PNAS PERSPECTIVE OPEN ACCESS

A journey to your self: The vague definition of immune self and its practical implications

Balázs Koncz[a,b,c], Gergő Mihály Balogh[a,b,c], and Máté Manczinger[a,b,c,1]

Edited by Philippa Marrack, National Jewish Health, Denver, CO; received July 7, 2023; accepted April 1, 2024

The identification of immunogenic peptides has become essential in an increasing number of fields in immunology, ranging from tumor immunotherapy to vaccine development. The nature of the adaptive immune response is shaped by the similarity between foreign and self-protein sequences, a concept extensively applied in numerous studies. Can we precisely define the degree of similarity to self? Furthermore, do we accurately define immune self? In the current work, we aim to unravel the conceptual and mechanistic vagueness hindering the assessment of self-similarity. Accordingly, we demonstrate the remarkably low consistency among commonly employed measures and highlight potential avenues for future research.

immunogenicity | neoantigen | T cell repertoire | immune self | infectious diseases

Understanding the molecular properties that influence adaptive immune recognition is crucial across various medical domains, including cancer immunotherapy (1), infectious diseases (2), vaccine design (3, 4), allergy (5), and autoimmune diseases (6). Despite the revolutionary impact of cancer immunotherapy on treating cancer patients, the response to treatment exhibits considerable variability, largely contingent on the quality of mutated cancer peptides (7). Properly characterizing these peptides is key to personalizing treatment and enhancing its efficacy. Effective vaccine development during the COVID-19 pandemic played a crucial role in reducing casualties (8). The protein sequences guiding the adaptive immune response in these vaccines are pivotal for efficacy (9). Thoughtful selection of these sequences not only impacts response rates but also determines the durability and resistance to emerging variants. The prevalence of allergy and autoimmunity is sharply increasing in developed countries (10, 11). These disorders arise when the immune system erroneously attacks nonharmful or our own proteins (5, 6). Characterizing these peptides aids in understanding disease development and identifying triggering factors.

A specific property of peptides stands out as particularly important in adaptive immune recognition. Similarity to self has been proposed as a fundamental determinant of immune recognition (12–27). This concept has been actively utilized in various studies to identify peptides that elicit an effective immune response against mutated cancer peptides (12, 20, 21, 27). Mutations leading to peptides with low similarity to self-proteins are more likely to trigger a powerful immune response (1, 12). Tumors carrying many such mutations are more likely to be destroyed by the immune system, particularly under immune checkpoint blockade immunotherapy, emphasizing the importance of selecting such peptides for neoantigen vaccines (1, 12). In the context of vaccines against pathogens, the similarity of peptides to our self-proteins holds significance for two key reasons. First, pathogen-associated peptides resembling our self-proteins are less likely to be targeted by the immune system (19, 26). Second, if a destructive immune response does occur against these peptides, cross-reactivity can result in severe autoimmune side effects (28). Similarly, infections can induce autoimmune diseases if the pathogens carry sequences similar to self-proteins (6). Consequently, the similarity of pathogen-associated peptides to self-proteins has been extensively studied (6). While the concept of self-similarity is widely used in immunology, the straightforward definition of "self" remains elusive. This perspective article seeks to illustrate the intricate nature of immune self and the challenges in defining similarity to self. Additionally, we propose an approach to accurately define self-similarity. We find it important to note that our primary focus is on cell-mediated immunity, while humoral immunity is beyond the scope of this perspective.

A Brief History of Immune Self

The interpretation of the immune self has evolved significantly, experiencing substantial changes or even complete omission throughout its history (29)

> *"On several occasions already it has been noted that no ordinary component of the body will provoke an immunological response."*
>
> Frank M. Burnet, 1962 (30)

In 1949, Frank MacFarlane Burnet introduced the concept of immune self, and launched the self-nonself theory (31), which was largely inspired by transplantation experiments carried out by Medawar (32). The theory suggested that elements of the body are self and do not induce an immune response, while foreign molecules are nonself and, thus, immunogenic.

Author affiliations: [a]Synthetic and Systems Biology Unit, Institute of Biochemistry, Hungarian Research Network (HUN-REN) Biological Research Centre, Szeged 6726, Hungary; [b]Hungarian Centre of Excellence for Molecular Medicine - Biological Research Centre (HCEMM-BRC) Systems Immunology Research Group, Szeged 6726, Hungary; and [c]Department of Dermatology and Allergology, University of Szeged, Szeged 6720, Hungary

Author contributions: B.K., G.M.B., and M.M. designed research; B.K. and M.M. performed research; B.K. analyzed data; B.K. and M.M. acquired funding; and B.K., G.M.B., and M.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: manczinger.mate@brc.hu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2309674121/-/DCSupplemental.

Published May 9, 2024.

Burnet suggested that the immune self is determined in the genetic material of the individual (33). Notably, the self-nonself theory inspired Burnet to elaborate his clonal selection theory. While the self-nonself model was unable to explain certain common phenomena, such as autoimmunity, feto-maternal tolerance, and tolerance to microbiota in the skin, gut, and other mucosal surfaces (34), it still dominates the thinking of many immunologists.

> *"...the immune system, even in the absence of antigens that do not belong to the system, must display an eigen-behaviour..."*
>
> Nils K. Jerne, 1974 (35)

In 1974, Nils K. Jerne proposed a groundbreaking idea (35). He suggested that our immune system is inherently autoreactive (it displays an eigen-behavior). This stems from the assumption that every antibody acts as an antigen, leading to the generation of specific antibodies against them. These antibody-specific antibodies, in turn, are antigenic and trigger the formation of more antibodies, creating a continuous cycle. According to Jerne, autoreactivity is a normal part of the immune system's functioning. Jerne's theory served as inspiration for many immunologists and played a crucial role in shaping other theories, such as the now-abandoned autopoiesis theory (36) or subsequent works by Irun Cohen and Henri Atlan (37). These later theories consider autoimmunity as a normal phenomenon, but their originality and validity remain subjects of debate (38). Nevertheless, since Jerne's proposal, it has been repeatedly demonstrated that autoreactivity in the immune system is common and normal (39), which challenges the self-nonself theory in its original form.

> *"...the immune system does not care about self and non-self, that its primary driving force is the need to detect and protect against danger..."*
>
> Polly Matzinger, 1994 (40)

Polly Matzinger's danger theory dismissed the concept of self-nonself discrimination. Matzinger proposed that immunological decisions are orchestrated at the level of antigen-presenting cells (APCs), and an immune response is triggered when danger signals are detected (40). These danger signals consist of evolutionarily conserved molecules such as lipopolysaccharide, which have receptors on the surface of APCs. Matzinger's theory effectively explained tolerance to commensal bacteria and feto-maternal tolerance, where the fetus and microbiota do not cause harm. Initially met with enthusiasm, the danger theory also faced criticism later on. Some critiques focused on the imprecise definition of danger and raised questions about its originality (41, 42).

Although the concept of immune self has undergone significant transformations, it continues to be a prevailing perspective among immunologists (12, 17, 43, 44). The definition of self has been refined over time as our understanding of how adaptive immune recognition functions has grown. For instance, Waldmann and his colleagues proposed that only a subset of peptides presented by MHC molecules should be considered as self from the perspective of T lymphocytes (45). Mitchison suggested that a protein could be deemed self only if it surpasses a certain concentration threshold within the body (46). Building on the two-signal theory, which posits that lymphocyte activation requires at least two signals, Janeway proposed in 1989 that the immune system differentiates between infectious nonself, where both signals are present, and noninfectious self (47). According to Pradeau, immunologists hold at least five different meanings of self (48). Intriguingly, in a significant portion of current studies, the "genetic" self proposed by Burnet is employed to evaluate self-similarity (see later).

Adaptive Immune Recognition and T Cell Response

To arrive at the most accurate definition of immune self, it is necessary to provide a concise overview of how adaptive immune recognition operates.

The immunological synapse (Fig. 1*A*) is formed between antigen-presenting cells and T cells (49). This is the fundamental unit of adaptive immune recognition and consists of the major histocompatibility complex (MHC) molecules, the presented peptide, and the T cell receptor (TCR) (49, 50). Concurrently, other proteins like costimulatory and adhesion molecules participate in the intricate interactions between the antigen-presenting cell and T cell. Note that in humans, MHC molecules are referred to as human leukocyte antigen (HLA) molecules. We will use the term HLA throughout our work, except when specifically discussing studies that focus on nonhuman MHC molecules or MHC molecules in general.

HLA molecules present short peptides on the surface of cells (52). Their allelic diversity ranks among the highest within the human genome (53). The different variants display substantial disparities in amino acid specificities (54–57). The HLA molecules form two major classes. HLA-I molecules are found on all nucleated cells and typically present peptides ranging from 8 to 12 amino acids in length (58). These peptides originate from the host cell, where endogenous, viral, and bacterial proteins undergo proteasomal cleavage (59). The resulting peptides are then loaded onto HLA-I molecules by the peptide loading complex within the endoplasmic reticulum (60). HLA-II molecules are exclusively expressed on professional antigen-presenting cells, such as dendritic cells, B cells, and macrophages, and bind longer peptides ranging between 12 and 30 amino acids in length (61, 62). These peptides derive from extracellular proteins, which can be either harmless or associated with pathogens. They are internalized via endocytosis and degraded in lysosomes (63). HLA-I and HLA-II-presented peptides are recognized by CD4+ and CD8+ T cells, respectively (52).

