

VDJdb in 2026: boosting T-cell receptor recognition evidence using paratope embeddings and AI-based structure prediction

Daniil V. Luppov^{1,†}, Anna E. Koneva^{1,2,†}, Dmitry V. Bagaev³, Anastasiia V. Alexandrova¹, Elizaveta K. Vlasova^{1,4}, Dmitry M. Chudakov^{1,2,5,6}, Chihiro Motozono⁷, Andrew K. Sewell^{7,8}, Mikhail Shugay^{1,2,9,*}

¹Institute of Translational Medicine, Russian National Medical State University, 117513 Moscow, Russia

²Department of Genomics of Adaptive Immunity, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, 117997 Moscow, Russia

³BIASlab, Department of Electrical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands

⁴ITMO University, 197101 Saint-Petersburg, Russia

⁵Center for Molecular and Cellular Biology, 121205 Moscow, Russia

⁶Abu Dhabi Stem Cells Center (ADSCC), 00000 Abu Dhabi, United Arab Emirates

⁷Division of Human Molecular Immunology, Joint Research Center for Human Retrovirus Infection, Kumamoto University, 860-0811 Kumamoto, Japan

⁸Division of Infection and Immunity, Cardiff University School of Medicine, CF14 4XN Cardiff, Wales, United Kingdom

⁹Institute of Molecular Biology NAS RA, 0014 Erevan, Armenia

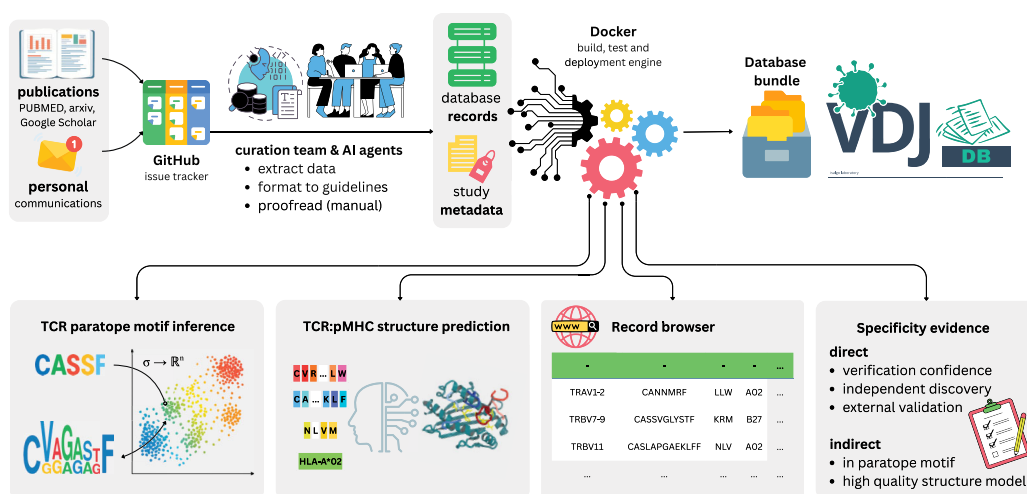
*To whom correspondence should be addressed. Email: mikhail.shugay@gmail.com

[†]The first two authors should be regarded as Joint First Authors.

Abstract

We present a significant update to VDJdb, introducing substantial enhancements to both the data content and the technical infrastructure. The integration of steadily accumulating T-cell receptor (TCR):epitope recognition data, together with advances in high-throughput experimental techniques, has expanded the landscape for machine learning-based prediction of TCR specificity, covering foreign antigens, neoantigens, and autoimmunity-associated epitopes. However, these advances have introduced new challenges, most notably in data quality—large-scale assays have heightened concerns about measurement reliability and exacerbated existing issues such as HLA and epitope coverage biases. To address these issues, we have adopted state-of-the-art artificial intelligence approaches, including protein sequence embeddings and AI-driven structure prediction. The updated VDJdb resource now features TCR specificity records annotated with embedding-derived paratope features, advanced noise filtering capabilities, and predicted three-dimensional protein structures. These improvements enable more comprehensive interrogation of TCR–epitope recognition and provide novel lines of evidence to support the reliability of high-throughput assay records, which are frequently limited by insufficient independent validation. VDJdb can be accessed at <https://vdjdb.com> and <https://github.com/antigenomics/vdjdb-db>.

Graphical abstract



Received: June 11, 2026. Revised: July 24, 2026. Accepted: August 21, 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

The adaptive immune system is characterized by an extraordinary repertoire of T-cell receptor (TCR) rearrangements, sampled from an immensely diverse sequence space exceeding 10^{15} in humans [1]. These TCRs recognize epitopes of diverse origin—short peptides presented by major histocompatibility complex (peptide-MHC, or pMHC) molecules on the cell surface. The theoretical diversity of an individual's immunopeptidome is estimated to surpass 10^9 for humans [2], rendering the combined TCR:pMHC interaction space astronomical in scale. While the vast TCR repertoire is designed to recognize virtually any foreign pathogen, factors such as biophysical constraints, negative selection against self-reactive TCRs, and the extreme rarity of certain TCRs result in a highly sparse landscape for TCR:pMHC binding events.

Extensive efforts have been devoted to identifying TCR:pMHC binders, given their critical roles in vaccine development, autoimmune disease, and cancer immunology. Nevertheless, these efforts have yielded $\sim 10^5$ TCR specificity records to date, each exhibiting highly variable measurement reliability [3]. Machine learning approaches, particularly recent advances in artificial intelligence, offer the potential to generalize and extrapolate from this narrow and biased dataset—one that is skewed towards medically relevant epitopes and MHC alleles with higher population frequencies. However, the advent of high-throughput, automated assays has revealed a sobering limitation: even for the most extensively studied epitopes, the proportion of records that can be independently validated may be as low as $\sim 50\%$, as recently reported by Messemaker *et al.* [4].

Since both machine learning strategies and our broader understanding of TCR:pMHC binding mechanisms are fundamentally shaped by the quality and reliability of available data, it is imperative to critically assess and refine our current data collection approach (Fig. 1a). The original VDJdb, released in 2017 [5], prioritized the collection of highly reliable TCR:pMHC epitope records, applying stringent inclusion criteria and rigorous confidence scoring. The 2019 update [6] introduced additional features, such as paratope motif collection, and greatly expanded the database by incorporating high-throughput, multiplexed assays. However, this increase in data volume came at a cost: most high-throughput datasets lack sufficient independent validation, precluding consistent application of the original confidence scoring system. Although the COVID-19 pandemic prompted efforts to demonstrate the reproducibility of assays and motifs across laboratories [7], systematic analyses and independent validations remain scarce, with the most comprehensive attempt discussed here emerging only in 2025.

In the current VDJdb update, we place a central emphasis on data reliability and interpretability. Notably, the original paratope motif feature, based on TCR sequence similarity, remains a valuable tool for noise filtering, especially in the absence of detailed assay information. Recent studies have further demonstrated that the quality scores of TCR:pMHC structures, as predicted by AI-based structural modelling algorithms, can serve as independent indicators of record reliability, as shown by Messemaker *et al.* Accordingly, we have integrated a novel motif-detection and noise-filtering algorithm based on protein sequence embeddings, alongside the provision of structural modelling results for VDJdb records. A dedicated web interface facilitates the assessment of

record reliability using multiple lines of evidence and scoring methodologies.

Materials and methods

Data acquisition and processing

Relevant publications and preprints are systematically tracked using the GitHub issue system (<https://github.com/antigenomics/vdjdb-db/issues>), with additional entries sourced through PubMed searches on relevant terms, manual monitoring of Google Scholar alerts, and continuous review of bioRxiv and arXiv feeds. Personal communications and author submissions are similarly managed via the issue system and processed on a case-by-case basis. All data undergo formatting and proofreading in accordance with VDJdb guidelines and database specifications (<https://github.com/antigenomics/vdjdb-db/blob/master/README.md>). A dedicated set of agentic AI skills is employed to extract, format, and proofread submissions (<https://github.com/antigenomics/vdjdb-db/tree/master/skills>). Mandatory manual curation and review by the VDJdb team were required for all submission types. Proofreading issues identified by database users are also monitored and resolved individually. Final database assembly and validation are performed using a custom Docker-based pipeline, maintained alongside input data fragments in the VDJdb-db repository. This build pipeline is hard-coded and version-controlled, and has been used and progressively refined since the inception of the database; it validates every submitted data chunk for missing fields, field syntax and controlled-vocabulary naming (IMGT gene and HLA allele nomenclature), canonical CDR3 amino acid sequences, and correct mapping of the V and J segments onto the CDR3, rejecting any record that fails these checks. Each build additionally emits a human-readable summary report (<https://vdjdb.com/overview>) that is manually inspected for CDR3 spectratype and V–J usage anomalies and for spurious species, antigen-source-species, gene, and allele names. As a further safeguard, every imported chunk is independently audited by a curator other than the one who performed the import.

