ImmuneWatch

W H I T E P A P E R

How Reliable Are Antigen-Specificity Annotations for T-Cell Receptors?

An overview of the state-of-the-art and validation of computational tools for TCR-antigen specificity annotations

immunewatch.com

ImmuneWatch BV

Table of Contents

Summary

Despite their potential, **concerns regarding the biological "ground truth" of computational TCR-specificity predictions remain**. This white paper addresses those concerns head-on. We will explore the current state of the art in TCR-peptide annotation tools, providing a data-driven defense of their reliability and demonstrating how integrating computational intelligence into your pipeline is now a competitive necessity.

Background

In the era of precision medicine, the T-cell receptor (TCR) sits at the center of our most promising therapeutic frontiers. Whether you are engineering the next generation of adoptive cell therapies, monitoring patient responses to novel vaccines, or unmasking the drivers of autoimmune pathologies, one fundamental question remains: **What does this specific T-cell actually recognise?**

Identifying the "cognate peptide", the specific antigen target of a TCR, is the "Holy Grail" of modern immunology. However, as the field has moved from small scale lab experiments to personalised, repertoire-scale data, the industry has hit a significant throughput ceiling. In other words, **"wet-lab" technologies for T-cell specificity are not scalable enough for today's throughput.**

The strategic importance of antigen specificity

The ability to accurately annotate TCR specificity is the difference between a high-affinity hit and a costly clinical failure. This capability is currently the primary driver in three critical sectors:

- **T-cell therapy quality control:** Ensuring that the induced T-cells after treatment or in a product have the targeted specificity, and aren't contaminated with bystanders.
- **Precision immunomonitoring:** Tracking the evolution of a patient's immune response in real-time during clinical trials to validate drug efficacy.
- **Target discovery in autoimmunity and oncology:** Deconvoluting the complex TCR repertoires of patients to identify the "self-antigens" that elicit an immune response.

The scalability of experimental specificity determination

We currently possess a robust arsenal of wet-lab technologies designed to bridge the gap between T-cell receptors and the cognate peptide they recognise. Techniques such as MHC-multimers, peptide stimulation assays, and high-throughput TCR cloning or yeast/phage display have served as the gold standards for decades. However, **these methods face an inherent scaling paradox.** Wet-lab assays are **labor-intensive**, **require significant biological material**, and are prohibitively **expensive** when applied to entire repertoires or the full antigen space. In addition, the "design-test-learn" cycle in a purely experimental environment is too slow to keep pace with the rapid clinical timelines required for personalised medicine.

The credibility of computational specificity annotation

To solve this, the industry has turned toward **computational annotations**. By leveraging state-of-the-art machine learning techniques and TCR-epitope databases, these tools can scan millions of sequences and **predict antigen binding in a fraction of the time and cost of an in vitro assay.**

Despite their potential, concerns regarding the biological "ground truth" of digital predictions remain. This white paper addresses those concerns head-on. We will explore the current state of the art in TCR-peptide annotation tools, providing a **data-driven defense of their reliability** and demonstrating how **integrating computational intelligence into your pipeline is now a competitive necessity**.

Three Major Categories of TCR-Antigen Annotation

Not all algorithms are created equal; they vary in their underlying logic, their data requirements, and their ultimate predictive power. We can categorize the current landscape of TCR annotation into three distinct tiers, each representing an increase in both computational complexity and clinical utility:

1) Matching TCRs with curated databases
2) Seen epitope algorithms
3) Unseen epitope algorithms

Strategy 1: Matching TCRs with curated databases

The most straightforward annotation system is **looking up if the TCRs in your datasets have an annotation in a curated database**. However this seemingly simple task is hindered by several TCR-specific challenges:

Challenge 1: What even counts as a match?

While it may seem trivial to search for the "same" TCR sequence, each TCR complex is composed of a heterodimer (αβ or γδ) with three diverse complementary determining regions (CDR1, CDR2, CDR3). Many TCR datasets are derived from a targeted sequencing approach that only captures a part of the full TCR complex. **So with this partial availability of data, when is it enough to be considered a match?** Is it enough to conclude from solely the beta CDR3, which is the most diverse but not the whole complex either?

Furthermore, the **likelihood of finding an exact match between your own data and the small number of TCRs with known epitope targets is small** as the potential TCR space is large, with estimates up to 10^{26} (not accounting for public TCRs as discussed below). It can be tempting to allow some mismatch to expand the number of hits. However, even conservatively allowing just a single amino acid mismatch has a massive impact on the reliability of TCR annotations, with a minimum of 30% false positives even on small data sets [1].