In the sequence of the presented peptides, anchor residues primarily govern HLA binding, while TCR-contact residues predominantly establish chemical interactions with the TCR (64). In the case of an HLA-I-bound 9-mer peptide, primary anchor amino acids are frequently found at the second and ninth positions, while amino acids between the fourth and eighth positions constitute the T cell exposed motif (65–67) (Fig. 1*A*). It is important to emphasize the allele-specific nature of these positions, noting the presence of secondary anchor residues at the third, fifth, and sixth positions in peptides bound by certain HLA-I alleles (68). In the case of HLA-II-bound peptides, a core segment of nine amino acids fits

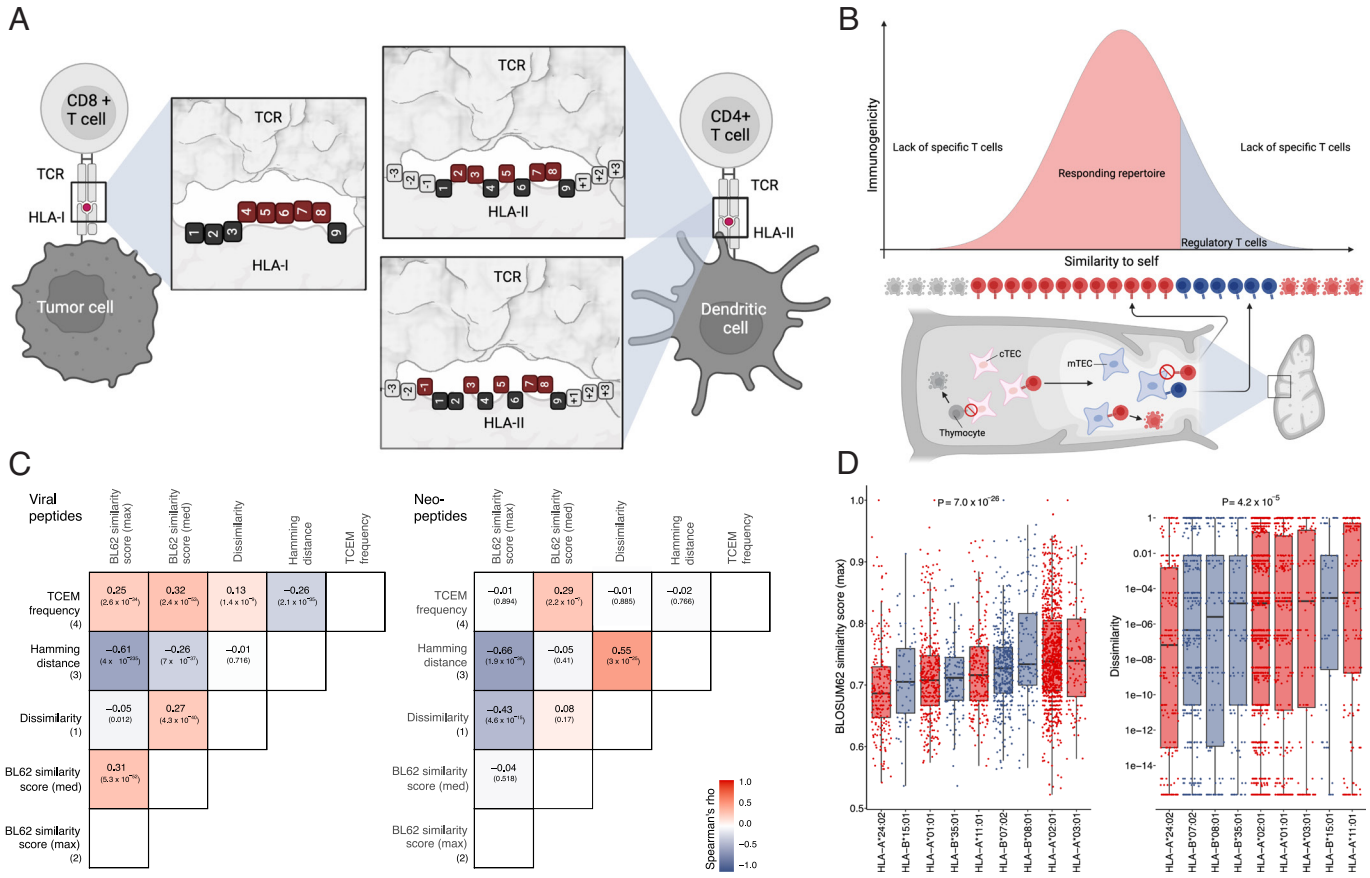

Fig. 1. The immunological synapse, similarity to the immune self, and inconsistency in self-similarity measures. (*A*) The immunological synapse and the positions of T cell exposed amino acids of HLA-presented peptides. For HLA-II molecules, the positions are indicated relative to the bound core sequence. The figures of TCEMs were created based on ref. 51. (*B*) The relationship between peptide similarity to self and immunogenicity. The underlying thymic processes defining this relationship are depicted in the schematic figure (see text for detailed explanation). (*C*) Spearman's correlation coefficients and the level of significance are indicated between different self-similarity measures. The strength of correlations is shown color-coded. The numbers below the name of measures indicate their index in Table 1. The calculations were performed separately for viral peptides (*Left*) and neopeptides (*Right*). (*D*) Systematic differences were found in BLOSUM62 similarity (*Left*) and dissimilarity (*Right*) when calculated for different HLA alleles. Only alleles with a minimum of five reported peptides in the IEDB are shown. Dissimilarity values are presented on a $\log_{10}$-transformed vertical axis (the minimal nonzero value was added to all values to handle the presence of zero dissimilarity values). *P*-values from Kruskal–Wallis tests are shown.

within the peptide binding groove of the HLA molecule, with the N- and C-terminal parts of the peptides overhanging (Fig. 1*A*). The positions of the anchoring and contacting amino acids are often variable and challenging to be identified (69). Furthermore, in the case of both HLA classes, the peptide sequence alone is insufficient to fully explain the interaction with the TCR; the conformation of the peptide must also be taken into account (70, 71).

While the presentation of a peptide by HLA is crucial for eliciting an immune response, its nature is also influenced by the binding strength and the stability of the peptide–HLA complex (54, 72). The complex is bound by the TCR, a heterodimeric protein composed of either an α and a β or a γ and a δ chain (73, 74). Importantly, HLA-presented peptides are bound by αβ TCRs on αβ T cells. Both chains consist of constant and variable domains. The variable domain includes the highly variable complementarity-determining region 3 (CDR3), which plays a crucial role in recognizing the presented peptide, while CDR1 and CDR2 make contact with the HLA molecule (75).

The strength of interaction between TCRs and peptide–HLA complexes has been demonstrated to influence T cell differentiation, both within the thymus and in peripheral tissues (76, 77). On the other hand, TCR avidity explains the strength of multiple interactions between peptide–HLA complexes and TCRs and considers the effect of positive and negative costimulatory molecules (78). Among the most well-known costimulatory receptors on T cells are CD28 and its related family members. CD28 plays a pivotal role in promoting T cell proliferation and cytokine production (79). The interaction between CD28 and its ligands (CD80, CD86) activates a phosphorylation cascade, culminating in a complex cellular response. To enhance T cell activation, antigen-presenting cells up-regulate the expression of CD28 ligands upon encountering microorganisms (80). Numerous other surface receptors, such as CD2, CD5, and ICOS, among others, have also been identified as having costimulatory functions (78). In addition to positive regulatory mechanisms, inhibitory molecules play a crucial role in modulating the immune response. Notably, CTLA4 and PD1, extensively studied in this context, serve as key "checkpoints" to mitigate T cell hyperactivation (81). These molecules represent primary targets for immune checkpoint inhibitors employed in cancer immunotherapy (81).

The initiation of TCR signaling in response to antigenic stimulation constitutes a multifaceted intracellular signaling pathway (82). Signals derived from the TCR induce rearrangements in the actin cytoskeleton, a process deemed crucial for the functional capacity of T cells (83). These signals also modulate

the expression of genes that are essential for the effector functions of T cells, including proliferation, cytokine secretion, and cytotoxicity (84). Following exposure to antigens and concurrent costimulation, naïve T cells undergo proliferation and differentiation, leading to the generation of specific effector cell populations. CD4+ cells can differentiate into distinct effector cell types, namely T-helper (Th) 1, Th2, Th9, and Th17 cells, depending on the cytokine milieu in the environment (85). Upon encountering their specific antigen, naive CD8+ T cells undergo clonal expansion, a process characterized by rapid and extensive multiplication (86). This is pivotal for the generation of an adequate reservoir of effector T cells, which is essential for the elimination of infected cells. After pathogen clearance, a substantial proportion of activated CD4+ and CD8+ T cells undergo apoptosis, marking a phase of contraction (87). Simultaneously, a subset of effector cells transitions into memory T cells, contributing to long-term immunological memory (88, 89).