Embedding-based paratope motif detection and noise filtering

We refined the VDJdb motif annotation by recomputing database-wide TCR clusters using TCREMP embeddings [8]. In contrast to the previous TCRNET annotation [9], which relied on clustering TCRs based on a Hamming distance-1 graph of complementarity-determining region 3 (CDR3) sequences, TCREMP represents full TCR amino acid sequences as high-dimensional numerical vectors (dimension $\sim 10^3$), enabling direct clustering or dimensionality reduction using principal component analysis. Aggregating variable gene segments (encompassing CDR1 and CDR2 of the paratope) with variable CDR3 lengths increased clustering coverage across both TRA and TRB chains and yielded larger, less fragmented motif clusters. Clustering was performed using the DBSCAN algorithm, with the neighbourhood radius ϵ selected automatically from the knee of the sorted k -nearest-neighbour distance curve by the Kneedle method [10]; in the rare degenerate case in which this curve is near-linear and no knee is resolved, we fall back to the mean k -distance, so that no per-dataset threshold tuning is required, and synthetic controls were used to distinguish true clusters from those arising from convergent

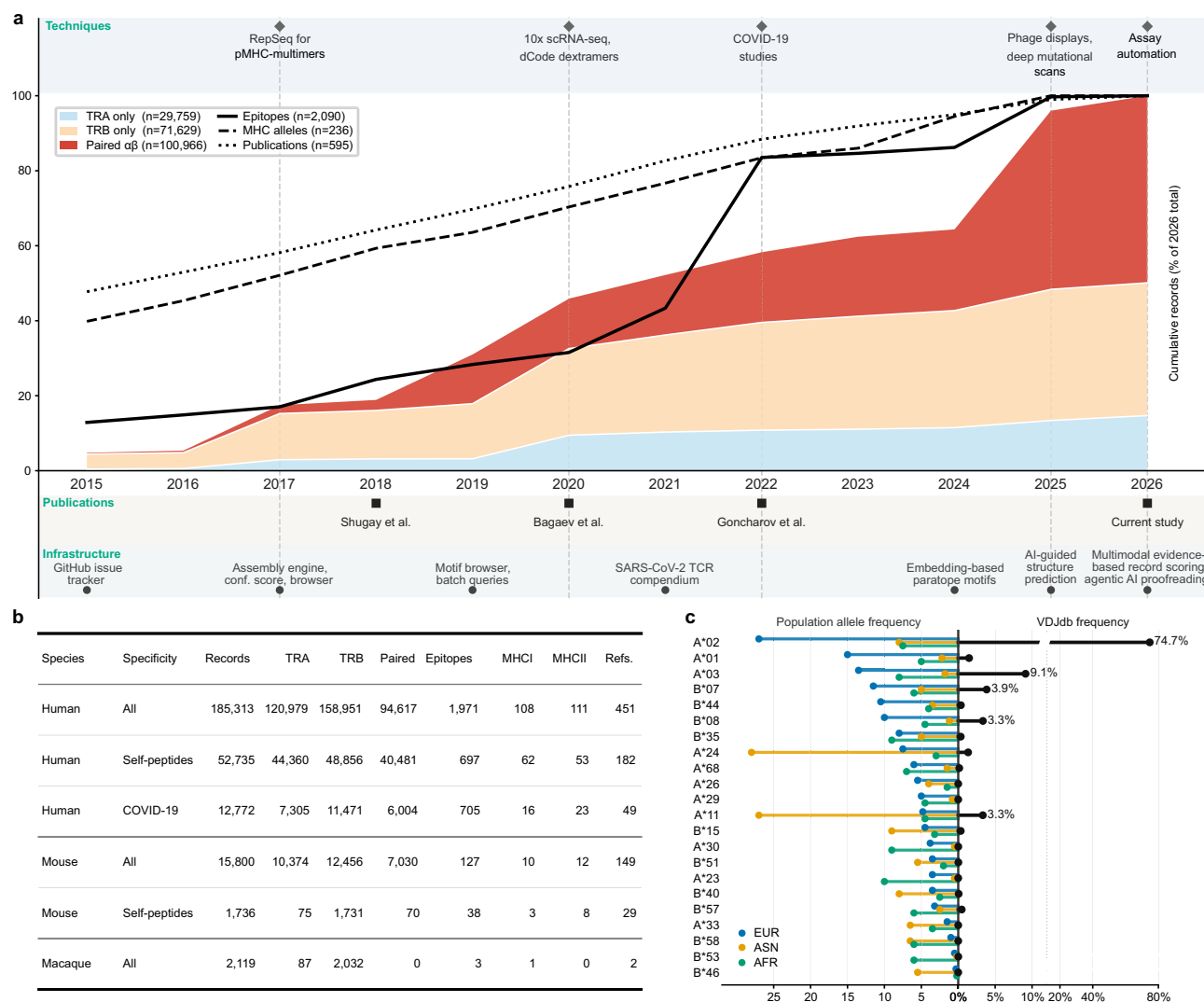


Figure 1. VDJdb database expansion, coverage, and HLA bias. **(a)** Cumulative composition of VDJdb TCR:pMHC records (unique entries) by year of publication, 2015 and earlier to 2026, present. Normalized to 100% as of 2026. Coloured areas indicate TRA-only (light blue), TRB-only (light orange), and paired (having both TRA and TRB) $\alpha\beta$ chain TCRs (red). Overlaid black lines (right axis) represent cumulative unique epitopes (solid), MHC alleles (dashed), and publications (dotted). Vertical dashed lines denote key technological milestones: amplicon RepSeq for MHC-multimer-sorted T cells (2017), 10x scRNA-seq and dCODE dextramer multiplexing (2020), large-scale COVID-19 T-cell studies (2022), deep epitope mutational scans and phage displays (2025, [20, 21]), and the implementation of automated TCR cloning/validation platforms TCRvdb (Messemaker *et al.*) and TCRAFT [22] emerging by 2026. Below the axis, squares indicate VDJdb key publications: Shugay *et al.* [5], Bagaev *et al.* [6], Goncharov *et al.* [7], and the present study (2026); circles mark VDJdb infrastructure updates: submission tracking via GitHub issue tracker (2015), assembly engine for automated database builds with record confidence scoring and web browser (2017), motif viewer and batch query interface (2019), embedding-based paratope analysis (2024), AI-guided structure prediction (2025), and multimodal evidence scoring with agentic AI proofreading (2026). **(b)** Summary counts for VDJdb TCR:pMHC records. Rows correspond to species—human (*Homo sapiens*), mouse (*Mus musculus*), and macaque (*Macaca mulatta*)—stratified by antigen specificity (all peptides, self-peptides including neoantigens and autoimmune peptides, and SARS-CoV-2 epitopes). Columns indicate total records; records with TRA sequence; records with TRB sequence; paired $\alpha\beta$ records; unique epitopes; unique MHCII α alleles; unique MHCII α/β pairs; and unique publications. **(c)** Representation of MHC class I alleles among human MHCII records in VDJdb (black bars) as a percentage of all human MHCII records. Alleles are ordered by their frequency in European populations; the top 22 alleles are shown. European (EUR), Asian (ASN), and African (AFR) population frequencies were sourced from <https://www.allelefrequencies.net/> [23]. The y-axis is broken at 10% to display the overrepresentation of HLA-A*02.

V(D)J rearrangements, based on statistical enrichment of VDJdb sequences. Because TCREMP encodes each receptor by alignment to a fixed set of reference prototypes rather than by pairwise CDR3 identity, it groups CDR3 motifs of differing lengths together and extends naturally to paired $\alpha\beta$ receptors by concatenating the per-chain embeddings. The updated motifs were validated on an in-house set of TCR specificity prediction benchmarks, confirming that precision was maintained while recall and F1 scores improved. Quantita-

tively, on the full human VDJdb release TCREMP clustering matched the legacy TCNET on cluster purity (0.94 versus 0.95) while more than doubling both retention (0.57 versus 0.23, the fraction of receptors assigned to a motif cluster) and the coverage-aware F1 score (0.71 versus 0.37), and it outperformed the protein language-model embeddings TCR-BERT [11] and ESM-2 at both 35 and 650M parameters [12] on every metric under an identical clustering pipeline (Supplementary Fig. S1a).

