# Data science approach for the exploration of HLA antigenicity based on 3D structures and molecular dynamics

Diego Amaya-Ramirez

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

# Data science approach for the exploration of HLA antigenicity based on 3D structures and molecular dynamics

## THÈSE

Présentée et soutenue publiquement le 10 juillet 2024
pour l'obtention du

**Doctorat de l'Université de Lorraine**
**(mention informatique)**

par

Diego Alfredo Amaya Ramirez

**Composition du jury :**

| | | |
|---|---|---|
| *Président :* | Olivier Toutirais | Professeur, Université de Caen |
| *Rapporteurs :* | Frédéric Cazals | DR INRIA, Centre Sophia Antipolis–Méditerranée |
| | Juliette Martin | DR CNRS, ENS Lyon |
| *Examinateurs :* | Manuel Dauchez | Professeur, Université de Reims Champagne-Ardenne |
| | Malika Smaïl-Tabbone | MC HDR, Université de Lorraine |
| *Encadrants:* | Marie-Dominique Devignes | CR CNRS (HDR), LORIA Nancy |
| | Jean-Luc Taupin | Professeur, Université Paris-Cité |

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

# Acknowledgements

As I conclude this transformative chapter of my life, I find myself reflecting on the incredible support I've received from the beginning to the present day. This PhD journey has been shaped by the contributions of many, and I wish to express my profound gratitude to each person who has played a part in this achievement.

First and foremost, I extend my sincerest gratitude to my thesis supervisors, Marie-Dominique Devignes and Jean-Luc Taupin, for giving me the chance to work under their guidance. The knowledge I have acquired and the valuable insights I gained from both of you have been invaluable. Your unwavering support, thoughtful guidance, and constant encouragement have been crucial in helping me navigate this challenging journey. Thank you both for playing a key role in my personal and professional development.

I would also like to express my appreciation to the CAPSID team. Your camaraderie, intellectual collaboration, and shared experiences have created a stimulating and supportive research environment. I am especially grateful to Antoine Moniot, Dominique Mias-Lucquin and Athénaïs Vaginay for their intellectual collaborations and shared experiences. A warm thank you also goes to Malika Smail-Tabbone and Bernard Maigret, your diverse insights and our stimulating discussions have made this experience both enlightening and enjoyable. I'm truly fortunate to have been part of such an exceptional team.

I would also like to express my appreciation to Magali, Cedric, Romain, and Constantin from Jean-Luc Taupin's team at Saint Louis Hospital for their invaluable help throughout this project.

The Opailleur team also deserves recognition for the convivial lunches and moments we shared. A special mention goes to Laura and Gui - your friendship and support have been a constant source of encouragement.

My sincere thanks to the jury members - Prof. Olivier Toutirais, Dr. Frédéric Cazals, Dr. Juliette Martin, and Prof. Manuel Dauchez - for accepting the invitation to evaluate my work.

I'm grateful to the engineering team at LORIA, particularly Patrice Ringot, for their technical support with grid5000. I also extend my thanks to the laboratory staff, from reception to canteen, especially Isabelle, Florianne, and Caro, whose support and assistance behind the scenes were greatly appreciated.

Lastly, but most importantly, I owe an immense debt of gratitude to my family, especially Viviana and Kaoru, your unconditional love and support have been my foundation throughout this journey. Without you, this achievement would not have been possible.

To everyone who lent a hand, whether directly or indirectly, in helping me achieve this milestone—thank you.

# Table of contents

# Introduction

## I. Thesis context

Organ transplantation is the unique option to treat patients suffering from end-stage organ failure. In France, in 2022, 5 493 organs were transplanted, mostly kidneys (n=3 376), then livers (n=1 294), hearts (n=411), lungs (n=334), pancreases (n=70), and cardiopulmonary (n=8), with ~11,000 patients on the waiting list. However, the duration of function for transplanted organs remains limited to a median of around 12, 15, 8, and 11 years for kidney, liver, lung, and heart, respectively. The majority of transplants are lost due to the recipient's alloimmune response directed to donor-specific antigens recognized as non-self because they are absent from the recipient, and essentially to the humoral response, leading to the production of donor-specific antibodies (DSA). Indeed, the cell-mediated response is far better controlled by the mandatory immunosuppressive drug regimen that accompanies the patient throughout the graft's life.

The main antigenic system triggering DSA is the Human Leukocyte Antigen (HLA) system, with its 11 major genes codominantly expressed by every individual. The HLA system was first deciphered between the 1950s and 1980s, pioneered by Nobel Prize laureate Professor Jean Dausset at Saint-Louis Hospital in Paris, using antigen/antibody techniques as molecular biology of the gene did not exist yet, leading to an official list of around 120 antigens. DNA molecular techniques have since revealed an extreme genetic polymorphism, currently reaching ~37 000 distinct alleles leading to about 20 000 proteins differing by at least one amino acid (AA).

Nevertheless, compatibility definition still relies on the historical serological classification of HLA, which is not adapted to the immune system's finesse capable of discriminating differences of as little as one AA between two antigens, provided that this difference can be presented to T-lymphocytes or is accessible at the antigen surface for the antibody produced by B-lymphocytes. Locally, each transplant team/HLA lab uses antigen-level information on the recipient's immunization, generated from the analysis of nearly 200 common HLA antigens in a "Single Antigen Beads" (SAB) anti-HLA antibody assay, but no interpretation rules are validated. Indeed, the cartography of AA differences that have a real immunological impact is not known: the HLA Eplet Registry[1] simply lists 560 potential polymorphisms called eplets, deduced from allele sequence alignments.

A simple approach involving eplets in the clinical understanding of the alloimmune humoral response is to measure the donor eplet load provided to the recipient, also called eplet mismatch load (MML). Many articles indeed show that the higher the eplet MML, the higher the risk of developing DSA or undergoing acute or chronic rejection, or even of losing the graft. However, the eplet MML simply means that among the load, at least one has a high risk of triggering DSA emergence, and the higher the load, the higher this risk or the higher the number of eplets bearing this risk. Therefore, this must only be the initial step toward a better understanding of the essence of antigenicity/immunogenicity and eplet/epitope/antibody relationships. Especially, besides this qualitative aspect (presence/absence of an eplet), a quantitative one is desirable linked to each eplet's own stimulation strength, which is constrained by the recipient's HLA antigens, as an antibody will not only interact with the eplet but also with the surrounding area, both parts constituting the epitope. There are indeed clues suggesting that within the large panel of very close allelic variants, promiscuity between self and non-self can de facto limit antibody affinity for the non-self eplet through tolerance to the surrounding self. This dilemma currently renders prediction of antibody clinical toxicity for the transplant very hard to determine from the mere knowledge of donor/recipient (D/R) HLA groups. In the era of personalized medicine, compatibility rules need to be adapted to match the real-life requirements of the immune system in each particular D/R pair. A lead to follow for this objective is to address the problem of antigenicity from the point of view of *in-silico* determination of structural

[1] HLA Eplet Registry is a website-based database that gathers data on eplets reported in the scientific literature

differences (i.e. what would generate eplets) between recipient and donor HLA antigens. The above is with a view to improving the definition of D/R compatibility to help allocate a transplant to a recipient with a better chance to not or poorly respond to donor mismatches. This would increase organ duration of function and delay retransplantation, representing major progress in a permanent context of organ and donor shortage.

## II. Thesis aims and contributions

The present thesis tackles the problem of defining antigenicity in the HLA system in the context of organ transplantation from a structural point of view using data science and machine learning methods. To address this challenge, an unprecedented set of high-quality 3D structures and molecular dynamics (MD) simulations was generated for 207 HLA antigens and analyzed to better understand the differences in structure and antigenicity between HLA antigens. After a first chapter describing the necessary notions and state of the art, this thesis presents the results of this analysis as four main contributions.

- Chapter 2: Contribution of MD simulations to the exploration of structural properties of HLA antigens. This exploration concerns properties such as solvent accessibility, physico-chemical properties and side-chain flexibility explored at AA and patch level distinguishing between different types of AA positions such as, on the one hand, conserved or polymorphic residues and, on the other hand, positions belonging to confirmed or non-confirmed eplets or positions non registered as eplets. Dynamic analysis was performed on MD trajectories and the results were compared to static PDB structure analysis. These comparisons revealed an important and relevant contribution of MD simulations in such structural analyses. Moreover, comparison of flexibility coefficients (N-RMSF) calculated from MD simulations for amino acids belonging to confirmed eplets and those not belonging to any eplet showed a significantly reduced flexibility for the former ones in good accordance with recent reports about epitope flexibility. This comparison was presented as a poster at the international conference ECCB 2022 in Sitges (Spain). https://inria.hal.science/hal-03924018/document.

- Chapter 3: 3D shape descriptor for dynamic comparison of protein surfaces. Zernike polynomials are a mathematical tool for representing and analyzing 3D shapes. They provide a compact, rotation- and translation-invariant descriptor of a protein shape (i.e., a series of coefficients) that facilitates tasks such as protein structure comparison and classification. The working hypothesis here was that HLA antigenicity is dependent on the structural similarity of HLA antigens between donor and recipient. Indeed, if the donor HLA epitopes are structurally similar (although not identical) to the recipient ones, they will not be recognized as antigenic by the recipient's immune system. Zernike polynomials were applied to protein surface patches representing HLA epitopes and have led to model MD trajectories of epitopes as clouds of multi-dimensional points. Then, the Wasserstein distance metrics was used to compare two such clouds of points between recipient and donor HLA antigens, in order to distinguish between structurally close and distant pairs of epitopes. The possibility to use Zernike polynomials for clustering the frames of an MD trajectory was shown on a peptide and presented as a poster at the 2021 joint workshop of the GGMM (Groupe de Graphisme et Modélisation Moléculaire) and SFCI (Société Française de Chémoinformatique) in Lille. https://hal.science/hal-04578212].

- Chapter 4: Training of a HLA epitope predictor on HLA antigens (HLA-EpiCheck). Leveraging both dynamic and static structural features such as solvent accessibility, side-chain flexibility, hydrophobicity and charge, a Machine Learning (ML) dataset consisting of 18 descriptors computed on 6886 surface patches derived from 207 HLA antigens was constructed. The "ground truth" was obtained from information on confirmed eplets contained in the HLA Eplet Registry. The ML dataset allowed the training of a high-performance HLA epitope predictor called HLA-EpiCheck, using the ExtraTrees algorithm. HLA-EpiCheck was used to predict the antigenicity of a subset of non-confirmed eplets. The results obtained were compared with experimental data

for eplet confirmation, obtained at the Immunology-Histocompatibility laboratory of Hôpital Saint-Louis. This comparison revealed a remarkable consistency between prediction and experiments. This work is currently under evaluation for publication in Bioinformatics Advances.

- Chapter 5: Implementation of an HLA 3D database and a user-friendly graphical tool called HLA-3D-Diff. This tool is intended to facilitate visualization and superposition of 3D HLA antigens to highlight their structural differences along MD simulations, and can be used as an aid to understand complex or unexpected immunization patterns. Source code and information for installing a local image of HLA-3D-Diff is available from the following gitlab repository: https://gitlab.inria.fr/capsid.public_codes/hla-3d-diff_public.

In conclusion, the present thesis provides, thanks to diversified and machine-learning oriented data science approaches, a deep insight into the 3D structural features of HLA antigens and their contribution to HLA antigenicity. Future work will involve clinical studies, retrospective and prospective, to evaluate the inclusion of these 3D structural elements in clinical practices of donor-recipient matching for transplantation as well as their contribution to graft retention or rejection.

# Chapter 1

# Context and state of the art

## Summary



## 1.1 Basics of transplantation and HLA system

### 1.1.1 Overview of transplantation

Organ transplantation is a medical procedure used to treat pathologies causing vital organ failure. This treatment is typically used when the patient's prognosis is at risk without additional medical intervention. Despite advances in healthcare and disease management, the number of transplanted patients is steadily increasing. This is mainly due to factors such as a growing and aging population, the increasing prevalence of chronic diseases, and the broadening of transplantation indications as a result of medical research (see Figure 1).

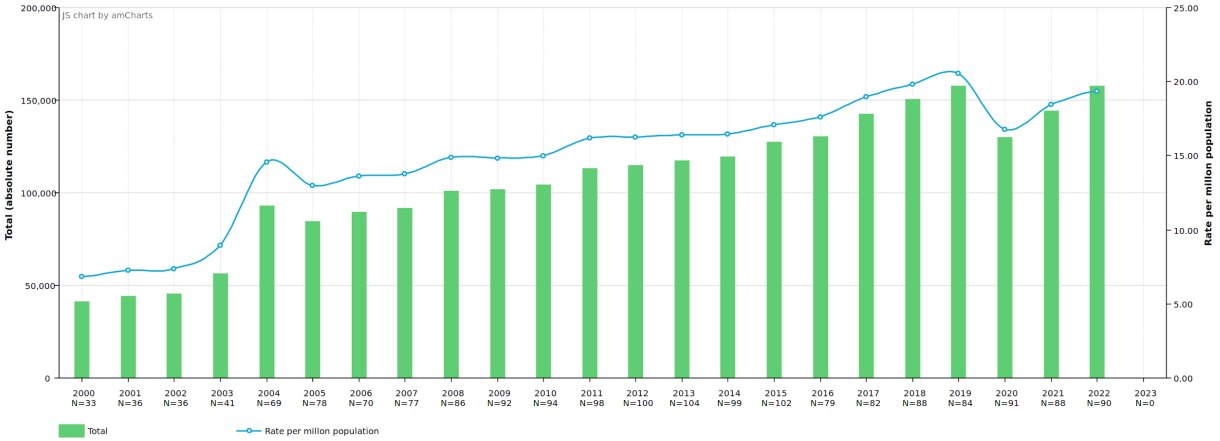



Figure 1: Annual transplant sum and rate per million population in the world (Kidney, Heart, Lungs, Liver, Pancreas, Small Bowel) from 2000 to 2022. Data of the WHO-ONT Global Observatory on Donation and Transplantation [1].

In France, as worldwide, there is a growing trend in the number of patients receiving transplants (see Figure 2). However, the number of patients on the active waiting list (i.e. those immediately eligible for an organ transplant) far exceeds the number of transplants performed. For instance, on January 1st 2023, there were 10 810 patients on the waiting list, while only 5 493 transplants were performed in 2022 [2, 3]. Therefore, it is crucial to optimize transplant allocation to reduce graft rejection.

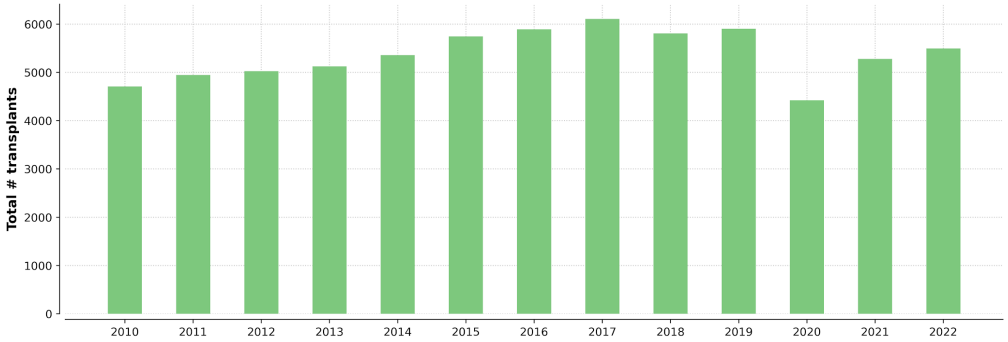



Figure 2: Annual transplant sum in France (Kidney, Heart, Lungs, Liver, Pancreas, Bowel) from 2010 to 2022. Data of the « Agence de la Biomédecine » [4].

Organ transplants can be classified into solid organ transplants and hematopoietic stem cell (HSC) transplants. Solid organ transplants include kidney, liver, intestines, heart, lung and pancreas. HSC transplantation is considered separately from solid organ transplantation because this type of grafting has several unique features that are not encountered with solid organ transplantation.

Kidney transplantation is by far the most common, with 3 376 transplants in 2022. It is the treatment of choice for chronic end-stage renal disease, avoiding the need for dialysis sessions several times a week. The main causes of chronic renal failure are chronic glomerulonephritis and diabetes. Liver transplantation comes second with 1 294 transplants in 2022. Hepatocellular carcinoma has been the main indication for liver transplants followed by alcoholic cirrhosis, retransplants, and fulminant hepatitis. Heart and lung transplants are much less frequent with 411 heart transplants, 334 lung transplants and 8 cardiopulmonary transplants performed in 2022. The primary indications for heart transplants were dilated cardiomyopathy and ischaemic heart disease. The main indications for lung transplants were emphysema and chronic obstructive pulmonary disease, pulmonary fibrosis, and cystic fibrosis. Finally, double heart-lung transplants mainly involve patients with pulmonary arterial hypertension. Pancreas transplants are much rarer with only 70 performed in 2022. The majority of these transplants are combined kidney-pancreas transplants and are performed on patients with type 1 diabetes. There is a high

rate of removal from the waiting list for a pancreas graft and transfer to the waiting list for an isolated kidney transplant [2, 3].

In France, the « Agence de la Biomédecine » manages the distribution and allocation of transplants according to the rules set by the Ministry of Health. The rules aim to ensure fair distribution while finding the best recipient. Technical constraints in organ procurement and transportation, graft viability, and tissue compatibility are considered in the rules. Priority is given to children, recipients with life-threatening conditions, and those with low probability of obtaining a graft due to specific morphological or immunological characteristics. If no priority recipient is identified, allocation is made at successive geographical levels: local, regional, and national. Each organ has its own set of allocation rules. Efforts are made to minimize the distance the graft has to travel in order to reduce the time between removal and transplantation. This is done to preserve transplant quality and maximize transplant success. If a transplant is not suitable for a recipient waiting in France, a very rare situation, it is offered to the « Agence de la Biomédecine »'s European counterparts. From an immunological point of view, the human leukocyte antigen (HLA) system, the human version of the Major Histocompatibility Complex (MHC) found in vertebrates, is at the heart of donor-recipient compatibility. Consequently, the following sections will provide a comprehensive overview of this system and explain the basis of donor-recipient matching.

## 1.1.2 General principles of transplantation immunology

### 1.1.2.1 Basics concepts

| Concept | Definition |
|---|---|
| Antigen | An antigen is a molecule that is recognized by the immune system as foreign or non-self. It triggers an immune response, leading to the production of antibodies or the activation of immune cells |
| Antibody | An antibody is the exclusive property that an immunoglobulin has to be able to recognize and bind to an antigen. Immunoglobulins are proteins produced by B cells in response to the presence of antigens. Antibodies mark them for destruction or neutralization by other components of the immune system. B cells are the cells of the humoral immune response. |
| Autologous graft | A graft transplanted from one individual to the same individual |
| Syngeneic graft | A graft transplanted between two genetically identical individuals |
| Allogeneic graft (allograft) | A graft transplanted between two genetically different individuals of the same species |
| Xenogeneic graft (xenograft) | A graft transplanted between individuals of different species |
| Alloantigens | The molecules that are recognized as foreign in allografts |
| Xenoantigens | The molecules that are recognized as foreign in xenografts |
| Alloreactivity | Reactivity of lymphocytes or antibodies against alloantigens |
| Xenoreactivity | Reactivity of lymphocytes or antibodies against xenoantigens |

The current work falls within the scope of allografts, alloantigens and alloreactivity.

### 1.1.2.2 Introduction to transplantation immunology

Graft compatibility in humans is mainly governed by the genes and antigens of the ABO erythrocyte system and the HLA (Human Leukocyte Antigen) leukocyte system (also known as HLA antigens), which are inducers of alloreactivity. The ABO system was discovered and understood in the first decades of the XXth century [5, 6] and is defined by the presence or absence of certain oligosaccharides in some glycoproteins and glycolipids present on erythrocyte membranes [7–9]. The blood transfusion rules put in place at that time are still valid and strict adhesion to them is required in order to avoid serious immunological events caused by anti-ABO antibodies (Ab) when they are present in the blood of the transfused person. For example, individuals of group A all have anti-B antibodies and therefore can only accept blood from groups A and O (= no ABO antigens) but not from groups B and AB. These rules also apply to organ transplantation because these antigens are not only present on erythrocytes, but also on endothelial cells, which are the cells that cover blood vessels and are in direct contact with the blood. However, when ABO compatibility is respected, transplants can still be lost because of incompatibilities between donor and recipient on HLA antigens, which are ubiquitous but are not expressed on erythrocytes [10].

Nowadays, the majority of graft losses are due to anti-HLA Ab generated by the recipient against one or more HLA antigens of the donor and called DSA (Donor-Specific Antibody). They are responsible for hyperacute (immediate post-transplant), acute (early or late but rapidly evolving) or chronic (most frequent, progressive and slow evolving) rejection, depending on the kinetics of appearance of DSA after the transplant (high rate, rapidly increasing rate, or slowly increasing rate, respectively), when they are not already present before the transplantation (at any level, from very weak to very high). More precisely, hyperacute rejection manifests within minutes or hours of transplantation. It arises from the presence of pre-existing DSAs at a very high level in the recipient's blood. These antibodies trigger a cascade of inflammation and ischemia[2] that rapidly damages the graft, turning it non-functional. To avoid this life-threatening scenario, pre-transplant plasmapheresis is sometimes employed. This procedure involves circulating the recipient's blood through a device that removes DSAs, effectively reducing antibody levels and minimizing the risk of hyperacute rejection. Ideally, recipients with DSAs should avoid receiving organs from donors against whom their antibodies react. However, urgent organ needs and limited donor availability may necessitate transplantation despite the presence of DSAs. In these cases, strict immunosuppression is crucial to control the immune response and minimize the risk of rejection. Acute rejection typically occurs within the first weeks to several months post-transplantation. It can also manifest later but always progresses rapidly, causing a rapid decline in graft function. The primary cause is a transient interruption in immunosuppressive treatment intake by the patient, allowing the immune system to mount an attack against the transplanted organ. Acute rejection can be further categorized into two subtypes: acute cellular rejection (ACR) and acute humoral rejection (AHR). In ACR, activated T cells release cytokines that attract inflammatory cells (such as other T lymphocytes, macrophages, NK cells, polymorphonuclear cells, etc.), ultimately leading to graft tissue necrosis, and also help B cells to produce DSAs. In AHR, DSAs directly bind to and destroy graft cells via antibody-dependent cellular cytotoxicity (involving NK cells), antibody-dependent inflammation (macrophages, polymorphonuclear cells) or classical complement activation. Fortunately, acute rejection is often manageable with immunosuppressive therapies. Post-transplant plasmapheresis can also be employed specifically to target AHR. Chronic rejection emerges in months to years following transplantation and represents the primary cause of long-term graft loss. Unlike acute rejection, chronic rejection progresses slowly, initiating irreversible changes in the transplanted organ long before any noticeable functional decline. Chronic rejection involves a combination of alloantibodies and cell-mediated responses, with a slowly increasing serum level of one or several antibodies often observed [11, 12].

[2] Ischemia happens when a tissue doesn't receive enough blood flow, which reduces the oxygen supply to that tissue.

The humoral response is therefore particularly important and is in fact poorly controlled by immunosuppressive treatments. Approximately 20% of recipients develop at least one *de novo* DSA after 3 years after transplantation, and the incidence then increases with time [13].

HLA antigens are much more complex than the 2 ABO antigens (i.e. A and B, O designing the absence of both, and AB the presence of both), which are actually small 6-carbon sugar radicals carried by surface proteins. Each sugar characterizes a group in the system, and this sugar represents the direct target recognized by the Ab. The HLA system has been known for less time, it was discovered in the early 1950s by Pr. Dausset, Nobel Prize in Medicine 1980, at the Saint-Louis Hospital, and is much more complex: HLA genes are located on chromosome 6p21.31 and are grouped into class I and class II loci. Genes HLA-A, -B and -C form the class I loci while genes HLA-DRA, -DRB1, -DRB3/4/5, -DQA1, -DQA2, -DQB1, -DPA1 and -DPB1 form the class II loci. There are others that are not analyzed for transplantation purposes (-E, -F and -G in class I and DOA). It is worth noting that genes HLA-DRB3, -DRB4, and -DRB5 are paralogues of DRB1, and are believed to have originated from a gene duplication event in the past [14]. An individual can have up to two HLA-DRB3/4/5 alleles, although it is also possible to have only one or none at all. HLA genes are expressed codominantly in each individual. This means that for a given HLA gene, an individual expresses the alleles inherited from both parents. This maximizes the number of HLA molecules available for the immune response to successfully play its role (the function of HLA as a presenter of peptides to T cells will be further explained in subsequent sections). But in the case of organ transplantation, it involves a larger number of potential incompatibilities between donor and recipient. Even worse, some of these genes are highly polymorphic (notably HLA-A, -B, -C, -DRB1, -DRB3/4/5, -DQA1 and -DQB1), which means that these genes have a huge number of alleles within a population. For instance, in clinical practice, about 200 different HLA antigens are tested in anti-HLA antibody detection and identification assays (the most frequent in the population) but the number of potential genetic variants is as high as 37 516 (26 341 for class I and 11 175 for class II alleles) encoding 21 611 HLA proteins which differ from each other by at least one amino acid[3] (AA). This extreme polymorphism of the HLA system means that most organ transplants are performed despite multiple differences between the sequences of donor and recipient HLA molecules, known as 'mismatches'. Indeed, these mismatches are the targets of DSA and are associated with an increased risk of graft rejection [16]. Therefore, the goal in transplant matching is to identify donor-recipient pairs that minimize the number of mismatches and are as similar as possible to the recipient's self antigens in order to reduce the risk of graft rejection. However, determining the relative contribution of each gene and establishing an appropriate metric to assess the severity of a mismatch remains an unresolved challenge.

Detection of anti-HLA antibodies in an individual's serum indicates a state of HLA immunization. This immunization usually arises from exposure to non-self HLA molecules. Transplantation, pregnancy and transfusion represent the three most potent triggers for HLA immunization [17]. Additionally, cross-reactivity with certain environmental non-HLA antigens can, in some cases, contribute to this immune response [18].

Although humoral-mediated immunity (antibodies/B lymphocytes) is the focus of attention when assessing graft compatibility, it is important to note that cell-mediated immunity (T lymphocytes) also plays a role in the immune response against grafts. In fact, humoral and cell-mediated immunity work together and are closely intertwined. For example, CD4+ T lymphocytes are responsible for stimulating antibody production by B lymphocytes (also known as B cells) and inducing their differentiation into memory B cells. However, cell-mediated immunity receives less attention due to the availability of highly effective immunosuppressive treatments against this immune response. The mechanisms of immunization in organ transplantation are as follows: some recipient's T lymphocytes are alloreactive, i.e. they are able to recognize the HLA differences between donor and recipient. This alloreactivity can be direct, i.e. by recognizing donor peptides presented by the donor HLA, or indirect, after donor HLA molecules have been phagocytosed by the recipient's antigen-presenting cells (APCs) and presented as peptides by the recipient's HLA molecules. The activated T lymphocytes stimulate the production of antibodies by B cells,

[3] Data obtained from [15] and accessed on January 13th 2024

which must also be stimulated by direct recognition of the donor HLA molecule by their membrane IgM[4] [10, 19].

Finally, although the HLA system is the main cause of transplant rejection (mainly due to its extreme polymorphism and near-ubiquity in the body's nucleated cells), recent studies have revealed the role of other immunological factors such as the SIRPα (Signal Regulatory Protein alpha) molecule, a receptor on the surface of monocytes. CD47, often referred to as the "don't eat me" signal, is its ligand and is expressed on multiple cells. When CD47 on a cell binds to SIRPα on a phagocytic cell, it transmits a negative regulatory signal that inhibits the phagocytic activity of that phagocytic cell. This is an essential mechanism for self-recognition, as it prevents the body's own cells from being inappropriately phagocytosed by macrophages. If the transplanted cells have a CD47 that does not interact effectively with the recipient's SIRPα, then these transplanted cells may be at increased risk of phagocytosis [20]. Missing self-induced NK cell activation is another immunological aspect that is gaining attention [21] but is much more complex than CD47, as it associates a dozen of genes that are significantly polymorphic.

### 1.1.3 HLA system

#### 1.1.3.1 Nomenclature

After the gene name (A, B, C, etc.) each HLA allele has 4 sets of numbers (usually composed of 2 to 4 digits per set) separated by a colon (see Figure 3). The first set of numbers typically indicate the allele group, which can be determined through serological (i.e. antibody/cell interaction) or lower-resolution DNA typings in urgent situations (such as deceased donors because their organs cannot be stored more than 24 hours). The second set of numbers allows the designation of unique amino acid sequences and is typically the level where one stops when matching HLA alleles in the context of organ transplantation, as the immune system copes with proteins and not the DNA that encodes them. The third set of numbers allows distinguishing between alleles that code for the same amino acid sequence but have different DNA sequences in the exons (coding regions in the DNA); the locations where this type of polymorphism occurs are known as silent mutations. Finally, the fourth set of numbers allows differentiating between alleles that have the same amino acid sequence and the same DNA sequence in the exons, but differ in the introns (non-coding regions in the DNA).

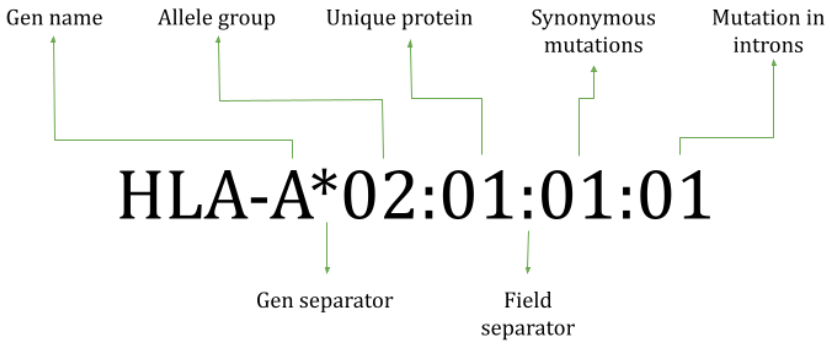



Figure 3: Nomenclature of HLA typings

Serological groups are named either by using the letter of the gene in the case of class I loci, or by the first two letters of the gene in the case of class II loci and then the allele group number, so for example the group corresponding to B*44 is called B44 and the DRB1*13 group would be DR13. However, there are some variations, such as the groups from the C locus[5], where the letter 'w' is inserted between the locus name and the number (e.g. Cw7), mainly to distinguish them from factors of the complement system.

[4] IgM is a type of antibody involved in the early stages of an immune response.
[5] The terms 'gene' and 'locus' (and then 'genes' and 'loci') are used here interchangeably.

### 1.1.3.2 The role of the HLA system in the immune response

The HLA system is essential in an individual's immune response, as it presents peptide antigens to T cells. Two pathways mediated by the HLA system activate the immune system. On one hand, the class I HLA proteins mediate a pathway present in almost all nucleated cells of the organism and whose function is to present antigens derived from the intracellular environment, i.e. antigens derived from a viral infection or a cancerous process. These molecules are solely responsible for activating CD8+ lymphocytes, which have the ability to induce the death of infected/cancerous cells. On the other hand, there is a pathway mediated by HLA proteins called class II, which are only present on "Antigen Presenting Cells" (APCs) such as mainly dendritic cells, but also B cells, macrophages, and others, and their function is to display antigens derived from the extracellular environment. Class II molecules are solely responsible for activating CD4+ helper lymphocytes, which in turn activate macrophages or B cells [10].

### 1.1.3.3 The HLA molecules

Each HLA molecule consists of extracellular, transmembrane and cytoplasmic regions. Class I HLA molecules (Figure 4A and Figure 4B) are composed of two-chain heterodimers, one chain encoded by one the HLA gene (called chain $\alpha$) and a second one encoded by a non-HLA gene (β2 - microglobulin). Class II HLA molecules (Figure 4C and Figure 4D) are composed of two-chain heterodimers, both chains encoded by the HLA genes (called $\alpha$ and β chains) and both being polymorphic. Within the extracellular region of the HLA molecules, a cleft (also known as "peptide pocket") is formed close to the amino termini of the HLA-encoded proteins. It is composed of two α-helices forming the walls of the cleft and about eight antiparallel β-strands forming the floor. The peptide antigen settles in this cleft and this HLA-peptide complex is the portion recognized by a T Cell Receptor (TCR).

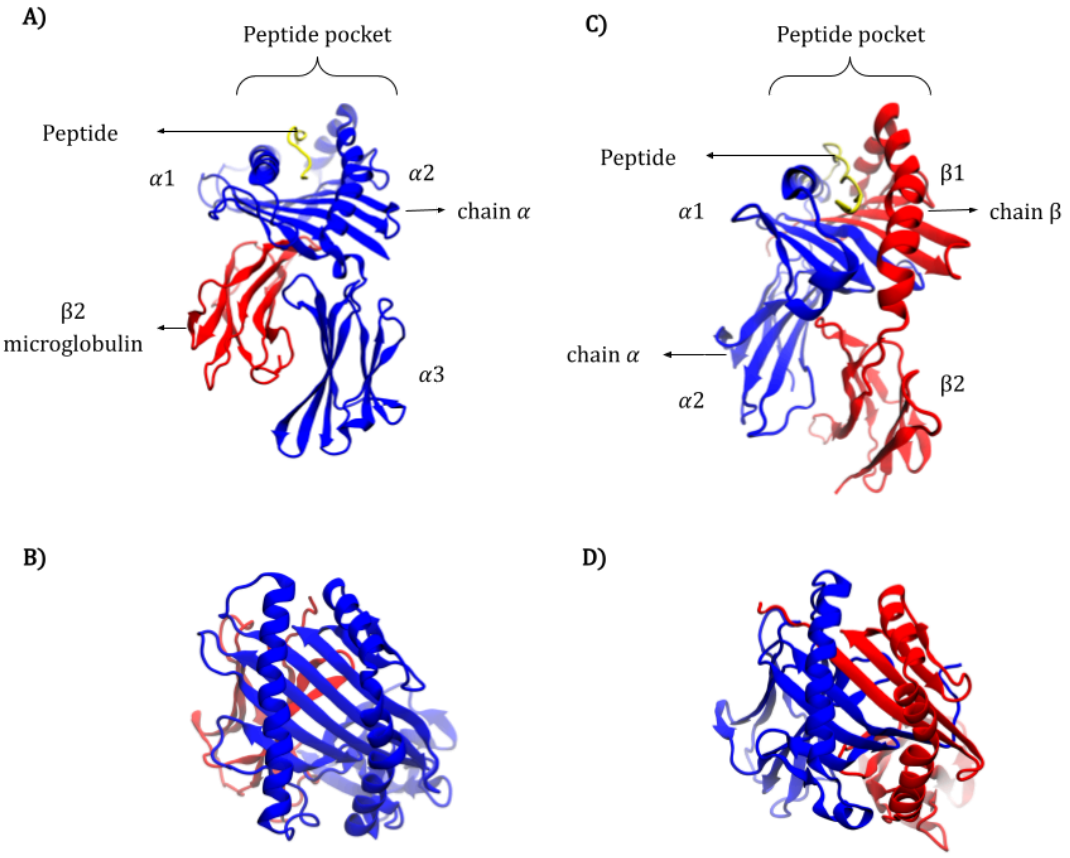



Figure 4: Example structures of proteins HLA class I and class II. A) Backbone representation of the class I protein HLA-A*02:01 (PDB entry: 6TRN). B) Top view of the peptide pocket in the class I protein

HLA-A*02:01 (peptide not shown). C) Backbone representation of the class II protein HLA-DQA1*01:02-DQB1*06:02 (PDN entry: 1UVQ). D) Top view of the peptide pocket in the class II protein HLA-DQA1*01:02-DQB1*06:02 (peptide not shown).

In class I HLA proteins, the extracellular region of chain $\alpha$ consists of 284-285 AA divided into 3 domains named $\alpha$1, $\alpha$2, and $\alpha$3 of about 90 AA each one whereas the β2 - microglobulin has 99 AA. Only the $\alpha$ chain enters into the membrane, involving approximately 25 hydrophobic AA. The intracellular region comprises approximately 30 AA with a small cluster of basic amino acids interacting with the inner side of the membrane, thereby anchoring the protein to the membrane [10]. As previously mentioned, the β2-microglobulin is encoded by a gene outside the HLA ones and is conserved among all class I HLA proteins

In class II HLA proteins, the extracellular region of the chain $\alpha$ is composed of 191-194 AA divided into 2 domains named $\alpha$1 and $\alpha$2 of about 90 AA each one whereas the chain β of 196-198 AA is divided into 2 domains named β1 and β2 of about 92 AA each one. Both chains enter into the membrane, comprising approximately 25 hydrophobic AAs of each one. Analogous to the class I case, the intracellular region of each chain has a small cluster of basic amino acids playing the role of protein anchoring to the inner face of the membrane.

Finally, the immune response is mediated by T lymphocytes. The TCR of T cells recognize the peptide-HLA complex (the exposed portion of the peptide in the HLA pocket, as well as some amino acids from the $\alpha$ helices of the HLA molecule), with the peptide being responsible for the fine specificity of antigen recognition.

### 1.1.3.4 HLA Typing Methods

#### 1.1.3.4.1. Sequence-Specific Oligonucleotide (SSO)

HLA-SSO typing emerged as one of the pioneering PCR-based HLA typing methods, introduced around 1990 [22]. Oligonucleotides are short nucleotide polymers, typically DNA fragments. "Sequence-specific" implies these oligonucleotides are designed to bind selectively to a particular nucleotide sequence. Following a locus-specific PCR using primers capable of binding to all known HLA alleles, the PCR amplification product hybridizes with oligonucleotides designed to recognize an extensive panel of known polymorphisms for any locus. Specifically, when exposed to a mixture containing their complementary sequence, these SSOs bind (or hybridize) to that sequence, forming a double-stranded DNA molecule. If the target sequence is absent, the SSO will not hybridize. Each probe can hybridize with a certain number of alleles, and the comparison of positive and negative hybridizations allows deduction of the alleles present in the starting DNA. Consequently, the greater the number and optimal selection of probes, the higher the test resolution. The goal is to identify alleles at the two-field level, which serological methods could not provide and correspond to the protein. As explained in Section 1.1.3.1, three fields for silent mutations of the genetic code, or four fields incorporating non-coding region polymorphisms, are currently of little utility. As many PCRs as necessary for the required loci are performed, and currently, up to several hundred color-coded probes can be analyzed simultaneously [23, 24]. This approach typically returns a list of ambiguities (e.g., A*03:01/A*03:01N/A*03:03N/A*03:04/...) despite utilizing several hundred probes due to the highly polymorphic nature of HLA (see Section 1.2.1). SSO is now commonly employed as a confirmatory typing method, performed on a few highly polymorphic loci (such as HLA-A, -B, and -DRB1) to limit costs while maintaining its efficacy as a confirmation approach. Nevertheless, it has been the routine method for over 20 years in most laboratories. Due to the necessity of two lengthy steps: PCR followed by hybridization/reading of each locus of interest, this method is unsuitable for urgent situations (approximately 2 days total [25]) but more appropriate for batch processing (88-184 samples can be readily typed within that time frame [25]). SSO can be adapted for emergency settings, but technical details will not be discussed in this manuscript.

#### 1.1.3.4.2. Sequence-Specific Primers (SSP)

Shortly after the development of SSO, the SSP method was introduced by Ole Olerup and Henrik Zetterquist [26]. This approach involves PCR amplification using primers specific to certain sequences. If the target sequence is present in the sample, the primer binds and allows PCR amplification. Amplification is detected by electrophoresing the PCR products through an agarose gel containing a DNA-binding dye, such as ethidium bromide [25]. A photograph taken under ultraviolet light reveals the presence or absence of amplified HLA alleles. In contrast to SSO, multiple PCR amplifications specific to certain alleles are performed. Similar to SSO, the comparison of positive and negative PCRs (instead of positive and negative hybridizations in SSO) allows the deduction of alleles present in the starting DNA. SSP can be carried out in two steps. The first step, with a reasonable number of PCRs (often 96, including controls), allows typing at a level equivalent to serology for the HLA-A, HLA-B, HLA-DRB1, and HLA-DQB1 loci. Then, if necessary, each inferred serological group can be studied in more detail using a dedicated reagent kit, which includes a variable number of PCRs (depending on the known polymorphism extent of each locus) designed to optimally specify the present alleles. Although both SSO and SSP utilize PCR amplification, SSP's second step (electrophoresis) only takes approximately 15 minutes and was, for a long time, the only method to approach a high-resolution level in an emergency setting (the entire process, including DNA extraction, can be completed within three hours [25]). The trade-off is that, due to the number of PCRs required for each sample, the SSP method is not well-suited for typing many patients [25]. Around 10 years ago, high throughput SSP approaches using gel-free approaches based on fluorescence read-outs of DNA intercalating agents in solution, and 384-well PCR with additionally a progressive implementation of multiplex reactions in the same well, have rendered SSP suitable for high resolution typing of all clinically relevant genes in two hours. This approach is widely used for emergency typing of organ donors, principally.

#### 1.1.3.4.3. DNA Sequencing

The SSO and SSP strategies, though differing in principles and applications, both lack the ability to analyze the complete HLA gene sequence to provide unambiguous second-field typing. Despite continuous efforts by reagent manufacturers to improve kit resolution, maximum resolution down to the nucleotide level has never been achieved, even when limited to protein-coding regions, except for a few alleles with highly specific sequences in restricted gene areas. Furthermore, SSO and SSP can only identify known alleles by comparing the obtained profile to a reference list. A novel allele with an undescribed sequence will either go unnoticed if the kits lack coverage of the region harboring its unique polymorphism or result in a failed typing, as one or more reactions will not yield the expected outcome. Consequently, DNA sequencing remains the only truly effective approach [25]. The initial method employed was the traditional Sanger sequencing applicable to any DNA [27]. The DNA is amplified by PCR, and the amplified fragments are analyzed base by base in polyacrylamide gel. Akin to SSO and SSP, both alleles of an individual for a specific locus are analyzed simultaneously, leading to ambiguities. A heterozygous site appears as an overlapping peak on the chromatogram. Although the HLA genes harbor numerous polymorphisms, there are far fewer identical regions. Hence, it is sometimes impossible to attribute a specific polymorphism to a particular allele among the two present. The PCR fragments analyzed by sequencing are short, precluding definitive linkage of distant polymorphisms to a single allele. This issue is especially pronounced when the sequences of the two present alleles have multiple distinguishing polymorphisms, some of which are partially present in other allele sequences [28]. Thus, typing ambiguities persist, as with SSO and SSP. Mono-allelic sequencing can circumvent this problem but is more demanding to implement and will not be detailed here. In 2009, a revolutionary technology termed Next-Generation Sequencing (NGS) emerged in the HLA field [29]. This approach is fundamentally different: each fragment amplified from the template DNA is analyzed individually, generating as many sequences as fragments (termed reads). If these fragments overlap, reconstructing the entire gene and assigning a unique sequence to each of the two present alleles without ambiguity is presumably straightforward. Amplification length is no longer a significant constraint, although alignments can

sometimes be complex due to high inter-allelic similarity, and the decisions made by the software tools of these kits can occasionally be modified by biologists validating the analyses. Another feature of NGS is the ability to identify dozens of different samples by a short nucleotide sequence (an index, akin to a barcode). This allows, after labeling, the mixing of amplifications for the sequencing stage. The process proceeds as follows: the initial DNA is amplified for all genes of interest, then labeled, with each DNA separated from the others at this stage. Subsequently, the PCR products are mixed and fragmented to produce fragments compatible with the sequencing step (around 300 base pairs). This forms the "library" that will be read on the sequencing chip in a dedicated, often expensive, device. NGS has become standard for routine analysis in recent years. However, it is unsuitable for emergencies, requiring at least two days of work [30] and prohibitively costly for single DNA samples given the technical process design. Its cost has significantly decreased over time, and it is now possible to analyze up to 96 DNA samples simultaneously with theoretically maximum resolution (every base of all introns and exons).

### 1.1.4 Antigen-Antibody recognition

#### 1.1.4.1 Structural Basis of Antigen-Antibody Interaction

Antigen-antibody recognition is a pivotal aspect of the immune system, playing a crucial role in defending the body against pathogens but becoming a major barrier in the context of organ transplantation. DSA recognition of HLA antigens is mainly responsible for organ rejection. Antibodies, also known as immunoglobulins (Ig), consist of two identical heavy chains ($C_H$) and two identical light chains ($C_L$). These chains form distinct regions, such as the variable (V) and constant (C) domains. The V regions contribute to antigen recognition, exhibiting high diversity due to genetic recombination. The antigen-binding portion of an antibody molecule is the Fab region, and the C-terminal end that is involved in effector functions is the $F_c$ region (see Figure 5).

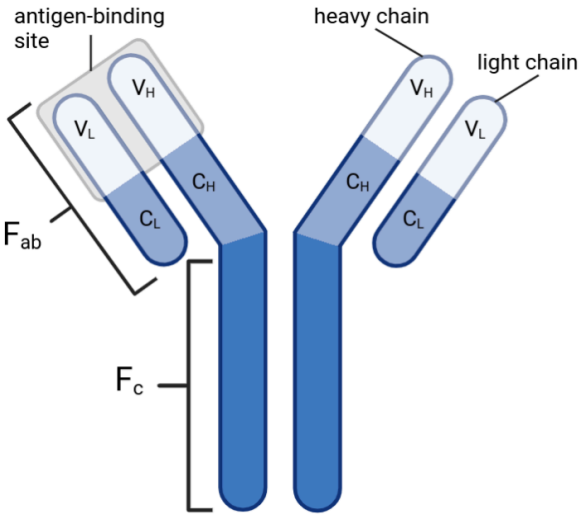



Figure 5: Structure of an antibody molecule. The antigen-binding sites (also known as paratopes) are formed by the juxtaposition of $V_L$ and $V_H$ domains.

The binding between an antibody and its cognate antigen is highly specific and relies on the complementarity between their three-dimensional structures. Antigens, on the other hand, expose specific epitopes recognized by antibodies. Epitopes are small, accessible regions on the antigen surface. The structural and physicochemical complementarity between the paratope on the antibody (the antigen-binding site) and the epitope on the antigen is crucial for a strong and specific interaction. The recognition of antigens by antibodies involves noncovalent, reversible binding. Various types of noncovalent interactions may contribute to antibody binding of antigens, including electrostatic forces, hydrogen bonds, van der Waals forces, and hydrophobic interactions.

There are two types of epitopes: linear and conformational. Linear epitopes consist of continuous amino acids in the protein sequence and are typically presented by APCs to T cells but also to B cells. Conformational epitopes result from protein folding and consist of discontinuous amino acids in the protein sequence. Antibodies typically recognize these epitopes. It is important to note that antibodies must directly interact with their target epitope on the antigen. This requires the epitope to be accessible at the cell surface. In contrast, T-lymphocytes do not need to interact with the antigen directly. Their TCR binds to the peptide/HLA complex, with the peptide coming from any part of the original molecule. In the present work, we will focus on B cells epitopes, hence on conformational epitopes.

#### 1.1.4.2 HLA epitopes and eplets

The sensitivity of antibodies can be astoundingly high. A single polymorphism, i.e. a single amino acid variation in the sequence between the donor and recipient HLA molecules can serve as an immunogenic epitope, capable of eliciting an antibody response. This incredible sensitivity is amplified by the immense diversity of HLA polymorphisms, resulting in a vast array of HLA epitopes. A consequence of this polymorphism is the phenomenon of Cross-Reactive Groups (CREGs), where an HLA epitope is shared by multiple HLA molecules. This means that immunization against a given antigen can lead to the recognition of other, unknown antigens for the individual who will make the antibody, and thus complicates transplant compatibility. To define HLA epitopes, one employs a combination of experimental and computational approaches that lean on modern molecular typing techniques and the 3D structures of HLA molecules. Experimentally, antibodies in patient's sera are tested against a panel of HLA antigens and those that react positively are analyzed to identify shared amino acid residues that differ from the negative ones at the corresponding positions (these correspond to the so-called polymorphisms). Then using computational methods, these polymorphisms (also known as eplets) are utilized to define HLA epitopes. More precisely, an HLA epitope is usually defined by the region encompassing 15Å around an eplet (see Figure 6) [31]. The surface covered by HLA epitopes typically spans 650 to 900 $Å^2$ and involves between 15 and 22 AAs [32]. This approach has aided in the identification of close to 300 HLA epitopes [33].

The term "eplet" was initially coined by Duquesnoy et al. in [32]. The HLA Eplet Registry [34] is a website-based database that gathers data on putative and confirmed eplets reported in the scientific literature. Nowadays[6] this database holds 224 Class I eplets (of which 72 are confirmed), 123 HLA-DRB (of which 36 are confirmed), 83 HLA-DQ eplets (of which 27 are confirmed), 62 HLA-DP (of which 11 are confirmed) and 21 interlocus eplets (of which 5 are confirmed), i.e. eplets shared across antigens of different locus. The term 'antibody-verified eplet' (also known as 'confirmed eplet') is applied when there is experimental evidence confirming the epitope's ability to elicit an immune response (methods for verifying an eplet will be addressed in more depth in a subsequent section), while putative eplets (also known as non-confirmed eplets) have mainly been deduced from alignments of HLA proteins sequences, and therefore have not been confirmed yet. Exploring confirmed eplets on genes beyond the commonly studied HLA-A, -B, -C, -DRB1, and -DQB1 has gained significant attention in recent years. This interest arises because mismatches in these additional genes can lead to the development of DSA and contribute to antibody-mediated rejection [35]. While the role of different genes in this context is still in its early phases, recent research has indicated notable findings. Elevated risk of formation of *de novo* DSA (dnDSA) and graft loss in kidney transplant have been linked to the mismatch load (i.e. the number of sequence mismatches/differences) of confirmed eplets, notably the ones from DQA1 and DQB1 genes [36, 37].

An eplet notation has been developed using the amino acid residues in polymorphic sequence positions. For example, the eplet 40ERV corresponds to positions 40, 41, and 45 displaying the residues E (GLU), R (ARG) et V (VAL) (see Figure 6).

[6] HLA Eplet Registry accessed on January 15th 2024

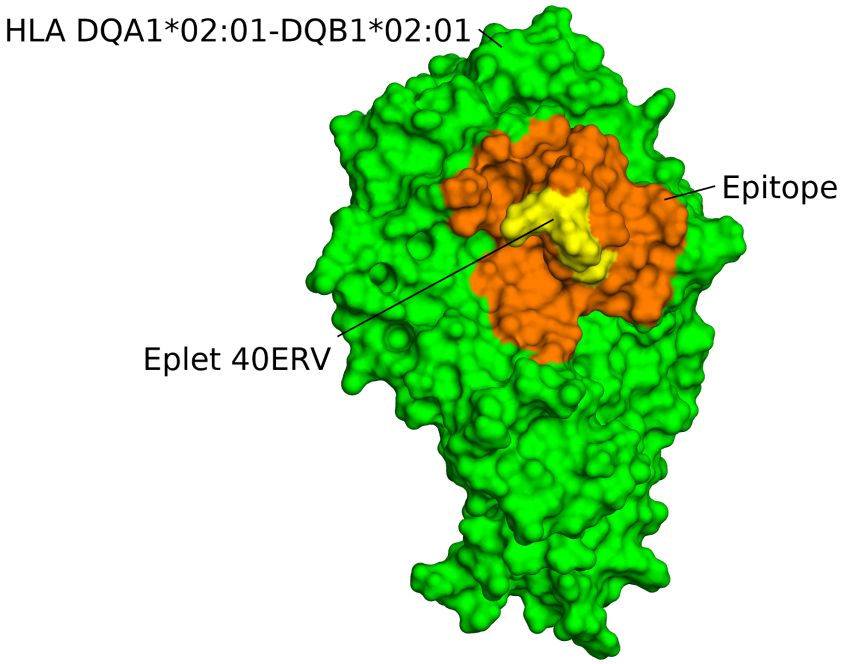



Figure 6: Visualization of the eplet 40ERV and its associated epitope on the antigen HLA-DQA1*02:01-DQB1*02:01. Surface representation of antigen HLA-DQA1*02:01-DQB1*02:01. Eplet 40ERV (composed of three residues: 40E, 41R and 45V) is shown in yellow. The orange zone (patch) corresponds to residues that are 15Å at most away from residue 41R of the eplet 40ERV. This zone represents the putative epitope.

### 1.1.5 Anti-HLA antibody detection and eplet confirmation

Currently, the most widely used technique for the detection of anti-HLA antibodies in patient's serum is the Luminex® technique [38]. This technology uses polystyrene microspheres coated with purified HLA antigens, which are distinguished by their internal fluorescence due to the combination of two fluorochromes excited by a red laser at a wavelength of 635 nm, in proportions that vary according to the type of bead. This technology has a version called "Single Antigen Beads" (SAB), which allows the specificities of the anti-HLA antibodies circulating in the receptor to be defined precisely and semi-quantitatively [38, 39]. In fact, a cell expresses several different HLAs, making it impossible to unambiguously identify the target antigen of an antibody among all the possible targets of the cell. The SAB version of the Luminex technique separately assays 97 and 95 class I and II antigens, respectively, which are the most common in the North-American population, at a rate of one antigen per bead. These beads are exposed to a patient's serum. If the patient's serum contains antibodies targeting the HLA antigen on a particular bead, these antibodies will bind to the bead. After several washing steps to remove any unbound antibodies, a secondary antibody, specifically anti-human IgG (γ-chain specific) conjugated with R-Phycoerythrin, is introduced. These detection antibodies are designed to bind to the patient's antibodies that are already bound to the HLA proteins on the beads. The presence and quantity of the patient's antibodies directed against the specific HLA molecules on the bead can then be determined by measuring the fluorescence emitted by the Phycoerythrin-conjugated detection antibodies. In other words, the SAB technique measures the mean fluorescence intensity (MFI) of each bead used, a value that is proportional to the number of antibodies that recognize the antigen on the bead and can be interpreted as a semi-quantitative measure of the "strength" of the antibody's interaction with the HLA antigen (see Figure 7).

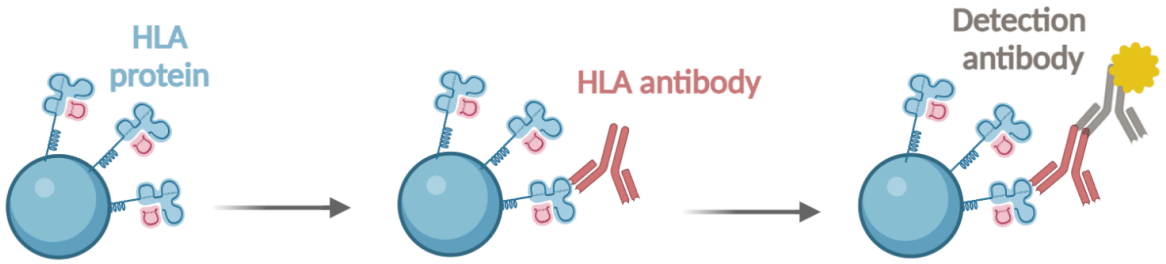



Figure 7: Overview of the SAB assay for anti-HLA antibody detection. Image taken from [40].

Verification methods for eplets in the HLA Eplet Registry range from polyclonal sera[7] from multi- and uni-parous women[8] to murine and human monoclonal antibodies[9], as well as antibodies purified by adsorption and elution from sera of HLA-immunized individuals. However, the classification of antibody verification based on different validation methods is considered problematic, as not all approaches provide the same level of evidence. For instance, murine monoclonal antibodies that target HLA antigens may identify certain eplets as foreign from the mouse's perspective. However, this does not necessarily mean that another human would produce antibodies against the same eplet. This is because the eplet may not be considered non-self from another human's perspective, even if there is a mismatch. This is due to the fact that the HLA antigens will be more similar to each other than to mouse MHC antigens. Bezstarosti et al. classified the level of evidence for antibody verification in Table 1 and assigned all current confirmed eplets to one of those categories. They also analyzed several unpublished human HLA-specific mAbs using the Luminex SAB assay to verify their HLA reactivity for eplet antibody verification. Most of the analyzed mAbs verified their aimed eplets, but some did not [41].

Table 1: Level of evidence of confirmed eplets as defined by Bezstarosti [41]

| Level of evidence | Description |
|---|---|
| A1 | Human monoclonal antibody + SAB assay, possibly supported by complement-dependent cytotoxicity assay (CDC) with high-resolution HLA typed cells (second field). |
| A2 | Adsorption and elution studies + SAB assay, possibly supported by CDC with high-resolution HLA-typed cells. |
| B | Patient serum tested in SAB assay and/or CDC with high-resolution HLA typed cells. |
| C | Human monoclonal antibody or adsorption and elution studies or patient sera tested with low-resolution HLA typed cells only (first field or serological typing). |
| D | Any reactivity analysis with antibodies from other species (e.g. murine monoclonal antibody). |
| X | Level of evidence is not satisfying. |

[7] Sera containing a mixture of antibodies, each produced by a different B cell in response to potentially different epitopes.
[8] Multiparous women have had two or more viable births, whereas uniparous women have had only one.
[9] Laboratory-produced antibodies that are designed to target specific antigens.

## 1.2 Basics of structural bioinformatics

### 1.2.1 A gentle introduction to proteins

Proteins are essential macromolecular components of life, playing a central role in a multitude of cellular functions across all living organisms. These versatile molecules are responsible for a wide range of tasks, including maintaining cell shape and inner organization, manufacturing products, cleaning up waste, and performing routine maintenance. Moreover, proteins can receive signals from the extracellular environment and mobilize intracellular responses. As the workhorse macromolecules of the cell, proteins exhibit a diversity that matches the variety of functions they perform.

Proteins are composed of a sequence of amino acids, which are translated from the nucleotide sequence of a gene. Living organisms utilize 21 naturally occurring amino acids to synthesize all proteins. Each amino acid corresponds to a triplet of nucleotides, as dictated by the genetic code [42]. Every amino acid contains a common base (an α-carbon to which an amine group, a carboxyl group, and a hydrogen are attached) and a side chain that differentiates them. The chemical properties of the side chain determine the characteristics of the amino acid and its role in the protein. Amino acids are connected by peptide bonds, in which the carboxyl group of one amino acid joins the amino group of another, releasing a water molecule in the process. The backbone of proteins involves the α-carbon, the amino, and carboxyl groups, while the side chains are not involved in the peptide bond (see Figure 8).

A) B)

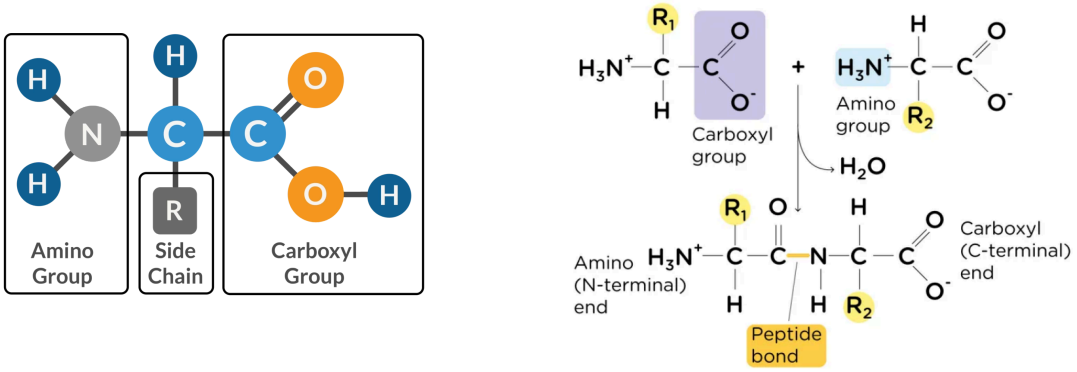



Figure 8: A) The amino acid structure. B) The formation of a peptide bond between two AAs.

The structure of a protein at different levels (primary, secondary, tertiary, and quaternary) is determined by its specific amino acid sequence. The primary structure refers to the linear sequence of amino acids in a protein. The secondary structure describes regular, local structures of the protein backbone, stabilized by intramolecular hydrogen bonds within the protein backbone. The two main types of secondary structures are α-helices and β-sheets. α-helices are coils held in place by hydrogen bonds every 4 amino acids, involving consecutive amino acids. β-sheets are formed by strands that unfold side by side in the same plane, and can be parallel or antiparallel. Different strands are not necessarily formed by consecutive amino acids or by amino acids from the same polypeptide chain. The remaining amino acids between these structures form flexible loops. The tertiary structure represents the overall three-dimensional shape of the protein, determined by interactions such as hydrogen bonding, ionic bonds, van der Waals interactions, disulfide bridges, and stacking interactions between side chains or with the backbone. During 3D folding, hydrophobic amino acids are typically found inside the protein, while hydrophilic amino acids are found on the surface, due to the hydrophilic nature of the cellular environment. The quaternary structure of a protein refers to the arrangement of multiple polypeptide chains in a structural assembly (protein complex), formed by the interactions between these subunits. Many proteins consist of two or three polypeptide chains assembled to form a functional macromolecule

(e.g. antibodies). The structuring of the protein gives it its 3D shape, which is crucial for its biological activity. Figure 9 presents the four levels of protein organization.

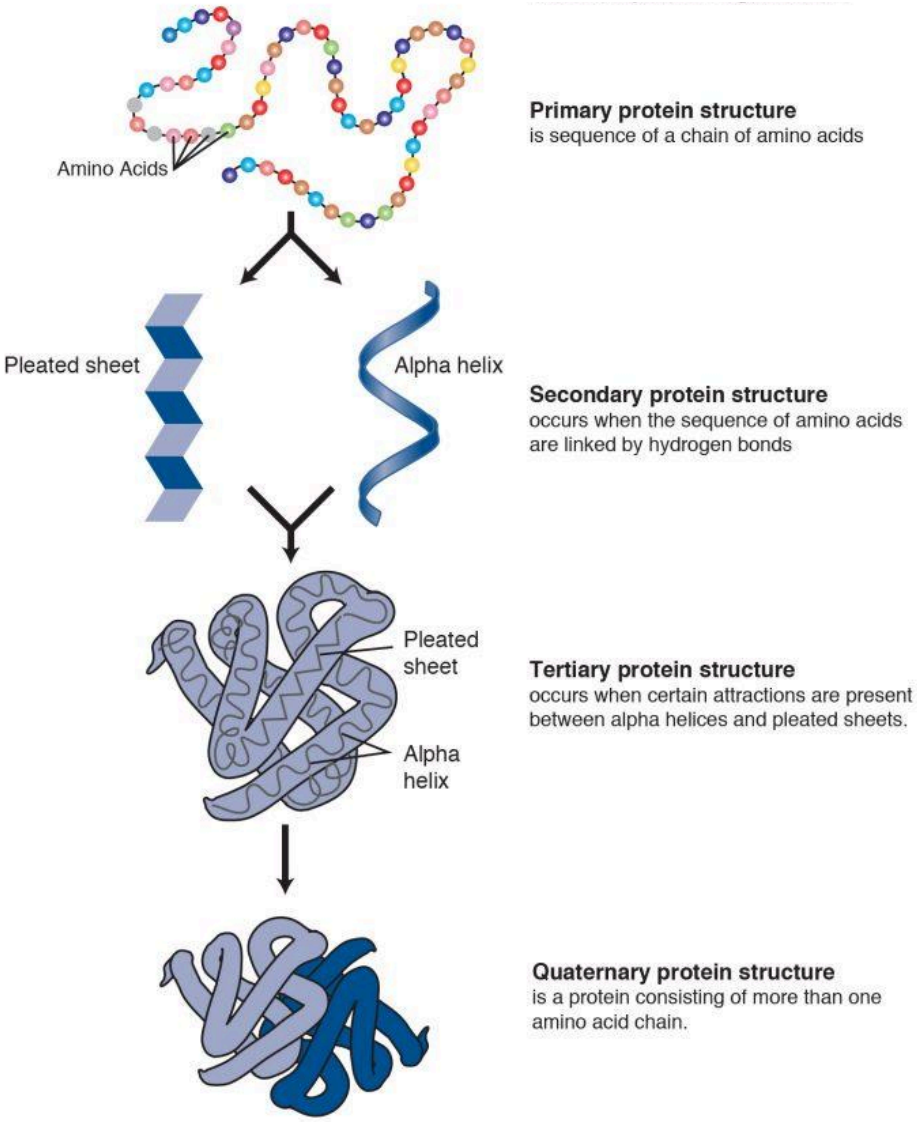



Figure 9: The 4 levels of protein organization. From primary structure to quaternary structure. Image taken from [43].

### 1.2.2 Physico-chemical properties of amino acids

AA properties are essential for understanding the structure, function, and behavior of proteins. Here, the AA properties as described by [44] are introduced, namely: hydropathy, volume, chemical characteristics, charge, polarity and Hydrogen donor/acceptor atoms.

#### 1.2.2.1 Hydropathy

The hydropathy of an AA, which is determined by the physico-chemical properties of its side chain, influences the orientation of the side chain in the 3D protein structure (inside the protein, on its surface, or neutral). Three hydropathy classes are defined using the Kyte and Doolittle AA hydropathy index [45]. AAs with a hydropathy index greater than or equal to 1.8 were classified as hydrophobic, those with an index less than or equal to -3.3 were classified as hydrophilic, and those with an index between -3.3 and 1.8 were classified as neutral. Consequently, the three classes made up as follows (with decreasing hydropathy

index in parentheses): hydrophobic (I, V, L, F, C, M, A, W), neutral (G, T, S, Y, P, H), and hydrophilic (D, N, E, Q, K, R) (see Table 2). Tryptophan (W) was included in the hydrophobic class, although its hydropathy index varies from -0.9 (Kyte and Doolittle index) to 1.9 (Goldman-Engelman-Steitz -GES- hydrophobicity scale [46, 46]), depending on the study.

Table 2: Kyte and Doolittle hydropathy index of AAs and the corresponding hydropathy classes.

| I | V | L | F | C | M | A | W | G | T | S | Y | P | H | N | D | Q | E | K | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4.5 | 4.2 | 3.8 | 2.8 | 2.5 | 1.9 | 1.8 | -0.9 | -0.4 | -0.7 | -0.8 | -1.3 | -1.6 | -3.2 | -3.5 | -3.5 | -3.5 | -3.5 | -3.9 | -4.5 |
| **Hydrophobic** | | | | | | | | **Neutral** | | | | | | **Hydrophilic** | | | | | |

### 1.2.2.2 Volume

The AA volume classes are defined based on the AA volumes in angstrom$^3$ (Å$^3$) [47]. Five classes were defined (with increasing amino acid volume in parentheses): very small (60–90 Å$^3$; G, A, S), small (108–117 Å$^3$; C, D, P, N, T), medium (138–154 Å$^3$; E, V, Q, H), large (162–174 Å$^3$; M, I, L, K, R), and very large (189–228 Å$^3$; F, Y, W) (Table 3).

Table 3: Approximate volume of AAs according to [47] and the corresponding volume classes.

| G | A | S | C | D | P | N | T | E | V | Q | H | M | I | L | K | R | F | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 60.1 | 88.6 | 89 | 108.5 | 111.1 | 112.7 | 114.1 | 116.1 | 138.4 | 140 | 143.8 | 153.2 | 162.9 | 166.7 | 166.7 | 168.6 | 173.4 | 189.9 | 193.6 | 227.8 |
| **Very small** | | | **Small** | | | | | **Medium** | | | | **Large** | | | | | **Very large** | | |

### 1.2.2.3 Chemical characteristics

The AA chemical characteristics classes are defined based on the principal chemical property of the AA side chain. Seven classes are defined: aliphatic (A, G, I, L, P, V), sulfur (C, M), hydroxyl (S, T), acidic (D, E), amide (N, Q), basic (H, K, R) and aromatic (F, W, Y). The side chains of F and W are typically buried in the hydrophobic interior of proteins. Tyrosine (Y) differs from phenylalanine by a para-hydroxyl group, resulting in a polar side chain that is weakly acidic. Glycine (G) has the simplest structure among all amino acids, with its side chain being merely a hydrogen atom. The small size of this side chain allows glycine to fit into niches that cannot accommodate any other amino acid, giving it a unique function in the structure of many proteins. The structure of proline (P) differs significantly from that of other amino acids, with its side chain bonded to both the nitrogen and the α-carbon in a pyrrolidine ring, which restricts the geometry of the backbone chain of the protein containing it and introduces abrupt changes in the direction of the chain. P is not chemically reactive and is an imino (—NH—) rather than an amino (—NH2—) acid due to the bond to nitrogen.

### 1.2.2.4 Charge

AA side chains can be classified as charged or uncharged. Charged polar side chains are either basic (H, K, R) or acidic (D, E) and are hydrophilic. Histidine (H) is in the neutral class under physiological conditions, although it is weakly basic. Uncharged side chains can be nonpolar (aliphatic, sulfur) or polar (hydroxyl, amide). Nonpolar side chains are hydrophobic, while uncharged polar side chains are neutral (hydroxyl) or hydrophilic (amide). Tyrosine (Y) is in the neutral class, although its OH is weakly acidic and polar.

#### 1.2.2.5 Polarity

AA side chains can be classified as polar or nonpolar based on their ability to interact with water. Polar side chains are hydrophilic and can form favorable interactions with water molecules. These include the charged side chains (basic and acidic) and the uncharged polar side chains (hydroxyl and amide). Nonpolar side chains, such as those in the aliphatic and sulfur classes, are hydrophobic and tend to avoid interactions with water, preferring to be buried in the hydrophobic core of the protein.

#### 1.2.2.6 Hydrogen Donor or Acceptor Atoms

AA side chains can also be classified based on their ability to act as hydrogen bond donors or acceptors. Side chains containing nitrogen (N, Q, H, K, R) or oxygen (S, T, Y, D, E) atoms can participate in hydrogen bonding. The amide group in asparagine (N) and glutamine (Q) can act as both a hydrogen bond donor and acceptor. The hydroxyl group in serine (S), threonine (T), and tyrosine (Y) can also serve as both a hydrogen bond donor and acceptor. The basic side chains of histidine (H), lysine (K), and arginine (R) can act as hydrogen bond donors, while the acidic side chains of aspartic acid (D) and glutamic acid (E) can act as hydrogen bond acceptors.

### 1.2.3 Experimental methods in structural biology

Experimental methods for determining the 3D structure of macromolecules have not only deepened our understanding of a wide range of biological processes but have also been the fundamental basis for the development of a large number of bioinformatics tools that have enabled major advances in, for example, molecular simulation, rational drug design, 3D structural prediction of molecules, etc. Experimentally determined structures are typically collected in the Protein Data Bank (PDB) [48], the largest structure database in existence. The main experimental methods for structure determination are presented below.

#### 1.2.3.1 X-Ray Crystallography

To date, X-ray crystallography remains the most popular method of obtaining a detailed 3D structure of a macromolecule. The method is based on X-ray diffraction through crystals and involves three basic steps. The first step is to obtain an X-ray diffraction crystal. This step is usually the bottleneck in the use of this method. The second step is to obtain the X-ray diffraction pattern of the crystal. Structure determination by X-ray crystallography depends entirely on the diffracted rays produced when X-rays are deflected in directions determined by the position of the atoms in the crystal lattice and by the wavelength of the X-rays. By measuring the angle and intensity of the diffracted rays, a three-dimensional lattice can be obtained. From this lattice, it is possible to determine the average position of the atoms in the crystal. The quality of this lattice determines the resolution. With this method, crystal structures of very good quality (between 1-3 Å) can be obtained [49]. However, one of the main limitations of this method is that the structures obtained may differ from those that exist under physiological conditions, since protein crystallization requires very specific medium conditions, e.g. supersaturation of the protein, as well as temperature, pH and solvent. Another limitation of X-ray crystallography is the difficulty in determining overly flexible regions such as loops [50].

#### 1.2.3.2 Nuclear Magnetic Resonance:

This method is based on the property of certain atomic nuclei that have a nuclear spin. These nuclei absorb electromagnetic radiation energy and emit it back at a precise frequency that depends on the atomic environment. Measuring these frequencies allows the structure of molecules to be reconstructed. The method is particularly effective for small molecules. Unlike X-ray crystallography, this method does not require the molecule to be crystallized, which makes it possible to obtain structures of poorly structured

or very flexible objects. In fact, NMR works in soluble media, making it possible to obtain multiple structures of the same object and thus understand the flexibility. NMR structure determination still presents some considerable challenges: the method is limited to systems of relatively low molecular mass, data acquisition times are long, data analysis remains a lengthy process, and it is difficult to evaluate the quality of the final structures [51].

#### 1.2.3.3 Cryogenic Electron Microscopy

Although X-ray crystallography and NMR are the fundamental techniques for structure determination at atomic resolution, Cryogenic Electron Microscopy (Cryo-EM) has been intrinsically gaining popularity in the field of structural biology. This is mainly because in Cryo-EM large multi-subunit complexes of viruses, bacterial appendages, eukaryotic ribosomes, and cellular organelles can be visualized without the need for crystallization. Currently, structures of small proteins, or other antigens of similar size cannot be analyzed using Cryo-EM. In this manner structures of complexes as small as 65 kDa can be determined.

In Cryo-EM, a small amount of sample suspended in buffer solution is quickly frozen to form a non-crystalline glass-like specimen called vitreous ice. This vitrified sample is then exposed to high frequency electrons which scatter through the specimen due to electrostatic interactions with the sample's atoms. Scattered and unscattered electrons form an interference pattern that results in 2D projection images which are recorded. These 2D projections are recorded by viewing the specimen from different angles. For structure determination, the combined 2D projections are further digitally processed and analyzed taking into account all orientation parameters. Unfortunately, quite often this technique provides low-resolution structural information, and it has to be combined with other methods to obtain more details [52].

#### 1.2.3.4 Limitations of the experimental methods

Experimental methods have enabled the understanding of the function of biomolecules and their interactions. However, these methods have several limitations that can lead to uncertainties in the resulting structural models, influencing the interpretation and applications derived from such data.

One major limitation of experimentally solved structures is the possibility of discrepancies between the theoretical model and the actual experimental data. These discrepancies may be due to errors made during the model's construction. When building a model to represent an electron density map, it is crucial to avoid incorrect assumptions about the identity or conformation of residues/nucleotides. Such assumptions can lead to a model that inadequately fits the data or, even worse, misrepresents the molecular structure.

Another challenge arises when experimental data is unavailable for specific regions of the model. In certain regions of a protein or nucleic acid, inherent disorder or movement can result in areas with low electron density. This can cause inefficient scattering of X-rays or electron beams, leading to poor resolution or even a complete absence of structural information. As a result, accurately modeling these regions becomes difficult due to the lack of data. Additionally, during the process of building and refining a model, inaccuracies may cause deviations from idealized bond lengths, bond angles, or dihedral angles. These geometric aberrations can be subtle and may require careful assessment against known physicochemical constraints or higher-resolution data to rectify.

Clashes between atoms, where the model predicts atoms to be closer than physically possible, are not uncommon and may indicate errors in model building or refinement. Validation tools can identify steric clashes caused by physically improbable configurations.