Finally there is every TCR bioinformatician's worst nightmare, the V gene annotations. There are several conflicting V gene reference nomenclatures, without a clear mapping between them. What's more is that even when using the same nomenclature, alignment pipelines will make widely different calls on the often ambiguous mappings. This means that **what might be considered as a specific V gene in the database, is not the same as what is considered that same V gene in your data**. Again while it is

tempting here to simply disregard the V gene with its ambiguity, this has a large impact on reliability, doubling the amount of noise in the annotations [2].

Challenge 2: False positives in the databases

The experiments used to determine TCR-epitope specificity are inherently noisy. As the amount of true epitope-specific TCRs are rare, **this often results in a large fraction of false positives in the current databases**. Experimental validation of select pairs has revealed that at least 50% of even the best curated datasets are likely false positives [3]. This is a high error rate that signifies that even an exact full TCR match has a high chance to be incorrect.

Challenge 3: TCR cross-reactivity

The potential binding space of a single TCR has been estimated to be around one million epitopes. This is because the TCR complex only makes contact with some of the residues of the epitope, leaving the others free to be different. For this reason, TCRs can be cross-reactive against different epitopes from diverse origins. Currently the extent of this cross-reactivity is poorly understood for most TCRs. **Therefore even annotating correctly that a TCR binds epitope A, doesn't mean that it wasn't activated by a similar — but distinct — epitope B in your dataset.**

Challenge 4: TCR publicity

Not all TCRs are created equally: it is well known that some TCRs have a higher chance to be generated during the recombination process. Notably those that are close to germline, with few if any nucleotide additions and deletions, have a significantly higher probability and are known to occur more frequently, often acting as hubs within the repertoire. And because of their high chance of generation, they are more often than not shared between individuals, i.e. public clonotypes. Therefore exact database matching often results in matching the public clonotypes. This creates a problematic bias in the annotation output. As public clonotypes are highly frequent, they have a high chance of being picked up as a false positive (a common prevalent clone has a higher background level than a rare clone). As exact matching will mostly work for public clonotypes and these have some of the least reliable annotation, this signifies that annotating a TCR repertoire will result in very few true annotations, except for the most common antigens triggering the most common TCRs.

ImmuneWatch has established the largest database of TCR-epitope pairs

The ImmuneWatch database collected more than half a million reported TCR-epitope pairs, which has been refined to a high quality dataset of **134,476 interactions**.

Through this curated effort, we ensure that only the most reliable data forms the basis of our annotations with careful notation of the underlying confidence of each record.

A dedicated curation team within ImmuneWatch continuously updates with the newest information, enabling a new database release every three months.

This database is integrated in ImmuneWatch DETECT, allowing for full repertoire annotation using the best practices.

Strategy 2: Seen epitope algorithms

Clearly, exact matching your TCRs to a database has its problems. To address these issues, specific algorithms have been developed that smartly leverage all the data collected in a database to annotate TCRs. **These aid in deciding when a TCR is considered a hit, reduce the amount of false positives, and extend the annotation space beyond the public clonotypes.** These types of algorithms have been an active research field in the past ten years, and numerous solutions have been developed going from dedicated distance measures to full machine learning models to distinguish binders from non-binders.

As each method claims to have the best performance in their own publications, independent benchmarking competitions have been set up. At the end of 2023, the IMMREP competition provided a blinded dataset wherein teams around the world were asked to annotate the epitope specificity [4]. This dataset contained 598 previously undocumented TCRs binding one of 20 epitopes across 6 different HLA backgrounds. **Around 50 different models participated, and was eventually won by ImmuneWatch DETECT.**

> In the words of the organisation: *"Most methods in the G2 group have comparable performance with the exception of **IMW DETECT** that shows a **substantial predictive advantage**."*
> — IMMREP23 competition report [4]