Structural modelling

TCR structural modelling was performed using the locally installed TCRmodel2 tool [13], selected for its high predictive performance, ease of installation, and open-source distribution. Default parameters were used with the authors' custom AlphaFold database. Only VDJdb records containing both TCR α and β chains were considered for modelling and subsequent analyses. For each record, complete TCR sequences were generated using the stitchr Python library (v1.3.1; [14]), and corresponding MHC sequences were retrieved from the IMGT database (<https://www.ebi.ac.uk/ipd/imgt/hla/>, accessed 12 August 2025). Modelling was executed on a high-performance computing cluster and managed via the SLURM workload manager. For each VDJdb record, five structures were generated, and only the highest-ranked model was retained based on AlphaFold2 confidence [15]. The average wall-clock time to generate a single structure was ~ 10 min on a single node equipped with an NVIDIA Tesla V100 GPU with 32 GB VRAM, 32 CPU cores, and 128 GB of RAM. Owing to computational constraints, comprehensive modelling of all $\sim 10^5$ VDJdb records in a relatively short time was unfeasible; instead, a pilot set of $\sim 10^4$ structures was prioritized, focusing on records associated with identified human paratope motifs for MHC class I (MHCI)-restricted epitopes.

To facilitate model comparability, all generated structures underwent a two-stage alignment using the PyMOL Python API (<https://github.com/schrodinger/pymol-open-source>). The initial alignment was performed across all structures sharing the same MHC class, followed by alignment among structures recognising the same epitope. After alignment, the 3D C_α coordinates of the CDR3 loops and the peptide were projected onto a uniform 2D plane for all structures. Inter-chain amino acid contacts were annotated when the distance between any atoms of two residues was < 5 Å. The resulting 2D projections, annotated with contact information (aka 'complementarity maps'), are now accessible in the VDJdb structure browser. Each VDJdb record with a generated structure was further annotated with its AlphaFold2 ranking confidence, the number of TCR-peptide contacts, and scanning and pitch angles computed with the STCRpy library (v1.0.6; [16]). PDB files, contact tables, and C_α coordinate data are made available via the structure browser interface and on HuggingFace (https://huggingface.co/datasets/isalgo/vdjdb_structure_models).

Web interface

The legacy VDJdb web browser [6] (<https://github.com/antigenomics/vdjdb-web>) was updated to include an updated paratope motif browser, a TCR:pMHC structure browser, and integrated evidence filtering within the main search interface.

Results

Timeline of database evolution and data accumulation

Since 2017, the accumulation of published TCR specificity data has accelerated considerably, propelled by advances in experimental methodologies that have expanded coverage of TCRs, epitopes, and MHC alleles (Fig. 1). Tetramer-based enrichment followed by deep repertoire sequencing has markedly increased the number of single-chain TCR records [17]. The deployment of single-cell sequencing platforms (e.g.,

10X Genomics, <https://www.10xgenomics.com/>) in conjunction with multiplexed tetramer assays has enabled comprehensive cataloguing of paired TCR chains and facilitated the identification of numerous new epitopes.

The COVID-19 pandemic precipitated a substantial and diverse influx of epitope records, predominantly centred on SARS-CoV-2 antigens. Notably, SARS-CoV-2 datasets provide a compelling case study for robust orthogonal validation in TCR specificity research [7]. For several immunodominant HLA-A*02- and HLA-A*24-restricted Spike epitopes, multiple independent laboratories, donor cohorts, and assay platforms have reproducibly identified convergent TCR motifs and clonotypes, in agreement with those previously described in VDJdb [18, 19]. This remarkable consistency across distinct studies and technical approaches represents one of the most compelling real-world validations of TCR-epitope assignment reproducibility to date. Furthermore, as discussed below (see Table 1, HLA-A*02:YLQPRFTLL case), VDJdb records substantiated by cross-laboratory evidence remain robust even under the most rigorous and meticulous experimental scrutiny of Messemaker *et al.*

Emerging methodologies, including deep epitope mutational scanning [20] and phage display-based selection of epitope binders from pools of TCRs with degenerate CDR3 sequences [21], have enabled the generation of extensive TCR:pMHC datasets, thereby expanding coverage of rare TCR variants and epitopes and facilitating in-depth analysis of epitope recognition hotspots and paratope motifs (Fig. 1a). Nonetheless, these approaches can introduce pronounced biases during classifier training, as they may yield thousands of highly similar TCR sequences for a single epitope or enhance epitope diversity without a commensurate increase in supporting TCR counts, ultimately limiting their utility for machine learning applications. Laboratory automation has further improved throughput and reproducibility and continues to advance [22] and Messemaker *et al.*; however, each methodology produces a distinct dataset structure and introduces sampling biases that must be carefully considered.

The current version of VDJdb comprises $\sim 203\,230$ records (compared to 5491 at initial publication, representing a ~ 37 -fold increase over 8 years) and now provides broad coverage of autoimmune peptides and neoantigens (Fig. 1b). Nevertheless, significant compositional biases remain—most notably, a pronounced overrepresentation of HLA-A*02 allele records compared to the allele population frequency [23], as reported by multiple groups and highlighted in Fig. 1c. We anticipate that the growing requirement for TCR specificity assessment in donors with diverse HLA backgrounds, particularly in the context of personalized cancer vaccine development, coupled with advances in high-throughput pMHC multiplexing assays, will facilitate the mitigation of these biases in the near future.

Despite the implementation of stringent inclusion criteria, VDJdb is now comparable (Fig. 2) to the IEDB Epitope Specific T Cell Receptors database [3] in terms of record count, based on the latest releases at the time of writing (accessed 7 June 2026). Notably, certain datasets—such as MIRA (Multiplexed Identification of T cell Receptor Antigen specificity)-based assays [24, 25], which utilize a variation of the combinatorial peptide pool approach—are valuable for exploratory analyses but lack precise pMHC specificity information and are therefore not considered for inclusion into VDJdb (Fig. 2a–c). Other datasets are likewise ex-

Table 1. Independent corroboration in VDJdb predicts experimental validation of paired $\alpha\beta$ TCRs in the Messemaker *et al.* screen

Epitope	Independent validation (VDJdb)	Validated (TCRvdb)	Not validated (TCRvdb)	Total	OR (95% CI)	Fisher <i>P</i>
GLCTLVAML	Not independently validated	43	30	73	—	—
	TRA only	25	8	33	2.10 (0.85–5.19)	0.1
	TRB only	22	17	39	0.90 (0.41–1.96)	0.8
	Both TRA and TRB	44	4	48	6.93 (2.37–20.30)	6.5×10^{-5}
	Total	134	59	193		
YLQPRTFLL	Not independently validated	65	223	288	—	—
	TRA only	55	10	65	18.04 (8.82–36.86)	7.9×10^{-21}
	TRB only	23	13	36	5.94 (2.88–12.25)	9.7×10^{-7}
	Both TRA and TRB	29	3	32	28.76 (9.18–90.09)	2.6×10^{-14}
	Total	172	249	421		

Receptors from the Messemaker *et al.* TCRvdb dataset (snapshot 01-05-2025, provided by authors) were screened against their cognate HLA-A*02:01-restricted epitope (YLQPRTFLL, SARS-CoV-2 spike, $n = 421$; GLCTLVAML, EBV BMLF1, $n = 193$) and considered validated if the Benjamini–Hochberg adjusted $P < 1 \times 10^{-5}$, as per author recommendations. Independent validation was assigned when each chain matched the current VDJdb snapshot by epitope, V-gene (allele-collapsed), and CDR3 amino acid sequence, and was supported by more than one unique reference (≥ 2 independent studies). Receptors are cross-tabulated by independent validation class (rows) and experimental validation status (columns). Statistical association was tested for each epitope using Pearson's χ^2 test (YLQPRTFLL: $\chi^2 = 132.1$, $P = 1.9 \times 10^{-28}$; GLCTLVAML: $\chi^2 = 18.7$, $P = 3.1 \times 10^{-4}$). Odds ratios (OR), their confidence intervals (95% CI), and two-sided Fisher exact test *P*-values compare the odds of validation for each class relative to the baseline (not independently validated).

cluded when T-cell stimulation to a given epitope can only be inferred indirectly, or critical pMHC information is missing, thereby limiting their suitability for high-confidence TCR specificity annotation. Notably, after excluding such indirect assays, the IEDB TCR database ($\sim 140\,000$ records) is considerably smaller than VDJdb (Fig. 2d–f).