T cells are also able to respond to antigens with tolerance. Regulatory T cells (Treg cells), comprising approximately 10% of peripheral CD4+ T cells, play a vital role in maintaining tolerance to harmless antigens and preventing autoimmune diseases (90, 91). These cells express CD4, CD25, and FOXP3. While a substantial proportion of these cells originate from the thymus, the remaining subset undergoes differentiation following exposure to harmless agonist antigens in the periphery (92, 93). CD8+ T cells can exist in four hyporesponsive states: tolerance, ignorance, anergy, and exhaustion (94). In the tolerant state, encountering the specific antigen does not induce activation but rather leads to the apoptosis of the T cell (95) or initiates a cell-intrinsic tolerance program (96). Remarkably, a subset of CD8+ T cells, known as CD8+ Treg cells, plays an active role in executing immunosuppressive functions (97).

In summary, the quality of the T cell response is collectively influenced by various factors, including the interaction between the peptide and the HLA, the sequence and 3D structure of the presented peptide, TCR affinity and avidity, the involvement of costimulatory and inhibitory receptors, the dosage of the antigen, and the cytokine milieu. Moreover, T cells make collective and not individual decisions, which is contingent on quorum sensing and mediated by cytokines received from other T cells in the surrounding environment (98).

The Formation of the T Cell Repertoire

It is a fundamental concept in immunology, that adaptive immune response is triggered exclusively by peptides, for which specific T cells exist in the repertoire (99). A comprehensive understanding of T cell repertoire development clarifies why the similarity to self-proteins plays a crucial role in shaping the nature of the immune response.

The formation of the T cell repertoire takes place in the thymus and involves two primary steps. Initially, lymphoid progenitor cells called thymocytes undergo positive selection, which is mediated by self-peptides presented by HLA molecules on the surface of cortical thymic epithelial cells (cTECs) (76). Thymocytes that fail to bind any peptide–HLA complex on cTECs undergo apoptosis due to neglect (76). Subsequently, thymocytes undergo negative selection if they strongly bind self-peptides presented by HLA molecules on medullary thymic epithelial cells (mTECs). These thymocytes are either eliminated or skewed toward alternative differentiation paths, giving rise to Tregs, which mediate immune tolerance (76, 100). Given the imperfect nature of negative selection, T cells that manage to survive this process may potentially exhibit self-reactivity. It is proposed that quorum sensing plays a crucial role in preventing autoimmunity, as the activation and proliferation of a particular T cell depend on a sufficient number of activated T cells in the surrounding environment (98). Notably, an increasing body of evidence supports the involvement of epitopes and metabolites derived from gut microbes in the intrathymic development of T cells (101).

The two-stage thymic selection process eliminates more than 95% of T cell precursors (102). Considering the mechanism of repertoire formation, one might expect the presence of "holes" in the T cell repertoire, given the elimination of cells binding self-peptides with high affinity (19, 103, 104). However, T cells exhibit cross-reactivity, allowing them to bind a range of somewhat similar peptides, introducing a level of controversy (105, 106). Beyond this issue, due to the positive and negative selection, the nature of immune response strongly depends on the similarity of presented peptides to self-proteins (Fig. 1*B*). Central tolerance mechanisms reduce the likelihood of destructive immune response against peptides that are similar to self-proteins (14, 17, 107) (Fig. 1*B*). Additionally, positive selection contributes to a T cell repertoire biased toward peptides sharing a certain level of similarity with our self-proteins (43, 107, 108). Accordingly, peptides that are highly dissimilar to our self-proteins are less likely to provoke an immune response (14, 17, 109) (Fig. 1*B*). Supporting this, it was recently reported that point mutations in tumors often yield sequences uncommon in the human proteome and commensal microbes, suggesting a reduced immune recognition of these peptides (18).

It is important to underscore that while the development of the T cell repertoire implies the dependence of the immune response on self-similarity, this process is intricately shaped by various factors discussed in earlier sections. The interaction between the peptide–HLA complex and the TCR alone is insufficient for triggering an effective immune response, its quality is determined by the presence of costimulatory molecules on cells and cytokines in the environment (78–80). For instance, in the case of a commensal bacterium, even if the T cell repertoire encompasses specific effector T cells for many of its peptides, these T cells remain inactive in a healthy state due to the absence of inflammatory cytokines and positive costimulatory molecules (110). A similar scenario is observed in numerous cancer samples where, despite the potential expression of mutated self-peptides recognized by effector T cells in the repertoire, the presence of negative costimulatory molecules along with a tolerogenic microenvironment prevent their activation (111). Conversely, in autoimmune diseases, the inflammatory cytokine milieu and positive costimulation can induce destructive immune responses to self-peptides (112). In essence, the impact of self-similarity on the immune response is invariably context-dependent, necessitating careful consideration in studies.

The Definition of Immune Self

Considering that i) the prevalence of specific T cells in the repertoire is essential for responding to a given peptide and ii) the development of the T cell repertoire relies on peptides expressed in the thymus, aligning with the perspective

of many immunologists (48, 113), we define immune self based on peptide–HLA complexes in the thymus. Specifically, we focus on those parts of the complexes that come into contact with receptors on thymocytes, ultimately determining their fate. We also take into account the different roles of cTECs and mTECs in repertoire formation. Following this concept, we introduce various relevant factors that should be considered when identifying the molecules constituting self. As our primary focus is on cell-mediated immunity, we will now place greater emphasis on HLA-I molecules and CD8+ T cells.

Our proteome comprises a minimum of 20,000 proteins. These proteins can be further divided into approximately 10 million overlapping 9-mer peptides (114). Due to various factors, including low protein expression, absence of proteasomal cleavage, inadequate HLA-binding, peptide-loading, and others, roughly 90% of the peptides are not presented on the cell surface by HLA-I molecules (115). A similar scenario may be observed in the thymus, with some differences. First, thymoproteasomes in cTECs generate a special set of peptides for mediating CD8+ T cell positive selection (116–118). Thymoproteasomes exhibit a reduced propensity to cleave peptides after hydrophobic amino acids. Consequently, this unique enzymatic activity results in the production of peptides with weaker binding affinities to HLA molecules (119–121). Moreover, these peptide–HLA complexes are characterized by weaker binding interactions with TCRs (120). Second, protein expression is cell-type-specific in general (122). This specificity is particularly notable in mTECs, which flexibly mimic the expression of diverse peripheral tissues driven by AIRE molecules (123).

The presented peptides are contingent upon the specificity of HLA molecules carried by an individual. Even identical peptides presented by different HLA alleles can evoke distinct immune responses (68, 124). Additionally, TCRs make contact not only with the peptide, but also with the HLA (75). Consequently, the HLA genotype profoundly influences the development of the T cell repertoire, making the immune self specific to each individual (45, 125).

As demonstrated earlier (Fig. 1*A*), only certain segments of the presented peptide come into contact with TCRs, as the amino acids anchoring the peptide to HLA molecules remain concealed from TCRs (64). While the sequence of these contacting amino acids could be regarded as constituting the immune self, its sole reliance proves inadequate in certain instances. This limitation became apparent in studies focusing on heteroclitic peptides. For these peptides, altering the HLA-anchoring amino acids modified the structure of TCR-contacting residues, potentially leading to different immune responses mediated by distinct T cells (126). Although the same amino acid sequences were presented to the TCR, their conformation differs, resulting in altered immune recognition. Another limitation of solely considering the sequence of T cell-exposed amino acids is the oversight of the regions on HLA molecules that interact with the TCR (75).

The role of cTECs and mTECs is markedly different in shaping the immune self. Peptides presented on cTECs mediate positive selection, forming the foundation of a responding repertoire (Fig. 1*B*). On the other hand, those presented on mTECs mediate negative selection and constitute the core of the immune self. Based on this, the peptides most similar to self are those that bear the same T cell contacting segments as the ones presented on the surface of mTECs. At the other extreme are peptides that bear no resemblance to those presented on cTECs.

Finally, a significant complicating factor in defining the immune self is the cross-reactivity of T cells, referring to the capability of a given TCR to recognize multiple different peptide sequences. While the level of cross-reactivity is reported to be around 10^6 peptides/TCR (127, 128), it is important to note that these data are available for only a limited number of TCRs. Nonetheless, T cell cross-reactivity exerts a profound impact on the development of the T cell repertoire and, consequently, shapes the formation of the immune self (129). Additionally, it has the potential to modify the actual extent of holes created in the repertoire by thymic selection (105, 106).

The Evolutionary Constraints on Immune Self

Gaining insight into the evolution of the adaptive immune system helps us understand the factors that shaped immune responses based on the similarity to self-proteins. The adaptive immune recognition in mammals relies on an exceptionally diverse array of TCRs generated through somatic recombination (121, 130). This diversity was made possible by the emergence of the recombination-activating gene (RAG) transposon in jawed vertebrates around 500 Mya (121, 130). However, this extensive receptor repertoire posed the risk of self-reactivity (130). To mitigate this risk, thymocytes undergo a two-step quality control process (76, 130).