Furthermore, crystallographic refinement processes can sometimes produce models that are overly influenced by the starting model or refinement restraints, leading to bias in the resulting structure. This issue is of particular concern when the resolution of the data is low, as the electron density provides less guidance to the correct structure.

It is important to note that methods such as X-ray crystallography or Cryo-EM typically provide a static image of a molecule, which may exist in several different conformations due to its dynamic nature.

This 'crystallographic snapshot' often fails to capture the range of conformations present in solution, overlooking the fluidity with which biological macromolecules operate within a cellular context. Advanced computational methods and molecular dynamics simulations are increasingly used to complement X-ray crystallography and Cryo-EM data, providing insights into the flexibility and dynamics of molecules beyond the static crystal structures.

### 1.2.4 Repair of experimentally solved structures

As previously mentioned, due to the different intrinsic limitations of experimental methods for solving structures, various types of errors can occur in the structures obtained in this way. Addressing these errors is crucial for obtaining high-quality structures, as they can lead to incorrect interpretations of structural features or can cause a molecular dynamics simulation to fail. For example, clash errors can induce extremely high van der Waals repulsion forces, which can generate very large atomic motions, breaking the topology of the structure. However, there are computational methods to alleviate most of these errors. The following is a description of the most important types of errors to be addressed and which will be tackled in the present work.

#### 1.2.4.1 Refinement of the crystallographic structure model

The refinement and validation of a crystallographic structure model is the final step before submitting the coordinates and associated data to the PDB. This process aims for optimal interpretation of the electron density generated by the X-ray diffraction. For example, each amino acid side chain is examined to identify if there is an alternative rotameric conformation that fits the electron density equally or better. In numerous cases, this thorough search results in a significant enhancement of the model's geometric quality [53]. Other parameters can be also optimized such as finding appropriate restraint weights or selecting a most suitable B-factor model. PDB_REDO [54] is a recognized tool for this type of tasks and also proposes a databank that contains optimized versions of existing PDB entries.

#### 1.2.4.2 Clashes

The most straightforward sort of contact problem in a structural model is a physically impossible overlap, or clash, of nonbonded atoms [55]. The type of error can arise when atoms are positioned too close together, violating the principles of atomic radii and van der Waals interactions. These steric clashes represent regions of high strain and are physically implausible in a stable protein structure. Clash errors may occur due to poor data resolution or incomplete modeling of flexible regions. Alternatively, they can result from errors during the refinement process of the crystallographic model.

#### 1.2.4.3 NQH flips

Most H atoms can be placed by simple geometry with adequate accuracy, but the orientations of groups such as OH, NH3, and even side-chain amides need to be optimized within entire (local) H-bond networks, including their interactions with nonpolar as well as polar atoms and with ordered water molecules [56, 57]. This optimization process has the beneficial side effect of diagnosing incorrect 180° flips of ASN, GLN, and HIS (NQH) side chains, and of robustly correcting them from H-bond and all-atom contact evidence without affecting agreement to the diffraction data (Word et al., 1999b, Higman et al., 2004, Arendall et al., 2005). The NQH flip corrections provide a no-cost route to local but often important improvements to model accuracy, especially at resolutions lower than ~3Å [55].

### 1.2.4.4 Chirality

Chirality is a fundamental property of many biological molecules. It refers to a chemical property of a molecule being non-superimposable on its mirror image. This means that a chiral molecule can exist in two distinct forms, called enantiomers, which are mirror images of each other (see Figure 10).

In proteins, chirality is most commonly observed at the $C_\alpha$ of amino acids but threonine and isoleucine have an additional chiral center at $C_\beta$. In the case of the $C_\alpha$ chiral center, the four C bonds are filled by a carboxyl group (-COOH), an amino group (-NH2), a hydrogen atom (H), and a side chain (R). The spatial configuration of the side chain R defines each enantiomer. The prefixes L- and D- are typically used to differentiate them. L-amino acids are typically found in proteins. This is because in several organisms the enzymes that synthesize proteins, known as aminoacyl-tRNA synthetases, have a proofreading mechanism that ensures that only L-amino acids are incorporated into the polypeptide chain [58, 59].

### 1.2.4.5 Peptide bonds

The bond that connects the carboxy end of one amino acid to the amino end of the next in a protein is called a peptide bond. Peptide bonds can be distinguished as cis ($\omega \approx 0°$) or trans ($\omega \approx 180°$) isomers, depending on the value of the dihedral angle ω described by $C_{\alpha, n}$, $C_n$, $N_{n+1}$ and $C_{\alpha, n+1}$ (see Figure 10 ). The trans isomer is energetically more stable due to steric reasons and is therefore the prevalent form in proteins, except in the case of peptide bonds involving prolines. The proline residue's distinctive five-membered ring leads to steric hindrance in both the cis and trans positions, reducing the transition barrier and resulting in a cis bond probability of up to 40% [60].

Stereochemical errors can influence the folding and stability of proteins. These errors can lead to the formation of incorrect structures and artifacts. For example, a simulation with a chirality error may result in the formation of a protein with a different fold than the native fold. Additionally, a simulation with a cis peptide bond may result in the formation of an incorrect hydrogen bonding network.

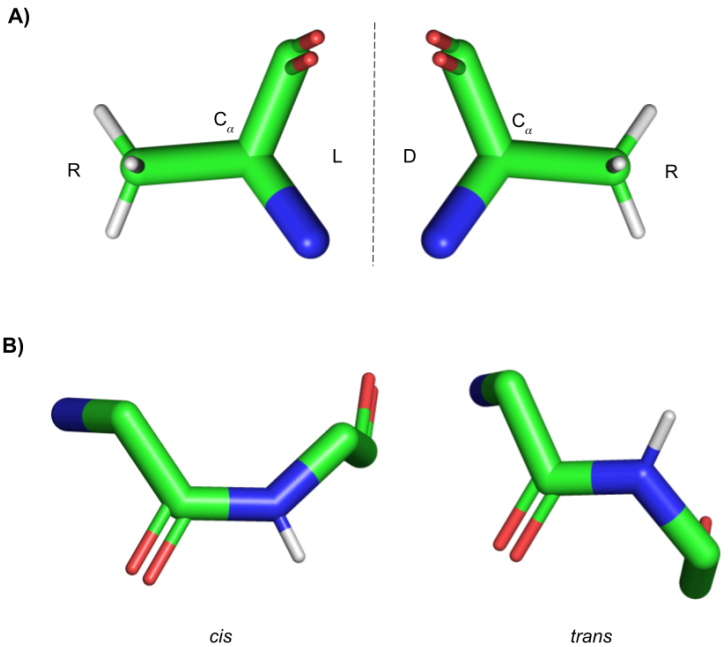



Figure 10: A) Chirality configuration at $C_\alpha$ of an alanine. R corresponds to the side-chain of the amino acid and $C_\alpha$ to the chiral center. B) *cis* and *trans* isomers of a peptide bond. Carbon atoms are shown in green, nitrogen in blue, oxygen in red, and hydrogen in white.

#### 1.2.4.6 Missing atoms or side chains

Missing side chains or atoms are mainly due to highly flexible regions or loops with low electron density in X-ray crystallographic analysis. In addition, hydrogen atoms are not included in many PDB files because they are not easily detectable by X-ray crystallography, although they are essential for analyzing non-covalent interactions and refining atomic-level structures.

### 1.2.5 Protein structure prediction

Until recently, the most effective 3D structure prediction methods in terms of accuracy were limited to homology-based modeling, more accurately called comparative modeling [61]. Tools like Phyre2 [62] and HHPred [63] allow searching for similarities between the sequence to be modeled and those present in 3D structure databases (e.g., PDB, CATH, SCOPe). The MODELLER program [64], a reference in the field of comparative modeling, allows building 3D structure models satisfying spatial constraints based on the alignment of a sequence with one or more known 3D structure sequence(s) used as template(s), before refining them by energy minimization (detailed steps in Figure 11). Tools like SWISSMODEL [64] automate all these steps, from template search to the proposal of refined 3D structure models.

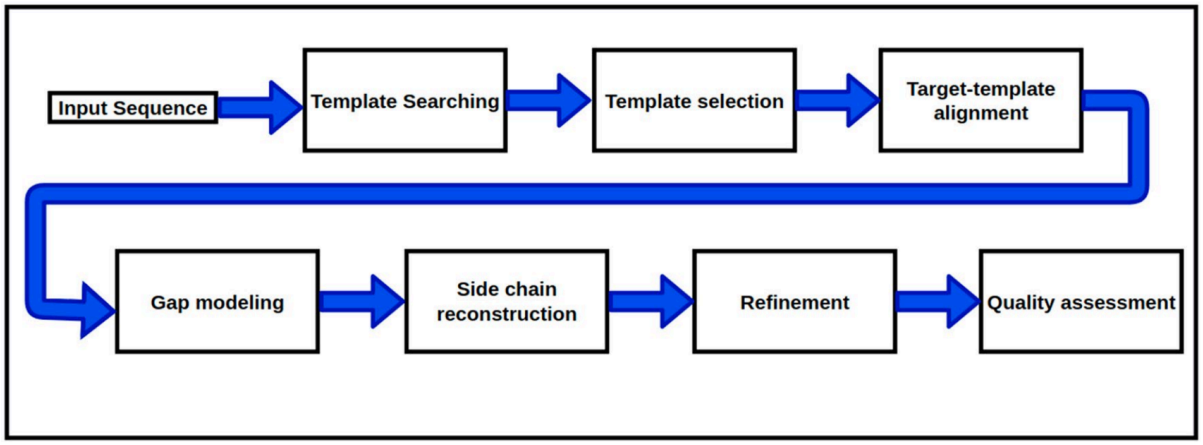



Figure 11: Typical homology-based modeling protocol. Figure taken from [61].

In recent years, deep learning methods based on evolutionary information have been recognized as the most effective for 3D structure predictions [65, 66] . Several predictors of this type can be mentioned: DMPFold [67], RoseTTAFold [68], and AlphaFold2 [69]. Below AlphaFold2 is presented in detail as its predictions were used in this thesis.

AlphaFold2, developed by DeepMind, revealed unprecedented levels of efficiency during the CASP14 competition, as can be seen in Figure 12.

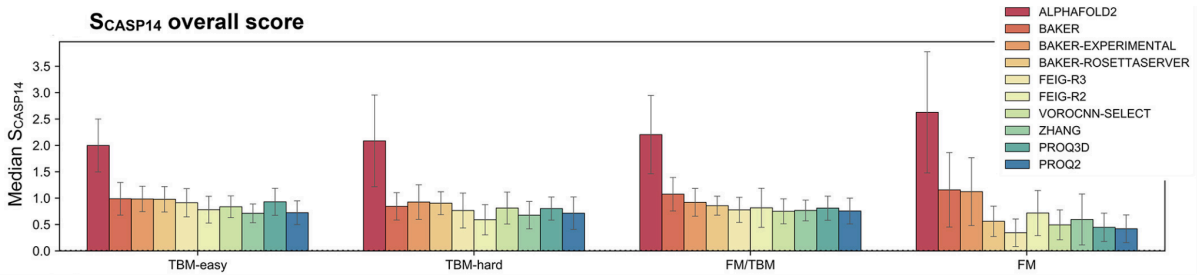



Figure 12:Barplots representing the ranking of the 10 methods that obtained the best scores (on the y-axis) during the CASP14 prediction competition. Each category on the x-axis corresponds to the difficulty level of the sequences to predict. TBM means Template Based Model. FM means Free Modelling, indicating that there were no homologs in the 3D structure databases. Figure taken from [70].

AlphaFold2 uses deep neural networks to predict the structure of proteins from their amino acid sequence. Its operation is outlined in Figure 13. Two representation modes are derived from the sequence: (i) a multiple alignment, constructed after searching the sequence in various sequence databases (BFD, MGnify, PDB, UniRef, and UniProt), identifying co-variation (co-evolution) motifs in this representation allows predicting contacts between each pair of residues; and (ii) a pairwise residue distance matrix, which can be completed by using structural information when available. These representation modes are used by the two modules composing AlphaFold2 to predict the protein structure. In a simplified way, the *evoformer* module refines the distance matrix and the multiple sequence alignment by confronting them with each other. These two representations are then used by the structure module to build the three-dimensional model of the protein.

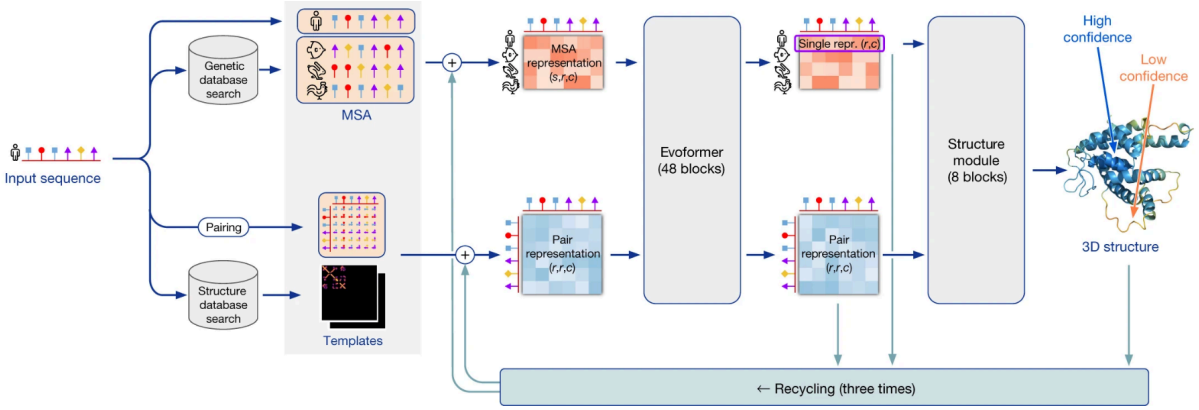



Figure 13: Alphafold model architecture. Arrows show the information flow among the various components. Array shapes are shown in parentheses with s, number of sequences; r, number of residues; c, number of channels. Image taken from [69].

The resulting 3D structure model is evaluated by a per-residue score, the "predicted local-distance difference test" (pLDDT), which is a prediction of the lDDT-Cα score [71], developed to evaluate the effectiveness of structure predictors by comparing predicted protein structures to experimentally determined structures. It is obtained by calculating the distances between the C$\alpha$ of neighboring residues in the experimental structure on one hand, and in the predicted structure on the other hand, then comparing these values. The ratio of conserved distances in the predicted structure compared to the experimental structure is calculated, which is the lDDT-Cα. The prediction of this value by AlphaFold2 is integrated into the neural network of the structure module that has been trained on good quality PDB structures (resolution between 0.1 and 3.0 Å, no NMR structures). A pLDDT greater than 90 corresponds to a very good quality prediction, a pLDDT between 70 and 90 to good quality, between 50 and 70 to low quality, and below 50 to very low quality. AlphaFold2 is very effective even for sequences that have no homologs in structural databases (scores for the "FM" category in Figure 12).

### 1.2.6 Molecular Dynamics Simulations

#### 1.2.6.1 Overview

The dynamic behavior of molecules at the atomistic level cannot be directly studied using current experimental methods. However, *in-silico* simulations provide a means to investigate the motion of atoms. These simulations can treat molecular behavior with varying levels of precision (Figure 14). In the quantum mechanics (QM) formalism, the motion of electrons is explicitly considered, and the system's energy is obtained by solving the Schrödinger equation [72]. While an analytical solution is only feasible for the hydrogen atom, numerical solutions for atoms with more electrons can be achieved using various algorithms. Nevertheless, these QM methods have a prohibitively high computational cost, limiting their applicability to relatively small systems and very short timescales, given the current state of computational

resources.

To overcome these limitations, the molecular mechanics (MM) formalism introduces approximations, such as the Born-Oppenheimer approximation, which decouples nuclei and electron motions. By neglecting electron motions, the energy can be treated as a function of nuclear coordinates [72]. In the MM framework, atoms are represented as spheres with constant volume, and bonds are modeled as springs. This simplification enables the study of larger systems, such as proteins or nucleic acids, at a more affordable computational cost and over increased timescales.

Within the MM framework, several approaches exist to describe molecular systems. All-atom simulations treat each atom in the system explicitly, providing the most precise description of interatomic interactions at this level of theory. To further reduce the computational cost, atoms can be grouped into coarse-grained representations, where multiple atoms are united to form larger entities. This approach sacrifices some level of detail but allows for the simulation of even larger systems and longer timescales [73]. All-atom model was used in the present work to simulate HLA proteins.

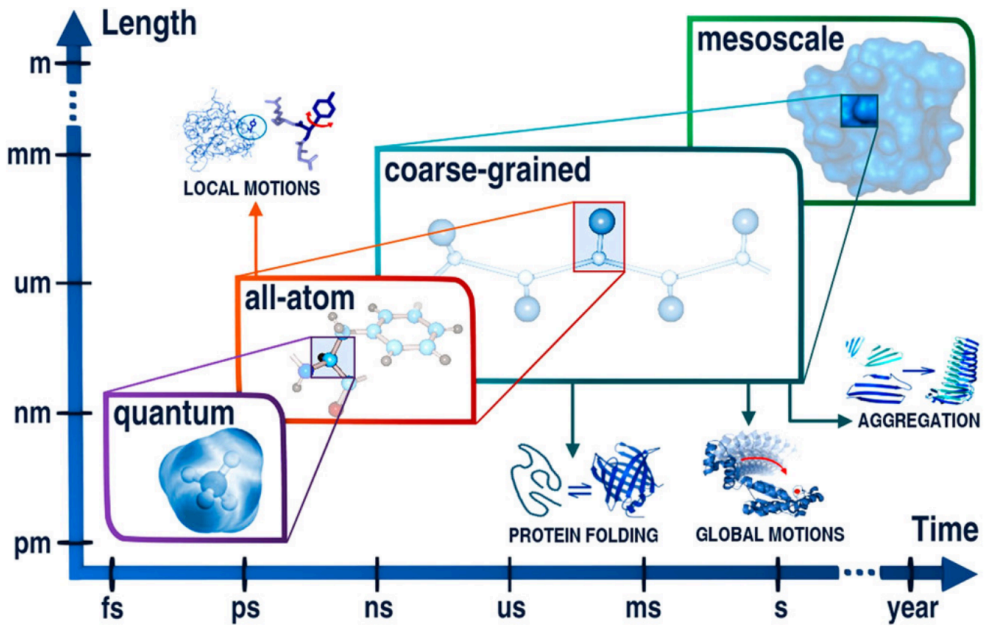



Figure 14: Ranges for molecular modeling, spanning from quantum to mesoscale, dictates the accessible time scales and system sizes. The plot illustrates the approximate ranges of these parameters for each resolution level. Image taken from [73].

### 1.2.6.2 Fundamentals of Molecular Dynamics simulations

Molecular Dynamics (MD) is a key methodology employed in this thesis work, enabling the simulation of the dynamic behavior of molecules. Since its initial applications in the late 1950s [74], MD has evolved alongside advancements in computational resources, becoming a standard approach for studying biological macromolecules, such as proteins and nucleic acids, over the past few decades [75].

The fundamental principle behind MD simulations is Newton's second law of motion [72]:

$$\vec{F}_i = m_i \cdot \vec{a}_i \qquad 1)$$

where Fi is the force applied to an atom i, mi is the mass of atom i, and ai is the acceleration of atom i.

Given that acceleration is the derivative of velocity (vi) with respect to time (t), and velocity is the derivative of position (ri) with respect to time, acceleration can be expressed as:

$$\vec{a}_i = \frac{\partial \vec{v}_i}{\partial t} = \frac{\partial^2 \vec{r}_i}{\partial t^2} \quad 2)$$

Consequently, the force applied to a particle can be related to its position through Newton's equation of motion:

$$\frac{\vec{F}_i}{m_i} = \frac{\partial^2 \vec{r}_i}{\partial t^2} \quad 3)$$

Thus, by knowing the force and mass of a particle, its position after a certain time step (δt) can be computed. Iterative integration of Equation 3 generates the trajectory, which is a sequential set of atomic positions in the system.

#### 1.2.6.3 Force field

To solve Equation 3, knowledge of the system's potential energy (U) is essential, as it allows for the computation of the force acting on each atom, as described in Equation 4:

$$\vec{F}_i = -\frac{\partial U}{\partial \vec{r}_i} \quad 4)$$

The force field (FF) is a potential energy function and a set of parameters used in this function. In empirical FFs, the total potential energy is the sum of potential energies of bonded ($U_{bonded}$) and non-bonded ($U_{non\text{-}bonded}$) interactions:

$$U_{total} = U_{bonded} + U_{non-bonded} \quad 5)$$

Each term can be further decomposed into a sum of simpler functions describing specific interactions. Bonded interactions (Eq. 6-8) include bonds between two consecutive atoms, angles between three consecutive atoms, dihedrals between four consecutive atoms, and improper dihedrals, which are described by three atoms attached to one central atom.

$$U_{bonded} = \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{impropers} K_\phi(\phi-\phi_0)^2 + \sum_{dihedral} \sum_{n=1}^{6} K_{\psi,n}(1+cos(n\psi-\delta_n)) \quad 6)$$

## 1.3 Basics of Data Science and Machine Learning

### 1.3.1 Data science definition and overall process

#### 1.3.1.1 The KDD (Knowledge Discovery from Data) process

Today, data science is a pivotal field for extracting meaningful insights from vast amounts of data, enabling informed decision-making and value creation across various domains. At its core, data science encompasses a spectrum of machine learning algorithms aimed at uncovering patterns, predicting trends,

and extracting actionable knowledge from data. Actually, the application of learning algorithms is just one step of a comprehensive KDD (Knowledge Discovery from Data) process for which several models were proposed. Figure 15 presents the model proposed by [76]. Hence, the process encompasses several stages contributing to the overall success of the process. The process is generally iterative as it permits backtracking to previous steps in order to explore various alternatives. Besides, the process is interactive as it is driven by a data analyst ideally accompanied by a domain expert or skilled with such expertise. The rest of this section is a brief description of the different steps.

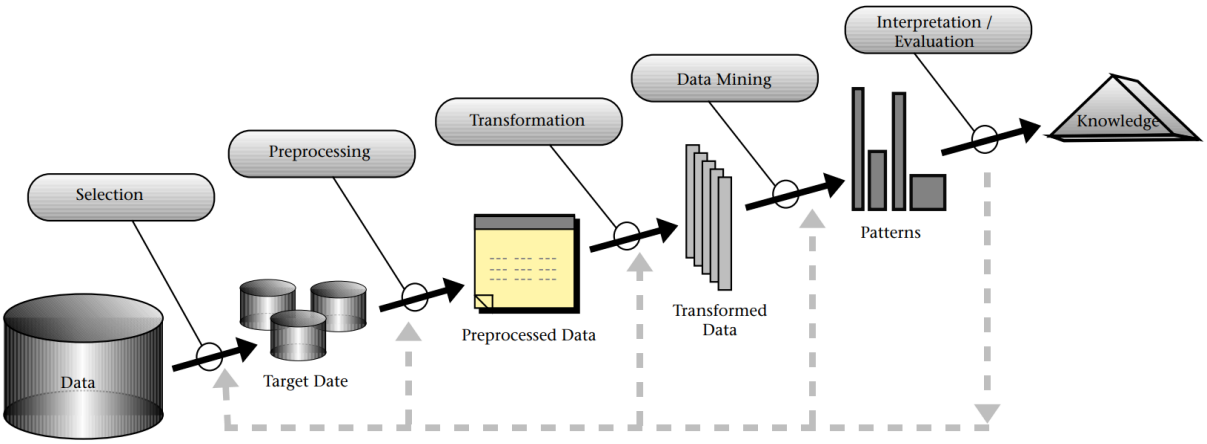



Figure 15: Knowledge Discovery from Data process model. Image taken from [76].

### 1.3.1.2 Data preparation: selection, preprocessing and transformation

The first phase of data preparation is the selection of the data source(s) to use given the task. This is followed by the necessary preprocessing of the data, which includes operations such as data cleaning (handling errors, correlated data, missing data, normalization...), feature construction, and feature selection. Preprocessing the data also involves adapting to the ML algorithm that will be applied using appropriate encodings. Data preparation may include other operations such as pre-computing distances (or similarities) for applying unsupervised ML algorithms or determining train/validation/test sets for supervised ML algorithms. Such dataset splitting must be performed upstream of any data preparation operation in order to avoid any data leakage.

Data preparation is certainly the most time-consuming part of the KDD process, but the rigor and creativity deployed therein determine the quality of the resulting models and knowledge.

Recent work in deep learning allows for the consideration of feature construction as a learning problem through the concept of representation learning which is based on the principle of automatically discovering and extracting meaningful representations or features from raw data. Unlike traditional machine learning approaches where feature engineering is often a manual and labor-intensive process, representation learning aims to learn latent representations directly from the data, thereby capturing intricate patterns and structures. The price to pay is the loss of the link between the learned features and the initial data.

### 1.3.1.3 Data mining or machine learning

This step is the core of the KDD process and corresponds to the application of an ML algorithm to the prepared data. There are several types of algorithms depending on the task to be performed:

- Classification algorithms: the task consists of predicting the class of an instance based on the

value of a set of characteristics of that instance.

- Clustering algorithms (or unsupervised classification): the task here is to group objects (or instances) into clusters so that each object in a cluster is close in terms of a similarity measure to the other objects in the cluster and is far from the objects in the other clusters.
- Pattern and association rule mining algorithms: the task in this case is to identify regularities in the data in the form of frequent itemsets and strong associations between itemsets.
- Outlier detection algorithms: the task is to identify objects that are far away from most other objects.

Several paradigms or types of learning exist. The best known are supervised learning and unsupervised learning. Indeed, a supervised classification algorithm requires a set of labeled data in order to build a model capable of predicting the class for new instances. Conversely, an algorithm is qualified as unsupervised if it does not require such a set of labeled data. Two examples: clustering algorithms and association rule mining algorithms.

Two other forms of learning have gained significant momentum in recent years: RL (Reinforcement Learning) and SSL (Self-Supervised Learning). RL is a learning method where an agent (which can be a robot or a chess player) learns to make the right decisions in a given environment by receiving rewards or penalties [77]. Finally, self-supervised learning consists of learning without resorting to supervision simply by masking a part of the input data and learning to correctly predict the missing part [78]. This form of learning is used in image or text analysis. These same principles have been applied to protein sequences to learn a latent representation of these sequences [79].

For supervised learning algorithms, the interpretable or non-interpretable nature of the produced model is an important differentiating factor and contributes to the explainability of the predictions made by the model.

Moreover, the task of predicting the class of an instance can be done directly by transduction or by induction of a general model and then by deduction (see Figure 16). The concept of transductive learning is closely related to semi-supervised learning. In this setting, specific training cases are used to directly make inferences on test cases. Unlike most common inductive methods, which have two stages (training and testing), transductive methods have only one stage, as shown in Figure 16. In the inductive scheme (a), the learning algorithm uses labeled training samples to fit a model. In the second stage, the fitted model is used to make predictions on unknown samples. In the transductive scheme (b), both labeled and unlabeled samples are processed together in a single stage. As a result, predictions are obtained for the unlabeled samples, but no decision function, trained model, or classification rule is returned as output. While an inductive method infers a decision function that can be used to predict labels of any sample, the transduction model directly estimates the set of labels for the unknown samples. However, a transductive classifier can be analyzed after training in a manner similar to an inductive classifier [80]. For example, in a graph-based method, the adjacencies among the samples labeled in the transductive stage could be used for knowledge extraction [81]. In the case of transduction, the prediction is referred to as instance-based prediction as opposed to model-based prediction.

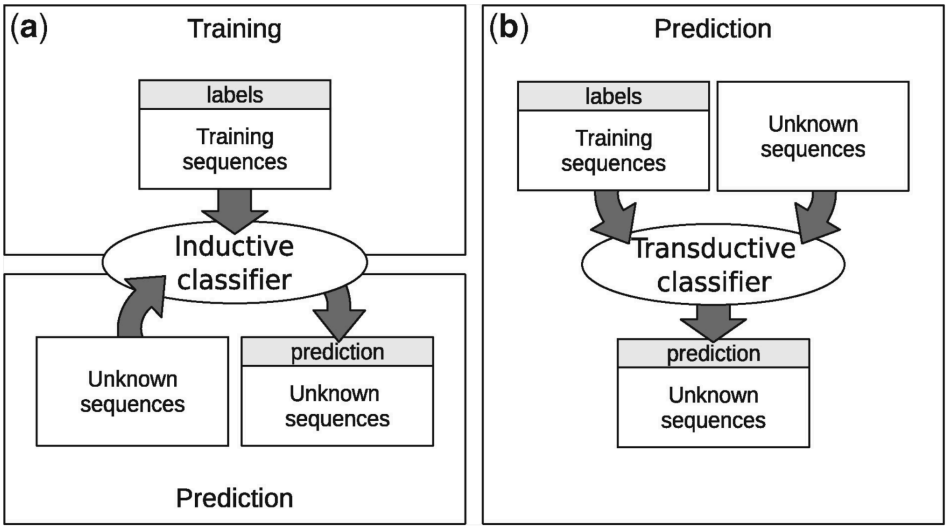



Figure 16: Inductive and transductive learning. (a) Traditional inductive scheme with two separate stages; (b) Transductive scheme with only one stage. Image taken from [80].

#### 1.3.1.4 Interpretation/Evaluation

The final step of the KDD process involves evaluating the results produced by the ML algorithm and interpreting these results. The evaluation protocols and metrics that can be used depend on the type of learning and the family to which the algorithm belongs. In favorable situations, an expert interpretation of the results can yield potentially useful knowledge for problem-solving.

### 1.3.2 Tree-based Predictive models and algorithms

This section describes a few supervised ML algorithms, focusing on tree-based models, such as decision trees and extremely randomized trees, as they are mainly used in the present thesis.

#### 1.3.2.1 Decision trees

Decision trees are a fundamental and widely used supervised learning algorithm in the field of machine learning. They are employed for both classification and regression tasks, with the objective of creating a model that predicts the value of a target variable based on a set of input variables. The structure of a decision tree consists of internal nodes, branches, and leaf nodes. Each internal node represents a test on an attribute, each branch denotes the outcome of the test, and each leaf node holds a class label or a numerical value in the case of regression [82] (see Figure 17).

The process of constructing a decision tree involves recursively partitioning the input space into subsets based on the most informative features. The selection of the best feature to split on is determined by a splitting criterion, which aims to maximize the homogeneity or purity of the resulting subsets with respect to the target variable [83]. Common splitting criteria for classification trees include information gain, gain ratio, and Gini impurity, while for regression trees, the reduction in variance or mean squared error is often used [84]. The recursive partitioning continues until a stopping criterion is met, such as reaching a maximum depth, a minimum number of instances per leaf, or when further splitting does not yield a significant improvement in the model's performance [82]. Pruning techniques, such as pre-pruning or post-pruning, can be applied to reduce overfitting and improve the generalization ability of the decision tree [85].

One of the main advantages of decision trees is their interpretability. The hierarchical structure of the tree allows for easy visualization and understanding of the decision-making process (see Figure 17). Additionally, decision trees can handle both categorical and numerical data and are robust to outliers and missing values [83].

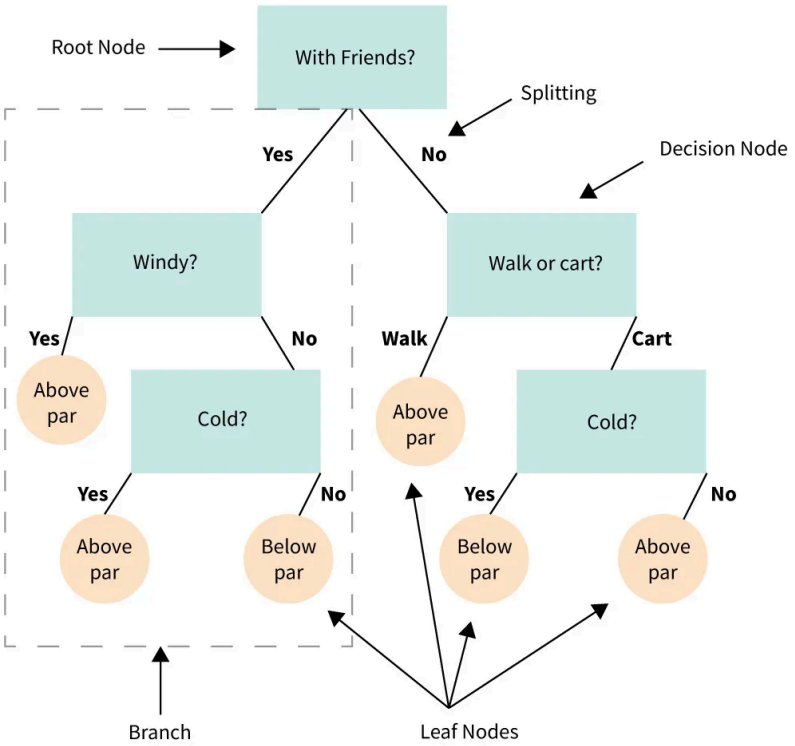



Figure 17: Structure of a decision tree. Image taken from [86].

However, decision trees also have some limitations. They are prone to overfitting, especially when the tree becomes deep and complex. Moreover, decision trees can be sensitive to small variations in the training data, leading to instability [87]. To address these issues, ensemble methods such as random forests [88], extremely randomized trees [89] and gradient boosting [90] have been developed, which combine multiple decision trees to improve predictive performance and robustness.

#### 1.3.2.2 Extremely randomized trees

Extremely randomized trees, also known as ExtraTrees, are an ensemble learning method that extends the concept of random forests [89]. This algorithm combines the principles of bagging [87] and random subspace method [91] to create a collection of decision trees, which are then aggregated to make predictions. The key difference between ExtraTrees and random forests lies in the way the trees are constructed and the splitting of nodes is determined .

In the ExtraTrees algorithm, each tree is built using the entire training dataset, rather than a bootstrap sample as in random forests. This means that all trees in the ensemble have access to the same training instances, reducing the variance in the model. Additionally, when selecting the best split at each node, ExtraTrees introduces further randomization compared to random forests [89].

While random forests consider a random subset of features at each node and choose the best split among them based on a splitting criterion (e.g., Gini impurity or information gain), ExtraTrees takes the randomization one step further. For each feature in the random subset, ExtraTrees selects a random split point. The best split among these randomly generated splits is then chosen according to the splitting criterion. This extra level of randomization helps to reduce the variance of the model even further and can lead to improved generalization performance [89].

One advantage of ExtraTrees is its computational efficiency. By randomly selecting both the features and split points, the algorithm avoids the need to exhaustively search for the best split at each node. This can lead to faster training times, especially when dealing with high-dimensional datasets. Moreover, the increased randomization in ExtraTrees can help to reduce overfitting and improve the model's ability to generalize to unseen data [89].

ExtraTrees have been successfully applied to various machine learning tasks, including classification, regression, and feature selection [89, 92]. They have shown competitive performance compared to other ensemble methods, such as random forests and gradient boosting machines, in terms of performance and robustness [93].

### 1.3.3 Evaluation procedures and metrics

#### 1.3.3.1 Confusion matrix

Confusion matrix is a key tool for evaluating the performance of classification models in machine learning. It provides a tabular summary of the model's predictions compared to the actual class labels, enabling a detailed analysis of the model's successes and failures. The confusion matrix is particularly useful for binary classification problems but can be extended to multi-class classification as well [94].

In a binary classification setting, the confusion matrix categorizes the model's predictions into four possible outcomes: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From the matrix, various evaluation metrics can be derived, such as accuracy, precision, recall, and F1-score among others [95] (see Figure 18). However, accuracy alone may not be sufficient, especially when dealing with imbalanced datasets [96].

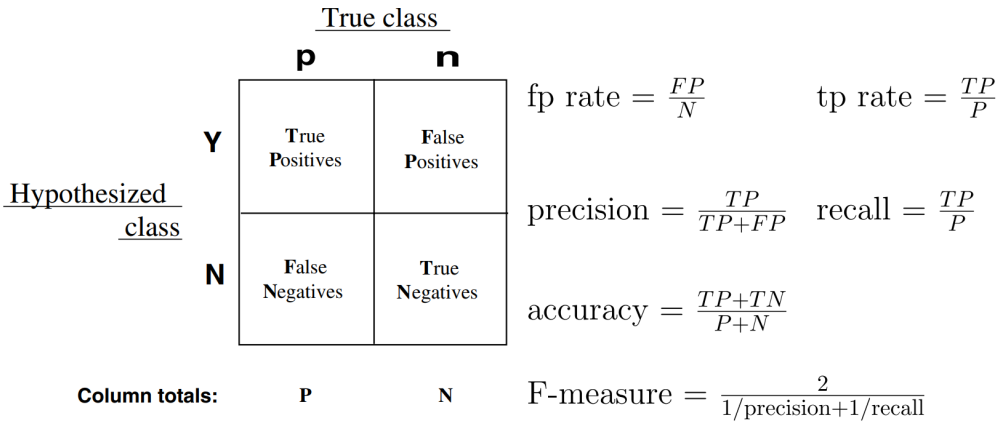



Figure 18: Confusion matrix and common performance metrics calculated from it. Image taken from [94].

When interpreting the confusion matrix, it is essential to consider the context and the costs associated with different types of errors [97]. The confusion matrix allows for a nuanced understanding of the model's performance.

#### 1.3.3.2 Accuracy, Precision, Recall, F1 score

Accuracy, precision, recall, and F1 score are essential evaluation metrics in machine learning, particularly for classification tasks. These metrics provide an assessment of a model's performance, each focusing on different aspects of the model's predictions compared to the actual class labels. Understanding these metrics is crucial for selecting and optimizing machine learning models based on the specific requirements of the application [95].

Accuracy is the most intuitive metric, measuring the overall correctness of the model's predictions. It is calculated as the ratio of correct predictions to the total number of predictions, expressed as (TP + TN) / (TP + TN + FP + FN) (see Figure 18). While accuracy provides a general overview of the model's performance, it can be misleading when dealing with imbalanced datasets, where the distribution of classes is skewed [96].

Precision, also known as positive predictive value, measures the proportion of true positive predictions among all instances predicted as positive. It is calculated as TP / (TP + FP) (see Figure 18). High precision indicates that when the model predicts an instance as positive, it is highly likely to be

correct. Precision is particularly important in applications where false positives have a high cost [95].

Recall, also known as sensitivity or true positive rate, measures the model's ability to identify positive instances correctly. It is calculated as TP / (TP + FN) (see Figure 18). High recall indicates that the model successfully captures a large portion of the positive instances in the dataset. Recall is crucial in applications where false negatives have a high cost [95].

The F1-score (also known as F-score or F-measure) is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. It is calculated as 2 * (precision * recall) / (precision + recall) (see Figure 18). The F1-score is particularly useful when the dataset has an imbalance class distribution and when both false positives and false negatives are considered equally important. A high F1-score indicates that the model has a good balance between precision and recall [95].

It is important to note that there is often a trade-off between precision and recall. Increasing precision may lead to a decrease in recall, and vice versa. In some cases, a custom metric can be used to assign different weights to precision and recall based on their relative importance [98].

#### 1.3.3.3 AUC-ROC

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a widely used evaluation metric for binary classification problems in machine learning. The AUC-ROC provides a single scalar value that summarizes the performance of a classifier across all possible classification thresholds, making it a valuable tool for comparing different models and assessing their overall discriminative power. The AUC-ROC is derived from the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds (see Figure 18). The TPR, also known as recall or sensitivity, measures the proportion of correctly classified positive instances, while the FPR measures the proportion of negative instances incorrectly classified as positive [94] (see Figure 19).

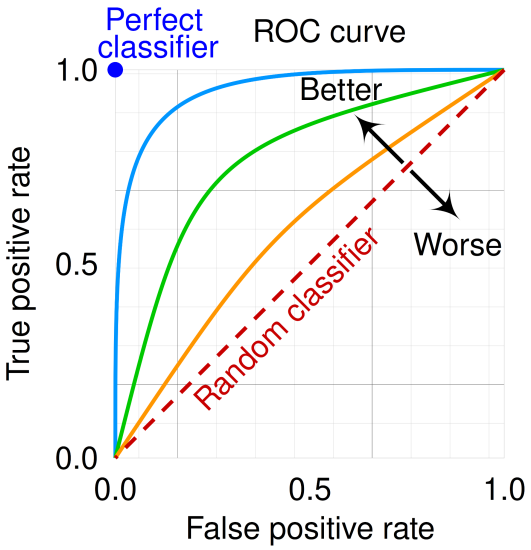



Figure 19: Receiver Operating Characteristic (ROC) curve with TPR and FPR. A diagonal shows the performance of a random classifier. 3 curved lines from (0, 0) to (1, 1) that get progressively closer to (0,1) show improving classifiers[10].

The AUC-ROC is calculated by integrating the area under the ROC curve. An AUC-ROC value of 1.0 indicates a perfect classifier that correctly ranks all positive instances above negative instances, while a value of 0.5 corresponds to a random classifier that performs no better than chance. In practice, AUC-ROC values between 0.7 and 0.8 are considered acceptable, values between 0.8 and 0.9 are considered excellent, and values above 0.9 are considered outstanding.

[10] Image taken from https://commons.wikimedia.org/wiki/File:Roc_curve.svg

One of the key advantages of the AUC-ROC metric is its ability to evaluate the classifier's performance independently of the classification threshold. This makes it particularly useful when comparing classifiers in situations where the optimal classification threshold is unknown or when the cost of false positives and false negatives is not well-defined. Additionally, the AUC-ROC is robust to class imbalance, as it does not depend on the relative proportions of positive and negative instances in the dataset [99].

#### 1.3.3.4 AUC-PR

The Area Under the Precision-Recall Curve (AUC-PR) is an evaluation metric used in machine learning, particularly for binary classification problems with imbalanced datasets [96] (see Figure 20). The AUC-PR summarizes a classifier's performance by considering the trade-off between precision and recall at various classification thresholds, providing a single scalar value that reflects the classifier's ability to correctly identify positive instances while minimizing false positives. The AUC-PR is calculated by integrating the area under the PR curve. An AUC-PR value of 1.0 indicates a perfect classifier that achieves both high precision and high recall, while a value equal to the prevalence of the positive class in the dataset corresponds to a random classifier [100].

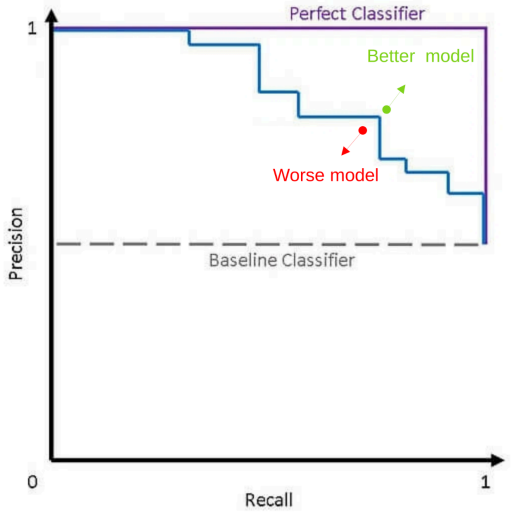



Figure 20: Precision-Recall (PR) curve. The horizontal line shows the performance of a random classifier. The line from (0, 1) to (1, 0) that gets closer to (1,1) shows a better performance[11].

The AUC-PR is particularly useful when evaluating classifiers on imbalanced datasets, where the number of positive instances is significantly smaller than the number of negative instances . In such cases, the AUC-ROC metric may provide an overly optimistic assessment of the classifier's performance, as it is less sensitive to changes in the number of false positives. By focusing on precision and recall, the AUC-PR provides a more informative measure of the classifier's ability to correctly identify positive instances in the presence of class imbalance [100].

#### 1.3.3.5 MCC

The Matthews correlation coefficient (MCC) is a performance metric used in machine learning to evaluate the quality of binary classification models [101]. The MCC takes into account all four elements of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). It provides a balanced measure of a classifier's performance, even when the classes are of different sizes [102].

The MCC is calculated using the following formula:

[11] Image adapted from https://miro.medium.com/v2/resize:fit:720/format:webp/1*6QPLsDvjo4H6OZrxEBI8Fg.png

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}} \quad 7)$$

The MCC ranges from -1 to +1, where +1 represents a perfect classifier, 0 indicates a classifier that performs no better than random guessing, and -1 corresponds to a classifier that always makes incorrect predictions. One of the main advantages of the MCC is its robustness to class imbalance [ref]. Unlike accuracy, which can be misleading when the classes are not equally represented, the MCC provides a more reliable assessment of the classifier's performance. It considers both the correctly classified instances (TP and TN) and the misclassified instances (FP and FN), giving a comprehensive view of the classifier's ability to discriminate between classes [102].

Furthermore, the MCC is a single scalar value that summarizes the confusion matrix, making it easy to compare different classifiers or to track a classifier's performance over time. However, it is important to note that the MCC should be used in conjunction with other metrics, such as precision, recall, and F1 score, to gain a complete understanding of a classifier's performance [102].

# Chapter 2

# Contribution of molecular dynamics simulations to the exploration of structural properties of HLA antigens

## Summary



This chapter presents various results calculated from MD simulations performed on our set of 207 HLA antigens. This set of chosen HLA antigens include the antigens tested in LSA flow bead assays, i. e. antigens from the loci A, B, C, DP, DQ, and DR (see Table 4 for an overview and Table S1 for the complete list of modeled antigens). Firstly the motivation for the use of molecular dynamics simulations is presented. Then, the protocol implemented for the simulation of HLA molecules is described. The results of the quality checks of the predicted (from Alphafold) and repaired (from the PDB) structures as well as the MD trajectories are shown. Finally, a structural characterization of the HLA molecules from the MD data will be performed. This characterization will cover properties such as solvent accessibility, physicochemical features and side-chain flexibility at AA and patch level. In addition, it will distinguish on

the one hand between conserved and polymorphic sequence positions, and on the other hand between positions that are part of confirmed or non confirmed eplets, or not registered as eplets in HLA Eplet Registry [34].

Table 4: Number of modeled antigens per locus. The number of distinct allele sequences is also indicated.

| Locus | # of modeled antigens | # of distinct allele sequences |
|---|---|---|
| A | 34 | 34 A |
| B | 59 | 59 B |
| C | 17 | 17 C |
| DP | 31 | 7 DPA1 and 19 DPB1 |
| DQ | 30 | 13 DQA1 and 14 DQB1 |
| DR | 36 | 19 DRB1, 3 DRB3, 2 DRB4 and 2 DRB5 |

## 2.1. Motivation for the use of molecular dynamics simulations

The initial interest in MD simulations relies on their ability to refine the 3D structures of proteins, especially to correct stereochemical errors. This technique has proven its usefulness by being integrated into the protocols for predicting the 3D structure of proteins [103–107] although it has also been used in the refinement of structures obtained experimentally by X-ray diffraction [108, 109]. Secondly, MD simulations allow to study the dynamic properties of proteins such as side-chain flexibility and solvent accessibility. A recent study suggests that side-chain flexibility is a key element in antigen-antibody recognition [110]. Additionally, MD simulations give access to conformations that are closer to those that occur under ambient conditions. Indeed, the structures obtained by X-ray diffraction present the protein in a crystallized conformation, which is not necessarily exactly the same under ambient conditions. This also applies to the structures predicted by tools such as AlphaFold v2 [69], as they have all been trained using structures from the PDB [48], which mostly consist of X-ray diffraction-resolved structures.

## 2.2. Molecular dynamics simulation protocol

Two procedures were implemented to generate the initial structures of the HLA antigens to run the MD simulations. One procedure corresponds to the antigens for which a structure exists in the PDB and the second one corresponds to antigens lacking a PDB structure (see Figure 21).

### 2.2.1. Identification of crystallographic structures in the PDB

The « ncbiblast » API [111] was used to find structures in the PDB using the HLA sequences in the Immuno Polymorphism Database (also known as IPD-IMGT/HLA database) [15]. Then, the availability of the structure in PDB_REDO [54] is checked, otherwise, the structure from the PDB[12] is downloaded. The resolution values of the structures were obtained from the pdbe/EBI API [112] to finally choose a single template per antigen according to the best resolution.

[12] The structures were collected on June 23rd, 2020.

### 2.2.2. Refinement of previously identified PDB structures

PDBFixer [113] was used to identify and correct missing atoms or amino acids, remove small molecules and groove peptides. Pymol [114] was then used to remove signal peptides (if present) and trim the protein to keep only the extracellular globular part (see Table S2 for details). Loops and side chains were refined with MODELLER [115]. The Visual Molecular Dynamics (VMD) plugins CHIRALITY and CISPEPTIDE [59] were then used to check for chirality and cis-peptide bond errors. In cases where cis peptide bond errors not involving prolines were identified, structures were generated for hydrogen and oxygen rotations, then MD simulations were performed for both cases and an additional minimization stage was run to choose the conformation that minimized the energy of the system. Otherwise, a single MD simulation was run on the structure (see Figure 21).

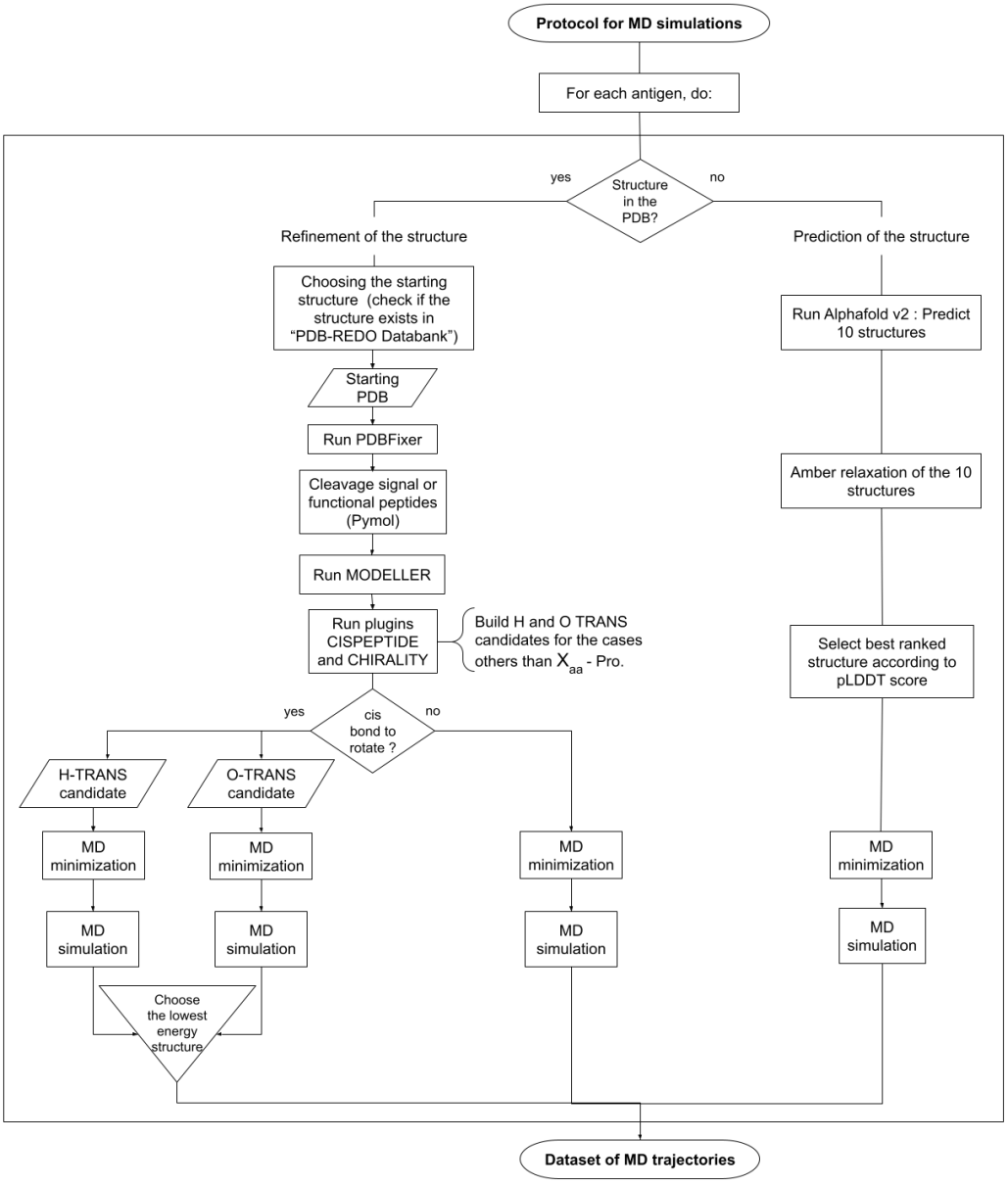



Figure 21: Protocol for MD simulations. There are two procedures, depending on whether the initial structure used to run the MD simulation comes from the PDB or from Alphafold.

### 2.2.3. Generation of structures for antigens lacking any structure in the PDB

A local instance of Alphafold v2.1.0 was used to generate the structures of antigens absent from the PDB. Default parameters were utilized. Ten structures were generated per antigen to which an Amber relaxation was applied and a single structure was chosen based on the best confidence score (pLDDT).

### 2.2.4. Molecular Dynamics simulations

MD simulation was performed for each antigen using the VMD [116] and NAMD3 [117] tools. For the generation of the system topology, the VMD AutoPSF tool was used together with the CHARMM36 force field [118]. The system was solvated using the TIP3P explicit solvent model and then neutralized using the VMD AutoIonize tool. The Particle Mesh Ewald (PME) method was used to calculate the electrostatic energy with a distance truncation of 11 Å. The simulation was carried out under NPT conditions, that is, constant pressure (1 atm) and temperature (310 K) and was composed of 2 stages: (i) a minimization of 150.000 steps, in which the system was brought into room temperature (310K), (ii) system equilibration of 10ns. The 500 frames from the last 5ns of simulations were used for subsequent analyses. It is worth mentioning that short MD simulations were performed because as demonstrated in [110], side-chain motions are fast enough to be studied through short simulations. See Supplementary materials for data availability.

## 2.3. Structure quality check of the predicted/refined structures and simulations

### 2.3.1. Structure quality check of the predicted/refined structures

To assess the quality of the structures used as a starting point for the MD simulations, a comparative analysis was conducted focusing on the quantity of stereochemical errors in the initial structures and those obtained after the MD minimization step. The WHATCHECK tool [119] available through the SAVES website [120] and the CISPEPTIDE plugin of VMD were used. Through WHATCHECK 3 types of errors were checked: Chirality (the presence of D-configurations), Van der Waals overlaps (also known as clashes) and irregular side chain flips (NQH flips). The CISPEPTIDE plugin was used to check Cis peptide bonds. Table 5 summarizes the results obtained per locus and across all antigens while Figure S1 shows the results per antigen. Remarkably, all chirality and clash errors were successfully resolved post MD minimization (originally 2 chirality errors and 2368 clashes). Notably, NQH flips saw a significant decrease, from 604 to 264. However, improvements in Cis bond errors were minimal, reducing only from 20 to 16. Indeed, in 7 cases, at least one CIS bond persisted from the initial structure (A*03:01, B*40:01, B*44:03, DPB1*28:01-DPA1*04:01, DQB1*02:01-DQA1*05:08, DQB1*06:02-DQA1*01:02 and DRB1*04:04). Intriguingly, 6 cases (DQB1*02:01-DQA1*03:01, DQB1*03:02-DQA1*03:01, DRB1*04:05, DRB1*15:01, DRB3*01:01, DRB3*03:01) exhibited new CIS bonds post MD minimization. Detailed examination revealed that these CIS bonds, initially corrected by the CISPEPTIDE plugin, reemerged after MD minimization. This suggests that such configurations might be energetically favorable considering the stereochemical constraints imposed by the molecular environment in which each CIS peptide bond is found.

Table 5: Stereochemical errors in HLA antigens for the initial structures and those obtained after MD minimization. Number of concerned antigens is in parentheses.

| | Initial structures | | | | Post MD Minimization | | | |
|---|---|---|---|---|---|---|---|---|
| | **Chirality** | **Clashes** | **NQH flips** | **Cis bond** | **Chirality** | **Clashes** | **NQH flips** | **Cis bond** |
| **A (34)** | 0 | 367 (34) | 65 (23) | 1 | 0 | 0 | 41 (26) | 1 |
| **B (59)** | 0 | 659 (59) | 178 (45) | 5 (4) | 0 | 0 | 75 (45) | 2 (2) |
| **C (17)** | 0 | 214 (17) | 44 (16) | 0 | 0 | 0 | 16 (13) | 0 |
| **DP (31)** | 0 | 324 (31) | 60 (28) | 1 | 0 | 0 | 51 (23) | 1 |
| **DQ (30)** | 1 | 335 (30) | 121 (29) | 9 (3) | 0 | 0 | 38 (21) | 6 (4) |
| **DR (36)** | 1 | 469 (36) | 136 (34) | 4 (2) | 0 | 0 | 43 (25) | 6 (5) |
| **ALL** | 2 | 2368 | 604 | 20 | 0 | 0 | 264 | 16 |

### 2.3.2. Structural equilibration along molecular dynamics trajectories

The coefficient of variation (CV, Equation 1) indicates the degree of variability of a set of values expressed as the ratio of the standard deviation (σ) to the absolute value of the mean (μ)

$$\mathcal{CV} = \frac{\sigma}{|\mu|} \quad 8)$$

This value is particularly useful in the context of MD simulations as it allows for the evaluation of the stability of a structure along the MD trajectory. Based on a visual inspection of the Root Mean Square Distance (RMSD) plots of MD trajectories (see data availability in supplementary materials), it was observed that most trajectories exhibited reasonable stability after 5ns of simulation. To verify this perception, the CV for the entire trajectory (referred to as "global" CV), the first 5ns of simulation (referred to as "0-5ns" CV), and the last 5ns of simulation (referred to as "5-10ns" CV) were computed for each antigen. Results were aggregated for all antigens at a certain locus as the mean value ± standard deviation and are presented in Table 6. They indicate a reduced structural variability for the last 5ns of simulations ("5-10ns") across all the groups considered, which confirms what was perceived during the visual inspection (Figure S2 displays the results per antigen). In the rest of the thesis, all MD data analyses will be performed on the last 5 ns of the simulations.

Table 6: Mean and standard deviation (std) of CV values aggregated by locus and across all the antigens for the 3 simulation intervals considered, i. e. "global" for the full 10ns trajectory, "0-5ns" for the first 5ns of the trajectory and "5-10ns" for the last 5ns of the trajectory.

| | CV | | |
|---|---|---|---|
| | **global (mean ± std)** | **0-5ns (mean ± std)** | **5-10ns (mean ± std)** |
| **A** | $0.133 \pm 0.038$ | $0.134 \pm 0.035$ | $0.084 \pm 0.033$ |
| **B** | $0.128 \pm 0.044$ | $0.124 \pm 0.034$ | $0.076 \pm 0.027$ |
| **C** | $0.110 \pm 0.027$ | $0.117 \pm 0.038$ | $0.070 \pm 0.018$ |
| **DP** | $0.110 \pm 0.031$ | $0.103 \pm 0.025$ | $0.065 \pm 0.018$ |
| **DQ** | $0.120 \pm 0.029$ | $0.112 \pm 0.026$ | $0.069 \pm 0.019$ |
| **DR** | $0.135 \pm 0.060$ | $0.139 \pm 0.060$ | $0.073 \pm 0.035$ |
| **ALL** | $\mathbf{0.125 \pm 0.043}$ | $\mathbf{0.123 \pm 0.040}$ | $\mathbf{0.074 \pm 0.028}$ |

## 2.4. Exploration of the molecular dynamics data at the amino acid level

### 2.4.1. Solvent accessibility

#### 2.4.1.1. Overview

Solvent accessibility of AAs plays a crucial role in antigen-antibody recognition, as only surface-exposed AAs can interact directly with antibodies. The RSASA ("Relative Solvent Accessible Surface Area") measure is widely used to assess AA accessibility (see above, Chapter 1). While no definitive threshold exists, a 20% RSASA value is broadly accepted in scientific literature as a threshold above which AAs are said to be accessible to the solvent [110, 121]. Utilizing this threshold offers valuable insights into the prevalence of solvent-accessible AAs, particularly in relation to specific attributes such as conserved and polymorphic AA positions, as well as AAs members of confirmed or non-confirmed eplets, or not member of any eplet. However, in MD simulations, solvent accessibility can fluctuate over time, to the extent that an AA may oscillate around the accessibility threshold during a simulation. Therefore, in this work, an AA is deemed solvent-accessible if its median RSASA throughout the MD trajectory is equal to or greater than 20%. Table 7 offers a summary of the percentage of solvent-accessible AA positions per locus and considering each chain (α and β) separately, while Figure S3 provides the results per antigen. Solvent accessibility was calculated from the MD trajectories and from the static 3D structures (the ones issued from the refinement process and corresponding to the initial frame of MD simulations) to make a comparison between the two and elucidate the contribution of MD data. The SASA calculator implemented in VMD [122] was utilized to compute the absolute solvent accessibility of residues and the theoretical « ALL » maximum solvent accessibility of each residue defined in [123] (see Table S3) was used to calculate RSASA values for each frame in the trajectory.

When considering MD data, all chains have a similar percentage of solvent accessible AA positions varying between 53% and 56%. Similarly, when considering static 3D data, all chains have a similar percentage of solvent accessible AA positions but the range of values varies between 48% and 51%. Interestingly, the percentages derived from MD data are always higher than those of the static 3D structure, revealing that MD simulation reflects a larger variety of solvent-accessible residues than static

3D structures. Also, when looking at the results by antigen (see Figure S3D, S3E and S3F), it appears that differences between MD and static data can vary greatly depending on HLA antigens considered, especially in class II loci DP, DQ and DR.

Table 7: Counts and percentages of solvent accessible AAs (RSASA threshold value: 20%) in HLA antigens for MD trajectories and static 3D structures aggregated by chain for each locus . The mean μ and standard deviation σ are computed over all antigens of a given locus.

| | Locus | | A | B | C | DP | DQ | DR |
|---|---|---|---|---|---|---|---|---|
| | Total AAs in chain α | | 276 | 276 | 276 | 183 | 186 | 182 |
| | Total AAs in chain β | | 99 | 99 | 99 | 191 | 192 | 192 |
| MD | Chain α | Surface AAs (μ ± σ) | 152.7 ± 4 | 151.2 ± 3.5 | 154.7 ± 3.4 | 97.8 ± 2.2 | 98 ± 2.7 | 98.7 ± 3.1 |
| | | % total AAs | 55.4 | 54.8 | 56.1 | 53.4 | 52.7 | 54.2 |
| | Chain β | Surface AAs (μ ± σ) | 54.4 ± 1.4 | 53.8 ± 1.6 | 53.2 ± 1.4 | 107.6 ± 3.4 | 107.7 ± 3.4 | 106.8 ± 3.1 |
| | | % total AAs | 55 | 54.4 | 53.8 | 56.4 | 56.1 | 55.6 |
| Static 3D | Chain α | Surface AAs (μ ± σ) | 141.7 ± 2.8 | 138.7 ± 3.2 | 141.7 ± 2.7 | 91.1 ± 1.6 | 92.8 ± 1.8 | 87.2 ± 2.3 |
| | | % total AAs | 51.4 | 50.3 | 51.4 | 49.8 | 49.9 | 48 |
| | Chain β | Surface AAs (μ ± σ) | 49.2 ± 1.2 | 49.3 ± 1.2 | 49 ± 0.5 | 99.3 ± 2.1 | 98.6 ± 2.3 | 96.3 ± 2.4 |
| | | % total AAs | 49.7 | 49.8 | 49.5 | 52 | 51.4 | 50.2 |

### 2.4.1.2. Solvent accessibility and polymorphic AA positions

Polymorphic AA positions are at the heart of HLA matching between donor and recipient and it is therefore important to have an insight into their prevalence among solvent-accessible AAs. Table 8 gives an overview of the number and percentage of polymorphic AA positions in the HLA antigen (i.e. two-field allele) sequences and among the solvent-accessible AAs, computed either from MD trajectories or from static 3D structures (the refined ones corresponding to the first frame of MD trajectories). The values are aggregated for all antigen structures grouped by gene. To complete, Figure S4 provides the count for polymorphic and conserved surface positions for each antigen individually. It is important to note that the $\beta$ chain of antigens from loci A, B and C is $\beta$2-microglobulin, which is completely conserved in all antigens from these loci. In addition, the $\alpha$ chain of the antigens from locus DR is also conserved in all the antigens of this locus (allele DRA1*01:01). For this reason, these genes are not considered in Table 8.

From Table 8, it appears that both genes of locus DP present a remarkably low percentage of polymorphisms at the sequence level (7.1 % and 9.9 % for DPA1 and DPB, respectively) and also among solvent-accessible AAs (8.9 % and 7.4 % in MD data and 9.1 % and 7.7 % in static data). On the contrary, the DQA1 gene displays a huge amount of polymorphisms (76.3% of sequence data and 79.6 % or 76.4 %

of solvent-accessible AAs for MD or static data, respectively). The second highest percentage of polymorphic positions is found for genes DRB1 to DRB5 (30.2% of sequence data and 30.5 % or 30.6 % of solvent-accessible AAs for MD or static data, respectively). Genes A, B and DQB1 occupy the 4th, 3rd and 5th rank, respectively, with respect to sequence data (21.7 %, 22.1 % and 20.8%), the 3rd, 4th, and 5th rank with respect to MD data (26.3 %, 21.0 % and 20.9 %) and the 3rd, 5th and 4th rank with respect to static data (25.9 %, 20.0 % and 21.1%). Altogether, when comparing MD data against static 3D data, we observe a rather similar prevalence of polymorphic positions.

When comparing MD data against static 3D data, an average difference of ~4% in the percentage of surface AAs is observed in both polymorphic and conserved positions. These percentages are always higher for MD data. The A, B and DRB genes show the largest differences between the percentages derived from static and MD data (5.8, 6.7 and 5.8 respectively), and the DPA1 and DPB1 the smallest differences (2.4 and 1.8 respectively). Altogether, the increase in percentage of surface AAs among polymorphic AA provided by MD data may appear small. However, it is worth remembering that the sensitivity of the immune system is such that even a single polymorphic AA can be sufficient to cause recognition by an antibody and thus trigger an immune response. Thus, MD data can reveal the role of additional AA(s) to the one or the few involved in the targeted eplet, in the formation of the full epitope, a role that was not detected from the static 3D structure.

In summary, these results are in line with the current knowledge about the differences in AA polymorphism observed in the allele sequences of various HLA genes. They just reveal that no specific enrichment in polymorphic positions is observed in the subgroups of solvent-accessible AAs for each gene type. The differences between HLA genes, known at the sequence level, also exist in a similar range at the level of surface AAs.

Table 8: Count and percentage of polymorphic positions across HLA genes for all antigens considered in this study and relative to AA sequence present in the 3D structures (Seq data), solvent-accessible ("surface") AAs computed from MD trajectories (MD) and (3) surface AAs computed from static 3D structures (Static 3D).

| | Gene | A | B | C | DPA1 | DPB1 | DQA1 | DQB1 | DRB1/3/4/5 |
|---|---|---|---|---|---|---|---|---|---|
| **Seq data** | **Total AAs present in 3D structures** | 276 | 276 | 276 | 183 | 191 | 186 | 192 | 192 |
| | **Polymorphic positions** | 60 | 61 | 45 | 13 | 19 | 142 | 40 | 58 |
| | **% of polymorphic AAs in 3D structure** | 21.7 | 22.1 | 16.3 | 7.1 | 9.9 | 76.3 | 20.8 | 30.2 |
| **MD** | **Total Surface AAs ($\mu \pm \sigma$)** | 152.7 ± 4 | 151.2 ± 3.5 | 154.7 ± 3.4 | 97.8 ± 22 | 107.6 ±3.4 | 98 ±2.7 | 107.7 ± 3.4 | 98.6 ±3.1 |
| | **Surface polymorphic positions ($\mu \pm \sigma$)** | 40.2 ± 2.2 | 31.8 ± 1.8 | 27.1 ± 2.2 | 8.7 ± 0.5 | 8 ± 1 | 78 ± 1.9 | 22.5 ± 1 | 30.1 ± 1.6 |
| | **% of polymorphic surface AAs** | 26.3 | 21.0 | 17.5 | 8.9 | 7.4 | 79.6 | 20.9 | 30.5 |
| **Static 3D** | **Total Surface AAs ($\mu \pm \sigma$)** | 141.7 ± 2.8 | 138.7 ± 3.2 | 141.7 ± 2.7 | 91.1 ± 1.6 | 99.3 ± 2.1 | 92.8 ± 1.8 | 98.6 ± 2.3 | 87.2 ± 2.3 |
| | **Surface polymorphic positions ($\mu \pm \sigma$)** | 36.7 ± 1.3 | 27.7 ± 1.4 | 25.4 ± 1.3 | 8.3 ± 0.4 | 7.6 ± 0.8 | 74 ± 1.7 | 20.8 ± 1.2 | 26.7 ± 1.1 |
| | **% of polymorphic surface AAs** | 25.9 | 20.0 | 17.9 | 9.1 | 7.7 | 79.7 | 21.1 | 30.6 |

### 2.4.1.3. Solvent accessibility in eplets

As mentioned in §1.1.4.2, eplets correspond to polymorphic AA positions that can often be identified on the basis of the SAB profiles obtained with sera. However, traditional definition of eplets overlook the solvent accessibility of these polymorphisms, mainly because 3D structures are not available for all HLA antigens on which these eplets are located. For this reason, the solvent accessibility of AA positions reported as being part of an eplet have been explored. A comparison is conducted here between AA positions that are part of confirmed eplets (also known as antibody-verified) or non-confirmed eplets, and those that have never been reported to be part of any eplet, referred to here as "non-eplet". These three groups cover the 207 antigens studied in this work. They are named "Confirmed", "Non-confirmed" and "Non-eplet" and contain 2 985, 6 108 and 69 684 AAs, respectively, including for each antigen the alpha and beta chains. Results of solvent accessibility in the three groups are presented in Table 9, Table 10 and Table 11 respectively.

Regarding the AA positions involved in confirmed eplets (see Table 9), the DPA1 gene has the lowest number (only one) of eplet positions but this position is solvent-accessible in all antigens where it is present. The DPB1 gene has an average of 10 eplet positions and displays the lowest percentage of solvent accessible AAs among them (48.8 % in MD data and 48.5 % in static data). For a similar average number of eplet positions (10.9), gene C displays the highest percentage of solvent accessible AAs among them (95.1 % in MD data and 90.8 % in static data, not counting the 100 % computed on a single position for gene DPA1). These findings are in line with what is known about the minor importance of the locus DP for both donor/recipient matching and antigen-antibody recognition risk, at the level of a population of patients. The DQB1 and DRB1/3/4/5 genes stand out for their low percentage of surface AAs among eplet positions compared to the remaining genes other than DPB1 (~61% and ~62%, respectively). Conversely, gene A presents the highest number of surface AA positions (~13 AAs with a solvent accessibility percentage of ~91%), which is also in line with what is known about importance of this gene in both donor/recipient matching and antigen-antibody recognition risk. Regarding the comparison between MD data and static 3D data, an average difference of ~4% in the percentage of surface AAs is observed. The values are always higher for MD data, showing the importance of exploiting MD data to identify surface AAs as putative eplets.

Table 9: Number of solvent accessible AAs at confirmed eplet positions aggregated by gene for all antigens considered in this study. Counting was made for MD trajectories and static 3D structures. HLA Eplet Registry [34] accessed in February, 2023.

| | Confirmed | | | | |
|---|---|---|---|---|---|
| | | MD | | Static 3D | |
| Gene | Confirmed eplet positions ($\mu \pm \sigma$) | Surface AA positions ($\mu \pm \sigma$) | % | Surface AA positions ($\mu \pm \sigma$) | % |
| A | $14 \pm 2.8$ | $12.7 \pm 2.5$ | 90.6 | $12 \pm 2.3$ | 85.2 |
| B | $12.1 \pm 2.1$ | $8.9 \pm 1.6$ | 74 | $7.7 \pm 1.7$ | 64.2 |
| C | $10.9 \pm 2.9$ | $10.4 \pm 2.4$ | 95.1 | $9.9 \pm 2.5$ | 90.8 |
| DPA1 | 1 | 1 | 100 | 1 | 100 |
| DPB1 | $10 \pm 2.9$ | $4.9 \pm 1.7$ | 48.8 | $4.9 \pm 1.3$ | 48.5 |
| DQA1 | $9.4 \pm 1.6$ | $6.9 \pm 1.6$ | 73.1 | $6.7 \pm 1.5$ | 71 |
| DQB1 | $14.9 \pm 2.2$ | $9 \pm 1.6$ | 60.8 | $8.3 \pm 1.4$ | 56.1 |
| DRB1/3/4/5 | $13.7 \pm 3.3$ | $8.5 \pm 1.4$ | 61.9 | $7.8 \pm 1.1$ | 56.7 |

Table 10 presents the results for AA positions involved in non-confirmed eplets. Interestingly, all percentages of surface AAs are decreased when compared to those obtained for confirmed eplets (see Table 10). This suggests that the non-confirmed group of AAs is heterogeneous and represents a mix of possible eplet (likely on surface) and possible non eplet (not on surface) AAs. As before, the numbers of AA positions on surface are the lowest for the DPA1 and DPB1 genes (~4 AAs and ~7 AAs, respectively). The most striking decrease is observed for the C gene which displays the lowest percentage of surface AAs (~39% from MD data and 34% from static 3D data), when it was at ~95% and ~90%, respectively for confirmed eplet positions. Regarding the comparison between MD and static 3D data, the observation is the same as for the confirmed case: an average difference of ~4% in the percentage of surface AAs with values always higher for MD data.

Table 10: Number of solvent accessible AAs at non-confirmed eplet positions aggregated by gene. Counting was made for MD trajectories and static 3D structures.

| | Non-confirmed | | | | |
|---|---|---|---|---|---|
| | | MD | | Static 3D | |
| Gene | Non-confirmed eplet positions ($\mu \pm \sigma$) | Surface AA positions ($\mu \pm \sigma$) | % | Surface AA positions ($\mu \pm \sigma$) | % |
| A | $24.9 \pm 2.5$ | $12.5 \pm 2$ | 50.2 | $10.3 \pm 1.6$ | 41.5 |
| B | $21.7 \pm 2.1$ | $8.9 \pm 1.6$ | 41 | $8.1 \pm 1.5$ | 37.3 |
| C | $17.7 \pm 2$ | $6.8 \pm 1.7$ | 38.7 | $6 \pm 1.5$ | 34 |
| DPA1 | $5.9 \pm 1$ | $3.8 \pm 1$ | 64.4 | $3.7 \pm 1$ | 63.9 |
| DPB1 | $13.7 \pm 2.1$ | $6.7 \pm 1$ | 48.9 | $5.9 \pm 0.9$ | 43.2 |
| DQA1 | $9.8 \pm 2.1$ | $6.3 \pm 1.9$ | 64.5 | $6.3 \pm 1.9$ | 64.2 |
| DQB1 | $19.6 \pm 2.4$ | $11.6 \pm 1$ | 59.2 | $10.9 \pm 1.2$ | 55.8 |
| DRB1/3/4/5 | $29.1 \pm 2.6$ | $13.9 \pm 1.9$ | 47.8 | $12.3 \pm 1.7$ | 42.3 |

Finally, results obtained for non-eplet AA positions are presented in Table 11. Here, very similar solvent accessibility percentages, ranging from 51.2% to 56.9%, are observed across all genes. As expected, a larger number of surface AA positions are present in class I genes (A, B, C) due to their longer sequences (276 AAs) compared to class II genes (DPA1, DPB1, DQA1, DQB1, DRB1/3/4/5 which have lengths of 183, 191, 186, 192 and 192 AAs, respectively). When comparing MD data against static 3D data, an average difference of ~5% in the percentage of surface AAs is observed. The values are always higher for MD data.