Paratope motifs and noise filtering

It is widely accepted that TCRs recognizing the same pMHC complex often share similar paratope sequences, exemplified by well-characterized conserved *k*-mers such as the ‘RS’ motif for HLA-A*02:GILGFVFTL of Influenza A virus or ubiquitous Variable gene biases like TRBV20-1 for HLA-A*02:GLCTLVAML of Epstein-Barr Virus, TRAV12-2 for HLA-A*02:LLWNGPMAV of Yellow Fever Virus, and HLA-A*02:ELAGIGILTV, an artificial heteroclitic variant of naturally presented self-peptide HLA-A*02:EAAGIGILTV from MART-1 gene. However, as VDJdb has grown, it has become apparent that this relationship is more complex: a single epitope may be associated with multiple paratope motifs or recognition modes, and the fine structure of paratope sequence space can vary dramatically between epitopes. For example, some epitopes present a single dominant motif, while others, such as HLA-A*02:NLVPMVATV of Cytomegalovirus, display multiple disjoint motifs, complicating motif discovery and modelling [26]. Although the exact molecular determinants governing these differences remain incompletely elucidated, current evidence suggests that ‘featureless’ epitopes tend to be recognized by a broader and more diverse TCR repertoire [27, 28]. Additionally, rare sequence variants or insufficient sampling may impede the unambiguous delineation of certain motif architectures, as seen for TCRs with low V(D)J generation probabilities, such as autoimmunity-associated variants (e.g. TCR motif in HLA-B*27+ ankylosing spondylitis donors [29]).

Given that a substantial portion of assay-derived records may represent false positives, it is essential to distinguish genuine ‘orphan’ sequences from methodological artefacts [30]. The TCRNET algorithm, previously used for robust paratope motif detection [31], can also serve as a noise filter by flag-

ging sequences not associated with any statistically significant motif as likely false positives and eliminating numerous false-positive hits that dominate cross-repertoire comparisons [9]. However, TCRNET is highly stringent—requiring only single amino acid differences and limited handling of CDR1 and CDR2 regions—which restricts sensitivity. To address these limitations, we developed the TCREMP embedding-based clustering approach [8] (see the ‘Materials and methods’ section), enabling a more flexible and comprehensive delineation of paratope motifs.

Transitioning to embedding-based annotation has markedly increased both the number and size of paratope motif clusters for TRA and, especially, TRB chains (Fig. 3a and b). This approach reduces fragmentation of related sequence neighbourhoods and yields more coherent motif clusters. Notably, CDR3 sequences in VDJdb are frequently associated with multiple epitopes, reflecting both genuine cross-reactivity and false positives (Fig. 3c and d). Comparing CDR3:epitope associations to TCREMP cluster incidence reveals that most shared CDR3 sequences are limited to a few motif contexts, CDR3s associated with several epitopes are distinguishable from more specific motifs, and those with an unexpectedly high number of associations do not belong to any motif and fall into the ‘noise’ category. Thus, embedding-based motifs provide a refined annotation layer that maintains biologically plausible cross-reactivity while filtering out spurious records that are likely misannotated.

Furthermore, the new annotation consolidates structurally similar motifs that were previously split among smaller clusters, particularly for clonotypes with divergent CDR3 lengths but conserved paratope features (Fig. 3e and f). For instance, the TRB CSAR/GELFF-16 motif associated with HLA-A*02:YLQPRTFLL forms a single TCREMP cluster, merging previously separate TCRnet clusters. Collectively, TCREMP-based clustering delivers a more comprehensive and robust paratope motif annotation layer for VDJdb and aids in database denoising (see Fig. 5a). To quantify these improvements and to benchmark TCREMP against modern alternatives, we compared antigen-specific clustering across the full human VDJdb release and a high-quality subset of records whose specificity is corroborated by at least two in-