Positive selection favors general T cell functionality and orchestrates the formation of a functional repertoire (76, 107, 130). cTECs express the thymoproteasome, which is found only in jawed vertebrates suggesting its co-occurrence with adaptive immunity (121). The peptides generated by the thymoproteasome are thought to possess a more "foreign" character, potentially contributing to the development of a repertoire that is more specific to nonself (118). Thymocytes that survive positive selection go through negative selection, which deletes clones binding self-peptide–HLA complexes with high affinity (76). The process is mediated by mTECs expressing tissue-restricted antigens, which is controlled by the AIRE transcription factor. The emergence of AIRE during evolution correlated with the divergence of T cells (131), suggesting a strong interdependence of negative selection and T cell-mediated immunity.

HLA molecules have the highest genetic diversity, which is maintained by heterozygous advantage (i.e., being heterozygous is advantageous as the peptide-binding repertoire of two different alleles is larger) and negative frequency-dependent selection (i.e., pathogens adapt to avoid HLA-presentation by common alleles) (132). While HLA alleles exhibit variable peptide specificities, peptides derived from housekeeping proteins are typically bound with high affinity by HLA-I molecules (133). Housekeeping proteins are encoded by hyperconserved genes with abundant expression across various tissue types. Notably, their elevated expression is proposed to play a role in fostering self-tolerance within the thymus (133).

The interpersonal variability and the global pattern of HLA alleles are shaped by infectious diseases, which are proposed to exert significant selection pressure on the immune system (134). Certain common HLA variants exhibit evidence of robust positive selection by pathogens (135). Interestingly, these alleles often carry a risk for autoimmune disorders, suggesting an evolutionary trade-off between the two types of diseases (135) and significant transformations in the immune self for individuals harboring these alleles. Similarly, high pathogen diversity is suggested to drive the selection of HLA alleles with a broad peptide-binding repertoire (136). However, carrying these alleles is associated with a reduced capacity of the immune system to discriminate between self and mutated cancer peptides (57).

The impact of pathogen-driven selection pressure extends beyond HLA alleles and influences the diversity of HLA-presented peptides through various components of the antigen-presentation machinery (137). For instance, ERAP2 is responsible for trimming peptides for HLA-I molecules, which subsequently present them to CD8+ T cells (138). Recent findings indicate that genomic variants associated with increased ERAP2 expression were positively selected during the plague pandemic (139). Remarkably, like HLA alleles mentioned before, this high-expression variant is also linked to autoimmunity.

Finally, the immune self is potentially influenced by components of human cells that share significant sequence similarity with pathogen-associated proteins: endogenous retroviruses and mitochondria. Endogenous retroviruses integrated into the genome over the past 100 My (140). Remarkably, the epitopes encoded by these retroviruses still exhibit higher similarities with viral epitopes than human ones (141). They are abundantly expressed in mTECs in an AIRE-independent manner, potentially contributing to the negative selection of T cells (141). However, immune responses targeting endogenous retroviruses have been reported, suggesting incomplete tolerance formation (142, 143). Nevertheless, the resemblance of endogenous retroviruses to pathogenic viruses is proposed to contribute to autoimmune diseases through molecular mimicry (144, 145). Mitochondria are organelles originating from the integration of the endosymbiotic bacterium into the host cell (146). HLA-I molecules are likely to present peptides derived from mitochondrial proteins (147), and these peptides are suggested to be more immunogenic (148). While the contribution of these proteins to the establishment of self-tolerance remains unexplored, the potential cross-reactivity between the mitochondrial protein PDC-E2 and proteins of *Escherichia coli* may play a role in the pathogenesis of the autoimmune disease primary biliary cholangitis (6).

The Spatiotemporal Variability of Immune Self

In our definition of immune self, we traditionally perceive it as a static or stable concept. However, even with precise identification of molecular patterns guiding the formation of the T cell repertoire, the repertoire undergoes dynamic changes throughout our lives. Various factors, including aging, changing environments, diseases, and their treatments, contribute to this dynamism (149). It is well exemplified by specific T cells that are present in the repertoire during younger ages but diminish in older ages due to the decreasing diversity of the repertoire (150). On the other hand, T cells undergo transformation into induced Treg cells at the periphery throughout life, thereby altering the immune self (90). It is important to acknowledge that in current studies, accounting for the life-long variability of the immune self (which is specific for each individual) would be especially challenging, if not unfeasible.

Inconsistency in Self-Similarity among Studies

The accurate definition of the immune self presents significant challenges, as outlined in the previous sections. Despite the obvious vagueness of its definition, the concept of similarity to self is widely applied across various studies. Table 1 provides a nonexhaustive list of standard methods along with example use-cases for calculating self-similarity. It is notable that most approaches focus on the amino acid sequence of presented peptides, rather than the detailed structure of peptide–HLA-I complexes, as the latter is not yet feasible for large-scale studies (20). Additionally, the utilization of the BLOSUM similarity matrix is prevalent among these methods, which is based on the evolutionary divergence of amino acids and does not directly account for structural differences (21). Importantly, a majority of these methods treat the entire proteome as self, aligning with Burnet's concept of "genetic self," but this approach may carry inherent inaccuracies, as discussed in previous sections.

As the interaction between the peptide–HLA complex and TCRs is the primary determinant of cellular adaptive immune recognition, an important question is what to consider as contacting regions in an HLA-presented peptide. While most methods consider the entire peptide sequence, only one focuses specifically on the segments that potentially interact with T cells. Notably, the former approach is highlighted as a potential limitation in one of the original papers (19). In this context, it is critical to acknowledge the findings of structural studies which indicate that the distinction between TCR-facing and non-TCR-facing residues is not always clear and can also vary between different TCRs and HLA alleles (66, 68). Moreover, even primary anchor residues can influence peptide conformation (126).

The consistency of values derived from different methods has not been examined to date. To explore this, we calculated multiple measures of self-similarity using two distinct sets of peptides. The first set consisted of 2,261 viral peptides, while the second set comprised 301 neopeptides presented by various HLA alleles.

We observed weak to moderate correlations among the calculated measures for both the viral and the neopeptide datasets (Fig. 1*C*), which can be attributed to several factors. First, four of the measures focused on the entire peptide sequence, whereas only one considered the specific amino acids in contact with TCRs. In the former case, the similarity values may be affected by the amino acids at anchor positions, responsible for HLA binding and subject to variation among alleles. Indeed, the similarity values displayed systematic differences across peptides presented by different alleles (Fig. 1*D*). Second, even

Table 1. A collection of self-similarity measures

	Which positions?	To what does it compare?	Similarity measure	Calculation	Refs.	Focus
1	Whole peptide	Whole proteome	BLOSUM62 (binding energy)	All BLAST hits are involved	(12)	Neopeptide
2	Whole peptide	Whole proteome	BLOSUM62	Maximum	(13)	Microbiota
3	Whole peptide	Whole proteome	Number of mismatched amino acids (Hamming distance)	Minimum	(14–16)	Microbiota
4	TCEM (4 to 8)	Whole proteome	Exact match	Count	(17, 18)	Microbiota
5	Whole peptide?	Selected self-antigens	BLOSUM35	Maximum	(19)	Microbiota
6	Whole peptide	Original counterpart of a neopeptide	3D structure, crystallography, molecular dynamics simulation	See ref.	(20)	Neopeptide
7	Whole peptide	Original counterpart of a neopeptide	BLOSUM62	See ref.	(21)	Neopeptide
8	Whole peptide	A subset of the human proteome	BLOSUM62; binding energy	Number of peptides in the proteome above cutoff	(22)	Microbiota
9	Whole peptide	Retinal proteins	LALIGN	Overlap	(23)	Microbiota
10	Any 5-mer along the peptide	Whole proteome	Exact match	Count	(24, 25)	
11	Different N-mers	Whole proteome	Varying	Count	(26)	Microbiota
12	Whole peptide	Whole proteome	BLOSUM62	Minimum (distance)	(27)	Neopeptide

when two methods consider the complete peptide sequence, their approaches to calculating similarity differ. For instance, in the case of neopeptides, the dissimilarity value calculated in ref. 12 exhibited only a moderate correlation with the maximum BLOSUM62 similarity score proposed in ref. 13, and this relationship completely disappeared for viral peptides. Both measures utilized BLOSUM62 similarity to compare peptide sequences, but the former considered multiple similarity values, while the latter took into account only the highest value during calculation. In sum, the analysis revealed inconsistent correlations among measures, underscoring the complexity of assessing self-similarity.

Enhancing Similarity Measures

The limited agreement among existing self-similarity measures emphasizes the necessity for refining and standardizing the definition of self-similarity across different studies. What factors should be considered, and how should they be integrated to precisely determine the similarity to self-proteins?