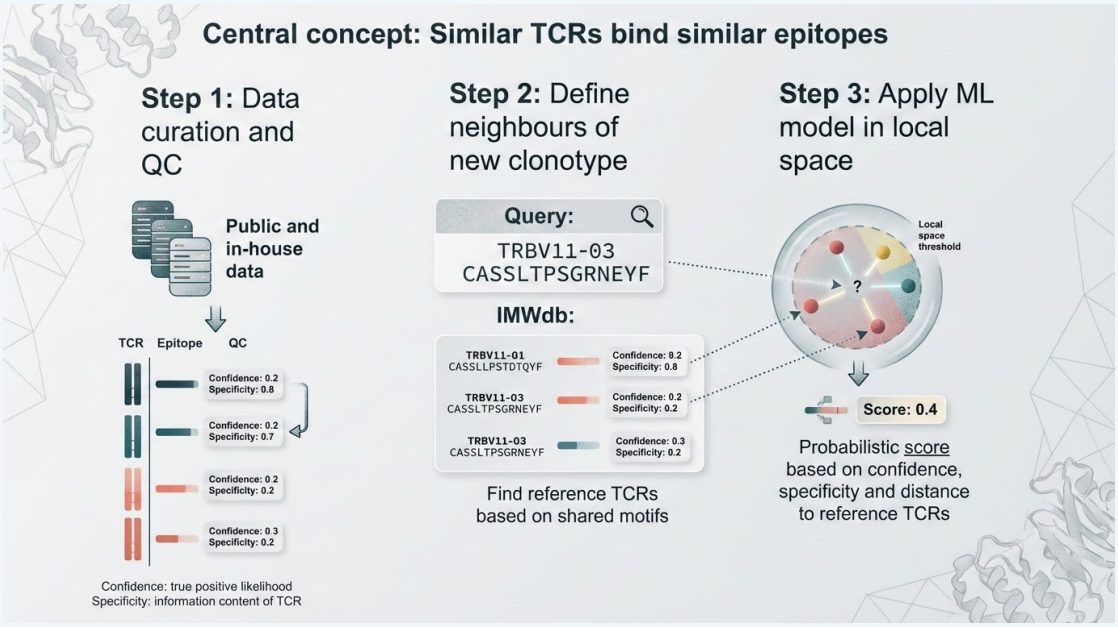
Since then, ImmuneWatch DETECT has become the golden standard for epitope TCR annotation.

In addition, the results of this international benchmark challenge demonstrated that **seen epitope algorithms had an exceptionally high performance** that outperformed any kind of traditional lookup methods, even when faced with complex repertoire data. When a model achieves an AUC higher than 0.85 on a fully blinded independent dataset, such as is the case for DETECT, **it can be seen as a highly-confident filtering and annotation tool that is ready for integration into research and discovery pipelines.**

How ImmuneWatch DETECT works in a nutshell

T-cell receptor annotation based on a custom **probabilistic framework**: DETECT is built on a dedicated machine learning framework designed from the ground up for T-cell receptor sequence data. It judges for any input TCR sequence if it has enough information available to accurately assign an epitope. This is based on a combination of metrics, such as the confidence and specificity of TCR-epitope pairs, that judge the false positive rate. DETECT has been tuned to be extremely strict in its predictions, so that high scores (>0.2) are accurate for more than half of the annotations, even when screening a full repertoire with millions of T-cell receptor sequences.

Full transparency of the prediction: The central concept behind the algorithm is that similar T-cell receptor sequences bind similar epitopes. All DETECT predictions are accompanied by a list of reference TCRs, which are the TCR with shared motif features that are driving the prediction, along with the literature evidence that supports this interaction.



Beyond the IMMREP23 validation, which considered predicting a small number of TCRs with a known ground truth, ImmuneWatch DETECT has been validated in the following ways:

1) **External validation** of annotations with dedicated experiments: Our recent study on annotating the epitope specificity of TIL products featured an extensive *in vitro* validation of specific TCR-epitope pairs [5]. This resulted in a **100% accuracy rate** of all selected annotations, despite the enormity of the potential epitope space. This proves that DETECT isn't just the best performing algorithm but that it can identify high-affinity therapeutic candidates from an enormous search space with surgical precision.

2) Monitoring of annotation robustness through **best practices MLops**: Every change to DETECT or its underlying database is benchmarked across over 400 independent *in silico* validation experiments, where the performance is quantified and the false positive rate is assessed. This is the equivalent of completing over a hundred thousand validation experiments with each new line of code.