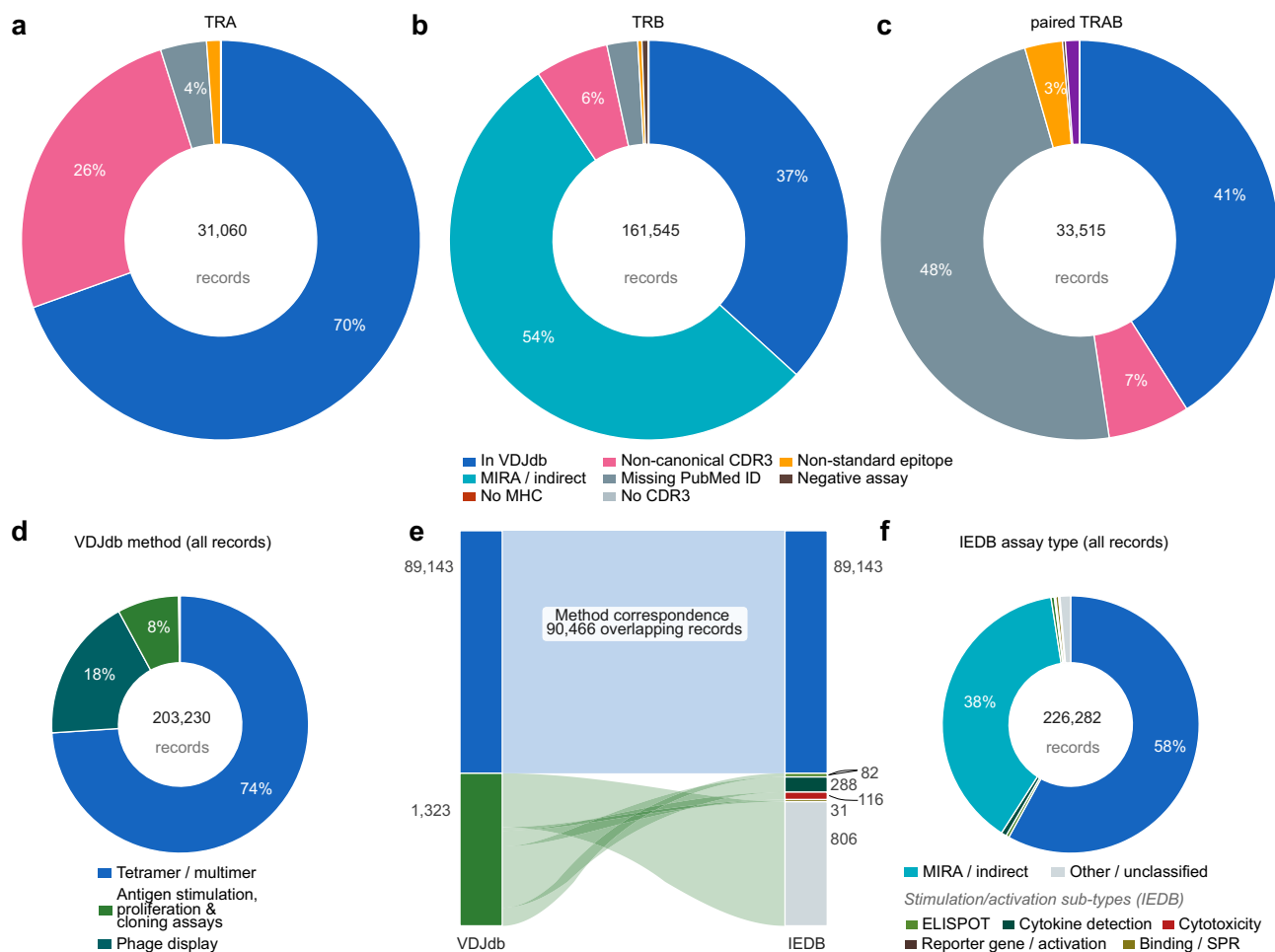


Figure 2. Comparison of IEDB epitope-specific TCRs and VDJdb databases. **(a–c)** Proportional breakdown of all IEDB TCR records (unique entries) by their classification according to VDJdb database specification and submission standards, shown separately for TRA-only, TRB-only, and paired (having both TRA and TRB) $\alpha\beta$ chain TCRs. Each donut segment reports the percentage of records in one of the following categories; the centre displays the total record count: In VDJdb (blue), records present in VDJdb; MIRA/indirect (cyan), indirect TCR reactivity association from combinatorial epitope pools, e.g. ImmuneCODE results based on MIRA assay [25]; Non-canonical CDR3 (pink), CDR3 sequences containing stop codons, non-standard amino acids, or biologically implausible lengths/flanks that fail VDJdb quality filters; No CDR3 (light grey), records without a CDR3 amino acid sequence; MHC-unvalidated epitope (amber), epitope–MHC combinations lacking orthogonal experimental support for the stated restriction element; Non-standard epitope (gold), peptides with non-canonical residues, post-translational modifications, or unclear sequence identity; No MHC (dark red), records lacking a defined MHC restriction element; Negative assay (dark brown), records explicitly annotated as non-reactive or negative; Missing PubMed ID (grey-blue), records citing URLs, DOIs, not formally published at the time of IEDB curation, e.g. 10x Genomics assay application notes. **(d)** Assay-method breakdown across VDJdb and IEDB. The following VDJdb metadata groups are shown: Tetramer / multimer, records from MHC-multimer staining experiments, including tetramer, dextramer, pentamer, streptamer and related reagents; Antigen stimulation, proliferation & cloning assays, records identified via antigen-loaded target cells, co-culture assays, limiting-dilution cloning, T-cell culture, and cytokine-capture enrichment, also includes measurements from deep epitope mutational scans [20]; Phage display, records from *in vitro* panning of TCR-like binders with randomized CDR3 against pMHC-coated beads [21]. Fewer than 0.001% of records fall into the ‘other’ category (e.g. structural assays) and are absent from the figure. **(e)** Alluvial diagram showing the correspondence between the VDJdb and IEDB assay methods for records from publications present in both databases. The y-axis uses a symmetric-log scale (linear threshold 200 records) to resolve minority methods alongside the dominant tetramer category. **(f)** Assay-type distribution across all IEDB TCR records. Categories with detailed assay types—Cytokine detection, ELISPOT, T-cell cloning/proliferation, Cytotoxicity, Reporter gene/activation marker, Binding/SPR, Structural—together account for fewer than 5% of records.

dependent references (Supplementary Fig. S1). TCREMP recovered antigen-pure clusters while assigning far more of the repertoire to motifs than TCRNET, and it dominated the purity–retention frontier of both TCR-BERT—a language model pretrained on TCR sequences, including those in VDJdb—and untuned ESM-2 embeddings across parameter scales. The advantage persisted on the high-quality subset, where stringent noise filtering is less critical, indicating that the result is not an artefact of the noise threshold (Supplementary Fig. S1b and c).

Exploring TCR:pMHC contacts in structural models

Previous studies by our group and others [5, 32] have demonstrated a strong concordance between structural contact patterns and paratope motif annotations. Recent results further indicate that the confidence scores of structural models are effective predictors of true TCR:pMHC binding. This was first observed indirectly in the IMMREP25 challenge: while overall classifier accuracy for predicting specificity against previously unseen epitopes was moderate to low, only structure-based methods—leveraging AlphaFold-

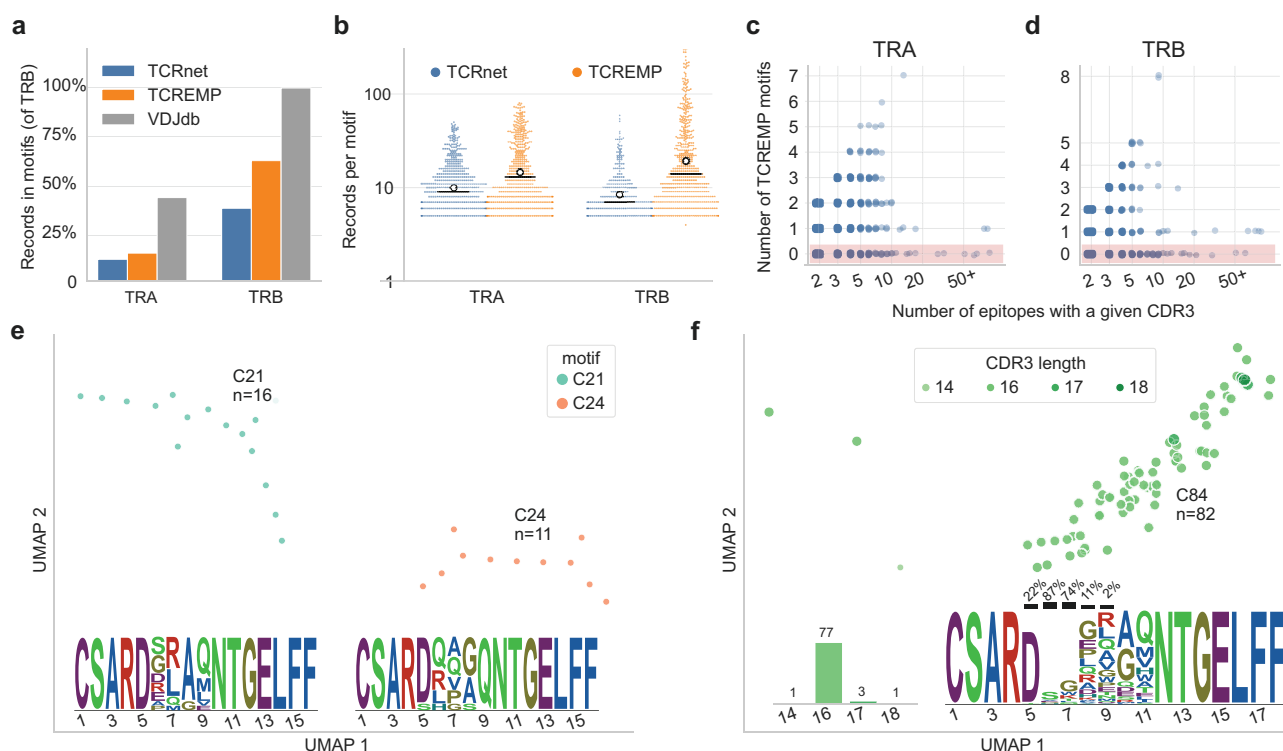


Figure 3. TCREMP-based paratope motif inference improves coverage, reduces motif fragmentation, and filters false positives. **(a)** Fraction of clustered records for TRA and TRB chains for TCRnet and TCREMP algorithms, scaled to the total number of TRB records in VDJdb. Grey bars denote unclustered VDJdb records. **(b)** Cluster-size distributions for TCRnet and TCREMP annotations by chain. Each point is a cluster. Open circles and horizontal lines indicate average cluster size for each method and chain. TCREMP yields larger clusters, reflecting reduced motif fragmentation. Incidence of TRA **(c)** and TRB **(d)** CDR3 sequences in distinct TCREMP clusters (y-axis) compared to the number of unique epitopes it is associated with (x-axis). Each point represents a unique CDR3; colour shows point density. Note that most CDR3s detected across an extremely high number of epitopes do not have an associated TCREMP cluster ($y=0$, labelled as noise) and therefore represent false positives. **(e)** Fragmented paratope motif annotation in TCRnet for HLA-A*02:YLQPRFTLL-associated TRB CDR3 sequences: two TCRnet clusters (C21, C24) have highly similar length-16 sequence logos but remain separated. UMAP coordinates are calculated from the embedding of the CDR3 similarity graph with a Hamming distance threshold of 1. **(f)** TCREMP-based annotation of the same motif merges into a single cluster (C84) and includes CDR3 variants with different raw lengths. Points are coloured by CDR3 length; the inset bar plot shows length distribution within C84. Sequence logo displays the shared embedded CSAR/GELFF-16 motif. Black bars above the sequence logo indicate the fraction of clonotypes with a deletion/gap at a given position aligned to a 18-aa sequence logo. UMAP coordinates are calculated from TCREMP embeddings of TCR sequences.

generated TCR:pMHC models—produced substantial gains over random classifiers, highlighting the utility of accurate structural modelling for identifying bona fide binders, as reported by Richardson *et al.* [33]. More recently, studies such as Messemaker *et al.* have demonstrated that integrating AlphaFold model confidence, as implemented in approaches like TCRbridge, can reliably distinguish true binders from noise. Such model-confidence scores are most informative when adequate structural templates are available—as is the case for the well-characterized, paired-chain records selected here for modelling—and are correspondingly less reliable for entirely novel epitopes or receptors that lack close templates; per-model confidence, interface ipTM, and outlier-binding-mode flags are therefore surfaced for every modelled structure so that reliability can be assessed on a case-by-case basis.

