A functionally validated TCR-pMHC database for TCR specificity model development.

Marius Messemaker[1,2], Bjørn P.Y. Kwee[1,2,#], Živa Moravec[1,#], Daniel Álvarez-Salmoral[2,3], Jos Urbanus[1,2], Sam de Paauw[1,2], Jeroen Geerligs[1,2], Rhianne Voogd[1,2], Ben Morris[4], Aurélie Guislain[1,2], Maike Mußmann[1,2], Yaël Winkler[1,2], Maxime Steinmetz[1,2], Matyas Iras[1,2], Eric Marcus[4,5], Jonas Teuwen[4,5], Anastassis Perrakis[2,3], Roderick L. Beijersbergen[4], Wouter Scheper[1,^,], and Ton N. Schumacher[1,2,6,^,*].*

[1]Division of Molecular Oncology & Immunology, The Netherlands Cancer Institute, Amsterdam, The Netherlands.

[2]Oncode Institute, Utrecht, The Netherlands.

[3]Division of Biochemistry, The Netherlands Cancer Institute, Amsterdam, The Netherlands.

[4]Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, The Netherlands.

[5] AI for Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands.

[6]Department of Hematology, Leiden University Medical Center, Leiden, The Netherlands.

[#] These authors contributed equally.

[^] These authors contributed equally.

**Correspondence to: w.scheper@nki.nl, t.schumacher@nki.nl*

Summary

Accurate prediction of TCR specificity forms a holy grail in immunology and large language models and computational structure predictions provide a path to achieve this. Importantly, current TCR-pMHC prediction models have been trained and evaluated using historical data of unknown quality.

Here, we develop and utilize a high-throughput synthetic platform for TCR assembly and evaluation to assess a large fraction of VDJdb-deposited TCR-pMHC entries using a standardized readout of TCR function. Strikingly, this analysis demonstrates that claimed TCR reactivity is only confirmed for 50% of evaluated entries. Intriguingly, the use of TCRbridge to analyze AlphaFold3 confidence metrics reveals a substantial performance in distinguishing functionally validating and non-validating TCRs even though AlphaFold3 was not trained on this task, demonstrating the utility of the validated VDJdb (TCRvdb) database that we generated. We provide TCRvdb as a resource to the community to support training and evaluation of improved predictive TCR specificity models.

Keywords

T cell receptor, major histocompatibility antigen, peptide, predictive models, functional genetic screening

Introduction

The T cell-based adaptive immune system provides the body with the capacity to eliminate infected or malignant cells. T cells specifically detect antigenic peptides presented by major histocompatibility complex (MHC) molecules through their heterodimeric αβ T cell receptor (TCR). During thymic development, T cells acquire their TCR α and β chains through somatic recombination of V(D)J gene segments, along with random nucleotide insertion and deletion at the resulting gene junctions, generating an estimated 10^8-10^{10} different TCR clones with distinct reactivity patterns in an individual[1]. This 'distributed recognition capacity' equips the human T cell compartment with the capacity to e.g., detect and respond to pathogens, including those yet to emerge, and to eliminate malignant cells that present cancer (neo-)antigens[2]. Ever since the discovery of MHC restriction by Doherty and Zinkernagel[3] and the cloning of the first T cell receptors by the groups of Mak and Davis[4–6] immunologists across scientific fields have aimed to identify the antigens and T cell receptors that drive autoimmune disease[7] and cancer regression[8], and that can inform vaccine design[9–12].

Given the essential role of T cell recognition in a variety of human conditions, and given the inability to experimentally map this vast interaction space, *in silico* models that can predict TCR-peptide-MHC reactivity hold a major promise for our understanding of adaptive T cell immunity. As an example, in immuno-oncology, such predictive models may enable the selective boosting of tumor-reactive T cell responses, or may facilitate patient stratification based on the breadth of the tumor-reactive T cell pool. Beyond cancer, predictive models of TCR specificity could advance our understanding of autoimmune disease by revealing which self-antigens are recognized by autoreactive T cells, or aid in rational vaccine design. Ultimately, robust prediction models of TCR specificity may enable the *in silico* design and iterative optimization of synthetic TCRs that are free from the constraints of the natural TCR pool.

Recent advances in deep-learning-based approaches have resulted in protein structure prediction models such as Alphafold and RoseTTAFold that show unprecedented accuracy[13–15]. The remarkable progress in these protein structure prediction models has relied heavily on the large volume of experimental structural information available in the Protein Data Bank (PDB)[16], serving as a training data resource for deep learning approaches. Importantly, PDB provides multiple standardized validation measures to assess the quality of deposited structures, such as R_{free} and the proportion of Ramachandran outliers[17,18], and development of this standardized