3) Confirmation of annotations at a **patient level**: DETECT is meant to be used on full repertoires, and go beyond simply recapitulating dextramer staining. If its annotations are correct, it should work on a full repertoire level for diagnostics. When applied to a large cohort (160 individuals) of Covid patients versus healthy controls, half of all patients could be picked out solely based on the amount of SARS-CoV-2 annotations with a zero percent false positive rate. Similarly the predictions made by DETECT can be used to HLA type the individuals of origin with high accuracy (0.93 AUC) [7]. As DETECT can accurately diagnose a patient based on repertoire-wide annotations, it confirms that the underlying molecular predictions are robust enough for diagnostic and monitoring applications.

4) **User-validations.** ImmuneWatch DETECT (or its predecessor TCRex*) has aided in the discovery of novel immunological insights in the following papers:

a. *Ha, My, et al. "Generation of a T cell receptor, cytokine and cell repertoire synovial fluid atlas to define commonalities and dissimilarities between arthritic diseases through systems immunology approaches." bioRxiv (2025): 2025-01.*

b. *Llaó-Cid, L., et al. "Integrative multi-omics reveals a regulatory and exhausted T-cell landscape in CLL and identifies galectin-9 as an immunotherapy target." Nature Communications 16.1 (2025): 7271.*

c. *Gielis, Sofie, et al. "Analysis of Wilms' tumor protein 1 specific TCR repertoire in AML patients uncovers higher diversity in patients in remission than in relapsed." Annals of Hematology 104.1 (2025): 317-333.**

d. *Mellors, Patrick, et al. "Cytotoxic, natural killer-like ex-tissue resident memory T-cells circulate in human chronic graft-versus-host disease at diagnosis." Blood 146 (2025): 821.*

e. *Brand, Eelco C., et al. "Clonal overlap and convergent clustering of T-cell receptor signatures in Crohn's disease in monozygotic twins." bioRxiv (2025): 2025-10.*

f. *Khare, Kriti, et al. "Temporal TCR dynamics and epitope diversity mark recovery in severe COVID-19 patients." Frontiers in Immunology 16 (2025): 1582949.*

g. *Moghbeli, Kaveh, et al. "NKG2D blockade impairs tissue-resident memory T cell accumulation and reduces chronic lung allograft dysfunction." JCI insight 10.4 (2025): e184048.*

h. *Mittl, Kristen, et al. "Antigen specificity of clonally-enriched CD8+ T cells in multiple sclerosis." bioRxiv (2024).**

i. *Flumens, Donovan, et al. "Training of epitope-TCR prediction models with healthy donor-derived cancer-specific T cells." Methods in Cell Biology. Vol. 183. Academic Press, 2024. 143-160.**

j. *Paran, Faith Jessica, et al. "BCR, not TCR, repertoire diversity is associated with favorable COVID-19 prognosis." Frontiers in Immunology 15 (2024): 1405013.**

k. *Mullan, Kerry A., et al. "T cell receptor–centric perspective to multimodal single-cell data analysis." Science Advances 10.48 (2024): eadr3196.*

l. *Postovskaya, Anna, et al. "Leveraging T-cell receptor–epitope recognition models to disentangle unique and cross-reactive T-cell response to SARS-CoV-2 during COVID-19 progression/resolution." Frontiers in Immunology 14 (2023): 1130876.**

m. *Kashima, Yukie, et al. "Intensive single-cell analysis reveals immune-cell diversity among healthy individuals." Life Science Alliance 5.7 (2022).**

n. *Johnson, Douglas B., et al. "A case report of clonal EBV-like memory CD4+ T cell activation in fatal checkpoint inhibitor-induced encephalitis." Nature Medicine 25.8 (2019): 1243-1250.**

Strategy 3: Unseen epitope models

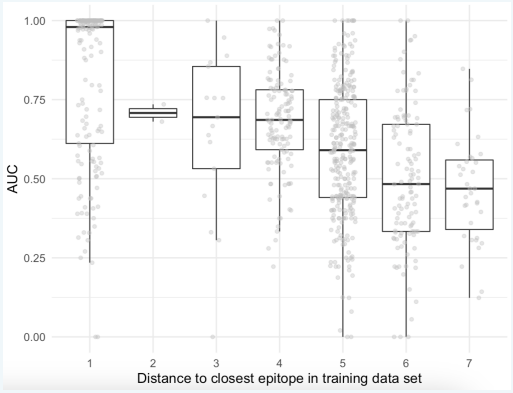
While identifying T-cells for well-characterised antigens (like common viral or cancer proteins) is now highly reliable, the ultimate goal is generalisable prediction. This means predicting the binding of a TCR to a completely novel “unseen” epitope, one that has never appeared in any database or training set. Once developed, it would allow for the instant identification of TCRs for rare patient-specific neoantigens or orphan autoimmune triggers without any prior lab data.