First, the identification of HLA-I-presented peptides on the surface of cTECs and mTECs is crucial (Fig. 2*A*). The most accurate data on these peptides could currently be obtained through immunopeptidomics. However, such data for these cells are unavailable, necessitating the determination of peptide presentation by considering the entire proteome and accounting for specific factors. It is essential to leverage transcriptomic or proteomic data for these cells to select proteins expressed at sufficient levels (151). Notably, mTECs express approximately 95.9% of protein-coding genes, with a minimum level of 1 transcript per million (TPM) (123). Additionally, these cells express tissue-restricted antigens and numerous alternative splice variants, further expanding the pool of potential proteins to be considered. Furthermore, determining the proteasomal cleavage of these proteins is crucial. While prediction algorithms exist for estimating cleavage by constitutive and immunoproteasomes (152), the absence of a prediction method for thymoproteasomes complicates the accurate determination of the peptide pool from which HLA alleles can bind peptides in cTECs. Last, the use of HLA-binding prediction methods is required to predict which peptides have the potential to be presented on the surface of mTECs and cTECs.

Second, it is necessary to identify the segments of the presented peptides that come into contact with or face TCRs (Fig. 2*A*). Here, both a sequence-based and a conformation-based approach are viable. The sequence-based method has been extensively employed in previous studies (12–19, 21–27) but considers only the order of amino acids. It is also important to highlight that the amino acid positions exposed to TCRs may vary between HLA alleles (68), necessitating careful consideration. On the other hand, the conformation-based approach could leverage advanced AI-based algorithms such as AlphaFold, which has been utilized to ascertain the 3D structure of peptide–HLA–TCR complexes (71, 153). These methods offer several advantages by accounting for subtle variations in the structure of TCR-exposed amino acids and taking also the T cell contacting regions of the HLA molecules into account. They are becoming increasingly practical and could address numerous challenges highlighted in this article.

Finally, treating the definition of similarity to self as a classification problem is a viable approach (Fig. 2*B*). The data on molecules mediating positive selection (on cTECs) or constituting the immune self (on mTECs) can be leveraged to train contemporary machine learning models. Similarly, data on peptides found in pathogens but absent in the human

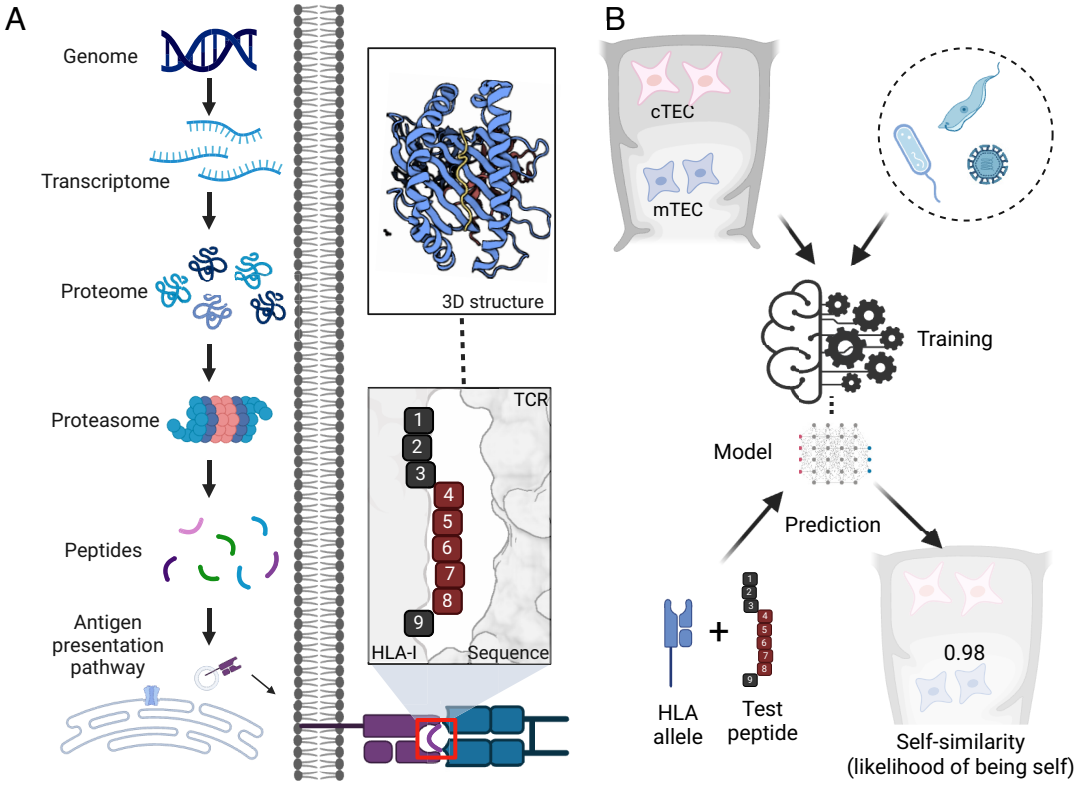

Fig. 2. An approach to enhance similarity measures. (*A*) Identification of the sequences and/or the 3D structures of peptide segments that come into contact with TCRs on the surface of thymic epithelial cells. (*B*) Utilizing machine learning to define the self-similarity of custom peptides (see text for a detailed explanation).

proteome can serve as another class. Notably, integrating T cell cross-reactivity into the models could significantly enhance their accuracy. A recent method has introduced an innovative approach to measure the distance between mutated cancer peptides and their original counterparts. The method approximates the ability of a TCR to discriminate between the original and the mutated peptide based on its cross-reactivity (154, 155). Moreover, it assigns different weights to peptide positions, considering their impact on cross-reactivity. Another study recently demonstrated accurate prediction of TCR reactivity within the sequence space around its agonist peptide (156). These advancements open avenues for systematically explaining TCR cross-reactivity and its integration into models predicting self-similarity in future studies. Such models could offer the probability of a given peptide belonging to a specific class (e.g., mTEC-presented peptides), with this probability interpreted as the degree of similarity to self. Crucially, these models should undergo rigorous cross-validation and testing on independent datasets to ensure accuracy. A highly precise model could potentially discern self-similarity at a personalized level by considering the HLA genotype.

In sum, the vague definition of immune self presents a considerable challenge in vaccine design and neoantigen identification, as similarity to immune self is a crucial factor to consider. The complexity of adaptive immune recognition and the vast range of potential peptide sequences make it difficult to accurately assess self-similarity using traditional methods. Nevertheless, ongoing technological advancements hold the potential to expedite the future development of precise and personalized measures.

Methods

Raw T cell assay results were downloaded from the Immune Epitope Database (157) as of April 11, 2023. Nine amino acid–long dengue virus and SARS-CoV-2 peptides were selected (n = 2,261). Nine amino acid–long neopeptides were obtained from ref. 1. The human proteome was downloaded from the UniProt database (51) (SwissProt proteins only, download date: April 11, 2023).

BLAST 2.12.0 (158) was utilized to determine the most similar sequences in the human proteome for each peptide. We employed ungapped alignment and set an E value cutoff of 10^8 to ensure a sufficient number of hits. The maximum number of target sequences was limited to 100. Subsequently, we estimated the similarity between all pairs of peptides, following the approach outlined in ref. 13. For each viral peptide and neopeptide, we determined the maximum and the median of the 100 similarity values. The Hamming distance was defined as the minimum number of differing amino acids between the given peptide and its most similar match in the human proteome.

The dissimilarity score was calculated according to the methodology published in ref. 12. We utilized the dissimilarity score function from antigen.garnish workflow (https://github.com/andrewrech/antigen.garnish) with keeping the k and a parameters unchanged. The TCEM frequency in the human proteome was determined following the procedure outlined in ref. 17.

Data, Materials, and Software Availability. Code and similarity data of peptides data have been deposited in Github (https://github.com/immunoteam/journey_to_your_self) (159). Previously published data were used for this work (https://www.iedb.org; https://doi.org/10.1016/j.cell.2020.09.015) (160). All other data are in the article and/or supporting information.

ACKNOWLEDGMENTS. The work was supported by the European Union's Horizon 2020 research and innovation program grant No. 739593 (B.K., G.M.B., and M.M.). This work was supported by Hungarian Scientific Research Fund grant FK-142312 (M.M.) and PD-146654 (B.K.). M.M. and B.K. were supported by the Bolyai János Research Fellowship of the Hungarian Academy of Sciences. M.M. and B.K. were supported by the ÚNKP-23-5, and G.M.B. was supported by the

ÚNKP-22-4 and ÚNKP-23-4 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101136582 Acronym ID-DarkMatter-NCD. Figs. 1 *A* and *B* and 2 were created with https://BioRender.com.