104 validation process has been associated with an increase in the quality of deposited protein
105 structures over time[18].
106
107 As compared to the prediction of protein structures and protein complexes in other fields,
108 progress in predicting reactive (i.e., inducing TCR signaling) TCR-pMHC pairs has to date been
109 modest[19,20]. Part of the complexity of this task is related to the low (μM range) affinity of TCR-
110 pMHC interactions and the fact that the binding site of TCRs is composed of loops with
111 considerable conformational flexibility. However, we speculated that the quality of datasets that
112 have been used to train and evaluate TCR-pMHC predictors could also form a major factor that
113 has limited progress in this field. The primary training and evaluation[21] data resource for models
114 of TCR specificity has to date been formed by the VDJdb[22] and IEDB[23] databases. Importantly, the
115 TCR-pMHC pairs that make up these databases have been identified in a variety of studies using
116 diverse assays and sample sources. With the goal to increase data quality, the TCR-pMHC pairs
117 that are deposited in these databases are curated using assay information. In addition, VDJdb
118 data contain confidence scores based on presumed indicators of assay quality. However,
119 whether such confidence scores hold similar value as e.g., PDB validation metrics such as R_{free}
120 has not been established.
121
122 With the aim to provide standardized metrics of TCR reactivity, we here develop a scalable semi-
123 synthetic platform to evaluate antigen recognition. We then utilize this platform to demonstrate
124 that the quality of data used to create TCR-pMHC prediction models is concerningly low. We also
125 demonstrate that the performance of established sequence- and structure-based models of TCR
126 specificity has been underestimated because of this issue. Finally, we provide a validated
127 dataset that should allow the more effective training of novel TCR-pMHC prediction models.
128
129
130 **Results**
131 ***Exploration of data quality in VDJdb***
132 The VDJdb MHC-I database contains approximately 6,300 TCRαβ entries that are dominated by
133 a modest number of well-studied viral epitopes, with ten epitopes making up about half of the
134 database (Supplementary Table 1). To assess the quality of TCR-pMHC pairs contained in this
135 database, we focused on the HLA-A*02:01-restricted GLCTLVAML (GLC) Epstein–Barr virus (EBV)
136 and YLQPRTFLL (YLQ) severe acute respiratory syndrome coronavirus 2 (SARS-Cov2) epitopes,
137 representing epitopes for which TCR discovery has either been pursued for decades or is

138 restricted to the past few years. For each epitope, a set of TCRs with more similar or distinct
139 sequences was selected. Following manual synthesis and arrayed introduction into $CD8^+$ Jurkat
140 T cells, capacity of these TCRs to recognize either the cognate GLC or YLQ epitope was evaluated
141 in co-cultures using CD69 expression as a measure of TCR signaling. Strikingly, of the 49 YLQ-
142 annotated and 41 GLC-annotated TCRs evaluated, the claimed reactivity was only observed in
143 34.7% and 63.4% of cases, respectively (Figure 1A and 1B). To understand whether this result
144 was influenced by the assay system used, a subset of validating and non-validating TCRs was
145 introduced into primary human peripheral blood $CD8^+$ T cells. Using CD137 expression as a
146 measure of TCR signaling in human $CD8^+$ T cells, a high level of agreement between the two assay
147 systems was observed (Figures 1C and 1D). Comparison of validating and non-validating TCRs
148 based on tcrdist3 sequence similarity suggests that sequence conservation may be used to
149 predict reactivity for the majority of YLQ TCRs but less so for GLC-reactive TCRs (Figure 1E).
150 Jointly, these analyses provide evidence that the data quality of VDJdb is poor, potentially limiting
151 its value for TCR specificity model training and evaluation.
152
153 ***Scalable synthetic TCR library assembly and characterization***
154 To allow the evaluation of a major fraction of TCR-pMHC entries in VDJdb or in future datasets,
155 we set out to create a high-throughput synthetic system for TCR generation and functional
156 screening. Value of such a synthetic system is formed by its sole requirement for TCR sequence
157 and epitope sequence information as input. In addition, use of such a synthetic system removes
158 variation in T cell (dys)function as a confounder. Assembly of large synthetic TCRαβ libraries
159 requires the precise combination of defined TCRα and β chains into gene segments of
160 approximately 1.8 kilobase that are difficult to synthesize in a cost-efficient and hence scalable
161 manner. Previous approaches have aimed to increase scalability by arrayed TCR assembly of
162 TRAV- and TRBV-encoding vectors in combination with TCR-specific CDR3α/β-Jα/β-encoding
163 oligonucleotides[1,2], but throughput remains limited by the requirement for individual custom
164 oligonucleotides. As an alternative, recent work has described a scalable TCR assembly
165 approach that relies on oligonucleotide pools that encode all required CDR3α/β-Jα/βs segments
166 in combination with re-usable TRAV- and TRBV-encoding vector stocks in a one-pot reaction[26].
167 However, relatively low sequence accuracy and imbalanced composition of resulting TCR
168 libraries make it difficult to ensure sensitive screening of all TCRs within libraries, especially for
169 those TCRs that are present at low frequencies.
170