Given the current scarcity of experimentally resolved TCR:pMHC structures, imputing missing data with high-quality computational models offers a unique opportunity to investigate the molecular mechanisms underlying antigen recognition. To address this gap, and to complement paratope motif-based evidence for VDJdb record confidence, we un-

dertook large-scale modelling of TCR:pMHC complexes (see the ‘Materials and methods’ section). In total, we generated over 10 000 structural models spanning a range of quality scores (Fig. 4a–c), with predicted docking angles that closely matched those observed in experimental structures [16]. Although there were cases where TCRmodel2 failed to produce a docked complex (yielding zero contacts), the inter-chain contacts were normally distributed and, in general, contact, pitch, and scanning angle distributions were similar for native and modelled structures (Fig. 4d and e). Interestingly, there was a significant but extremely weak correlation between the number of contacts in the complex and the Interface Predicted Template-Modelling Score (ipTM; Fig. 4f). This observation suggests that the nature and positioning of specific residue interactions—rather than their absolute number—are central to the stability of the TCR:pMHC complex.

To facilitate exploration and interpretation, we have integrated a dedicated structure browser into the VDJdb web interface. This resource enables users to visualize contacts between the TRA and TRB CDR3 loops and the epitope across paratope motifs using a minimalist, intuitive ‘complementarity maps’ visualization (Fig. 5b).

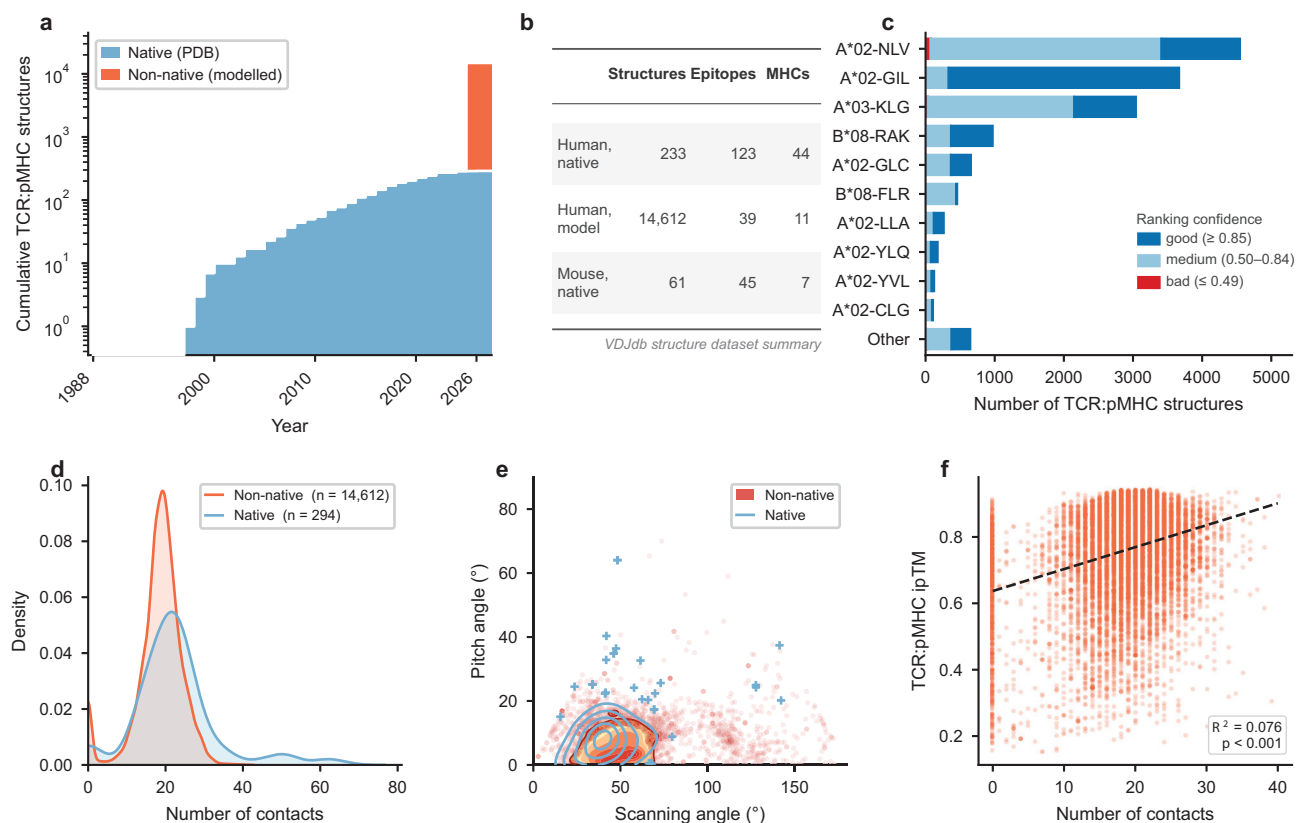


Figure 4. Structural coverage and quality of TCR:pMHC complexes in VDJdb. All complexes shown are MHC class I-restricted TCR:pMHC structures. **(a)** Cumulative TCR:pMHC structure records (log scale) over time. Blue: experimentally determined native structures from Protein Data Bank published prior to 2025 that were annotated and deposited to VDJdb ($n = 294$). Orange: structures predicted for database records using TCRmodel2 in 2025–2026 ($n = 14,612$). **(b)** Summary by species and structure type, showing counts of unique TCR:pMHC complexes, epitope sequences, and MHC alleles. **(c)** Distribution of TCRmodel2 ranking confidence scores for the 10 most frequently modelled epitope-MHC pairs (e.g., A02-NLV is NLVPMVATV in HLA-A*02:01). Bars are coloured by confidence: bad (<0.50 , red), medium ($0.50-0.84$, light blue), and good (≥ 0.85 , blue). **(d)** Kernel density-smoothed histograms of inter-chain contacts (residues from distinct complex chains with closest atoms <5 Å) per structure for native (blue) and non-native (orange) complexes. **(e)** Two-dimensional kernel density of TCR docking geometry. Contours: non-native structures (filled, yellow to red, low to high density); native PDB structures (blue lines). Outlier points: $<5\%$ of peak density. Axes: $0^\circ-180^\circ$, scanning angle, $0^\circ-90^\circ$, and pitch angle calculated by the STCRpy tool. **(f)** Correlation between inter-chain contacts and AlphaFold's interface predicted TM-score (ipTM) in non-native structures. Dashed line: least-squares fit; inset: R^2 and P -value.

Model-based evidence for confidence scoring

Rigorous independent validation remains a prerequisite for confirming findings from high-throughput assays; however, despite advances in automation and efficiency [22] and Messemaker *et al.*, such validation currently exists for only a limited fraction of epitopes represented in VDJdb, notably those associated with extensively studied pathogens such as SARS-CoV-2. Legacy VDJdb confidence scores are predominantly supported by records in which TCR:pMHC interactions have been investigated in depth, often leveraging structural biology data or comprehensive biochemical characterization from earlier studies. In contrast, many more recent entries derived from large-scale screening platforms are unlikely to undergo such detailed, targeted validation.

To address this gap and enhance the database's reliability, we have implemented an evidence imputation framework that draws on two complementary modalities: paratope motif analysis and structural modelling. The updated VDJdb interface facilitates record selection and prioritization by combining legacy confidence scores with model-based quality metrics (Fig. 5a and b), thereby supporting more nuanced evidence filtering (Fig. 5c). For example, independent validation

of records across multiple studies—particularly when both TRA and TRB chains of paired-chain TCRs in the VDJdb compendium are validated—is a strong predictor of successful experimental validation. This relationship is clearly illustrated by Table 1, which presents a re-analysis of the benchmarking study by Messemaker *et al.* These features enable researchers to maximize the reliability of specificity assignments and support robust benchmarking and model development for the next generation of TCR specificity prediction tools.