Table 11: Number of solvent accessible AAs at non-eplet positions aggregated by gene. Counting was made for MD trajectories and static 3D structures.

| | Non-eplet | | | | |
|---|---|---|---|---|---|
| | | MD | | Static 3D | |
| Gene | Non-eplet positions ($\mu \pm \sigma$) | Surface AA positions ($\mu \pm \sigma$) | % | Surface AA positions ($\mu \pm \sigma$) | % |
| A | $236.9 \pm 2.5$ | $129.1 \pm 3$ | 54.5 | $119.4 \pm 3$ | 50.3 |
| B | $242.1 \pm 2.6$ | $133.9 \pm 3.7$ | 55.3 | $122.8 \pm 3.3$ | 50.7 |
| C | $247.2 \pm 2.1$ | $138.8 \pm 4.4$ | 56.1 | $125.7 \pm 2.1$ | 50.8 |
| DPA1 | $176 \pm 1$ | $93.7 \pm 1.9$ | 53.2 | $86.4 \pm 1.7$ | 49 |
| DPB1 | $167.1 \pm 1.9$ | $96.8 \pm 3.5$ | 57.9 | $88.4 \pm 2.5$ | 52.9 |
| DQA1 | $166.7 \pm 1$ | $85.5 \pm 3.2$ | 51.2 | $79.8 \pm 2.2$ | 47.8 |
| DQB1 | $157.4 \pm 1.3$ | $87.8 \pm 3.3$ | 55.7 | $79.3 \pm 2.3$ | 50.3 |
| DRB1/3/4/5 | $149.1 \pm 2.7$ | $84.9 \pm 2.7$ | 56.9 | $76.2 \pm 2.8$ | 51.1 |

The AA positions involved in eplets used to build the results described in Table 9 and Table 10 suggest that some of them are recurrent across multiple antigens. For this reason, the most frequent solvent-accessible AA positions in confirmed eplets were examined, as this would be an indicator of

increased propensity for antibody recognition. Table 12 summarizes the most recurrent AA positions in confirmed eplets across all genes, while Figure S5 shows the detailed results and Figures S6-S13 show 3D visualizations of AA positions highlighted in Table 12.

In Class I genes, various regions can be identified on the antigen 3D structures for solvent-accessible AAs matching with eplet positions in most antigens. For genes A and B, they mostly occur in the region of the peptide-binding groove (Figures S6 and S7). In gene C, two regions are found along the peptide-binding groove and one region is further away, close to the part of the antigen that crosses the membrane (Figure S8). The accessibility of this zone to an antibody is certainly reduced.
For Class II genes, the situations are also very varied. Not surprisingly, due to the low number of surface polymorphisms (~9, see Table 8) and the low number of confirmed eplets (2 eplets, see Table 12), the DPA1 gene has only one polymorphic AA position and thus only one zone of interest for antibody recognition (relative to confirmed eplets), located on one edge of the peptide groove (Figure S9). Conversely, gene DPB1 has a greater number of confirmed eplets, which translates into three zones that are apparently favorable for antibody recognition (Figure S10). Surprisingly, the DQA1 gene has a low number of solvent-accessible confirmed eplet positions even though it is the gene with the greatest number of surface polymorphisms (~79 AAs, see Table 8) and these positions correspond to 3 regions of interest for antibody recognition (Figure S11). The DQB1 gene has the highest diversity of zones of interest (6 in total, Figure S12) although it does not have a large number of surface polymorphisms (~23 AAs, see Table 8) or confirmed eplets (21 eplets, see Table 12). Finally, genes DRB1/3/4/5, which have the greatest number of confirmed eplets (41 eplets, see Table 12), show 4 zones of interest, indicating a high concentration of confirmed eplets in these zones of the concerned antigens (Figure S13).

Table 12: List of most frequent solvent accessible AA positions across antigens within confirmed eplets and number of confirmed eplets corresponding to these positions. The number of occurrences is shown in parentheses in the column "Most frequent surface positions across antigens". Occurrences are counted as the number of antigens registered for a given eplet at a given position (HLA Eplet Registry, accessed in February, 2023). The number of antigens associated with each gene is shown in parentheses in the column "Gene".

| Gene | Most frequent surface positions across antigens | # confirmed eplets |
|---|---|---|
| A (34) | 62 (33), 65 (28), 66 (33), 69 (24), 79 (29), 80 (31), 138 (25), 142 (34), 145 (23) | 35 |
| B (59) | 69 (59), 76 (35), 80 (57), 131 (46), 163 (43), 167 (35) | 37 |
| C (17) | 65 (16), 66 (15), 69 (16), 80 (17), 193 (16), 194 (16) | 20 |
| DPA1 (31) | 50 (31) | 2 |
| DPB1 (31) | 57 (18), 58 (18), 85 (23), 86 (31), 87 (17), 96 (23) | 12 |
| DQA1 (30) | 2 (30), 40 (30), 47 (20) | 9 |
| DQB1 (30) | 46 (21), 52 (30), 55 (30), 77 (30), 84 (21), 125 (22), 182 (30) | 21 |
| DRB1/3/4/5 (36) | 4 (35), 25 (28), 70 (36), 73 (35), 77 (36), 96 (25), 98 (28) | 41 |

### 2.4.1.4. Solvent accessibility and physico-chemical properties of AAs

To explore the solvent accessibility of AAs with respect to their physicochemical properties, the 20 types of AAs were divided into 4 sets according to the IMGT classification[13], as explained in Chapter 1 (1.2 Basics of structural bioinformatics). The "non-polar non hydrophobic" set comprises Glycine (GLY) and Proline (PRO) which are particular AAs mostly associated with turns in the polypeptide chain. Glycine has the shortest side chain among all AAs, limited to a single hydrogen atom, and the side chain of Proline forms a ring with the nitrogen atom of its amino group, leading to forced curvature of the peptide chain. The "non-polar hydrophobic" set comprises Valine (VAL), Alanine (ALA), Leucine (LEU), Isoleucine (ILE), Methionine (MET), Cysteine (CYS), Phenylalanine (PHE) and Tryptophan (TRP). The "polar uncharged" set comprises Serine (SER), Threonine (THR), Tyrosine (TYR), Glutamine (GLN), Asparagine (ASN) and Histidine (HSD). The "polar charged" set comprises Glutamate (GLU), Aspartate (ASP), Lysine (LYS) and Arginine (ARG). As a reminder, Histidine shows a mostly neutral (non-charged) protonation state under physiological conditions, and in structural bioinformatics the 3-letter name HSD is often used for this state instead of HIS. For simplicity, the two non-polar sets are merged into a single non-polar (hydrophobic or not) set in the following.

Here, the RSASA values are collected either from MD trajectories or from static 3D data, for all AAs of the same type within a given physico-chemical set across all HLA antigens considered in this study. Each of the three AA groups defined above ("Confirmed", Non-confirmed" and "Non-eplet) are analyzed separately. The "All" group is added as a reference. It contains all the AAs present in the 207 structures considered in this work (total number of AAs: 77 648, including $\alpha$ and $\beta$ chains for each antigen). The results are plotted as the count of observations in bins of RSASA values (one bin per percent unit) in Figure S17 and Figure S18. A supplementary table provides all median values corresponding to these plots (Table S9).

In the first step, Figure S17 presents a comparison between the plots obtained with RSASA values calculated from MD trajectories (panels A-D) and those obtained from static 3D structures (panels E-H), for the set of non-polar AAs. MD trajectories provide 500 RSASA values for each AA (one per frame) while static 3D structures only provide one value. The highly redundant plots obtained with MD data make it possible to visualize trends in the data that cannot be detected with static data. This is particularly obvious for the "Confirmed" group (panel B) in which a peak of ALA emerges with MD data (median RSASA value: 33.79%) that goes completely unnoticed with static data. This peak is absent from the "All", "Non-confirmed" and "Non-eplet" groups. Similar observation occurs, although with less prevalence, for a peak of VAL in the "Non-confirmed" group (median RSASA value 25.94%), visible with MD data and undetectable with static data, but apparently characteristic of the "Non-confirmed" group when compared with the "All" and "Non-eplet" groups. Due to the better readability of plots derived from MD data, those obtained with static data will not be represented for the analysis of the other two sets of AAs.

Figure S18 represents the RSASA values distribution collected from MD trajectories for the set of polar and uncharged AAs (panels A-D) and the set of polar and charged AAs (panels E-H). In both cases, the plots obtained with the "All" group look similar to the one obtained with the "Non-eplet" group, which is an expected result. Plots corresponding to the "Confirmed" group are clearly different and reveal which types of AAs will be mostly represented as solvent accessible AAs in confirmed eplets, namely THR and to a lower extent HSD, ASN and GLN for polar uncharged AAs, ARG and to a lower extent ASP and GLU for polar charged AAs. Interestingly, the plots corresponding to the "Non-confirmed" group are different, indicating that the "Non-confirmed" eplets are likely composed of a mixture of true and false eplets.

When comparing the "Non-eplet" group with the "Confirmed" group for the MD data, a remarkable difference is observed at the level of non-polar AAs (Figure S17, panels B and D) as in the "Confirmed" plot RSASA values above the 20% threshold are highly represented for various types of AAs, while in the "Non-eplet" case it is the opposite. Huge changes in median RSASAs are observed for GLY, ALA, LEU, ILE, PRO, MET and TRP AAs (Table S9: ~28%, ~15%, ~10%, ~12%, ~34%, ~9%, ~10% in the "Non-eplet" group and ~18%, ~34%, ~30%, ~26%, ~56%, ~36%, ~27% in the "Confirmed" group, respectively) and

[13] see https://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/IMGTclasses.html

huge differences in the shape of distributions are noted for ALA, VAL, LEU and TRP. This trend towards increased solvent accessibility for non-polar AAs in the “Confirmed” group reflects the fact that this group mostly comprises surface AAs, including non-polar AA types that are less frequently encountered on the surface in the rest of the structure.

For polar and uncharged AAs (Figure S18, panels B and D), the “Confirmed” and “Non-eplet” groups show similar trends, although relevant variations in median RSASA for HSD is observed (Table S9: ~22% and ~18% in the "Non-eplet" group and ~49% and ~48% in the "Confirmed" group, respectively) and a considerably different distribution shape is observed for THR. CYS is a particular case as no CYS type is present in the “Non-eplet” group.

For polar and charged AAs (Figure S18, panels F and H), although in both cases all AAs have a median above the 20% threshold, differences in the shapes of the ARG and GLU distributions are evident. Similarly, changes are seen in the median RSASA values of the 4 charged AAs, especially for ARG (Table S9: ~33% for the "Non-eplet" group and ~51% for the "Confirmed" group).

In summary, this analysis illustrates not only the increase in solvent accessibility of AAs present in confirmed eplets but also the specific types of AAs that can be mostly expected in eplets compared with non-eplets, in particular for those types of AAs that are frequently found on the surface of proteins. A more detailed study of AA distribution in the four groups of AA defined in this work will be presented in next section (2.4.2). Before that, a concrete case-study showing how solvent accessibility data can be used to characterize two distinct serological groups of antigens is presented.

#### 2.4.1.5. Case-study: Serological groups BW4/BW6 and position 80

The BW4 and BW6 groups define two mutually exclusive well-known serologic phenotypes (CREG more precisely, but this acronym is not used much so far in the thesis) in the field of transplantation, that were initially considered as serologic groups per se, but were removed from this classification when it was shown in the late 1960s that they “only” represented antibody targets (what became 35 years later eplets). The BW4 group contains 23 antigens in the LSA assay from locus B, while the BW6 group contains 32 antigens also from locus B (see Table S4 for a detailed list in each group). Actually, all the B antigens that exist in HLA belong to one or the other, but the focus will be on the few that are represented in the LSA assay. Both groups are disjunct and position 80 is considered to be the main AA defining the eplet defining these groups (see multiple alignment for BW4 and BW6 groups around position 80 in Figure S15). To try to understand a little better the aspects that characterize the epitope around this position in each of them, the solvent accessibility of this position and its surroundings was explored during MD trajectories in all concerned antigens. Figure 22 shows the solvent accessibility at position 80 along the MD trajectory for groups BW4 and BW6, while in Figure S14 the solvent accessibility is shown also for surrounding positions 76, 77, 78, 79, 81, 83, 84 and for position 138 (conserved position along all antigens in BW4 and BW6, this last position is included as control). In Figure 22 and S14 , the value of the median RSASA from MD data and the value of the RSASA from static structure have been plotted for each position and serological group considered. It is worth noting that in several cases a large gap between these two values can be observed. This reinforces the importance of the contribution of MD to protein structural studies.

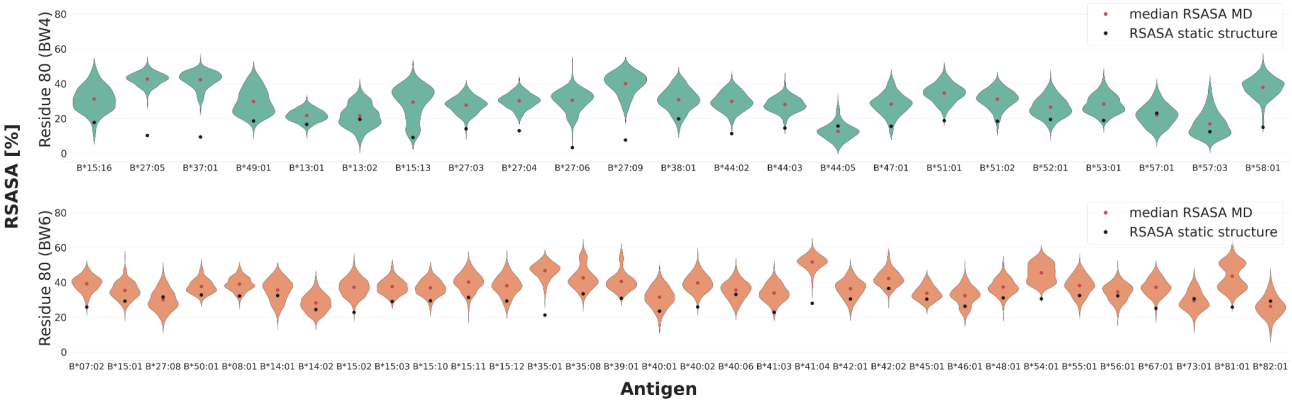



Figure 22: Solvent accessibility of position 80 in antigens from BW4 and BW6 serological groups. Violins represent the RSASA distributions along the MD trajectories for each antigen. Red dot corresponds to the median RSASA of the violin while black dot corresponds to the RSASA value derived from the static structure.

Upon visual inspection of the RSASA distributions in Figure 22 and S14, it is not possible to identify any pattern that distinguishes one group from the other. The Kolmogorov-Smirnov test (K-S test) is a statistical test sensitive to differences in both location and shape of the cumulative distribution functions of two samples. Therefore, the K-S test statistic was used as a distance score between the RSASA distributions in BW4 and BW6 groups in order to compare them. The results are shown as a heatmap in Figure 23 for position 80. Interestingly, it is possible to distinguish two fuzzy clusters displaying low distance values (majority of pale blue pixels), one for BW4 x BW4 and one for BW6 x BW6 comparison. By contrast, the BW4 x BW6 pairwise comparison shows higher distance values (darker blue pixels) This result suggests the existence of a distinct pattern or trend in the dynamics of solvent accessibility in each group. Similar clusters can also be seen in the heatmaps for positions 82 and 83, especially in the former (see Figure S16, panel E and F). Although these results are promising, RSASA alone is not sufficient to fully distinguish the BW4 and BW6 clusters. Therefore, it is important to integrate other structural descriptors to describe epitopes. This will be further explored in this chapter and will allow the creation of an ML dataset for B-cell epitope prediction in Chapter 4.

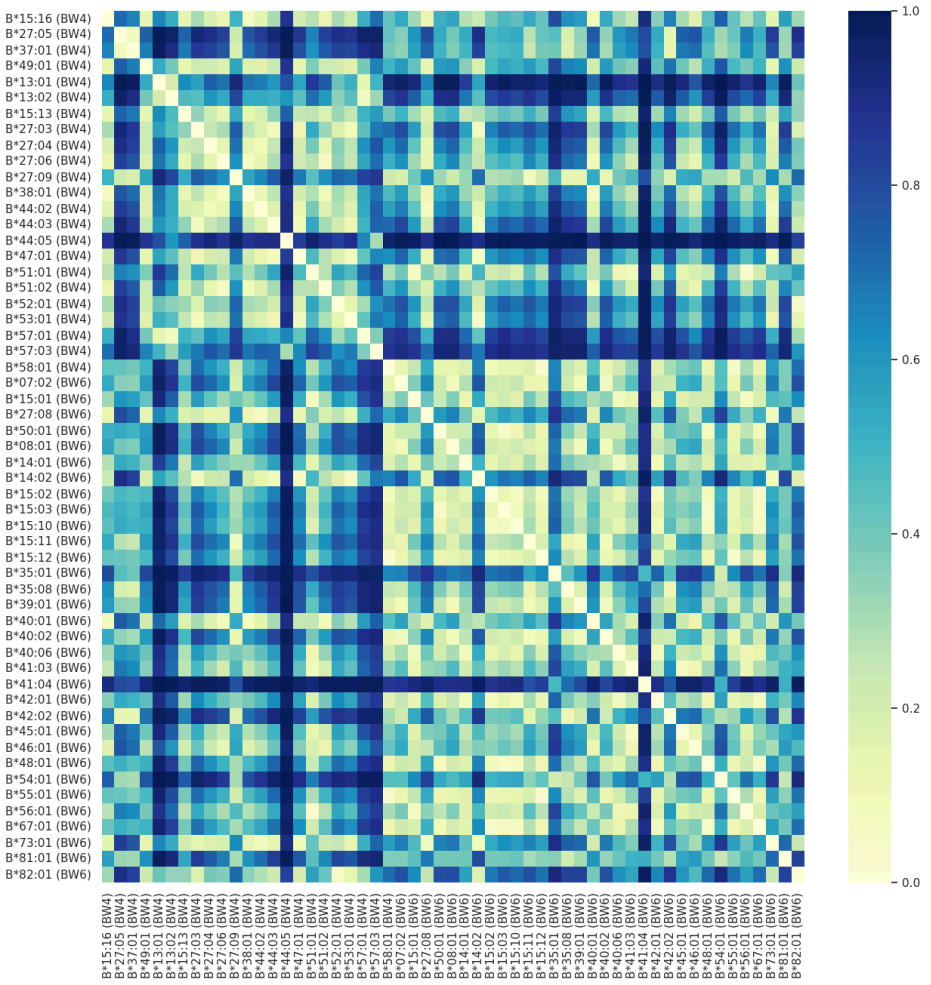



Figure 23: Heatmap of the K-S test statistic between RSASA distributions at position 80 of BW4 and BW6 antigens.

## 2.4.2. Frequency of occurrence and properties at eplet and non eplet positions

This section examines the frequencies of occurrence and properties of the 20 types of AA in each of the same four groups defined in the previous subsections, i.e. "All", "Confirmed", "Non-confirmed" and "Non-eplet". The frequencies of occurrence were calculated from sequence and from solvent-accessible AAs and results are summarized in Table 13.

A first observation is that, as expected, the frequencies of occurrence in the sequence is significatively different from the frequencies of occurrence on the surface of the protein for all groups of AA, with p-values from a chi-squared ($X^2$) test ranging from 0 for the “All” and “Non-eplet” groups to $2x10^{-6}$ for the “Confirmed” group. Variations of the percentages between sequence and surface are reported for each group and each type of AA in Table S5. Highest differences (> 3% in at least one group) are found for two polar charged AAs (ARG and GLU) which are more represented on the surface, and for two non-polar AAs (LEU and PHE) which are less represented on the surface. Two polar uncharged AAs also stand out in the “Non-confirmed” group with opposite behaviors: THR more represented on the surface and TYR less represented on the surface. Interestingly, the differences are very small (ranging from -1.8% to 1.5%) for the “Confirmed” group, except for ARG which is more represented (3.3%) on the surface. This observation is in line with the fact that most AAs in the “Confirmed” group belong to polymorphic eplets often present on the surface of the antigen.

Table 13: Frequency of occurrence of AAs (in %) for the "All", "Confirmed", "Non-confirmed", and "Non-eplet" groups. Percentages are calculated from the sequence or from solvent-accessible AAs calculated from MD trajectories. AA counts in each AA group are shown in parentheses. AAs are grouped according to their physico-chemical properties. Cells with values higher than 10% are highlighted in green.

| | | All | | Confirmed | | Non-confirmed | | Non-eplets | |
|---|---|---|---|---|---|---|---|---|---|
| **AA properties** | **AA** | **Seq (77 648)** | **Surf (42 626)** | **Seq (2 985)** | **Surf (2 110)** | **Seq (6 108)** | **Surf (3 323)** | **Seq (69 684)** | **Surf (38 100)** |
| Non-polar non-hydrophobic | GLY | 5.5 | 6.1 | 3.7 | 3.1 | 1.6 | 0.9 | 5.7 | 6.3 |
| | PRO | 5.6 | 7 | 1.7 | 2.4 | 2.3 | 4.3 | 5.7 | 7 |
| Non-polar hydrophobic | ALA | 5.6 | 4.6 | 10 | 10.3 | 9.1 | 9.2 | 5.5 | 4.2 |
| | VAL | 7.2 | 4.8 | 5.6 | 3.8 | 9.9 | 10.6 | 7.5 | 4.8 |
| | LEU | 7.3 | 4.3 | 4.1 | 3.1 | 6.2 | 3.5 | 7.3 | 4.5 |
| | ILE | 3.3 | 2.1 | 3.5 | 3.5 | 5.4 | 7.5 | 3.2 | 1.9 |
| | MET | 1.2 | 0.5 | 2.1 | 1.9 | 2.4 | 0.6 | 1 | 0.5 |
| | PHE | 4.5 | 1.1 | 1.5 | 0.6 | 3.4 | 1.5 | 4.4 | 1 |
| | TRP | 2.6 | 1 | 2.2 | 2.8 | 0.6 | 0.3 | 2.7 | 0.9 |
| | CYS | 1.7 | 0.1 | 0.4 | 0.3 | 0.4 | 0.4 | 1.8 | 0 |
| Polar uncharged | SER | 5.6 | 6.2 | 6.7 | 5.5 | 5.9 | 3.8 | 5.6 | 6.3 |
| | THR | 7.1 | 7.9 | 13.8 | 14.2 | 6.8 | 9.8 | 7 | 7.9 |
| | TYR | 4.6 | 1.9 | 1.5 | 0.6 | 9.2 | 4.2 | 4.5 | 2 |
| | ASN | 3.7 | 4.2 | 6.4 | 7.4 | 5.7 | 3.9 | 3.6 | 4.2 |
| | GLN | 5 | 6.5 | 5 | 6.5 | 1.7 | 2.3 | 5.1 | 6.6 |
| | HSD | 3.5 | 3.4 | 2.1 | 2.8 | 5.9 | 5.8 | 3.5 | 3.5 |
| Polar charged | ASP | 6.2 | 7.6 | 4.8 | 4.3 | 4.8 | 3.5 | 6.3 | 7.9 |
| | GLU | 8.8 | 13.9 | 8.5 | 7.9 | 6.7 | 10.2 | 8.7 | 13.9 |
| | LYS | 3.8 | 6 | 5.5 | 5.2 | 2.9 | 4.2 | 3.8 | 6.1 |
| | ARG | 7.2 | 10.7 | 10.7 | 14 | 9 | 13.4 | 7.1 | 10.6 |

Regarding surface residues, the AA types presenting high percentages of occurrence (>10 %) are ARG (in all groups), GLU (in all groups except for the “Confirmed” group), ALA and THR in the “Confirmed” group, and VAL in the “Non-confirmed” group. Statistical analyses based on the $X^2$ test reveal that the surface AA-type distribution in the “Confirmed” group significantly differs from each of the three other groups (p-value below $10^{-80}$). The same is true for the ”Non-Confirmed” group, while the comparison of surface AA-type distribution between the “All” and “Non-Eplet” group does not reveal any significant difference (p-value = 0.177). These results show that eplets, either confirmed or not, do not display the same AA-type composition as the rest of the surface of the antigen, whereas the “Non-eplet” group is similar to the “All” group from this point of view.

At the AA-type level (see the differences in % in Table S6), one can observe the largest variations between the “Confirmed” and “Non-eplet” groups for THR, ALA, GLU, PRO, ASP, ARG, ASN and GLY (6.2%,

6.1%, -6%, -4.6%, -3.6%, 3.3%, 3.2% and -3.1%, respectively). This list is slightly different when comparing the “Non-confirmed” and “Non-eplet” groups: VAL, ILE, GLY, ALA, ASP, GLN and GLU (5.8%, 5.7%, -5.4%, 5%, -4.4%, -4.4% and -3.7%, respectively). A higher prevalence of THR seems to be characteristic of the “Confirmed” group and a higher prevalence of VAL characteristic of the “non-confirmed” group. When further comparing the frequencies of occurrence of surface AAs for the "Non-confirmed" and "Confirmed" groups, the AA types with the largest variations are VAL, THR, GLN, ILE, TYR, ASN, and HSD (-6.8%, 4.3%, 4.2%, -4.1%, -3.6%, 3.5%, and -3%, respectively.).

It should be noted that the surface AAs are calculated here from MD trajectories. The same frequencies of occurrence of surface AAs were calculated also from static 3D structures. The results are presented in Table S7, while Table S8 shows the differences between the MD and static data for surface AAs. A $X^2$ test performed between static and MD data shows that the frequencies of occurrence are not significantly different (p-value > 0.96).

### 2.4.3. Side-chain flexibility along MD simulation

The Normalized-Root Mean Square Fluctuation (N-RMSF) allows the evaluation of the flexibility of the side-chain of an AA during MD simulation. This variable has been proposed as a determining factor in antigen-antibody recognition [110]. Equation 9 presents the formula used to calculate the RMSF, where i indicates a given amino acid and $r_i(t)$ represents the set of atomic coordinates for a given amino acid for frame t. Equation 10 introduces the normalization method used, known as the Z-score method, where μ corresponds to the mean and σ to the standard deviation of the RMSF values.

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\langle \left(r_i'(t) - r_i(t_{ref})^2\right)\right\rangle} \quad 9)$$

$$\text{N-RMSF}_i = \frac{RMSF_i - \mu}{\sigma} \quad 10)$$

The N-RMSF values were calculated for the AAs of the "Confirmed" and "Non-eplet" groups for all 207 antigens considered in this thesis (the number of N-RMSF values considered are 2 117 and 38 346 for the Confirmed and Non-eplet groups respectively). To compare the N-RMSF value distribution in the two AA groups, we decided to compare their cumulative distribution functions. The cumulative distribution function, or probability distribution function F(x), is the mathematical equation that describes the probability that a variable X is less than or equal to x, i.e. $F(x) = P(X \leq x)$ for all x, where $P(X \leq x)$ means the probability of the event $X \leq x$. Here, the probability $P(\text{N-RMSF} \leq x)$ is given by the ratio of the number of observed N-RMSF values less than or equal to x to the total number of observed N-RMSF values, in each group of AAs. Figure 24 shows the cumulative probability plots of the N-RMSF variable in the two groups of AA considered here. Interestingly, the "Confirmed" group has an overall tendency to be less flexible than the "Not Confirmed" group. However, in the interval [-1.7, -1] the trend is reversed. To evaluate whether there is a statistically significant difference between the two groups, the K-S test was performed. This statistical test was chosen because it is sensitive to differences in both location and shape of the cumulative distribution functions of two samples and doesn't require equal sample size. A p-value of $4.6^{-06}$ was found, indicating that there is a significant difference between the "Confirmed" and "Non-eplet" groups. This result confirms what has been reported in the literature [110] but never validated in the case of HLA antigen-antibody. This work was presented as a poster at ECCB 2022 in Sitgès [124].

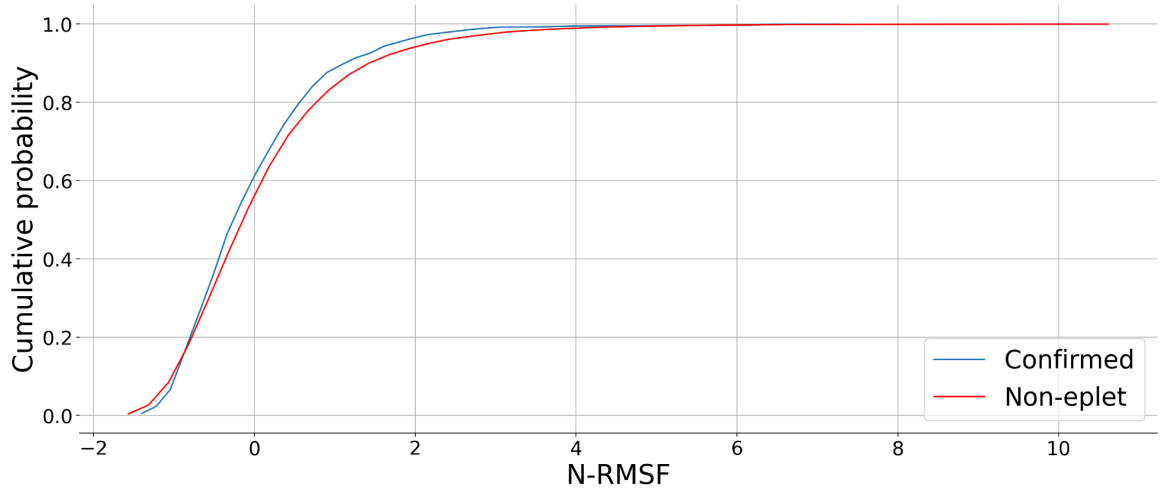



Figure 24: Cumulative probability plots representing the distribution function of N-RMSF values observed for AAs in Confirmed (2 117 values) and Non-eplet (38 346 values) groups.

## 2.5. Exploration of molecular dynamics data at the patch level

### 2.5.1. General overview

This section introduces the concept of "patch" which is essential to address the characterization of the HLA epitopes encompassing the eplets (Figure 6). In this work, a patch is a region of the protein surface which contains all solvent-accessible AAs within a given radius from a central solvent-accessible AA (distances are computed between the centers of mass of the AAs). Thus, patches are identified by the HLA antigen structure from which they are derived and by the type and sequence position of their central AA. As already mentioned above, a median RSASA of 20% along the MD trajectory is the threshold used to define solvent accessibility of an AA. Three radius values were explored (15Å, 12Å and 9Å) and various sets of patches were generated for all solvent-accessible AAs of our set of 207 HLA antigens, taking into account their MD trajectories. Pymol and Python scripts were created for this purpose.

For sake of clarity, the results of data exploration at the patch level in terms of AA distribution, physico-chemical properties, solvent accessibility and flexibility, are presented in this section only for "epitope" and "non-epitope" patches. Epitope patches are centered on a solvent-accessible AA known as a confirmed eplet (or part of such eplet) in HLA Eplet Registry. Non-epitope patches are centered on a solvent-accessible AA that is listed neither as an eplet nor as part of an eplet (either confirmed or not) in HLA Eplet Registry and they must fulfill two additional conditions. Firstly, none of the other solvent-accessible AAs present in the patch should be an eplet or part of an eplet (either confirmed or not). Secondly, no patch can be composed exclusively of locus-conserved positions. Indeed, locus-conserved positions could display B-cell epitope behavior that cannot be recognized as such because they are conserved throughout all human individuals and therefore cannot raise immune response between individuals of human species. In total, were computed as many epitope patches as solvent-accessible confirmed eplet residues (2 117) for the three radii (15Å, 12Å and 9Å). For the non-epitope patches, numbers are varying depending on the radius, due to their strict definition. A total of 4 769 non-epitope patches were obtained for a patch radius of 15Å, 6 332 for a patch radius of 12Å and 7 866 for a patch radius of 9Å. It is important to stress here that each patch is actually represented by its 500 conformers computed for each of the 500 frames representing the MD trajectory of the concerned antigen (Section 2.2.4).

### 2.5.2. Solvent accessibility

As in section 2.4.1.4, the solvent accessibility of AAs is examined here but at the patch level, with respect to the physico-chemical properties of AAs composing the epitope and non-epitope groups. The same three sets of AAs are considered (non-polar hydrophobic or not, polar-uncharged and polar-charged, see section 2.4.1.4 above). Two groups of AAs are built for each radius value: the "Epitope" and the "Non-epitope" ones, comprising all AAs belonging to epitope and non-epitope patches, respectively. We define that an AA belongs to a patch if this AA remains inside the patch for at least 70% of the 500 frames representing the MD trajectory. Indeed, due to structure fluctuations during MD simulation, several AAs located on the border of the patch may not fulfill from time to time the distance requirement to the central AA. The total numbers of AAs considered for epitope patches are 12 534, 21 831 and 36 545 for radii of 9Å, 12Å and 15Å, respectively, and for non-epitope patches 43 695, 65 259 and 84 107 for the same three sizes of patch, respectively.

Figure S19, Figure S20 and Figure S21 shows the RSASA distribution derived from MD trajectories for the various sets of AAs in epitope (left panels) versus non-epitope (right panels) patches of size 15Å, 12Å and 9Å, respectively. As in Figure S18, it should be reminded here that each AA here is represented by 500 frames, leading to a highly redundant plot in which frames rather than AAs are counted for each RSASA percent bin. However, these plots are useful to derive trends for each AA type, that can possibly differentiate the two groups of patches (epitope and non-epitope). The median values represented by vertical lines on the figures are listed in Table S10. As we are dealing with AAs present in patches, these AAs are all solvent-accessible by construction. Therefore, most RSASA values in the plots are above the 20% threshold.

Regarding the different patch sizes, consistent results are obtained over all radii and no systematic decrease or increase in the median values of RSASA is observed. Important variations are only observed for a few AA types either in the "Epitope" or in the "Non-epitope" group. In the former group, the RSASA median value increases for VAL from 31.0 to 38.1 % when the radius decreases from 15Å to 9 Å. In the latter group, the RSASA median value increases for ALA from 34.7 to 42.3 % when the radius decreases from 15Å to 9 Å and the RSASA median value for MET decreases from 37.9 to 22.4 when the radius decreases from 15Å to 9 Å.

When comparing the "Epitope" versus the "Non-epitope" group of AAs, important differences can be observed on Figure S19, Figure S20 and Figure S21, and are confirmed when looking at median RSASA values in Table S10. Some examples are provided here for the 15Å radius as this size of patch will be used later for machine learning protocols (Chapter 4). In Figure S19, in the set of non-polar AAs (panels A and D), the peak corresponding to ALA is specific of the "Epitope" group (RSASA median value 42.2 and 34.7% for the "Epitope" and "Non-epitope" groups, respectively). In the set of polar uncharged AAs (panels B and E), the peak corresponding to ASN around 70% RSASA is visible only in the "Epitope" group. The RSASA median values are not very different (38 and 34.9 % for the "Epitope" and the "Non-epitope groups, respectively) indicating that the number of AAs rather than their solvent accessibility makes the difference here. The two peaks corresponding to HSD are shifted to lower RSASA values in the "Non-epitope" group (RSASA median value 43.8 and 35.5% for the "Epitope" and "Non-epitope" groups, respectively). Similar shift (although less visible on the plots) is observed in Table S10 for THR (RSASA median values from 41.2 to 35.7 %) and to a lower extent for GLN (RSASA median values from 41.2 to 38.1 %) and TYR (RSASA median values from 29.1 to 23.8 %). In this set of AA types, the case of CYS is particular as no CYS is observed in the "Non-epitope" group (see below Table 14: Frequency of occurrence of AAs). Finally, in the set of charged AAs (panels C and F), the peaks corresponding to GLU and LYS appear shifted towards high RSASA values for the "Non-epitope" group (RSASA median values vary from 41.7 to 49.2% for GLU and from 50.7 to 55.7% for LYS in the "Epitope" and "Non-epitope"groups respectively), while the reverse is true for ARG (RSASA median values: 39.9 and 35.9 % in the "Epitope" and "Non-epitope" groups respectively).

The results obtained for the "Epitope" group can be aligned with the trends observed in section 2.4.1.4 at the AA level in the "Confirmed" group of AAs which includes all the central AAs of "Epitope"

patches (see Figure S18 and Table S9). Indeed, an important prevalence and high level of solvent accessibility were already observed in this group at the AA level for ALA, THR, GLN, HSD and ARG. To my knowledge, such results have never been reported so far for HLA epitopes. The prevalence of certain types of AA in epitopes will be further studied in the next section.

### 2.5.3. Frequency of occurrence of AAs in epitope versus non-epitope patches

As in section 2.4.2, the frequency of occurrence of AAs, but the AA considered are members of epitope patches on the one hand, non-epitope patches on the other hand. These two groups of AAs are defined for each of the three radius sizes (15Å, 12Å and 9Å) explored in this study and the number of AAs from each type is expressed as the percentage of the total number of AAs in the group.

Table 14 summarizes the obtained results. When comparing the AA-type distribution in the "Epitope" group, the p-values obtained for a $X^2$ test are all below the threshold of 0.05, indicating significant differences. However, the difference is less significant between 15 and 12Å (p-value $3.4x10^{-6}$) than between 12 and 9Å (p-value $5.2x10^{-9}$). The most affected AA types (differences 15 versus 9Å ranging from 0.5 to 2 in absolute value) are GLU, HSD and TRP that display decreasing percentages with decreasing patch size and LEU, ASN and GLN that increase in percentages when patch size decreases. In the "Non-epitope" group, the differences are even more significant with p-values of $5.7x10^{-61}$ between 15 and 12Å and of $1.6x10^{-17}$ between 12 and 9Å. The very small p-values are likely due to the large number of AA analyzed in this group. The most affected AA types (differences 15 versus 9Å ranging from 0.5 to 0.8 in absolute value) are ALA and THR that display decreasing percentages with decreasing patch size and ASN, GLU and LYS that increase in percentages when patch size decreases.

More importantly, when comparing the results in the "Epitope" versus "Non-epitope" groups, bigger differences are observed. The p-values obtained with a $X^2$ test are given as 0 for radius sizes on 15 and 12Å and $9.8x10^{-117}$ for radius size of 9Å, showing that AA-type distributions are significantly different between the "Epitope" and "Non-epitope" groups, whatever the patch size. For example, ARG displays a percentage of 14.4% in the "Epitope" group versus 9.4% in the "Non-epitope" one, 14.9% versus 9.7% and 14.8% versus 11%, for the 15Å, 12Å and 9Å radii, respectively, showing that ARG is more represented in epitope than in non-epitope patches. Similarly, VAL displays a percentage of 5.1% in the "Epitope" group versus 2.6% in the "Non-epitope" one, 4.8% versus 2.3% and 4.5% versus 2.5% for the 15Å, 12Å and 9Å radii respectively. In the other way round, THR displays a percentage of 6.1% in the "Epitope" group versus 12.1% in the "Non-epitope" one, 6% versus 11.3%, and 6% versus 9.3%, for the 15Å, 12Å and 9Å radii respectively. Similarly, PRO displays a percentage of 5.2% in the "Epitope" group versus 10% in the "Non-epitope" one, 5.2% versus 11%, and 5% versus 11.7%, for the 15Å, 12Å and 9Å radii respectively, and GLY a percentage of 4.9% in epitope patches versus 7% in non-epitope patches, 4.5% versus 7% and 4.5% versus 8.6% for the 15Å, 12Å and 9Å radii respectively. It should be recalled that due to their side-chain properties, these last two types of AA are mostly involved in the turns of the protein backbone.

Finally, we can compare these results with those obtained in section 2.4.2 at the AA level (Table 13), in particular the "Confirmed" group of solvent-accessible AAs as these AAs map to the central AAs of patches in the "Epitope" group, and the "Non-eplet" group which contains AA represented in the "Non-epitope" group of patches. Interestingly, all of these comparisons yield very significant p-values in $X^2$ tests (all p-values less than $10^{-100}$) indicating that the two levels of observation (AA or patch) are significantly different. Indeed, when comparing the AA-type distribution obtained for the "Confirmed" surface AAs and for the "Epitope" group (averaged % across the three radii), important differences (greater than 3% in absolute value) are observed for THR (14.2% and ~6.2%, respectively), GLU (7.9% and ~12.4%), ALA (10.3% and ~5.4%), TYR (0.6% and ~4.5%), and ASN (7.4% and ~4.0%). This reveals that AA composition in epitope patches differs from the one of their central AAs (eplets), and suggest that peripheric AAs (not all are eplets) play a yet unstudied role in the characterization of epitope patches. On the contrary in the case of the comparison between "Non-eplet" surface AAs and the "Non epitope" group of AAs (averaged % across all radii), no differences in percentage greater than 2.5% in absolute value are observed, except for GLN (6.6% and ~10.7%, respectively). These results reflect the fact that the

"Non-eplet" group of surface AAs contains AAs that can be anywhere in a patch and not only at the central position. Thus, the differences between this group and the "Non-epitope" group do not display so many high values as in the preceding comparison. The particular behavior observed for GLN can be an unexpected consequence of the constraints used for selecting the non-epitope patches.

Table 14: Frequency of occurrence of AAs (in %) in epitope and non-epitope patches computed from MD trajectories for three patch sizes (15Å, 12Å and 9Å). AA types are grouped according to their physicochemical properties as in Table 13. Cells with percentages > 10 % are highlighted in green. Total number of AAs is indicated in parentheses below the Epitope and Non-epitope headers

| | | 15Å | | 12Å | | 9Å | |
|---|---|---|---|---|---|---|---|
| **AA properties** | **AA** | **Epitope (36 545)** | **Non-epitope (84 107)** | **Epitope (21 831)** | **Non-epitope (65 259)** | **Epitope (12 534)** | **Non-epitope (43 695)** |
| Non-polar non-hydrophobic | GLY | 4.9 | 4.7 | 4.9 | 4.8 | 4.5 | 4.9 |
| | PRO | 5.2 | 6.2 | 5.2 | 6.3 | 5 | 5.8 |
| Non-polar hydrophobic | ALA | 5.4 | 6.6 | 5.3 | 6.8 | 5.5 | 6.1 |
| | VAL | 5.1 | 5.4 | 5.2 | 5.3 | 4.7 | 5.4 |
| | LEU | 3.4 | 2.6 | 3.9 | 2.3 | 4.1 | 2.5 |
| | ILE | 2.6 | 1.1 | 2.5 | 1.2 | 2.3 | 1.3 |
| | MET | 0.4 | 0.4 | 0.4 | 0.2 | 0.4 | 0.1 |
| | PHE | 1.2 | 0.7 | 0.9 | 0.7 | 1 | 0.8 |
| | TRP | 1.8 | 1 | 1.7 | 1.1 | 1.2 | 1.1 |
| | CYS | 0.1 | 0 | 0.1 | 0 | 0.1 | 0 |
| Polar uncharged | SER | 6 | 7 | 6 | 7 | 6 | 7 |
| | THR | 6.1 | 9.4 | 5.9 | 8.2 | 6.5 | 8.6 |
| | TYR | 4.6 | 3.2 | 4.5 | 4.1 | 4.5 | 3.6 |
| | ASN | 3.5 | 2.6 | 4.0 | 3.3 | 4.4 | 3.1 |
| | GLN | 8 | 10 | 9 | 11 | 10 | 11 |
| | HSD | 3 | 1 | 3 | 1 | 2 | 1 |
| Polar charged | ASP | 6.9 | 9.7 | 6.3 | 9.7 | 6.8 | 9.3 |
| | GLU | 12.6 | 11.2 | 12.4 | 11.1 | 12.1 | 11.7 |
| | LYS | 4.7 | 4.5 | 4.8 | 4.2 | 4.5 | 3.7 |
| | ARG | 14.4 | 12.1 | 14.9 | 11.3 | 14.8 | 12.0 |

## 2.5.4. Side-chain flexibility

As in section 2.4.3, side-chain flexibility is examined here but in order to compare the N-RMSF values observed during MD trajectories for AAs belonging to the "Epitope" and "Non-epitope" groups. Similarly to what was done in section 2.4.3, the K-S test was used to evaluate the existence of a statistically significant difference in N-RMSF cumulative probability function between the two groups of AAs. This evaluation was performed for the three patch sizes considered and p-values of $4^{-104}$, $8.3^{-149}$, and $1.1^{-267}$ were found for the 9Å, 12Å, and 15Å patches, respectively. This result extends at the patch level the observations performed at the AA level (section 2.4.5) showing that N-RMSF values for AAs in the "Confirmed" eplet group are significantly smaller than those found for AAs in the "Non-eplet" group. Here too, the "Epitope" group of patches has an overall tendency to be less flexible than the "Non-epitope" group. Figure 25 shows the

cumulative probability plots of the N-RMSF variable in the two groups of AAs and for each patch size considered.

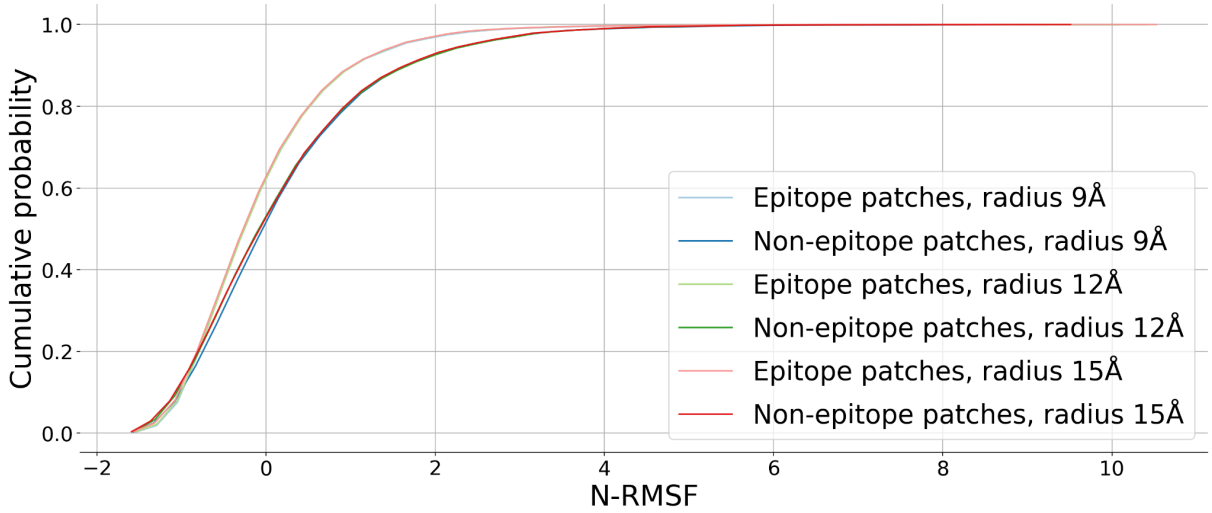



Figure 25: Cumulative probability plots representing the distribution function of N-RMSF values observed for AAs in Epitope (12 534, 21 831 and 36 545 values for radii 9Å, 12Å, and 15Å, respectively) and Non-epitope (43 695, 65 259 and 84 107 values for radii 9Å, 12Å, and 15Å, respectively) groups.

## 2.6. Chapter summary

In this chapter, the study of the structural properties of HLA antigens has been addressed, with particular emphasis on the comparison between the results obtained from MD trajectories and 3D static structures in order to highlight the contribution of the former.

First, the motivations for using MD simulations were presented. For example, MD simulations allow the correction of various stereochemical errors introduced by experimental methods. Also, MD data allows the analysis of dynamic properties of the structures. MD simulations allow the modeling of molecules under conditions closer to physiological conditions. This is different from the case of structures obtained by experimental methods such as X-ray crystallography, where in particular the crystallization conditions of the molecules modify the environment and the packing of the molecule. This point has been highlighted throughout the different structural properties studied in this chapter, where in some cases significant gaps were observed between the properties calculated from MD trajectories and 3D static structures. These results made it possible to highlight the contribution of molecular dynamics to the study of the structural properties of HLA antigens. Then, the protocol implemented to perform the MD simulations was presented in detail. Subsequently, the results of the structure quality check of the predicted/refined structures used as starting point for the MD simulations were presented. These results highlighted the usefulness of MD simulations for correcting stereochemical errors as mentioned above.

Regarding the study of the structural properties of HLA antigens, solvent accessibility was extensively addressed by considering several criteria. For example, different types of positions were distinguished, such as conserved positions, polymorphic positions, positions belonging to confirmed eplets, non-confirmed eplets or not listed as eplet at all. Similarly, results from MD trajectories or from static 3D structures were distinguished, and were aggregated by locus or by gene. All of these analyses provided a much deeper insight into the structure of HLA antigens. For example, the results reveal that a significant number of polymorphic positions are not solvent-accessible.

With respect to eplets, it was unexpectedly found that not all positions listed as belonging to confirmed eplets (also called antibody-verified) are solvent accessible. Since the number of confirmed eplet mismatches is a known risk factor for transplant rejection, a perspective of this work is to assess

whether mismatches of solvent accessible confirmed eplet AAs would be a better indicator of the risk of transplant rejection.

The loci A and C were also found to have the highest number of solvent-accessible positions repertorized as confirmed eplets. However, this result may be biased by the fact that class I loci (i.e. loci A, B, and C) have a higher number of confirmed eplets.

Also, the most recurrent solvent-accessible AA positions were identified among the confirmed eplets. This would be an indicator of an increased propensity for antibody recognition by these positions and could therefore be considered when assigning grafts.

Regarding AA-type distribution and their physico-chemical properties, the results are clearly different between the "Confirmed", "Non-confirmed" and "Non-eplet" groups of AAs. The different patterns observed in the “Non-confirmed” group when compared to the “Confirmed” one indicate that the former group is not only composed of true eplet candidates awaiting experimental confirmation but is likely a mixture of true and false eplets. This observation raises the need of a computational method capable of distinguishing among non-confirmed eplets those that may constitute true functional epitopes.

Finally, the side-chain flexibility of AAs during MD trajectories was investigated, and it was found that AAs belonging to confirmed eplets or to epitope patches tend to be less flexible than those considered as"non-eplet" or members of non-epitope patches.

All the structural properties explored in this chapter form the basis for building a machine-learning dataset that will allow the development of an HLA-epitope predictor capable of ascertaining the status of non-confirmed epitopes ahead of possible future experimental validation. This will be discussed in detail in Chapter 4.

# Chapter 3

# 3D shape descriptor for dynamic comparison of protein surfaces

## Summary



## 3.1. Comparative approach using Zernike polynomials

The shape complementarity between a ligand and a receptor (in the case of the present thesis, between an HLA antigen and an antibody) at the binding site is an essential element for a stable interaction [125]. In this Chapter, we address the problem of shape comparison between recipient and donor for improving graft allocation. In particular, we aim to compare 3D patches on the surface of HLA antigens taking into account shape variations during MD simulation. Thus, the challenge here is to compare sets of 3D shapes corresponding to the same epitope and extracted from sets of frames in MD trajectories.

### 3.1.1. Rotation- and translation-invariant shape representation for 3D surfaces

Moment-based representations, a class of mathematical descriptors originally developed for pattern recognition [126], describe a structure using a quantitative numerical representation that can be expanded as a series of orthogonal polynomials. The level of detail in the description can be controlled by selecting the appropriate limit on the order of the expansion [127]. Zernike polynomials, first introduced by Zernike [128], have been widely used in image retrieval. Canterakis generalized these polynomials to a 3D space [129], subsequently, Novotni and Klein adapted them for shape retrieval [130] and Pozo et al. introduced an efficient implementation on surface meshes [131]. The Zernike description has been effectively applied in structural biology for various purposes, including protein tertiary structure retrieval and comparison [132, 133], protein-protein docking [134], and shape-based ligand similarity searching [135]. Among the advantages that make Zernike polynomials (also known as Zernike moments) an attractive shape

descriptor are their invariance under rotation and translation, as well as their compact description as a vector of coefficients (or moments). In other words, this representation furnishes an ordered set of numbers that describes the geometrical shape of a surface, thus allowing for easy and rapid comparison of surfaces.

To obtain 3D Zernike moments, the 3D function $f(r,\ \theta,\ \varphi)$ describing a surface can be represented by a series expansion in an orthonormal sequence of polynomials :

$$f(r,\ \theta,\ \varphi) = \sum_{n=0}^{\infty} \sum_{l=0}^{n} \sum_{m=-l}^{l} C_{nlm} Z_{nl}^{m}(r,\ \theta,\ \varphi) \qquad 11)$$

where $r$, θ, φ are the three variables defining spherical coordinates in a 3D space (see https://mathinsight.org/spherical_coordinates for more information), $Z_{nl}^{m}(r,\ \theta,\ \varphi)$ are the 3D Zernike polynomials and $C_{nlm}$ are the Zernike moments. The indices $n$, $m$ and $l$ are integers (with $-\ l\ <\ m\ <\ l$, $0 \leq l \leq n$ and $n\ -\ l$ even) and are called order, degree and repetition, respectively. The equality in equation 11 only holds when the sum over n goes to ∞, but it can be truncated at the desired level of approximation at the expense of describing the surface at different levels of detail.

The Zernike polynomials can be written as:

$$Z_{nl}^{m}(r,\ \theta,\ \varphi) = R_{nl}(r)\, Y_{l}^{m}(\theta,\ \varphi) \qquad 12)$$

where R only depends on the radius r and is given by:

$$R_{nl}(r) = \sum_{k=0}^{\frac{n-l}{2}} N_{nlk} r^{n-2k} \qquad 13)$$

N being a normalization factor. The Y functions are complex Spherical Harmonics depending on both θ and φ. The 3D Zernike moments of $f(r,\ \theta,\ \varphi)$ are defined as the coefficients of the expansion of $f(r,\ \theta,\ \varphi)$ in the Zernike polynomial basis, i.e.:

$$C_{nlm} = \int_{|r| \leq 1} f(r)\, \overline{Z_{nl}^{m}}(r)\, dr \qquad 14)$$

where $\overline{Z_{nl}^{m}}(r)$ is the complex conjugate of the polynomial and $C_{nlm}$ are the Zernike moments. To obtain the descriptors invariant under translation and rotation, it is necessary to compute the norm (the sum over the index m) of the Zernike moments. Therefore, the 3D Zernike descriptors are defined as:

$$D_{nl} = \|C_{nlm}\| = \sqrt{\sum_{m=-l}^{l} (C_{nlm})^2} \quad 15)$$

### 3.1.2. Wasserstein distance

The problem of comparing two sets of 3D shapes each represented by a vector of coefficients can be reduced to comparing two distributions of points. Such a problem is related to Optimal Transport (OT), which is a mathematical problem that aims to find the most efficient way to move a mass between two distributions of points while minimizing the overall cost [136]. The cost of moving a unit of mass between two positions is called the ground cost. The optimization problem can be formulated for two distributions of points $X$ and $Y$ , where the objective is to find a mapping $m$: $X \rightarrow Y$ that completely transports the mass from one distribution to the other (this is accomplished when $m \sharp X = Y$, $with\ m \sharp X\ called\ the\ push - forward\ of\ X$) while minimizing the total cost. This constraint can be expressed as follows:

$$\min_{m, m\#X=Y} \int c(x, m(x)) dX(x) \quad 16)$$

where x is a mass unit and c(x, m(x)) the ground cost of pushing x from its position in X to its position in Y.

The Wasserstein distance, also known as the optimal value of the OT problem, measures a global distance between two distributions, based on the distance of each point of one distribution to each point of the other distribution. In the case of discrete distributions, the Wasserstein distance is expressed as:

$$W_p(X, Y) = \left(\sum_i \|X_i - Y_{m^*(i)}\|_p\right)^{1/p} \quad 17)$$

where $m^*$ denotes the mapping that minimizes the sum and p is the dimension of the metric space used (p=2 imply a 2D euclidean space for example).

One of the key features of the Wasserstein distance is the ability to compute meaningful sub-gradients, making it an efficient tool for machine learning applications that require measuring and optimizing similarity between empirical distributions [136].

In addition to the Wasserstein distance, the OT problem also provides an optimal mapping, known as the Monge mapping or OT matrix (see Figure 26). This mapping finds correspondences between the samples in each distribution, ensuring to have the minimal cost or distance between each point of one distribution to any point in the other distribution.. The OT matrix is estimated in a non-supervised way, making it valuable for problems involving transfer between datasets, such as color transfer between images or domain adaptation.

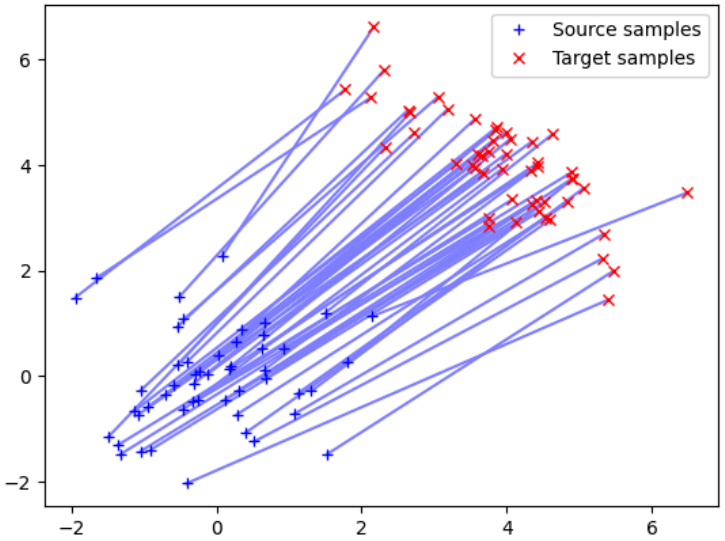



Figure 26: Plot of an example of OT matrix. The mapping between the points of the two distributions corresponds to the minimal cost of moving from one point in a distribution to any other in the other distribution. This mapping serves to compute the Wasserstein distance as the sum of these minimal distances.

In our case, the Wasserstein distance is computed between two sets of frames representing two 3D patches along MD simulation The implementation of the distance uses the POT python module.

### 3.1.3. Silhouette analysis

Silhouette analysis, introduced by Rousseeuw [137], is a widely used method for assessing the quality of clustering results and determining the optimal number of clusters in a dataset.

The silhouette coefficient is a measure of how well a data point fits into its assigned cluster compared to other clusters. For each data point *i*, the silhouette coefficient $s(i)$ is calculated as follows:

$$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))} \quad 18)$$

where a(i) is the average distance between data point i and all other points in the same cluster, and b(i) is the minimum average distance from data point i to all points in any other cluster [137].

The silhouette coefficient ranges from -1 to 1:

- A value close to 1 indicates that the data point is well-matched to its assigned cluster and poorly-matched to neighboring clusters.
- A value close to 0 suggests that the data point is on or very close to the decision boundary between two neighboring clusters.
- A negative value indicates that the data point may have been assigned to the wrong cluster.

## 3.2. Case-study of peptide [Pyr1]apelin-13: clustering of conformations during MD trajectory

Peptides are highly flexible, leading to diverse three-dimensional (3D) conformers that can modify their specificity and efficacy in binding to their targets. Identifying and understanding such 3D conformers

is therefore fundamental. This can be achieved by analyzing molecular dynamics (MD) runs. However, classical methods such as RMSD maps are difficult to process automatically, and manual interpretation can be subjective. For this reason, we wanted to explore the use of Zernike descriptors for the identification of conformers. As a case study, the [Pyr1]apelin-13 [138] peptide (with

sequence pGLU-ARG-PRO-ARG-LEU-SER-HIS-LYS-GLY-PRO-MET-PRO-PHE, see Figure 27) was used. A 100ns MD simulation was performed using NAMD and CHARMM force-field, from which a 1000-frame trajectory was extracted to represent the conformational space. For each frame, a 3D Zernike descriptor was computed to represent the 3D shape of the peptide. The 3D shape is the solvent-accessible surface, constructed from the Van der Waals radius sphere associated with each atom type and obtained with Chimerax [139] in STL format (STL from STereo-Lithography) and transformed to VTK format (VTK from Visualization Toolkit) using the python module *meshio*. The MindBoggle [140] python library is then adapted to calculate the 3D Zernike descriptor relying on 121 moments.

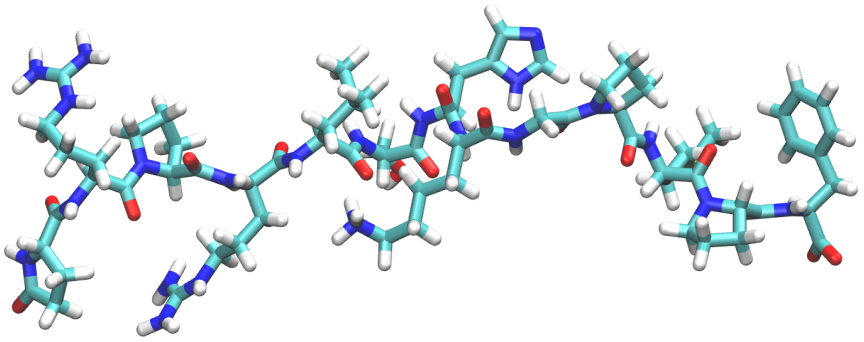


Figure 27: Licorice visualization of the peptide [Pyr1]apelin-13.

Conformer identification was performed by clustering the previously calculated 3D Zernike descriptors, using the K-means and silhouette analysis algorithms in the scikit-learn implementation to identify the optimal number of clusters. For each cluster, the conformation closest to its centroid is chosen as a representative. According to silhouette analysis, the optimal number of clusters is 2 (see Figure 28).

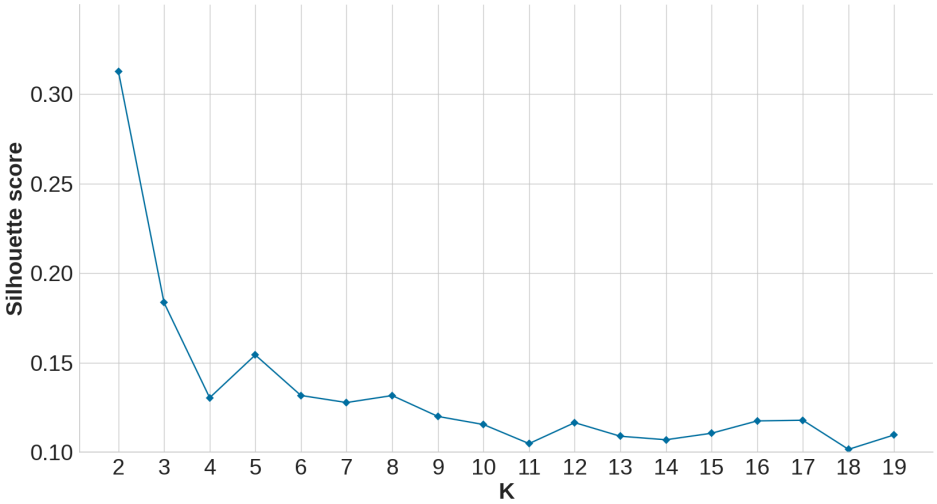



Figure 28: Silhouette scores for different K in K-means clustering of peptide MD frames using Zernike descriptors. Each value corresponds to the average of silhouette coefficients for all peptide frames.

The quality of the clustering for these two clusters is shown in Figure 29. A good definition of the clusters is obtained, where it is worth highlighting the very small number of peptide frames that present a poor classification (negative silhouette coefficients).

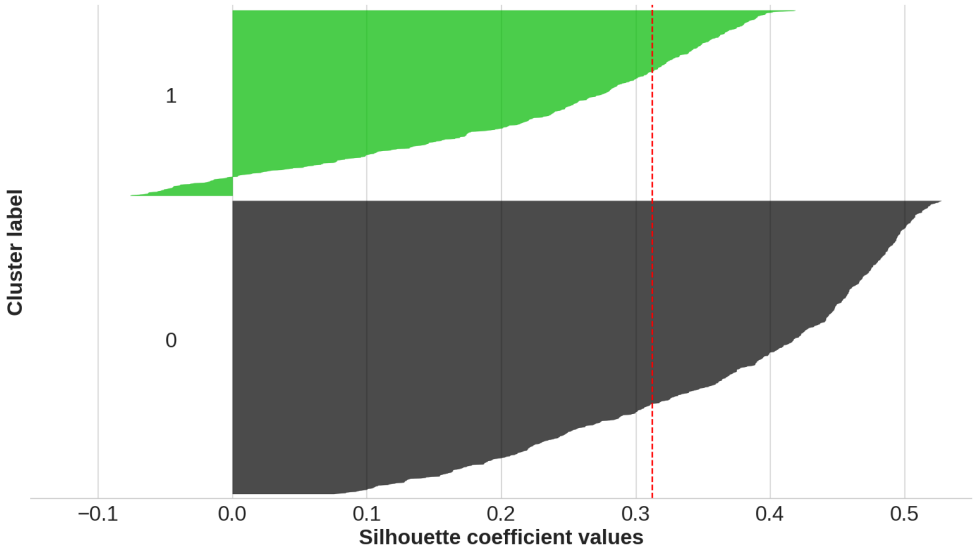



Figure 29: Silhouette analysis for K-means (K=2) clustering of peptide MD frames using their 3D Zernike descriptors. The x-axis represents the silhouette coefficient values. The y-axis represents the data points, ordered by their silhouette coefficient values within each cluster. The width of each silhouette is proportional to the number of data points in the corresponding cluster. The dashed line corresponds to the average silhouette score of 0.312.

The representative conformers of each cluster correspond to frames 748 and 840 and are shown in Figure 30.

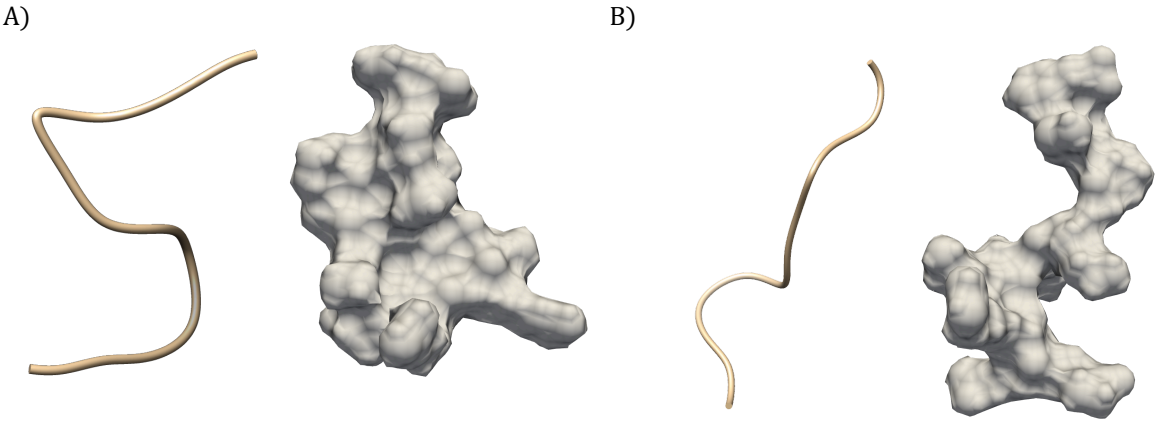



Figure 30: A) Backbone (on the left) and surface (on the right) visualization of the representative conformer for the Cluster #0 (frame 748). B) Backbone (on the left) and surface (on the right) visualization of the representative conformer for the Cluster #1 (frame 840).

The 3D Zernike descriptors have already been used for protein tertiary structure comparison but to our knowledge, not to identify 3D-conformers from MD runs. The Zernike vector format allows using a variety of clustering methods to process MD frames and to automatically determine the optimal number of clusters/conformers. This strategy could be promising for other biomolecules such as proteins or nucleic acids.

This work was presented as a poster at GGMM (Groupe de Graphisme Moléculaire et Modélisation) and SFCI (Société Française de ChémoInformatique) join conference in Lille in 2021.

## 3.3. Case-study of patches of BW4/BW6 serological groups

As mentioned in section 2.4.1.5, the BW4 and BW6 groups are two well-known serologic groups, which is why these groups will be used as case-study. In particular, the focus will again be on position 80 and its surroundings, which is considered to be the origin of the epitope defining these groups. For this purpose, Zernike descriptors will be used to compare the shape of the patches centered at a given position along the MD trajectories. To do this, 50 frames were taken from the last 500 frames of the trajectory (1 frame every 10 frames), i.e., from the last 5ns of the simulation. Patches of 7Å size (approximate size of an eplet) were generated, centered at positions 80, 82, 83 (suspected eplet residues), 138, and 226 (control positions corresponding to conserved regions, see Figure S22). Chimerax was used to calculate the patch surfaces, while the python module *meshio* was used to transform them from the STL format generated by Chimerax to VTK format. For the calculation of Zernike moments, the implementation of Mindboggle was adapted, truncating the order of the series to 40 (n=40) and therefore yielding 441 coefficients for each patch. In this way, each antigen considered will have 50 Zernike descriptors for each position studied, such that when making a comparison between two antigens at a given position, two distributions of Zernike descriptors must be compared (each distribution representing an MD trajectory). For this, the Wasserstein distance was used, which, as explained in section 3.1.2, is a method for the evaluation of the similarity between two sets/distributions of samples. In the present case study, 8 antigens will be considered, 4 from the BW4 group and 4 from the BW6 group (see Table 15).

Table 15: Antigens considered in the present case study.

| **BW4** | **BW6** |
|---|---|
| B*15:16 | B*07:02 |
| B*27:05 | B*15:01 |
| B*37:01 | B*27:08 |
| B*49:01 | B*50:01 |

In order to have an idea of the shape of the patches that were to be analyzed, a visual inspection of them was performed. In the case of position 82 (Figure 31), differences in 3D surface shapes are clearly detected visually between the BW4 and BW6 groups, which is in agreement with the results presented in Chapter 2 where evidence was found that there is a difference at the RSASA level between both groups for this position during MD trajectories (clusters could be distinguished in the heatmap representing the K-S test statistics between RSASA distributions calculated in BW4 and BW6 HLA antigens, see Figure S16E). At positions 80 and 83 (see Figure S23 and Figure S24), such shape differences are more difficult to identify visually, despite the fact that the results of analyses with the K-S test statistics suggested the existence of a difference in terms of RSASA (see heatmaps in Figure 15 and Figure S16F). Concerning the conserved positions 138 and 226, no noticeable difference can be detected visually (see Figure S25 and Figure S26).

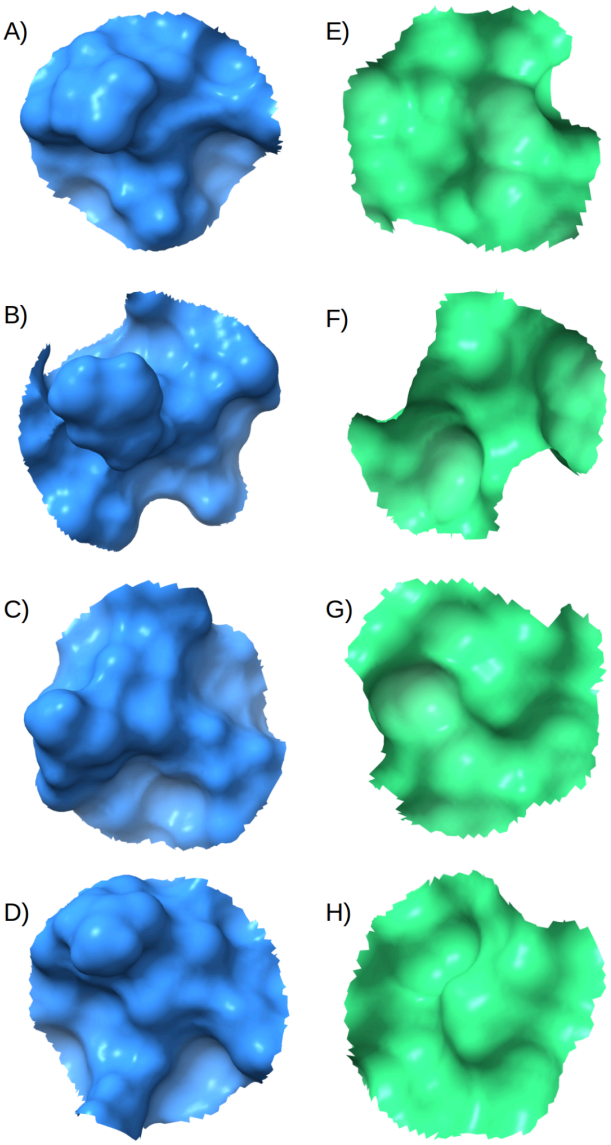



Figure 31: Examples of 7Å-sized patches centered at position 82 used for the calculation of Zernike descriptors. A) Antigen B*15:16. B) Antigen B*27:05. C) Antigen B*37:01. D) Antigen B*49:01. E) Antigen B*07:02. F) Antigen B*15:01. G) Antigen B*27:08. H) Antigen B*50:01. The mesh of each patch consisted of approximately 13K points.

The results of Wasserstein distance calculation between the Zernike descriptors of 3D patches computed at positions 80 and 82 are presented in Figure 32, panels A and B respectively. In the case of position 80 (panel A), it appears that the distances calculated within the Bw6 group of antigens are all rather low, indicating a rather homogenous shape proximity. By contrast, major differences are observed within the BW4 group with two antigens B*27:05 and B*37:01 showing a distance to antigens B*49:01 and B*15:16 from the same group (BW4) as large as the distance to the BW6 antigens. In line with this observation, an apparent similarity of antigens B*15:16 and B*49:01 with BW6 antigens is observed. In the case of position 82 (panel B), all distances are very low suggesting that our method does not detect any difference between the patches at this position, whatever the serological group. In the case of position 83 (Figure S28), only antigen B*15:01 and to a lower extent B*07:02 show large distances to all other antigens, whatever the serological group. These results go against what would be expected, that is, small distances between antigens of the same group and larger distances between antigens of different groups. Moreover, regarding the control conserved positions, at position 138 (Figure S29), large distances are observed to all other antigens for B*15:01 and albeit to a lower extent for B*15:16. ). At conserved

position 226 (Figure S30), the situation is less extreme with reasonably low distances between all antigens, whatever the serological group, except for antigen B*49:01 that displays a slightly higher distance to the rest of the antigens, although this time less pronounced than at position 138. These results are difficult to interpret as one would expect small distances between all antigens at conserved positions.