171 Building on these approaches, we reasoned that a scalable robotics-based TCR assembly
172 method could be developed that combines high assembly accuracy with the scalability of
173 oligonucleotide pools (Figures 2A and S1A). In brief, to accurately demultiplex oligonucleotide
174 pools to achieve assembly of unique TCRs in a TCR library of interest, we selected (see Methods)
175 and assigned primers from a set of highly orthogonal (non-interacting) primers[27] to wells in
176 plates, allowing the generation of PCR-amplified CDR3α-Jα and CDR3β-Jβ segments of defined
177 TCRs in each well. Joining of the resulting fragments by Golden Gate assembly with TRAV and
178 TRBV fragments and constant regions could then be used to yield full-length TCRβ-P2A-TCRα
179 gene segments encoded by viral vectors (for details see Methods). To ensure the intended
180 ligation during Golden Gate assembly, highly orthogonal four-base overhangs using a
181 comprehensive ligation fidelity dataset[28] were utilized (Figure S1B and S1C).
182
183 To evaluate the performance of this high-throughput TCR assembly method we created a diverse
184 set of 12,078 MHC class I- and II-restricted TCRs in five synthesis runs at a cost of €2.40/ TCR : i)
185 all 3,693 TCRs in VDJdb that are annotated as reactive with one of the top 10 most densely
186 sampled epitopes ("VDJdb-10" library), and ii-v) the TCR-s1 (3,510), TCR-s2 (983), TCR-s3
187 (1,000), and TCR-s4 (2,892) libraries that cover a range of tumor-derived and designed TCRs. To
188 estimate the percentage of sequence-perfect TCR assemblies, defined as TCR sequences with
189 correctly joined TCRβ-P2A-TCRα fragments with no mutations, insertions, or deletions, we
190 performed full-length TCR Sanger sequencing of 82 randomly selected individual TCR assembly
191 products. Sequence analysis of these bulk TCR assembly products from each of the five libraries
192 showed that >98% of TCRs (81 out of 82 TCRs evaluated) was successfully assembled as
193 sequence-perfect DNA molecules (Figure 2B). To extend this analysis of full-length TCR
194 sequences to entire TCR library assemblies, we developed an Oxford Nanopore Technologies
195 (ONT) full-length TCR sequencing method with sequencing error correction[29] to be able to count
196 sequence-perfect unique TCR molecules (Figure 2C, 2D, S2A, S2B). In brief, the read length of
197 ONT sequencing data in principle enables accurate estimation of Vα–CDR3α-Jα and Vβ–CDR3-
198 Jβ chain pairings. However, due to high sequence similarity between different TCR chains (Figure
199 2C), its raw base-call accuracy is insufficient to confidently call correctly paired or mispaired
200 TCRs, nor detect mutations or indels, and comprehensive QC metrics of TCR libraries have to
201 date not been available. To create a QC pipeline with single base error detection capacity, we
202 designed an ONT sequencing method that utilizes unique molecular identifiers (UMIs) to bin
203 reads, thereby enabling creation of highly accurate consensus sequences for individual TCR
204 molecules (Figure 2D, S2A and S2B). Resulting consensus sequences allow for precise

205 identification of sequence variations including mutations and indels, as well as correct and
206 incorrect chain pairings, as shown by a 'leave one plate out' repeat assembly experiment (Figure
207 S2C). Analysis of unique TCR molecules in each of the five TCR libraries showed successful
208 assembly of 96.9% (range: 96.2% - 98.6%) of all TCRs, with on average 74.2% sequence-perfect
209 unique TCR molecules (Figure 2E and S2D). Importantly, in line with the strategy for saturation
210 amplification of individual CDR3α/β-Jα/β segments (see Methods), library composition was
211 highly uniform, with abundance between the 5th and 95th percentiles in the five TCR libraries
212 varying on average only 5.6-fold (Figure 2E and S2D). Taken together, these data describe a TCR
213 assembly method that is highly accurate at the single-molecule and single-nucleotide level. In
214 addition, the UMI-based ONT approach for full-length TCR sequencing also provides a strategy
215 to distinguish reactive and non-reactive TCR sequences with single base resolution in genetic
216 screens (see below).
217
218 ***Pooled functional genetic screening of the VDJdb-10 TCR library***
219 To generate a high-confidence dataset for TCR specificity modeling, we set out to validate the
220 reactivity of the VDJdb-10 TCR library against the YLQ and GLC epitopes (Figure 3A), building on
221 a recently developed functional genetic screening approach to identify tumor-reactive TCRs in
222 human tumor-infiltrating lymphocyte (TIL) samples[26]. Following introduction of the VDJdb-10 TCR
223 library in TCR-null $CD8^+$ Jurkat T cells, library composition and uniformity was confirmed by ONT
224 sequencing of TCR^+ cells (Figure 3B). Subsequently, the VDJdb-10 Jurkat library was exposed to
225 either YLQ- or GLC-expressing target cells, and reactive T cells were isolated for ONT sequencing
226 by cell sorting of $CD69^+$ cells (Figure S3A). Using TCR abundance in the $CD69^+$ cell pools as a
227 measure of TCR signaling, reactivity could be confidently called for 171 out of 421 (40.6%) YLQ-
228 annotated TCRs and for 134 out of 193 (69.4%) GLC-annotated TCRs (Figure 3C). Importantly,
229 lack of claimed TCR reactivity was not explained by TCR abundance in the VDJdb-10 Jurkat library
230 (Figure S3B and S3C), nor by collapse of non-*01 TRAV and TRBV alleles to *01 alleles (Figure
231 S3D). Furthermore, TCR reactivity as measured by pooled functional screening was highly
232 consistent with the results obtained by the initial arrayed analyses, with a concordance of 96%
233 (Figure S3E). Interestingly, evaluation of features that were associated with lack of claimed
234 reactivity revealed a substantial variation in data quality between studies (Figure S3F).
235 Furthermore, the fraction of samples obtained from donors with a recent or ongoing COVID-19
236 infection across studies was highly predictive of the fraction of truly reactive TCRs as identified
237 here (Figure 3D), suggesting that antigen-driven expansion may help enrich for truly reactive
238 TCRs. Importantly, VDJdb confidence scores showed a low specificity and also low sensitivity in

