al. may be informative (36). Following the hypothesis, antibodies against antibody-verified Eplets with a wide distribution of surface accessibility (e.g. 138MI or 163R, Figure 11) are expected to yield into a binding pattern proportional to the respective alleles' Eplet-accessibility. Linked mismatched amino acid residues may however interfere with the complete antibody footprint, which has to be considered as a potential confounder of such experiments.

To fully evaluate the predictions of Snowflake, it is necessary to consider potential limitations. Our solvent accessibility prediction considered structural data generated by a protein folding predictor. Such an approach is essential due to limited and biased data of experimental HLA structures. Despite advances in the field of protein folding predictors (37), it must be noted protein folding is one of the most challenging problems in bioinformatics with predictions still lacking precision (38). Our data indicate considerable structural variance between experimental structures of the same allele, which may be explained by the varying specific experimental setups on the one part and by the flexibility of HLA molecules on the other (39, 40). Though predicted solvent accessibility varies to a greater extent across different alleles, protein flexibility and dynamics may still alter the protein surface *in vivo*.

Second, validation of AlphaFold-predicted structures did not account for structures being part of the training set and thus may overestimate the suitability of rendered structures for epitope analyses. Predicted positions of peptides in predicted HLA structures only considered nonameric peptides. The method may be further refined by including peptides of varying lengths depending on length distributions of reported binders, to evaluate the role of peptide length on solvent accessibility in the groove-surrounding region.

Training data for surface prediction is biased towards HLA-A*02:01 (Figure 4). Excluding structures based on sequence similarity as suggested previously (13, 14) would, however, significantly limits the training set. Considering the training data bias being not only sequence-specific but also position-specific, as positions have individual amino acid distributions across HLA alleles, training data aggregation at the allele level is insufficient. Lacking a robust aggregation option and accepting the potential bias, the complete data sets were used for training.

The applied surface area calculation considered a probe radius of 1.4 Å based on the size of water molecules. Given the size of HLA-binding antibodies and the interaction area between HLA and antibody (e.g. in PDB 6ID4 (41)), it may however be considered to restrict antibody accessibility to larger probe sizes or aggregate solvent accessibility with predicted ellipsoid protrusion.

Similar to all previously described HLA epitope matching algorithms, the Snowflake core algorithm also requires protein-level HLA typing data. By applying the previously described haplotype frequency-based multiple imputation (42), Snowflake scores can however be calculated with low- or intermediate resolution HLA typing data. To the best of our knowledge, Snowflake is the first HLA B cell epitope matching algorithm natively supporting low-resolution typing inputs. The Snowflake matching algorithm for batch processing or detailed matching is available at pirche.com.

In summary, we herein presented the Snowflake B-cell epitope matching algorithm, a deep-learning based software evaluating allele-specific surface-accessible amino acid mismatches applicable to all HLA Class I alleles. It refines protein surface-restricted amino acid matching in the context of transplantation and may support analyses on the immunogenicity and antigenicity of HLA B-cell epitopes.

Clinical studies are warranted to indicate correlation between the Snowflake score and immunological events, such as development of donor-specific antibodies.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: RCSB PDB: https://www.rcsb.org, Epitope Registry: https://www.epregistry.com.br, AlphaFold predictor: https://alphafold.ebi.ac.uk, Immune Epitope DataBase: https://www.iedb.org.

Author contributions

MN and ES contributed to conception and design of the study. BM and MN extracted data. MN developed the prediction pipeline and wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This study is supported by the German Federal Ministry for Economic Affairs and Climate Action (grant 01MJ21002B).

Conflict of interest

MN works for PIRCHE AG, which develops and operates the PIRCHE web service. PIRCHE AG and UMC Utrecht have filed a patent application on the prediction of an alloimmune response against allele-specific solvent-accessible amino acid mismatches. MN and ES are listed as inventors on this patent.