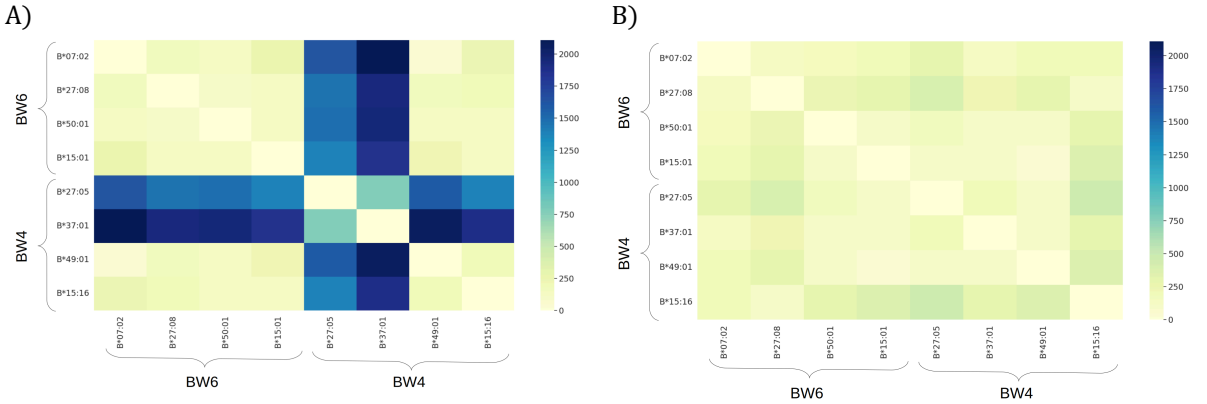



Figure 32: Heatmap of the Wasserstein distances between the patches of positions 80 (panel A) and 82 (panel B) of the antigens from the BW4 and BW6 groups.

## 3.4. Discussion

The first case-study reported in this chapter shows how successful it can be to analyze conformation diversity during MD simulation with Zernicke descriptors. In the case of the peptide studied here, two conformers were clearly identified. Classical methods use RMSD heatmaps that may reveal more difficult to cluster than a set of Zenicke descriptors. In particular, the very rapid and efficient K-means algorithm does not apply on a distance matrix. The cost of computing Zernicke coefficients on peptide or protein 3D surfaces, followed by K-means and silhouette calculation should be compared with RMSD distance matrix calculation and heatmap production.

The second use-case presented in this chapter concerns the use of 3D shape descriptors for characterizing/quantifying the differences existing between recipient and donor antigens at the level of 3D surfaces around eplets during MD simulations. Despite testing various parameters (results not presented here), such as the way to generate the patches (using a fixed number of neighboring AAs to the central AA or using closed surfaces), patch sizes, mesh densities, Zernike descriptor sizes, and other serological case-studies, it was never possible to obtain results consistent with biological knowledge. Several clues arise to explain these negative results. Firstly, generating consistent surfaces is a challenge in itself. Despite the various strategies used to generate surfaces, artifacts or abnormal behaviors were always encountered (see Figure S23E, S24B, S24D, and S25C). Secondly, it is possible that shape alone is not sufficient to explain serological groups, as it is well known that physicochemical properties play a key role in antigen-antibody recognition. Due to these difficulties, the approach using Zernike descriptors was abandoned in this thesis.

## 3.5. Chapter summary

In this chapter, the use of shape descriptors for characterizing eplet patches was explored, motivated by the importance of shape complementarity between ligands and receptors in stable interactions. The main objectives were to compare protein surfaces and to develop a shape descriptor that could be incorporated into an HLA B-cell epitope predictor (see Chapter 4). Zernike moments, a type of moment-based representation, were introduced as a promising shape descriptor, offering advantages such

as invariance under rotation and translation, and a compact description as a vector of coefficients. The Wasserstein distance, derived from the Optimal Transport (OT) problem, was presented as a measure of similarity between shape distributions during MD trajectories. Silhouette analysis was introduced as a method for assessing the quality of clustering results and determining the optimal number of clusters in a dataset, with the silhouette coefficient measuring how well a data point fits into its assigned cluster compared to other clusters.

Two case-studies were presented. The first one focuses on the identification of conformers during MD trajectories. It was applied to a peptide, [Pyr1]apelin-13, using Zernike descriptors, during a 100ns MD simulation. The 3D Zernike descriptors were calculated for 1000 frames to represent the 3D conformations of the peptide during MD trajectory. K-means clustering and silhouette analysis revealed clearly the existence of two main conformers. The second case-study involves using Zernike descriptors to compare the shape of patches centered at positions 80, 82, 83, 138, and 226 along the MD trajectories of 8 antigens in the context of the BW4/BW6 serological groups (4 antigens from the BW4 group and 4 antigens from the BW6 group), with the Wasserstein distance employed to evaluate the similarity between the Zernike descriptor distributions of different antigens at each position during MD trajectories. While the peptide [Pyr1]apelin-13 conformer identification case study successfully demonstrated the potential of Zernike descriptors for automatically determining the optimal number of clusters/conformers from MD runs, the BW4/BW6 serological group patch comparison did not yield the expected differentiation between the groups, despite visual differences in the patches at position 82. Inconsistencies were also observed at positions 80, 83, and the control positions (138 and 226), attributed to challenges in generating consistent 3D surfaces and the potential insufficiency of shape alone in explaining serological groups. Future research could explore alternative mesh/surface generation techniques or explore different descriptors able to catch physicochemical properties to improve the characterization of HLA B-cell eplets/epitopes.

# Chapter 4

# HLA-EpiCheck

## Summary



## 4.1. Motivations for an HLA B-cell epitope predictor

Recognition of HLA proteins by Donor-Specific Antibodies (DSA) is the main cause of organ transplant loss. This is due, firstly, to the fact that these membrane proteins are expressed in almost all nucleated cells of the organism, i.e. they are practically omnipresent [10]. Secondly, the genes encoding these proteins are the most polymorphic genes in humans. As a consequence, it is extremely difficult to find HLA identical donors and recipients outside siblings, and therefore transplantations are almost always performed across HLA polymorphisms that may later on trigger an immune response (both humoral and cellular) against the donor. However, the cellular response is usually well controlled by immunosuppressive treatments, leaving the humoral response as the major risk for graft rejection (i.e. the process of recognition of HLA antigens by antibodies). B-cell epitopes are protein regions recognized and bound by B-cell produced antibodies and eplets are considered the key components of these epitopes in the case of HLA antigens. Thus, understanding the characteristics of HLA B-cell epitopes and predicting their antigenicity could lead to a better definition of HLA matching between donor and recipient to reduce *de novo* DSA formation and graft rejection.

In this chapter HLA-EpiCheck is presented, a machine learning predictor for B-cell epitopes on HLA proteins that leverages the unprecedented set of molecular dynamics simulations of 207 HLA proteins introduced in chapter 2.

## 4.2. 3D-surface patch definition of HLA epitopes

The HLA epitope predictor presented in this chapter relies on the use of 3D surface patches. Each of these patches is centered on a single solvent-accessible AA, and 18 molecular descriptors are calculated from both static and dynamic properties of this patch. These patches and their descriptors form the ML dataset used to train HLA-EpiCheck.

The aforementioned patches are fragments of the protein surface which contain all solvent-accessible AA within a given radius from the patch's central AA (distances are computed between the centers of mass of the AAs). The effect of radius size on patch composition and properties has been described in detail in Chapter 2 and the results shown in this chapter only deal with the larger size of 15Å. Also, and as already mentioned in chapter 2, a median RSASA of 20% along the MD trajectory is the threshold used to define solvent accessibility of an AA.

HLA-EpiCheck is a binary classifier where the two classes to be predicted are "epitope" or "non-epitope". In other words, HLA-EpiCheck will predict whether a 3D-surface patch is an HLA epitope or not. The labeling of each patch in the ML dataset was done as follows. Any patch whose central AA belongs to a confirmed eplet in the HLA Eplet Registry is labeled as epitope patch. By contrast, the non-epitope patches are centered on a solvent-accessible AA that is not listed as part of an eplet in the HLA Eplet Registry (either confirmed or not). However, not all such patches are eligible as negative counterparts of epitope patches. A first condition is that, in addition to the central AA, none of the solvent-accessible AAs composing a non-epitope patch should be part of an eplet (either confirmed or not). Another condition is that no non-epitope patch can be composed exclusively of locus-conserved positions. Indeed, considering locus-conserved positions could be misleading as these positions could by chance display B-cell epitope behavior but cannot be recognized as such because they cannot raise immune response as they are conserved throughout all individuals. Thus, considering these locus-conserved positions could bias the training by introducing « silencious » positive examples among the non-epitope patches. Homemade Pymol and Python scripts were written to compute the patches as described in Chapter 2. In total, 2 117 epitope and 4 769 non-epitope 3D surface patches were computed from the 3D structure and MD trajectories of the 207 antigens considered in this study.

## 4.3. Model selection for HLA-EpiCheck

### 4.3.1. Generation and description of the machine learning datasets

The 18 descriptors used to describe the 3D-surface patches of the ML dataset can be classified into two types, static and dynamic (Table 16).

Table 16: Descriptor dictionary of the ML dataset.

**Patch Descriptor Dictionary**

| Name | Description (per patch) | Descriptor type | | Value range for the dataset |
|---|---|---|---|---|
| | | Static | Dynamic | |
| H_central | Hydrophobicity of central AA | x | | [-4.5, 4.5] ☩ |
| H_patch_min | Minimum AA hydrophobicity | x | | [-4.5, -3.5] ☩ |
| H_patch_max | Maximum AA hydrophobicity | x | | [-0.8, 4.5] ☩ |
| H_patch_avg | Average of all AA hydrophobicities | x | | [-3.2, 0.3] ☩ |
| Pos_central | Positive charge of central AA | x | | [0, 1] |
| Pos_patch | Sum of positive charges of all AAs | x | | [0, 11] |
| Neg_central | Negative charge of central AA | x | | [-1, 0] |
| Neg_patch | Sum of negative charges of all AAs | x | | [-11, 0] |
| S_central_min | Minimum RSASA of central AA over MD | | x | [0, 82.6] % |
| S_central_max | Maximum RSASA of central AA over MD | | x | [21.6, 133] % |
| S_central_median | Median RSASA of central AA over MD | | x | [10.2, 105.6] % |
| S_patch_min | Weighted* average of all AAs minimum RSASAs over MD | | x | [9.2, 30.3] % |
| S_patch_max | Weighted average of all AAs maximum RSASAs over MD | | x | [26.4, 62.6] % |
| S_patch_avg | Weighted average of all AAs median RSASAs over MD | | x | [18.8, 43.2] % |
| F_central | N-RMSF of central AA | | x | [-2, 11.3] |
| F_patch_min | Minimum AA N-RMSF | | x | [-2, 0.4] |
| F_patch_max | Maximum AA N-RMSF | | x | [-2, 17.7] |
| F_patch_Avg | Weighted average of all AA N-RMSFs | | x | [-0.91, 1.85] |

* Weights (of residue values) correspond to the frequency of presence of each residue in the patch over MD (% of frames).

☩ Kyte-Doolittle hydrophobicity scale was used.

#### 4.3.1.1. Static descriptors

Static descriptors do not vary along MD simulation. AA hydrophobicity is computed according to the Kyte-Doolittle scale [45] and leads to 4 descriptors for a patch: the hydrophobicity of the central AA, the minimum, maximum and average AA hydrophobicity over all AAs of the patch.

The other static descriptors are derived from the AA electrostatic charges, namely a negative charge for aspartate (D) and glutamate (E) AAs, and a positive charge for lysine (K) and arginine (R) AAs. First, the charge of the central AA is described: the descriptor "Pos_central" takes the value +1 for a positively charged AA, otherwise 0, and the same for the descriptor Neg_central except that the value is -1 for a negatively charged AA, Second, the sum of the charges (either positive or negative) of the AAs within the patch is calculated (integer $\geq 0$ for the descriptor "Pos_patch", integer $\leq 0$ for the descriptor "Neg_patch").This gives four descriptors for each patch, two for the negative charges and two for the positive charges.

#### 4.3.1.2. Dynamic descriptors

Dynamic descriptors measure conformational changes of the patch structure during MD simulation. A first group of 6 descriptors reflects the variation of the RSASA percentage during the run. RSASA values are calculated per residue for each of the 500 frames considered. For the central AA of a patch, the minimum,

maximum and median RSASA values along the run are kept. For the entire patch, the minimum, maximum and median RSASA values of each residue in the patch along the run are used and these values are aggregated according to the type (all minima, all maxima, all medians) as three weighted averages in which the weights correspond for each residue to the % of frames in which this residue is really present in the patch during the MD trajectory (distance to the central residue within the patch radius).

A second group of 4 descriptors reflects side-chain flexibility along the run. These descriptors are derived from the N-RMSF values. One corresponds to the N-RMSF of the central AA of a patch and the others to the minimum, maximum and weighted average N-RMSF over all residues of the patch (weights are the same as defined above). RMSF values are calculated from the 500 frames considered for each trajectory using VMD [116]. The Z-score method is used for N-RMSF normalization and includes all solvent-accessible AAs except for five residues at each terminus because of their artifactual flexibility.

#### 4.3.1.3. Dataset composition and evaluation metrics

HLA-EpiCheck is a binary classifier that predicts whether the patch centered on a solvent-accessible residue is an epitope or not. It is important to highlight that the ML datasets generated here present important imbalances between labels as can be seen in Table 17 and Table 18. Therefore, the precision and recall metrics are more pertinent to evaluate the ability of a binary classifier to predict a label. The F1 score (harmonic mean of the precision and recall) and Matthews Correlation Coefficient (MCC) metrics are used to evaluate the predictor [141]. Scikit-Learn [142] was used to compute all the metrics. Exploratory experiments (not presented here) with various learning algorithms using datasets derived from the 3 different patch sizes considered in Chapter 2 showed that the best performances are obtained with the 15Å patches. Therefore, all the results presented in this chapter were obtained using this patch size.

Two ML datasets have been generated: the complete dataset called “Redundant” dataset (Table 18) and a reduced "Non-redundant" dataset (Table 17) which corresponds to the methodology typically used in bioinformatics where a reduction of redundancy with respect to sequences is performed. Here, sequences that have a sequence identity higher than 90% are removed . However, due to the enormous loss of samples (~90%) in the "Non-redundant" case, the complete redundant dataset was also explored in hopes of getting better performance results. K-Nearest Neighbor scales well with redundant data because it is not designed to learn a model from the data. This method rather uses the “instance-based learning” paradigm and classifies new examples on the basis of their similarity with existing examples in the dataset. It should be noted that the present case study has some peculiarities that distinguishes it from a typical machine learning problem. For example, the HLA proteins tested in the Luminex Bead Assay have not changed in over 10 years, largely because the current set extensively covers the alleles present in the population served. For this reason, it is not expected to change substantially in the future. This leads to the fact that in the case of the HLA epitope predictor presented here, a better ability to predict epitopes (despite the use of a redundant dataset) is more interesting than a predictor that manages to generalize over the data and thus build a learning model from it. This will be discussed in more detail later in this chapter (see Table 18).

Regarding the "Non-redundant" dataset, MMseqs2 [92] was used to perform the redundancy reduction using a similarity threshold of 90%. MMseqs2 clustered the 189 sequences of alleles present in the 207 modeled HLA antigens according to the given similarity threshold, i.e., sequences with a similarity equal to or greater than 90% were grouped into the same cluster. In this way, MMseqs2 found 17 clusters. Then, the sequence with the highest number of "Non-eplet" AAs was chosen as the representative of each cluster and all the “Non-epitope” patches centered on these AAs were integrated into the dataset. Since the antigens in the same cluster can carry different eplets, all the "Epitope" patches derived from all the antigens in the cluster were included. The 90% similarity threshold may appear high when compared to what is usually used in the literature, but this is due to the enormous similarity between HLA proteins. For example, in Table 17, it can be seen that with this threshold, the dataset is reduced to one-tenth of the initial dataset.

The split into training and test sets was performed in such a way that 90% of the samples from each label (Epitope/Non-epitope) were taken to form the training set, and consequently, the remaining 10% of each label formed the test set. This division is due to the small number of samples available and the fact that in the clusters generated by MMseqs2, certain clusters have a large number of antigens while others could have only one antigen. Therefore, performing a split based on clusters/antigens would generate a tremendously biased and unbalanced dataset in terms of the representativeness of the different antigens and loci. Additionally, 10 different splits (i.e., 10 different training/test set pairs) were generated in order to evaluate the generalization capacity of the predictors as well as to verify if the performances are split-dependent.

Table 17: Composition of "Non-redundant" ML dataset. Number of samples are aggregated by loci and by class (Epitope/Non-epitope). Due to the generation of 10 different splits, the mean and standard deviation in the composition by locus are presented.

| | | Dataset | | Training sets | | | | Test sets | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Epitope | | Non-epitope | | Epitope | | Non-epitope | |
| | | Epitope | Non-epitope | Mean | Std. | Mean | Std. | Mean | Std. | Mean | Std. |
| # samples | **A** | 42 | 45 | 37.80 | 1.81 | 40.70 | 1.06 | 4.20 | 1.81 | 4.30 | 1.06 |
| | **B** | 64 | 199 | 56.90 | 3.51 | 177.90 | 3.07 | 7.10 | 3.51 | 21.10 | 3.07 |
| | **C** | 11 | 0 | 10.20 | 0.92 | 0 | 0 | 0.80 | 0.92 | 0 | 0 |
| | **DP** | 15 | 89 | 12.80 | 0.92 | 80.20 | 2.39 | 2.20 | 1.03 | 8.80 | 2.39 |
| | **DQ** | 61 | 27 | 55.00 | 2.00 | 24.40 | 1.67 | 6.00 | 2.00 | 2.60 | 1.58 |
| | **DR** | 50 | 61 | 45.30 | 1.95 | 54.80 | 1.87 | 4.70 | 1.95 | 6.20 | 1.87 |
| | **Total** | **243** | **421** | **218** | | **378** | | **25** | | **43** | |

Regarding the "redundant" dataset, the division between training and test sets was made with 80%/20% proportions with respect to the antigens, i.e., the training set consists of all the 3D patches from 80% of the antigens (i.e., 163 antigens) which were chosen in a stratified manner with respect to the loci (see Table 18). In view of the large counts of samples in both Epitope and Non-epitope classes in the “redundant” dataset and given that some exploratory experiments (not presented here) indicated a very little impact of using different splits on the performance of various predictors, only one split was used with this dataset to rationalize computation times.

Table 18: Composition of "Redundant" ML dataset. Number of samples are aggregated by loci and by patch type (Epitope/Non-epitope) while the number of antigens is aggregated by locus.

| | | Dataset | | Training set | | Test set | |
|---|---|---|---|---|---|---|---|
| | | Epitope | Non-epitope | Epitope | Non-epitope | Epitope | Non-epitope |
| # samples | A | 434 | 656 | 353 | 506 | 81 | 150 |
| | B | 529 | 2822 | 427 | 2270 | 102 | 552 |
| | C | 177 | 359 | 134 | 281 | 43 | 78 |
| | DP | 184 | 637 | 138 | 529 | 46 | 108 |
| | DQ | 486 | 107 | 395 | 95 | 91 | 12 |
| | DR | 307 | 188 | 241 | 141 | 66 | 47 |
| | **Total** | **2117** | **4769** | **1688** | **3822** | **429** | **947** |
| # antigens | A | 34 | | 27 | | 7 | |
| | B | 59 | | 47 | | 12 | |
| | C | 17 | | 13 | | 4 | |
| | DP | 31 | | 24 | | 7 | |
| | DQ | 30 | | 24 | | 6 | |
| | DR | 36 | | 28 | | 8 | |
| | **Total** | **207** | | **163** | | **44** | |

### 4.3.2. Model selection

A broad model selection process was conducted to determine the most suitable learning algorithm for training the HLA-EpiCheck predictor. With the "Non-redundant" dataset, the following learning algorithms were used: Decision Tree (DT), Extremely Randomized Trees (here named ExtraTrees), Gradient Boosted Decision Trees (here named GradientTrees) and K-Nearest Neighbor (KNN). With the "Redundant" dataset, the following algorithms were employed: DT, Random Forest (RF), ExtraTrees, GradientTrees, Histogram-based Gradient Boosted Trees (here named HistoGradientTrees), KNN, and Multi-Layer Perceptron (MLP). To ensure a robust and reliable model selection process, a combination of gridsearch and cross-validation techniques using the scikit-learn library was implemented. Gridsearch was used to explore and tune the hyperparameters of each learning algorithm on the training sets of both datasets ("Non-redundant" and "Redundant"), while cross-validation was performed using the RepeatedStratifiedKFold function from Scikit-Learn with 10 folds and 10 repetitions also on the training set of both datasets. Table S11 shows the set of parameters explored in the gridsearch for each learning algorithm and highlights the best parameter values that were selected for testing. This approach helps to assess the performance and generalization ability of each predictor by evaluating them on multiple subsets of the data. The F1-score was chosen as the evaluation metric for model selection, as it provides a balanced measure of precision and recall, making it suitable for imbalanced datasets. Min-Max normalization on input data was used for the following learning algorithms : MLP and KNN (Figure 33 shows a graphical description of the model selection implementation).

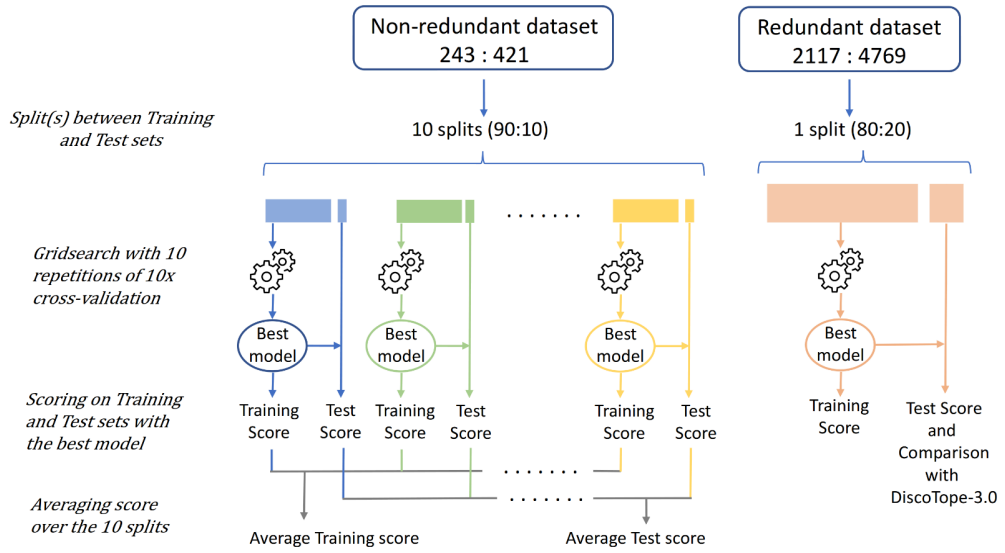



Figure 33: Machine learning setup for model selection. This setup was reproduced for every machine learning algorithm tested[14].

The performance obtained with the best hyperparameter values for each predictor on the training set are displayed in Table 19 for the "Non-redundant" and Table 20 for the "Redundant" predictors.

Table 19: Cross-validation results of the best model found through gridsearch for each learning algorithm using the "Non-redundant" dataset. The results correspond to the mean and standard deviation of each of the 10 splits. 10 repetitions calculated. The F1-score, precision and recall metrics were calculated for each label (Epitope/Non-epitope).

| | | F1 | | Precision | | Recall | |
|---|---|---|---|---|---|---|---|
| | | Mean | Std. | Mean | Std. | Mean | Std. |
| **Decision Tree** | **Epitope** | 0.63 | 0.07 | 0.67 | 0.08 | 0.61 | 0.10 |
| | **Non-epitope** | 0.80 | 0.04 | 0.79 | 0.05 | 0.82 | 0.06 |
| **ExtraTrees** | **Epitope** | 0.70 | 0.08 | 0.82 | 0.10 | 0.62 | 0.10 |
| | **Non-epitope** | 0.86 | 0.04 | 0.81 | 0.04 | 0.92 | 0.05 |
| **GradientTrees** | **Epitope** | 0.76 | 0.08 | 0.83 | 0.08 | 0.70 | 0.10 |
| | **Non-epitope** | 0.88 | 0.04 | 0.84 | 0.05 | 0.91 | 0.05 |
| **K-NN (K=3)** | **Epitope** | 0.56 | 0.09 | 0.67 | 0.09 | 0.49 | 0.10 |
| | **Non-epitope** | 0.80 | 0.04 | 0.75 | 0.04 | 0.86 | 0.05 |

For the “Non-redundant” dataset, we observe that the performance is always higher for the Non-epitope class, with recall values greater than precision values. This may be due to data imbalance in favor of this class. Regarding the “Epitope” class, the F1 score increases from 0.63 to 0.70 and 0.76 between simple decision tree and more sophisticated tree-based algorithms (ExtraTrees and GradientTrees). In all three cases the precision value is greater than the recall value, indicating a greater capacity of the model to discriminate between epitope and non epitope patches than to retrieve all epitope patches from the dataset. In other words, assuming that the “Epitope” class is the positive one, the three systems generate less false positive than false negative predictions. Precision is the proportion of predicted positive cases that are actually positive, while recall is the proportion of all positive cases that are correctly

[14] Credit icon cogwheels: https://www.flaticon.com/fr/icones-gratuites/roue-dentee

predicted. Therefore, a higher precision metric indicates that the model is less likely to make false positives and a lower recall that the model is more likely to leave positive examples as false negative.
The K-NN algorithm performs the worst with an F1 score at 0.56 and particularly poor recall value (0.49). It should be remembered that K-NN is an instance-based rather than a model-based classification method. Thus, the more examples in the dataset, the better performance. We will see below that it works better with the "Redundant" dataset. The results obtained on the 10 test sets of the 10 splits of the "Non-redundant" dataset are very consistent with those obtained on the training set and are shown in Table S12.

In Table 20, the results of the cross-validation step under optimal hyperparameters on the "Redundant" training set reveal that the tree-based algorithms and the K-NN method have much better performance here than when they were trained on the "Non-redundant" training sets. The MLP algorithm performs the worst, especially on the "Epitope" class (F1 score 0.56). The interpretation of the higher scores obtained with the "Non-epitope" class and of the precision values being greater than the recall ones for the "Epitope" class, is the same as for the "Non-redundant" dataset. Moreover, due to the high-redundancy of HLA sequences and 3D structures from which this dataset is derived, it is very likely that the models built here with the tree-based algorithms are overfitted. However, one should keep in mind that the purpose of the HLA epitope predictor we want to build is to predict HLA epitopes that are all very similar to each other. In this particular case, the redundancy does not harm as long as the models can accommodate it. This is the case of the tree-based models which can become very detailed to the point of describing almost every example. On the contrary, an algorithm such as MLP is less successful because it struggles to find trends allowing certain coefficients to be reinforced in the neural network. The acceptability of the tree-based models built on the "Redundant" dataset is fully confirmed by the fact that their performance does not drop when calculated on the test set (see Table S13).

Table 20: Cross-validation results of the best model found through the gridsearch for each learning algorithm using the "Redundant" dataset. The results correspond to the mean and standard deviation of each of the 10 folds and 10 repetitions calculated. The F1-score, precision and recall metrics were calculated for each label (Epitope/Non-epitope).

| | | F1 | | Precision | | Recall | |
|---|---|---|---|---|---|---|---|
| | | Mean | Std. | Mean | Std. | Mean | Std. |
| **Decision Tree** | **Epitope** | 0.70 | 0.03 | 0.71 | 0.03 | 0.71 | 0.03 |
| | **Non-epitope** | 0.87 | 0.01 | 0.87 | 0.01 | 0.87 | 0.02 |
| **Random Forest** | **Epitope** | 0.84 | 0.02 | 0.88 | 0.02 | 0.80 | 0.03 |
| | **Non-epitope** | 0.93 | 0.01 | 0.92 | 0.01 | 0.95 | 0.01 |
| **ExtraTrees** | **Epitope** | 0.88 | 0.02 | 0.92 | 0.02 | 0.83 | 0.03 |
| | **Non-epitope** | 0.95 | 0.01 | 0.93 | 0.01 | 0.97 | 0.01 |
| **GradientTrees** | **Epitope** | 0.84 | 0.01 | 0.87 | 0.02 | 0.82 | 0.02 |
| | **Non-epitope** | 0.93 | 0.01 | 0.92 | 0.01 | 0.95 | 0.01 |
| **HistoGradientTrees** | **Epitope** | 0.87 | 0.02 | 0.91 | 0.02 | 0.86 | 0.03 |
| | **Non-epitope** | 0.95 | 0.01 | 0.94 | 0.01 | 0.96 | 0.01 |
| **MLP** | **Epitope** | 0.60 | 0.02 | 0.58 | 0.03 | 0.61 | 0.02 |
| | **Non-epitope** | 0.81 | 0.01 | 0.82 | 0.01 | 0.81 | 0.02 |
| **KNN (K=3)** | **Epitope** | 0.81 | 0.02 | 0.85 | 0.03 | 0.77 | 0.03 |
| | **Non-epitope** | 0.92 | 0.01 | 0.90 | 0.01 | 0.94 | 0.01 |

In conclusion of this model selection step, the ExtraTrees model shows the best performance when trained on the "Redundant" dataset, which makes it the predictor that will be retained in the final version of HLA-EpiCheck. The average precision and recall over the 10 repetitions are $0.92 \pm 0.02$ and $0.83 \pm 0.03$ for the epitope label yielding average F1 score of $0.88 \pm 0.02$. The average Area Under the Curve of the Receiver Operating Characteristic curve (AUC-ROC), Area Under the Curve of the Precision-Recall curve (AUC-PR) and MCC were also computed, yielding $0.98 \pm 0.005$, $0.96 \pm 0.01$ and $0.83 \pm 0.02$, respectively, for the epitope label. Low standard deviation values indicate an homogeneous and robust performance of the predictor over the entire training set (Figure S31 and Figure S32). This indicates that the ExtraTrees model is performing well at distinguishing between the positive/epitope and negative/non-epitope labels. Similar to what was observed in the precision, recall and F1 metrics, the low values in the standard deviations on AUC-ROC and AUC-PR metrics suggest that HLA-EpiCheck is robust. Regarding the MCC metric, although the values are somewhat lower than AUC-PR and AUC-ROC, it is still very acceptable. It is important to remember that this metric is less « optimistic » when evaluating a predictor because it takes into account the 4 components of the confusion matrix (i.e. true positives, true negatives, false positives, false negatives). Thus, the MCC result (0.829) denotes a very good performance of the predictor. However, as suggested by the results on the recall metric, the predictor has room for improvement on false negatives.

Moreover, the complexity of the ExtraTrees predictor was explored. It was found that the average number of nodes per tree is 1 807.2, the average depth per tree is 23.3, and the total number of trees is 300. Additionally, it contains a total of 271 186 leaves, of which 190 016 (i.e., 70%) contain no more than 3 samples. These results indicate an enormous complexity of the model, reflecting the fact that the predictor likely memorizes the entire training set. Therefore, when making a prediction on a patch, the predictor acts by finding the tree and the path leading to the most similar patch handled during the training step. This behavior is reminiscent of the random forest-based dissimilarity largely exploited for unsupervised learning [143]. The basic principle here is that examples that look similar are frequently found in the same leaves of trees. The use of extremely randomized trees has also been proposed for clustering data of mixed types [144]. Thus, in our redundant dataset, our ExtraTrees predictor strongly resembles the KNN method (an instance-based learning method that accommodates data redundancy), except for the fact that it does not compute any distance between the samples but rather memorizes the path leading to a similar sample. This more sophisticated behavior clearly leads to enhanced performance.

### 4.3.3. Feature importances of HLA-EpiCheck on the training set

To evaluate the contribution of dynamic and static descriptors to the performance of HLA-EpiCheck, Mean Decrease in Impurity (MDI) [92] was calculated for each descriptor (see Figure 34). Simply put, the MDI is a normalized value that indicates the importance of a descriptor in the predictor performance. The static descriptor **H_patch_max** is the most important contributor to performance (~0.115, bottom bar in Figure 34). However, in second place is the dynamic descriptor **F_patch_avg** (~0.082) which represents the average side-chain flexibility of the patch. In addition, the descriptors **H_central** (hydrophobicity of the central residue) and **H_patch_avg** (average hydrophobicity of the patch) rank fifth and sixth respectively in the feature importance analysis (~0.072 and ~0.066 MDI values respectively). These results highlight the importance of hydrophobicity in the predictive ability of the tool (the sum of the MDI values of **H_patch_max**, **H_central** and **H_patch_avg** is 0.253) and is in agreement with other studies that have pointed out the importance of hydrophobicity in antigen-antibody recognition [145–148]. Although 4 of the 5 descriptors with the highest MDI values are static descriptors, the sum of the MDI values of this type of descriptor is 0.47 (with 1 being the sum of all MDIs), indicating that although the dynamic descriptors do not stand out for their high MDI values, these descriptors have a slightly greater contribution (0.53) to the model. Moreover, the descriptor **F_patch_avg** stands out among the dynamic descriptors by having the second-highest MDI value. This suggests, as mentioned in [110], that side-chain flexibility plays a key role in antigen-antibody recognition.

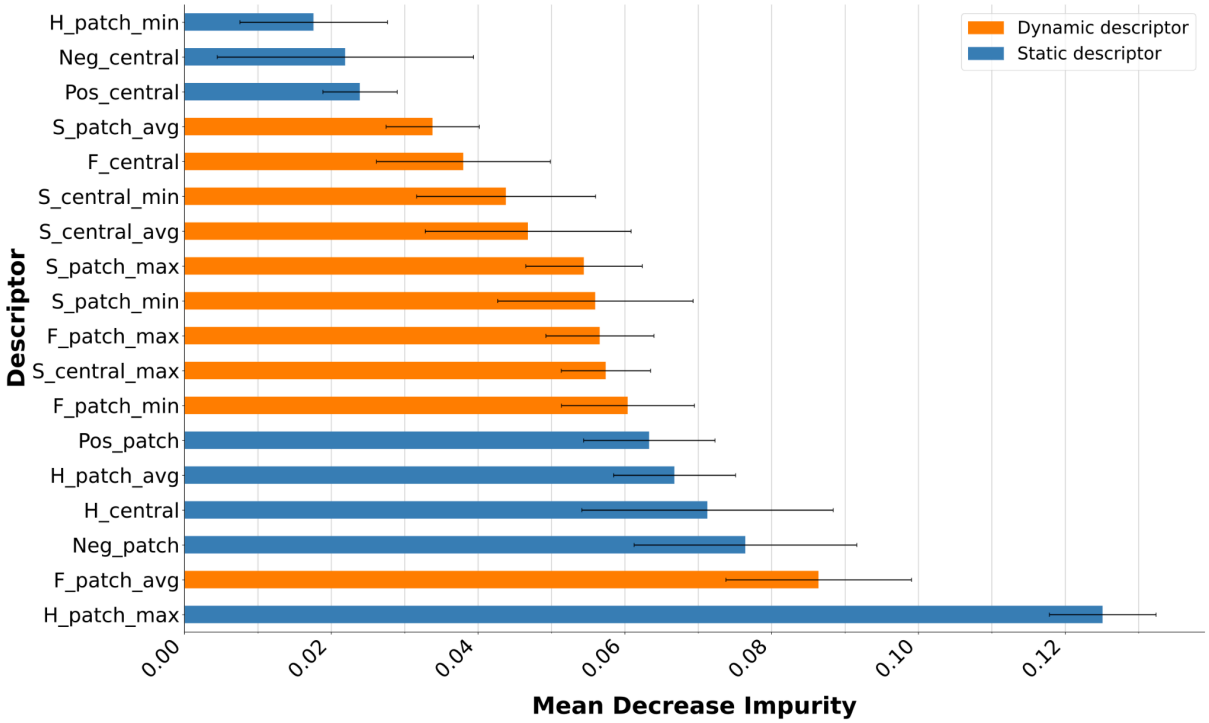



Figure 34: MDI feature importance of the HLA-EpiCheck predictor. Horizontal bar length corresponds to mean MDI values (each tree in the forest having a MDI value). Orange bars correspond to dynamic descriptors. Blue bars correspond to static descriptors. Whiskers correspond to standard deviations of MDI values over the 300 trees.

The relevance of dynamic descriptors in recognizing true epitope patches has also been studied in simpler but less efficient models obtained with non-redundant data. I decided to build a DT using the entire non-redundant dataset (i.e., 668 samples: 396 of “Non-epitope” label and 272 of “Epitope” label), as a descriptive model that could reveal interesting explicit rules leveraging dynamic descriptors. Interestingly this tree can be visualized (Figure S33). It has a total of 81 nodes, a maximal depth of 11 levels and among the 41 leaves, none has less than 3 samples. All pure “Epitope” leaves contain more than 10 samples each. One of them contains 36 “Epitope” samples (out of 272 Epitope samples), and 5 leaves have a number of “Epitope” samples between 10 and 16, for a total of 64 samples, leading to a total of 100 “Epitope” samples in pure “Epitope” leaves (about 37% of the “Epitope” count). Table 21 presents the classification rules that can be obtained from the most populated pure Epitope leaves. Most of the rule items involve descriptors associated with N-RMSF, indicating the relevance of dynamic descriptors representing side-chain flexibility for the recognition of epitope patches.

Table 21: Classification rules of the pure "Epitope" leaves with the largest number of samples in the DT trained with the non-redundant dataset.

| **Leaf with 36 Epitope samples** |
|---|
| F_patch_max ≤ 2.34<br>AND F_patch_max > 0.443<br>AND F_patch_min > -1.274<br>AND charge_patch_pos > 4.5<br>AND F_patch_avg_freq > -0.103<br>=> Epitope |
| **Leaf with 16 Epitope samples** |
| F_patch_max ≤ 2.34<br>AND F_patch_max > 0.443<br>AND F_patch_min > -1.274<br>AND Pos_patch ≤ 4.5<br>AND F_patch_max >1.79<br>=> Epitope |
| **Leaf with 14 Epitope samples** |
| F_patch_max ≤ 2.34<br>AND F_patch_max > 0.443<br>AND F_patch_min > -1.274<br>AND Pos_patch > 4.5<br>AND F_patch_avg_freq ≤ -0.103<br>AND S_patch_min >19.48<br>=> Epitope |

In the same vein, Figure 35 presents the feature importances of this DT, where the descriptors derived from N-RMSF sum up to 0.574 of MDI, and among which N_RMSF_patch_max stands out with an MDI of 0.393 (the highest among all descriptors) and N_RMSF_patch_min with an MDI of 0.125 (the third-highest value).

For comparison, the complexity of the tree trained with the “Redundant” dataset was also explored. It was found that the number of nodes is 1 029 (compared to 81 in the “Non-redundant” case), the maximal depth is 25 levels (versus 11), and out of the 515 leaves (versus 41), 296 contained 3 or fewer samples (versus 0 in the previous DT). These results indicate an huge difference in complexity between the trees trained with the “Non-redundant” and “Redundant” datasets, implying that in the latter case, a predictor that memorizes the training set is obtained, as pointed out in the case of ExtraTrees in section 4.3.2. This high complexity makes the redundant tree less interesting for the exploration of pure epitope leaves as they are mainly composed of very few samples. Regarding the feature importance (see Figure S34), significant changes are observed compared to what was obtained with the non-redundant DT, as the descriptors associated with N-RMSF are no longer in first place. However, their total contribution to MDI remains greater than 50% (check the exact value). This comparison confirms that it is better to use a DT trained on non-redundant data if the purpose is to extract explicit rules showing the involvement of specific descriptors in the paths leading to “Epitope” leaves.

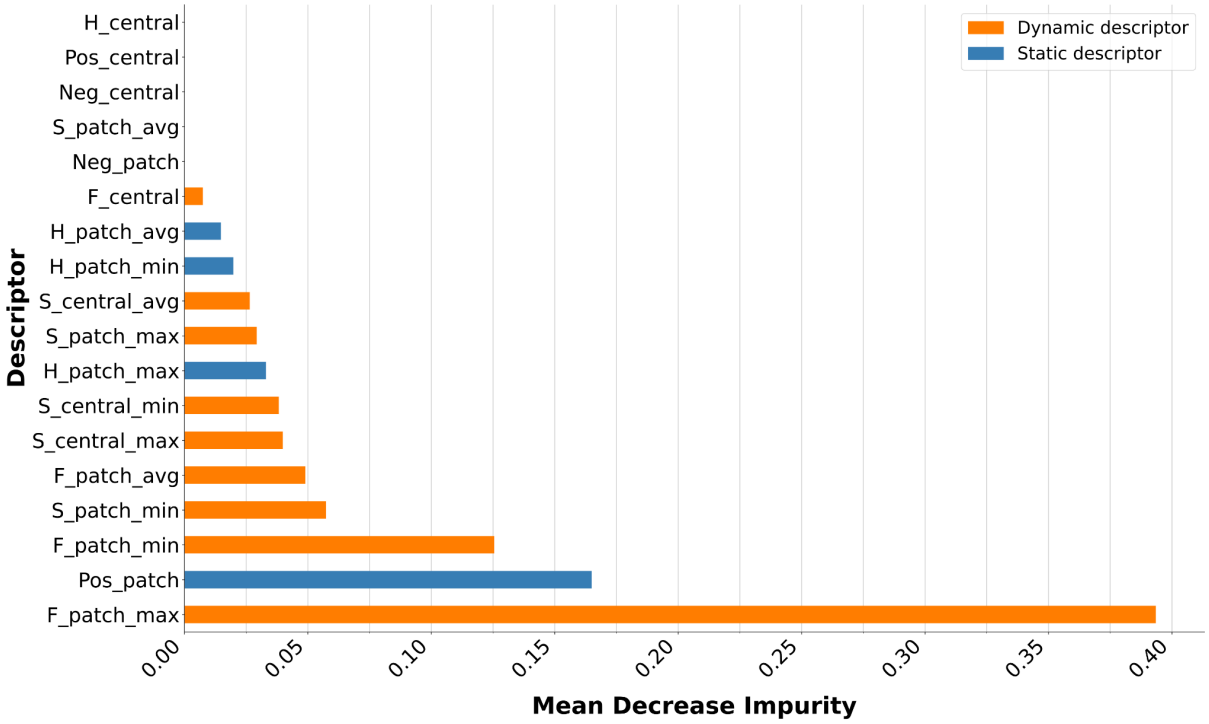



Figure 35: MDI feature importances of the DT trained on the whole Non-redundant dataset. Orange bars correspond to dynamic descriptors. Blue bars correspond to static descriptors. The sum of the MDI values of the dynamic descriptors is 0.77.

### 4.3.4. Performance evaluation of HLA-EpiCheck and comparison with DiscoTope-3.0

Exhaustive performance evaluation will only be performed with the final predictor chosen for HLA-EpiCheck (i.e. ExtraTrees trained with the redundant dataset). However, the results of the performance evaluations on the test sets of the other learning algorithms can be seen in Table S12 and Table S13. It is worth to mention that Table S12 shows high values in the standard deviation, particularly in the case of the Epitope label (0.11, 0.11, 0.14 for the DT; 0.05, 0.08, 0.05 for the ExtraTrees; 0.09, 0.10, 0.11 for the GradientTrees; 0.06, 0.12, 0.07 for the KNN on F1, Precision and Recall metrics respectively). This indicates a significant sensitivity of the predictors' performance to the dataset split, likely due to the reduced size of test sets (25 Epitope and 43 Non-epitope samples, see Table 17).

The results of HLA-EpiCheck evaluation on the test set ( 429 “Epitope” and 947 “Non-epitope”, see Table 18), are presented in more detail in Table 22 (column HLA-EpiCheck) with the values obtained per locus. When considering all the loci, the precision and recall are 0.93 and 0.82 respectively for the epitope label, and 0.92 and 0.97 respectively for the non-epitope label yielding F1 scores of 0.87 and 0.95 for the two labels respectively. The AUC-ROC, AUC-PR and MCC are 0.98, 0.98 and 0.82 respectively. These results are similar to those obtained by cross-validation on the training set (Table 20). As with the training set, a better performance is observed on precision than on recall metric for the epitope label which indicates again that the tool is less likely to make false positive predictions and has room for improvement on false negative predictions. Concerning the results per locus, HLA-EpiCheck shows a high performance for loci A, B, C, DP and DQ with F1 scores not lower than 0.86 for epitope prediction. The F1 score is reduced for DR (0.79) for a yet unclear reason.

Table 22: Performance evaluation of HLA-EpiCheck and DiscoTope-3.0 on test set.

| | | HLA-EpiCheck | | DiscoTope-3.0 | |
|---|---|---|---|---|---|
| **Locus** | **Metric** | **Epitope** | **Non-epitope** | **Epitope** | **Non-epitope** |
| **A** | **Precision** | 0.89 | 0.91 | 0.31 | 0.63 |
| | **Recall** | 0.83 | 0.95 | 0.32 | 0.62 |
| | **F1** | 0.86 | 0.93 | 0.32 | 0.62 |
| **B** | **Precision** | 0.91 | 0.97 | 0.19 | 0.85 |
| | **Recall** | 0.82 | 0.99 | 0.26 | 0.79 |
| | **F1** | 0.87 | 0.98 | 0.22 | 0.82 |
| **C** | **Precision** | 1 | 0.91 | 0.5 | 0.75 |
| | **Recall** | 0.81 | 1 | 0.58 | 0.68 |
| | **F1** | 0.9 | 0.95 | 0.54 | 0.71 |
| **DP** | **Precision** | 0.97 | 0.94 | 0.32 | 0.72 |
| | **Recall** | 0.85 | 0.99 | 0.46 | 0.59 |
| | **F1** | 0.91 | 0.96 | 0.38 | 0.65 |
| **DQ** | **Precision** | 1 | 0.48 | 0.89 | 0.15 |
| | **Recall** | 0.86 | 1 | 0.54 | 0.54 |
| | **F1** | 0.92 | 0.65 | 0.67 | 0.23 |
| **DR** | **Precision** | 0.83 | 0.7 | 0.66 | 0.44 |
| | **Recall** | 0.76 | 0.79 | 0.29 | 0.79 |
| | **F1** | 0.79 | 0.74 | 0.4 | 0.56 |
| **All loci** | **Precision** | 0.93 | 0.92 | 0.56 | 0.8 |
| | **Recall** | 0.82 | 0.97 | 0.55 | 0.8 |
| | **F1-score** | 0.87 | 0.95 | 0.55 | 0.8 |
| | **AUC-PR** | 0.98 | | 0.67 | |
| | **AUC-ROC** | 0.98 | | 0.76 | |
| | **MCC** | 0.82 | | 0.35 | |

Table 22 also reports the results of the comparison between HLA-EpiCheck and DiscoTope-3.0 (a state-of-the-art B-cell epitope prediction tool). Similar to HLA-EpiCheck, DiscoTope-3.0 is a binary classifier that predicts whether a residue is part of an epitope or not. However, DiscoTope-3.0 differs from HLA-EpiCheck as it does not consider the surrounding residues forming a 3D patch around the tested residue and it is only trained with descriptors extracted from static 3D structures. Moreover, the training dataset covers a large variety of proteins while HLA EpiCheck is dedicated to HLA epitopes. Here, the predictions of DiscoTope-3.0 were made on structures corresponding to the first frame of MD simulations. Although HLA-EpiCheck classifies patches rather than single residues, for comparison purposes, the predicted label of the patch was imputed to its central residue. The comparison was performed on the test set. Table 22 summarizes the performances of both tools. Globally, HLA-EpiCheck largely outperforms DiscoTope-3.0 with an MCC value of 0.83 for HLA-EpiCheck and only 0.35 for DiscoTope-3.0. The difference is also observed for F1-score in the prediction of the epitope label since the F1 score is 0.87 for

HLA-EpiCheck and 0.55 for DiscoTope-3.0. The Precision-Recall and ROC curves are shown in Figures S35 and Figure S36. Concerning the results per locus, DiscoTope-3.0 works the best for locus DQ with a remarkably high precision value for the epitope label (0.89) leading to the highest F1-score among all loci (0.67). The lowest F1 score is for locus B (0.22). This great variability in the performance of DiscoTope-3.0 over all HLA loci suggests that the model learnt by DiscoTop-3.0 does not cover properly all the confirmed eplet residues selected in this study. The superiority of HLA-EpiCheck may rely, on the one hand, on the use of dynamic descriptors in addition to the static ones and on the other hand, as explained above, HLA-EpiCheck was trained on a highly redundant set of samples derived solely from HLA antigens as its purpose is only to predict HLA epitopes. In future work, it can be envisaged to train a more generalist predictor using 3D structures and MD trajectories of a more diverse set of proteins for which epitopes are known. This could answer the question whether our set of descriptors, used in the appropriate setting, could improve B-cell epitope prediction in general compared to DiscoTope-3.0.

### 4.3.5. Comparison with experimental results on a subset of non-confirmed eplets

#### 4.3.5.1. HLA-EpiCheck scores for comparison with experimental results

Since HLA-EpiCheck acts on patches centered on a unique residue while some eplets are composed of more than one solvent-accessible residue, and because the prediction can be performed on all HLA antigens displaying an eplet of interest, the results were aggregated by patch prediction for each eplet residue and for all antigens displaying this eplet. This led to the definition of two HLA-EpiCheck scores: one for single-residue eplets and the other one for composite eplets. For a given eplet residue, the prediction returned by HLA-EpiCheck for this residue is the probability that the patch centered on this residue is an epitope patch, as the ExtraTrees algorithm returns the predicted class probability as the fraction of samples of the same class in a leaf.

In a first step the distance between residues was inspected in the case of eplets composed of more than one solvent-accessible residue ("composite" eplets). This verification should not be necessary as the eplet definition implies that eplet residues are located at a 3-3.5Å distance from each other. However, when checking non-confirmed eplets of the locus DQ, I found an exception (75I) composed of residues 75I, 161D and 163I (74I, 160D and 162I in alleles DQA1*02, DQA1*04, DQA1*05 and DQA1*06 because of a deletion of one AA in these sequences). It is the only eplet with two completely different candidate epitope zones. This is because position 75/74 and positions 161/160 - 163/162 are approximately 50Å apart in the 3D structure.

For eplets composed of a unique amino acid, the HLA-EpiCheck single prediction is aggregated over all antigens containing this eplet according to Equation 10:

$$\mathcal{S}_{e\_resid} = \frac{\sum_{a \in A_e} pred(e\_resid_a)}{|A_e|} \qquad 19)$$

where $S_{e\_resid}$ is the score for eplet residue e_resid of eplete, pred($e\_resid_a$) is the HLA-EpiCheck prediction for eplet residue e_resid on antigen a, and $A_e$ is the set of antigens carrying eplet e.

For composite eplets, the HLA-EpiCheck predictions are first aggregated at the patch level for each antigen level using the maximal value over all solvent accessible residues members of the eplet and the resulting score is then averaged over all antigens containing this eplet, according to Equation 11.

$$\mathcal{S}_e = \frac{\sum_{a \in A_e} \mathcal{P}_{a,e}}{|A_e|} \quad with \quad \mathcal{P}_{a,e} = max_{r \in R_{a,e}}(pred(r)) \qquad 20)$$

where $S_e$ is the score for composite eplet e, $A_e$ is the set of antigens carrying eplet e, $P_{a,e}$ is the prediction score aggregated on all solvent-accessible residues members of eplet e for antigen a, $R_{a,e}$ is the set of solvent-accessible residues members of eplet e on antigen a, pred(r) is the HLA-EpiCheck prediction on residue r .

These two aggregated HLA-EpiCheck scores range from 0 to 1 and a non-confirmed eplet is considered as determining an epitope if its aggregated score is greater than 0.5.

#### 4.3.5.2. Comparison with experimental results

The HLA Eplet Registry contains 492 eplets (accessed in February, 2023), of which only 146 had the antibody-verified status. For locus DQ, 56 non-confirmed eplets are found (see Table 23), of which 40 eplets have at least one solvent-accessible residue on at least one HLA antigen of this study. The HLA-EpiCheck scores $S_e$ and $S_{e\_resid}$ were computed for each of these 40 eplets (see Table 23) depending if they are single-residue or composite eplets. Moreover, single residue scores were also computed for each eplet residues in the case of composite eplets

In Table 23, 11 of the 40 eplets have a score in [0.8, 1], 12 eplets have a score in [0.6, 0.8[, 7 eplets have a score in [0.4, 0.6[, 9 eplets have a score in [0.2, 0.4[ and 1 eplet has a score in [0, 0.2[. The HLA-EpiCheck scores have been compared with experimental results obtained for 15 non-confirmed eplets in order to establish their epitope status (antibody-verified). These experimental results consist of adsorption of patients' sera on normal spleen mononuclear cells (SMCs) and DQ transfected murine cell clones [Manuscript submitted for publication].

The 15 validated eplets are highlighted in the lightgreen rows. It appears that 11 of the 15 validated eplets have an HLA-EpiCheck score greater than 0.5, representing a remarkable consistency between HLA-EpiCheck predictions and experimental results. Out of these 11 eplets, 3 eplets have a score in [0.8, 1], and 8 eplets have a score in ]0.5, 0.8[. Regarding the remaining 4 evaluated eplets, they have a score in [0.3, 0.5].

The 4 eplets with low prediction scores were further investigated. To do this, the HLA-EpiCheck predictions of solvent-accessible neighboring residues were checked. In the case of eplet 23L, residues 22, 25, 43 and 80 were predicted as epitopes in the two antigens that display the eplet (see Table S14). These results suggest that the epitope is not centered on the eplet residue but in its vicinity. However, for the other 3 low scoring eplets, no neighboring residue favorable for being predicted as epitope was found.

Eplet 75I is composed of residues 75I, 161D and 163I (74I, 160D and 162I in alleles DQA1*02, DQA1*04, DQA1*05 and DQA1*06). It is the only eplet with two completely different candidate epitope zones. This is because position 75/74 and positions 161/160 - 163/162 are approximately 50Å apart. Looking in detail at the HLA-EpiCheck predictions for this eplet (see Table S15), the position 75/74 was systematically predicted as epitope on all the antigens displaying the eplet. These results suggest that the patch centered on position 75/74 is the epitope actually recognized by antibodies.

Table 23: HLA-EpiCheck scores for 40 non-confirmed eplets of locus DQ. Scores are also computed for each single eplet residue in the case of composite eplets. "Max distance" column corresponds to the maximum distance between eplet residues. Mean center of mass along the trajectory of residues is used for computing distances. Allele families DQA1*02, DQA1*04, DQA1*05 and DQA1*06 have a deletion relative to the remaining DQA1*01 and DQA1*03 allele families at position 56, so the residue numbering changes for these alleles. Eplets concerned with this deletion are marked with a *. Values for the HLA-EpiCheck score per eplet residue and Max distance are not shown for eplets composed of only one solvent-accessible residue. In $S_{e\_resid}$ , missing residues of composite eplets are not solvent-accessible. Non-confirmed eplets experimentally validated are highlighted in the lightgreen rows. Number of antigens showing an eplet are presented in "# antigens" column.

| No | Eplet name | Se | Se_resid | Chain | Max distance [Å] | # antigens |
|---|---|---|---|---|---|---|
| 1 | 55PPD | 0.997 | 55P:0.98; 56P:0.99 | β | 6.4 | 9 |
| 2 | 55PPA | 0.995 | 55P:0.98; 56P:0.99 | β | 5.8 | 4 |
| 3 | 55RPD | 0.992 | 55R:0.99; 56P:0.84; 57D:0.51 | β | 6.5 | 5 |
| 4 | 74EL | 0.99 | 74E:0.52; 77T:0.99 | β | 8.6 | 18 |
| 5 | 40ERV | 0.976 | 40E:0.97; 41R:0.58; 45V:0.64 | α | 8.3 | 15 |
| 6 | 66DR | 0.946 | 66D:0.42; 67I:0.77; 70R:0.94 | β | 8.2 | 8 |
| 7 | 56PD | 0.945 | 56P:0.94; 57D:0.51 | β | 6.3 | 13 |
| 8 | 56PS | 0.906 | - | β | - | 2 |
| 9 | 56P | 0.902 | - | β | - | 27 |
| 10 | 56PA | 0.895 | 56P:0.89; 57A:0.80 | β | 5.5 | 10 |
| 11 | 135G | 0.835 | - | β | - | 2 |
| 12 | 70GT | 0.798 | 67V:0.79; 70G:0.63 | β | 7.3 | 4 |
| 13 | 67VG | 0.782 | 66E:0.36; 67V:0.76; 70G:0.69 | β | 8.2 | 7 |
| 14 | 66ER | 0.755 | 66E:0.33; 67V:0.73; 70R:0.70 | β | 6.8 | 14 |
| 15 | 67VT | 0.748 | - | β | - | 17 |
| 16 | 66EV | 0.744 | 66E:0.34; 67V:0.74 | β | 7.3 | 20 |
| 17 | 23R | 0.721 | - | β | - | 29 |
| 18 | 185T | 0.712 | - | β | - | 21 |
| 19 | 70RT | 0.701 | - | β | - | 15 |
| 20 | 75I* | 0.683<br>0.418 | 74I/75I:0.66; 160D/161D:0.43; 162I/163I:0.40 | α<br>α | 49.9 | 26<br>26 |
| 21 | 185I | 0.66 | - | β | - | 11 |
| 22 | 66D | 0.649 | 66D:0.42; 67I:0.70 | β | 5.2 | 12 |
| 23 | 75IL | 0.646 | - | α | - | 11 |
| 24 | 125G | 0.599 | - | β | - | 2 |
| 25 | 129QS | 0.528 | 129Q:0.52; 130S:0.47 | α | 6.3 | 8 |
| 26 | 129H* | 0.525 | 128H:0.54; 129H:0.50 | α | - | 24 |
| 27 | 23L | 0.466 | - | β | - | 2 |
| 28 | 160AD* | 0.446 | 159A:0.31; 160A:0.37; 160D:0.35; 161D:0.46 | α | 5.3 | 24 |
| 29 | 135D | 0.401 | - | β | - | 30 |
| 30 | 3P | 0.4 | - | β | - | 2 |
| 31 | 167H | 0.379 | - | β | - | 7 |
| 32 | 160S | 0.369 | - | α | - | 2 |
| 33 | 130R | 0.361 | - | β | - | 29 |
| 34 | 3S | 0.348 | - | β | - | 29 |
| 35 | 160D | 0.344 | - | α | - | 4 |
| 36 | 167R | 0.336 | - | β | - | 25 |
| 37 | 160A* | 0.33 | 159A:0.31; 160A:0.34 | α | - | 27 |
| 38 | 130A | 0.278 | - | α | - | 3 |
| 39 | 130Q | 0.261 | - | β | - | 3 |
| 40 | 175E* | 0.174 | 174E:0.16; 175E:0.18 | α | - | 18 |

## 4.4. Discussion

There are a few limitations in the HLA-EpiCheck approach that can be considered in further studies to improve the HLA epitope predictions. On the one hand, although according to [110], short MD simulations, like the ones implemented here, could be sufficient to capture the flexibility of the side chains, wider range movements involving several AAs (for example, torsion movements) might require longer simulation times. On the other hand, HLA-EpiCheck lacks geometric descriptors that capture the shape of the surface patches. Although other tools [149, 150] have made contributions in this regard, they are applied to static structures, so it would be necessary to find this type of descriptor that also manages to capture the dynamic behavior during an MD trajectory. In this sense, embedding approaches using DL could be a lead to follow. For example, dMaSIF [151] is a DL approach that generates embeddings of surface patches that simultaneously capture geometric and physicochemical properties.

Regarding the perspectives of using HLA-EpiCheck when defining the donor-recipient matching, one could consider generating an estimate of MML (mismatch load). The idea would be to analyze the recipient and donor HLA antigens pairwise and locuswise. In other words, for a given locus, all pairs of alleles (max 4 if both recipient and donors are heterozygotes) should be considered. For instance, if the recipient's locus A has A1 and A2, and the donor's locus A has A3 and A4, then one should examine A1 versus A3, A1 versus A4, A2 versus A3, and A2 versus A4. Subsequently, for each pair, the two 3D structures and MD trajectories should be found in the database. Then, for each HLA antigen, all solvent-accessible residues should be extracted, and the corresponding patches centered on these surface residues should be built. Afterward, HLA-EpiCheck should be run on all the patches, and positions corresponding to putative epitope patches should be listed. The two lists should then be compared, and the mismatches should be counted. It is worth noting that all pairs of alleles could be pre-computed for a defined list of HLA antigens (e.g., our list of 207 antigens). Returning to the calculation of MML, the mismatches of the relevant pairs of HLA antigens should be aggregated at the locus level. The MML scores for each locus can be aggregated to form a global score across loci. Weights could be added at this step depending on the importance of the locus. Retrospective studies should be conducted to evaluate the relevance of the HLA-EpiCheck MML score on DSA production and graft rejection.

## 4.5. Chapter summary

Donor-Specific Antibodies (DSA) recognition of Human Leukocyte Antigen (HLA) proteins is the primary cause of organ transplant loss. Predicting the antigenicity of HLA B-cell epitopes, key components of which are eplets, could improve donor-recipient HLA matching and reduce de *novo* DSA formation and graft rejection. HLA-EpiCheck, a ML predictor for B-cell epitopes on HLA proteins, was developed leveraging both static and dynamic molecular descriptors derived from molecular dynamics simulations.

HLA-EpiCheck is a binary classifier that predicts whether a patch, centered on a solvent-accessible amino acid (AA), is an epitope or not. The predictor was trained using 18 molecular descriptors, both static (e.g., hydrophobicity, electrostatic charges) and dynamic (e.g., relative solvent accessible surface area, side-chain flexibility), calculated for each patch. Two ML datasets were generated: "Non-redundant" with reduced sequence redundancy using MMseqs2 and "Redundant" using the entire dataset. The datasets were split into training and test sets, and various learning algorithms (e.g., DT, RF, ExtraTrees, GradientTrees, HistoGradientTrees, KNN and MLP) were evaluated using gridsearch and cross-validation techniques. Performance was assessed using precision, recall, F1-score, and MCC metrics.

The Extremely Randomized Trees (ExtraTrees) algorithm trained with the "Redundant" dataset showed the best performance and was chosen for the final version of HLA-EpiCheck. Indeed, redundancy does not harm in our study as HLA-EpiCheck is dedicated to work solely on HLA antigens. Moreover, the model complexity of HLA-EpiCheck suggests that this classifier acts more on an instance-based than on a

model-based learning paradigm. The HLA-EpiCheck predictor achieved high precision and recall for epitope prediction, both upon training (0.92 and 0.83, respectively) and upon testing(0.93 and 0.82, respectively, leading to an F1-score of 0.87.. HLA-EpiCheck outperformed DiscoTope-3.0, a state-of-the-art B-cell epitope prediction tool, with an MCC value of 0.83 compared to 0.35. Mean Decrease in Impurity (MDI) analysis revealed that static descriptors related to hydrophobicity and dynamic descriptors reflecting side-chain flexibility were the most important contributors to HLA-EpiCheck's performance. The DT model, trained with the "Non-redundant" dataset, provided interpretable classification rules and further highlighted the relevance of side-chain flexibility descriptors (e.g., N_RMSF_patch_max, N_RMSF_patch_min) for identifying HLA epitopes.

HLA-EpiCheck scores were computed for 3D patches covering 40 non-confirmed eplets of the DQ locus and compared with experimental results for 15 of these eplets. 11 out of the 15 experimentally validated eplets had an HLA-EpiCheck score greater than 0.5, demonstrating remarkable consistency between prediction and wet lab experiment. Further investigation of the four low-scoring eplets suggested that the actual epitopes might be centered on neighboring residues of the reported eplet.

A surface scanning strategy could be envisaged to leverage HLA-EpiCheck predictions in an alternative or complementary mismatch load (MML) estimation.

# Chapter 5

# HLA-3D-Diff

## Summary



The **HLA-3D-Diff** tool, developed as part of the eponymous project, collects the data generated by this thesis as well as certain domain knowledge on HLA antigens. This project, led by Marie-Dominique Devignes and myself, received funding for an engineer (Louane Sigrist) for its implementation. My exact role in the project was divided into three steps. In a first step, I provided the engineer the general vision of the project and various use cases that motivated the project. I also indicated the relevant data sources and the data to collect. I provided the 3D structures, 3D-surface patches, SASA profiles and MD runs produced during this thesis with all explanations about the methods used. In a second step, I had regular meetings with the engineer and I gave my feedback about her propositions for the database model and for the visualization interface features. In a third step, I extensively tested the HLA-3D-Diff database and visualization interface, leading to the latest bug corrections and improvements. I also organized demo sessions with my thesis supervisors and other colleagues in Nancy and in Paris. During these sessions, the HLA-3D-Diff system was installed thanks to a Docker image on their laptop and several use cases (already known or on demand) were run.

The database contains public data from PDB [48], PDB-REDO [54], HLA Eplet Registry [34], PD-IMGT/HLA database [15], and Allele Frequency Net database [152].

The main objective of **HLA-3D-Diff** is to provide a user-friendly graphical tool to facilitate the comparison of HLA antigens at the sequence/eplet level by leveraging certain domain knowledge (HLA Eplet Registry, IPD-IMGT/HLA, Allele Frequency Net) and at the structural level by exploiting the 3D structures and MD data generated in this thesis. The purpose is to serve as an aid to understand complex or unexpected immunization patterns.

**HLA-3D-Diff** consists of two user interfaces. The first is called **HLA-3D-Diff database**, which aims to provide easy access to integrated data from multiple public sources and data generated by members of the project. The second is called **HLA-3D-Diff visualization interface**, which is a web app for visualizing and comparing structures and MD simulations of HLA antigens.

## 5.1. HLA-3D-Diff database

The **HLA-3D-Diff database** allows access to and exploration of data from either public sources such as the HLA Eplet Registry, IPD-IMGT/HLA, Allele Frequency Net, or locally generated data derived from MD data such as SASA values or patch composition.

The conceptual schema (entity-relationship) of the database is presented in Figure 36. For example, the database integrates information about serological groups associated with HLA antigens, the frequency of HLA antigens in the European population, the protein sequence of HLA antigens, the eplets (confirmed or not) associated to HLA antigens, the Ellipro score of each eplet, the 3D structures used as starting points for MD simulations of HLA antigens, the MD trajectories of HLA antigens, as well as the metadata associated with these structural data (for example, whether the structure comes from the PDB or was predicted with AlphaFold-2, the structure repair protocol, the resolution of the structure if it comes from the PDB, the parameters used for MD, etc.). The complete dictionary of the database can be consulted in its documentation (accessible through https://gitlab.inria.fr/capsid.public_codes/hla-3d-diff_public).

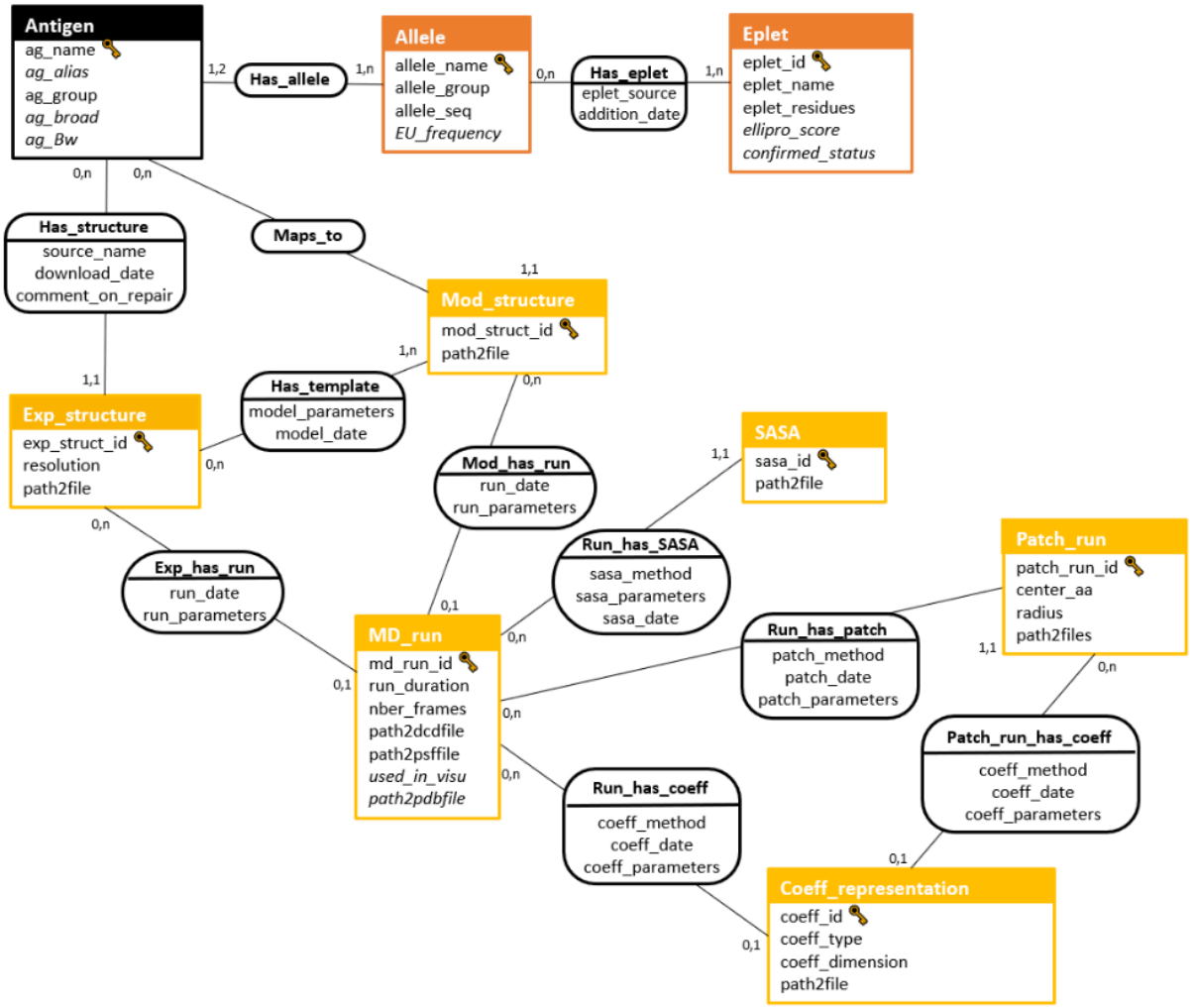



Figure 36: Conceptual schema of the HLA-3D-Diff database.

In order to illustrate the usefulness of the database, some use cases will be presented below:

- List eplets shared by a set of alleles.
- List eplets shared by a set of alleles and absent from another set of alleles.
- List eplets present at a given position for a set of alleles.
- List all alleles that carry a given eplet.
- List alleles that carry a given eplet but do not carry another one.
- List confirmed eplets present in an allele having a given Ellipro exposition level.

- List of SASA values for residues in the patch surrounding a given AA.

To query the database, queries are performed using the SQL language through a web interface implemented with phpMyAdmin (see Figure 37).

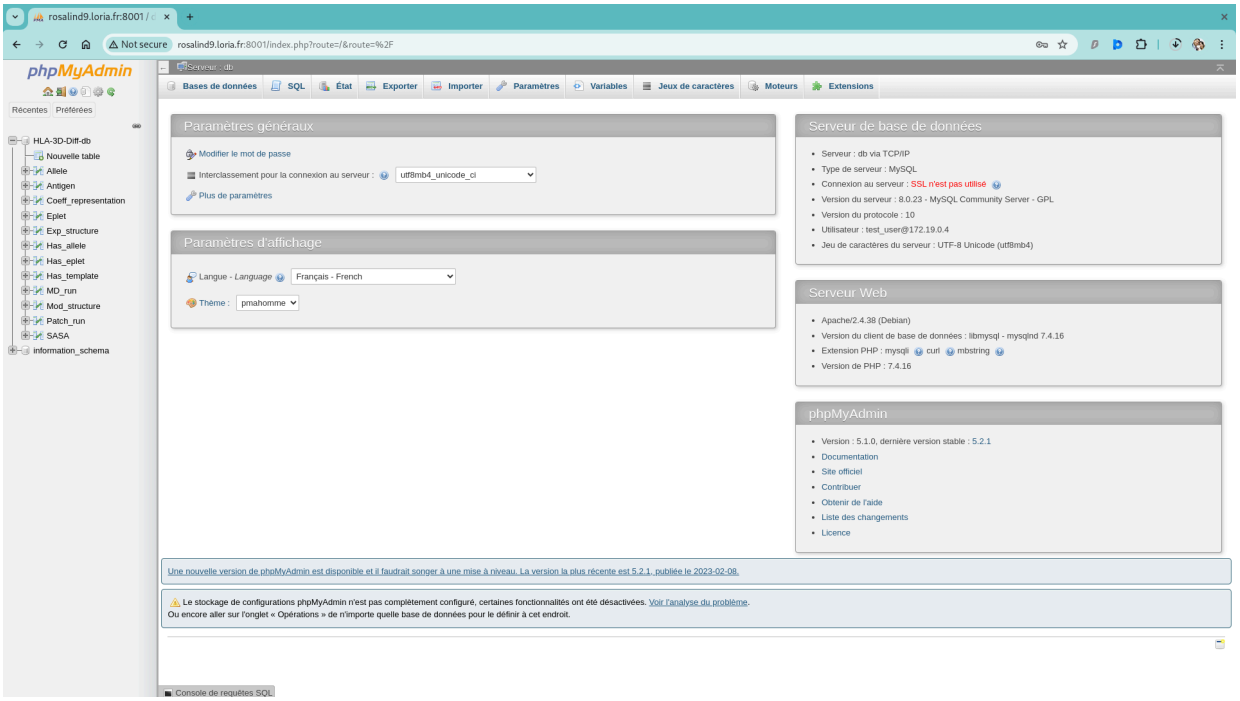



Figure 37: User interface of the **HLA-3D-Diff database.**

## 5.2. HLA-3D-Diff visualization interface

**HLA-3D-Diff visualization interface** is a web application that allows visualizing, superimposing, and comparing 3D structures of HLA antigens to highlight their structural differences along MD simulations. This application can be used by biologists or clinicians without computer science expertise as an aid to understand complex or unexpected immunization patterns. The interface was developed with the help of NGL viewer [153], a collection of tools for web-based molecular graphics.

The main features of the interface are described here.

- Representations. Five types of representations are available: Cartoon, licorice, ball & stick, spacefill, and surface (see Figure 38).

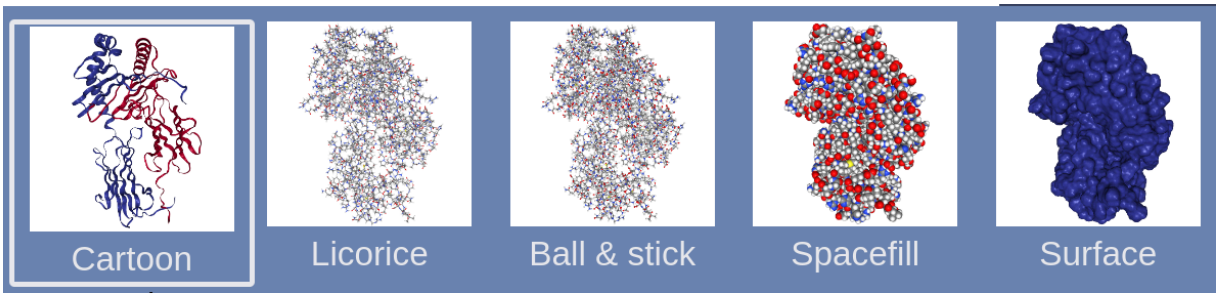



Figure 38: Representations available in **HLA-3D-Diff visualization interface**.