239 identifying truly reactive TCRs for both epitopes (Figure 3E and 3F), highlighting the lack of reliable
240 quality control metrics in currently used TCR-pMHC databases. Collectively, these functional
241 screening data demonstrate that the key property of TCRs for predictive model development, i.e.,
242 their ability to signal upon encounter of endogenously presented antigen, is lacking in a large
243 fraction of annotated VDJdb TCRs. In addition, we provide a functionally validated VDJdb
244 database ("TCRvdb") for model evaluation, and a standardized evaluation platform to rapidly
245 expand this dataset to other TCR – antigen collections.
246
247 ***Functionally validated VDJdb improves TCR-pMHC reactivity model evaluation***
248 To understand whether the use of TCRvdb alters our interpretation of the precision and sensitivity
249 of existing TCR-pMHC reactivity prediction models, we evaluated the tcrdist3[30], STAPLER[19], and
250 Alphafold3[14] models (Figure 4A). tcrdist3 exploits sequence similarity to group TCRs with similar
251 sequences at defined thresholds. While, by its nature, tcrdist3 cannot be trained to predict
252 epitope reactivity, sequence similarity to one or multiple user-defined 'anchor' TCRs with known
253 epitope reactivities can be utilized to identify additional TCRs that may recognize the same
254 epitope. In contrast to tcrdist3, STAPLER is a supervised language model that was trained to
255 predict the epitope reactivity of TCRs, using the (frequently incorrect, see Figure 1 and 3)
256 reactivity labels provided by VDJdb (see Methods). Finally, Alphafold3 was utilized as a general
257 protein complex structure prediction model that has not explicitly been trained to predict the
258 epitope reactivity of TCRs. To predict TCR reactivity from the AlphaFold3 output, we developed
259 the TCRbridge model, which uses AlphaBridge[31] to combine AlphaFold3's three per-residue
260 predicted confidence metrics at the TCRα–peptide and TCRβ–peptide interfaces.
261 Importantly, analysis of the output of the functional genetic screening data (Figure 3) using these
262 three models showed that each of these displayed a moderate to substantial capacity to
263 distinguish validating and non-validating TCRs, but with important nuances (Figure 4 and S4).
264 First, while tcrdist3 showed a high capacity to distinguish truly YLQ-reactive TCRs from non-
265 validating TCRs with claimed YLQ-reactivity, performance in distinguishing validating and non-
266 validating TCRs with claimed GLC reactivities was modest (Figure 4B and S4A). Like tcrdist3, the
267 masked language-based sequence model STAPLER showed some performance in distinguishing
268 truly reactive from non-reactive TCRs, particularly for the YLQ epitope (Figure S4A). Notably, this
269 performance was observed in spite of the fact that STAPLER was fine-tuned using a dataset in
270 which both the validating and non-validating TCRs as identified here were all labeled as positives.
271 We conclude that noise in datasets can only efficiently be identified with these sequence-based
272 models in cases in which validating and non-validating TCRs are sufficiently distinct in sequence

space. In contrast, TCRbridge demonstrated substantial performance in distinguishing reactive from non-reactive TCRs for both the YLQ and GLC epitopes (Figures 4C, 4D, and S4A), and outperformed tcrdist3 and STAPLER in identifying GLC-reactive TCRs (Figure 4D). Predictive capacity was also obtained when using a single confidence measure (Figure S4B) or interface confidence between TCR subunits and MHC (Figure S4C and S4D). These data suggest that, despite not being trained to predict TCR-pMHC reactivity in any way, AlphaFold3 has a superior capacity to correctly predict pMHC reactivity in case of more complex TCR repertoires.