1. D. K. Wells *et al.*, Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* **183**, 818–834.e13 (2020).
2. A. Grifoni *et al.*, SARS-CoV-2 human T cell epitopes: Adaptive immune response against COVID-19. *Cell Host Microbe* **29**, 1076–1092 (2021).
3. M. Enayatkhani *et al.*, Reverse vaccinology approach to design a novel multi-epitope vaccine candidate against COVID-19: An in silico study. *J. Biomol. Struct. Dyn.* **39**, 2857–2872 (2021).
4. S. Parvizpour, M. M. Pourseif, J. Razmara, M. A. Rafi, Y. Omidi, Epitope-based vaccine design: A comprehensive overview of bioinformatics approaches. *Drug Discov. Today* **25**, 1034–1042 (2020).
5. H. Matsuo, T. Yokooji, T. Taogoshi, Common food allergens and their IgE-binding epitopes. *Allergol. Int. Off. J. Jpn. Soc. Allergol.* **64**, 332–343 (2015).
6. M. Rojas *et al.*, Molecular mimicry and autoimmunity. *J. Autoimmun.* **95**, 100–123 (2018).
7. N. McGranahan, C. Swanton, Neoantigen quality, not quantity. *Sci. Transl. Med.* **11**, eaax7918 (2019).
8. Q. Liu, C. Qin, M. Liu, J. Liu, Effectiveness and safety of SARS-CoV-2 vaccine in real-world studies: A systematic review and meta-analysis. *Infect. Dis. Poverty* **10**, 132 (2021).
9. E. Ong, M. U. Wong, A. Huffman, Y. He, COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. *Front. Immunol.* **11**, 1581 (2020).
10. L. Moroni, I. Bianchi, A. Lleo, Geoepidemiology, gender and autoimmune disease. *Autoimmun. Rev.* **11**, A386–A392 (2012).
11. R. J. Doll *et al.*, "Epidemiology of allergic diseases" in *Allergy and Asthma: The Basics to Best Practices*, M. Mahmoudi, Ed. (Springer International Publishing, 2019), pp. 31–51.
12. L. P. Richman, R. h. Vonderheide, A. J. Rech, Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade. *Cell Syst.* **9**, 375–382.e4 (2019).
13. A. Bresciani *et al.*, T-cell recognition is shaped by epitope sequence conservation in the host proteome and microbiome. *Immunology* **148**, 34–39 (2016).
14. A. Mayer, C. J. Russo, Q. Marcou, W. Bialek, B. D. Greenbaum, How different are self and nonself? arXiv [Preprint] (2022). https://arxiv.org/abs/2212.12049 (Accessed 30 January 2024).
15. D. Vergni, R. Gaudio, D. Santoni, The farther the better: Investigating how distance from human self affects the propensity of a peptide to be presented on cell surface by MHC class I molecules, the case of Trypanosoma cruzi. *PLoS One* **15**, e0243285 (2020).
16. D. Santoni, Viral peptides-MHC interaction: Binding probability and distance from human peptides. *J. Immunol. Methods* **459**, 35–43 (2018).
17. B. Koncz *et al.*, Self-mediated positive selection of T cells sets an obstacle to the recognition of nonself. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2100542118 (2021).
18. E. J. Homan, R. D. Bremel, Determinants of tumor immune evasion: The role of T cell exposed motif frequency and mutant amino acid exposure. *Front. Immunol.* **14**, 1155679 (2023).
19. S. Frankild, R. J. de Boer, O. Lund, M. Nielsen, C. Kesmir, Amino acid similarity accounts for T cell cross-reactivity and for "holes" in the T cell repertoire. *PLoS One* **3**, e1831 (2008).
20. J. R. Devlin *et al.*, Structural dissimilarity from self drives neoepitope escape from immune tolerance. *Nat. Chem. Biol.* **16**, 1269–1276 (2020).
21. A.-M. Bjerregaard *et al.*, An analysis of natural T cell responses to predicted tumor neoepitopes. *Front. Immunol.* **8**, 1566 (2017).
22. A. Gao *et al.*, Learning from HIV-1 to predict the immunogenicity of T cell epitopes in SARS-CoV-2. *iScience* **24**, 102311 (2021).
23. I. K. Karagöz *et al.*, Using bioinformatic protein sequence similarity to investigate if SARS CoV-2 infection could cause an ocular autoimmune inflammatory reactions? *Exp. Eye Res.* **203**, 108433 (2021).
24. D. Kanduc, Immunogenicity in peptide-immunotherapy: From self/nonself to similar/dissimilar sequences. *Adv. Exp. Med. Biol.* **640**, 198–207 (2008).
25. D. Kanduc, Homology, similarity, and identity in peptide epitope immunodefinition. *J. Pept. Sci. Off. Publ. Eur. Pept. Soc.* **18**, 487–494 (2012).
26. M. Rolland *et al.*, Recognition of HIV-1 peptides by host CTL is related to HIV-1 similarity to human proteins. *PloS One* **2**, e823 (2007).
27. J. Schmidt *et al.*, Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Rep. Med.* **2**, 100194 (2021).
28. Y. Chen *et al.*, New-onset autoimmune phenomena post-COVID-19 vaccination. *Immunology* **165**, 386–401 (2022).
29. T. Pradeu, "The self-nonself theory" in *The Limits of the Self: Immunology and Biological Identity* (Oxford University Press, 2011), pp. 49–84.
30. F. M. Burnet, "Self-recognition" in *The Integrity of the Body: A Discussion of Modern Immunological Ideas*, E. Mayr, K. V. Thimann, D. R. Griffin, Eds. (Harvard University Press, 1962), p. 68.
31. F. M. Burnet, F. Fenner, *The Production of Antibodies* (Macmillan, 1949).
32. P. B. Medawar, The behaviour and fate of skin autografts and skin homografts in rabbits: A report to the War Wounds Committee of the Medical Research Council. *J. Anat.* **78**, 176–199 (1944).
33. F. M. Burnet, "Historical outline" in *The Integrity of the Body: A Discussion of Modern Immunological Ideas*, E. Mayr, K. V. Thimann, D. R. Griffin, Eds., (Harvard University Press, 1962), p. 13.
34. T. Pradeu, "The unity of the individual: Self-nonself, autoimmunity, tolerance, and symbiosis" in *Philosophy of Immunology, The Phylosophy of Biology* (Cambridge University Press, 2020), pp. 15–20.
35. N. K. Jerne, Toward a network theory of immune system. *Ann Immunol. Paris* **125**, 373–389 (1974).
36. H. R. Maturana, F. J. Varela, *Autopoiesis and Cognition: The Realization of the Living* (Springer Science & Business Media, 1991).
37. I. R. Cohen, "Chapter 5--on Autoimmunity" in *Tending Adam's Garden*, I. R. Cohenstring-name>, Ed. (Academic Press, 2000), pp. 197–239.
38. T. Pradeu, "Comparing the continuity theory to other immunological theories" in *The Limits of the Self: Immunology and Biological Identity* (Oxford University Press, 2011), pp. 200–204.
39. T. Pradeu, "Critique of the self-nonself theory" in *The Limits of the Self: Immunology and Biological Identity* (Oxford University Press, 2011), pp. 85–94.
40. P. Matzinger, Tolerance, danger, and the extended family. *Annu. Rev. Immunol.* **12**, 991–1045 (1994).
41. T. Pradeu, "Comparing the continuity theory to other immunological theories" in *The Limits of the Self: Immunology and Biological Identity* (Oxford University Press, 2011), pp. 208–209.
42. A. M. Silverstein, Immunological tolerance. *Science* **272**, 1405–1405 (1996).
43. R. B. Fulton *et al.*, The TCR's sensitivity to self peptide–MHC dictates the ability of naive CD8(+) T cells to respond to foreign antigens. *Nat. Immunol.* **16**, 107–117 (2015).
44. H. Jiang, L. Chess, How the immune system achieves self-nonself discrimination during adaptive immunity. *Adv. Immunol.* **102**, 95–133 (2009).
45. H. Waldmann, S. Cobbold, R. Benjamin, S. Qin, A theoretical framework for self-tolerance and its relevance to therapy of autoimmune disease. *J. Autoimmun.* **1**, 623–629 (1988).
46. N. A. Mitchison, A walk round the edges of self tolerance. *Ann. Rheum. Dis.* **52**, S3–S5 (1993).
47. C. A. Janeway, Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb. Symp. Quant. Biol.* **54**, 1–13 (1989).
48. T. Pradeu, "Immunology, self and nonself" in *The Limits of the Self: Immunology and Biological Identity* (Oxford University Press, 2011), pp. 44–46.
49. W. E. Paul, R. A. Seder, Lymphocyte responses and cytokines. *Cell* **76**, 241–251 (1994).
50. M. L. Dustin, The immunological synapse. *Cancer Immunol. Res.* **2**, 1023–1033 (2014).
51. The UniProt Consortium, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
52. K. L. Rock, E. Reits, J. Neefjes, Present yourself! By MHC class I and MHC class II molecules. *Trends Immunol.* **37**, 724–737 (2016).
53. J. Robinson *et al.*, IPD-IMGT/HLA database. *Nucleic Acids Res.* **48**, D948–D955 (2020).
54. S. Paul *et al.*, HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *J. Immunol.* **191**, 5831–5839 (2013).
55. D. Gfeller, M. Bassani-Sternberg, Predicting antigen presentation-what could we learn from a million peptides? *Front. Immunol.* **9**, 1716 (2018).
56. J. Racle *et al.*, Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* **37**, 1283–1286 (2019).
57. M. Manczinger *et al.*, Negative trade-off between neoantigen repertoire breadth and the specificity of HLA-I molecules shapes antitumor immunity. *Nat. Cancer* **2**, 950–961 (2021).
58. D. Gfeller *et al.*, The length distribution and multiple specificity of naturally presented HLA-I ligands. *J. Immunol. Baltim. Md* **1950**, 3705–3716 (2018).
59. P. M. Kloetzel, F. Ossendorp, Proteasome and peptidase function in MHC-class-I-mediated antigen presentation. *Curr. Opin. Immunol.* **16**, 76–81 (2004).
60. A. Blees *et al.*, Structure of the human MHC-I peptide-loading complex. *Nature* **551**, 525–528 (2017).
61. J. G. Abelin *et al.*, Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity* **51**, 766–779.e17 (2019).
62. M. Stražar *et al.*, HLA-II immunopeptidome profiling and deep learning reveal features of antigenicity to inform antigen discovery. *Immunity* **56**, 1681–1698.e13 (2023).
63. J. Neefjes, M. L. M. Jongsma, P. Paul, O. Bakke, Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11**, 823–836 (2011).
64. J. J. A. Calis *et al.*, Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* **9**, e1003266 (2013).
65. R. D. Bremel, E. J. Homan, Frequency patterns of T-cell exposed amino acid motifs in immunoglobulin heavy chain peptides presented by MHCs. *Front. Immunol.* **5**, 541 (2014).
66. M. G. Rudolph, R. L. Stanfield, I. A. Wilson, How Tcrs bind Mhcs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24**, 419–466 (2006).
67. J. J. A. Calis, R. J. de Boer, C. Keşmir, Degenerate T-cell recognition of peptides on MHC molecules creates large holes in the T-cell repertoire. *PLoS Comput. Biol.* **8**, e1002412 (2012).
68. A. T. Nguyen, C. Szeto, S. Gras, The pockets guide to HLA class I molecules. *Biochem. Soc. Trans.* **49**, 2319–2331 (2021).
69. H. B. Taylor *et al.*, MS-based HLA-II peptidomics combined with multiomics will aid the development of future immunotherapies. *Mol. Cell. Proteomics MCP* **20**, 100116 (2021).
70. J. M. Custodio *et al.*, Structural and physical features that distinguish tumor-controlling from inactive cancer neoepitopes. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2312057120 (2023).
71. V. Mikhaylov *et al.*, Accurate modeling of peptide-MHC structures with AlphaFold. *Structure* **32**, 228–241.e4 (2024).
72. M. Rasmussen *et al.*, Pan-specific prediction of peptide–MHC class I complex stability, a correlate of T cell immunogenicity. *J. Immunol.* **197**, 1517–1524 (2016).
73. C. Chothia, D. R. Boswell, A. M. Lesk, The outline structure of the T-cell alpha beta receptor. *EMBO J.* **7**, 3745–3755 (1988).
74. D. h. Raulet, The structure, function, and molecular genetics of the gamma/delta T cell receptor. *Annu. Rev. Immunol.* **7**, 175–207 (1989).
75. K. C. Garcia, E. J. Adams, How the T cell receptor sees antigen--a structural view. *Cell* **122**, 333–336 (2005).
76. L. Klein, B. Kyewski, P. M. Allen, K. A. Hogquist, Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377–391 (2014).
77. K. A. Hogquist, S. C. Jameson, The self-obsession of T cells: How TCR signaling thresholds affect fate "decisions" and effector function. *Nat. Immunol.* **15**, 815–823 (2014).
78. L. Chen, D. B. Flies, Molecular mechanisms of T cell co-stimulation and co-inhibition. *Nat. Rev. Immunol.* **13**, 227–242 (2013).