The remaining author 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

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.937587/full#supplementary-material

References

1. Süsal C, Opelz G. Current role of human leukocyte antigen matching in kidney transplantation. *Curr Opin Organ Transplant* (2013) 18(4):438–44. doi: 10.1097/MOT.0b013e3283636ddf

2. Tambur AR, Claas FHJ. HLA epitopes as viewed by antibodies: what is it all about? *Am J Transplant* (2015) 15(5):1148–54. doi: 10.1111/ajt.13192

3. Duquesnoy RJ. HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. *I Description Algorithm Hum Immunol* (2002) 63(5):339–52. doi: 10.1016/S0198-8859(02)00382-8

4. Duquesnoy RJ. Update of the HLA class I eplet database in the website based registry of antibody-defined HLA epitopes: HLA class I eplet database update. *Tissue Antigens* (2014) 83(6):382–90. doi: 10.1111/tan.12322

5. Duquesnoy RJ, Marrari M, Marroquim MS, Borges AG, da Mata Sousa LCD, Socorro A, et al. Second update of the international registry of HLA epitopes. *I HLA-ABC Epitope Database Hum Immunol* (2019) 80(2):103–6. doi: 10.1016/j.humimm.2018.11.007

6. Mallon DH, Kling C, Robb M, Ellinghaus E, Bradley JA, Taylor CJ, et al. Predicting humoral alloimmunity from differences in donor and recipient HLA surface electrostatic potential. *JI* (2018) 201(12):3780–92. doi: 10.4049/jimmunol.1800683

7. Bezstarosti S, Bakker KH, Kramer CSM, de Fijter JW, Reinders MEJ, Mulder A, et al. A comprehensive evaluation of the antibody-verified status of eplets listed in the HLA epitope registry. *Front Immunol* (2022) 12:800946. doi: 10.3389/fimmu.2021.800946

8. Tan J, Qiu J, Tang X. HLA amino acid residue matching in 2575 kidney transplants. *Transplant Proc* (2007) 39(5):1429–31. doi: 10.1016/j.transproceed.2007.03.063

9. Kosmoliaptsis V, Sharples LD, Chaudhry A, Johnson RJ, Fuggle SV, Halsall DJ, et al. HLA class I amino acid sequence-based matching after interlocus subtraction and long-term outcome after deceased donor kidney transplantation. *Hum Immunol* (2010) 71(9):851–6. doi: 10.1016/j.humimm.2010.06.003

10. Kramer CSM, Koster J, Haasnoot GW, Roelen DL, Claas FHJ, Heidt S. HLA-EMMA: A user-friendly tool to analyse HLA class I and class II compatibility on the amino acid level. *HLA* (2020) 96(1):43–51. doi: 10.1111/tan.13883

11. Berman HM. The protein data bank. *Nucleic Acids Res* (2000) 28(1):235–42. doi: 10.1093/nar/28.1.235

12. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA database. *Nucleic Acids Res* (2019) gkz950:D948–55. doi: 10.1093/nar/gkz950

13. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK, et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins* (2019) 87(6):520–7. doi: 10.1002/prot.25674

14. Mirabello C, Pollastri G. Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics* (2013) 29(16):2056–8. doi: 10.1093/bioinformatics/btt344

15. Jumper J, Evans R, Pritzel A, Green T, Figureurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* (2021) 596 (7873):583–9. doi: 10.1038/s41586-021-03819-2

16. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* (2021) 50 (D1):gkab1061. doi: 10.1093/nar/gkab1061

17. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-multimer. *Bioinformatics* (2021). doi: 10.1101/2021.10.04.463034

18. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinf* (2009) 10(1):421. doi: 10.1186/1471-2105-10-421

19. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* (2009) 25(11):1422–3. doi: 10.1093/bioinformatics/btp163

20. Golub GH, Van Loan CF. *Matrix computations. fourth edition*. Baltimore: The Johns Hopkins University Press (2013). 756 p. (Johns Hopkins studies in the mathematical sciences).

21. 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(D1): D339–43. doi: 10.1093/nar/gky1006

22. Abella J, Antunes D, Clementi C, Kavraki L. APE-gen: A fast method for generating ensembles of bound peptide-MHC conformations. *Molecules* (2019) 24 (5):881. doi: 10.3390/molecules24050881

23. Kufareva I, Abagyan R. Methods of protein structure comparison. In: AJW Orry and R Abagyan, editors. *Homology modeling*, vol. vol. 857 . Totowa, NJ: Humana Press (2011). p. 231–57. doi: 10.1007/978-1-61779-588-6_10

24. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *J Mol Biol* (1973) 79(2):351–71. doi: 10.1016/0022-2836(73)90011-9

25. Kunzmann P, Hamacher K. Biotite: a unifying open source computational biology framework in Python. *BMC Bioinf* (2018) 19(1):346. doi: 10.1186/s12859-018-2367-z

26. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Trans Signal Process* (1997) 45(11):2673–81. doi: 10.1109/78.650093

27. Kingma DP, Ba J. *Adam: A method for stochastic optimization* (2017). Available at: http://arxiv.org/abs/1412.6980.

28. Matern BM, Mack SJ, Osoegawa K, Maiers M, Niemann M, Robinson J, et al. Standard reference sequences for submission of HLA genotyping for the 18th international HLA and immunogenetics workshop. *HLA* (2021) 97(6):512–9. doi: 10.1111/tan.14259

29. Garrett TPJ, Saper MA, Bjorkman PJ, Strominger JL, Wiley DC. Specificity pockets for the side chains of peptide antigens in HLA-Aw68. *Nature* (1989) 342 (6250):692–6. doi: 10.1038/342692a0

30. Geneugelijk K, Spierings E. Matching donor and recipient based on predicted indirectly recognizable human leucocyte antigen epitopes. *Int J Immunogenet* (2018) 45(2):41–53. doi: 10.1111/iji.12359

31. Ponomarenko J, Bui HH, Li W, Fusseder N, Bourne PE, Sette A, et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinf* (2008) 9(1):514. doi: 10.1186/1471-2105-9-514

32. Kringelum JV, Lundegaard C, Lund O, Nielsen M. Reliable b cell epitope predictions: Impacts of method development and improved benchmarking. peters b, editor. *PloS Comput Biol* (2012) 8(12):e1002829. doi: 10.1371/journal.pcbi.1002829

33. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based b-cell epitope prediction using conformational epitopes. *Nucleic Acids Res* (2017) 45(W1):W24–9. doi: 10.1093/nar/gkx346

34. Mohammadhassanzadeh H, Oualkacha K, Zhang W, Klement W, Bourdiec A, Lamsatfi J, et al. On path to informing hierarchy of eplet mismatches as determinants of kidney transplant loss. *Kidney Int Rep* (2021) 6(6):1567–79. doi: 10.1016/j.ekir.2021.03.877

35. Menezes Teles e Oliveira D, Brandão RMSd, Sousa LCDd, Lima FdCA, do Monte SJH, Marroquim MSC, et al. pHLA3D: An online database of predicted three-dimensional structures of HLA molecules. *Hum Immunol* (2019) 80 (10):834–41. doi: 10.1016/j.humimm.2019.06.009

36. Bezstarosti S, Kramer CSM, Franke-van Dijk MEI, Vergunst M, Bakker KH, Uyar-Mercankaya M, et al. HLA-DQ-Specific recombinant human monoclonal antibodies allow for in-depth analysis of HLA-DQ epitopes. *Front Immunol* (2022) 12:761893. doi: 10.3389/fimmu.2021.761893

37. Alexander LT, Lepore R, Kryshtafovych A, Adamopoulos A, Alahuhta M, Arvin AM, et al. Target highlights in CASP14 : Analysis of models by structure providers. *Proteins* (2021) 89(12):1647–72. doi: 10.1002/prot.26247

38. Unger R, Moult J. Finding the lowest free energy conformation of a protein is an NP-hard problem: Proof and implications. *Bltn Mathcal Biol* (1993) 55 (6):1183–98. doi: 10.1016/S0092-8240(05)80169-7

39. Wieczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S, Noé F, et al. Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. *Front Immunol* (2017) 8:292/full. doi: 10.3389/fimmu.2017.00292/full

40. Natarajan K, Jiang J, May NA, Mage MG, Boyd LF, McShan AC, et al. The role of molecular flexibility in antigen presentation and T cell receptor-mediated signaling. *Front Immunol* (2018) 9:1657. doi: 10.3389/fimmu.2018.01657

41. Gu Y, Wong YH, Liew CW, Chan CEZ, Murali TM, Yap J, et al. Defining the structural basis for human alloantibody binding to human leukocyte antigen allele HLA-A*11:01. *Nat Commun* (2019) 10(1):893. doi: 10.1038/s41467-019-08790-1

42. Geneugelijk K, Wissing J, Koppenaal D, Niemann M, Spierings E. Computational approaches to facilitate epitope-based HLA matching in solid organ transplantation. *J Immunol Res* (2017) 2017:1–9. doi: 10.1155/2017/9130879