In future prediction tasks, TCRs that share the same HLA restriction element (e.g., HLA-A*02:01) may be compared for their potential reactivity with a given peptide-MHC. However, the HLA restriction element of the non-validating TCRs as described here is unknown and may not be HLA-A*02:01. Furthermore, because of the bias towards HLA-A2*02:01-restricted TCRs in the AlphaFold3 training dataset, TCRbridge could potentially assign higher confidence to complexes that contain HLA*02:01-restricted TCRs. To understand whether the performance of TCRbridge on the TCRvdb dataset could be due to a learned HLA-A*02:01-restriction class imbalance, we predicted interfaces between all validated GLC-reactive TCRs and either the cognate GLC-HLA-A*02:01 complex or the control YLQ-HLA-A*02:01 complex (for which these TCRs are experimentally confirmed true negatives, see Figure 3). Likewise, interfaces were predicted between all validated YLQ-reactive TCRs and either the cognate YLQ-HLA-A*02:01 complex or the control GLC-HLA-A*02:01 complex. Notably, TCRbridge still assigned meaningful confidence scores in this setting, indicating that model performance was not due to a learned HLA-A*02:01 restriction class imbalance (Figure 4E, 4F, S4E, and S4F). In contrast, this ability to distinguish TCR-pMHC complexes that differ by only the 9-mer epitope sequence was not seen with non-validating TCRs for either epitope (Figure S4G), further highlighting how TCRbridge performance estimates are dependent on data quality.

Discussion

When low quality datasets are used to evaluate predictive models, this may lead to an underestimate of model performance, especially when these models have not seen, and hence been biased by, these data during training. We here demonstrate that data quality is a significant concern for the VDJdb database that has been the starting point of TCR-pMHC prediction evaluation efforts[21]. We also share a technology to synthetically recreate TCR collections and validate their function in a standardized fashion. Finally, we exploit this platform to yield

307 validated datasets for two epitopes and then use these datasets to demonstrate that
308 performance of established sequence- and structure-based prediction pipelines is higher than
309 assumed[19,32,33].
310
311 The VDJdb database was created through the joint effort of many research groups that examine
312 T cell-mediated immune responses against pathogens, cancer antigens and other antigen
313 classes. In line with its bottom-up nature, contributors have significant freedom with respect to
314 assay systems (e.g., using different types of MHC multimers, using fresh or restimulated cell
315 populations, using material from acutely infected or healthy donors) and thresholds used to call
316 reactive TCR-pMHC pairs. Equally important, unlike in the field of structural biology, where well-
317 established metrics[34,35] are used to unambiguously describe data quality, it has been difficult to
318 define standardized validation metrics that comprehensively describe data quality of TCR-pMHC
319 pairs. We set out to evaluate previously proposed TCR-pMHC pairs in a well-defined
320 experimental setting to create harmonized and validated datasets for model development. A key
321 component in this effort has been the use of immortalized synthetic systems for both antigen and
322 TCR expression, in order to remove the variability associated with finite biological samples.
323 Importantly, as the goal of TCR-pMHC prediction efforts is to predict productive T cell signaling
324 upon encounter of antigen-expressing cells, we define TCR reactivity based on T cell activation
325 marker expression, rather than pMHC binding. Output of these pooled functional genetic screens
326 was well-correlated with arrayed pilot screening of the same TCRs using distinct functional
327 readouts. In addition to activation marker expression as used here, productive signaling through
328 the T cell receptor may also be analyzed by e.g., antigen-induced proliferation or degranulation,
329 or using distinct cell systems. We note though that the threshold for productive TCR signaling
330 can depend on e.g., costimulatory signals, and it will thus be important to calibrate such
331 alternative systems.
332
333 For both the datasets evaluated here, the fraction of non-validating TCRs was remarkably high.
334 Notably, the fraction of TCR-pMHC pairs that could be experimentally validated varied
335 substantially between studies, and, in case of SARS-CoV-2, was positively correlated with the
336 fraction of samples from donors with recent or acute infections. We conclude that MHC
337 multimer-based isolation of T cells from healthy donors frequently results in the isolation of TCRs
338 that fail to signal upon antigen encounter, and this may either reflect random assay noise or may
339 point towards the occurrence of TCRs that bind the relevant pMHC but do not productively signal,
340 an issue that remains to be addressed.

For all three prediction models evaluated we identified an – in some cases substantial – capacity to distinguish those TCR-pMHC pairs that do or do not pass experimental validation. However, only for AlphaFold3 substantial performance was observed in the more demanding situation in which valid and non-valid TCRs show high sequence similarity (Fig. 4). Interestingly, recent evaluations of AlphaFold2 PAE confidence-based approaches observed an only modest performance in TCR-pMHC prediction for well-studied epitopes in VDJdb[32,33]. Based on the data presented here, we speculate that the seemingly modest performance of these models may in fact reflect their capacity to also partly distinguish validating and non-validating TCR-pMHC complexes. The PDB version that was used to train AlphaFold3 included structures of two TCR-YLQ-HLA-A2 complexes (7N1F and 7N6E) and one TCR-GLC-HLA-A2 complex (3O4L) (training cutoff date: 30 September 2021[14]). Prediction confidence of TCRbridge for these three structures did not exceed average prediction confidence (Figure S4F), arguing against data leakage as a driver of the observed performance. Notably, TCRbridge not only performed well in distinguishing validating and non-validating TCRs, but also in assigning epitope reactivity when analyzing two sets of TCRs that share the same HLA-A*02:01 restriction element. The TCRbridge reactivity score that we describe here combines the three AlphaFold3 confidence measures at the TCR-pMHC interface and appears to yield slightly better performance as compared to the use of single confidence measures (Figure S4B). With the upcoming availability of validated TCR reactivity datasets for a larger number of epitopes, it will be attractive to utilize supervised machine learning models trained on the full range of features extractable from AlphaFold3 to identify those feature combinations that are most predictive of TCR-pMHC reactivity across epitopes. By the same token, it will also be attractive to use these data to investigate alternative biophysical scoring approaches, for instance extending TCRbridge with force field interaction energies[36], as force fields were recently shown to outperform supervised models trained on AlphaFold2-derived peptide, MHC, and TCR structural features in predicting the recognition of unseen epitopes[33]. Next to these approaches to optimize extraction of the relevant information from structure predictions that are obtained using the generic version of structure-based algorithms, it will be valuable to use large validated TCR reactivity datasets to directly fine-tune AlphaFold3 model parameters for TCR specificity prediction, as was recently demonstrated for the prediction of peptide-MHC binding[37].