While the choice of filtering criteria ultimately resides with the end user, we recommend prioritizing the most stringent evidence—specifically, records with strong paratope motif support and high-confidence structural models. It is also critical to consider HLA matching between the donor and the epitope, as this can substantially reduce background noise [31], especially when searching for rare, antigen-specific TCRs in large-scale repertoires, where even a low false-positive rate can yield a high number of spurious hits. For foundational studies of TCR:pMHC interaction rules, we advise pre-filtering for well-supported, true positive records based on paratope features, and subsequently incorporating negative controls in a manner tailored to the study design, for example, by

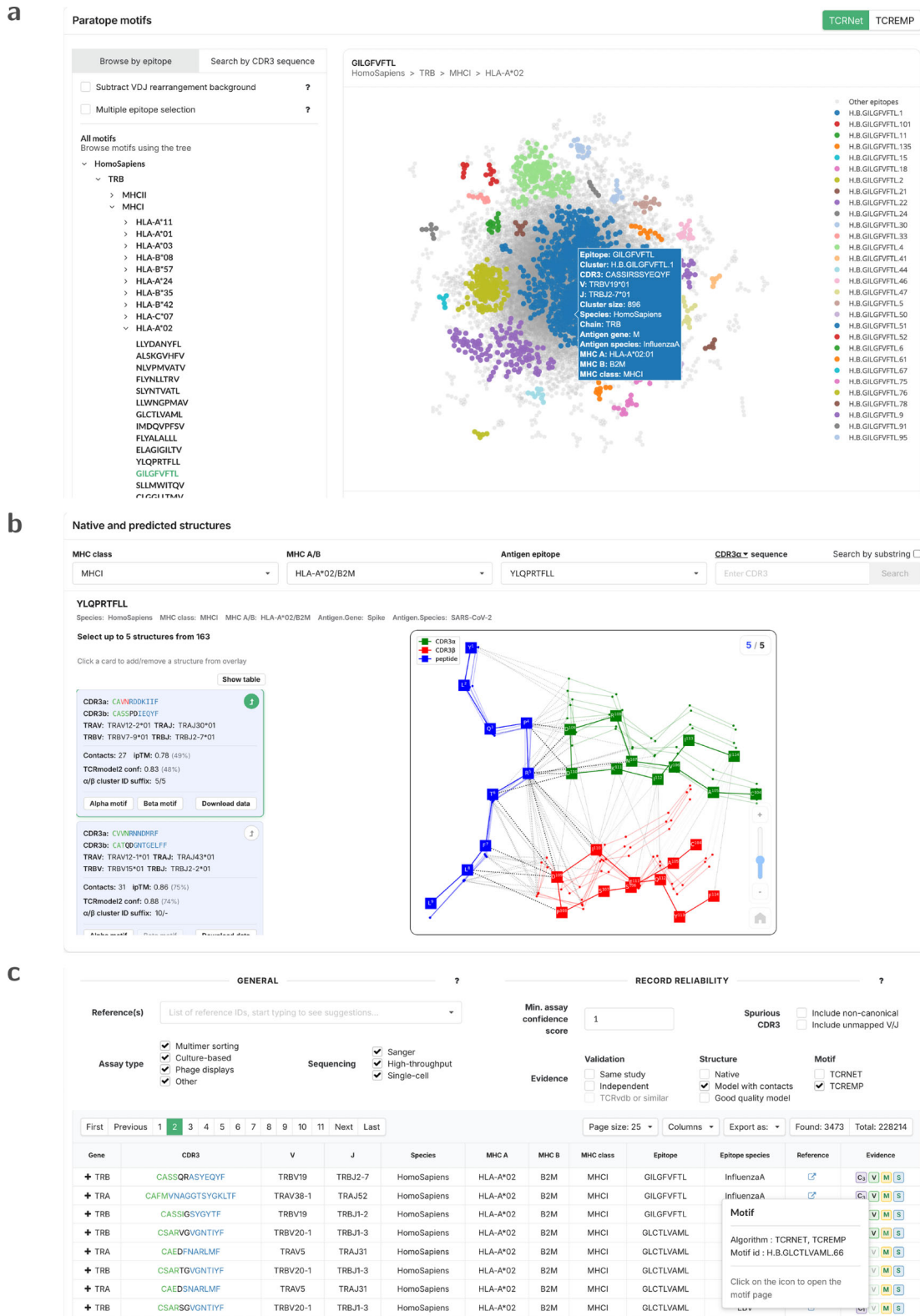


Figure 5. Overview of the VDJdb web interface. **(a)** Paratope motif browser tab. Users can select epitopes with associated paratope motifs from the left menu. Upon selection, a UMAP visualisation (TCRNET or TCREMP) displays clusters representing motifs, with background TCRs from healthy donors shown in grey. Sequence logos for each motif are provided below (not shown). **(b)** Predicted structure complementarity map tab. The interface enables browsing of structures by epitope, with information on TCR chains, MHC alleles, and paratope motifs. The interactive plot visualizes complementarity maps, showing CDR3 α , CDR3 β , and epitope residues, with loops and peptides indicated by bold lines and contacts (<5 Å) by dashed lines. Multiple structures can be overlaid to compare loop conformations. **(c)** Main database browser with evidence-based record reliability filters and quality badges. Users can filter records by VDJdb confidence score, independent validation (see also Table 1), and database cross-references (other databases enabled upon official TCRvdh release). Records can also be filtered by high-quality structure or paratope motif assignment. Reliability metadata are shown as badges and tooltips in the 'Evidence' column of the main table.

generating random V(D)J rearrangements from probabilistic models or pre-selection repertoires [31].

Discussion

Technological advances in TCR specificity measurement continue at an unprecedented pace, exemplified by the development of high-throughput and automated assays capable of systematically identifying TCR–epitope pairs across a wide array of biological contexts, including infectious diseases, cancer, and autoimmunity [34]. In parallel, the field has seen a rapid expansion of computational approaches—particularly machine learning and artificial intelligence methods—designed to predict TCR specificity *in silico*, with many models relying on VDJdb as a principal source of training and validation data [35].

This dual progress underlines VDJdb's critical responsibility as a community resource: to aggregate and curate emerging data efficiently, uphold rigorous standards for data quality, and provide transparent, reproducible workflows for data processing. By maintaining full public access to both datasets and processing pipelines, we enable the scientific community to independently scrutinize, benchmark, and validate specificity records—advantages that proprietary or closed datasets cannot offer. The recent independent benchmarking by Messesmaker *et al.* highlights the value of such transparency for rigorous, community-driven validation.

To further support the integration of experimental and computational immunology, we now provide auxiliary evidence—including paratope motif embeddings and AI-guided structure predictions—that complements high-confidence specificity records and enhances the accuracy and interpretability of TCR–epitope assignments. This approach enables systematic identification of ambiguous or challenging cases, facilitating iterative improvements to both the database and machine learning models.

While the inclusion of AI-based structure prediction and paratope motif embeddings provides valuable orthogonal evidence, it is important not to overstate the biological certainty these methods confer. High-confidence models are not equivalent to definitive proof of molecular recognition, and the use of AI for data curation introduces the risk of pre-filtering training datasets with methods later evaluated on them. These caveats should be considered when interpreting results and benchmarking model performance.

A major limitation of current datasets is the pronounced bias in pMHC coverage, with a substantial overrepresentation of MHCI-restricted epitopes, particularly those associated with frequent alleles such as HLA-A*02. As a result, MHCII and unconventional T-cell systems are markedly underrepresented. Existing MHCII data are often limited to imprecisely defined epitopes and incomplete mapping of peptide regions within the MHC groove. Resolving these biases and data gaps will require new experimental strategies, improved MHC binding prediction tools, and expanded data collection, particularly for MHCII and less-studied T-cell modalities. Addressing this challenge remains a central priority for the future development of antigen specificity databases.

We believe that this community-centric, multimodal strategy is essential for overcoming persistent challenges in the field, such as susceptibility to model overfitting on artefacts of the training set [36] and the limited generalizability of current algorithms to previously unseen epitopes, as reported by

Richardson *et al.* By promoting transparency, interoperability, and ongoing validation, VDJdb aims to accelerate the development of robust and clinically relevant TCR specificity prediction tools.

Acknowledgements

The authors would like to thank Denis Syrko for his help with HPC modelling infrastructure.

Author contributions: D.V.L. and A.E.K. collected and curated previously published TCR specificity data. D.V.B. and D.V.L. set up the web server and database assembly infrastructure. A.E.K. implemented agentic proofreading system and performed meta-analysis of database records. A.V.A. and D.V.L. performed structural modelling. E.K.V. performed paratope motif inference. A.K.S. and C.M. provided TCR specificity data. D.M.C., A.K.S., and M.S. wrote the manuscript. M.S. supervised the study.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the Russian Science Foundation (RSF) Grant No. 25-75-30013 to D.M.C. Funding to pay the Open Access publication charges for this article was provided by the personal savings of the corresponding author.

Data availability

VDJdb can be accessed at <https://vdjdb.com>; raw data can be found at <https://github.com/antigenomics/vdjdb-db> and <https://doi.org/10.5281/zenodo.22104776>.