79. O. Acuto, F. Michel, CD28-mediated co-stimulation: A quantitative support for TCR signalling. *Nat. Rev. Immunol.* **3**, 939–951 (2003).
80. A. h. Sharpe, G. J. Freeman, The B7–CD28 superfamily. *Nat. Rev. Immunol.* **2**, 116–126 (2002).
81. E. I. Buchbinder, A. Desai, CTLA-4 and PD-1 pathways: Similarities, differences, and implications of their inhibition. *Am. J. Clin. Oncol.* **39**, 98–106 (2016).
82. J. E. Smith-Garvin, G. A. Koretzky, M. S. Jordan, T cell activation. *Annu. Rev. Immunol.* **27**, 591–619 (2009).
83. J. K. Burkhardt, E. Carrizosa, M. h. Shaffer, The actin cytoskeleton in T cell activation. *Annu. Rev. Immunol.* **26**, 233–259 (2008).
84. A. Weiss, D. R. Littman, Signal transduction by lymphocyte antigen receptors. *Cell* **76**, 263–274 (1994).
85. J. Zhu, h. Yamane, W. E. Paul, Differentiation of effector CD4 T cell populations. *Annu. Rev. Immunol.* **28**, 445–489 (2010).
86. J. M. Curtsinger, C. M. Johnson, M. F. Mescher, CD8 T cell clonal expansion and development of effector function require prolonged exposure to antigen, costimulation, and signal 3 cytokine. *J. Immunol. Baltim. Md* **1950**, 5165–5171 (2003).
87. R. A. Seder, R. Ahmed, Similarities and differences in CD4+ and CD8+ effector and memory T cell generation. *Nat. Immunol.* **4**, 835–842 (2003).
88. J. S. Hale *et al.*, Distinct memory CD4+ T cells with commitment to T follicular helper- and T helper 1-cell lineages are generated after acute viral infection. *Immunity* **38**, 805–817 (2013).
89. B. J. Laidlaw, J. E. Craft, S. M. Kaech, The multifaceted role of CD4+ T cells in CD8+ T cell memory. *Nat. Rev. Immunol.* **16**, 102–111 (2016).
90. K. Wing, S. Sakaguchi, Regulatory T cells exert checks and balances on self tolerance and autoimmunity. *Nat. Immunol.* **11**, 7–13 (2010).
91. C.-S. Hsieh, h.-M. Lee, C.-W.J. Lio, Selection of regulatory T cells in the thymus. *Nat. Rev. Immunol.* **12**, 157–167 (2012).
92. L. Klein, E. A. Robey, C.-S. Hsieh, Central CD4+ T cell tolerance: Deletion versus regulatory T cell differentiation. *Nat. Rev. Immunol.* **19**, 7–18 (2019).
93. J. M. Weiss *et al.*, Neuropilin 1 is expressed on thymus-derived natural regulatory T cells, but not mucosa-generated induced Foxp3+ T reg cells. *J. Exp. Med.* **209**, 1723–1742, S1 (2012).
94. M. Philip, A. Schietinger, CD8+ T cell differentiation and dysfunction in cancer. *Nat. Rev. Immunol.* **22**, 209–223 (2022).
95. C. Kurts, h. Kosaka, F. R. Carbone, J. F. A. P. Miller, W. R. Heath, Class I–restricted cross-presentation of exogenous self-antigens leads to deletion of autoreactive CD8+ T cells. *J. Exp. Med.* **186**, 239–245 (1997).
96. A. Schietinger, J. J. Delrow, R. S. Basom, J. N. Blattman, P. D. Greenberg, Rescued tolerant CD8 T cells are preprogrammed to reestablish the tolerant state. *Science* **335**, 723–727 (2012).
97. S. Mishra, S. Srinivasan, C. Ma, N. Zhang, CD8+ regulatory T cell – A mystery to be revealed. *Front. Immunol.* **12**, 708874 (2021).
98. T. C. Butler, M. Kardar, A. K. Chakraborty, Quorum sensing allows T cells to discriminate between self and nonself. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11833–11838 (2013).
99. A. W. Goldrath, M. J. Bevan, Selecting and maintaining a diverse T-cell repertoire. *Nature* **402**, 255–262 (1999).
100. H. Takaba, h. Takayanagi, The mechanisms of T cell selection in the thymus. *Trends Immunol.* **38**, 805–816 (2017).
101. R. Hebbandi Nanjundappa, C. Sokke Umeshappa, M. B. Geuking, The impact of the gut microbiota on T cell ontogeny in the thymus. *Cell. Mol. Life Sci. CMLS* **79**, 221 (2022).
102. A. Yates, Theories and quantification of thymic selection. *Front. Immunol.* **5**, 13 (2014).
103. D. Vidović, P. Matzinger, Unresponsiveness to a foreign antigen can be caused by self-tolerance. *Nature* **336**, 222–225 (1988).
104. M. Wölfl *et al.*, Hepatitis C virus immune escape via exploitation of a hole in the T cell repertoire. *J. Immunol.* **181**, 6435–6446 (2008).
105. K. Ogasawara, W. L. Maloy, R. h. Schwartz, Failure to find holes in the T-cell repertoire. *Nature* **325**, 450–452 (1987).
106. W. Yu *et al.*, Clonal deletion prunes but does not eliminate self-specific αβ CD8(+) T lymphocytes. *Immunity* **42**, 929–941 (2015).
107. N. Vrisekoop, J. P. Monteiro, J. N. Mandl, R. N. Germain, Revisiting thymic positive selection and the mature T cell repertoire for antigen. *Immunity* **41**, 181–190 (2014).
108. J. N. Mandl, J. P. Monteiro, N. Vrisekoop, R. N. Germain, T cell-positive selection uses self-ligand binding strength to optimize repertoire recognition of foreign antigens. *Immunity* **38**, 263–274 (2013).
109. C. h. Lee *et al.*, A robust deep learning workflow to predict CD8+ T-cell epitopes. *Genome Med.* **15**, 70 (2023).
110. C. Sorini, R. F. Cardoso, N. Gagliani, E. J. Villablanca, Commensal bacteria-specific CD4+ T cell responses in health and disease. *Front. Immunol.* **9**, 2667 (2018).
111. X. Jiang *et al.*, Role of the tumor microenvironment in PD-L1/PD-1-mediated tumor immune escape. *Mol. Cancer* **18**, 10 (2019).
112. M. Wahren-Herlenius, T. Dörner, Immunopathogenic mechanisms of systemic autoimmune disease. *Lancet* **382**, 819–831 (2013).
113. P. Kourilsky, J. M. Claverie, The peptidic self model: A hypothesis on the molecular nature of the immunological self. *Ann. Inst. Pasteur Immunol.* **137D**, 3–21 (1986).
114. J. G. Abelin *et al.*, Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315–326 (2017).
115. H. Pearson *et al.*, MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* **126**, 4690–4701 (2016).
116. Y. Xing, S. C. Jameson, K. A. Hogquist, Thymoproteasome subunit-β5T generates peptide-MHC complexes specialized for positive selection. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6979–6984 (2013).
117. K. Takada, K. Kondo, Y. Takahama, Generation of peptides that promote positive selection in the thymus. *J. Immunol.* **198**, 2215–2222 (2017).
118. Y. Takahama, I. Ohigashi, S. Murata, K. Tanaka, Thymoproteasome and peptidic self. *Immunogenetics* **71**, 217–221 (2019).
119. S. Murata *et al.*, Regulation of CD8+ T cell development by thymus-specific proteasomes. *Science* **316**, 1349–1353 (2007).