The scalable 'sequence-synthesis-screen' approach presented here may be utilized for a number of purposes. First, in line with the strategy chosen here, the approach may be used to assess the

375 function of TCR-pMHC pairs that have been described by researchers in different fields over the
376 past decades. Second, the synthetic systems described here should form a valuable quality
377 control step for ongoing large-scale data collections efforts that e.g., aim to describe neoantigen-
378 specific T cell reactivity across cancer patient cohorts, or that aim to describe the TCR repertoire
379 for model antigens to saturation. Finally, the synthetic nature of the technology makes it well-
380 suited to explore reactivity beyond the TCRs that are observed in natural samples. For instance,
381 the predictive capacity of TCR-pMHC specificity models may be evaluated by the design of TCRs
382 with pre-specified criteria, and experimental testing of thousands of predicted TCRs may then be
383 utilized to understand features of predictive success and failure. Output from all three efforts
384 should contribute to the validated dataset initiated here. To support its use in model
385 development, the TCRvdb dataset will be hosted on https://github.com/schumacherlab/TCRvdb
386 as a resource to the community.
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408

Resource Availability

Lead Contact

Requests for further information and resources and reagents should be directed to W. Scheper (w.scheper@nki.nl) and T. Schumacher (t.schumacher@nki.nl).

Materials Availability

The VDJdb-10 TCR plasmid library is made available for academic research upon request.

Data and Code Availability

All ONT sequencing data will be deposited in the National Center for Biotechnology Information's Sequence Read Archive. The TCRtoolbox python package is available at: https://github.com/schumacherlab/TCRtoolbox. The ONT-TCRconsensus package is available at: https://github.com/schumacherlab/ONT-TCRconsensus. The validated VDJdb database is available at: https://github.com/schumacherlab/TCRvdb. The TCRbridge python package is available at: https://github.com/schumacherlab/TCRbridge. Code use to generate figures in this paper is available at: https://github.com/schumacherlab/schumacherlab_publications_code.

Acknowledgments

This work was supported by the Dutch Cancer Society Young Investigator Grant (no. 2020-1/12977) (to W.S.); ZonMw Translational Research Program 2 (no. 446002001) (to W.S.) and by the Louis Jeantet prize and Stevin Award (to T.N.S.). This work was also created as part of the MATCHMAKERS team, supported by the Cancer Grand Challenges partnership financed by CRUK (CGCATF-2023/100003) and the National Cancer Institute (OT2CA297206). RLB was supported by Oncode Accelerator (grant number NGFOP2201) and by the ScreeninC infrastructure (KWF 12539). Research at the Netherlands Cancer Institute is supported by institutional grants from the Dutch Cancer Society and the Dutch Ministry of Health, Welfare and Sport. We thank the Research High Performance Computing facility of the Netherlands Cancer Institute for providing and maintaining computation resources. We thank the flow cytometry core at the Netherlands Cancer Institute for technical support. We would like Benoit Nicolet and other members of the Schumacher group for scientific input.

443 **Author contributions**

444 T.N.S, W.S., and M.Me. conceived and designed the study. M.Me., W.S., J.U., B.P.Y.K., T.N.S.,
445 J.G., B.M., Ž.M., S.d.P., M.S., and R.L.B. developed the TCR assembly method. Ž.M., S.d.P., R.V.,
446 A.G., M.Me., Y.W., M.Mu., designed and performed T cell activation and functional genetic
447 screening experiments. M.Me., J.G., Ž.M., J.U., and B.P.Y.K. developed the full-length TCR ONT-
448 sequencing method. B.P.Y.K., M.Me., D.Á.S., Ž.M., M.I., A.P., J.T., E.M., performed data analysis
449 and modelling. M.Me., T.N.S. and W.S., wrote the manuscript. T.N.S. and W.S. supervised the
450 project. All authors reviewed and revised the manuscript.

451

452

453 **Declaration of interests**

454 W.S. is an advisor to BD Biosciences and Lumicks. T.N.S. is an advisor to Allogene Therapeutics,
455 Merus, Neogene Therapeutics, JNJ, and Scenic Biotech and is stockholder in Allogene
456 Therapeutics, Cell Control, Celsius, Merus, Polar Therapeutics and Scenic Biotech. T.N.S. is also
457 a venture partner at Third Rock Ventures, all outside the submitted work.