- Colors. Five types of coloring are possible: Uniform coloring, coloring by chain (when applicable), coloring by hydrophobicity (color gradient of the user's choice), coloring by electrostatic potential (gradient), and coloring by RSASA (gradient) (see Figure 39).

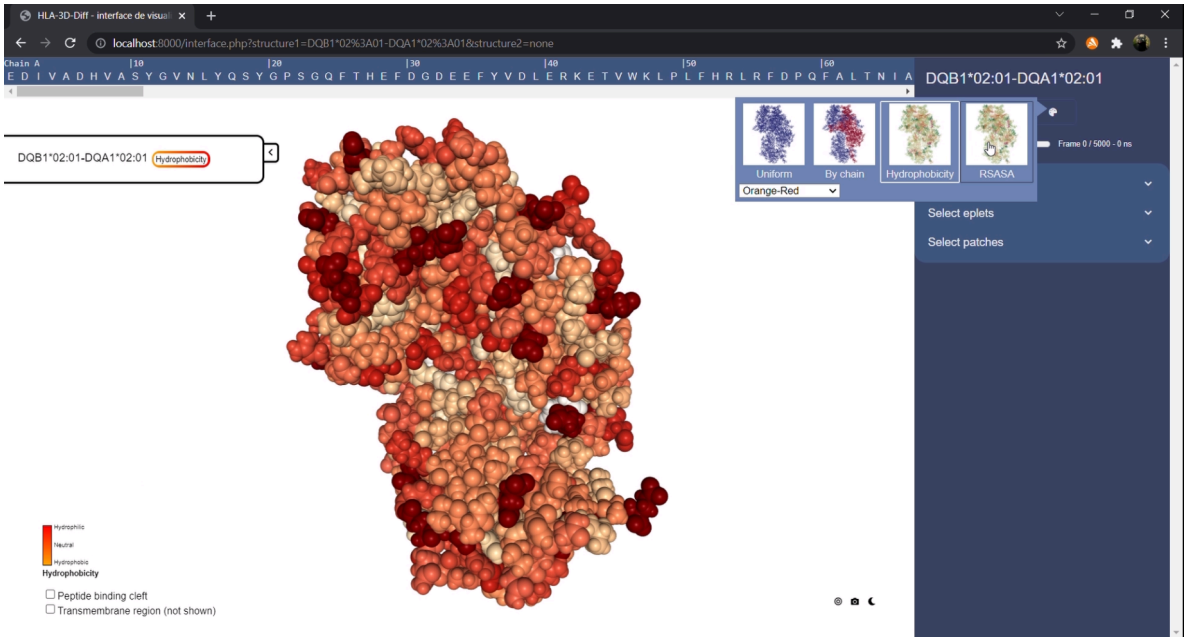



Figure 39: Example of the hydrophobicity gradient coloring for the DQB1*02:01-DQA1*02:01 antigen.

- Selections. Three types of selections are available: selection of residues by entering their positions, selection of eplets by choosing from the list of eplets of the structure and selection of patches by entering a central residue and a radius. All the eplets of an antigen will be listed, with their ElliPro score and their status (confirmed or not) (see Figure 40).

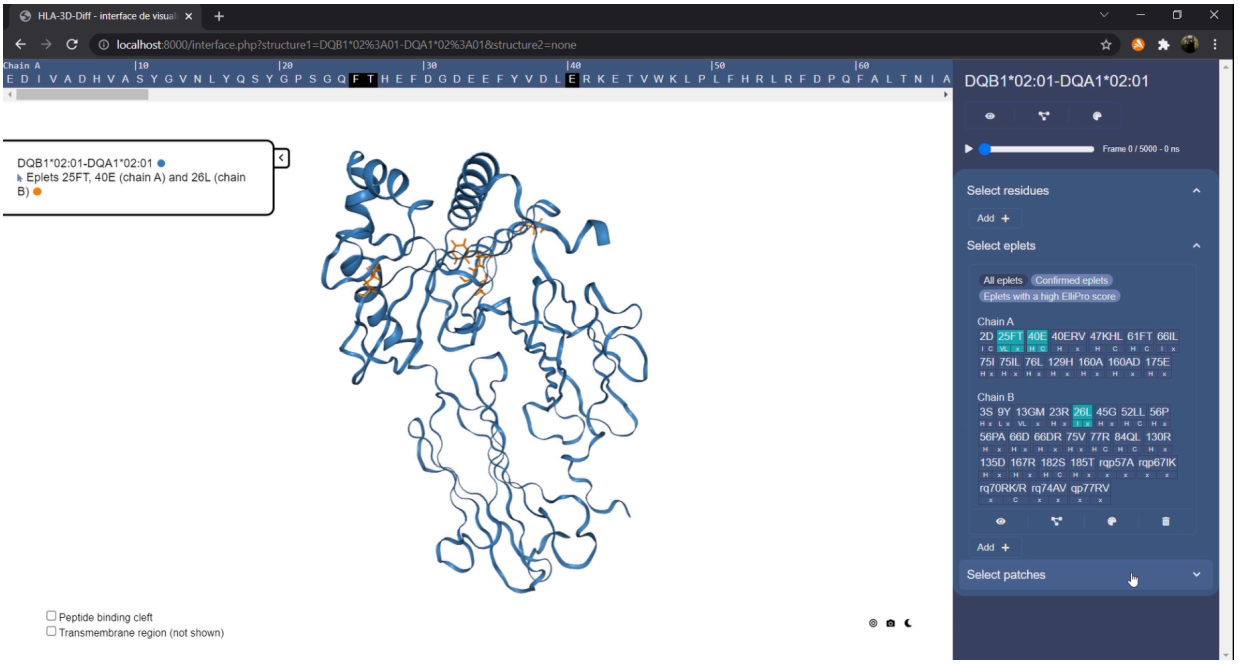



Figure 40: Example of the eplet selection feature. Antigen eplets are shown on the right panel. Eplets 25FT, 40E and 26L of the DQB1*02:01-DQA1*02:01 antigen are shown in licorice representation and colored in orange.

Finally, MD simulations can be played on the interface and allow the user to observe movements affecting the antigen and the regions or residues of interest.

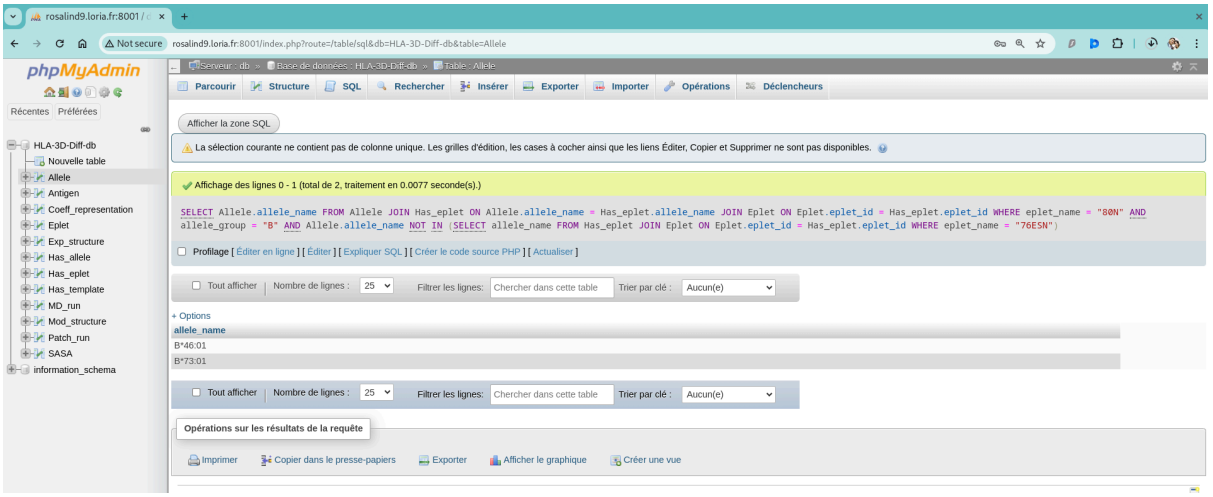
As was done with HLA-3D-Diff database, some use cases of the interface are listed below:

- Superposition of two HLA antigen structures and dynamic visualization of their MD trajectories.
- Differential visualization (by color and/or shape) of a list of positions/eplets.
- Visualization of a numerical value (RSASA, hydrophobicity, electrostatic potential, etc.) by color gradient applied to one or two superposed HLA antigens.

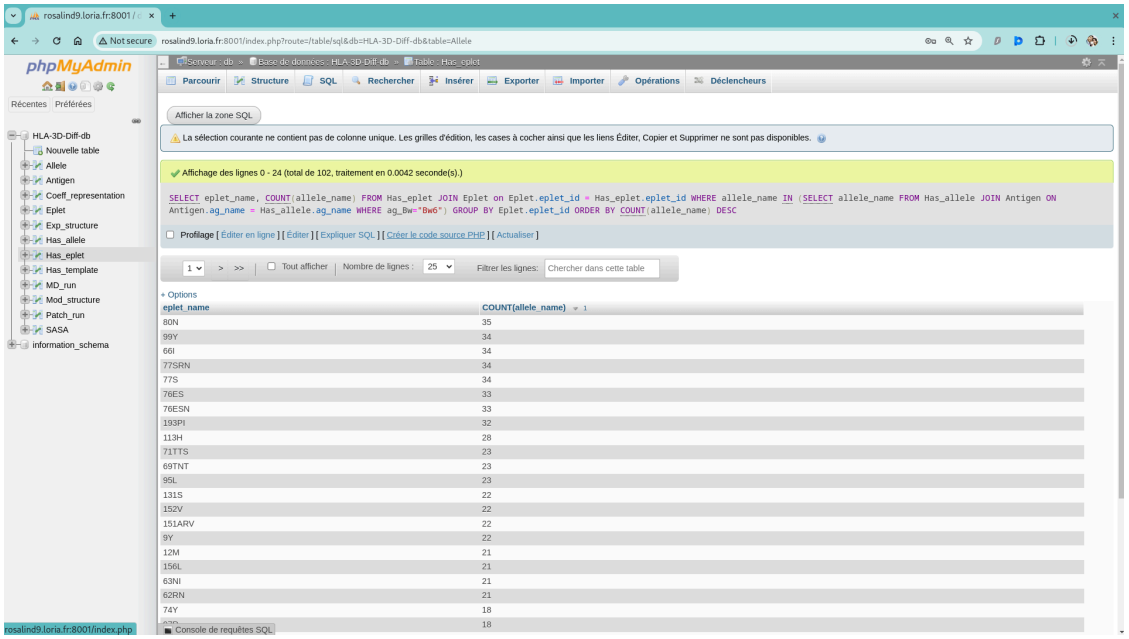
## 5.3. Examples

Here are some examples of use of **HLA-3D-Diff database**:

1. List alleles in locus B that possess the 80N eplet but not the 76ESN eplet.
   Answer: B*46:01 and B*73:01




2. List eplets shared by all alleles that present the Bw6 serotype.
   Answer : list of 102 eplets



Here are some examples of use of **HLA-3D-Diff visualization interface**:

1. Visualization of the polymorphism between antigens A*02:01 (in blue) and A*02:02 (in red) at position 43. In Figure 41, the structures of the antigens are presented in the backbone representation, while the side chains of position 43 (Alanine for antigen A02:01 and Arginine for antigen A*02:02) are presented in the "spacefill" representation. The difference in shape between both side chains can be appreciated. In Figure 42, the side chains are presented in the surface representation. The structure of antigen A*02:01 is also presented with the surface representation and a slight transparency effect. In this case, the change in solvent accessibility of the arginine of antigen A*02:02 compared to antigen A*02:01 can be appreciated. The unique feature here (compared with other visualization interface) is that MD simulations can be run for each or both antigens by clicking on the arrow left to the timeline proper to each antigen in the menu panel on the right.

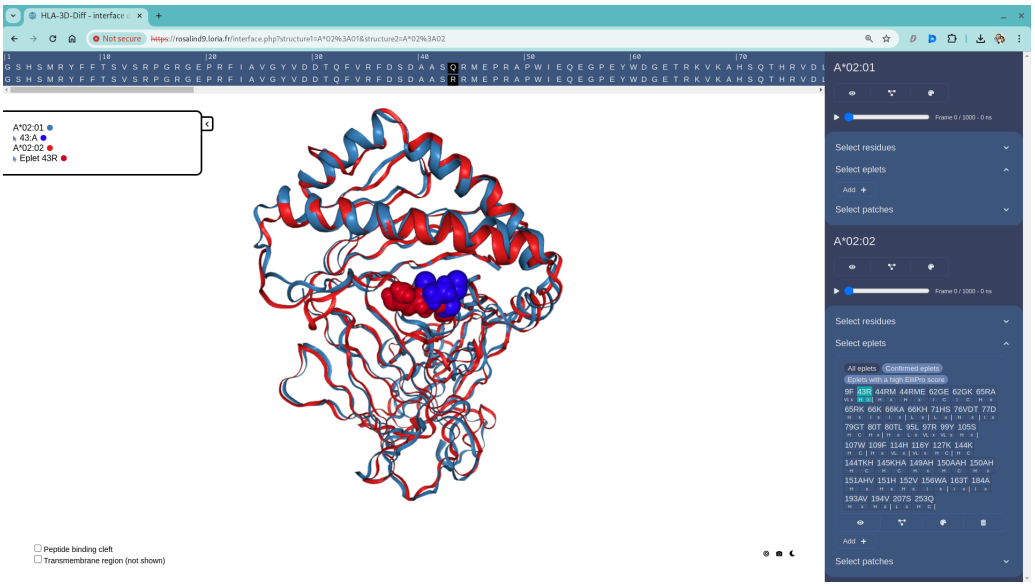



Figure 41: Visualization of antigens A*02:01 and A*02:02 comparing the difference in side chains at position 43 using the "spacefill" representation.

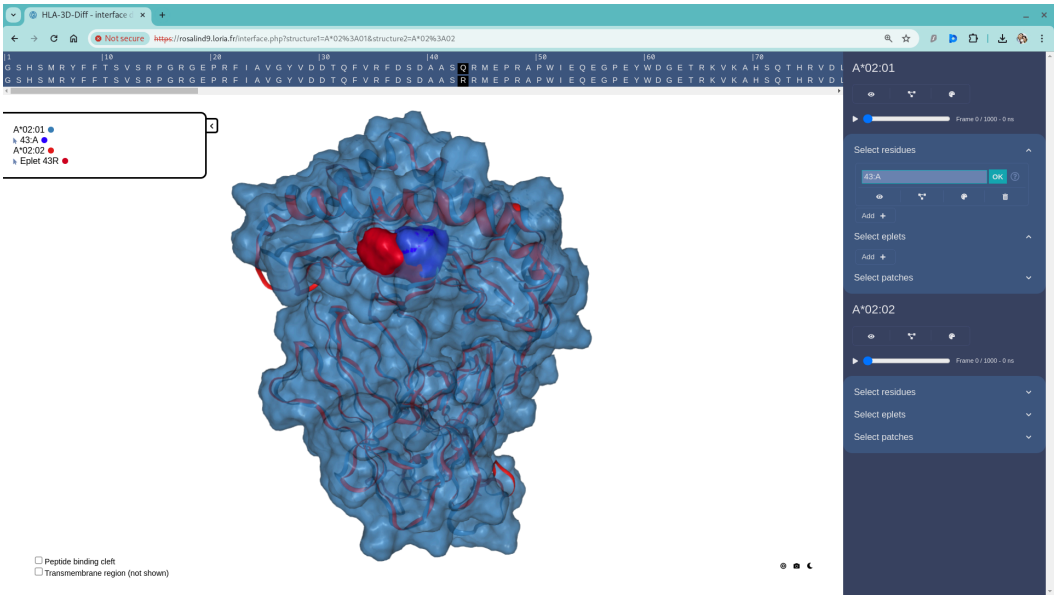



Figure 42: Visualization of antigens A*02:01 and A*02:02 comparing the difference in side chains at position 43 using the surface representation.

2. The QR code below will redirect you to a "User Guide" video (realized by Louane Sigrist under my supervision) that provides a more complete appreciation of the interface as well as its mode of use.

# Conclusions

The precise definition of D/R compatibility is a question that remains open in the field of organ transplantation. In particular, the real immunological importance of mismatches between donor and recipient is unknown; in other words, the antigenicity/immunogenicity of these mismatches is unknown. The present thesis addressed this problem from a structural point of view using data science and machine learning methods. To this end, I generated an unprecedented set of high-quality 3D structures and molecular dynamics (MD) simulations for 207 HLA antigens and analyzed them to provide a deep insight into the 3D structural features of HLA antigens and their contribution to HLA antigenicity. The results of this thesis are presented in 4 main contributions.

The first contribution, presented in Chapter 2, consists of an in-depth analysis of the 3D structural features of HLA antigens through MD simulations. Among the main findings derived from this analysis, the following can be highlighted:

- The difference in solvent accessibility of AAs between MD data and static PDB structures, where ~4% more solvent-accessible AAs were found in the MD data. This means that 5 to 12 AAs (depending on the antigen size) are overlooked when only static 3D structure is considered
- The composition trends of non-confirmed eplets indicate that they are likely composed of a mixture of true and false eplets. This makes it necessary to propose a prediction tool capable of identifying true eplet candidates to be tested in priority by experimental methods
- A tendency towards increased solvent accessibility for non-polar AAs in confirmed eplet AAs was identified. This makes sense as these AAs play an important role in binding to antibodies and it is in line with the use of solvent-accessibility as a descriptor for eplet/epitope prediction.
- It was unexpectedly found that a significant number of polymorphic positions, including eplets, are not solvent-accessible. This raises interesting questions about the role of eplets in HLA epitope recognition and fully justifies the use of 3D-surface patches for HLA epitope prediction.
- Eplets do not display the same AA-type composition as the rest of the surface of the antigen.
- An important prevalence and high level of solvent accessibility were observed in confirmed eplet AAs for ALA, THR, GLN, HSD, and ARG. To my knowledge, such results have never been reported until now for HLA epitopes.
- Loci A and C were found to have the highest number of solvent-accessible positions repertorized as confirmed eplets.
- AAs and patches associated with confirmed eplets present lower flexibility than their non-eplet counterparts, suggesting that areas with reduced side-chain flexibility are more favorable for antibody recognition. This result confirms what has been reported in the literature but never validated in the case of HLA antigen-antibody interaction.
- MD simulations allow the correction of various stereochemical errors introduced by experimental methods. Also, MD simulation presents itself as a valuable technique that allows considering dynamic properties that are relevant when defining the antigenicity of HLA antigens (e.g., side-chain flexibility).

The second contribution, presented in Chapter 3, consists of the exploration of Zernike descriptors for the comparison of 3D surfaces. Among the main findings, the following can be highlighted:

- Zernike descriptors present themselves as an interesting alternative for the automatic identification of peptide conformers along MD runs.
- The difficulties encountered in obtaining results consistent with ground truth when comparing patches on protein surfaces suggest that Zernike descriptors are highly sensitive to the generation of patch meshes.

The third contribution, presented in Chapter 4, introduces a machine learning predictor for B-cell epitopes on HLA antigens that leverages the MD data generated in Chapter 2. Among the main findings, the following can be highlighted:

- A high-performance B-cell epitope predictor (called HLA-EpiCheck) was successfully trained, which outperforms DiscoTope 3.0 (a state-of-the-art B-cell epitope prediction tool). Additionally, HLA-EpiCheck demonstrated remarkable consistency with experimental results of non-confirmed eplets.
- Due to the high redundancy of HLA sequences and 3D structures from which the dataset used is derived, it is very likely that HLA-EpiCheck is overfitted. However, redundancy does not harm the study as HLA-EpiCheck is dedicated to working solely on HLA antigens. Moreover, the model complexity of HLA-EpiCheck suggests that this classifier acts more on an instance-based than on a model-based prediction paradigm.
- Feature importance analysis of HLA-EpiCheck is in agreement with other studies that have pointed out the importance of hydrophobicity and side-chain flexibility in antigen-antibody recognition. This underlines the benefit of using MD simulation data for describing 3D-surface patches.
- Most important contribution in feature importance analysis concerns “patch” rather than “ central residue” descriptors, emphasizing the importance of dealing with 3D-surface patch for HLA epitope prediction
- Although less performant, the Decision Tree model trained with the "Non-redundant" dataset provided interpretable classification rules and further highlighted the relevance of side-chain flexibility and patch-level descriptors for predicting HLA epitopes.
- The exhaustive performance evaluation of HLA-EpiCheck revealed that the tool is less likely to make false positive predictions and has room for improvement on false negative predictions.

The fourth contribution, presented in Chapter 5, presents the implementation of an HLA 3D database and a user-friendly graphical interface called HLA-3D-Diff. This tool is intended to facilitate visualization and superposition of 3D HLA antigens to highlight their structural differences along MD simulations and can be used as an aid to understand complex or unexpected immunization patterns.

# Perspectives

The work carried out in this thesis offers several perspectives for future studies in the short or long term.

In this thesis, molecular dynamics has proven its usefulness as a technique that provides highly relevant information when it comes to better understanding HLA antigen-antibody recognition. Although the descriptors derived from MD data and used in this thesis are effective, they likely fail to extract all the information encoded in this type of data. Therefore, future work could be oriented towards the development of new descriptors derived from MD trajectories, allowing extraction of new protein characteristics. Capturing longitudinal temporal information could be envisaged. For instance, the maximal number of consecutive frames without any major N-RMSF fluctuation in the patch could be an indication of the relative stability of the patch, favorable for interaction with antibodies.

In the current context, it may seem relevant for future developments to try to use deep-learning (DL) models for HLA epitope prediction. In particular, I think that 3D embedding approaches using DL could be a lead to follow. For example, dMaSIF [151] is a DL approach that generates embeddings of surface patches that simultaneously capture geometric and physicochemical properties. However, several difficulties could arise. The first is that approaches such as dMaSIF have so far only been implemented with PDB data, i.e., static structures. An interesting challenge would be to adapt dMaSIF to the use of MD data. A second difficulty would be linked to the fact that DL-based approaches require enormous amounts of data for training and that MD simulations require enormous amounts of computation. This raises the challenge of sustainability of digital research. In this sense, a contribution of this thesis is the free availability of all MD simulations performed that have been deposited in the open repository of Recherche Data Gouv (doi:10.57745/GXZHH8). In future, dedicated repositories for MD data will likely be developed, continuing initiatives such as MDDB [154] and MDverse [155]. Large scale availability of MD simulation records will represent an invaluable source of data for DL experimentations, similar to what happened with PDB and AlphaFold, for example.

Regarding the use of shape descriptors, i.e., Zernike descriptors, a research direction could be to explore alternative mesh/surface generation techniques that manage to generate consistent meshes for 3D-surface patches. Discussions and collaboration with experts in 3D surfaces should be initiated for this purpose.

Concerning HLA-EpiCheck, there is a lot to be done in the short term to understand false negative predictions. In particular, some false negative predictions can be explained as in the example of eplet 23L by finding a positive patch in the close vicinity of the negative one. Systematic screening of eplet neighborhood at the patch level can be envisaged to improve HLA epitope prediction.

In the long term, it can be envisaged to train a more general-purpose predictor using 3D-surface patches and MD trajectories acquired from a more diverse set of proteins for which epitopes are known. This could answer the question of whether our set of descriptors, used in the appropriate setting, could improve B-cell epitope prediction in general compared to DiscoTope-3.0.

Regarding the perspectives of using HLA-EpiCheck when defining the D/R matching, one could consider generating an MML (mismatch load) score based on HLA epitope prediction as explained in section 4.4. Retrospective studies should then be conducted to evaluate the relevance of this HLA-EpiCheck MML score on DSA production and graft rejection.

Finally, concerning HLA-3D-Diff, this interactive tool could be enriched with HLA-EpiCheck predictions, the previously mentioned MML score and side-chain flexibility (N-RMSF) coloration of 3D

structures. In this way, HLA-3D-Diff could become the structural component of a future decision support tool for D/R matching.

# Supplementary materials

Table S1: List of modeled HLA antigens. DP and DQ antigens are composed of two polymorphic protein chains, produced from the DPA1 and DPB1, or DQA1 and DQB1 loci, respectively.

| Locus | Antigens | Total |
|---|---|---|
| A | A*01:01, A*02:01, A*02:02, A*02:03, A*02:05, A*02:06, A*02:07, A*03:01, A*11:01, A*11:02, A*23:01, A*24:02, A*24:03, A*25:01, A*26:01, A*29:01, A*29:02, A*30:01, A*30:02, A*30:03, A*31:01, A*32:01, A*33:01, A*33:03, A*34:01, A*36:01, A*43:01, A*66:01, A*66:02, A*68:01, A*68:02, A*69:01, A*74:01, A*80:01 | 34 |
| B | B*07:02, B*08:01, B*13:01, B*13:02, B*14:01, B*14:02, B*15:01, B*15:02, B*15:03, B*15:10, B*15:11, B*15:12, B*15:13, B*15:16, B*18:01, B*27:03, B*27:04, B*27:05, B*27:06, B*27:08, B*27:09, B*35:01, B*35:08, B*37:01, B*38:01, B*39:01, B*40:01, B*40:02, B*40:06, B*41:01, B*41:03, B*41:04, B*42:01, B*42:02, B*44:02, B*44:03, B*44:05, B*45:01, B*46:01, B*47:01, B*48:01, B*49:01, B*50:01, B*51:01, B*51:02, B*52:01, B*53:01, B*54:01, B*55:01, B*56:01, B*57:01, B*57:03, B*58:01, B*59:01, B*67:01, B*73:01, B*78:01, B*81:01, B*82:01 | 59 |
| C | C*01:02, C*02:02, C*03:02, C*03:03, C*03:04, C*04:01, C*05:01, C*06:02, C*07:02, C*08:01, C*08:02, C*12:03, C*14:02, C*15:02, C*16:01, C*17:01, C*18:02 | 17 |
| DP | DPA1*01:03-DPB1*01:01, DPA1*01:03-DPB1*02:01, DPA1*01:03-DPB1*03:01, DPA1*01:03-DPB1*04:01, DPA1*01:03-DPB1*04:02, DPA1*01:03-DPB1*06:01, DPA1*01:03-DPB1*11:01, DPA1*01:03-DPB1*19:01, DPA1*01:03-DPB1*23:01, DPA1*01:03-DPB1*28:01, DPA1*01:04-DPB1*18:01, DPA1*01:05-DPB1*03:01, DPA1*01:05-DPB1*18:01, DPA1*01:05-DPB1*28:01, DPA1*02:01-DPB1*01:01, DPA1*02:01-DPB1*03:01, DPA1*02:01-DPB1*05:01, DPA1*02:01-DPB1*06:01, DPA1*02:01-DPB1*09:01, DPA1*02:01-DPB1*13:01, DPA1*02:01-DPB1*14:01, DPA1*02:01-DPB1*15:01, DPA1*02:01-DPB1*17:01, DPA1*02:01-DPB1*18:01, DPA1*02:02-DPB1*05:01, DPA1*02:02-DPB1*10:01, DPA1*02:02-DPB1*11:01, DPA1*02:02-DPB1*13:01, DPA1*03:01-DPB1*13:01, DPA1*03:01-DPB1*20:01, DPA1*04:01-DPB1*28:01 | 31 |
| DQ | DQA1*01:01-DQB1*05:01, DQA1*01:01-DQB1*06:02, DQA1*01:02-DQB1*05:01, DQA1*01:02-DQB1*05:02, DQA1*01:02-DQB1*06:02, DQA1*01:02-DQB1*06:04, DQA1*01:02-DQB1*06:09, DQA1*01:03-DQB1*06:01, DQA1*01:03-DQB1*06:03, DQA1*02:01-DQB1*02:01, DQA1*02:01-DQB1*02:02, DQA1*02:01-DQB1*03:01, DQA1*02:01-DQB1*03:02, DQA1*02:01-DQB1*03:03, DQA1*02:01-DQB1*04:01, DQA1*02:01-DQB1*04:02, DQA1*03:01-DQB1*02:01, DQA1*03:01-DQB1*03:01, DQA1*03:01-DQB1*03:02, DQA1*03:01-DQB1*03:03, DQA1*03:02-DQB1*03:02, DQA1*03:02-DQB1*03:03, DQA1*03:03-DQB1*04:01, DQA1*04:01-DQB1*02:01, DQA1*04:01-DQB1*04:02, DQA1*05:01-DQB1*02:01, DQA1*05:03-DQB1*03:01, DQA1*05:05-DQB1*03:01, DQA1*05:08-DQB1*02:01, DQA1*06:01-DQB1*03:01 | 30 |
| DR | DRB1*01:01, DRB1*01:02, DRB1*01:03, DRB1*03:01, DRB1*03:02, DRB1*04:01, DRB1*04:02, DRB1*04:03, DRB1*04:04, DRB1*04:05, DRB1*07:01, DRB1*08:01, DRB1*09:01, DRB1*09:02, DRB1*10:01, DRB1*11:01, DRB1*11:04, DRB1*12:01, DRB1*12:02, DRB1*13:01, DRB1*13:03, DRB1*14:01, DRB1*14:02, DRB1*14:54, DRB1*15:01, DRB1*15:02, DRB1*15:03, DRB1*16:01, DRB1*16:02, DRB3*01:01, DRB3*02:02, DRB3*03:01, DRB4*01:01, DRB4*01:03, DRB5*01:01, DRB5*02:02 | 36 |

Table S2: Length of modeled sequences per locus. Only the extracellular globular part of the protein was modeled, so the signal peptide, the amino acids linking the extracellular globular part to the transmembrane part, the transmembrane part and the intracellular part were excluded

| **Locus** | **A, B, C** | | **DP** | | **DQ** | | **DR** | |
|---|---|---|---|---|---|---|---|---|
| **chain** | **chain α** | **β2-microglobulin** | **chain α** | **chain β** | **chain α** | **chain β** | **chain α** | **chain β** |
| **modeled length** | 276 | 99 | 183 | 191 | 186 | 192 | 182 | 192 |

Table S3: Maximum solvent accessibility of residues [123]

| **Residue** | **Max. accessibility [Å$^2$]** |
|---|---|
| ALA | 138 |
| ARG | 285 |
| ASN | 204 |
| ASP | 204 |
| CYS | 169 |
| GLU | 233 |
| GLN | 234 |
| GLY | 114 |
| HIS | 231 |
| ILE | 208 |
| LEU | 211 |
| LYS | 246 |
| MET | 227 |
| PHE | 251 |
| PRO | 166 |
| SER | 161 |
| THR | 182 |
| TRP | 295 |
| TYR | 274 |
| VAL | 184 |

Table S4: List of antigens belonging to serological groups BW4 and BW6.

| Serological group | Antigens | Total |
|---|---|---|
| BW4 | B*13:01 , B*13:02, B*15:13, B*15:16, B*27:03, B*27:04, B*27:05, B*27:06, B*27:09, B*37:01, B*38:01, B*44:02, B*44:03, B*44:05, B*47:01, B*49:01,B*51:01, B*51:02, B*52:01, B*53:01, B*57:01, B*57:03, B*58:01 | 23 |
| BW6 | B*07:02, B*08:01, B*14:01, B*14:02, B*15:01, B*15:02, B*15:03, B*15:10, B*15:11, B*15:12, B*27:08, B*35:01, B*35:08, B*39:01, B*40:01, B*40:02, B*40:06, B*41:03, B*41:04, B*42:01, B*42:02, B*45:01, B*46:01 , B*48:01, B*50:01, B*54:01, B*55:01, B*56:01, B*67:01, B*73:01, B*81:01, B*82:01 | 32 |

Table S5: Differences in the frequencies of occurrence (in %) between surface and sequence for "All", "Confirmed", "Non-confirmed" and "Non-eplet" groups. Cells highlighted in green correspond to the most important variations in each group considered ($|\Delta| \geq 2.5$).

| | Δ percentage surface - sequence) | | | |
|---|---|---|---|---|
| AA | All | Confirmed | Non-confirmed | Non-eplet |
| ALA | -1.0 | 0.3 | 0.2 | -1.3 |
| ARG | 3.5 | 3.3 | 4.4 | 3.5 |
| ASN | 0.5 | 1.0 | -1.8 | 0.5 |
| ASP | 1.4 | -0.5 | -1.4 | 1.5 |
| CYS | -1.6 | -0.1 | 0.0 | -1.8 |
| GLN | 1.5 | 1.5 | 0.6 | 1.5 |
| GLU | 5.1 | -0.7 | 3.5 | 5.1 |
| GLY | 0.6 | -0.6 | -0.7 | 0.6 |
| HSD | -0.1 | 0.7 | -0.2 | 0 |
| ILE | -1.2 | -0.1 | 2.2 | -1.3 |
| LEU | -3 | -1.1 | -2.7 | -2.8 |
| LYS | 2.2 | -0.2 | 1.3 | 2.3 |
| MET | -0.7 | -0.2 | -1.8 | -0.6 |
| PHE | -3.4 | -0.9 | -1.9 | -3.3 |
| PRO | 1.5 | 0.6 | 2.0 | 1.3 |
| SER | 0.5 | -1.2 | -2.1 | 0.7 |
| THR | 0.9 | 0.3 | 3.1 | 0.9 |
| TRP | -1.6 | 0.6 | -0.3 | -1.7 |
| TYR | -2.7 | -1.0 | -5 | -2.5 |
| VAL | -2.4 | -1.8 | 0.7 | -2.7 |

Table S6: Differences in the frequencies of occurrence of surface AAs (in %) among "Confirmed", "Non-confirmed" and "Non-eplet" groups. Surface AAs are computed from MD trajectories. Cells highlighted in green correspond to the most important variations for each comparison (|Δ| >=2.5).

| | **Δ frequency of occurrence** | | |
|---|---|---|---|
| **AA** | **Non-eplet vs Confirmed [%]** | **Non-confirmed vs Confirmed [%]** | **Non-eplet vs Non-confirmed [%]** |
| ALA | 6.1 | 1.1 | 5 |
| ARG | 3.3 | 0.6 | 2.8 |
| ASN | 3.2 | 3.5 | -0.3 |
| ASP | -3.6 | 0.8 | -4.4 |
| CYS | 0.3 | -0.1 | 0.4 |
| GLN | -0.2 | 4.2 | -4.4 |
| GLU | -6.0 | -2.3 | -3.7 |
| GLY | -3.1 | 2.3 | -5.4 |
| HSD | -0.7 | -3.0 | 2.3 |
| ILE | 1.6 | -4.1 | 5.7 |
| LEU | -1.4 | -0.4 | -1.0 |
| LYS | -0.8 | 1.0 | -1.9 |
| MET | 1.4 | 1.3 | 0.1 |
| PHE | -0.4 | -0.9 | 0.5 |
| PRO | -4.6 | -1.9 | -2.7 |
| SER | -0.8 | 1.7 | -2.5 |
| THR | 6.2 | 4.3 | 1.9 |
| TRP | 1.9 | 2.5 | -0.6 |
| TYR | -1.5 | -3.6 | 2.1 |
| VAL | -1.0 | -6.8 | 5.8 |

Table S7: Comparison of the frequency of occurrence of surface AAs (in %) calculated from static structure and from MD trajectories for the "All", "Confirmed", "Non-confirmed", and "Non-eplet" groups. For the "Confirmed" and "Non-confirmed" groups, the $X^2$ test revealed no significant difference in the surface AA-type distribution when calculated from the static structure or from MD trajectories (p-values of 0.490 and 0.051, respectively). However, for the "All" and "Non-eplet" groups the p-values were significant (less than $10^{-16}$) probably because the number of AAs handled is much higher.

| | | All | | Confirmed | | Non-confirmed | | Non-eplet | |
|---|---|---|---|---|---|---|---|---|---|
| | AA | Static | MD | Static | MD | Static | MD | Static | MD |
| Non-polar non-hydrophobic | GLY | 6.3 | 6.1 | 2.7 | 3.1 | 1.1 | 0.9 | 6.4 | 6.3 |
| | PRO | 7.0 | 7 | 2.6 | 2.4 | 4.8 | 4.3 | 7.0 | 7 |
| Non-polar hydrophobic | ALA | 4.9 | 4.6 | 10.9 | 10.3 | 10.2 | 9.2 | 4.5 | 4.2 |
| | VAL | 4.5 | 4.8 | 4.4 | 3.8 | 8.7 | 10.6 | 4.5 | 4.8 |
| | LEU | 3.6 | 4.3 | 2.9 | 3.1 | 3.1 | 3.5 | 3.7 | 4.5 |
| | ILE | 2.1 | 2.1 | 3.3 | 3.5 | 8.3 | 7.5 | 1.8 | 1.9 |
| | MET | 0.6 | 0.5 | 2.3 | 1.9 | 0.7 | 0.6 | 0.6 | 0.5 |
| | PHE | 0.8 | 1.1 | 0.7 | 0.6 | 1.6 | 1.5 | 0.8 | 1 |
| | TRP | 0.5 | 1 | 1.2 | 2.8 | 0.1 | 0.3 | 0.5 | 0.9 |
| | CYS | 0.1 | 0.1 | 0.3 | 0.3 | 0.4 | 0.4 | 0.0 | 0 |
| Polar uncharged | SER | 6.0 | 6.2 | 5.2 | 5.5 | 3.6 | 3.8 | 6.2 | 6.3 |
| | THR | 7.9 | 7.9 | 14.2 | 14.2 | 9.8 | 9.8 | 7.9 | 7.9 |
| | TYR | 1.6 | 1.9 | 0.5 | 0.6 | 3.3 | 4.2 | 1.7 | 2 |
| | ASN | 4.4 | 4.2 | 7.4 | 7.4 | 4.9 | 3.9 | 4.3 | 4.2 |
| | GLN | 6.6 | 6.5 | 6.6 | 6.5 | 1.7 | 2.3 | 6.8 | 6.6 |
| | HSD | 3.2 | 3.4 | 2.8 | 2.8 | 5.1 | 5.8 | 3.3 | 3.5 |
| Polar charged | ASP | 7.5 | 7.6 | 4.4 | 4.3 | 3.6 | 3.5 | 7.7 | 7.9 |
| | GLU | 14.6 | 13.9 | 7.8 | 7.9 | 10.0 | 10.2 | 14.6 | 13.9 |
| | LYS | 6.5 | 6 | 4.9 | 5.2 | 4.5 | 4.2 | 6.5 | 6.1 |
| | ARG | 11.3 | 10.7 | 15.1 | 14 | 14.7 | 13.4 | 11.2 | 10.6 |

Table S8: Difference in the frequency of occurrence of surface AAs between percentages obtained when surface AAs are computed from MD trajectories or from static structures for the “All”, "Confirmed", "Non-confirmed" and "Non-eplet" groups. Cells highlighted in green correspond to the most important variations in each group considered.

| | | Δ percentage (MD - static) | | | |
|---|---|---|---|---|---|
| | **AA** | **All** | **Confirmed** | **Non-confirmed** | **Non-eplet** |
| Non-polar non-hydrophobic | GLY | -0.2 | 0.4 | -0.2 | -0.1 |
| | PRO | 0 | -0.2 | -0.5 | 0 |
| Non-polar hydrophobic | ALA | -0.3 | -0.6 | -1 | -0.3 |
| | VAL | 0.3 | -0.6 | 1.9 | 0.3 |
| | LEU | 0.7 | 0.2 | 0.4 | 0.8 |
| | ILE | 0 | 0.2 | -0.8 | 0.1 |
| | MET | -0.1 | -0.4 | -0.1 | -0.1 |
| | PHE | 0.3 | -0.1 | -0.1 | 0.2 |
| | TRP | 0.5 | 1.6 | 0.2 | 0.4 |
| Polar non-charged | CYS | 0 | 0 | 0 | 0 |
| | SER | 0.2 | 0.3 | 0.2 | 0.1 |
| | THR | 0 | 0 | 0 | 0 |
| | TYR | 0.3 | 0.1 | 0.9 | 0.3 |
| | ASN | -0.2 | 0 | -1 | -0.1 |
| | GLN | -0.1 | -0.1 | 0.6 | -0.2 |
| | HSD | 0.2 | 0 | 0.7 | 0.2 |
| Polar charged | ASP | 0.1 | -0.1 | -0.1 | 0.2 |
| | GLU | -0.7 | 0.1 | 0.2 | -0.7 |
| | LYS | -0.5 | 0.3 | -0.3 | -0.4 |
| | ARG | -0.6 | -1.1 | -1.3 | -0.6 |

Table S9: Median RSASA values for different groups of AAs. RSASA values were computed for MD data (columns on the left)and for static 3D structures (columns on the right). In the case of MD data, RSASA values for each AA and for each frame of the trajectories were considered while for static 3D structures RSASA values were computed for each AA in the single structure corresponding to the first frame of each MD trajectory.

| | | Median RSASA [%] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MD | | | | Static 3D | | | |
| **Type AA** | **AA** | **All** | **Confirmed** | **Non-confirmed** | **Non-eplet** | **All** | **Confirmed** | **Non-confirmed** | **Non-eplet** |
| Non-polar non-hydrophobic | GLY | 27.57 | 17.70 | 8.17 | 28.43 | 24.52 | 12.24 | 8.55 | 25.08 |
| | PRO | 35.41 | 56.22 | 56.53 | 33.64 | 32.42 | 54.22 | 55.18 | 30.82 |
| Non-polar hydrophobic | ALA | 16.46 | 33.79 | 21.45 | 14.94 | 15.04 | 29.58 | 17.98 | 13.41 |
| | VAL | 12.75 | 16.94 | 25.94 | 11.88 | 11.83 | 16.15 | 20.58 | 10.90 |
| | LEU | 10.56 | 30.34 | 8.42 | 10.50 | 8.23 | 32.35 | 6.34 | 7.98 |
| | ILE | 13.70 | 25.68 | 35.63 | 12.36 | 12.45 | 20.08 | 32.76 | 10.57 |
| | PHE | 4.85 | 13.82 | 3.16 | 4.93 | 2.19 | 11.89 | 1.77 | 2.19 |
| | MET | 9.15 | 36.17 | 6.62 | 9.32 | 6.36 | 29.67 | 5.56 | 5.77 |
| | TRP | 10.38 | 26.68 | 15.45 | 9.92 | 8.14 | 14.46 | 12.83 | 7.69 |
| Polar uncharged | SER | 26.36 | 22.43 | 10.14 | 27.63 | 22.89 | 20.33 | 9.79 | 23.95 |
| | THR | 27.59 | 39.04 | 32.52 | 26.76 | 23.41 | 33.01 | 30.49 | 22.85 |
| | CYS | 0.00 | 15.71 | 15.39 | 0.00 | 0.00 | 8.29 | 5.62 | 0.00 |
| | TYR | 10.20 | 16.97 | 7.77 | 10.61 | 6.31 | 9.18 | 4.57 | 6.51 |
| | ASN | 26.79 | 34.10 | 7.36 | 27.88 | 24.39 | 29.77 | 6.39 | 25.08 |
| | GLN | 35.52 | 38.84 | 21.14 | 35.68 | 35.64 | 33.43 | 12.58 | 36.31 |
| | HSD | 22.41 | 49.22 | 19.03 | 22.12 | 17.43 | 47.79 | 12.86 | 17.59 |
| Polar charged | ASP | 31.84 | 36.96 | 14.08 | 32.42 | 28.38 | 39.68 | 10.44 | 28.94 |
| | GLU | 43.03 | 36.03 | 37.64 | 43.58 | 41.71 | 33.90 | 35.95 | 42.41 |
| | LYS | 50.71 | 42.02 | 38.83 | 51.37 | 47.86 | 42.42 | 36.37 | 48.45 |
| | ARG | 35.13 | 51.25 | 43.93 | 33.18 | 32.67 | 48.70 | 43.65 | 30.56 |

Table S10: RSASA median values grouped according three sets of AA types (non-polar, polar uncharged, polar charged) and computed from MD trajectories for all AAs present in the "Epitope" and "Non-epitope" groups for the three patch sizes. The total number of AAs in each group is indicated in parentheses. Each AA is represented by 500 frames. Cells with values yielding a difference of RSASA median values greater than 7% (absolute value) between the "Epitope" group and the "Non-epitope" group are highlighted in green. No significant differences were found depending on the patch sizes.

| | | **Radius 15Å** | | **Radius 12Å** | | **Radius 9Å** | |
|---|---|---|---|---|---|---|---|
| **AA set** | **AA** | **Epitope** (36 545) | **Non-epitope** (84 107) | **Epitope** (21 831) | **Non-epitope** (65 259) | **Epitope** (12 534) | **Non-epitope** (43 695) |
| Non-polar non-hydrophobic | GLY | 34.3 | 43.9 | 34 | 44.1 | 33.8 | 44.1 |
| | PRO | 46.5 | 44 | 45.6 | 45.1 | 44.7 | 49.1 |
| Non-polar hydrophobic | ALA | 42.2 | 34.7 | 40.2 | 37.1 | 40.3 | 42.3 |
| | VAL | 31.9 | 35.9 | 34.2 | 34.8 | 38.1 | 35.1 |
| | LEU | 30.3 | 31.2 | 31.4 | 31.6 | 32.7 | 32.5 |
| | ILE | 39.1 | 44.4 | 38.1 | 42.6 | 37.9 | 41.3 |
| | PHE | 28.9 | 26.3 | 29.1 | 27.1 | 29.5 | 26.1 |
| | MET | 55 | 37.9 | 61 | 23.4 | 60.3 | 22.4 |
| | TRP | 25.7 | 23.2 | 26.1 | 23.4 | 28.1 | 23.4 |
| Polar uncharged | CYS | 26.9 | 0 | 27 | 0 | 26.9 | 0 |
| | SER | 37 | 36.1 | 37.6 | 35.2 | 35.6 | 37.3 |
| | THR | 41.2 | 35.7 | 42 | 36.7 | 42.3 | 38.1 |
| | TYR | 29.1 | 23.8 | 28.9 | 24.8 | 29.8 | 27.6 |
| | ASN | 38 | 34.9 | 34.6 | 35.6 | 32.6 | 39.3 |
| | GLN | 41.2 | 38.1 | 41.2 | 37 | 42.3 | 37.7 |
| | HSD | 43.8 | 35.5 | 47.4 | 34.3 | 46.9 | 31.1 |
| Polar charged | ASP | 38.6 | 39.5 | 40.1 | 39.8 | 40.8 | 37.4 |
| | GLU | 41.7 | 49.2 | 40.5 | 48.5 | 40.7 | 47.4 |
| | LYS | 50.7 | 55.7 | 48.9 | 56.1 | 47.2 | 55.5 |
| | ARG | 39.9 | 35.9 | 40.3 | 37.6 | 42 | 38.7 |

Table S11: Parameters used in the gridsearch process for the redundant and nonredundant dataset. The best values found for the HLA-EpiCheck final model are highlighted in green.

| Learning algorithm | Parameters used in the gridsearch |
|---|---|
| **Decision Tree** | criterion:["gini", 'entropy', 'log_loss'], max_features:[None, 'sqrt', 'log2'], min_samples_leaf:[1, 2, 3, 5, 10], min_samples_split:[2, 3, 5, 10], splitter:['best', 'random'] |
| **Random Forest** | criterion:["gini", 'entropy', 'log_loss'], max_features:[None, 'sqrt', 'log2'], min_samples_leaf:[1,2,3,5], min_samples_split:[2,3,5], n_estimators:[50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000], warm_start:[False, True] |
| **ExtraTrees** | criterion:["gini", 'entropy', 'log_loss'], max_features:[None, 'sqrt', 'log2'], min_samples_leaf:[1,2,3,4] , min_samples_split:[2,3,4,5], n_estimators:[50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 900,1000], warm_start:[False, True], |
| **GradientTrees** | criterion:['friedman_mse', 'squared_error'], learning_rate:[0, 0.2, 0.3, 0.5, 0.7], loss:['log_loss', 'exponential'], max_features:[None, 'sqrt', 'log2'], min_samples_leaf:[1, 2, 3], min_samples_split:[2, 3, 5, 7, 10, 15], n_estimators:[100, 200, 300, 400, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1700], subsample:[0.2, 0.5, 0.7, 0.8, 1] , warm_start:[True, False], |
| **HistoGradientTrees** | l2_regularization:[0, 0.2, 0.5, 0.8], learning_rate:[0, 0.2, 0.5, 0.8], loss:['log_loss', 'auto'], max_bins:[255, 200, 155, 100, 55, 45, 35, 25, 15], min_samples_leaf:[2, 3, 5, 7, 10, 12, 15, 20], warm_start:[True, False], max_iter:[300, 500, 600] |
| **MLP** | batch:[32, 64, 128], units_layer_1:[16, 32, 64, 128], units_layer_2:[16, 32, 64, 128], dropout_1:[0.0, 0.1, 0.2] , dropout_2:[0, 0.1, 0.2] |

Table S12: Performance evaluation on the test set using the "Non-redundant" dataset. The results correspond to the mean and standard deviation of the 10 splits generated. The F1-score, precision and recall metrics were calculated for each label (Epitope/Non-epitope).

| | | F1 | | Precision | | Recall | |
|---|---|---|---|---|---|---|---|
| | | **Mean** | **Std.** | **Mean** | **Std.** | **Mean** | **Std.** |
| **Decision Tree** | **Epitope** | 0.66 | 0.11 | 0.70 | 0.11 | 0.64 | 0.14 |
| | **Non-epitope** | 0.82 | 0.05 | 0.80 | 0.06 | 0.84 | 0.06 |
| **ExtraTrees** | **Epitope** | 0.70 | 0.05 | 0.83 | 0.08 | 0.61 | 0.05 |
| | **Non-epitope** | 0.86 | 0.02 | 0.80 | 0.02 | 0.92 | 0.04 |
| **GradientTrees** | **Epitope** | 0.76 | 0.09 | 0.82 | 0.10 | 0.71 | 0.11 |
| | **Non-epitope** | 0.87 | 0.05 | 0.85 | 0.05 | 0.91 | 0.06 |
| **KNN** | **Epitope** | 0.57 | 0.06 | 0.69 | 0.12 | 0.50 | 0.07 |
| | **Non-epitope** | 0.80 | 0.04 | 0.75 | 0.03 | 0.86 | 0.07 |

Table S13: Performance evaluation on the test set using the "Redundant" dataset. The results correspond to the mean and standard deviation of the 10 splits generated. The F1-score, precision and recall metrics were calculated for each label (Epitope/Non-epitope).

| | | F1 | Precision | Recall |
|---|---|---|---|---|
| **Decision Tree** | **Epitope** | 0.72 | 0.74 | 0.70 |
| | **Non-epitope** | 0.88 | 0.87 | 0.89 |
| **Random Forest** | **Epitope** | 0.85 | 0.90 | 0.80 |
| | **Non-epitope** | 0.94 | 0.92 | 0.96 |
| **ExtraTrees** | **Epitope** | 0,87 | 0,93 | 0,82 |
| | **Non-epitope** | 0.95 | 0.92 | 0.97 |
| **GradientTrees** | **Epitope** | 0.83 | 0.87 | 0.78 |
| | **Non-epitope** | 0.93 | 0.91 | 0.95 |
| **HistoGradientTrees** | **Epitope** | 0.86 | 0.91 | 0.81 |
| | **Non-epitope** | 0.95 | 0.92 | 0.97 |
| **MLP** | **Epitope** | 0.58 | 0.69 | 0.50 |
| | **Non-epitope** | 0.85 | 0.80 | 0.90 |
| **KNN** | **Epitope** | 0.84 | 0.92 | 0.77 |
| | **Non-epitope** | 0.94 | 0.91 | 0.97 |

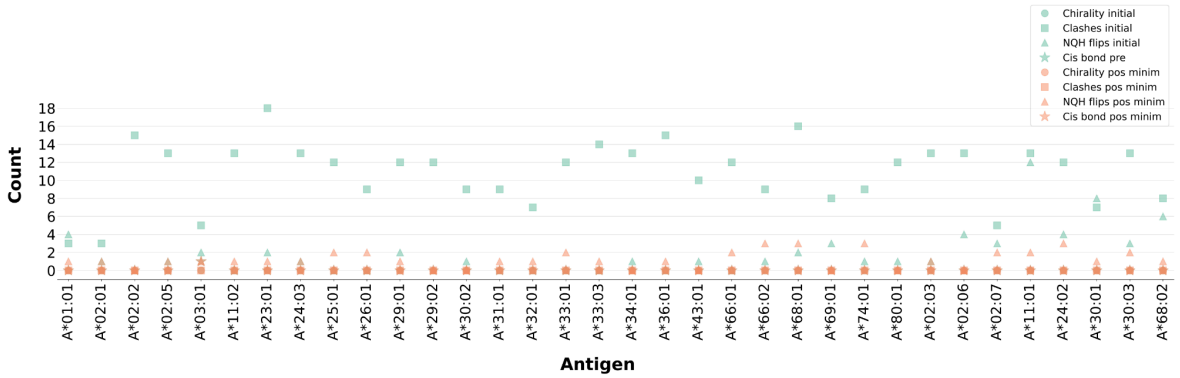
Table S14: HLA-EpiCheck predictions on the eplet 23L and its neighboring residues. Columns **Ag1** and **Ag2** correspond to predictions on antigens DQA1*02:01-DQB1*04:01 and DQA1*03:03-DQB1*04:01 respectively. Missing prediction value corresponds to a non solvent-accessible residue.

| Central residue | Chain | Ag1 | Ag2 |
|---|---|---|---|
| 23L | β | 1 | 0 |
| 3I | α | - | 0 |
| 4V | α | 0 | 0 |
| 19N | β | 1 | 0 |
| 22E | β | 1 | 1 |
| 25R | β | 1 | 1 |
| 43D | β | 1 | 1 |
| 80R | β | 1 | 1 |

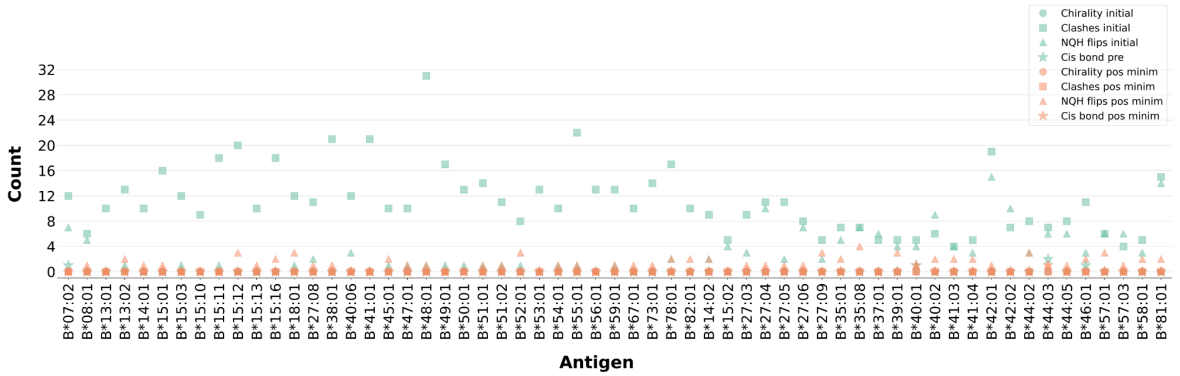
Table S15: HLA-Epicheck predictions before aggregation for all antigens displaying the eplet 75I. Ep=Epitope, Non-Ep=Non-Epitope and NA=Non-applicable.

| No | Antigen | 74I | 75I | 160D | 161D | 162I | 163I |
|---|---|---|---|---|---|---|---|
| 1 | DQA1*01:01-DQB1*05:01 | NA | Ep | NA | Non-Ep | NA | Non-Ep |
| 2 | DQA1*01:01-DQB1*06:02 | NA | Ep | NA | Non-Ep | NA | Non-Ep |
| 3 | DQA1*01:02-DQB1*05:01 | NA | Ep | NA | Ep | NA | Ep |
| 4 | DQA1*01:02-DQB1*05:02 | NA | Ep | NA | Non-Ep | NA | Non-Ep |
| 5 | DQA1*01:02-DQB1*06:02 | NA | Ep | NA | Ep | NA | Ep |
| 6 | DQA1*01:02-DQB1*06:04 | NA | Ep | NA | Ep | NA | Ep |
| 7 | DQA1*01:02-DQB1*06:09 | NA | Ep | NA | Non-Ep | NA | Non-Ep |
| 8 | DQA1*01:03-DQB1*06:01 | NA | Ep | NA | Ep | NA | Ep |
| 9 | DQA1*01:03-DQB1*06:03 | NA | Ep | NA | Ep | NA | Ep |
| 10 | DQA1*02:01-DQB1*02:01 | Ep | NA | Non-Ep | NA | Ep | NA |
| 11 | DQA1*02:01-DQB1*02:02 | Ep | NA | Ep | NA | Ep | NA |
| 12 | DQA1*02:01-DQB1*03:01 | Ep | NA | Ep | NA | Ep | NA |
| 13 | DQA1*02:01-DQB1*03:02 | Ep | NA | Non-Ep | NA | Non-Ep | NA |
| 14 | DQA1*02:01-DQB1*03:03 | Ep | NA | Non-Ep | NA | Non-Ep | NA |
| 15 | DQA1*02:01-DQB1*04:01 | Ep | NA | Non-Ep | NA | Non-Ep | NA |
| 16 | DQA1*02:01-DQB1*04:02 | Ep | NA | Non-Ep | NA | Non-Ep | NA |
| 17 | DQA1*03:01-DQB1*02:01 | NA | Ep | NA | Non-Ep | NA | Ep |
| 18 | DQA1*03:01-DQB1*03:01 | NA | Ep | NA | Non-Ep | NA | Ep |
| 19 | DQA1*03:01-DQB1*03:02 | NA | Ep | NA | Ep | NA | Non-Ep |
| 20 | DQA1*03:01-DQB1*03:03 | NA | Ep | NA | Non-Ep | NA | Ep |
| 21 | DQA1*03:02-DQB1*03:02 | NA | Ep | NA | Non-Ep | NA | Non-Ep |
| 22 | DQA1*03:02-DQB1*03:03 | NA | Ep | NA | Ep | NA | Ep |
| 23 | DQA1*03:03-DQB1*04:01 | NA | Ep | NA | Non-Ep | NA | Ep |
| 24 | DQA1*04:01-DQB1*02:01 | Ep | NA | Non-Ep | NA | Ep | NA |
| 25 | DQA1*04:01-DQB1*04:02 | Ep | NA | Ep | NA | Ep | NA |
| 26 | DQA1*06:01-DQB1*03:01 | Ep | NA | Non-Ep | NA | Ep | NA |

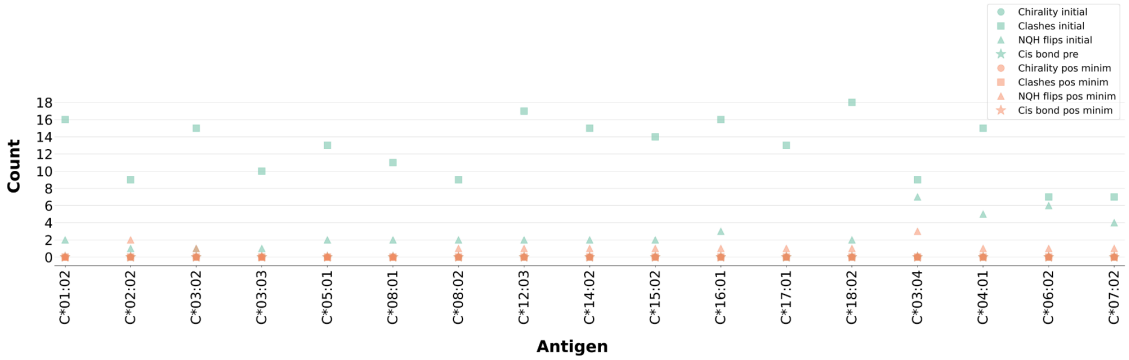
A)


Chirality initial
Clashes initial
NQH flips initial
Cis bond pre
Chirality pos minim
Clashes pos minim
NQH flips pos minim
Cis bond pos minim
Count
Antigen
A*01:01
A*02:01
A*02:02
A*02:05
A*03:01
A*11:02
A*23:01
A*24:03
A*25:01
A*26:01
A*29:01
A*29:02
A*30:02
A*31:01
A*32:01
A*33:01
A*33:03
A*34:01
A*36:01
A*43:01
A*66:01
A*66:02
A*68:01
A*69:01
A*74:01
A*80:01
A*02:03
A*02:06
A*02:07
A*11:01
A*24:02
A*30:01
A*30:03
A*68:02

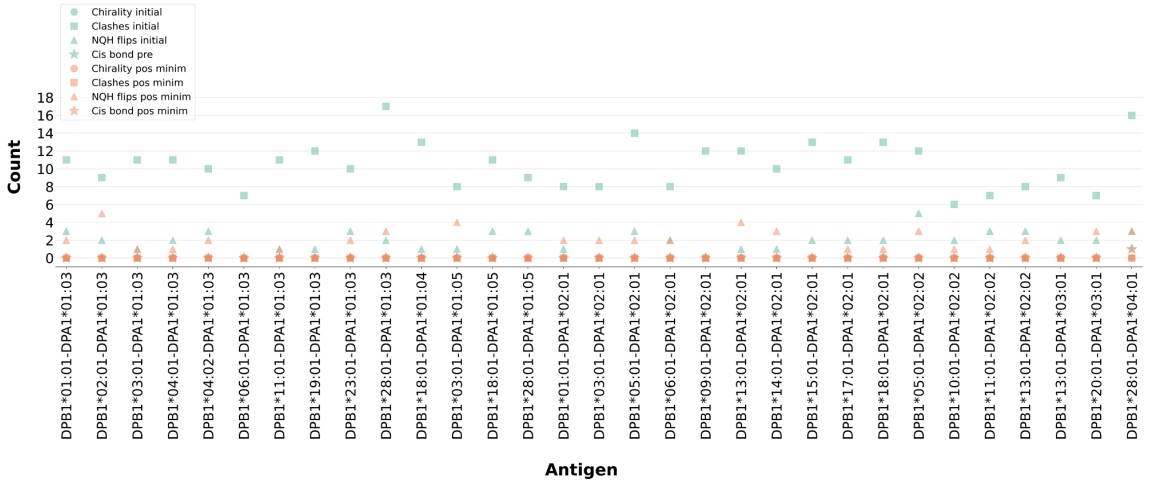
B)


Chirality initial
Clashes initial
NQH flips initial
Cis bond pre
Chirality pos minim
Clashes pos minim
NQH flips pos minim
Cis bond pos minim
Count
Antigen
B*07:02
B*08:01
B*13:01
B*13:02
B*14:01
B*15:01
B*15:03
B*15:10
B*15:11
B*15:12
B*15:13
B*15:16
B*18:01
B*27:08
B*38:01
B*40:06
B*41:01
B*45:01
B*47:01
B*48:01
B*49:01
B*50:01
B*51:01
B*51:02
B*52:01
B*53:01
B*54:01
B*55:01
B*56:01
B*59:01
B*67:01
B*73:01
B*78:01
B*82:01
B*14:02
B*15:02
B*27:03
B*27:04
B*27:05
B*27:06
B*27:09
B*35:01
B*35:08
B*37:01
B*39:01
B*40:01
B*40:02
B*41:03
B*41:04
B*42:01
B*42:02
B*44:02
B*44:03
B*44:05
B*46:01
B*57:01
B*57:03
B*58:01
B*81:01

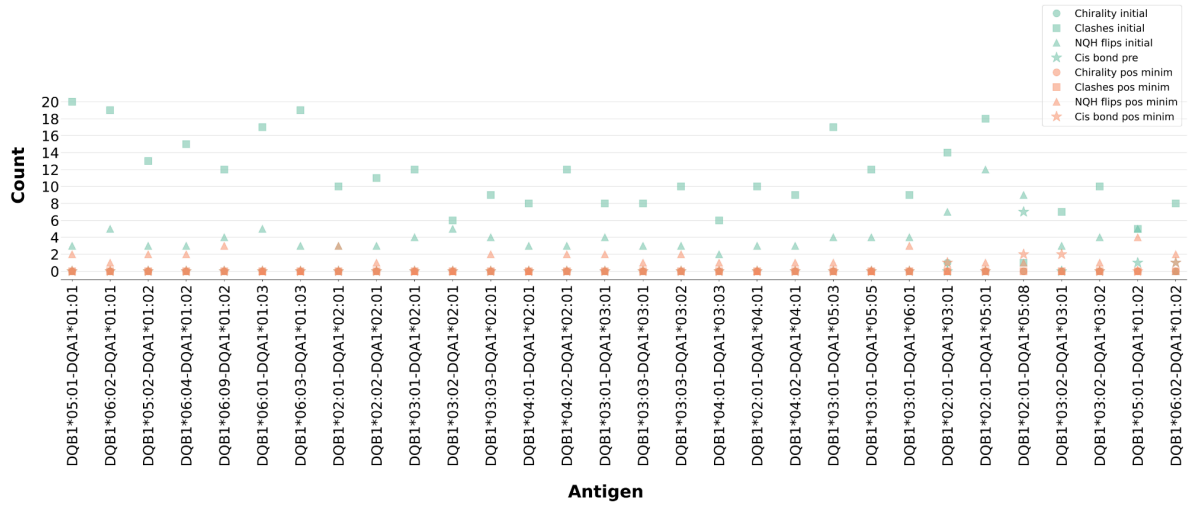
C)


Chirality initial
Clashes initial
NQH flips initial
Cis bond pre
Chirality pos minim
Clashes pos minim
NQH flips pos minim
Cis bond pos minim
Count
Antigen
C*01:02
C*02:02
C*03:02
C*03:03
C*05:01
C*08:01
C*08:02
C*12:03
C*14:02
C*15:02
C*16:01
C*17:01
C*18:02
C*03:04
C*04:01
C*06:02
C*07:02

D)


Chirality initial
Clashes initial
NQH flips initial
Cis bond pre
Chirality pos minim
Clashes pos minim
NQH flips pos minim
Cis bond pos minim
Count
Antigen
DPB1*01:01-DPA1*01:03
DPB1*02:01-DPA1*01:03
DPB1*03:01-DPA1*01:03
DPB1*04:01-DPA1*01:03
DPB1*04:02-DPA1*01:03
DPB1*06:01-DPA1*01:03
DPB1*11:01-DPA1*01:03
DPB1*19:01-DPA1*01:03
DPB1*23:01-DPA1*01:03
DPB1*28:01-DPA1*01:03
DPB1*18:01-DPA1*01:04
DPB1*03:01-DPA1*01:05
DPB1*18:01-DPA1*01:05
DPB1*28:01-DPA1*01:05
DPB1*01:01-DPA1*02:01
DPB1*03:01-DPA1*02:01
DPB1*05:01-DPA1*02:01
DPB1*06:01-DPA1*02:01
DPB1*09:01-DPA1*02:01
DPB1*13:01-DPA1*02:01
DPB1*14:01-DPA1*02:01
DPB1*15:01-DPA1*02:01
DPB1*17:01-DPA1*02:01
DPB1*18:01-DPA1*02:01
DPB1*05:01-DPA1*02:02
DPB1*10:01-DPA1*02:02
DPB1*11:01-DPA1*02:02
DPB1*13:01-DPA1*02:02
DPB1*13:01-DPA1*03:01
DPB1*20:01-DPA1*03:01
DPB1*28:01-DPA1*04:01

E)




F)

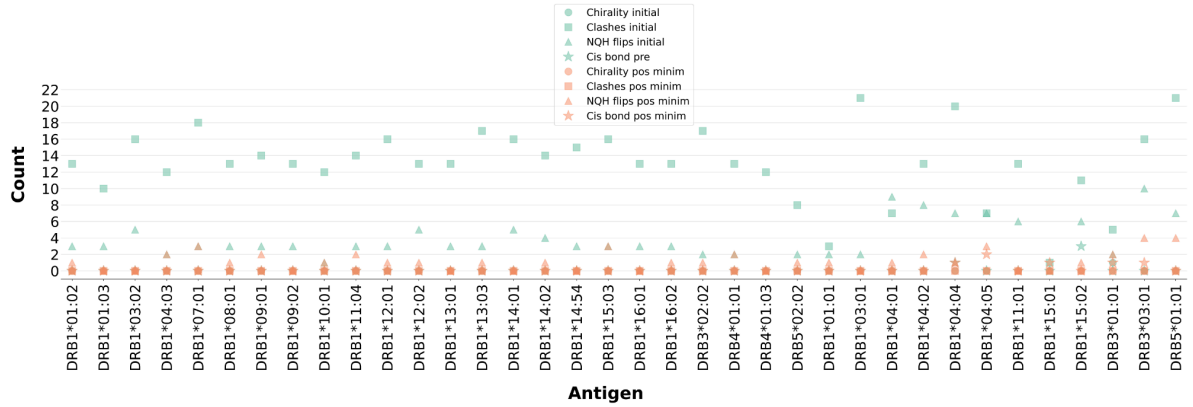
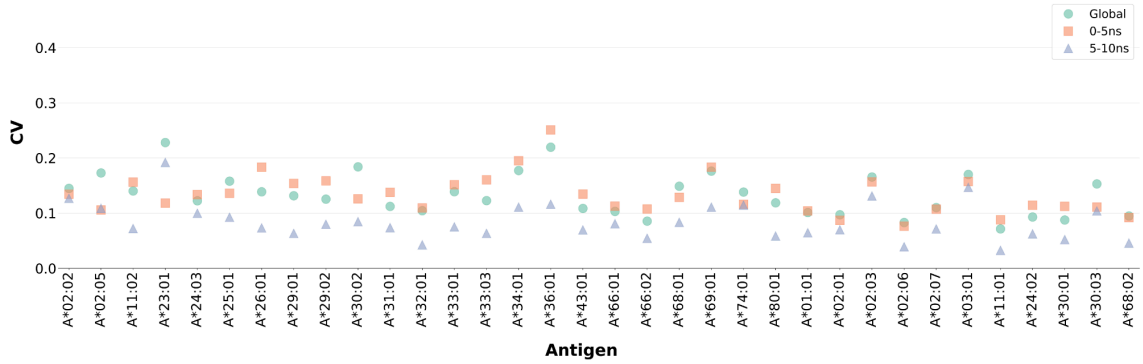
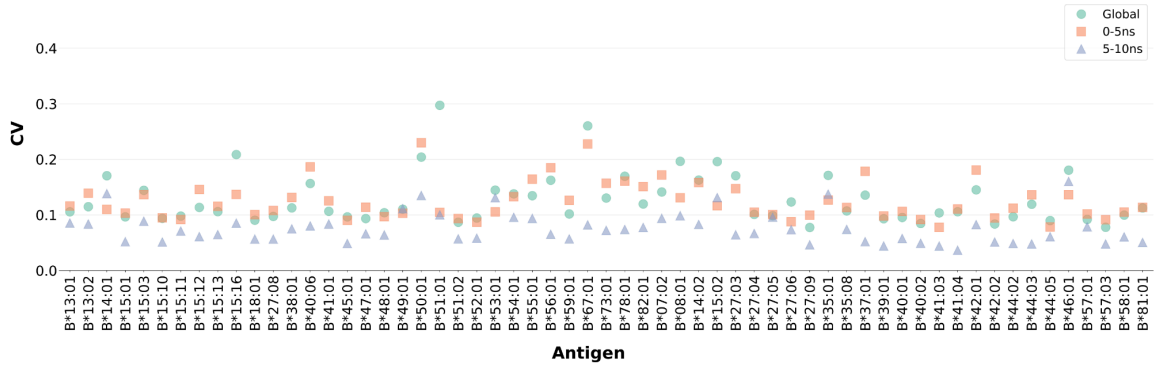



Figure S1: Stereochemical errors in HLA antigens. A) Antigens from locus A. B) Antigens from locus B. C) Antigens from locus C. D) Antigens from locus DP. E) Antigens from locus DQ and F) Antigens from locus DR.