The IMMREP competition in 2025 [6] highlighted that while we have made massive strides, unseen epitope prediction remains an unsolved challenge for the entire industry. There are two primary reasons for this. The first is the complexity of the problem, where even advanced molecular modelling approaches (such as AlphaFold) still barely manage to capture the complexity of the TCR-pMHC complex. The second is the lack of available data, which is growing year-on-year, but is still insufficient to train generalisable models.

Despite these hurdles, the industry is converging on a two-pronged solution that ImmuneWatch is actively pursuing:

- **Hybrid Modeling:** We are moving toward models that combine the speed of ImmuneWatch DETECT with the biological accuracy of structural physics. This allows us to pre-filter the repertoire at high speed and then use high-fidelity structural modeling on only the most promising candidates.
- **Massive Data Generation:** The biggest force multiplier for unseen prediction will be the continued generation of high-quality, paired-chain TCR data. As our internal databases grow, what is “unseen” today becomes “seen” tomorrow.

Performance of ImmuneWatch DETECT on unseen epitopes



Predictive performance for unseen epitopes based on the closest example in the training data set.

The framework of ImmuneWatch DETECT allows it to be extended to make **predictions for any given epitope**, even if they are not present in the database. While these are less reliable predictions, they still hold value in those instances where one must prioritise the most likely candidate for a given TCR.

To illustrate this usage, ImmuneWatch DETECT was benchmarked to monitor T-cell reactivity against neo-epitopes, which are by definition unseen. The predictions it gave matched those found in an ELISPOT assay [8].

Conclusion

We are at a turning point in immunology. **Relying exclusively on wet-lab technologies to annotate TCR repertoires is no longer a viable strategy for companies aiming to lead in the T-cell therapy or autoimmune space.**

The research field has progressed steadily to the point where now:

- Computational tools like ImmuneWatch DETECT are no longer experimental side-projects; they are **validated discovery engines** that have proven their accuracy in independent benchmarks (IMMREP) and real-world patient cohorts.
- By integrating these tools, **you can transform a needle in a haystack search into a prioritised list of high-affinity candidates**, reducing your R&D timelines from months to days.
- The frontier of "unseen" prediction is rapidly closing. **Early adopters who integrate these AI-driven workflows now will be best positioned to capitalise on the next generation of precision immunotherapies.**

References

[1] Meysman P, De Neuter N, Gielis S, Bui Thi D, Ogunjimi B, Laukens K. On the viability of unsupervised T-cell receptor sequence clustering for epitope preference. Bioinformatics. 2019 May 1;35(9):1461-8.

[2] Meysman P, Barton J, Bravi B, Cohen-Lavi L, Karnaukhov V, Lilleskov E, Montemurro A, Nielsen M, Mora T, Pereira P, Postovskaya A. Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. ImmunoInformatics. 2023 Mar 1;9:100024.

[3] Messemaker M, Kwee BP, Moravec Ž, Álvarez-Salmoral D, Urbanus J, de Paauw S, Geerligs J, Voogd R, Morris B, Guislain A, Mußmann M, et al. A functionally validated TCR-pMHC database for TCR specificity model development. BioRxiv. 2025 May 12.

[4] Nielsen M, Eugster A, Jensen MF, Goel M, Tiffeau-Mayer A, Pelissier A, Valkiers S, Martínez MR, Meynard-Piganeeau B, Greiff V, Mora T, et al. Lessons learned from the IMMREP23 TCR-epitope prediction challenge. ImmunoInformatics. 2024 Dec 1;16:100045.

[5] Van Houcke M, Wuyts S, Bosschaerts T, Harari A, Chiffelle J, Auger A, Coukos G, Meysman P. Applications of T-cell receptor specificity annotation models for quality control and immunomonitoring in adoptive T-cell therapies. bioRxiv. 2026:2026-01.

[6] Richardson E, et al. IMMREP25: Unseen Peptides. 2026

[7] https://immunewatch.gitlab.io/detect-docs/Walkthroughs/hla-typing-a-tcr-repertoire

[8] https://immunewatch.gitlab.io/detect-docs/Walkthroughs/monitoring-neoantigens-in-cancer-vaccination