458

459

460 **STAR Methods**

461 **Experimental model and subject details**

462 ***Cell lines***

463 FLY-RD18 packaging cells (Sigma) were used for production of retroviral supernatants. TCRα$^-$/β$^-$
464 CD8αβ$^+$ Jurkat cells were described previously[38]. Primary CD8$^+$ T cells were isolated from
465 peripheral blood mononuclear cells (PBMCs) of healthy donors (Sanquin, the Netherlands).
466 Immortalized B cell lines were generated by retroviral introduction of Bcl-6 and Bcl-xL into B cells
467 isolated from donor PBMCs, as previously described[26].

468

469 **Method details**

470 ***Antibodies***

471 The following antibodies were used for flow cytometry: anti-CD4-FITC (clone RPA-T4, BD
472 Biosciences, dilution 1:200); anti-CD19-FITC (clone 4G7, BD Biosciences, dilution 1:200); anti-
473 CD8α-APC (clone SK1, BioLegend, dilution 1:50); anti-CD69-BV421 (clone FN50, BioLegend,
474 dilution 1:50); anti-CD137-BV421 (clone 4B4, BioLegend, dilution 1:50). PE-conjugated anti-
475 mouse TCRβ constant domain (clone H57-597, BD Biosciences, dilution 1:150). Live/Dead
476 Fixable Near-IR Dead Cell Stain (Thermo Fisher Scientific), or DAPI was used to identify live cells.

477 Data from flow cytometry experiments were acquired using FACSDiva software (version 8.0.2,
478 BD Biosciences) and analyzed using FlowJo (version 10.7.1, BD Biosciences).
479
480 ***Robotics-assisted Golden Gate cloning of synthetic TCR libraries***
481 Individual synthetic TCRs were assembled and cloned into the pMX retroviral expression plasmid
482 in separate wells of 384-well plates. Briefly, Golden Gate cloning was used to join the CDR3α-Jα,
483 CDR3β-Jβ, TRAV, TRBV, and murine TRBC-P2A (muTRBC-P2A) fragments into a pMX retroviral
484 vector that also encodes murine TRAC and the puromycin N-acetyltransferase resistance gene
485 (Figure S1A). To enable PCR amplification of desired CDR3α/β-Jα/β fragments from complex
486 oligonucleotide pools in individual wells of 384-well plates, 384 unique and highly orthogonal
487 forward and reverse primer combinations ('well-specific' primers) were used (Supplementary
488 Table 2). Separate unique primer combinations ('plate-specific' primers) were utilized to enable
489 PCR amplification of 384-well-plate-specific sub-pools of CDR3α/β-Jα/β fragments from larger
490 oligonucleotide pools (Supplementary Table 3). Individual CDR3α/β-Jα/β fragments were
491 amplified by two rounds of targeted PCR: (i) amplification of 384-well-plate-specific
492 oligonucleotide sub-pools using plate-specific primers, followed by (ii) amplification of
493 individual CDR3α-Jα and CDR3β-Jβ fragments using the assigned 384-well orthogonal primer
494 combinations. As a means to reduce the differential abundance of individual CDR3α/β-Jα/β
495 fragments in synthetic oligonucleotide pools (previously observed to contribute to non-
496 uniformity of final TCR libraries[26]), PCR reactions in the second round were carried out into
497 plateau phase. The TRAV and TRBV genes required for assembly of each unique TCR were then
498 added to individual wells from TRAV and TRBV plasmid stocks (Supplementary Table 4). Finally,
499 a Golden Gate reagent mix that also contains muTRBC-P2A and pMX-muTRAC-T2A plasmids was
500 added to the wells, and all fragments were joined by Golden Gate assembly. All process steps,
501 further detailed below, were carried out using largely automated dispensers to ensure scalability
502 of synthetic TCR assembly.
503
504 *Well-specific CDR3α-Jα and CDR3β-Jβ fragment generation*
505 In order to enable PCR amplification of desired CDR3α/β-Jα/β fragments from complex
506 oligonucleotide pools in individual wells of 384-well plates with high specificity, we tested the
507 PCR efficiencies of a collection of highly orthogonal primers[27] and assigned the best performing
508 primers to 384 unique forward and reverse primer combinations (well-specific primers).
509 Additional efficient primer pairs were chosen to enable PCR amplification of plate-specific sub-
510 pools of oligonucleotides from larger pools. For each 384-well plate, a sub-pool of plate-specific