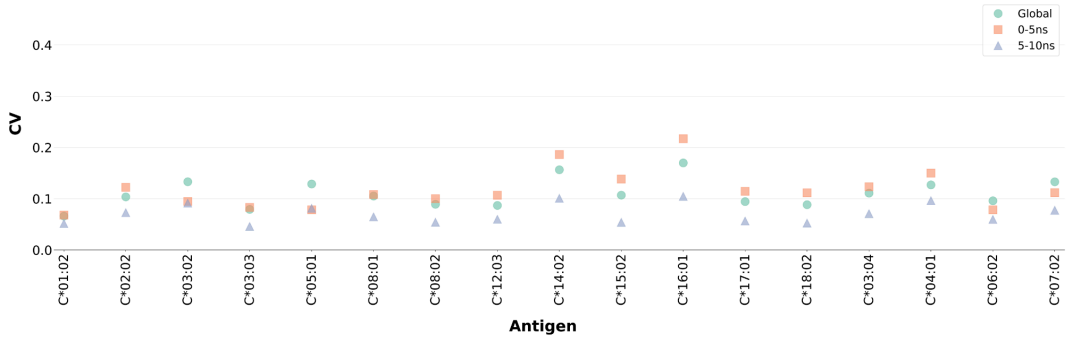
A)


Global
0-5ns
5-10ns
CV
0.0
0.1
0.2
0.3
0.4
A*02:02
A*02:05
A*11:02
A*23:01
A*24:03
A*25:01
A*26:01
A*29:01
A*29:02
A*30:02
A*31:01
A*32:01
A*33:01
A*33:03
A*34:01
A*36:01
A*43:01
A*66:01
A*66:02
A*68:01
A*69:01
A*74:01
A*80:01
A*01:01
A*02:01
A*02:03
A*02:06
A*02:07
A*03:01
A*11:01
A*24:02
A*30:01
A*30:03
A*68:02
Antigen

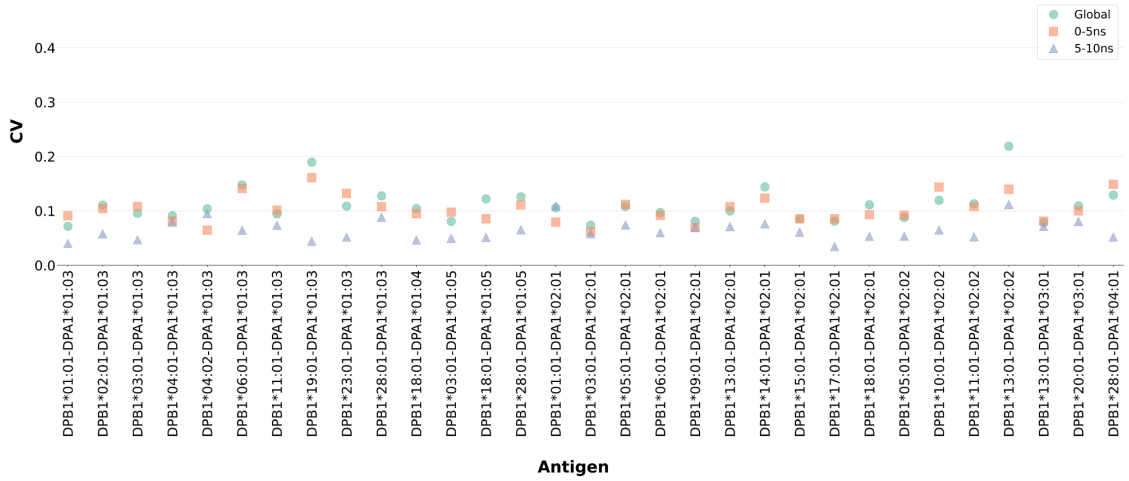
B)


Global
0-5ns
5-10ns
CV
0.0
0.1
0.2
0.3
0.4
B*13:01
B*13:02
B*14:01
B*15:01
B*15:03
B*15:10
B*15:11
B*15:12
B*15:13
B*15:16
B*18:01
B*27:08
B*38:01
B*40:06
B*41:01
B*45:01
B*47:01
B*48:01
B*49:01
B*50:01
B*51:01
B*51:02
B*52:01
B*53:01
B*54:01
B*55:01
B*56:01
B*59:01
B*67:01
B*73:01
B*78:01
B*82:01
B*07:02
B*08:01
B*14:02
B*15:02
B*27:03
B*27:04
B*27:05
B*27:06
B*27:09
B*35:01
B*35:08
B*37:01
B*39:01
B*40:01
B*40:02
B*41:03
B*41:04
B*42:01
B*42:02
B*44:02
B*44:03
B*44:05
B*46:01
B*57:01
B*57:03
B*58:01
B*81:01
Antigen

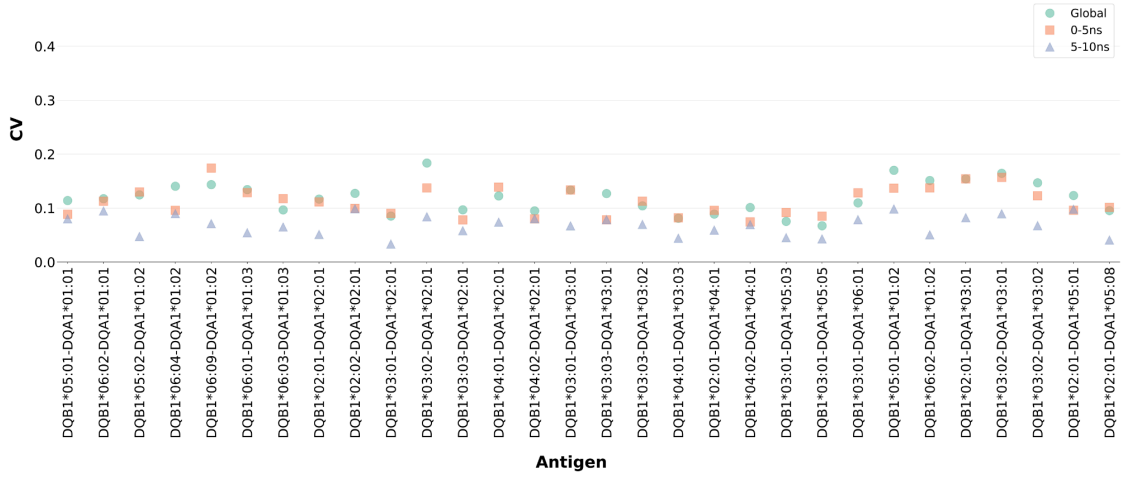
C)


Global
0-5ns
5-10ns
CV
0.0
0.1
0.2
0.3
0.4
C*01:02
C*02:02
C*03:02
C*03:03
C*05:01
C*08:01
C*08:02
C*12:03
C*14:02
C*15:02
C*16:01
C*17:01
C*18:02
C*03:04
C*04:01
C*06:02
C*07:02
Antigen

D)


Global
0-5ns
5-10ns
CV
0.0
0.1
0.2
0.3
0.4
DPB1*01:01-DPA1*01:03
DPB1*02:01-DPA1*01:03
DPB1*03:01-DPA1*01:03
DPB1*04:01-DPA1*01:03
DPB1*04:02-DPA1*01:03
DPB1*06:01-DPA1*01:03
DPB1*11:01-DPA1*01:03
DPB1*19:01-DPA1*01:03
DPB1*23:01-DPA1*01:03
DPB1*28:01-DPA1*01:03
DPB1*18:01-DPA1*01:04
DPB1*03:01-DPA1*01:05
DPB1*18:01-DPA1*01:05
DPB1*28:01-DPA1*01:05
DPB1*01:01-DPA1*02:01
DPB1*03:01-DPA1*02:01
DPB1*05:01-DPA1*02:01
DPB1*06:01-DPA1*02:01
DPB1*09:01-DPA1*02:01
DPB1*13:01-DPA1*02:01
DPB1*14:01-DPA1*02:01
DPB1*15:01-DPA1*02:01
DPB1*17:01-DPA1*02:01
DPB1*18:01-DPA1*02:01
DPB1*05:01-DPA1*02:02
DPB1*10:01-DPA1*02:02
DPB1*11:01-DPA1*02:02
DPB1*13:01-DPA1*02:02
DPB1*13:01-DPA1*03:01
DPB1*20:01-DPA1*03:01
DPB1*28:01-DPA1*04:01
Antigen

E)




F)

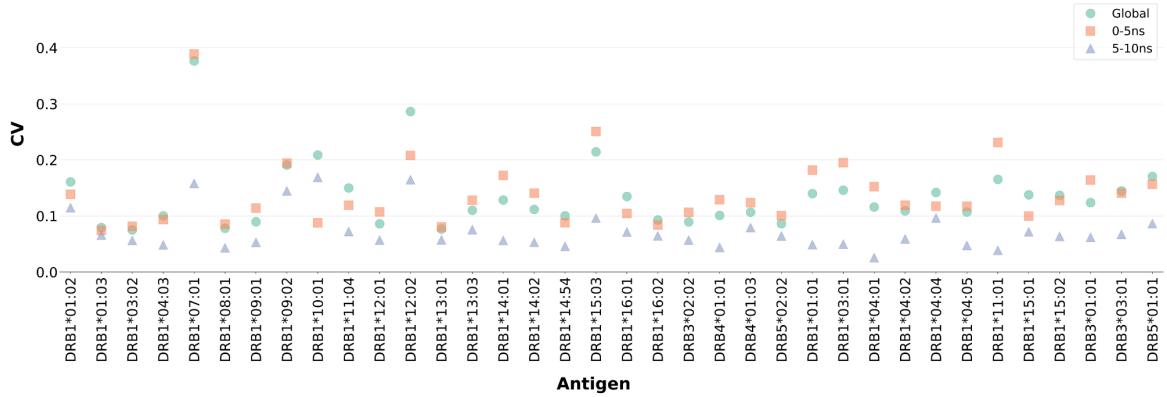
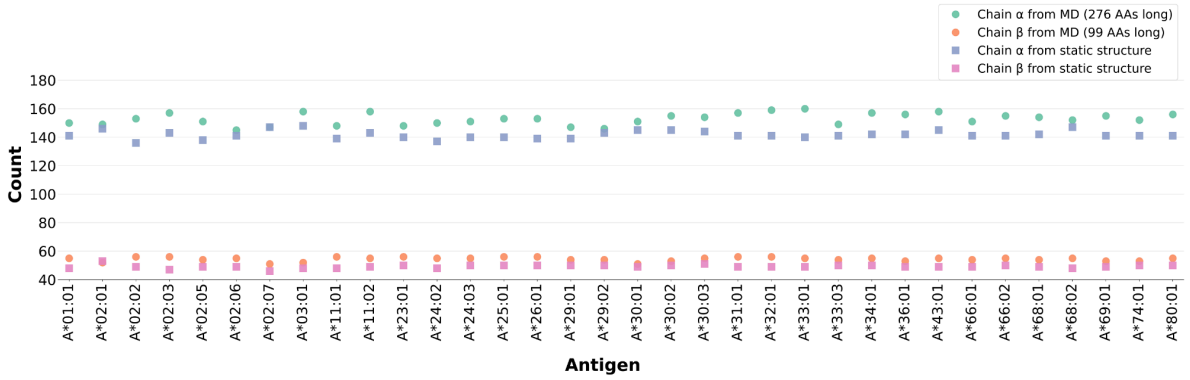
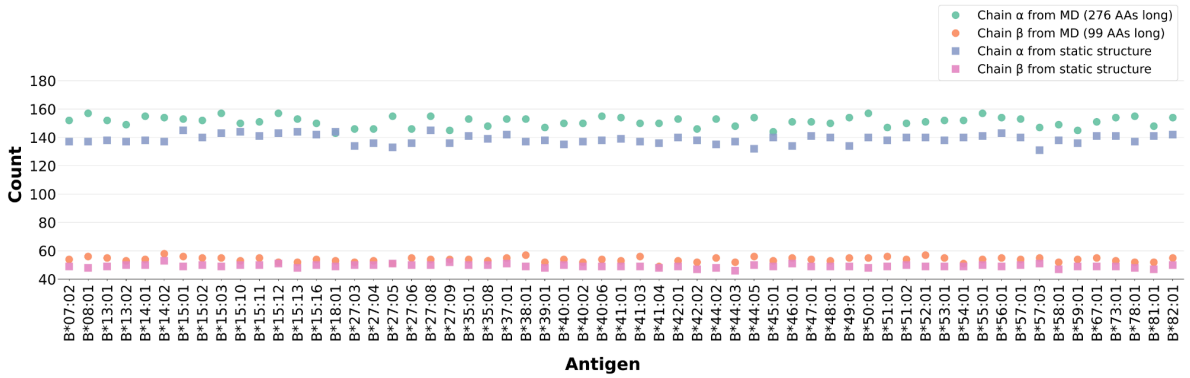



Figure S2: CV values per HLA antigen for the 3 simulation intervals considered. A) Antigens from locus A. B) Antigens from locus B. C) Antigens from locus C. D) Antigens from locus DP. E) Antigens from locus DQ and F) Antigens from locus DR.

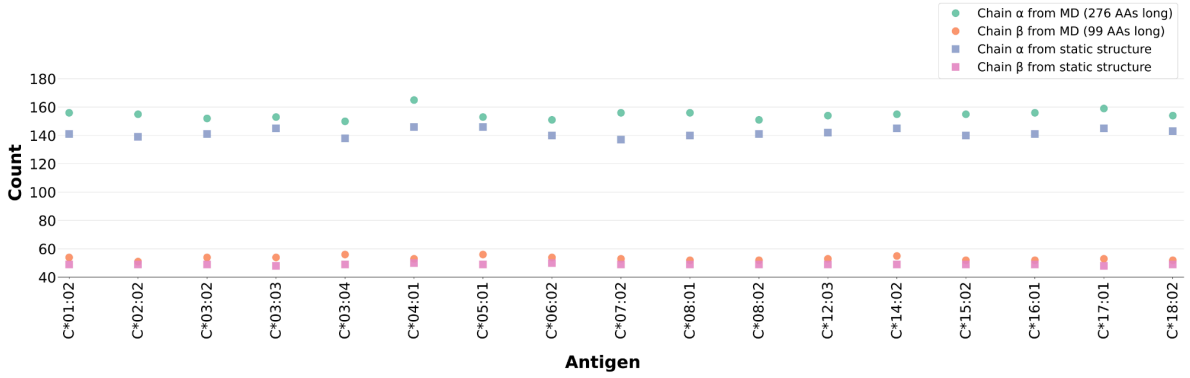
A)


Chain α from MD (276 AAs long)
Chain β from MD (99 AAs long)
Chain α from static structure
Chain β from static structure
Count
Antigen
A*01:01
A*02:01
A*02:02
A*02:03
A*02:05
A*02:06
A*02:07
A*03:01
A*11:01
A*11:02
A*23:01
A*24:02
A*24:03
A*25:01
A*26:01
A*29:01
A*29:02
A*30:01
A*30:02
A*30:03
A*31:01
A*32:01
A*33:01
A*33:03
A*34:01
A*36:01
A*43:01
A*66:01
A*66:02
A*68:01
A*68:02
A*69:01
A*74:01
A*80:01

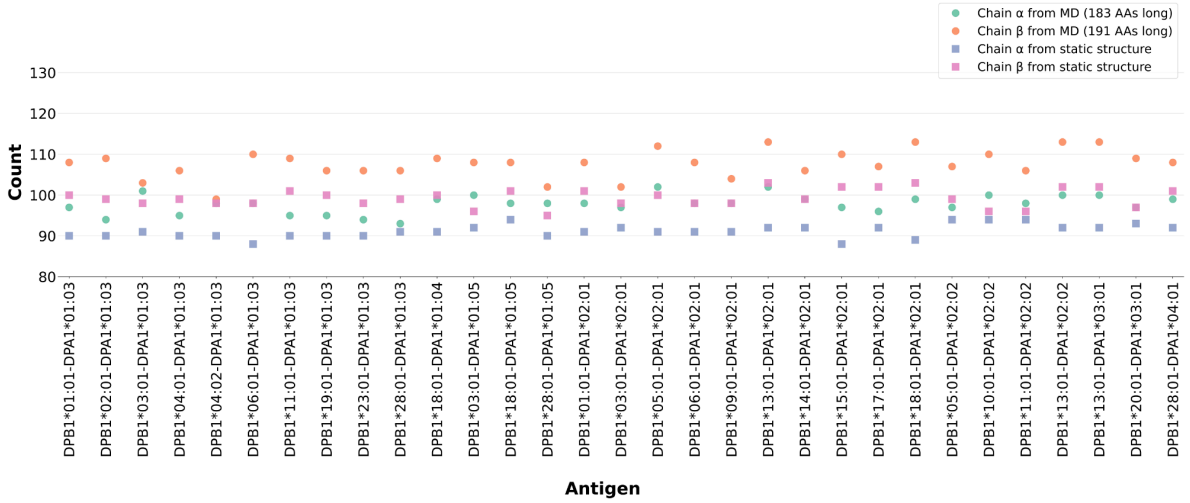
B)


Chain α from MD (276 AAs long)
Chain β from MD (99 AAs long)
Chain α from static structure
Chain β from static structure
Count
Antigen
B*07:02
B*08:01
B*13:01
B*13:02
B*14:01
B*14:02
B*15:01
B*15:02
B*15:03
B*15:10
B*15:11
B*15:12
B*15:13
B*15:16
B*18:01
B*27:03
B*27:04
B*27:05
B*27:06
B*27:08
B*27:09
B*35:01
B*35:08
B*37:01
B*38:01
B*39:01
B*40:01
B*40:02
B*40:06
B*41:01
B*41:03
B*41:04
B*42:01
B*42:02
B*44:02
B*44:03
B*44:05
B*45:01
B*46:01
B*47:01
B*48:01
B*49:01
B*50:01
B*51:01
B*51:02
B*52:01
B*53:01
B*54:01
B*55:01
B*56:01
B*57:01
B*57:03
B*58:01
B*59:01
B*67:01
B*73:01
B*78:01
B*81:01
B*82:01

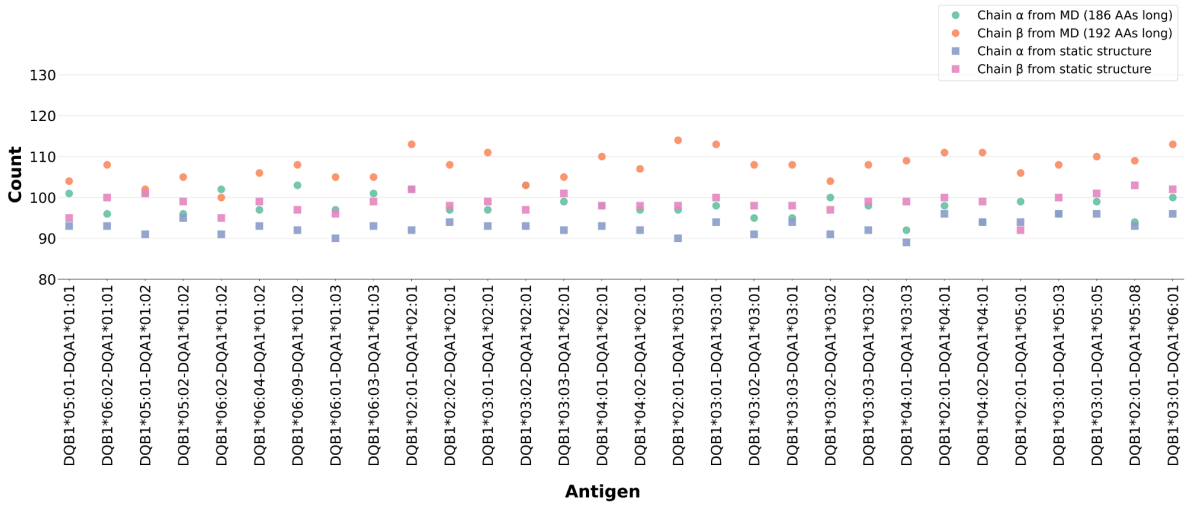
C)


Chain α from MD (276 AAs long)
Chain β from MD (99 AAs long)
Chain α from static structure
Chain β from static structure
Count
Antigen
C*01:02
C*02:02
C*03:02
C*03:03
C*03:04
C*04:01
C*05:01
C*06:02
C*07:02
C*08:01
C*08:02
C*12:03
C*14:02
C*15:02
C*16:01
C*17:01
C*18:02

D)




E)




F)

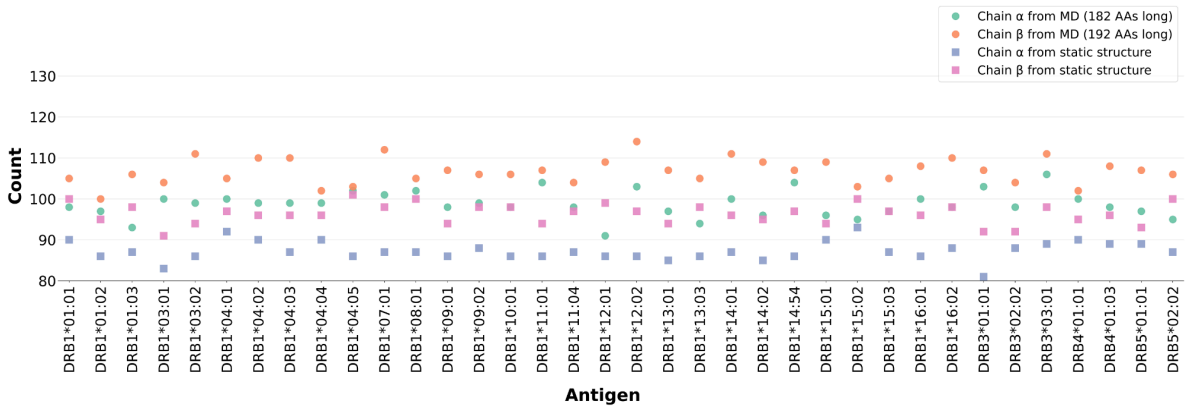
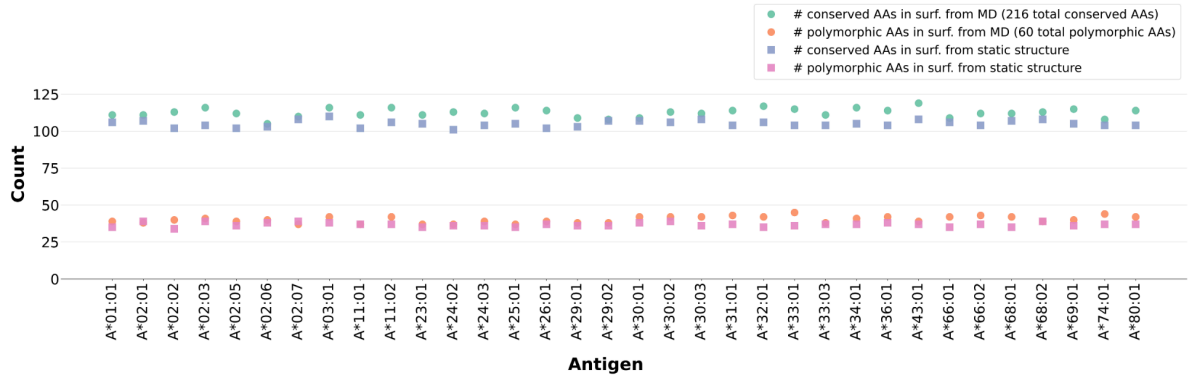
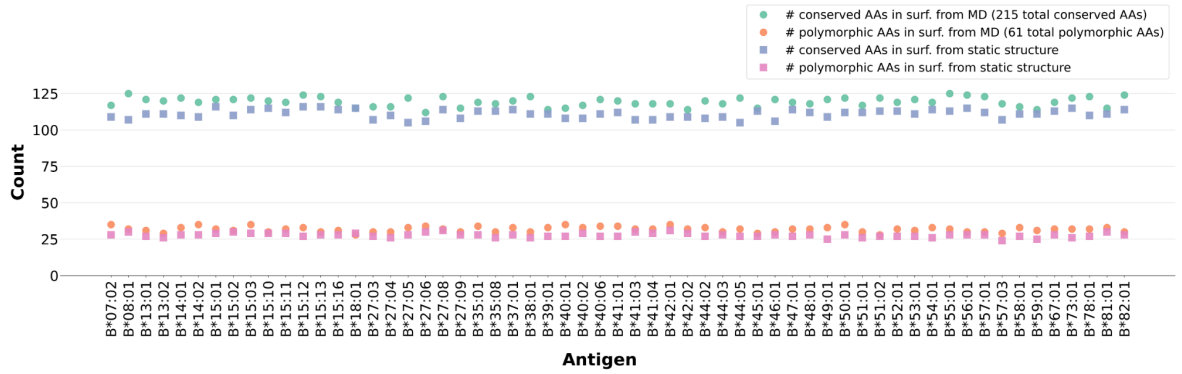



Figure S3: Number of surface AAs (with median RSASA greater than 20%) for each chain of each HLA antigen. Counting was made for MD and static PDB data (structures issued from the refinement process). Results are plotted for antigens from locus A in panel A, from locus B in panel B, locus C in panel C, locus DP in panel D, locus DQ in panel E and locus DR in panel F.

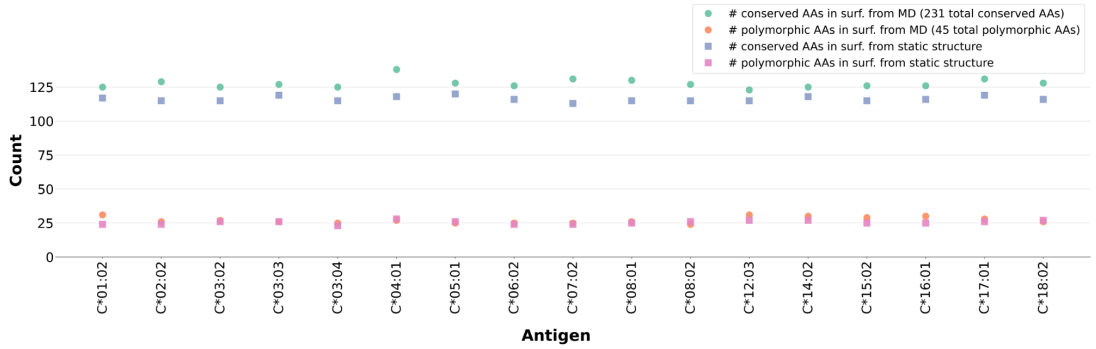
A)


# conserved AAs in surf. from MD (216 total conserved AAs)
# polymorphic AAs in surf. from MD (60 total polymorphic AAs)
# conserved AAs in surf. from static structure
# polymorphic AAs in surf. from static structure
Count
Antigen

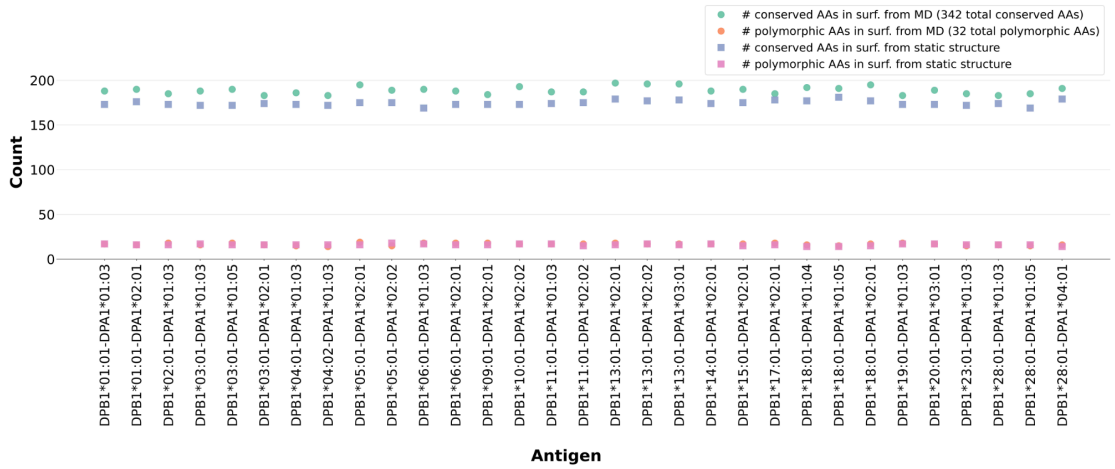
B)


# conserved AAs in surf. from MD (215 total conserved AAs)
# polymorphic AAs in surf. from MD (61 total polymorphic AAs)
# conserved AAs in surf. from static structure
# polymorphic AAs in surf. from static structure
Count
Antigen

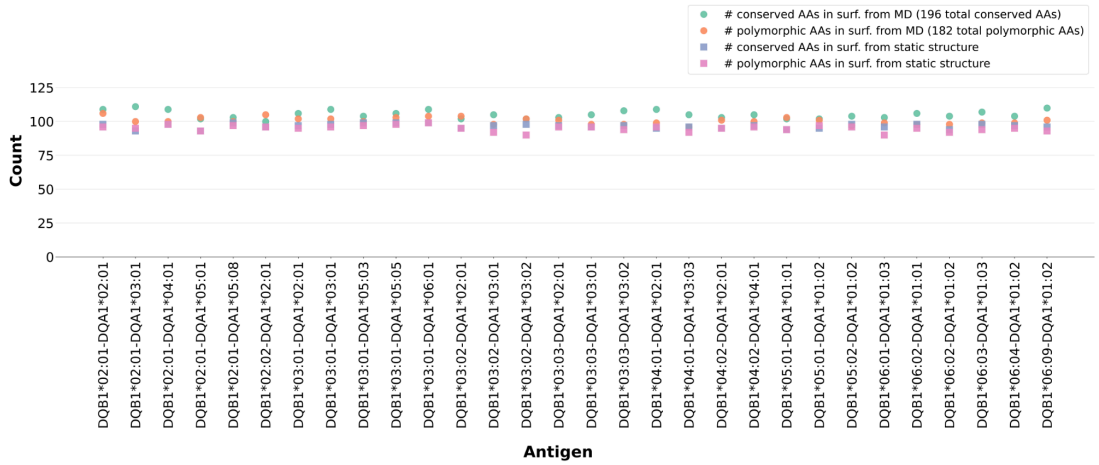
C)


# conserved AAs in surf. from MD (231 total conserved AAs)
# polymorphic AAs in surf. from MD (45 total polymorphic AAs)
# conserved AAs in surf. from static structure
# polymorphic AAs in surf. from static structure
Count
Antigen

D)


# conserved AAs in surf. from MD (342 total conserved AAs)
# polymorphic AAs in surf. from MD (32 total polymorphic AAs)
# conserved AAs in surf. from static structure
# polymorphic AAs in surf. from static structure
Count
Antigen

E)




F)

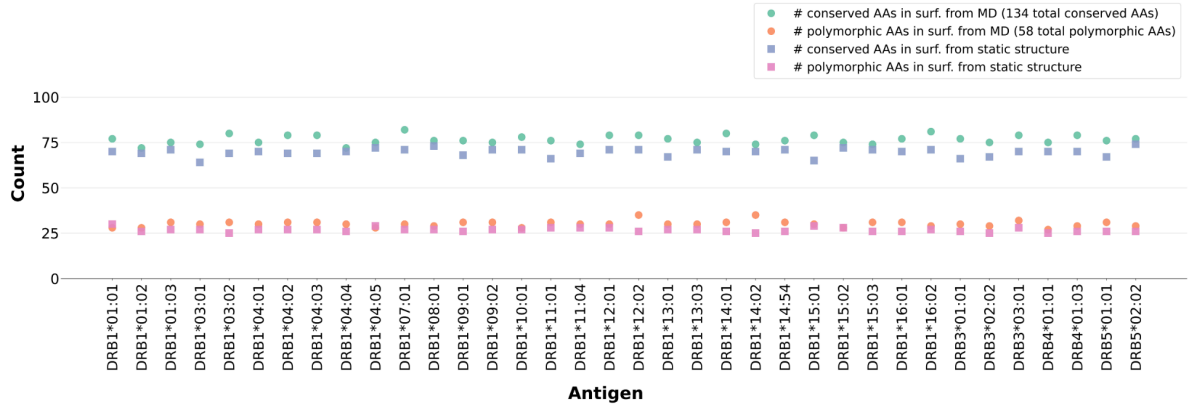
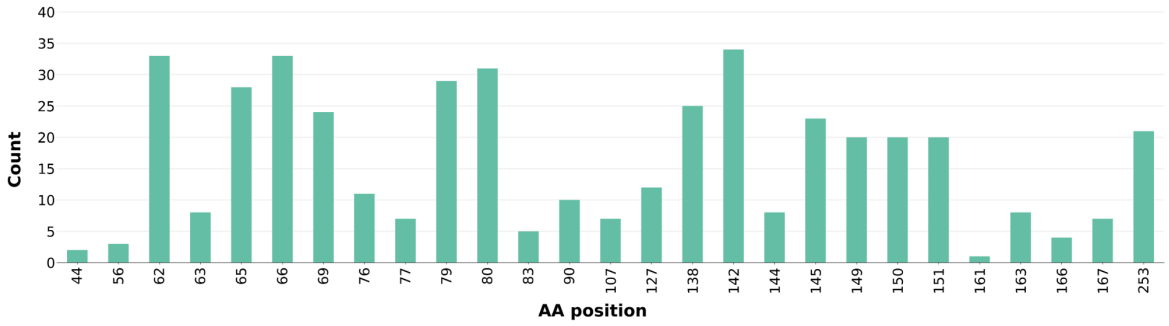
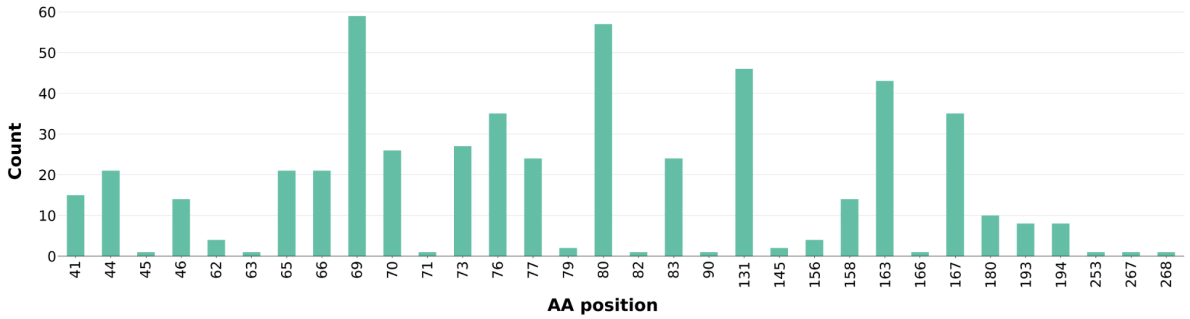



Figure S4: Conserved and polymorphic AA positions at surface in HLA antigen. Counting was made for MD and static PDB data (structures issued from the refinement process). A) Antigens from locus A (β2-microglobulin not included). B) Antigens from locus B (β2-microglobulin not included). C) Antigens from locus C (β2-microglobulin not included). D) Antigens from locus DP. E) Antigens from locus DQ. F) Antigens from locus DR (chain encoded from gene DRA1 not included).

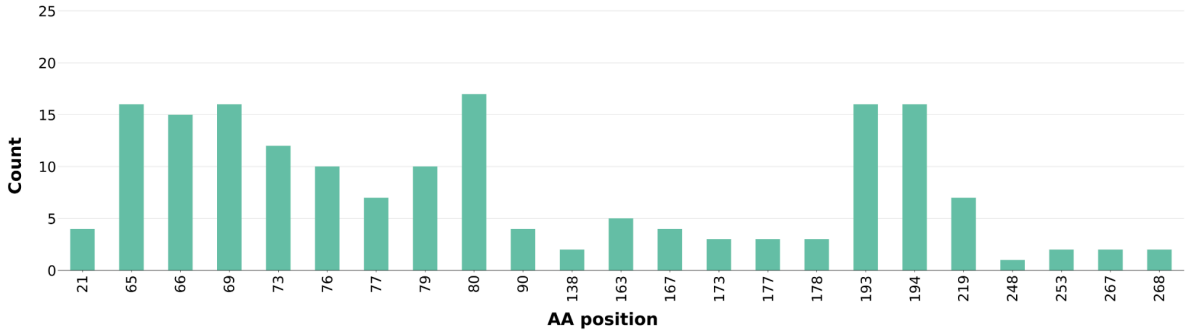
A)


Count
AA position
0 5 10 15 20 25 30 35 40
44 56 62 63 65 66 69 76 77 79 80 83 90 107 127 138 142 144 145 149 150 151 161 163 166 167 253

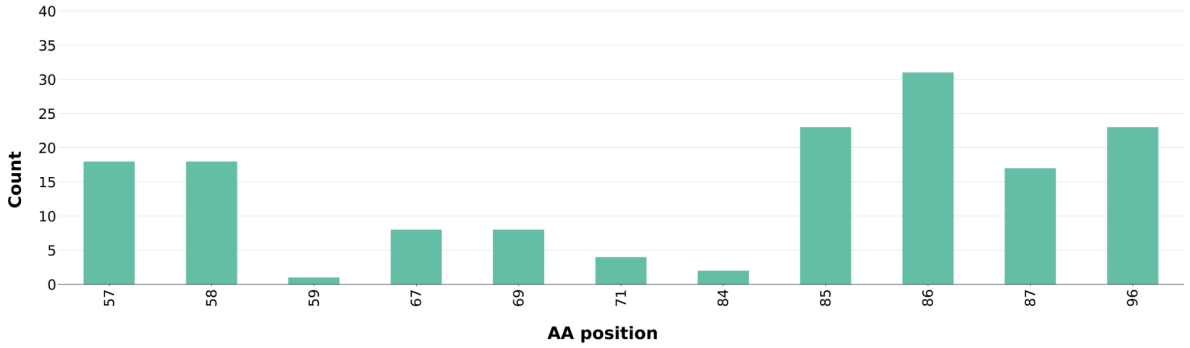
B)


Count
AA position
0 10 20 30 40 50 60
41 44 45 46 62 63 65 66 69 70 71 73 76 77 79 80 82 83 90 131 145 156 158 163 166 167 180 193 194 253 267 268

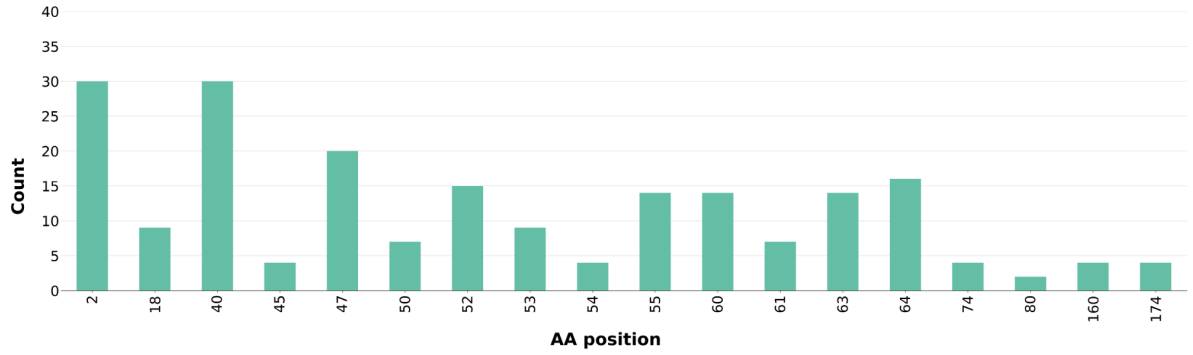
C)


Count
AA position
0 5 10 15 20 25
21 65 66 69 73 76 77 79 80 90 138 163 167 173 177 178 193 194 219 248 253 267 268

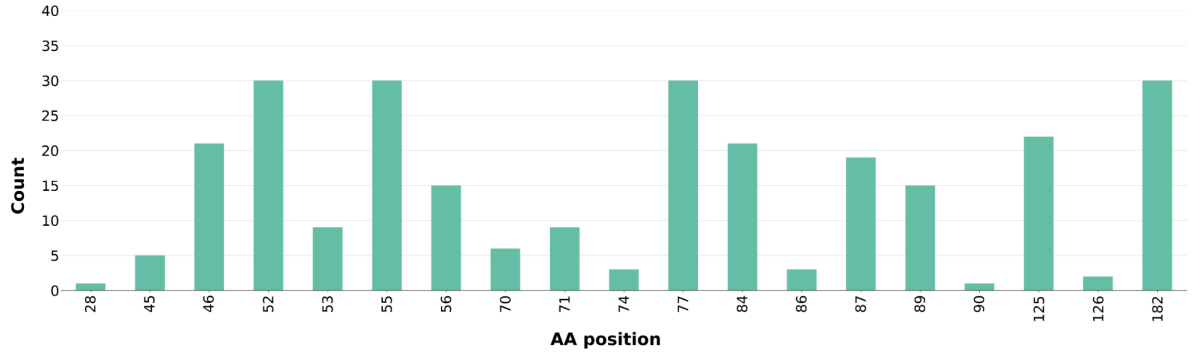
D)


Count
AA position
0 5 10 15 20 25 30 35 40
57 58 59 67 69 71 84 85 86 87 96

E)




F)




G)

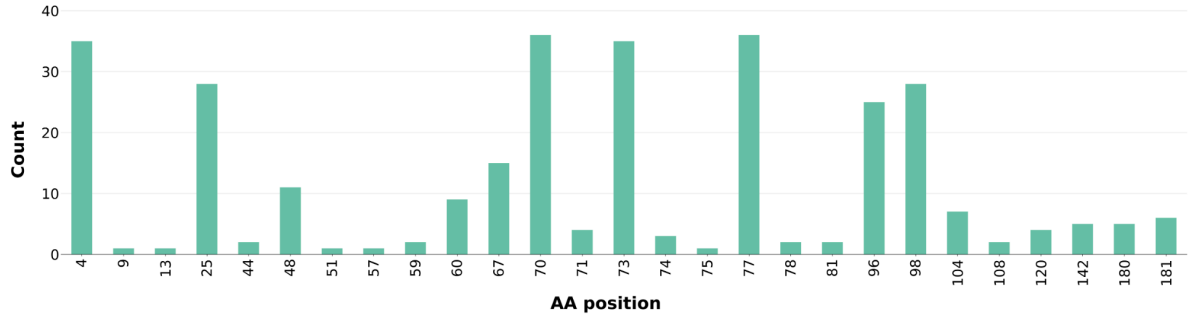



Figure S5: Number of occurrences of solvent accessible AA positions in confirmed eplets. Results are plotted for an antigens corresponding to A genes in panel A, B genes in panel B, C genes in panel C, DPB1 gene in panel D, DQA1 gene in panel E, DQB1 gene in panel F and DRB1/3/4/5 genes in panel G.

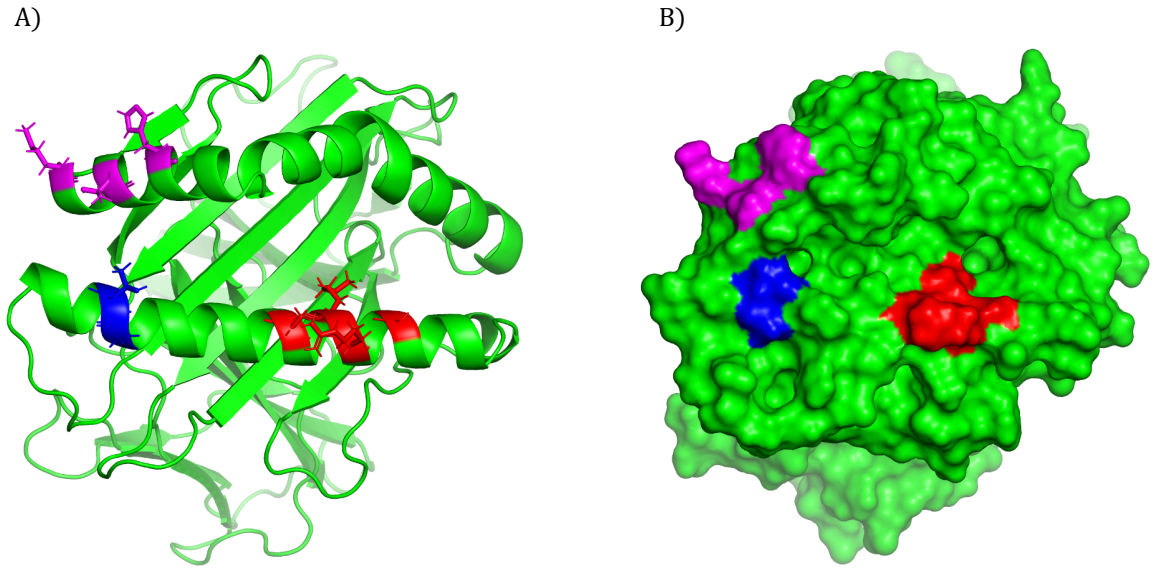



Figure S6: Visualization of most frequent solvent accessible AA positions within confirmed eplets across HLA antigens from locus A. Structure of antigen A*02:01 is used. AA Positions 62, 65, 66 and 69 are depicted in red, AA positions 79 and 80 in blue, and AA positions 138, 142 and 145 in magenta. A) Backbone representation for the molecule and licorice representation for side chains of concerned AA positions. B) Surface representation of the molecule with coloring for concerned AA positions.

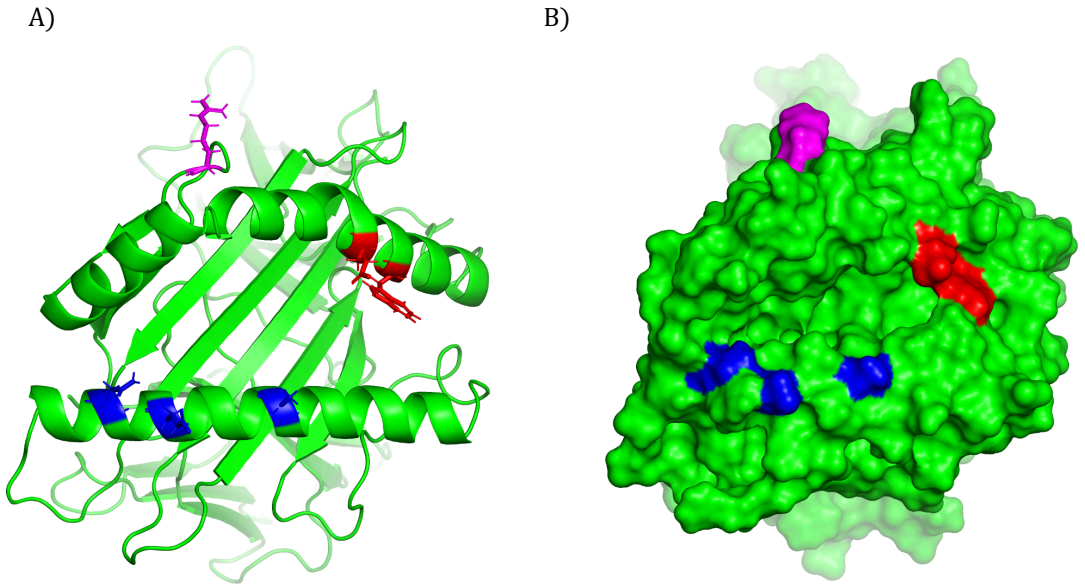



Figure S7: Visualization of most frequent solvent accessible AA positions within confirmed eplets across HLA antigens from locus B. Structure of antigen B*07:02 is used. AA positions 69, 76 and 80 are depicted in blue, AA position 131 in magenta, and AA positions 163 and 167 in red. A) Backbone representation for the molecule and licorice representation for side chains of concerned AA positions. B) Surface representation of the molecule with coloring for concerned AA positions.

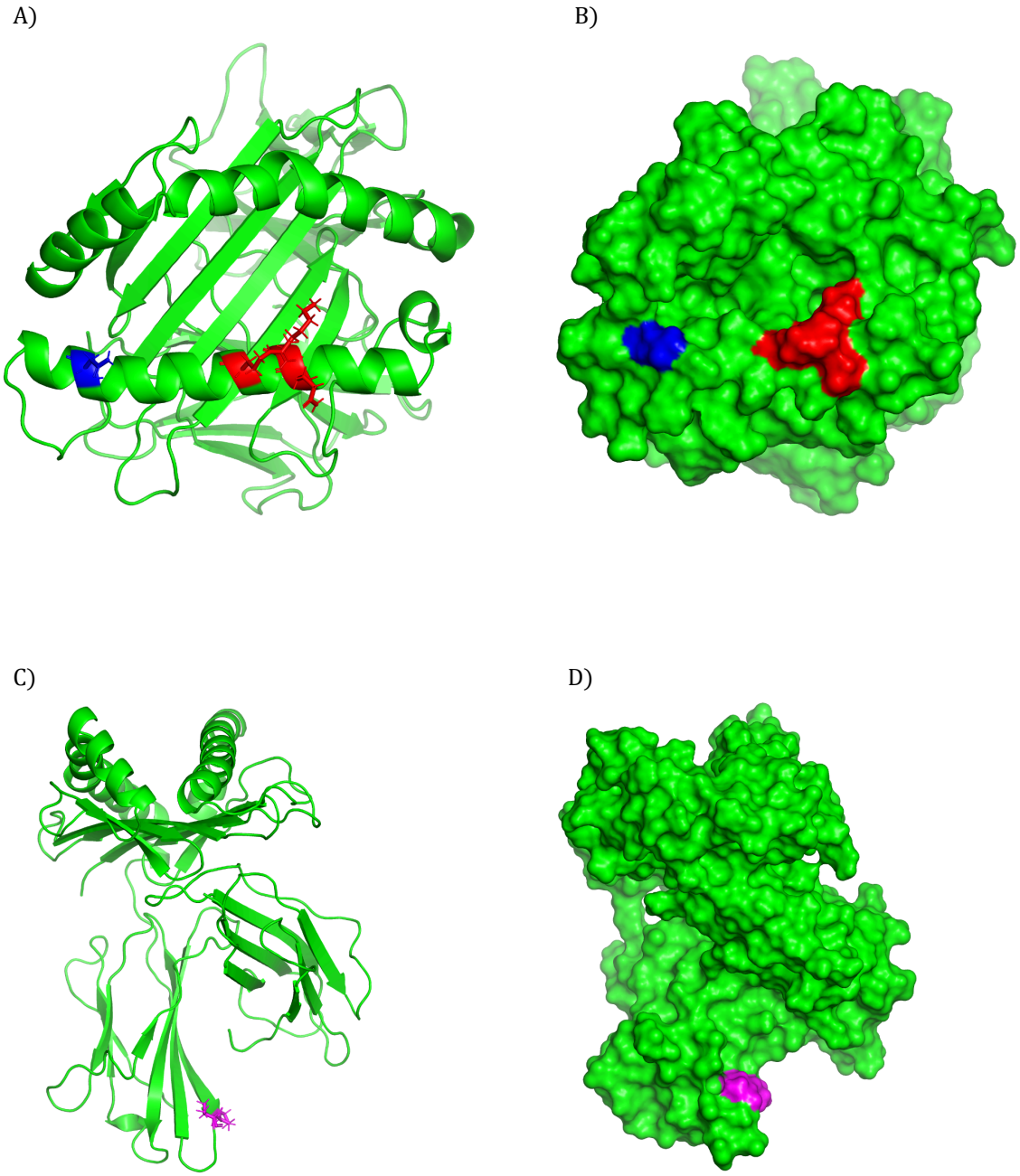



Figure S8: Visualization of most frequent solvent accessible AA positions within confirmed eplets across gene C. Structure of antigen C*01:02 is used. AA positions 65, 66 and 69 are depicted in red, AA position 80 in blue, and AA positions 193 and 194 in magenta. A) Backbone representation for the molecule and licorice representation for side chains of AA positions 65, 66, 69 and 80. B) Surface representation of the molecule with coloring for AA positions 65, 66, 69 and 80. C) Backbone representation for the molecule and licorice representation for side chains of AA positions 193 and 194. D) Surface representation of the molecule with coloring for AA positions 193 and 194.

A) B)

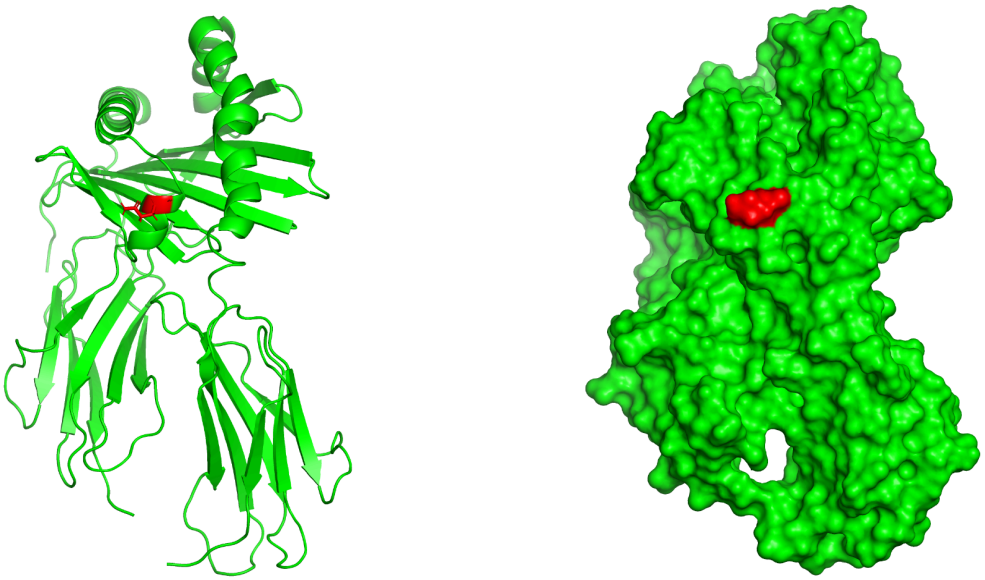


Figure S9: Visualization of the unique solvent accessible AA position within confirmed eplets present in all antigens corresponding to gene DPA1. Structure of antigen DPA1*01:03-DPB1*01:01 is used. AA position 50 is depicted in red. A) Backbone representation for the molecule and licorice representation for side chain of concerned AA position. B) Surface representation of the molecule with coloring for concerned AA position.

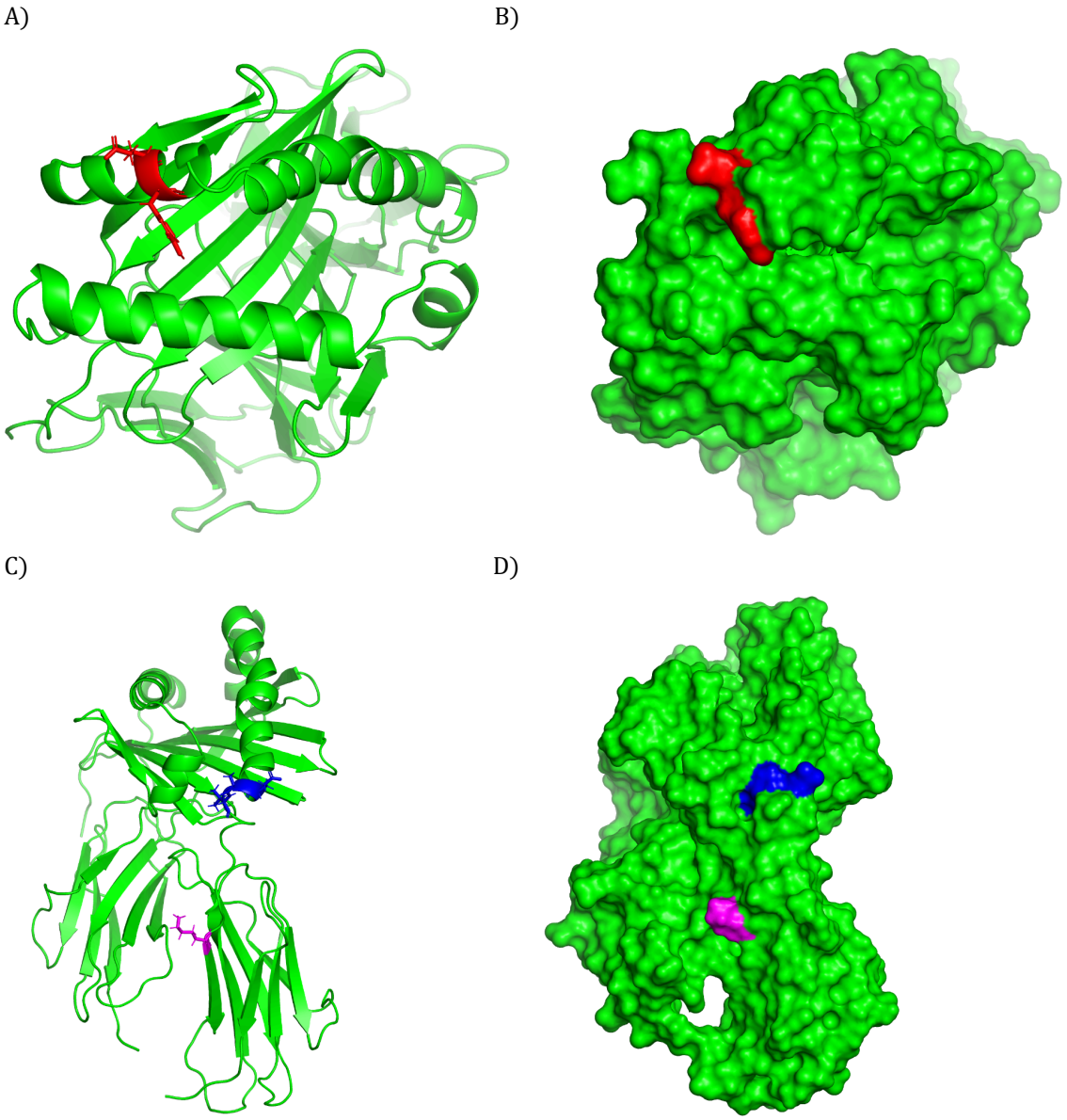



Figure S10: Visualization of most frequent solvent accessible AA positions within confirmed eplets across antigens corresponding to gene DPB1. Structure of antigen DPA1*01:03-DPB1*01:01 is used. AA positions 57 and 58 are depicted in red, AA positions 85, 86 and 87 in blue, and AA position 96 in magenta. A) Backbone representation for the molecule and licorice representation for side chains of AA positions 57 and 58. B) Surface representation of the molecule with coloring for AA positions 57 and 58. C) Backbone representation for the molecule and licorice representation for side chains of AA positions 85, 86, 87 and 96. D) Surface representation of the molecule with coloring for AA positions 85, 86, 87 and 96.

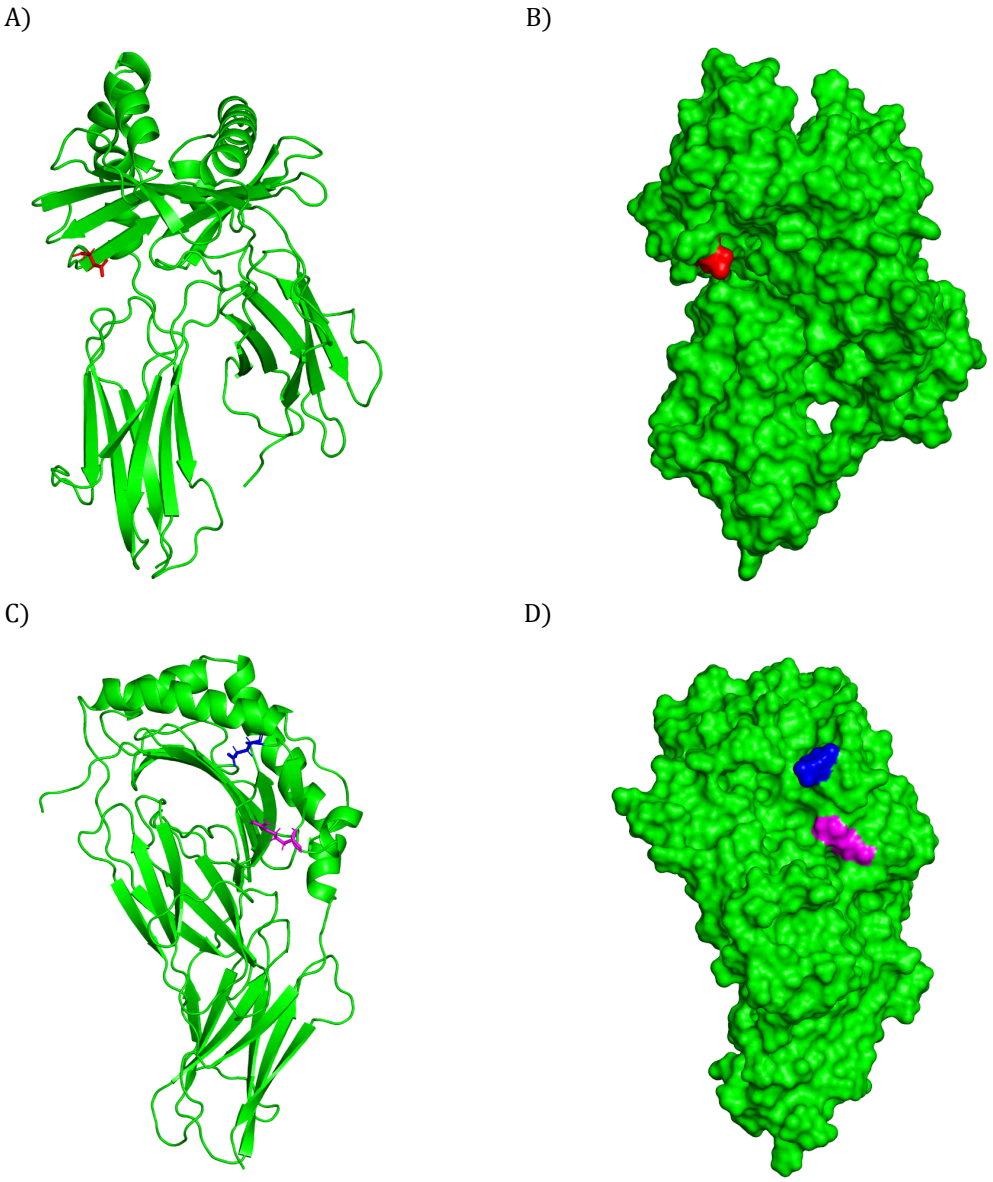



Figure S11: Visualization of most frequent solvent accessible AA positions within confirmed eplets across antigens corresponding to gene DQA1. Structure of antigen DQA1*01:01-DQB1*05:01 is used. AA position 2 is depicted in red, AA position 40 in blue, and AA position 47 in magenta. A) Backbone representation for the molecule and licorice representation for side chains of AA position 2. B) Surface representation of the molecule with coloring for AA position 2. C) Backbone representation for the molecule and licorice representation for side chains of AA positions 40 and 47. D) Surface representation of the molecule with coloring for AA positions 40 and 47.

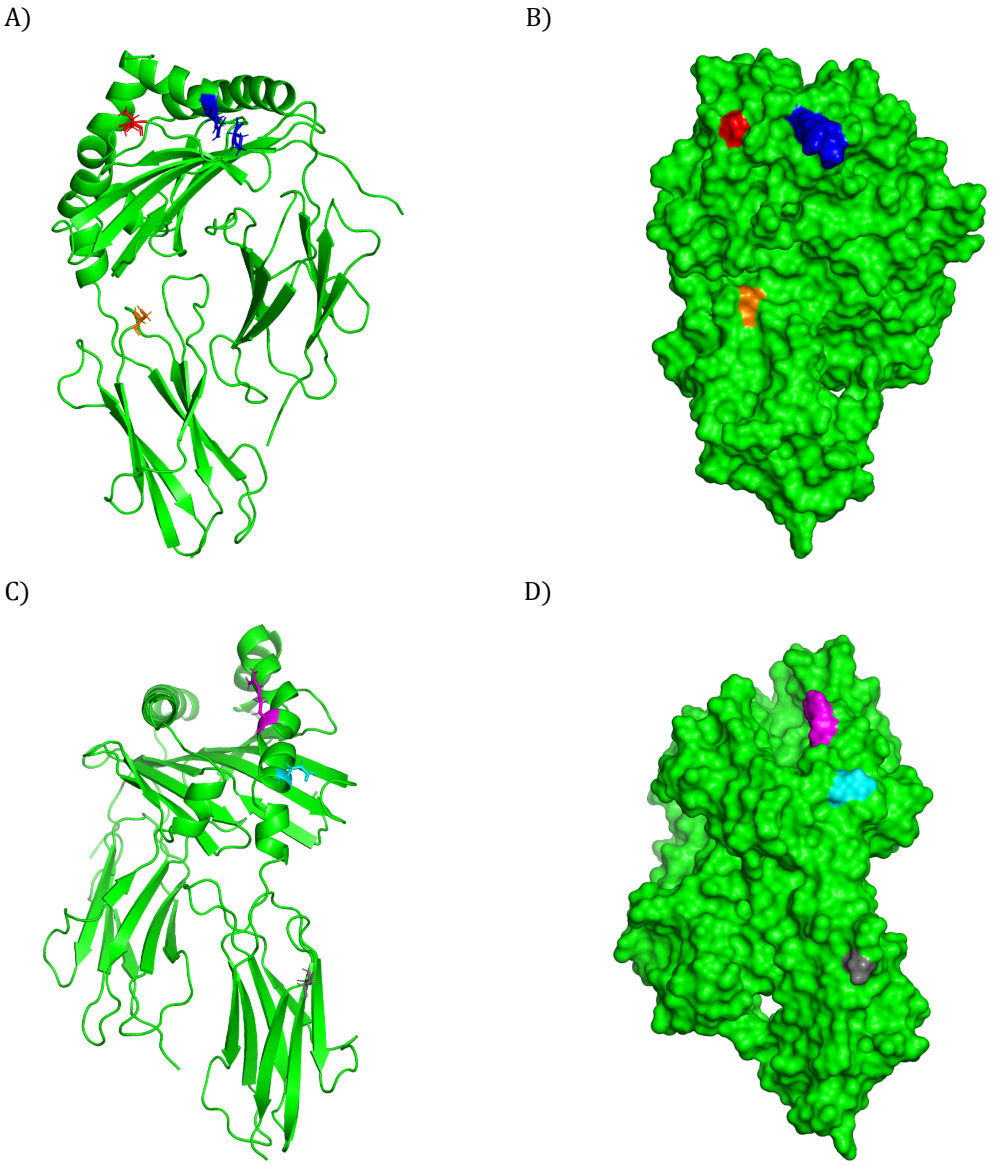



Figure S12: Visualization of most frequent solvent accessible AA positions within confirmed eplets across antigens corresponding to gene DQB1. Structure of the antigen DQA1*01:01-DQB1*05:01 is used. AA position 46 is depicted in red, AA positions 52 and 55 in blue, AA position 125 in orange, AA position 77 in magenta, AA position 84 in cyan and AA position 182 in gray. A) Backbone representation for the molecule and licorice representation for side chains of AA positions 46, 52, 55 and 125. B) Surface representation of the molecule with coloring for AA position 46, 52, 55 and 125. C) Backbone representation for the molecule and licorice representation for side chains of AA positions 77, 84 and 182. D) Surface representation of the molecule with coloring for AA positions 77, 84 and 182.

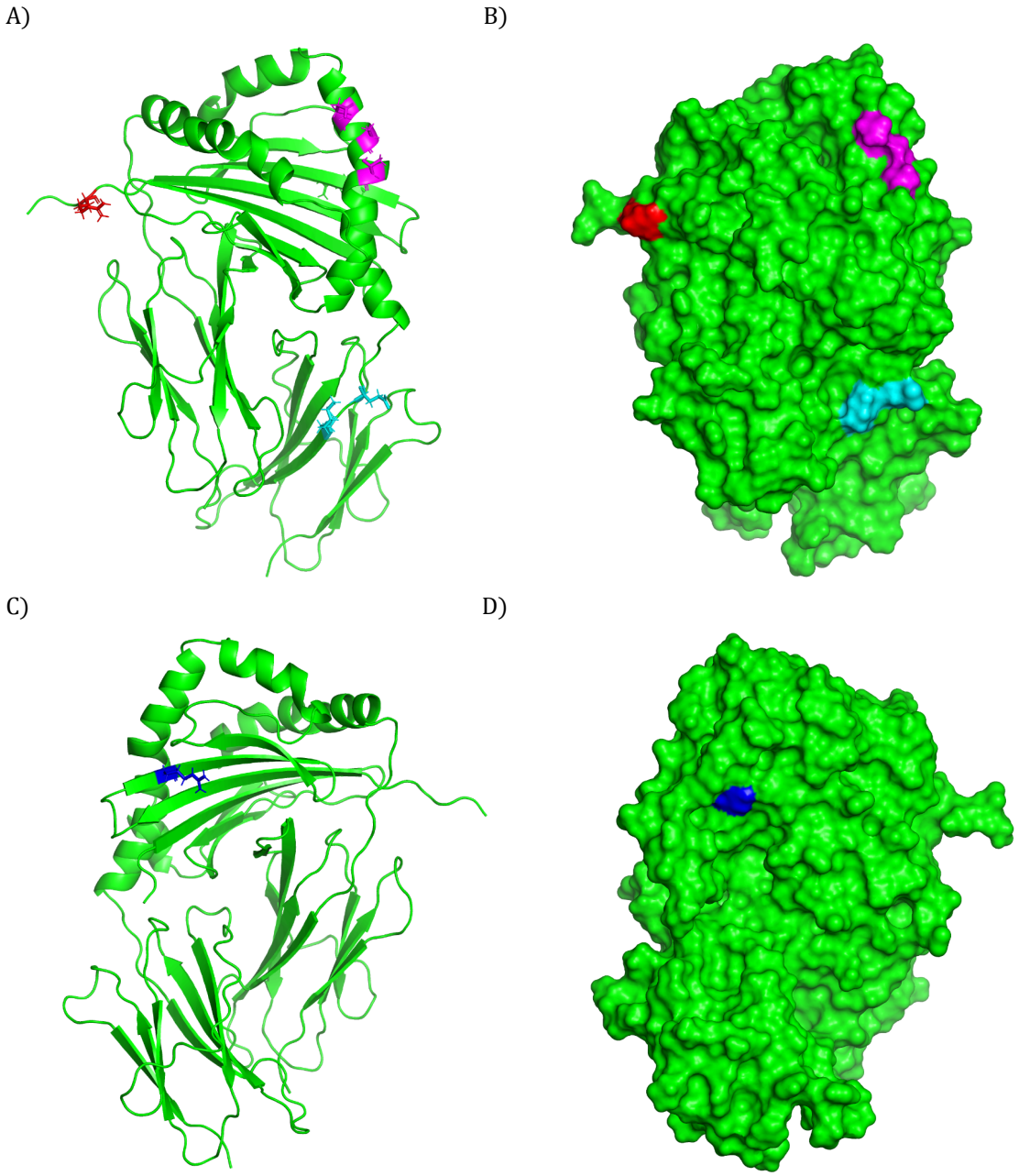
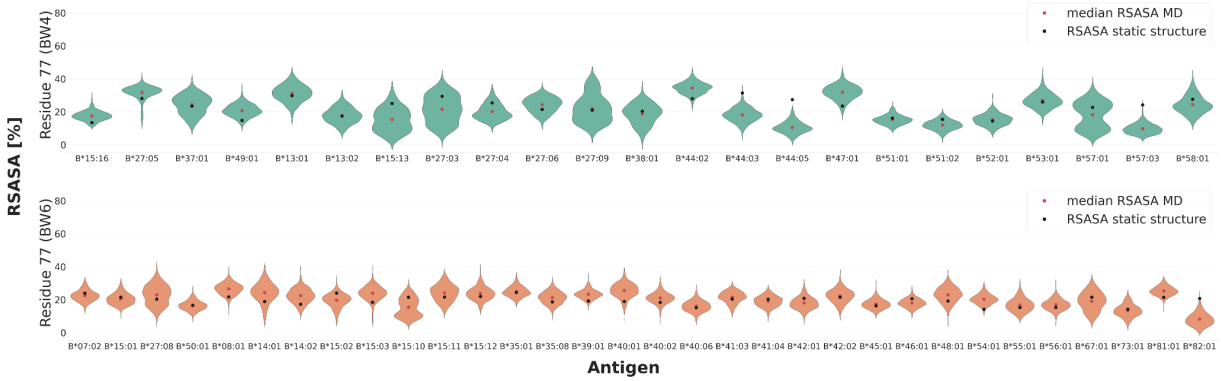
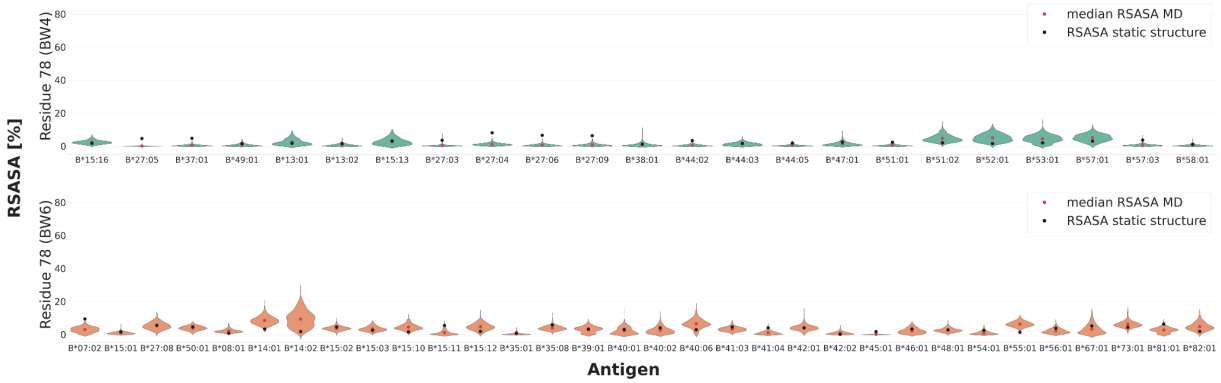



Figure S13: Visualization of the solvent accessible AA positions with the most occurrences within confirmed eplets in gene DRB1/3/4/5. Structure of the antigen DRB1*01:01 is used. AA position 4 is depicted in red, AA positions 70, 73 and 77 in magenta, AA positions 96 and 98 in cyan and AA position 25 in blue. A) Backbone representation for the molecule and licorice representation for side chains of AA positions 4, 70, 73, 77, 96 and 98. B) Surface representation of the molecule with coloring for AA position 4, 70, 73, 77, 96 and 98. C) Backbone representation for the molecule and licorice representation for side chains of AA position 25. D) Surface representation of the molecule with coloring for AA position 25.