References

1. Sewell AK. Why must T cells be cross-reactive? *Nat Rev Immunol* 2012;12:669–77. <https://doi.org/10.1038/nri3279>
2. Rock KL, Reits E, Neefjes J. Present yourself! by MHC class I and MHC class II molecules. *Trends Immunol* 2016;37:724–37. <https://doi.org/10.1016/j.it.2016.08.010>
3. Mahajan S, Vita R, Shackelford D *et al.* Epitope specific antibodies and T cell receptors in the immune epitope database. *Front Immunol* 2018;9:2688. <https://doi.org/10.3389/fimmu.2018.02688>
4. Messesmaker M, Kwee BPY, Moravec Ž *et al.* A functionally validated TCR–pMHC database for TCR specificity model development. *bioRxiv*, <https://doi.org/10.1101/2025.04.28.651095>, 12 May 2025, preprint: not peer reviewed.
5. Shugay M, Bagaev DV, Zvyagin IV *et al.* VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* 2018;46:D419–27. <https://doi.org/10.1093/nar/gkx760>
6. Bagaev DV, Vroomans RMA, Samir J *et al.* VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res* 2020;48:D1057–62. <https://doi.org/10.1093/nar/gkz874>

7. Goncharov M, Bagaev D, Shcherbinin D *et al*. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. *Nat Methods* 2022;19:1017–9. <https://doi.org/10.1038/s41592-022-01578-0>
8. Kremlyakova Y, Vlasova EK, Luppov D *et al*. TCREMP: A bioinformatic pipeline for efficient embedding of T-cell receptor sequences from immune repertoire and single-cell sequencing data. *J Mol Biol* 2025;437:169205. <https://doi.org/10.1016/j.jmb.2025.169205>
9. Pogorelyy MV, Shugay M. A framework for annotation of antigen specificities in high-throughput T-cell repertoire sequencing studies. *Front Immunol* 2019;10:2159. <https://doi.org/10.3389/fimmu.2019.02159>
10. Satopää V, Albrecht J, Irwin D *et al*. Finding a “Kneedle” in a haystack: detecting knee points in system behavior. In: *2011 31st International Conference on Distributed Computing Systems Workshops*. Minneapolis, MN, USA: IEEE, 2011, 166–71. <https://doi.org/10.1109/ICDCSW.2011.20>
11. Wu KE, Yost K, Daniel B *et al*. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-binding analyses. In: *Proceedings of the 18th Machine Learning in Computational Biology meeting*, Vol. 240 of *Proceedings of Machine Learning Research*. Seattle, WA, USA: PMLR, 2024, 194–229.
12. Lin Z, Akin H, Rao R *et al*. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;379:1123–30. <https://doi.org/10.1126/science.ade2574>
13. Yin R, Ribeiro-Filho HV, Lin V *et al*. TCRmodel2: high-resolution modeling of T cell receptor recognition using deep learning. *Nucleic Acids Res* 2023;51:W569–76. <https://doi.org/10.1093/nar/gkac356>
14. Heather JM, Spindler MJ, Alonso MH *et al*. Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information. *Nucleic Acids Res* 2022;50:e68. <https://doi.org/10.1093/nar/gkac190>
15. Jumper J, Evans R, Pritzel A *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–9. <https://doi.org/10.1038/s41586-021-03819-2>
16. Quast NP, Deane CM, Raybould MIJ. STCRpy: a software suite for T-cell receptor structure parsing, interaction profiling, and machine learning dataset preparation. *Bioinformatics* 2025;41:btaf566. <https://doi.org/10.1093/bioinformatics/btaf566>
17. Rius C, Attaf M, Tungatt K *et al*. Peptide-MHC Class I tetramers can fail to detect relevant functional T cell clonotypes and underestimate antigen-reactive T cell populations. *J Immunol* 2018;200:2263–79. <https://doi.org/10.4049/jimmunol.1700242>
18. Yang G, Wang J, Sun P *et al*. SARS-CoV-2 epitope-specific T cells: immunity response feature, TCR repertoire characteristics and cross-reactivity. *Front Immunol* 2023;14:1146196. <https://doi.org/10.3389/fimmu.2023.1146196>
19. Nakama T, Wall A, Dolton G *et al*. Position-5-driven reorientation of an immunodominant HLA-A*24:02 SARS-CoV-2 epitope drives universal T cell escape. *JCI insight* 2026;11: e202235. <https://doi.org/10.1172/jci.insight.202235>
20. Banerjee A, Pattinson DJ, Wincek CL *et al*. T cell receptor cross-reactivity prediction improved by a comprehensive mutational scan database. *Cell systems* 2025;16:101345. <https://doi.org/10.1016/j.cels.2025.101345>
21. Croce G, Lani R, Tardivon D *et al*. Phage display enables machine learning discovery of cancer antigen-specific TCRs. *Sci Adv* 2025;11:eads5589. <https://doi.org/10.1126/sciadv.ads5589>
22. Gaglione SA, Mukkamala RS, Krishna C *et al*. Scalable TCR synthesis and screening enable antigen reactivity mapping in vitiligo. *Immunity* 2026;59:477–93. <https://doi.org/10.1016/j.immuni.2026.01.001>
23. Gonzalez-Galarza FF, McCabe A, Santos EJMD *et al*. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res* 2020;48:D783–88. <https://doi.org/10.1093/nar/gkz1029>
24. Klinger M, Pepin F, Wilkins J *et al*. Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PLoS One* 2015;10:e0141561. <https://doi.org/10.1371/journal.pone.0141561>
25. Snyder TM, Gittelman RM, Klinger M *et al*. Magnitude and dynamics of the T-cell response to SARS-CoV-2 infection at both individual and population levels. *Front Immunol* 2024;15:1488860. <https://doi.org/10.3389/fimmu.2024.1488860>
26. Chen G, Yang X, Ko A *et al*. Sequence and structural analyses reveal distinct and highly diverse human CD8⁺ TCR repertoires to immunodominant viral antigens. *Cell Rep* 2017;19:569–83. <https://doi.org/10.1016/j.celrep.2017.03.072>
27. Song I, Gil A, Mishra R *et al*. Broad TCR repertoire and diverse structural solutions for recognition of an immunodominant CD8⁺ T cell epitope. *Nat Struct Mol Biol* 2017;24:395–406. <https://doi.org/10.1038/nsmb.3383>
28. Pogorelyy MV, Fedorova AD, McLaren JE *et al*. Exploring the pre-immune landscape of antigen-specific T cells. *Genome Med* 2018;10:68. <https://doi.org/10.1186/s13073-018-0577-7>
29. Komech EA, Koltakova AD, Barinova AA *et al*. TCR repertoire profiling revealed antigen-driven CD8⁺ T cell clonal groups shared in synovial fluid of patients with spondyloarthritis. *Front Immunol* 2022;13:973243. <https://doi.org/10.3389/fimmu.2022.973243>
30. Dolton G, Tungatt K, Lloyd A *et al*. More tricks with tetramers: a practical guide to staining T cells with peptide-MHC multimers. *Immunology* 2015;146:11–22. <https://doi.org/10.1111/imm.12499>
31. Pogorelyy MV, Minervina AA, Shugay M *et al*. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biol* 2019;17:e3000314. <https://doi.org/10.1371/journal.pbio.3000314>
32. Glanville J, Huang H, Nau A *et al*. Identifying specificity groups in the T cell receptor repertoire. *Nature* 2017;547:94–8. <https://doi.org/10.1038/nature22976>
33. Richardson E, Aarts YJM, Altin JA *et al*. IMMREP25: Unseen Peptides. *bioRxiv*, <https://doi.org/10.64898/2026.03.30.715276>, 1 April 2016, preprint: not peer reviewed.
34. Dratva LM, Teichmann SA, Kretschmer L. Decoding human T cell immunity with artificial intelligence and single-cell genomics. *Annu Rev Immunol* 2026;44:381–405. <https://doi.org/10.1146/annurev-immunol-082724-011631>
35. Drost F, Chernysheva A, Albahar M *et al*. Benchmarking of T cell receptor-epitope predictors with ePytope-TCR. *Cell Genom* 2025;5:100946. <https://doi.org/10.1016/j.xgen.2025.100946>
36. Nielsen M, Eugster A, Jensen MF *et al*. Lessons learned from the IMMREP23 TCR–epitope prediction challenge. *ImmunoInformatics* 2024;16:100045. <https://doi.org/10.1016/j.immuno.2024.100045>