120. K. Sasaki *et al.*, Thymoproteasomes produce unique peptide motifs for positive selection of $CD8^+$ T cells. *Nat. Commun.* **6**, 7484 (2015).
121. M. Kasahara, M. F. Flajnik, Origin and evolution of the specialized forms of proteasomes involved in antigen presentation. *Immunogenetics* **71**, 251–261 (2019).
122. M. Uhlén *et al.*, Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
123. J. A. Carter *et al.*, Transcriptomic diversity in human medullary thymic epithelial cells. *Nat. Commun.* **13**, 4296 (2022).
124. H. N. Kløverpris *et al.*, A molecular switch in immunodominant HIV-1-specific CD8 T-cell epitopes shapes differential HLA-restricted escape. *Retrovirology* **12**, 20 (2015).
125. C. Krishna, D. Chowell, M. Gönen, Y. Elhanati, T. A. Chan, Genetic and environmental determinants of human TCR repertoire diversity. *Immun. Ageing* **17**, 26 (2020).
126. D. K. Cole *et al.*, Modification of MHC anchor residues generates heteroclitic peptides that alter TCR binding and T cell recognition. *J. Immunol.* **185**, 2600–2610 (2010).
127. A. K. Sewell, Why must T cells be cross-reactive? *Nat. Rev. Immunol.* **12**, 669–677 (2012).
128. D. Mason, A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunol. Today* **19**, 395–404 (1998).
129. G. Petrova, A. Ferrante, J. Gorski, Cross-reactivity of T cells and its role in the immune system. *Crit. Rev. Immunol.* **32**, 349–372 (2012).
130. T. Boehm, Quality control in self/nonself discrimination. *Cell* **125**, 845–858 (2006).
131. M.-E. Mickael *et al.*, Fezf2 and Aire1 evolutionary trade-off in negative selection of T cells in the thymus. bioRxiv [Preprint] (2022). https://www.biorxiv.org/content/10.1101/2022.02.01.478624v1 (Accessed 31 January 2024).
132. J. Radwan, W. Babik, J. Kaufman, T. L. Lenz, J. Winternitz, Advances in the evolutionary understanding of MHC polymorphism. *Trends Genet. TIG* **36**, 298–311 (2020).
133. P. Kubiniok *et al.*, Understanding the constitutive presentation of MHC class I immunopeptidomes in primary tissues. *iScience* **25**, 103768 (2022).
134. L. Quintana-Murci, Human immunology through the lens of evolutionary genetics. *Cell* **177**, 184–199 (2019).
135. B. J. Crespi, M. C. Go, Diametrical diseases reflect evolutionary-genetic tradeoffs: Evidence from psychiatry, neurology, rheumatology, oncology and immunology. *Evol. Med. Public Health* **2015**, 216–253 (2015).
136. M. Manczinger *et al.*, Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. *PLoS Biol.* **17**, e3000131 (2019).
137. M. Sironi, R. Cagliani, D. Forni, M. Clerici, Evolutionary insights into host–pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* **16**, 224–236 (2015).
138. L. Saveanu *et al.*, Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nat. Immunol.* **6**, 689–697 (2005).
139. J. Klunk *et al.*, Evolution of immune genes is associated with the Black Death. *Nature* **611**, 312–319 (2022).
140. N. Bannert, R. Kurth, The evolutionary dynamics of human endogenous retroviral families. *Annu. Rev. Genomics Hum. Genet.* **7**, 149–173 (2006).
141. J.-D. Larouche *et al.*, Widespread and tissue-specific expression of endogenous retroelements in human somatic tissues. *Genome Med.* **12**, 40 (2020).
142. G. Kassiotis, The immunological conundrum of endogenous retroelements. *Annu. Rev. Immunol.* **41**, 99–125 (2023).
143. J. B. Sacha *et al.*, Vaccination with cancer- and HIV infection–associated endogenous retrotransposable elements is safe and immunogenic. *J. Immunol. Baltim. Md* **1950**, 1467–1479 (2012).
144. N. Tugnet, P. Rylance, D. Roden, M. Trela, P. Nelson, Human endogenous retroviruses (HERVs) and autoimmune rheumatic disease: Is there a link? *Open Rheumatol. J.* **7**, 13–21 (2013).
145. B. A. Nexø *et al.*, Are human endogenous retroviruses triggers of autoimmune diseases? Unveiling associations of three diseases and viral loci. *Immunol. Res.* **64**, 55–63 (2016).
146. A. J. Roger, S. A. Muñoz-Gómez, R. Kamikawa, The origin and diversification of mitochondria. *Curr. Biol. CB* **27**, R1177–R1192 (2017).
147. V. Karnaukhov *et al.*, HLA variants have different preferences to present proteins with specific molecular functions which are complemented in frequent haplotypes. *Front. Immunol.* **13**, 1067463 (2022).
148. G. Prota *et al.*, Enhanced immunogenicity of mitochondrial-localized proteins in cancer cells. *Cancer Immunol. Res.* **8**, 685–697 (2020).
149. A. Grignolio *et al.*, Towards a liquid self: How time, geography, and life experiences reshape the biological identity. *Front. Immunol.* **5**, 153 (2014).
150. X. Sun *et al.*, Longitudinal analysis reveals age-related changes in the T cell receptor repertoire of human T cell subsets. *J. Clin. Invest.* **132**, e158122 (2022).
151. R. Ragazzini *et al.*, Defining the identity and the niches of epithelial stem cells with highly pleiotropic multilineage potency in the human thymus. *Dev. Cell* **58**, 2428–2446.e9 (2023).
152. M. Gomez-Perosanz, A. Ras-Carmona, P. A. Reche, PCPS: A web server to predict proteasomal cleavage sites. *Methods Mol. Biol. Clifton NJ* **2131**, 399–406 (2020).
153. R. Yin *et al.*, TCRmodel2: High-resolution modeling of T cell receptor recognition using deep learning. *Nucleic Acids Res.* **51**, W569–W576 (2023).
154. L. A. Rojas *et al.*, Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. *Nature* **618**, 144–150 (2023).
155. M. Łuksza *et al.*, Neoantigen quality predicts immunoediting in survivors of pancreatic cancer. *Nature* **606**, 389–395 (2022).
156. E. Dorigatti *et al.*, Predicting T cell receptor functionality against mutant epitopes. bioRxiv [Preprint] (2023). http://biorxiv.org/lookup/doi/10.1101/2023.05.10.540189 (Accessed 7 November 2023).
157. R. Vita *et al.*, The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).
158. C. Camacho *et al.*, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
159. B. Koncz, G. Mihály Balogh, M. Manczinger, Code and data for "Journey to your self: the vague definition of immune self and its practical implications. GitHub. https://github.com/immunoteam/journey_to_your_self. Deposited 22 April 2024.
160. R. Vita, *et al.*, Database export. tcell_full_v3. IEDB. https://www.iedb.org/downloader.php?file_name=doc/tcell_full_v3.zip. Deposited 11 April 2023.