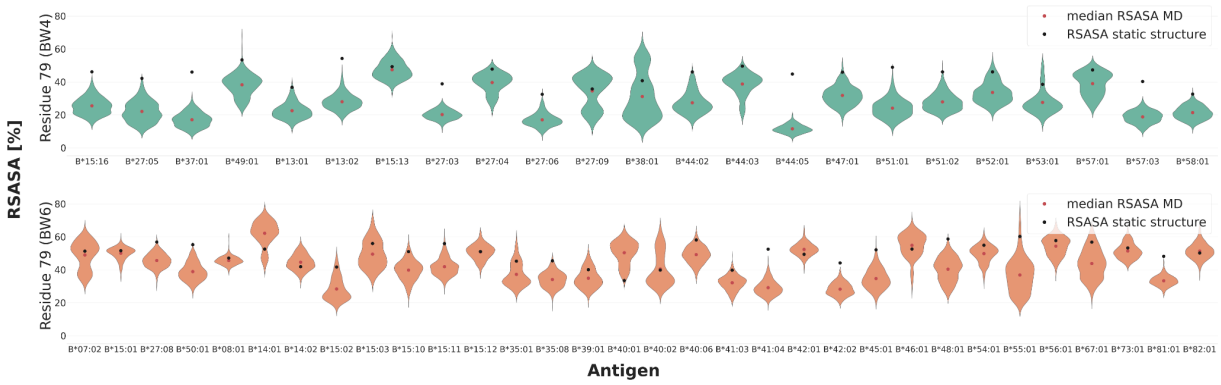
A)


RSASA [%]
Residue 77 (BW4)
Residue 77 (BW6)
median RSASA MD
RSASA static structure
B*15:16 B*27:05 B*37:01 B*49:01 B*13:01 B*13:02 B*15:13 B*27:03 B*27:04 B*27:06 B*27:09 B*38:01 B*44:02 B*44:03 B*44:05 B*47:01 B*51:01 B*51:02 B*52:01 B*53:01 B*57:01 B*57:03 B*58:01
B*07:02 B*15:01 B*27:08 B*50:01 B*08:01 B*14:01 B*14:02 B*15:02 B*15:03 B*15:10 B*15:11 B*15:12 B*35:01 B*35:08 B*39:01 B*40:01 B*40:02 B*40:06 B*41:03 B*41:04 B*42:01 B*42:02 B*45:01 B*46:01 B*48:01 B*54:01 B*55:01 B*56:01 B*67:01 B*73:01 B*81:01 B*82:01
Antigen

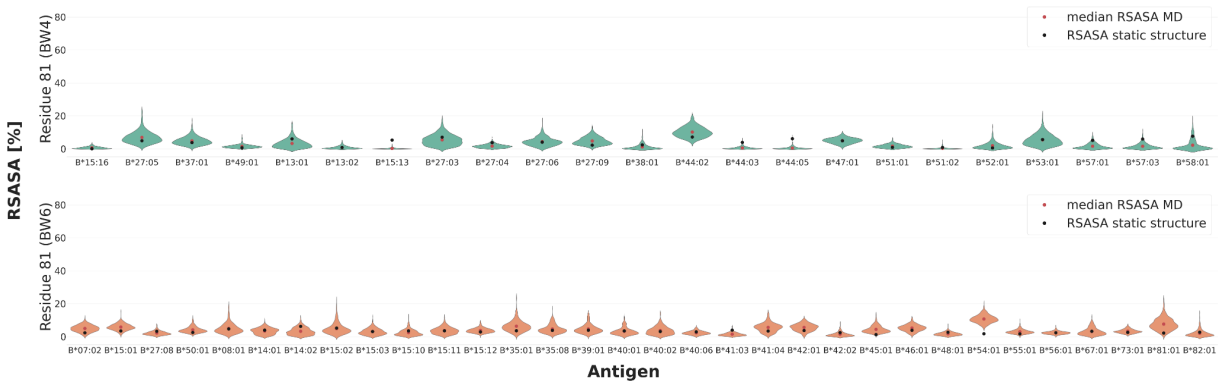
B)


RSASA [%]
Residue 78 (BW4)
Residue 78 (BW6)
median RSASA MD
RSASA static structure
B*15:16 B*27:05 B*37:01 B*49:01 B*13:01 B*13:02 B*15:13 B*27:03 B*27:04 B*27:06 B*27:09 B*38:01 B*44:02 B*44:03 B*44:05 B*47:01 B*51:01 B*51:02 B*52:01 B*53:01 B*57:01 B*57:03 B*58:01
B*07:02 B*15:01 B*27:08 B*50:01 B*08:01 B*14:01 B*14:02 B*15:02 B*15:03 B*15:10 B*15:11 B*15:12 B*35:01 B*35:08 B*39:01 B*40:01 B*40:02 B*40:06 B*41:03 B*41:04 B*42:01 B*42:02 B*45:01 B*46:01 B*48:01 B*54:01 B*55:01 B*56:01 B*67:01 B*73:01 B*81:01 B*82:01
Antigen

C)


RSASA [%]
Residue 79 (BW4)
Residue 79 (BW6)
median RSASA MD
RSASA static structure
B*15:16 B*27:05 B*37:01 B*49:01 B*13:01 B*13:02 B*15:13 B*27:03 B*27:04 B*27:06 B*27:09 B*38:01 B*44:02 B*44:03 B*44:05 B*47:01 B*51:01 B*51:02 B*52:01 B*53:01 B*57:01 B*57:03 B*58:01
B*07:02 B*15:01 B*27:08 B*50:01 B*08:01 B*14:01 B*14:02 B*15:02 B*15:03 B*15:10 B*15:11 B*15:12 B*35:01 B*35:08 B*39:01 B*40:01 B*40:02 B*40:06 B*41:03 B*41:04 B*42:01 B*42:02 B*45:01 B*46:01 B*48:01 B*54:01 B*55:01 B*56:01 B*67:01 B*73:01 B*81:01 B*82:01
Antigen

D)


RSASA [%]
Residue 81 (BW4)
Residue 81 (BW6)
median RSASA MD
RSASA static structure
B*15:16 B*27:05 B*37:01 B*49:01 B*13:01 B*13:02 B*15:13 B*27:03 B*27:04 B*27:06 B*27:09 B*38:01 B*44:02 B*44:03 B*44:05 B*47:01 B*51:01 B*51:02 B*52:01 B*53:01 B*57:01 B*57:03 B*58:01
B*07:02 B*15:01 B*27:08 B*50:01 B*08:01 B*14:01 B*14:02 B*15:02 B*15:03 B*15:10 B*15:11 B*15:12 B*35:01 B*35:08 B*39:01 B*40:01 B*40:02 B*40:06 B*41:03 B*41:04 B*42:01 B*42:02 B*45:01 B*46:01 B*48:01 B*54:01 B*55:01 B*56:01 B*67:01 B*73:01 B*81:01 B*82:01
Antigen

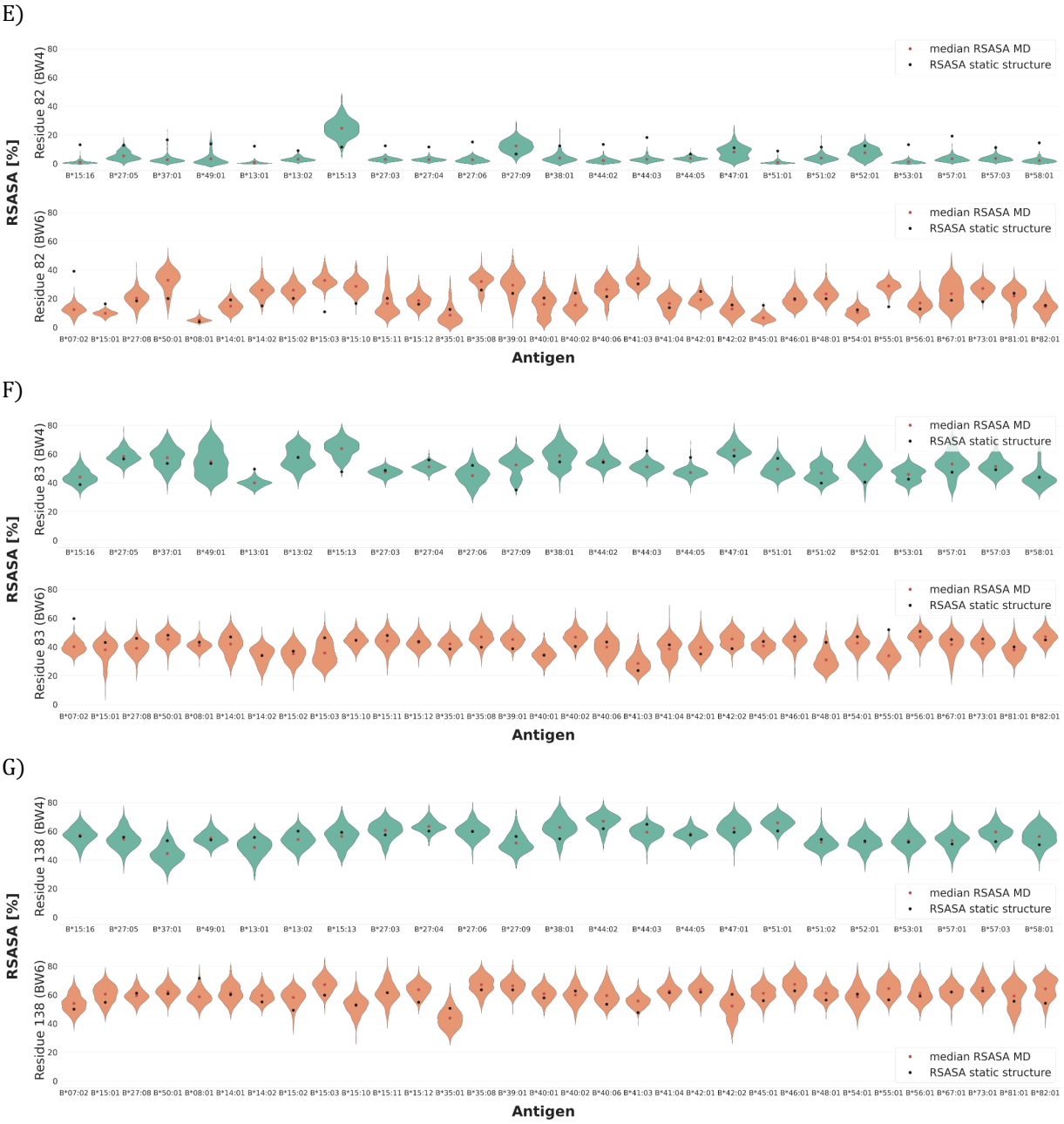



Figure S14: Solvent accessibility of surrounding positions to position 80 in antigens from BW4 and BW6 serological groups. Violins represent the RSASA distributions along the MD trajectories for each antigen. A) Violin plot for position 77. B) Violin plot for position 78. C) Violin plot for position 79. D) Violin plot for position 81. E) Violin plot for position 82. F) Violin plot for position 83. G) Violin plot for position 138. Red dot corresponds to the median RSASA of the violin while black dot corresponds to the RSASA value from the static structure.

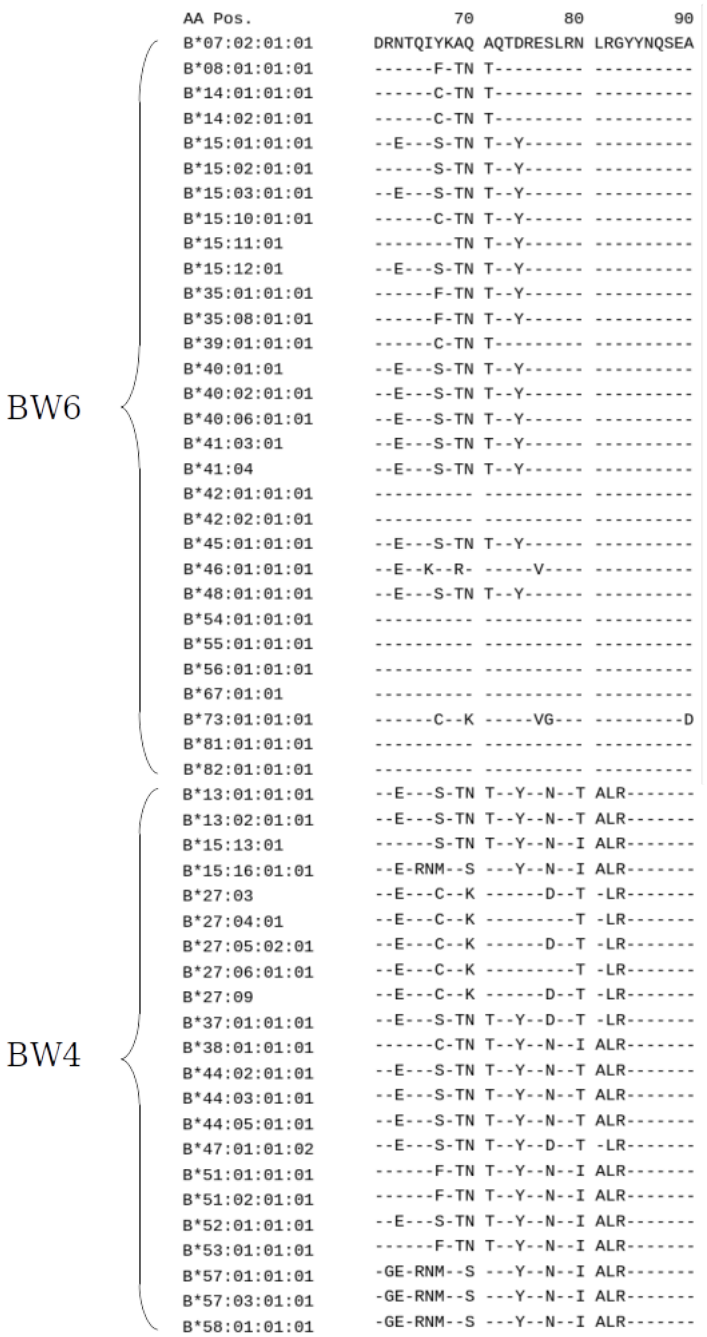
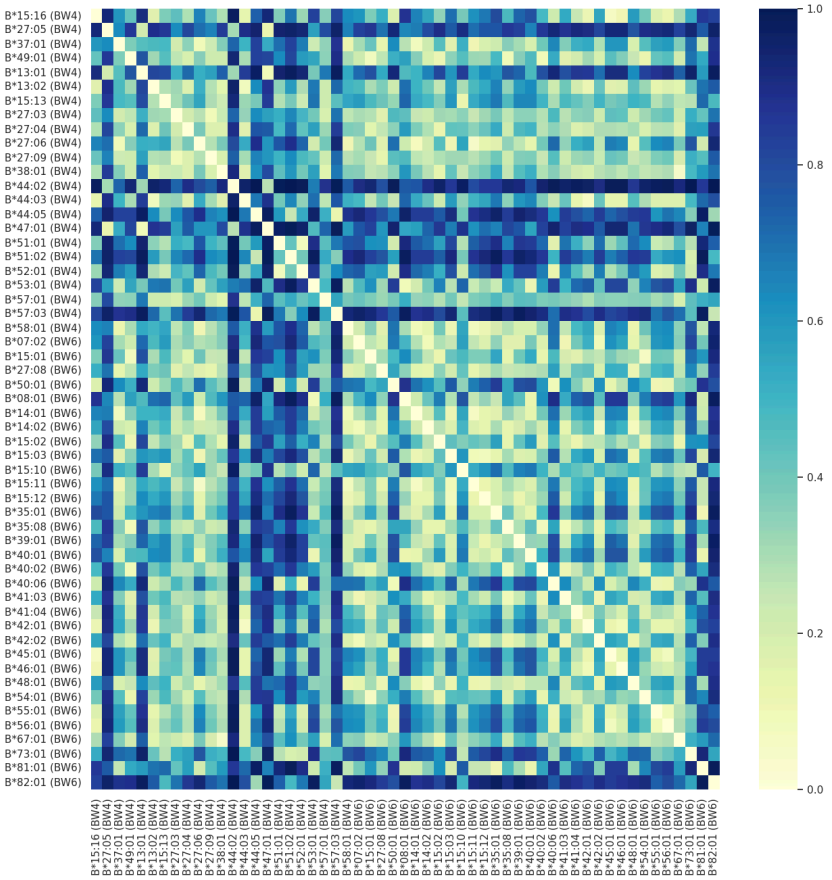
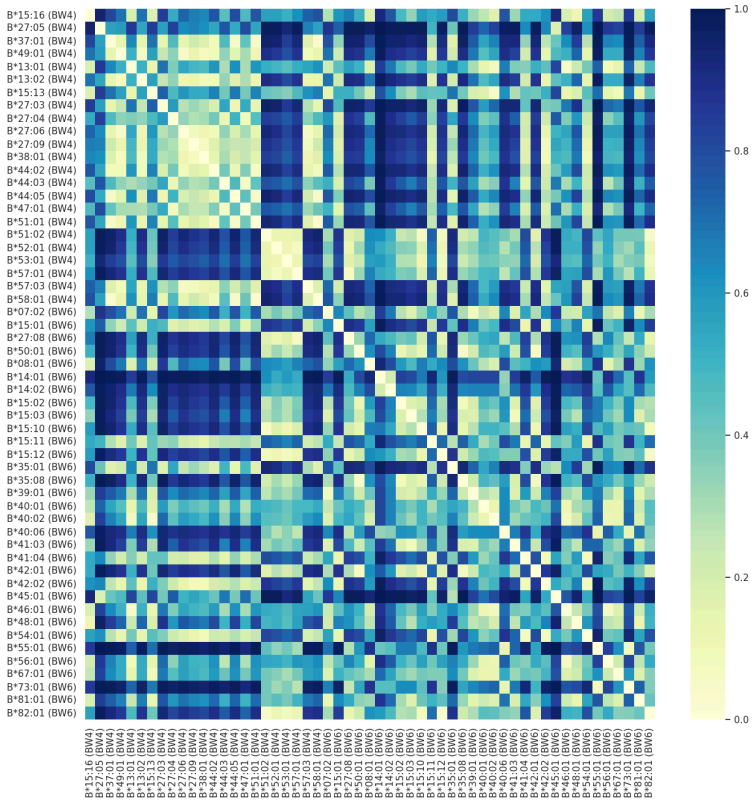



Figure S15: Sequence alignment for BW4 and BW6 antigens around position 80. Image obtained from [15].

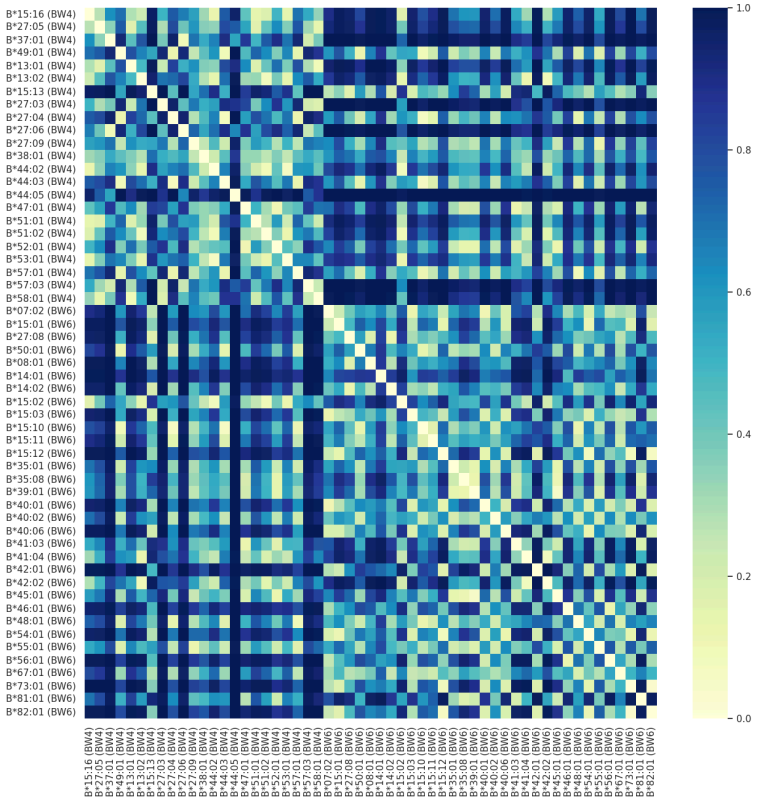
A)


B*15:16 (BW4)
B*27:05 (BW4)
B*37:01 (BW4)
B*49:01 (BW4)
B*13:01 (BW4)
B*13:02 (BW4)
B*15:13 (BW4)
B*27:03 (BW4)
B*27:04 (BW4)
B*27:06 (BW4)
B*27:09 (BW4)
B*38:01 (BW4)
B*44:02 (BW4)
B*44:03 (BW4)
B*44:05 (BW4)
B*47:01 (BW4)
B*51:01 (BW4)
B*51:02 (BW4)
B*52:01 (BW4)
B*53:01 (BW4)
B*57:01 (BW4)
B*57:03 (BW4)
B*58:01 (BW4)
B*07:02 (BW6)
B*15:01 (BW6)
B*27:08 (BW6)
B*50:01 (BW6)
B*08:01 (BW6)
B*14:01 (BW6)
B*14:02 (BW6)
B*15:02 (BW6)
B*15:03 (BW6)
B*15:10 (BW6)
B*15:11 (BW6)
B*15:12 (BW6)
B*35:01 (BW6)
B*35:08 (BW6)
B*39:01 (BW6)
B*40:01 (BW6)
B*40:02 (BW6)
B*40:06 (BW6)
B*41:03 (BW6)
B*41:04 (BW6)
B*42:01 (BW6)
B*42:02 (BW6)
B*45:01 (BW6)
B*46:01 (BW6)
B*48:01 (BW6)
B*54:01 (BW6)
B*55:01 (BW6)
B*56:01 (BW6)
B*67:01 (BW6)
B*73:01 (BW6)
B*81:01 (BW6)
B*82:01 (BW6)
1.0
0.8
0.6
0.4
0.2
0.0

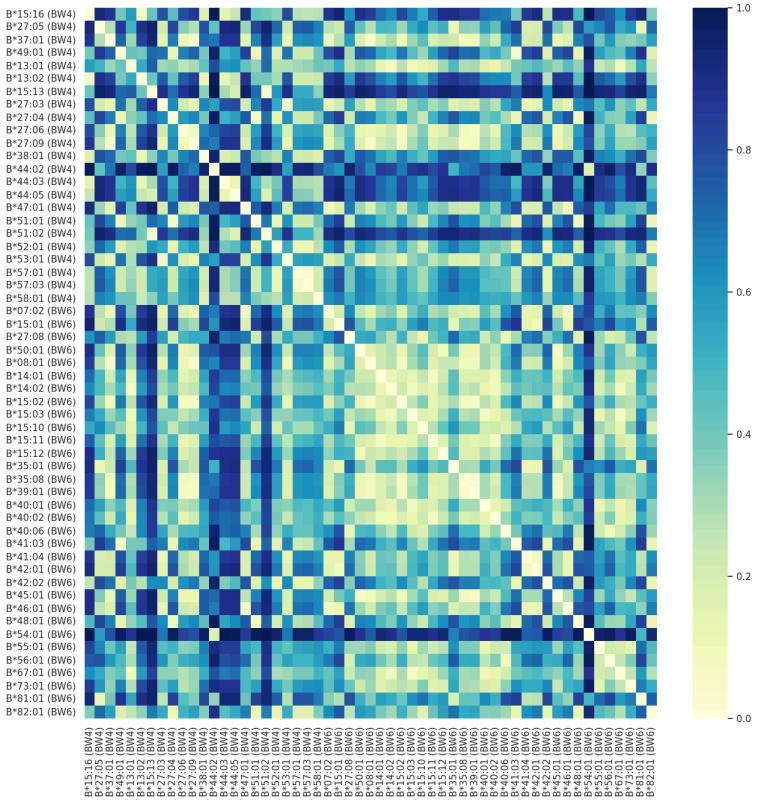
B)


B*15:16 (BW4)
B*27:05 (BW4)
B*37:01 (BW4)
B*49:01 (BW4)
B*13:01 (BW4)
B*13:02 (BW4)
B*15:13 (BW4)
B*27:03 (BW4)
B*27:04 (BW4)
B*27:06 (BW4)
B*27:09 (BW4)
B*38:01 (BW4)
B*44:02 (BW4)
B*44:03 (BW4)
B*44:05 (BW4)
B*47:01 (BW4)
B*51:01 (BW4)
B*51:02 (BW4)
B*52:01 (BW4)
B*53:01 (BW4)
B*57:01 (BW4)
B*57:03 (BW4)
B*58:01 (BW4)
B*07:02 (BW6)
B*15:01 (BW6)
B*27:08 (BW6)
B*50:01 (BW6)
B*08:01 (BW6)
B*14:01 (BW6)
B*14:02 (BW6)
B*15:02 (BW6)
B*15:03 (BW6)
B*15:10 (BW6)
B*15:11 (BW6)
B*15:12 (BW6)
B*35:01 (BW6)
B*35:08 (BW6)
B*39:01 (BW6)
B*40:01 (BW6)
B*40:02 (BW6)
B*40:06 (BW6)
B*41:03 (BW6)
B*41:04 (BW6)
B*42:01 (BW6)
B*42:02 (BW6)
B*45:01 (BW6)
B*46:01 (BW6)
B*48:01 (BW6)
B*54:01 (BW6)
B*55:01 (BW6)
B*56:01 (BW6)
B*67:01 (BW6)
B*73:01 (BW6)
B*81:01 (BW6)
B*82:01 (BW6)
1.0
0.8
0.6
0.4
0.2
0.0

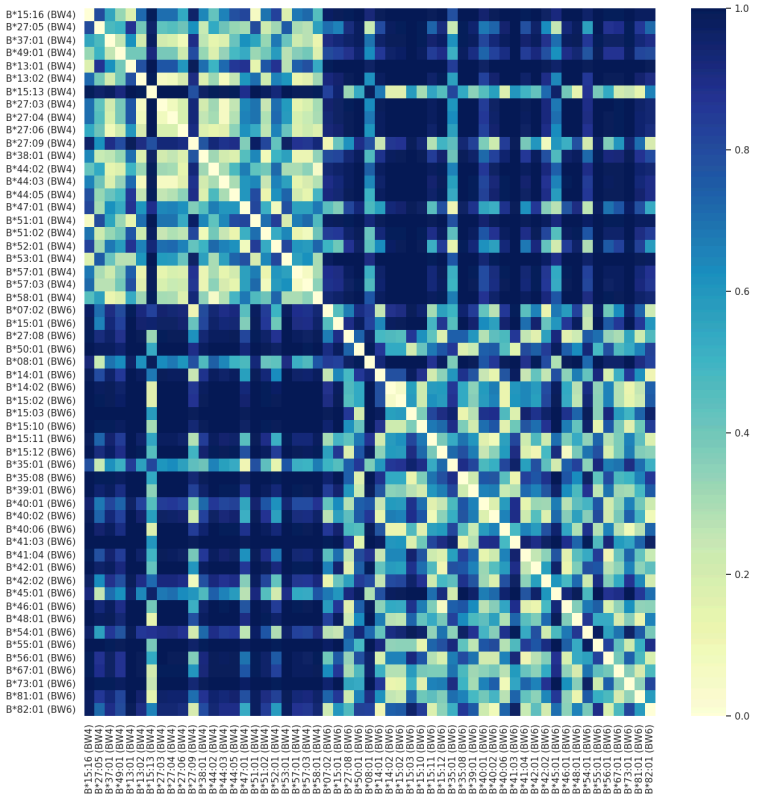
C)


B*15:16 (BW4)
B*27:05 (BW4)
B*37:01 (BW4)
B*49:01 (BW4)
B*13:01 (BW4)
B*13:02 (BW4)
B*15:13 (BW4)
B*27:03 (BW4)
B*27:04 (BW4)
B*27:06 (BW4)
B*27:09 (BW4)
B*38:01 (BW4)
B*44:02 (BW4)
B*44:03 (BW4)
B*44:05 (BW4)
B*47:01 (BW4)
B*51:01 (BW4)
B*51:02 (BW4)
B*52:01 (BW4)
B*53:01 (BW4)
B*57:01 (BW4)
B*57:03 (BW4)
B*58:01 (BW4)
B*07:02 (BW6)
B*15:01 (BW6)
B*27:08 (BW6)
B*50:01 (BW6)
B*08:01 (BW6)
B*14:01 (BW6)
B*14:02 (BW6)
B*15:02 (BW6)
B*15:03 (BW6)
B*15:10 (BW6)
B*15:11 (BW6)
B*15:12 (BW6)
B*35:01 (BW6)
B*35:08 (BW6)
B*39:01 (BW6)
B*40:01 (BW6)
B*40:02 (BW6)
B*40:06 (BW6)
B*41:03 (BW6)
B*41:04 (BW6)
B*42:01 (BW6)
B*42:02 (BW6)
B*45:01 (BW6)
B*46:01 (BW6)
B*48:01 (BW6)
B*54:01 (BW6)
B*55:01 (BW6)
B*56:01 (BW6)
B*67:01 (BW6)
B*73:01 (BW6)
B*81:01 (BW6)
B*82:01 (BW6)
1.0
0.8
0.6
0.4
0.2
0.0

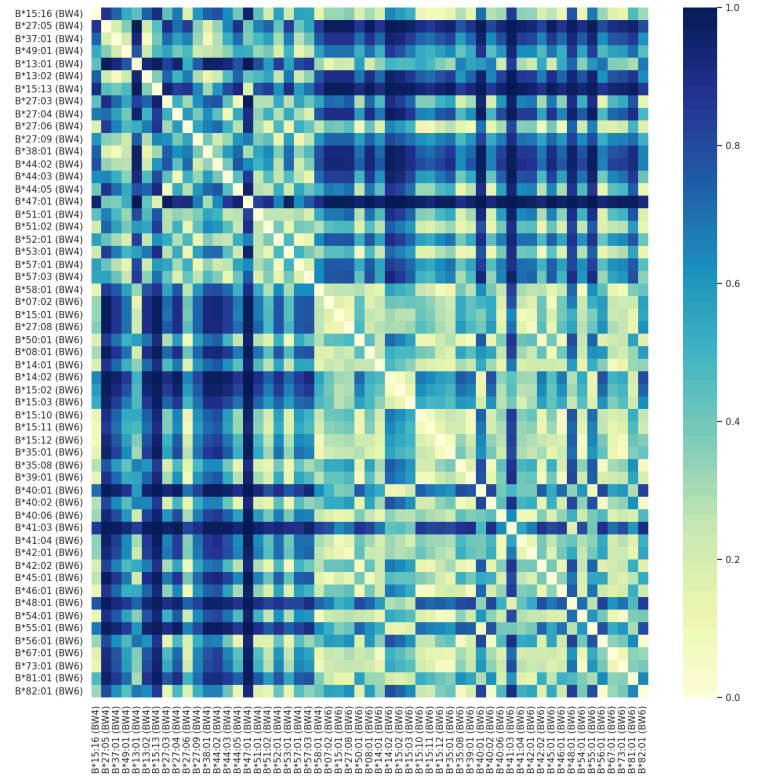
D)


B*15:16 (BW4)
B*27:05 (BW4)
B*37:01 (BW4)
B*49:01 (BW4)
B*13:01 (BW4)
B*13:02 (BW4)
B*15:13 (BW4)
B*27:03 (BW4)
B*27:04 (BW4)
B*27:06 (BW4)
B*27:09 (BW4)
B*38:01 (BW4)
B*44:02 (BW4)
B*44:03 (BW4)
B*44:05 (BW4)
B*47:01 (BW4)
B*51:01 (BW4)
B*51:02 (BW4)
B*52:01 (BW4)
B*53:01 (BW4)
B*57:01 (BW4)
B*57:03 (BW4)
B*58:01 (BW4)
B*07:02 (BW6)
B*15:01 (BW6)
B*27:08 (BW6)
B*50:01 (BW6)
B*08:01 (BW6)
B*14:01 (BW6)
B*14:02 (BW6)
B*15:02 (BW6)
B*15:03 (BW6)
B*15:10 (BW6)
B*15:11 (BW6)
B*15:12 (BW6)
B*35:01 (BW6)
B*35:08 (BW6)
B*39:01 (BW6)
B*40:01 (BW6)
B*40:02 (BW6)
B*40:06 (BW6)
B*41:03 (BW6)
B*41:04 (BW6)
B*42:01 (BW6)
B*42:02 (BW6)
B*45:01 (BW6)
B*46:01 (BW6)
B*48:01 (BW6)
B*54:01 (BW6)
B*55:01 (BW6)
B*56:01 (BW6)
B*67:01 (BW6)
B*73:01 (BW6)
B*81:01 (BW6)
B*82:01 (BW6)
1.0
0.8
0.6
0.4
0.2
0.0

E)


B*15:16 (BW4)
B*27:05 (BW4)
B*37:01 (BW4)
B*49:01 (BW4)
B*13:01 (BW4)
B*13:02 (BW4)
B*15:13 (BW4)
B*27:03 (BW4)
B*27:04 (BW4)
B*27:06 (BW4)
B*27:09 (BW4)
B*38:01 (BW4)
B*44:02 (BW4)
B*44:03 (BW4)
B*44:05 (BW4)
B*47:01 (BW4)
B*51:01 (BW4)
B*51:02 (BW4)
B*52:01 (BW4)
B*53:01 (BW4)
B*57:01 (BW4)
B*57:03 (BW4)
B*58:01 (BW4)
B*07:02 (BW6)
B*15:01 (BW6)
B*27:08 (BW6)
B*50:01 (BW6)
B*08:01 (BW6)
B*14:01 (BW6)
B*14:02 (BW6)
B*15:02 (BW6)
B*15:03 (BW6)
B*15:10 (BW6)
B*15:11 (BW6)
B*15:12 (BW6)
B*35:01 (BW6)
B*35:08 (BW6)
B*39:01 (BW6)
B*40:01 (BW6)
B*40:02 (BW6)
B*40:06 (BW6)
B*41:03 (BW6)
B*41:04 (BW6)
B*42:01 (BW6)
B*42:02 (BW6)
B*45:01 (BW6)
B*46:01 (BW6)
B*48:01 (BW6)
B*54:01 (BW6)
B*55:01 (BW6)
B*56:01 (BW6)
B*67:01 (BW6)
B*73:01 (BW6)
B*81:01 (BW6)
B*82:01 (BW6)
1.0
0.8
0.6
0.4
0.2
0.0

F)


B*15:16 (BW4)
B*27:05 (BW4)
B*37:01 (BW4)
B*49:01 (BW4)
B*13:01 (BW4)
B*13:02 (BW4)
B*15:13 (BW4)
B*27:03 (BW4)
B*27:04 (BW4)
B*27:06 (BW4)
B*27:09 (BW4)
B*38:01 (BW4)
B*44:02 (BW4)
B*44:03 (BW4)
B*44:05 (BW4)
B*47:01 (BW4)
B*51:01 (BW4)
B*51:02 (BW4)
B*52:01 (BW4)
B*53:01 (BW4)
B*57:01 (BW4)
B*57:03 (BW4)
B*58:01 (BW4)
B*07:02 (BW6)
B*15:01 (BW6)
B*27:08 (BW6)
B*50:01 (BW6)
B*08:01 (BW6)
B*14:01 (BW6)
B*14:02 (BW6)
B*15:02 (BW6)
B*15:03 (BW6)
B*15:10 (BW6)
B*15:11 (BW6)
B*15:12 (BW6)
B*35:01 (BW6)
B*35:08 (BW6)
B*39:01 (BW6)
B*40:01 (BW6)
B*40:02 (BW6)
B*40:06 (BW6)
B*41:03 (BW6)
B*41:04 (BW6)
B*42:01 (BW6)
B*42:02 (BW6)
B*45:01 (BW6)
B*46:01 (BW6)
B*48:01 (BW6)
B*54:01 (BW6)
B*55:01 (BW6)
B*56:01 (BW6)
B*67:01 (BW6)
B*73:01 (BW6)
B*81:01 (BW6)
B*82:01 (BW6)
1.0
0.8
0.6
0.4
0.2
0.0

G)

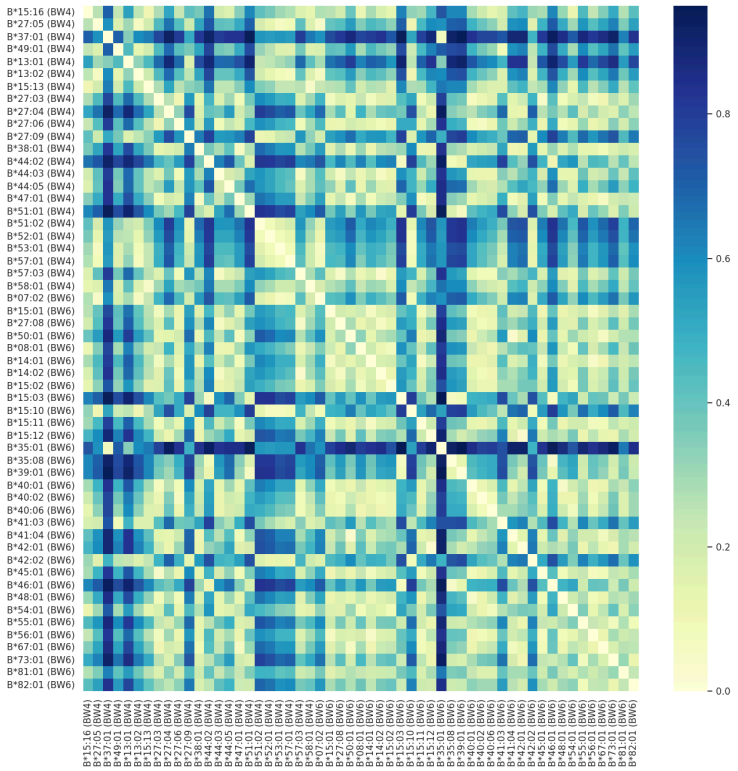



Figure S16: Heatmap of the Kolmogorov-Smirnov test statistic between RSASA distributions of BW4 and BW6 antigens at given positions along MD simulation.. A) Heatmap plot for position 77. B) Heatmap plot for position 78. C) Heatmap plot for position 79. D) Heatmap plot for position 81. E) Heatmap plot for position 82. F) Heatmap plot for position 83. G) Heatmap plot for position 138.

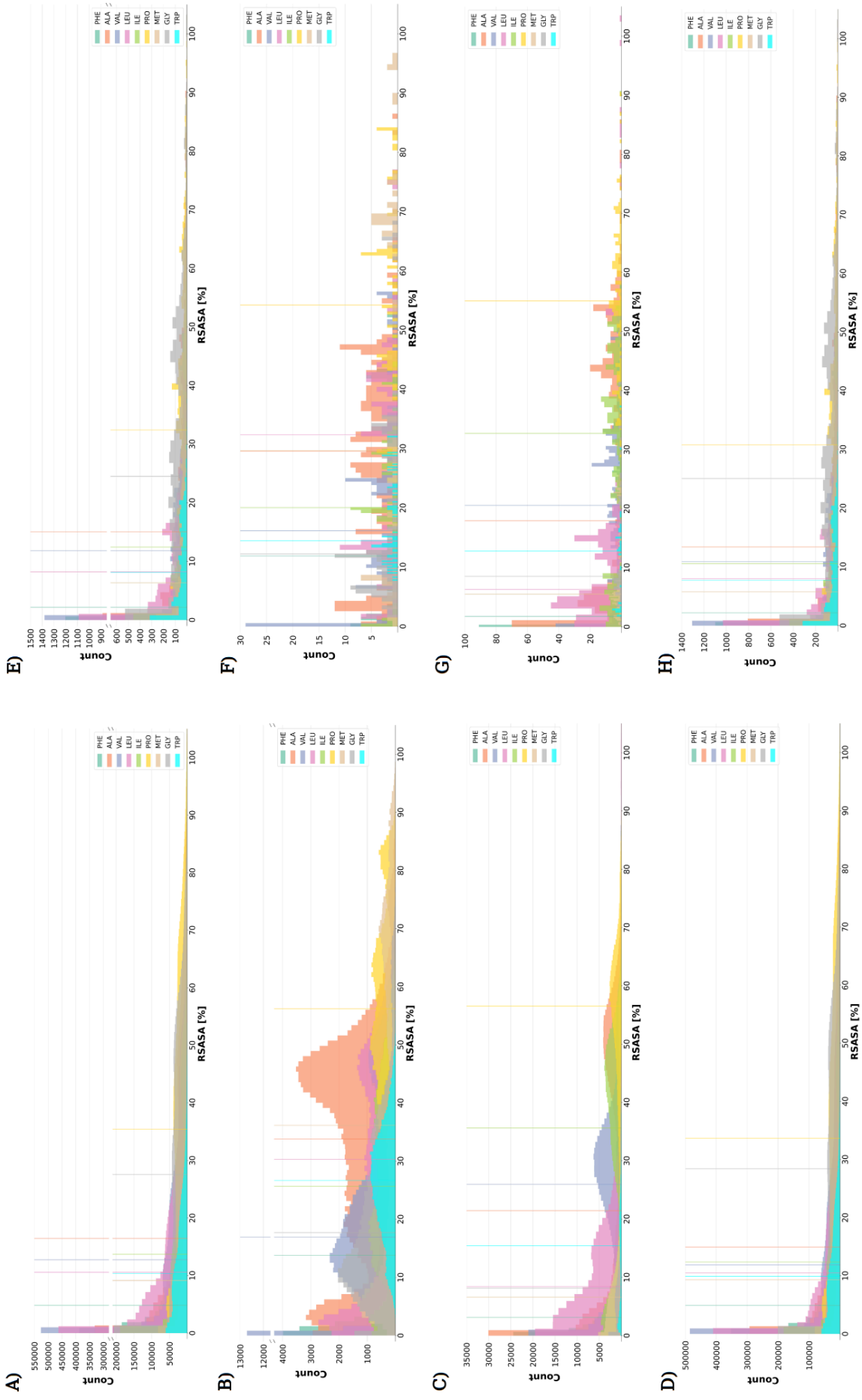



Figure S17: Comparison of RSASA values distributions derived from MD trajectories (panels A-D) with those derived from static 3D data (panels E-H) for the set of non-polar AAs (hydrophobic or not). The color code corresponding to the nine types of AA in the set is detailed on the right of the figure. RSASA values distributions are represented for the "All" group (panels A and E, 77 648 AAs), the "Confirmed" group (panels B and F, 2 985 AAs), the "Non-confirmed" group (panels C and G, 6 108 AAs), and the "Non-eplet" group (panels D and H, 69 684 AAs). Colored vertical lines indicate the median value for each AA type. Plots are generated with the Python package Matplotlib.

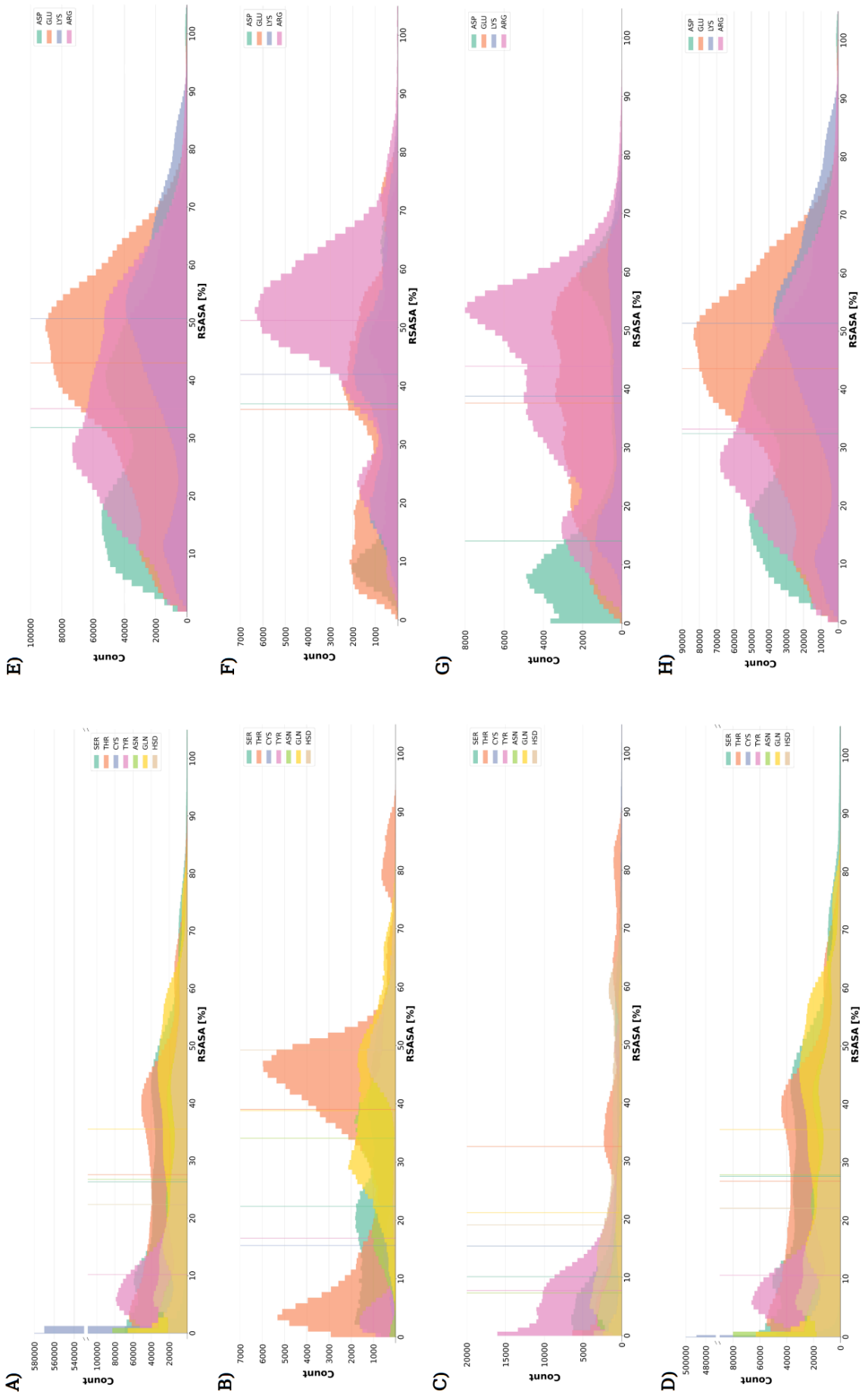



Figure S18: RSASA values distribution derived from MD trajectories for the set of polar and uncharged AAs (panels A-D) and the set of polar and charged AAs (panels E-H). The color codes corresponding to the 7 types of polar uncharged AAs and the 4 types of polar charged AAs are detailed on the left and right of the figure, respectively. RSASA values distributions are represented for the "All" group (panels A and E), the "Confirmed" group (panels B and F), the "Non-confirmed" group (panels C and G), and the "Non-eplet" group (panels D and H). Colored vertical lines indicate the median value for each AA type. Plots are generated with the Python package Matplotlib.

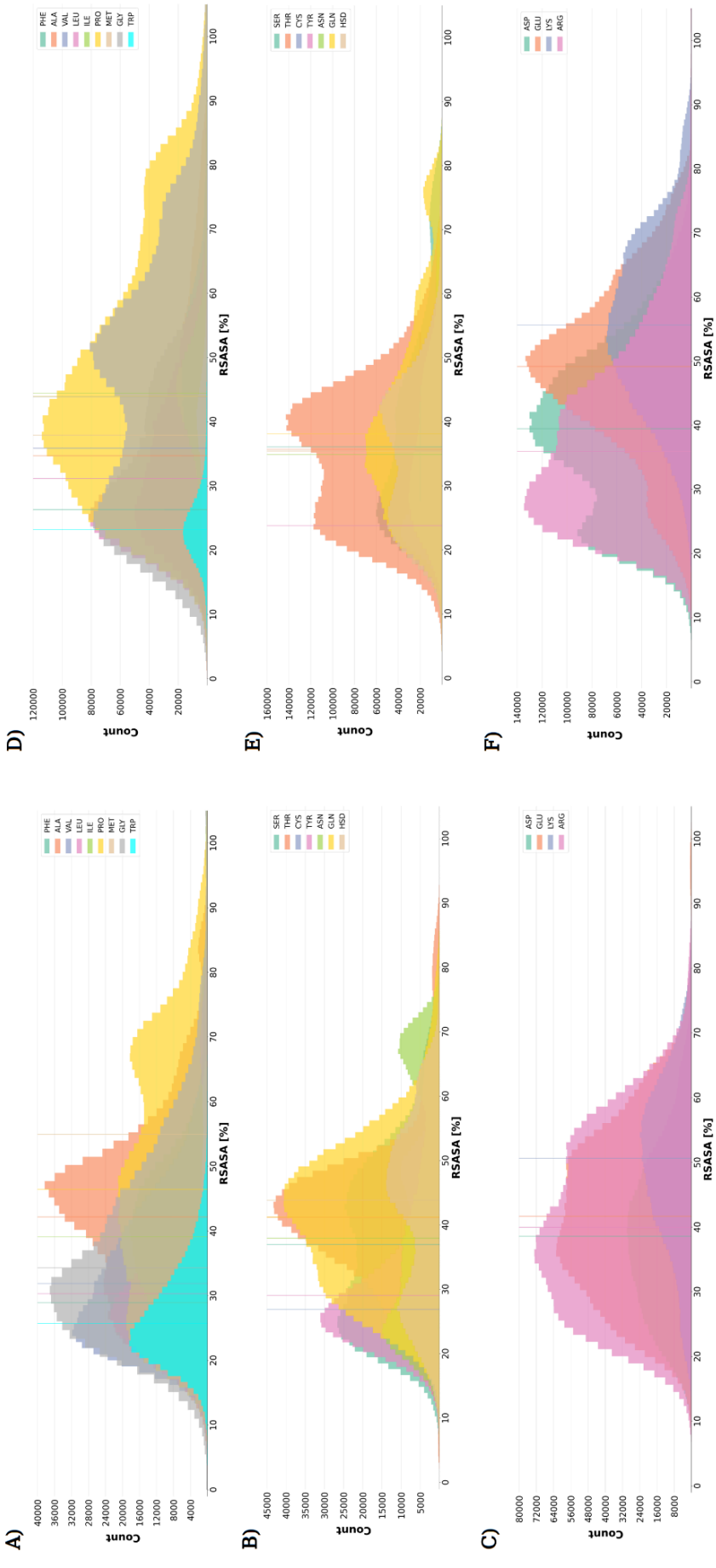



Figure S19: RSASA values distribution derived from MD trajectories for the "Epitope" (panels A-C) and the "Non-epitope"(panels D-F) groups of AAs. Plot corresponding to the three AA sets: Non-polar (panels A and D), polar-uncharged (panels B and E) and polar-charged (panels C and F), are grouped for the patch radius of 15Å. The color codes corresponding to the 9 types of non-polar AAs, 7 types of polar-uncharged AAs and 4 types of polar-charged AAs are detailed on the right of the figure. Colored vertical lines indicate the median value for each AA type (all median values can be found in Table S10). Plots are generated with the Python package Matplotlib.

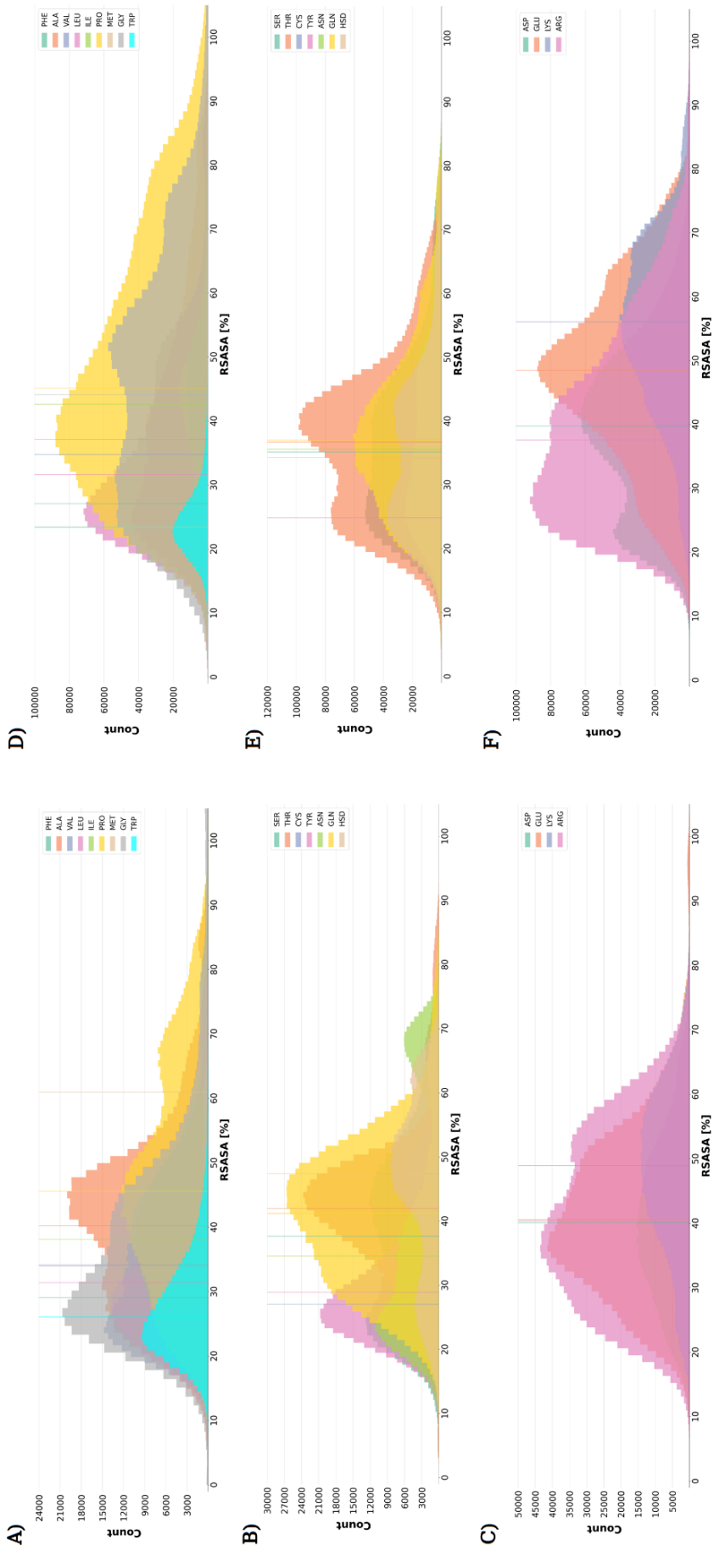



Figure S20: RSASA values distribution derived from MD trajectories for the "Epitope" (panels A-C) and the "Non-epitope"(panels D-F) groups of AAs. Plot corresponding to the three AA sets: Non-polar (panels A and D), polar-uncharged (panels B and E) and polar-charged (panels C and F), are grouped for the patch radius of 12Å. The color codes corresponding to the 9 types of non-polar AAs, 7 types of polar-uncharged AAs and 4 types of polar-charged AAs are detailed on the right of the figure. Colored vertical lines indicate the median value for each AA type (all median values can be found in Table S10). Plots are generated with the Python package Matplotlib.

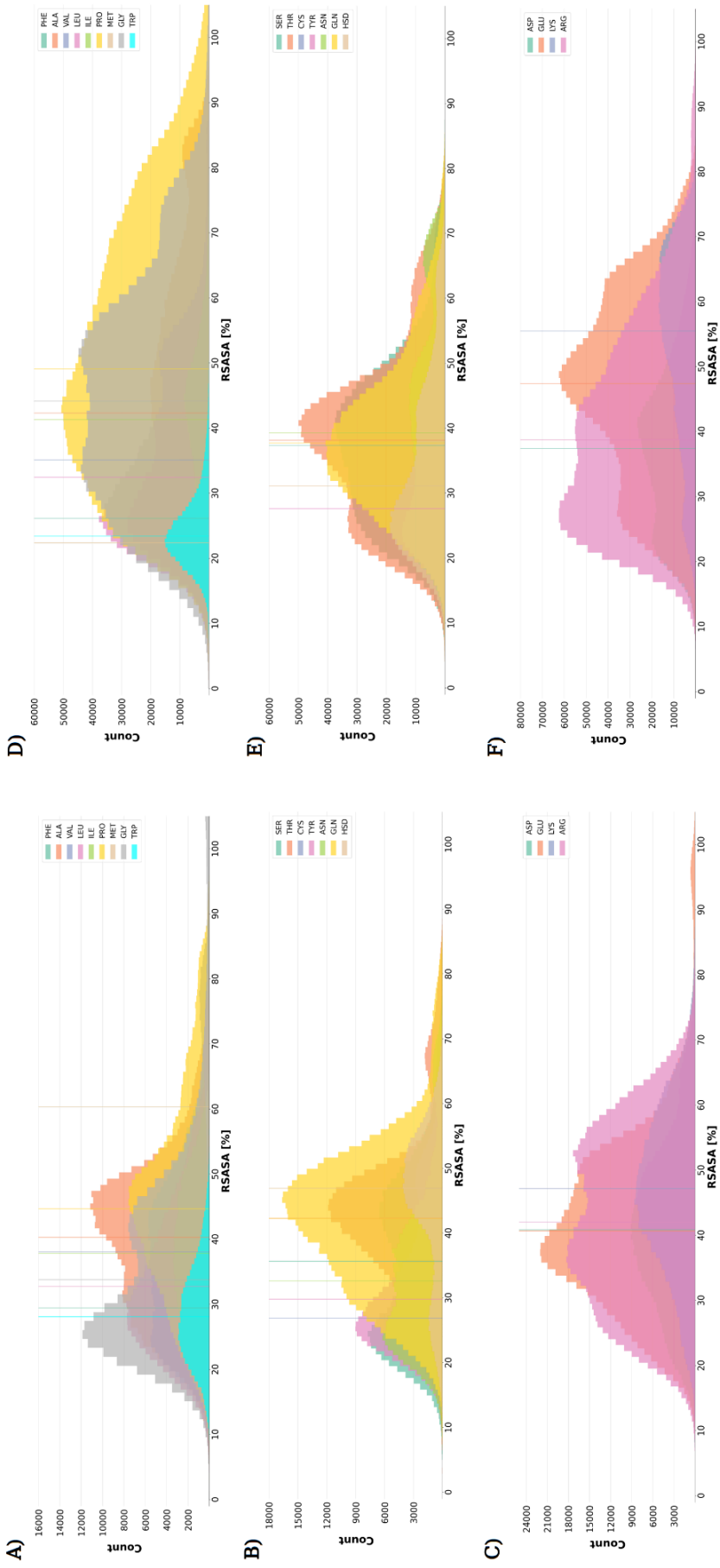



Figure S21: RSASA values distribution derived from MD trajectories for the "Epitope" (panels A-C) and the "Non-epitope"(panels D-F) groups of AAs. Plot corresponding to the three AA sets: Non-polar (panels A and D), polar-uncharged (panels B and E) and polar-charged (panels C and F), are grouped for the patch radius of 9Å. The color codes corresponding to the 9 types of non-polar AAs, 7 types of polar-uncharged AAs and 4 types of polar-charged AAs are detailed on the right of the figure. Colored vertical lines indicate the median value for each AA type (all median values can be found in Table S10). Plots are generated with the Python package Matplotlib.

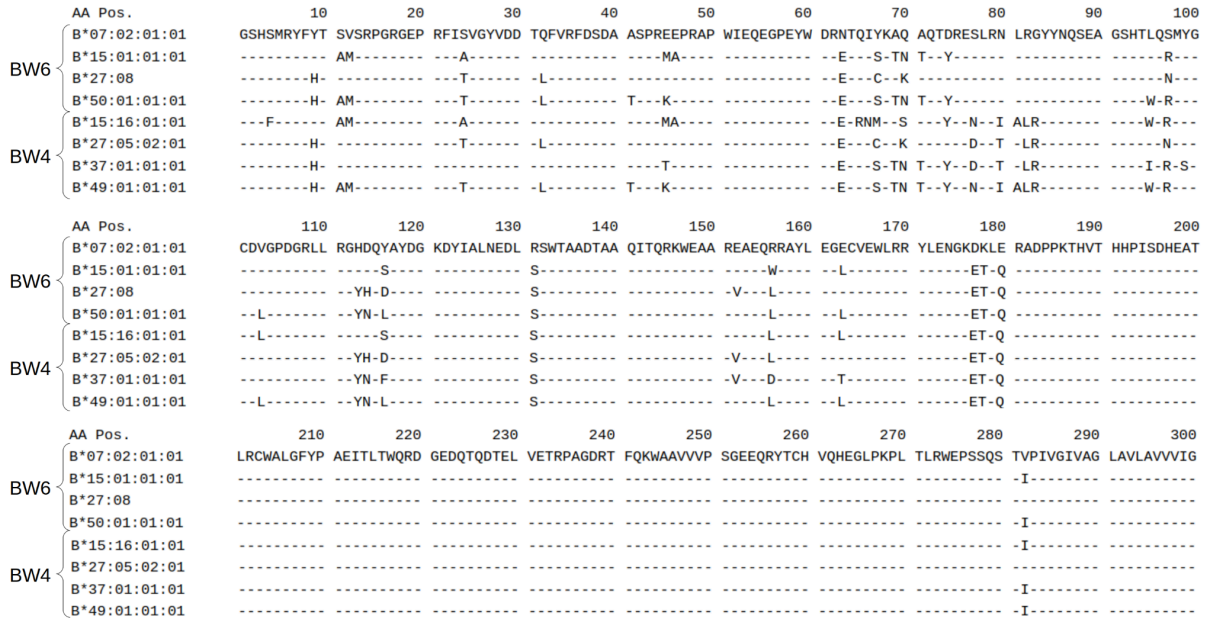



Figure S22: Sequence alignments of antigens in section 3.3. Antigens B*15:16, B*27:05, B*37:01 and B*49:01 correspond to the BW4 group, and antigens B*07:02, B*15:01, B*27:08 and B*50:01 correspond to the BW6 group. mage obtained from [[15]].

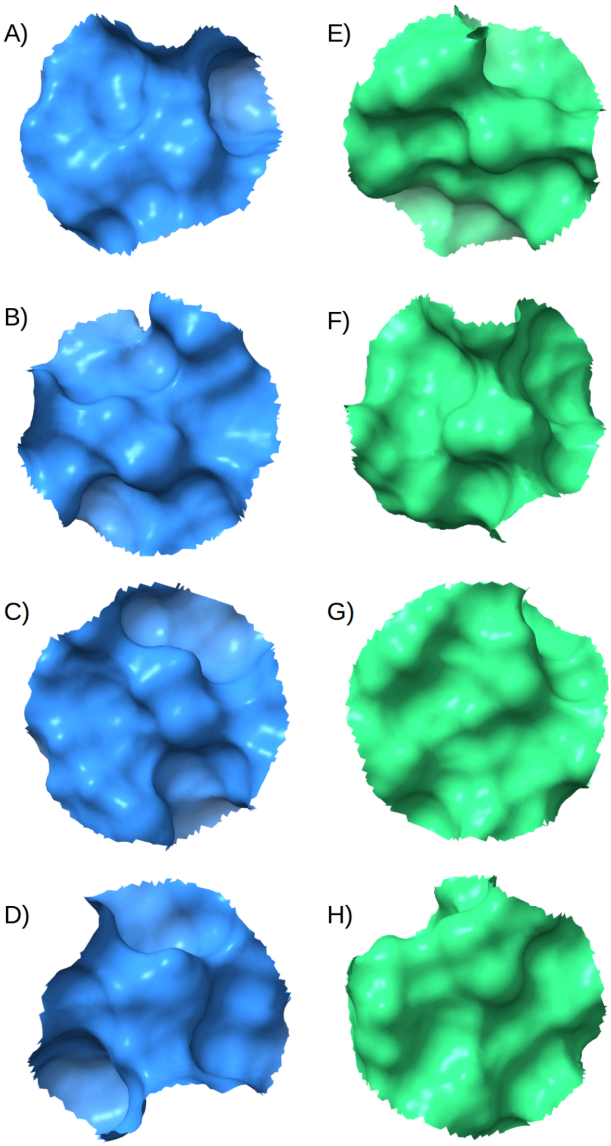



Figure S23: Examples of 7Å-sized patches centered at position 80 used for the calculation of Zernike descriptors. A) Antigen B*15:16. B) Antigen B*27:05. C) Antigen B*37:01. D) Antigen B*49:01. E) Antigen B*07:02. F) Antigen B*15:01. G) Antigen B*27:08. H) Antigen B*50:01.

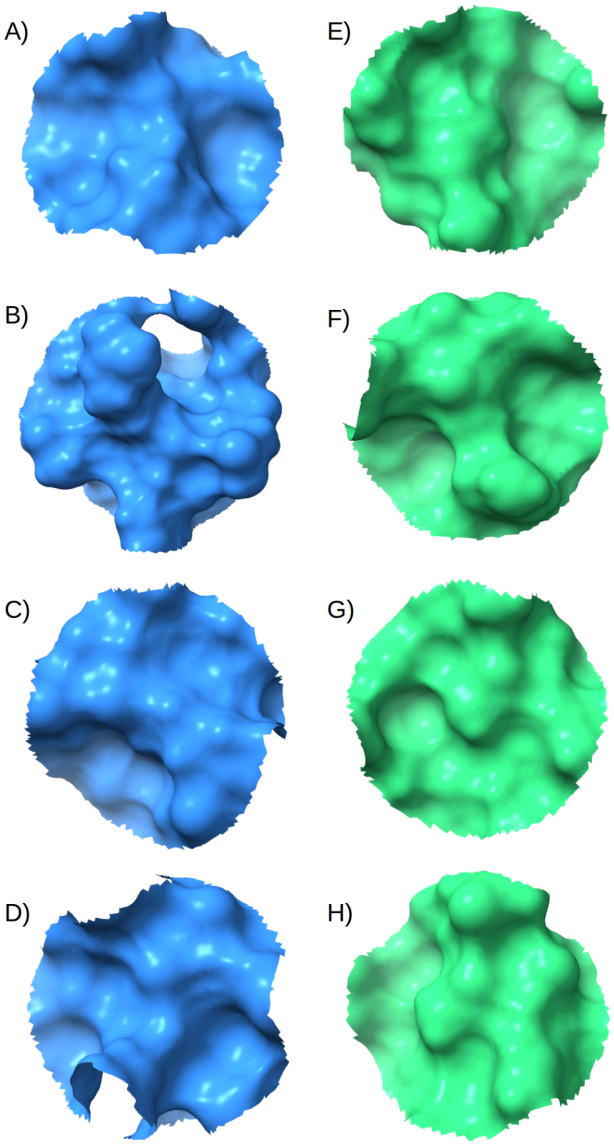



Figure S24: Examples of 7Å-sized patches centered at position 83 used for the calculation of Zernike descriptors. A) Antigen B*15:16. B) Antigen B*27:05. C) Antigen B*37:01. D) Antigen B*49:01. E) Antigen B*07:02. F) Antigen B*15:01. G) Antigen B*27:08. H) Antigen B*50:01.

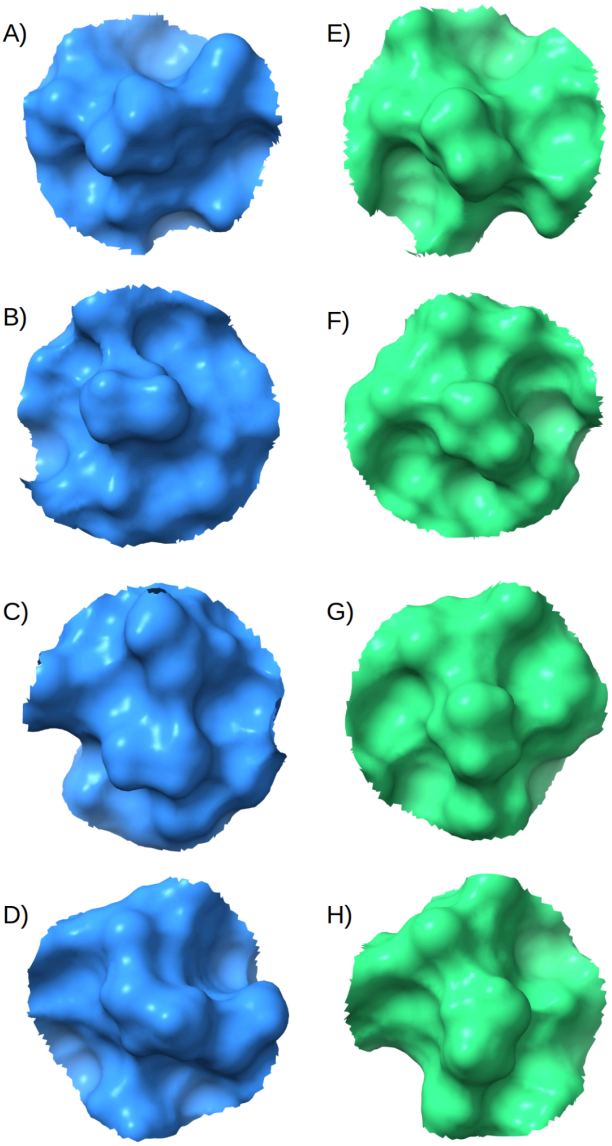



Figure S25: Examples of 7Å-sized patches centered at position 138 used for the calculation of Zernike descriptors. A) Antigen B*15:16. B) Antigen B*27:05. C) Antigen B*37:01. D) Antigen B*49:01. E) Antigen B*07:02. F) Antigen B*15:01. G) Antigen B*27:08. H) Antigen B*50:01.

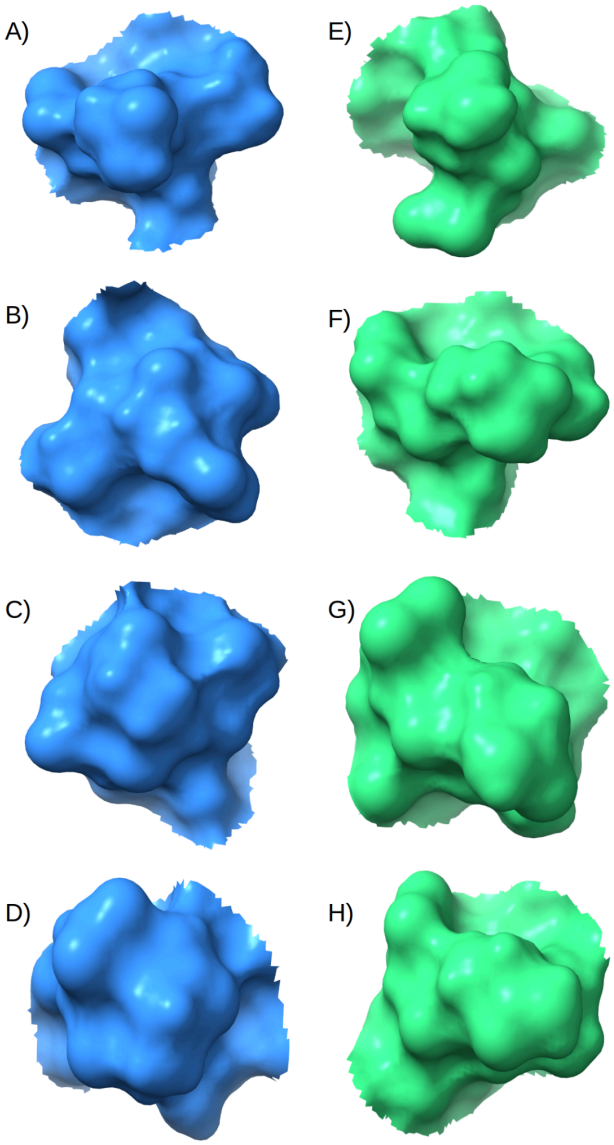



Figure S26: Examples of 7Å-sized patches centered at position 226 used for the calculation of Zernike descriptors. A) Antigen B*15:16. B) Antigen B*27:05. C) Antigen B*37:01. D) Antigen B*49:01. E) Antigen B*07:02. F) Antigen B*15:01. G) Antigen B*27:08. H) Antigen B*50:01.

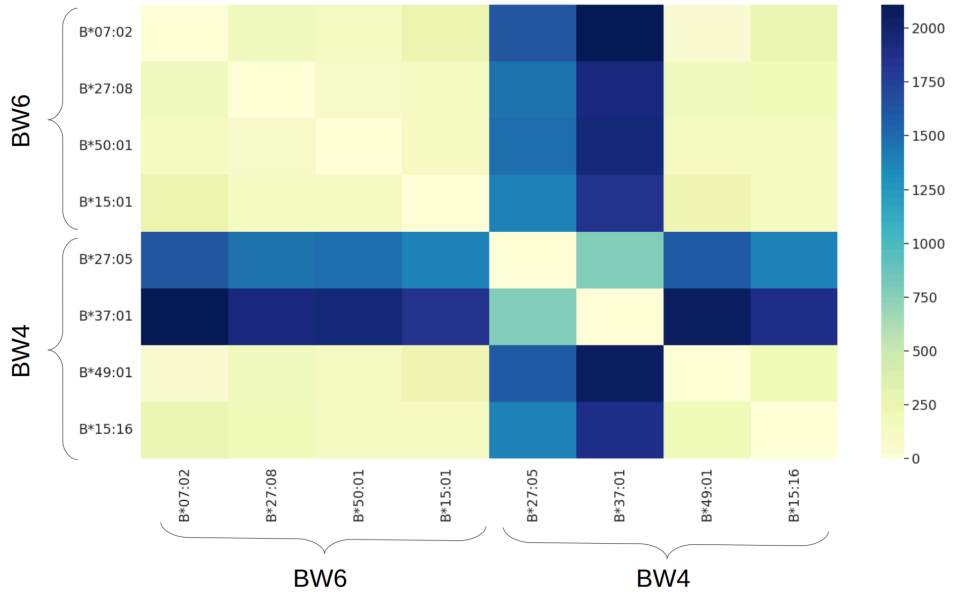



Figure S27: Heatmap of the Wasserstein distances between the patches of position 80 of the antigens from the BW4 and BW6 groups.

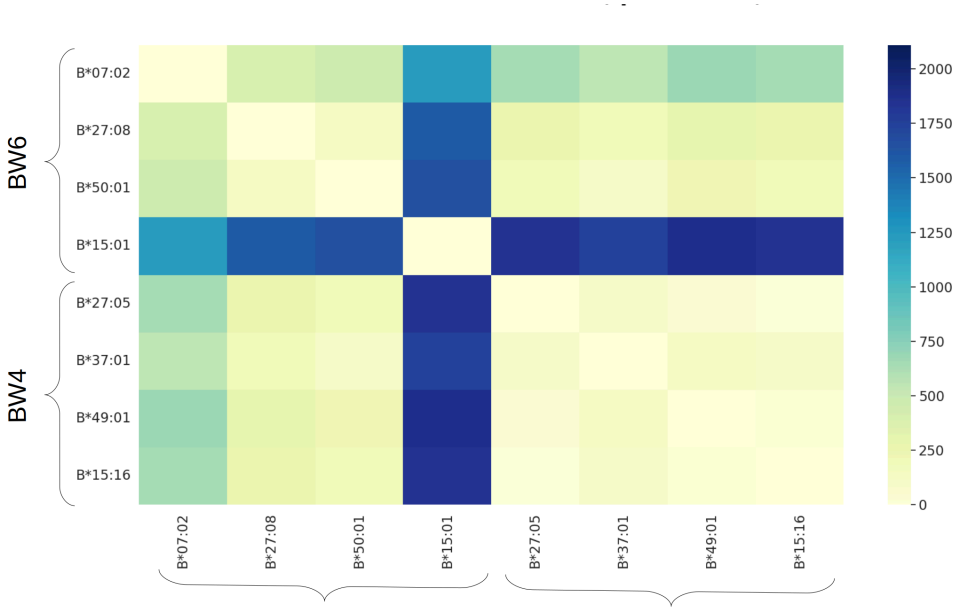



Figure S28: Heatmap of the Wasserstein distances between the patches of position 83 of the antigens from the BW4 and BW6 groups.

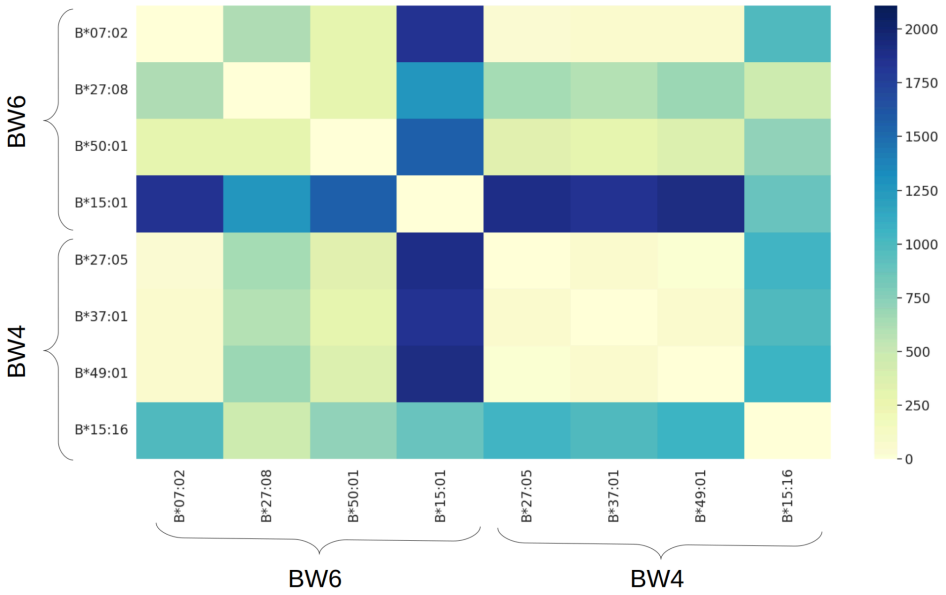



Figure S29: Heatmap of the Wasserstein distances between the patches of position 138 of the antigens from the BW4 and BW6 groups.

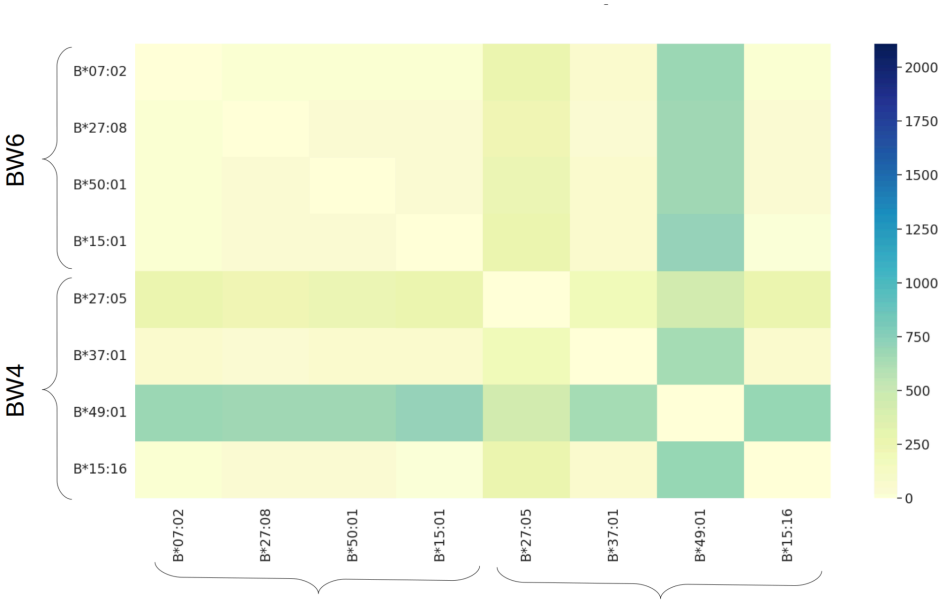



Figure S30: Heatmap of the Wasserstein distances between the patches of position 226 of the antigens from the BW4 and BW6 groups.

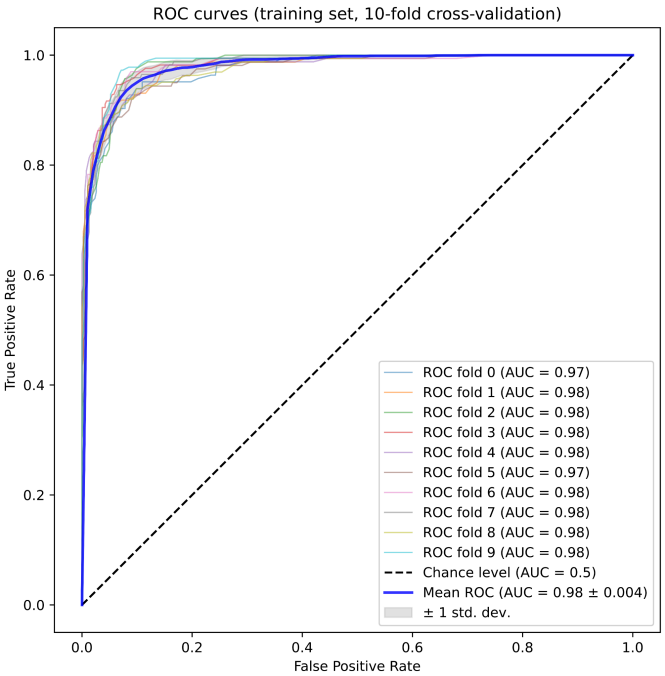



Figure S31: Receiver Operating Characteristic (ROC) curves of a 10-fold cross-validation on training set (repetition 0). AUC value was computed for each fold as well as for the mean ROC curve.

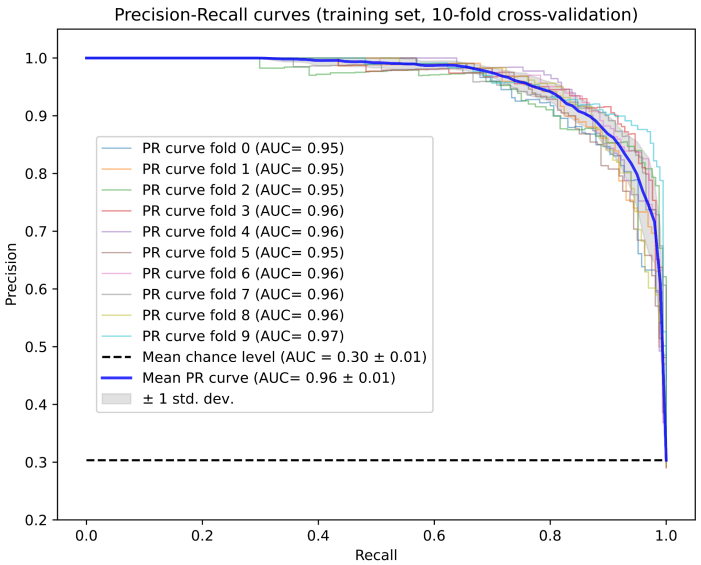



Figure S32: Prediction-Recall (PR) curves of a 10-fold cross-validation on training set (repetition 0). AUC value was computed for each fold as well as for the mean PR curve.

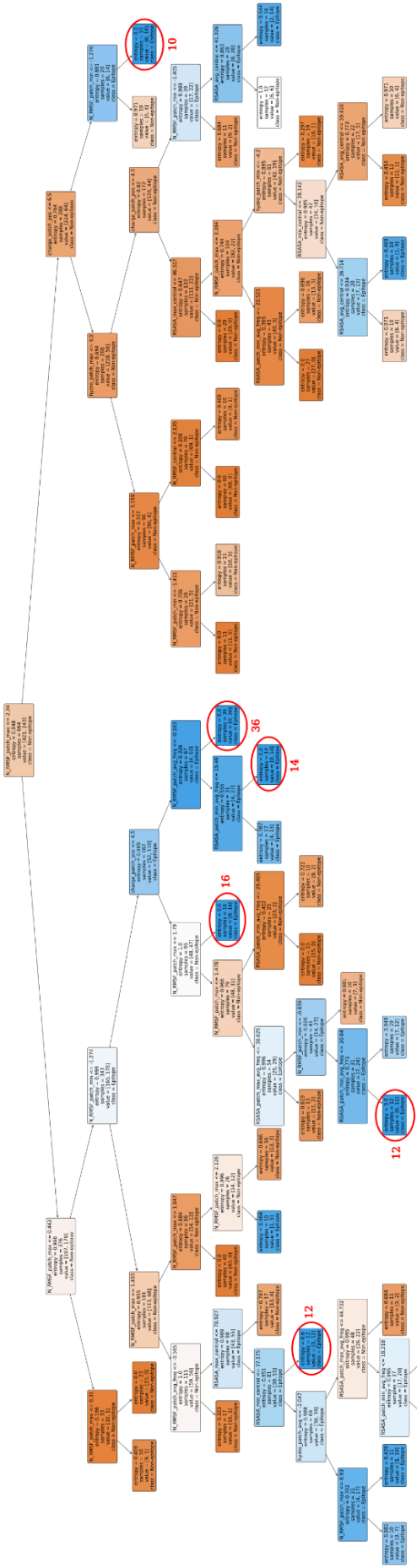



Figure S33: Visualization of the DT trained using the whole Non-redundant dataset. In red circles, the pure Epitope leaves are highlighted with the respective number of epitope samples also indicated in red. Right branch is for False and the left one for True.

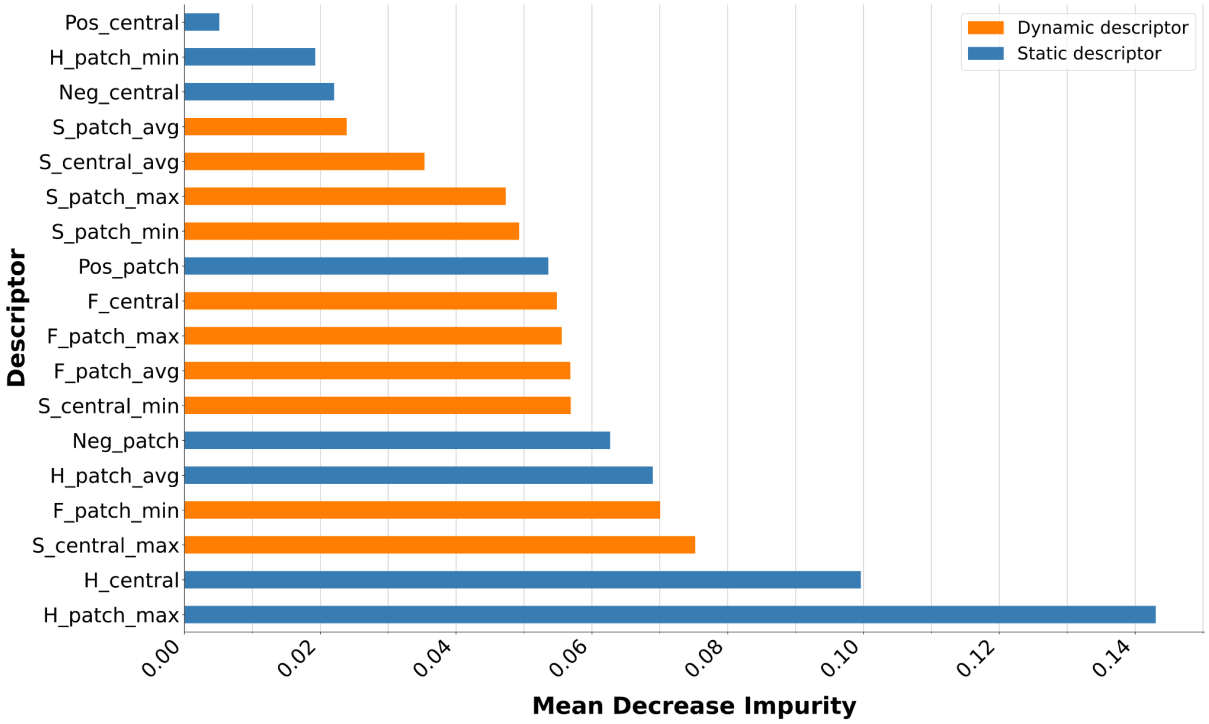



Figure S34: MDI feature importances of the DT trained on the whole Redundant dataset. Orange bars correspond to dynamic descriptors. Blue bars correspond to static descriptors.

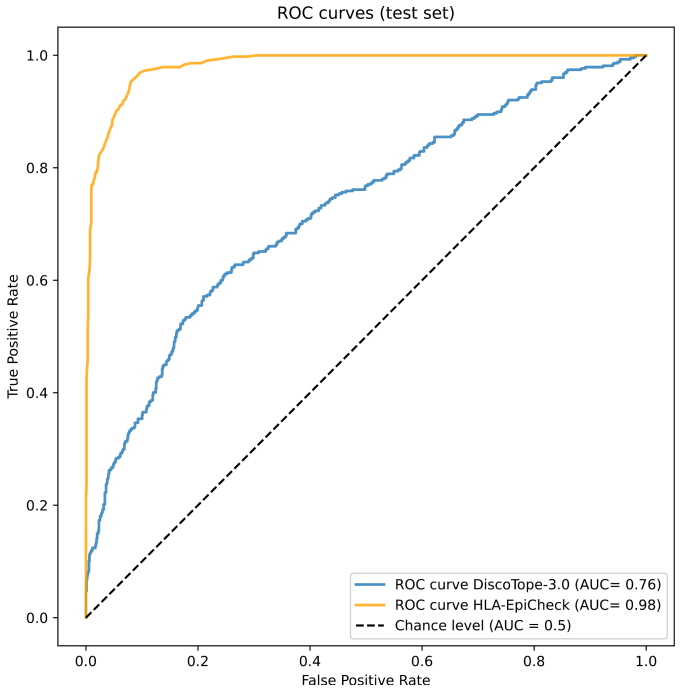



Figure S35: Receiver Operating Characteristic (ROC) curves for HLA-EpiCheck and DiscoTope-3.0 on test set.

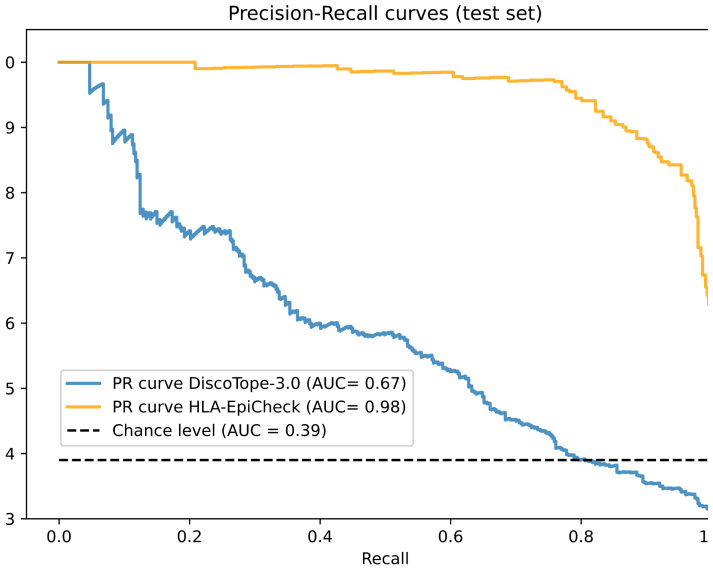



Figure S36: Prediction-Recall (PR) curves for HLA-EpiCheck and DiscoTope-3.0 on test set.

## Data availability:

All the data used to generate the results presented in this work are openly available at https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/GXZHH8

# Bibliography


1. WHO, ONT: Global observatory on donation and transplantation., https://www.transplant-observatory.org/data-charts-and-tables/chart/, last accessed 2024/01/11.
2. Agence de la biomédicine: Chiffres de l'activité du prélèvement et de la greffe d'organes en 2022, https://presse.agence-biomedecine.fr/chiffres-2022-de-lactivite-de-prelevement-et-de-greffe-dorganes-et-de-tissus-et-baromete-2023-sur-la-connaissance-et-la-perception-du-don-dorganes-en-france/, last accessed 2024/01/11.
3. Bonneau, A., Monchaud, C.: La transplantation d'organes en France. Actual. Pharm. 60, 18–20 (2021). https://doi.org/10.1016/j.actpha.2021.02.005.
4. Agence de la biomédicine: Don d'organes et de tissus, https://presse.agence-biomedecine.fr/section/don-dorganes-et-de-tissus/, last accessed 2024/01/11.
5. Landsteiner, K.: On Agglutination of Normal Human Blood. Transfusion (Paris). 1, 5–8 (1961). https://doi.org/10.1111/j.1537-2995.1961.tb00005.x.
6. Bugert, P., Klüter, H.: 100 Years after von Dungern & Hirschfeld: Kinship Investigation from Blood Groups to SNPs. Transfus. Med. Hemotherapy. 39, 161–162 (2012). https://doi.org/10.1159/000339263.
7. Watkins, W.M., Morgan, W.T.J.: Possible Genetical Pathways for the Biosynthesis of Blood Group Mucopolysaccharides. Vox Sang. 4, 97–119 (1959). https://doi.org/10.1111/j.1423-0410.1959.tb04023.x.
8. Krusius, T., Finne, J., Rauvala, H.: The Poly(glycosyl) Chains of Glycoproteins. Eur. J. Biochem. 92, 289–300 (1978). https://doi.org/10.1111/j.1432-1033.1978.tb12747.x.
9. Yamamoto, F., Clausen, H., White, T., Marken, J., Hakomori, S.: Molecular genetic basis of the histo-blood group ABO system. Nature. 345, 229–233 (1990). https://doi.org/10.1038/345229a0.
10. Abbas, A.K., Lichtman, A.H., Pillai, S.: Chapter 6. Antigen Presentation to T Lymphocytes and the Function of Major Histocompatibility Complex Molecules. In: Cellular and Molecular Immunology. Elsevier Health Sciences (2021).
11. Mak, T.W., Saunders, M.E., Jett, B.D. eds: Chapter 17: Transplantation. In: Primer to the Immune Response. pp. 457–486. Academic Cell, Boston (2014). https://doi.org/10.1016/B978-0-12-385245-8.00017-0.
12. Kloc, M., Ghobrial, R.M.: Chronic allograft rejection: A significant hurdle to transplant success. Burns Trauma. 2, 3–10 (2014). https://doi.org/10.4103/2321-3868.121646.
13. Wan, S.S., Chadban, S.J., Watson, N., Wyburn, K.: Development and outcomes of de novo donor-specific antibodies in low, moderate, and high immunological risk kidney transplant recipients. Am. J. Transplant. 20, 1351–1364 (2020). https://doi.org/10.1111/ajt.15754.
14. Rollini, P., Mach, B., Gorski, J.: Linkage map of three HLA-DR beta-chain genes: evidence for a recent duplication event. Proc. Natl. Acad. Sci. U. S. A. 82, 7197–7201 (1985).
15. Barker, D.J., Maccari, G., Georgiou, X., Cooper, M.A., Flicek, P., Robinson, J., Marsh, S.G.E.: The IPD-IMGT/HLA Database. Nucleic Acids Res. 51, D1053–D1060 (2023). https://doi.org/10.1093/nar/gkac1011.
16. Lim, W.H., Chadban, S.J., Clayton, P., Budgeon, C.A., Murray, K., Campbell, S.B., Cohney, S., Russ, G.R., McDonald, S.P.: Human leukocyte antigen mismatches associated with increased risk of rejection, graft failure, and death independent of initial immunosuppression in renal transplant recipients. Clin. Transplant. 26, E428–E437 (2012). https://doi.org/10.1111/j.1399-0012.2012.01654.x.
17. Pandey, P., Pande, A., Mandal, S., Devra, A.K., Sinha, V.K., Bhatt, A.P., Mishra, S.:

Effects of different sensitization events on HLA alloimmunization in renal transplant cases; a retrospective observation in 1066 cases. Transpl. Immunol. 75, 101680 (2022). https://doi.org/10.1016/j.trim.2022.101680.
18. Morales-Buenrostro, L.E., Terasaki, P.I., Marino-Vázquez, L.A., Lee, J.-H., El-Awar, N., Alberú, J.: “Natural” Human Leukocyte Antigen Antibodies Found in Nonalloimmunized Healthy Males. Transplantation. 86, 1111 (2008). https://doi.org/10.1097/TP.0b013e318186d87b.
19. Abbas, A.K., Lichtman, A.H., Pillai, S.: Chapter 17. Transplantation Immunology. In: Cellular and Molecular Immunology. Elsevier Health Sciences (2021).
20. Dai, H., Friday, A.J., Abou-Daya, K.I., Williams, A.L., Mortin-Toth, S., Nicotra, M.L., Rothstein, D.M., Shlomchik, W.D., Matozaki, T., Isenberg, J.S., Oberbarnscheidt, M.H., Danska, J.S., Lakkis, F.G.: Donor SIRPα polymorphism modulates the innate immune response to allogeneic grafts. Sci. Immunol. 2, eaam6202 (2017). https://doi.org/10.1126/sciimmunol.aam6202.
21. Koenig, A., Chen, C.-C., Marçais, A., Barba, T., Mathias, V., Sicard, A., Rabeyrin, M., Racapé, M., Duong-Van-Huyen, J.-P., Bruneval, P., Loupy, A., Dussurgey, S., Ducreux, S., Meas-Yedid, V., Olivo-Marin, J.-C., Paidassi, H., Guillemain, R., Taupin, J.-L., Callemeyn, J., Morelon, E., Nicoletti, A., Charreau, B., Dubois, V., Naesens, M., Walzer, T., Defrance, T., Thaunat, O.: Missing self triggers NK cell-mediated chronic vascular rejection of solid organ transplants. Nat. Commun. 10, 5350 (2019). https://doi.org/10.1038/s41467-019-13113-5.
22. Wordsworth, B.P., Allsopp, C.E., Young, R.P., Bell, J.I.: HLA-DR typing using DNA amplification by the polymerase chain reaction and sequential hybridization to sequence-specific oligonucleotide probes. Immunogenetics. 32, 413–418 (1990). https://doi.org/10.1007/BF00241635.
23. Immucor: LIFECODImmucorES® HLA SSO Typing Kits., https://www.immucor.com/en-gb/product/lifecodes-hla-sso-typing-kits/, last accessed 2024/03/11.
24. One Lambda: LABType: Reverse SSO DNA Typing Multiplex Assays, https://www.thermofisher.com/onelambda/wo/en/pre-transplant/hla-typing/labtype.html, last accessed 2024/03/11.
25. Dunckley, H.: HLA typing by SSO and SSP methods. Methods Mol. Biol. Clifton NJ. 882, 9–25 (2012). https://doi.org/10.1007/978-1-61779-842-9_2.
26. Olerup, O., Zetterquist, H.: HLA-DR typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours: an alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation. Tissue Antigens. 39, 225–235 (1992). https://doi.org/10.1111/j.1399-0039.1992.tb01940.x.
27. DNA sequencing with chain-terminating inhibitors | PNAS, https://www.pnas.org/doi/10.1073/pnas.74.12.5463, last accessed 2024/03/11.
28. Ambiguous allele combinations in HLA Class I and Class II sequence-based typing: when precise nucleotide sequencing leads to imprecise allele identification - PMC, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC517951/, last accessed 2024/03/11.
29. Bentley, G., Higuchi, R., Hoglund, B., Goodridge, D., Sayer, D., Trachtenberg, E.A., Erlich, H.A.: High-resolution, high-throughput HLA genotyping by next-generation sequencing. Tissue Antigens. 74, 393–403 (2009). https://doi.org/10.1111/j.1399-0039.2009.01345.x.
30. Rapid, scalable and highly automated HLA genotyping using next-generation sequencing: a transition from research to diagnostics | BMC Genomics | Full Text.
31. Duquesnoy, R.J.: Autobiographical perspectives on HLA epitopes: Past, present and future. Hum. Immunol. 83, 199–203 (2022). https://doi.org/10.1016/j.humimm.2022.01.001.
32. Duquesnoy, R.J.: A Structurally Based Approach to Determine HLA Compatibility at the Humoral Immune Level. Hum. Immunol. 67, 847–862 (2006). https://doi.org/10.1016/j.humimm.2006.08.001.
33. El-Awar, N., Jucaud, V., Nguyen, A.: HLA Epitopes: The Targets of Monoclonal and

Alloantibodies Defined. J. Immunol. Res. 2017, e3406230 (2017). https://doi.org/10.1155/2017/3406230.
34. Duquesnoy, R.J., Marrari, M., Sousa, L.C.D. da M., Barroso, J.R.P. de M., Aita, K.M. de S.U., da Silva, A.S., do Monte, S.J.H.: 16th IHIW: A Website for Antibody-Defined HLA Epitope Registry. Int. J. Immunogenet. 40, 54–59 (2013). https://doi.org/10.1111/iji.12017.
35. Youngs, D., Warner, P., Gallagher, M., Gimferrer, I.: New DQA1 allele specific antibody against epitope 2D (an exon 1 encoded amino acid). Considerations for alleles under the same P-group designation. Transpl. Immunol. 51, 32–39 (2018). https://doi.org/10.1016/j.trim.2018.08.007.
36. Senev, A., Coemans, M., Lerut, E., Van Sandt, V., Kerkhofs, J., Daniëls, L., Driessche, M.V., Compernolle, V., Sprangers, B., Van Loon, E., Callemeyn, J., Claas, F., Tambur, A.R., Verbeke, G., Kuypers, D., Emonds, M.-P., Naesens, M.: Eplet Mismatch Load and De Novo Occurrence of Donor-Specific Anti-HLA Antibodies, Rejection, and Graft Failure after Kidney Transplantation: An Observational Cohort Study. J. Am. Soc. Nephrol. 31, 2193–2204 (2020). https://doi.org/10.1681/ASN.2020010019.
37. Sapir-Pichhadze, R., Zhang, X., Ferradji, A., Madbouly, A., Tinckam, K.J., Gebel, H.M., Blum, D., Marrari, M., Kim, S.J., Fingerson, S., Bashyal, P., Cardinal, H., Foster, B.J.: Epitopes as characterized by antibody-verified eplet mismatches determine risk of kidney transplant loss. Kidney Int. 97, 778–785 (2020). https://doi.org/10.1016/j.kint.2019.10.028.
38. Lachmann, N., Todorova, K., Schulze, H., Schönemann, C.: Luminex® and Its Applications for Solid Organ Transplantation, Hematopoietic Stem Cell Transplantation, and Transfusion *. Transfus. Med. Hemotherapy. 40, 182–189 (2013). https://doi.org/10.1159/000351459.
39. Graham, H., Chandler, D.J., Dunbar, S.A.: The genesis and evolution of bead-based multiplexing. Methods. 158, 2–11 (2019). https://doi.org/10.1016/j.ymeth.2019.01.007.
40. Lhotte, Romain: Mathematical and computational Tools to Better Understand and Optimise Tissue Compatibility in Solid Organ Transplantation, (2023).
41. Bezstarosti, S., Bakker, K.H., Kramer, C.S.M., De Fijter, J.W., Reinders, M.E.J., Mulder, A., Claas, F.H.J., Heidt, S.: A Comprehensive Evaluation of the Antibody-Verified Status of Eplets Listed in the HLA Epitope Registry. Front. Immunol. 12, 800946 (2022). https://doi.org/10.3389/fimmu.2021.800946.
42. Crick, F.H.C.: The origin of the genetic code. J. Mol. Biol. 38, 367–379 (1968). https://doi.org/10.1016/0022-2836(68)90392-6.
43. Leja, Darryl: Levels of protein organization. (2006).
44. Pommié, C., Levadoux, S., Sabatier, R., Lefranc, G., Lefranc, M.: IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J. Mol. Recognit. 17, 17–32 (2004). https://doi.org/10.1002/jmr.647.
45. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132 (1982). https://doi.org/10.1016/0022-2836(82)90515-0.
46. Popot, J.-L., Vitry, C. de: On the Microassembly of Integral Membrane Proteins. Annu. Rev. Biophys. 19, 369–403 (1990). https://doi.org/10.1146/annurev.bb.19.060190.002101.
47. Zamyatnin, A.A.: Protein volume in solution. Prog. Biophys. Mol. Biol. 24, 107–123 (1972). https://doi.org/10.1016/0079-6107(72)90005-3.
48. wwPDB consortium: Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 47, D520–D528 (2019). https://doi.org/10.1093/nar/gky949.
49. Chruszcz, M., Wlodawer, A., Minor, W.: Determination of Protein Structures—A Series of Fortunate Events. Biophys. J. 95, 1–9 (2008). https://doi.org/10.1529/biophysj.108.131789.
50. McPherson, A., Gavira, J.A.: Introduction to protein crystallization. Acta Crystallogr. Sect. F Struct. Biol. Commun. 70, 2–20 (2013). https://doi.org/10.1107/S2053230X13033141.
51. Markwick, P.R.L., Malliavin, T., Nilges, M.: Structural Biology by NMR: Structure,

Dynamics, and Interactions. PLoS Comput. Biol. 4, e1000168 (2008). https://doi.org/10.1371/journal.pcbi.1000168.
52. Skiniotis, G., Southworth, D.R.: Single-particle cryo-electron microscopy of macromolecular complexes. Microscopy. 65, 9–22 (2016). https://doi.org/10.1093/jmicro/dfv366.
53. Joosten, R.P., Joosten, K., Cohen, S.X., Vriend, G., Perrakis, A.: Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. Bioinformatics. 27, 3392–3398 (2011). https://doi.org/10.1093/bioinformatics/btr590.
54. Joosten, R.P., Long, F., Murshudov, G.N., Perrakis, A.: The *PDB_REDO* server for macromolecular structure model optimization. IUCrJ. 1, 213–220 (2014). https://doi.org/10.1107/S2052252514009324.
55. Read, R.J., Adams, P.D., Arendall, W.B., Brunger, A.T., Emsley, P., Joosten, R.P., Kleywegt, G.J., Krissinel, E.B., Lütteke, T., Otwinowski, Z., Perrakis, A., Richardson, J.S., Sheffler, W.H., Smith, J.L., Tickle, I.J., Vriend, G., Zwart, P.H.: A New Generation of Crystallographic Validation Tools for the Protein Data Bank. Structure. 19, 1395–1412 (2011). https://doi.org/10.1016/j.str.2011.08.006.
56. Hooft, R.W.W., Sander, C., Vriend, G.: Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins Struct. Funct. Bioinforma. 26, 363–376 (1996). https://doi.org/10.1002/(SICI)1097-0134(199612)26:4<363::AID-PROT1>3.0.CO;2-D.
57. Word, J.M., Lovell, S.C., Richardson, J.S., Richardson, D.C.: Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1. J. Mol. Biol. 285, 1735–1747 (1999). https://doi.org/10.1006/jmbi.1998.2401.
58. Bruice, P.Y.: Organic Chemistry. Pearson/Prentice Hall (2004).
59. Schreiner, E., Trabuco, L.G., Freddolino, P.L., Schulten, K.: Stereochemical errors and their implications for molecular dynamics simulations. BMC Bioinformatics. 12, 190 (2011). https://doi.org/10.1186/1471-2105-12-190.
60. Stewart, D.E., Sarkar, A., Wampler, J.E.: Occurrence and role ofcis peptide bonds in protein structures. J. Mol. Biol. 214, 253–260 (1990). https://doi.org/10.1016/0022-2836(90)90159-J.
61. Jisna, V.A., Jayaraj, P.B.: Protein Structure Prediction: Conventional and Deep Learning Perspectives. Protein J. 40, 522–544 (2021). https://doi.org/10.1007/s10930-021-10003-y.
62. The Phyre2 web portal for protein modeling, prediction and analysis | Nature Protocols, https://www.nature.com/articles/nprot.2015.053, last accessed 2024/04/12.
63. HH-suite3 for fast remote homology detection and deep protein annotation | BMC Bioinformatics | Full Text, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3019-7, last accessed 2024/04/12.
64. Webb, B., Sali, A.: Protein Structure Modeling with MODELLER. In: Chen, Y.W. and Yiu, C.-P.B. (eds.) Structural Genomics: General Applications. pp. 239–255. Springer US, New York, NY (2021). https://doi.org/10.1007/978-1-0716-0892-0_14.
65. Laine, E., Eismann, S., Elofsson, A., Grudinin, S.: Protein sequence-to-structure learning: Is this the end(-to-end revolution)? Proteins Struct. Funct. Bioinforma. 89, 1770–1786 (2021). https://doi.org/10.1002/prot.26235.
66. Kandathil, S.M., Greener, J.G., Jones, D.T.: Recent developments in deep learning applied to protein structure prediction. Proteins Struct. Funct. Bioinforma. 87, 1179–1189 (2019). https://doi.org/10.1002/prot.25824.
67. Greener, J.G., Kandathil, S.M., Jones, D.T.: Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat. Commun. 10, 3977 (2019). https://doi.org/10.1038/s41467-019-11994-0.
68. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., Millán, C., Park, H., Adams, C., Glassman, C.R., DeGiovanni, A., Pereira, J.H., Rodrigues, A.V., van Dijk, A.A., Ebrecht, A.C., Opperman, D.J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M.K.,

Dalwadi, U., Yip, C.K., Burke, J.E., Garcia, K.C., Grishin, N.V., Adams, P.D., Read, R.J., Baker, D.: Accurate prediction of protein structures and interactions using a three-track neural network. Science. 373, 871–876 (2021). https://doi.org/10.1126/science.abj8754.
69. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with AlphaFold. Nature. 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2.
70. Pereira, J., Simpkin, A.J., Hartmann, M.D., Rigden, D.J., Keegan, R.M., Lupas, A.N.: High-accuracy protein structure prediction in CASP14. Proteins Struct. Funct. Bioinforma. 89, 1687–1699 (2021). https://doi.org/10.1002/prot.26171.
71. Mariani, V., Biasini, M., Barbato, A., Schwede, T.: lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 29, 2722–2728 (2013). https://doi.org/10.1093/bioinformatics/btt473.
72. Leach, A.R.: Molecular Modelling: Principles and Applications. Pearson Education (2001).
73. Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A.E., Kolinski, A.: Coarse-Grained Protein Models and Their Applications. Chem. Rev. 116, 7898–7936 (2016). https://doi.org/10.1021/acs.chemrev.6b00163.
74. Alder, B.J., Wainwright, T.E.: Studies in Molecular Dynamics. I. General Method. J. Chem. Phys. 31, 459–466 (1959). https://doi.org/10.1063/1.1730376.
75. Karplus, M., Petsko, G.A.: Molecular dynamics simulations in biology. Nature. 347, 631–639 (1990). https://doi.org/10.1038/347631a0.
76. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Mag. 17, 37–37 (1996). https://doi.org/10.1609/aimag.v17i3.1230.
77. Jia, J., Wang, W.: Review of reinforcement learning research. In: 2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC). pp. 186–191 (2020). https://doi.org/10.1109/YAC51587.2020.9337653.
78. Rani, V., Nabi, S.T., Kumar, M., Mittal, A., Kumar, K.: Self-supervised Learning: A Succinct Review. Arch. Comput. Methods Eng. 30, 2761–2775 (2023). https://doi.org/10.1007/s11831-023-09884-2.
79. Wu, L., Yin, C., Zhu, J., Wu, Z., He, L., Xia, Y., Xie, S., Qin, T., Liu, T.-Y.: SPRoBERTa: protein embedding learning with local fragment modeling. Brief. Bioinform. 23, bbac401 (2022). https://doi.org/10.1093/bib/bbac401.
80. Yones, C., Stegmayer, G., Milone, D.H.: Genome-wide pre-miRNA discovery from few labeled examples. Bioinformatics. 34, 541–549 (2018). https://doi.org/10.1093/bioinformatics/btx612.
81. Novák, P., Neumann, P., Macas, J.: Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics. 11, 378 (2010). https://doi.org/10.1186/1471-2105-11-378.
82. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Regression Trees. Taylor & Francis (1984).
83. Quinlan, J.R.: Induction of decision trees. Mach. Learn. 1, 81–106 (1986). https://doi.org/10.1007/BF00116251.
84. A survey of decision tree classifier methodology | IEEE Journals & Magazine | IEEE Xplore, https://ieeexplore.ieee.org/document/97458, last accessed 2024/05/12.
85. A comparative analysis of methods for pruning decision trees | IEEE Journals & Magazine | IEEE Xplore, https://ieeexplore.ieee.org/document/589207, last accessed 2024/05/12.
86. MastersInDataScience.org: What Is a Decision Tree?, https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/.
87. Breiman, L.: Bagging predictors. Mach. Learn. 24, 123–140 (1996).

https://doi.org/10.1007/BF00058655.
88. Breiman, L.: Random Forests. Mach. Learn. 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.
89. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. 63, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1.
90. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Ann. Stat. 29, 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451.
91. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844 (1998). https://doi.org/10.1109/34.709601.
92. Louppe, G.: Understanding Random Forests: From Theory to Practice, http://arxiv.org/abs/1407.7502, (2015).
93. Narassiguin, A., Bibimoune, M., Elghazel, H., Aussem, A.: An extensive empirical comparison of ensemble learning methods for binary classification. Pattern Anal. Appl. 19, 1093–1128 (2016). https://doi.org/10.1007/s10044-016-0553-z.
94. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874 (2006). https://doi.org/10.1016/j.patrec.2005.10.010.
95. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437 (2009). https://doi.org/10.1016/j.ipm.2009.03.002.
96. He, H., Garcia, E.A.: Learning from Imbalanced Data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284 (2009). https://doi.org/10.1109/TKDE.2008.239.
97. Elkan, C.: The Foundations of Cost-Sensitive Learning.
98. Buckland, M., Gey, F.: The relationship between Recall and Precision. J. Am. Soc. Inf. Sci. 45, 12–19 (1994). https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L.
99. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30, 1145–1159 (1997). https://doi.org/10.1016/S0031-3203(96)00142-2.
100. Saito, T., Rehmsmeier, M.: The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 10, e0118432 (2015). https://doi.org/10.1371/journal.pone.0118432.
101. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta BBA - Protein Struct. 405, 442–451 (1975). https://doi.org/10.1016/0005-2795(75)90109-9.
102. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 21, 6 (2020). https://doi.org/10.1186/s12864-019-6413-7.
103. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., Moult, J.: Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins Struct. Funct. Bioinforma. 87, 1011–1020 (2019). https://doi.org/10.1002/prot.25823.
104. Heo, L., Feig, M.: High-accuracy protein structures by combining machine-learning with physics-based refinement. Proteins Struct. Funct. Bioinforma. 88, 637–642 (2020). https://doi.org/10.1002/prot.25847.
105. Feig, M., Mirjalili, V.: Protein structure refinement via molecular-dynamics simulations: What works and what does not? Proteins Struct. Funct. Bioinforma. 84, 282–292 (2016). https://doi.org/10.1002/prot.24871.
106. Dutagaci, B., Heo, L., Feig, M.: Structure refinement of membrane proteins via molecular dynamics simulations. Proteins Struct. Funct. Bioinforma. 86, 738–750 (2018). https://doi.org/10.1002/prot.25508.
107. Mirjalili, V., Feig, M.: Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. J. Chem. Theory Comput. 9, 1294–1303 (2013). https://doi.org/10.1021/ct300962x.
108. Burnley, B.T., Afonine, P.V., Adams, P.D., Gros, P.: Modelling dynamics in protein crystal structures by ensemble refinement. eLife. 1, e00311 (2012). https://doi.org/10.7554/eLife.00311.

109. Brunger, A.T., Adams, P.D.: Molecular Dynamics Applied to X-ray Structure Refinement. Acc. Chem. Res. 35, 404–412 (2002). https://doi.org/10.1021/ar010034r.
110. Kim, D.-G., Choi, Y., Kim, H.-S.: Epitopes of Protein Binders Are Related to the Structural Flexibility of a Target Protein Surface. J. Chem. Inf. Model. 61, 2099–2107 (2021). https://doi.org/10.1021/acs.jcim.0c01397.
111. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990). https://doi.org/10.1016/S0022-2836(05)80360-2.
112. Armstrong, D.R., Berrisford, J.M., Conroy, M.J., Gutmanas, A., Anyango, S., Choudhary, P., Clark, A.R., Dana, J.M., Deshpande, M., Dunlop, R., Gane, P., Gáborová, R., Gupta, D., Haslam, P., Koča, J., Mak, L., Mir, S., Mukhopadhyay, A., Nadzirin, N., Nair, S., Paysan-Lafosse, T., Pravda, L., Sehnal, D., Salih, O., Smart, O., Tolchard, J., Varadi, M., Svobodova-Vařeková, R., Zaki, H., Kleywegt, G.J., Velankar, S.: PDBe: improved findability of macromolecular structure data in the PDB. Nucleic Acids Res. gkz990 (2019). https://doi.org/10.1093/nar/gkz990.
113. Eastman, P., Swails, J., Chodera, J.D., McGibbon, R.T., Zhao, Y., Beauchamp, K.A., Wang, L.-P., Simmonett, A.C., Harrigan, M.P., Stern, C.D.: OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput. Biol. 13, e1005659 (2017).
114. The PyMOL Molecular Graphics System, http://www.pymol.org.
115. Šali, A., Blundell, T.L.: Comparative Protein Modelling by Satisfaction of Spatial Restraints. J. Mol. Biol. 234, 779–815 (1993). https://doi.org/10.1006/jmbi.1993.1626.
116. Humphrey, W., Dalke, A., Schulten, K.: VMD: Visual molecular dynamics. J. Mol. Graph. 14, 33–38 (1996). https://doi.org/10.1016/0263-7855(96)00018-5.
117. Phillips, J.C., Hardy, D.J., Maia, J.D.C., Stone, J.E., Ribeiro, J.V., Bernardi, R.C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M.C.R., Radak, B.K., Skeel, R.D., Singharoy, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., Kalé, L.V., Schulten, K., Chipot, C., Tajkhorshid, E.: Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. 153, 044130 (2020). https://doi.org/10.1063/5.0014475.
118. Lee, J., Cheng, X., Swails, J.M., Yeom, M.S., Eastman, P.K., Lemkul, J.A., Wei, S., Buckner, J., Jeong, J.C., Qi, Y., Jo, S., Pande, V.S., Case, D.A., Brooks, C.L., MacKerell, A.D., Klauda, J.B., Im, W.: CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. J. Chem. Theory Comput. 12, 405–413 (2016). https://doi.org/10.1021/acs.jctc.5b00935.
119. Hooft, R.W.W., Vriend, G., Sander, C., Abola, E.E.: Errors in protein structures. Nature. 381, 272–272 (1996). https://doi.org/10.1038/381272a0.
120. University of California, Institute for Genomics and Proteomics: SAVES: Protein Structure Analysis and Verification Server, https://saves.mbi.ucla.edu/.
121. Kramer, C.S.M., Koster, J., Haasnoot, G.W., Roelen, D.L., Claas, F.H.J., Heidt, S.: HLA-EMMA: A user-friendly tool to analyse HLA class I and class II compatibility on the amino acid level. HLA. 96, 43–51 (2020). https://doi.org/10.1111/tan.13883.
122. Varshney, A., Brooks, F.P., Wright, W.V.: Linearly Scalable Computation of Smooth Molecular Surfaces.
123. Tien, M.Z., Meyer, A.G., Sydykova, D.K., Spielman, S.J., Wilke, C.O.: Maximum Allowed Solvent Accessibilites of Residues in Proteins. PLOS ONE. 8, e80635 (2013). https://doi.org/10.1371/journal.pone.0080635.
124. Amaya-Ramirez, D., Lhotte, R., Devriese, M., Hays, C., Taupin, J.-L., Devignes, M.-D.: Reduced structural flexibility of eplet amino acids in HLA proteins. Presented at the ECCB 2022 21st European Conference on Computational Biology Planetary Health and Biodiversity September 12 (2022).
125. Norel, R., Petrey, D., Wolfson, H.J., Nussinov, R.: Examination of shape complementarity in docking of Unbound proteins. Proteins Struct. Funct. Bioinforma. 36, 307–317 (1999).

https://doi.org/10.1002/(SICI)1097-0134(19990815)36:3<307::AID-PROT5>3.0.CO;2-R.
126. Hu, M.-K.: Visual pattern recognition by moment invariants. IRE Trans. Inf. Theory. 8, 179–187 (1962). https://doi.org/10.1109/TIT.1962.1057692.
127. Venkatraman, V., Sael, L., Kihara, D.: Potential for Protein Surface Shape Analysis Using Spherical Harmonics and 3D Zernike Descriptors. Cell Biochem. Biophys. 54, 23–32 (2009). https://doi.org/10.1007/s12013-009-9051-x.
128. Diffraction Theory of the Knife-Edge Test and its Improved Form, The Phase-Contrast Method | Monthly Notices of the Royal Astronomical Society | Oxford Academic, https://academic.oup.com/mnras/article/94/5/377/1084142, last accessed 2024/05/09.
129. Analysis, S.C. on I.: Proceedings of the 11th Scandinavian Conference on Image Analysis, Kangerlusssuaq, Greenland, June 7-11, 1999: Oral presentations. Pattern Recognition Society of Denmark (1999).
130. Shape retrieval using 3D Zernike descriptors - ScienceDirect, https://www.sciencedirect.com/science/article/abs/pii/S0010448504000077?via%3Dihub, last accessed 2024/05/09.
131. Efficient 3D Geometric and Zernike Moments Computation from Unstructured Surface Meshes | IEEE Journals & Magazine | IEEE Xplore, https://ieeexplore.ieee.org/document/5551145, last accessed 2024/05/09.
132. Sael, L., Li, B., La, D., Fang, Y., Ramani, K., Rustamov, R., Kihara, D.: Fast protein tertiary structure retrieval based on global surface shape similarity. Proteins Struct. Funct. Bioinforma. 72, 1259–1273 (2008). https://doi.org/10.1002/prot.22030.
133. Di Rienzo, L., Milanetti, E., Lepore, R., Olimpieri, P.P., Tramontano, A.: Superposition-free comparison and clustering of antibody binding sites: implications for the prediction of the nature of their antigen. Sci. Rep. 7, 45053 (2017). https://doi.org/10.1038/srep45053.
134. Protein-protein docking using region-based 3D Zernike descriptors | BMC Bioinformatics | Full Text, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-407, last accessed 2024/05/09.
135. Application of 3D Zernike descriptors to shape-based ligand similarity searching | Journal of Cheminformatics | Full Text, https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-1-19, last accessed 2024/05/09.
136. Flamary, R., Courty, N.: POT: Python Optimal Transport, https://pythonot.github.io/quickstart.html, last accessed 2024/05/17.
137. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis - ScienceDirect, https://www.sciencedirect.com/science/article/pii/0377042787901257?via%3Dihub, last accessed 2024/05/09.
138. Yang, P., Kuc, R.E., Brame, A.L., Dyson, A., Singer, M., Glen, R.C., Cheriyan, J., Wilkinson, I.B., Davenport, A.P., Maguire, J.J.: [Pyr1]Apelin-13(1–12) Is a Biologically Active ACE2 Metabolite of the Endogenous Cardiovascular Peptide [Pyr1]Apelin-13. Front. Neurosci. 11, (2017). https://doi.org/10.3389/fnins.2017.00092.
139. UCSF ChimeraX: Tools for structure building and analysis - Meng - 2023 - Protein Science - Wiley Online Library, https://onlinelibrary.wiley.com/doi/10.1002/pro.4792, last accessed 2024/05/09.
140. Klein, A., Ghosh, S.S., Bao, F.S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E.C., Keshavan, A.: Mindboggling morphometry of human brains. PLOS Comput. Biol. 13, e1005350 (2017). https://doi.org/10.1371/journal.pcbi.1005350.
141. Tharwat, A.: Classification assessment methods. Appl. Comput. Inform. 17, 168–192 (2020). https://doi.org/10.1016/j.aci.2018.08.003.
142. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
143. Shi, T., Horvath, S.: Unsupervised Learning With Random Forest Predictors. J.

Comput. Graph. Stat. 15, 118–138 (2006). https://doi.org/10.1198/106186006X94072.
144. Unsupervised extra trees: a stochastic approach to compute similarities in heterogeneous data | International Journal of Data Science and Analytics, https://link.springer.com/article/10.1007/s41060-020-00214-4, last accessed 2024/05/09.
145. Bush, D.B., Knotts, T.A., IV: Probing the effects of surface hydrophobicity and tether orientation on antibody-antigen binding. J. Chem. Phys. 146, 155103 (2017). https://doi.org/10.1063/1.4980083.
146. Qu, L., Qiao, X., Qi, F., Nishida, N., Hoshino, T.: Analysis of Binding Modes of Antigen–Antibody Complexes by Molecular Mechanics Calculation. J. Chem. Inf. Model. 61, 2396–2406 (2021). https://doi.org/10.1021/acs.jcim.1c00167.
147. Young, L., Jernigan, R. l., Covell, D. g.: A role for surface hydrophobicity in protein-protein recognition. Protein Sci. 3, 717–729 (1994). https://doi.org/10.1002/pro.5560030501.
148. Janin, J., Cherfils, J., Duquerroy, S.: Principles of Protein — Protein Recognition in Protease-Inhibitor and Antigen-Antibody Complexes. In: Soumpasis, D.M. and Jovin, T.M. (eds.) Computation of Biomolecular Structures. pp. 103–114. Springer, Berlin, Heidelberg (1993). https://doi.org/10.1007/978-3-642-77798-1_9.
149. Ponomarenko, J., Bui, H.-H., Li, W., Fusseder, N., Bourne, P.E., Sette, A., Peters, B.: ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics. 9, 514 (2008). https://doi.org/10.1186/1471-2105-9-514.
150. Niemann, M., Matern, B.M., Spierings, E.: Repeated local ellipsoid protrusion supplements HLA surface characterization. HLA. 103, e15260 (2024). https://doi.org/10.1111/tan.15260.
151. Sverrisson, F., Feydy, J., Correia, B.E., Bronstein, M.M.: Fast end-to-end learning on protein surfaces. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15267–15276. IEEE, Nashville, TN, USA (2021). https://doi.org/10.1109/CVPR46437.2021.01502.
152. Gonzalez-Galarza, F.F., McCabe, A., Santos, E.J.M. dos, Jones, J., Takeshita, L., Ortega-Rivera, N.D., Cid-Pavon, G.M.D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., Middleton, D., Jones, A.R.: Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res. 48, D783–D788 (2020). https://doi.org/10.1093/nar/gkz1029.
153. Rose, A.S., Bradley, A.R., Valasatava, Y., Duarte, J.M., Prlić, A., Rose, P.W.: NGL viewer: web-based molecular graphics for large complexes. Bioinformatics. 34, 3755–3758 (2018). https://doi.org/10.1093/bioinformatics/bty419.
154. MDDB: Molecular Dynamics Data Bank. The European Repository for Biosimulation Data, https://www.bsc.es/research-and-development/projects/mddb-molecular-dynamics-data-bank-the-european-repository, last accessed 2024/05/19.
155. Tiemann, J.K.S., Szczuka, M., Bouarroudj, L., Oussaren, M., Garcia, S., Howard, R.J., Delemotte, L., Lindahl, E., Baaden, M., Lindorff-Larsen, K., Chavent, M., Poulain, P.: MDverse: Shedding Light on the Dark Matter of Molecular Dynamics Simulations. eLife. 12, (2023). https://doi.org/10.7554/eLife.90061.1.

# Résumé étendu de la thèse en français

## Titre de la thèse : Approche de science des données pour l'exploration de l'antigénicité HLA basée sur les structures 3D et la dynamique moléculaire

## Introduction et Contexte

La transplantation d'organes demeure l'un des traitements les plus efficaces pour les patients souffrant d'une défaillance d'organe vital. En France, le nombre de greffes augmente chaque année, avec une majorité de transplantations rénales, hépatiques, cardiaques et pulmonaires. Cependant, la durée de vie des greffes reste limitée par la réponse allo-immune du receveur, qui se manifeste par la production d'anticorps spécifiques du donneur (DSA en anglais), dirigés principalement contre les antigènes HLA (Human Leukocyte Antigen en anglais). Ces anticorps peuvent entraîner un rejet aigu ou chronique de la greffe, ce qui souligne l'importance de la compatibilité entre donneur et receveur au niveau des antigènes HLA.

Le système HLA, hautement polymorphique, joue un rôle crucial dans la réponse immunitaire en présentant des peptides antigéniques aux lymphocytes T. La compatibilité entre le donneur et le receveur repose sur la minimisation des incompatibilités au niveau des antigènes HLA, afin de réduire le risque de rejet. Cependant, le système HLA présente une grande complexité, avec plus de 20 000 protéines HLA différentes identifiées, différant par au moins un acide aminé. Cette variabilité rend difficile la prédiction des réponses immunitaires de rejet de greffe et donc l'attribution donneur-receveur.

La présente thèse aborde la problématique de l'antigénicité des antigènes HLA dans le contexte de la transplantation d'organes, en adoptant une approche multidisciplinaire combinant la bioinformatique structurale et la science des données. L'objectif principal est de définir l'antigénicité des antigènes HLA du point de vue de la structure 3D de ces protéines, en utilisant des simulations de dynamique moléculaire (DM) pour explorer la dynamique des propriétés structurelles des antigènes HLA, pour mieux comprendre comment les variations de structure 3D influencent la reconnaissance des antigènes par les anticorps et pour développer un outil prédictif de l'antigénicité HLA basé sur l'apprentissage automatique.

Cette thèse se compose de 5 chapitres :

## Chapitre 1: Contexte et État de l'Art

Ce premier chapitre, qui regroupe les connaissances de base nécessaires à la compréhension de cette thèse, débute par une présentation des bases scientifiques et cliniques qui sous-tendent la transplantation d'organes et la compatibilité immunologique entre donneurs et receveurs. La transplantation d'organes implique des processus immunologiques complexes, dans lequel les antigènes HLA du donneur peuvent être reconnus comme étrangers par le système immunitaire du receveur, déclenchant ainsi une réponse allo-immune.

Le système HLA joue un rôle central dans la réponse allo-immune. Il est constitué de deux classes principales d'antigènes : les molécules HLA de classe I (HLA-A, -B, -C) qui sont exprimées par presque

toutes les cellules nucléées et les molécules HLA de classe II (HLA-DR, -DQ, -DP), exprimées principalement par les cellules présentatrices d'antigènes.

Actuellement, l'évaluation de la compatibilité donneur-receveur repose principalement sur la quantification des "mismatches" au niveau des séquences des antigènes HLA entre donneur et receveur. Ces "mismatches", également appelées éplets, peuvent être utilisés pour définir les épitopes HLA. Plus précisément, une épitope HLA est la région accessible au solvant d'un antigène HLA reconnue par un DSA et est généralement définie par la région englobant 15Å autour d'un eplet. Le terme "éplet vérifié par anticorps" (également appelé "éplet confirmé") est utilisé lorsqu'il existe des preuves expérimentales confirmant la capacité de l'épitope à déclencher une réponse immunitaire, tandis que les éplets putatifs (également appelés éplets non-confirmés) sont principalement déduits à partir d'alignements de séquences de protéines HLA et n'ont donc pas encore été confirmés.

Cependant, cette méthode pour définir la compatibilité donneur-receveur qu'à partir des séquences des antigènes ne prend pas en compte les structures 3D de ces protéines et leurs fluctuations dynamiques de ces molécules. Cette approche limitée nécessite une amélioration, car de nombreuses études montrent que les différences structurelles entre les antigènes du donneur et du receveur jouent un rôle majeur dans l'immunogénicité.

Ce chapitre aborde également les concepts fondamentaux de la bioinformatique structurale, une discipline qui permet de simuler et d'analyser les structures 3D des protéines. Cette discipline, couplée à des méthodes de science des données et d'apprentissage automatique, m'a permis de développer des outils prédictifs pour l'évaluation de l'antigénicité des antigènes HLA.

# Chapitre 2: Contribution des Simulations de Dynamique Moléculaire

Dans ce chapitre, l'étude des propriétés structurelles de 207 antigènes HLA issus des loci A, B, C, DP, DQ et DR est considérée en détail, avec une attention particulière portée à la comparaison entre les résultats obtenus à partir des trajectoires de DM et des structures statiques 3D, afin de mettre en évidence la contribution des premières.

Tout d'abord, les motivations pour l'utilisation des simulations de DM sont présentées. Par exemple, les simulations de DM permettent de corriger diverses erreurs stéréochimiques introduites par les méthodes expérimentales. De plus, les données de DM permettent l'analyse des propriétés dynamiques des structures. Les simulations de DM modélisent les molécules dans des conditions plus proches des conditions physiologiques, ce qui diffère des structures obtenues par des méthodes expérimentales telles que la cristallographie par rayons X, où les conditions de cristallisation modifient notamment l'environnement et l'organisation des atomes de la molécule. Ce point sera souligné tout au long des différentes propriétés structurelles étudiées dans ce chapitre, où dans certains cas, des écarts significatifs sont observés entre les propriétés calculées à partir des trajectoires de DM et des structures statiques 3D. Ces résultats permettent de mettre en lumière la contribution de la DM à l'étude des propriétés structurelles des antigènes HLA. Ensuite, le protocole mis en œuvre pour réaliser les simulations de DM est présenté en détail. Par la suite, les résultats de la vérification de la qualité des structures prédites/raffinées, utilisées comme point de départ pour les simulations MD, sont exposés. Ces résultats mettent en évidence l'utilité des simulations de DM pour corriger les erreurs stéréochimiques comme mentionné précédemment.

En ce qui concerne l'étude des propriétés structurelles des antigènes HLA, l'accessibilité au solvant est analysée en fonction de la position de chaque résidu (acide aminé), en prenant en compte plusieurs critères. Par exemple, différents types de positions ont été distingués, telles que les positions conservées, les positions polymorphes, les positions appartenant à des éplets confirmés, des éplets non-confirmés ou non répertoriés comme éplets (dites "non-éplet"). De même, les résultats issus des trajectoires de DM ou

des structures 3D statiques sont distingués et regroupés par locus ou par gène. Toutes ces analyses permettent une compréhension plus approfondie de la structure des antigènes HLA.

En ce qui concerne les éplets, il est constaté de manière inattendue que toutes les positions répertoriées comme appartenant à des éplets confirmés ne sont pas accessibles au solvant. Étant donné que le nombre de mismatches d'éplets confirmés est un facteur de risque connu pour le rejet de greffe, une perspective de ce travail serait d'évaluer si les mismatches d'éplets confirmés et accessibles au solvant constituent un meilleur indicateur du risque de rejet de greffe.

Il est également constaté que les loci A et C présentent le plus grand nombre de positions accessibles au solvant répertoriés comme éplets confirmés. Cependant, ce résultat pourrait être biaisé par le fait que les loci de classe I (c'est-à-dire les loci A, B et C) possèdent un plus grand nombre d'éplets confirmés.

De plus, les positions d'acides aminés (AA) les plus récurrentes et accessibles au solvant parmi les épitopes confirmés sont identifiées. Cela indiquerait une propension accrue à la reconnaissance par des anticorps à ces positions et pourrait donc être pris en compte lors de l'attribution des greffes.

En ce qui concerne la distribution des types d'AA et leurs propriétés physico-chimiques, les résultats sont nettement différents entre les groupes "Confirmé", "Non-confirmé" et "Non-éplet". Les différences observées dans le groupe "Non-confirmé" par rapport au groupe "Confirmé" indiquent que le premier groupe ne se compose pas uniquement de véritables candidats épitopes en attente de confirmation expérimentale, mais qu'il s'agit probablement d'un mélange de véritables et de faux éplets. Cette observation souligne le besoin d'une méthode informatique capable de distinguer, parmi les éplets non-confirmés, ceux qui pourraient constituer de véritables épitopes fonctionnels.

Enfin, la flexibilité des chaînes latérales des AA au cours des trajectoires de DM est étudiée, et il est constaté que les AA appartenant aux éplets confirmés ou aux patches "épitope" ont tendance à être moins flexibles que ceux considérés comme "non-éplets" ou appartenant à des patches "non-épitope" (un patch est la région accessible au solvant englobant 15Å autour d'un AA).

Toutes les propriétés structurelles explorées dans ce chapitre forment la base de la construction d'un ensemble de données pour l'apprentissage automatique, qui permettra le développement d'un prédicteur d'épitopes HLA capable de déterminer le statut des épitopes non-confirmés avant une validation expérimentale future.

# Chapitre 3: Descripteur de Forme 3D pour la Comparaison Dynamique des Surfaces Protéiques

Dans ce chapitre, l'utilisation de descripteurs de forme pour décrire les patches d'éplets est explorée, motivée par l'importance de la complémentarité de forme pour la reconnaissance des antigènes par des anticorps. Les objectifs principaux sont de comparer les surfaces protéiques et de développer un descripteur de forme pouvant être intégré dans un prédicteur d'épitopes HLA (voir Chapitre 4). Les moments de Zernike sont introduits en tant que descripteur de forme prometteur (nommé descripteur de Zernike), offrant des avantages tels que l'invariance sous rotation et translation, et une description compacte sous forme de vecteur de coefficients. La distance de Wasserstein, dérivée du problème du transport optimal (OT), est présentée comme une mesure de similarité entre les distributions de descripteurs de Zernike durant les trajectoires de DM. L'analyse de silhouette est utilisée comme une méthode permettant d'évaluer la qualité des résultats de clustering d'un ensemble de descripteurs de Zernike afin de déterminer le nombre optimal de clusters. En effet, le coefficient de silhouette mesure si un point (un descripteur de Zernike) s'intègre bien dans son cluster assigné par rapport aux autres clusters.

Deux études de cas ont été présentées. La première porte sur l'identification de conformères durant les trajectoires de DM. Elle est appliquée à un peptide, [Pyr1]apelin-13, en utilisant les descripteurs de Zernike, pendant une simulation de DM de 100ns. Les descripteurs de Zernike sont calculés pour 1000 structures afin de représenter les conformations 3D du peptide au cours de la trajectoire de DM. Le

clustering par K-means et l'analyse de silhouette révèle l'existence de deux conformères principaux. La deuxième étude de cas implique l'utilisation des descripteurs de Zernike pour comparer la forme des patches centrés sur les positions 80, 82, 83, 138 et 226 au cours des trajectoires de MD de 8 antigènes dans le contexte des groupes sérologiques BW4/BW6 (4 antigènes du groupe BW4 et 4 antigènes du groupe BW6), avec la distance de Wasserstein utilisée pour évaluer la distance entre les distributions des descripteurs de Zernike qui représentent les trajectoires de DM des 8 antigènes pour chaque position étudiée. Bien que l'étude de cas sur l'identification des conformères du peptide [Pyr1]apelin-13 ait démontré avec succès le potentiel des descripteurs de Zernike pour déterminer automatiquement le nombre optimal de clusters/conformères à partir des simulations de DM, la comparaison des patches des groupes sérologiques BW4/BW6 n'a pas donné la différenciation attendue entre les groupes, malgré des différences visuelles dans les patches à la position 82. Des incohérences ont également été observées aux positions 80, 83, ainsi qu'aux positions de contrôle (138 et 226), attribuées aux défis rencontrés lors de la génération de surfaces/maillages 3D cohérents et à l'insuffisance potentielle de la seule forme 3D pour expliquer les groupes sérologiques. Les recherches futures pourraient explorer des techniques alternatives de génération de surfaces/maillages pour les patches ou examiner différents descripteurs capables de capturer les propriétés physico-chimiques afin d'améliorer la caractérisation des éplets/épitopes HLA.

# Chapitre 4: HLA-EpiCheck

La reconnaissance des protéines HLA par des DSA est la principale cause de perte de greffe d'organes. Prédire l'antigénicité des épitopes HLA, dont les composants clés sont les éplets, pourrait améliorer la compatibilité HLA donneur-receveur et réduire la formation de DSA *de novo* ainsi que le rejet de greffe. HLA-EpiCheck, un prédicteur basé sur l'apprentissage automatique pour les épitopes HLA, a été développé en s'appuyant à la fois sur des descripteurs moléculaires statiques et dynamiques dérivés des simulations de DM.

HLA-EpiCheck est un classifieur binaire qui prédit si un patch, centré sur un AA accessible au solvant, est un épitope ou non. Le prédicteur a été entraîné en utilisant 18 descripteurs moléculaires, à la fois statiques (par exemple, hydrophobicité, charges électrostatiques) et dynamiques (par exemple, surface accessible au solvant relative, flexibilité des chaînes latérales), calculés pour 2117 patches Epitopes (centrés sur un AA appartenant à un éplet confirmé) et 4769 patches Non-épitopes (centrés sur un AA non-eplet) appartenant aux 207 antigènes HLA étudiés au Chapitre 2. Deux ensembles de données pour l'apprentissage automatique ont été générés : "Non-redondant" avec une réduction de la redondance des séquences (sous un seuil de 90% d'identité) à l'aide de MMseqs2, et "Redondant" en utilisant l'ensemble complet des données. Divers algorithmes d'apprentissage à base d'arbres, tels que l'arbre de décision (DT), les forêts aléatoires (RF), les forêts d'arbres extrêmement aléatoires (ExtraTrees), les forêts d'arbres avec gradient (GradientTrees), les forêts d'arbres avec histogradient (HistoGradientTrees, les K plus proches voisins (KNN) et le perceptron multicouche (MLP) ont été mis en oeuvre en utilisant une recherche par grille (en anglais 'gridsearch') des meilleurs paramètres et une méthode de validation croisée pour la mesure de la performance. La performance a été évaluée à l'aide des métriques de précision, rappel, F1-score et MCC.

L'algorithme "Extremely Randomized Trees" (ExtraTrees), entraîné avec l'ensemble de données "Redondant", a montré la meilleure performance et a été choisi pour la version finale de HLA-EpiCheck. En effet, la redondance n'est pas nuisible dans notre étude car HLA-EpiCheck est dédié à fonctionner uniquement sur les antigènes HLA. De plus, la complexité du modèle HLA-EpiCheck suggère que ce classifieur opère davantage sur un paradigme d'apprentissage basé sur les instances plutôt que sur un apprentissage basé sur un modèle. Le prédicteur HLA-EpiCheck a obtenu une haute précision et un bon rappel pour la prédiction des épitopes, à la fois lors de l'entraînement (0,92 et 0,83, respectivement) et lors des évaluations de performance (0,93 et 0,82, respectivement), ce qui a conduit à un F1-score de 0,87. HLA-EpiCheck a surpassé DiscoTope-3.0, un outil de pointe pour la prédiction des épitopes, avec une valeur de MCC (Coefficient de Corrélation de Matthews) de 0,83 contre 0,35. L'analyse de la diminution

moyenne des impuretés (MDI) a révélé que les descripteurs statiques liés à l'hydrophobicité et les descripteurs dynamiques reflétant la flexibilité des chaînes latérales étaient les contributeurs les plus importants à la performance de HLA-EpiCheck. Le modèle DT, entraîné avec l'ensemble de données "Non-redondant", a fourni des règles de classification interprétables et a mis en avant l'importance des descripteurs de flexibilité des chaînes latérales pour identifier les épitopes HLA.

Les scores HLA-EpiCheck ont été calculés pour des patches 3D couvrant 40 éplets non confirmés du locus DQ et comparés aux résultats expérimentaux pour 15 de ces éplets. 11 des 15 éplets validés expérimentalement avaient un score HLA-EpiCheck supérieur à 0,5, démontrant une remarquable cohérence entre la prédiction et les expériences en laboratoire. Une enquête plus approfondie sur les quatre éplets ayant de faibles scores a suggéré que les épitopes réels pourraient être centrés sur des résidus voisins de l'éplet signalé.

L'ensemble des résultats obtenus suggère que les prédictions de HLA-EpiCheck pourraient être utilisées pour les patches correspondant à des éplets non confirmés afin d'enrichir le calcul des mismatches entre les antigènes HLA du donneur et ceux du receveur (encore appelé MML pour 'MisMatch Load'). Jusqu'à présent le MML n'est calculé que pour les éplets confirmés. Des études rétrospectives devront être menées pour déterminer si la prise en compte des éplets non confirmés grâce à HLA-EpiCheck améliore la prédiction du rejet de greffe ou non.

# Chapitre 5: HLA-3D-Diff

Le cinquième chapitre de la thèse présente HLA-3D-Diff, un outil composé de deux interfaces graphiques ayant comme but faciliter la comparaison des antigènes HLA au niveau des séquences/éplets en s'appuyant sur certaines bases de données (HLA Eplet Registry, IPD-IMGT/HLA, Allele Frequency Net) et au niveau structurel en exploitant les structures 3D et les données de dynamique moléculaire générées dans cette thèse. HLA-3D-Diff offre des interfaces utilisateur intuitives permettant aux chercheurs et aux cliniciens de visualiser les antigènes HLA et de mieux comprendre certaines situations d'immunisation complexes ou inattendues.

# Conclusion et Perspectives

La présente thèse a permis de mieux comprendre l'antigénicité des antigènes HLA à travers une approche combinant bioinformatique structurale et science des données. Les résultats obtenus montrent que les différences de structure 3D entre les antigènes HLA du donneur et du receveur jouent un rôle crucial dans la réponse immunitaire post-transplantation. La flexibilité réduite des chaînes latérales associées aux épitopes confirmés, mise en évidence par les simulations de DM, souligne l'importance des mouvements moléculaires dans la reconnaissance antigénique.

L'un des apports majeurs de cette thèse est le développement de l'outil HLA-EpiCheck, qui offre une nouvelle méthode pour prédire l'antigénicité des épitopes HLA. Ce prédicteur a surpassé DiscoTope 3.0 et a montré une cohérence remarquable avec les résultats expérimentaux d'un ensemble d'éplets non-confirmés. Ces résultats suggèrent que HLA-EpiCheck pourrait être utilisé pour améliorer l'évaluation de la compatibilité entre donneurs et receveurs, et ainsi réduire le risque de rejet de greffe.

Les résultats obtenus dans cette thèse ont fourni des outils et des perspectives innovantes pour comprendre l'antigénicité des antigènes HLA. Cependant, plusieurs autres méthodes et approches méritent d'être explorées afin d'améliorer encore la précision et l'utilité des prédicteurs d'épitopes HLA. Par exemple, les travaux futurs pourraient être orientés vers le développement de nouveaux descripteurs dérivés des trajectoires de DM, permettant d'extraire de nouvelles caractéristiques des protéines. Cette thèse ouvre également des possibilités d'applications cliniques. La génération d'un score de MML basé sur les prédictions de HLA-EpiCheck peut être envisagée. La conduite d'études cliniques rétrospectives sera

nécessaire pour évaluer la pertinence de ce score pour la prédiction de la production de DSA et donc le pronostic du rejet de greffe, et ainsi confirmer l'utilisation de ce score dans la pratique clinique.

# Abstract

## Data science approach for the exploration of HLA antigenicity based on 3D structures and molecular dynamics

This thesis addresses the challenge of defining HLA antigenicity in organ transplantation through a multidisciplinary approach combining structural bioinformatics and data science. The primary objective was to explore HLA antigen antigenicity from a structural perspective, utilizing molecular dynamics (MD) simulations to study structural properties and develop a machine learning-based predictive tool for HLA antigenicity.

An extensive analysis of HLA antigen structural properties using MD simulations revealed significant insights. The simulations demonstrated marked differences in structural properties compared to static 3D structures, underscoring the importance of considering protein dynamics in antigenicity studies. Unexpectedly, it was found that not all positions listed as confirmed eplets were solvent-accessible, suggesting that solvent-accessible confirmed eplet mismatches might serve as a more accurate indicator of graft rejection risk. Furthermore, amino acids belonging to confirmed eplets or epitope patches exhibited less flexibility than those considered non-eplets or belonging to non-epitope patches, providing new perspectives on the structural basis of antigen recognition.

A major contribution of this research is the development of HLA-EpiCheck, a novel machine learning-based predictor for HLA epitopes. This binary classifier predicts whether a patch centered on a solvent-accessible amino acid is an epitope or not, utilizing 18 molecular descriptors derived from both static and dynamic properties. HLA-EpiCheck demonstrated impressive performance, achieving high precision (0.93) and good recall (0.82) for epitope prediction. Notably, it outperformed DiscoTope-3.0, a state-of-the-art tool, with a Matthews Correlation Coefficient of 0.83 compared to 0.35. Analysis of feature importance revealed that static descriptors related to hydrophobicity and dynamic descriptors reflecting side-chain flexibility were the most significant contributors to HLA-EpiCheck's performance. The tool's efficacy was further validated against experimental results for 15 non-confirmed eplets, showing remarkable consistency with 11 out of 15 experimentally validated eplets having an HLA-EpiCheck score above 0.5.

In conclusion, this thesis provides novel insights into HLA antigen antigenicity through an innovative combination of structural bioinformatics and data science approaches. The developed tool, HLA-EpiCheck, offers a new method for predicting HLA epitope antigenicity, potentially improving donor-recipient compatibility assessment and reducing graft rejection risk. Future work could focus on developing new descriptors derived from MD trajectories and conducting retrospective studies to evaluate HLA-EpiCheck's clinical relevance in predicting donor-specific antibody production and graft rejection prognosis. This research contributes significantly to the field of transplantation by offering innovative tools and perspectives for understanding HLA antigen antigenicity, with potential applications in improving organ transplant outcomes.

Keywords : Organ transplantation, HLA system, HLA antigenicity, molecular dynamics, molecular modeling, structural bioinformatics, data science, machine learning

# Résumé


## Approche de science des données pour l'exploration de l'antigénicité HLA basée sur les structures 3D et la dynamique moléculaire

Cette thèse aborde le défi de définir l'antigénicité HLA dans la transplantation d'organes à travers une approche multidisciplinaire combinant la bioinformatique structurale et la science des données. L'objectif principal était d'explorer l'antigénicité des antigènes HLA d'un point de vue structurel, en utilisant des simulations de dynamique moléculaire (DM) pour étudier les propriétés structurelles et développer un outil prédictif basé sur l'apprentissage automatique pour l'antigénicité HLA.

Une analyse approfondie des propriétés structurelles des antigènes HLA utilisant des simulations de DM a révélé des informations significatives. Les simulations ont démontré des différences marquées dans les propriétés structurelles par rapport aux structures 3D statiques, soulignant l'importance de considérer la dynamique des protéines dans les études d'antigénicité. De manière inattendue, il a été constaté que toutes les positions répertoriées comme éplets confirmés n'étaient pas accessibles au solvant, suggérant que les incompatibilités d'éplets confirmés accessibles au solvant pourraient servir d'indicateur plus précis du risque de rejet de greffe. De plus, les acides aminés appartenant aux éplets confirmés ou aux patches épitopes présentaient moins de flexibilité que ceux considérés comme non-éplets ou appartenant à des patches non-épitopes, offrant de nouvelles perspectives sur la base structurelle de la reconnaissance antigénique.

Une contribution majeure de cette recherche est le développement de HLA-EpiCheck, un nouveau prédicteur d'épitopes HLA basé sur l'apprentissage automatique. Ce classificateur binaire prédit si un patch centré sur un acide aminé accessible au solvant est un épitope ou non, en utilisant 18 descripteurs moléculaires dérivés de propriétés statiques et dynamiques. HLA-EpiCheck a démontré des performances impressionnantes, atteignant une haute précision (0,93) et un bon rappel (0,82) pour la prédiction d'épitopes. Notamment, il a surpassé DiscoTope-3.0, un outil de pointe, avec un coefficient de corrélation de Matthews de 0,83 contre 0,35. L'analyse de l'importance des caractéristiques a révélé que les descripteurs statiques liés à l'hydrophobicité et les descripteurs dynamiques reflétant la flexibilité des chaînes latérales étaient les contributeurs les plus significatifs à la performance de HLA-EpiCheck. L'efficacité de l'outil a été validée davantage par rapport aux résultats expérimentaux pour 15 éplets non confirmés, montrant une cohérence remarquable avec 11 des 15 éplets validés expérimentalement ayant un score HLA-EpiCheck supérieur à 0,5.

En conclusion, cette thèse fournit de nouvelles perspectives sur l'antigénicité des antigènes HLA grâce à une combinaison innovante d'approches de bioinformatique structurale et de science des données. L'outil développé, HLA-EpiCheck, offre une nouvelle méthode pour prédire l'antigénicité des épitopes HLA, améliorant potentiellement l'évaluation de la compatibilité donneur-receveur et réduisant le risque de rejet de greffe. Les travaux futurs pourraient se concentrer sur le développement de nouveaux descripteurs dérivés des trajectoires de DM et la conduite d'études rétrospectives pour évaluer la pertinence clinique de HLA-EpiCheck dans la prédiction de la production d'anticorps spécifiques du donneur et du pronostic de rejet de greffe. Cette recherche contribue significativement au domaine de la transplantation en offrant des outils et des perspectives innovants pour comprendre l'antigénicité des antigènes HLA, avec des applications potentielles dans l'amélioration des résultats de transplantation d'organes.

Mots-clés : Transplantation d'organes, système HLA, antigénicité HLA, dynamique moléculaire, modélisation moléculaire, bioinformatique structurale, science des données, apprentissage automatique