511 CDR3α/β-Jα/β fragments was PCR-amplified by combining 0.230 ng CDR3α/β-Jα/β
512 oligonucleotide pool (Twist Bioscience), 0.2 mM dNTPs (Sigma; 11969064001), 0.02 U/µL Q5
513 polymerase (NEB; M0493L), 0.25 µM of both forward and reverse plate-specific primers, and 1x
514 Q5 polymerase buffer in a total volume of 140 µL, followed by incubation in a thermocycler (30 s
515 at 98 °C, 16 cycles of 10 s at 98 °C, 20 s at 65 °C and 35 s at 72 °C, followed by 2 min at 72 °C).
516 Amplified CDR3α/β-Jα/β fragment sub-pools were purified using the NEB Monarch DNA Column
517 Kit (T1130L) and PCR product yield was measured using an Agilent Bioanalyzer 2100 (DNA 7500
518 kit, 5067-1506).
519 Using a Hamilton/I.DOT dispenser, well-specific CDR3α/β-Jα/β fragments were amplified in 384-
520 well plates from amplified CDR3α/β-Jα/β fragment sub-pools by combining 0.15 ng PCR-
521 amplified sub-pool with 0.2 mM dNTPs, 0.02 U/µL Q5 polymerase and 0.75 µM forward and
522 reverse primers, and 1x Q5 polymerase buffer in a total volume of 2.67 µL in each well. Forward
523 and reverse primer combinations were pre-added to 384-well-plates using a Hamilton Microlab
524 STAR 384 liquid handler system, and PCR amplification mix was dispensed into the 384-well-
525 plates using an I.DOT dispenser (Dispendix, BICO), followed by incubation in a thermocycler (30
526 s at 98 °C, 39 cycles of 10 s at 98 °C, 20 s at 63.5 °C and 25 s at 72 °C, followed by 2 min at 72 °C).
527 Amplified well-specific CDR3α/β-Jα/β fragments were diluted 5.3-fold with nuclease-free water
528 (Invitrogen; #4387936) using a Hamilton/I.DOT dispenser and PCR yield was routinely verified for
529 a small selection of wells using an Agilent Bioanalyzer 2100 (DNA 7500 kit). Lastly, PCR products
530 were diluted another 13.4-fold with nuclease-free water (to approximately 220 pg/µL) using a
531 Hamilton/I.DOT dispenser in order to reduce the concentrations of PCR reagents that inhibit
532 downstream Golden Gate reactions.
533
534 *Golden Gate assembly of synthetic TCRs*
535 To prevent mis-ligation of TCR fragments during Golden Gate assembly, we first designed a highly
536 orthogonal set of four-base overhangs using a comprehensive overhang ligation fidelity[28] dataset
537 (Figure S1B and S1C). Golden Gate assembly of synthetic TCRs was performed by automated
538 dispensing of 8.55 ng pMX-muTRAC-T2A retroviral plasmid, 14 ng of muTRBC-P2A plasmid, 11.88
539 ng of both TRAV and TRBV plasmids (Twist Bioscience), 840 pg of PCR-amplified well-specific
540 CDR3α/β-Jα/β fragments , 20 U/µL T4 ligase (NEB; M0202M), 0.5x BbsI (Thermo Fisher Scientific;
541 #FD1014) , and 1x T4 ligase buffer in each well of a 384-well plate, in a total volume of 4.75 µL,
542 followed by incubation of 384-well plates in thermocyclers (50 cycles of 10 min at 37 °C and 10
543 min at 16 °C, followed by 10 min at 55 °C and 10 min heat inactivation at 65 °C).

544 Following Golden Gate assembly, reaction products were pooled using only part [2 µL] of the
545 product in each well, to thereby retain a repository of individual TCRs . To purify DNA from pooled
546 Golden Gate products, DNA was precipitated with isopropanol (1:1 volume ratio), GlycoBlue
547 Coprecipitant (100 ng/µL), and 0.05 M NaCl. Precipitated DNA was washed twice with 80%
548 ethanol and eluted in 25 µL nuclease free water. Purified Golden Gate product pools were
549 transformed at >1,000x coverage into electrocompetent Endura *E.coli* (60242-2, Biosearch
550 Technologies) by electroporation (10 uF, 600 Ω, 1800 V), using a BioRad Gene Pulser Xcell.
551 Bacteria were subsequently grown overnight at 37 °C in LB medium containing kanamycin. The
552 use of distinct antibiotic resistance genes in the used plasmids (Fig. S1A) enables counter-
553 selection of remaining muTRBC-P2A, TRAV and TRBV plasmid material (Fig. S1A). 384-well plates
554 containing individual TCR archives were stored at -20 °C for use in, for instance, downstream
555 characterization of selected TCRs.
556
557 ***Preparation of long read TCR sequencing libraries***
558 *Dual-UMI-tagging and nested PCR amplification*
559 For each individual TCR in a library, we tagged 50,000 (plasmid DNA) or 100 (gDNA) template
560 molecule copies (Figure S2A). Each template molecule was tagged with terminal UMIs by PCR,
561 using primers nanop_GSP_umi_Fw and nanop_GSP_umi_Rev (Supplementary Table 5). Both
562 primers contain a universal primer site for downstream amplification and a patterned UMI
563 consisting of 18 V nucleotides (all nucleotides except T) and 14 patterned 'anchor' T nucleotides,
564 yielding a combined complexity of 1.5×10^{17} UMI sequences.
565 To estimate the average rate of UMI collisions per template molecule (i.e., the average fraction
566 of distinct molecules that are indistinguishable due to being tagged with the same UMIs), we
567 used an alternative formulation of the birthday problem[39]. For this calculation, the effective
568 number of UMIs was estimated using the 3-error Hamming bound, as 3-errors conservatively
569 corresponds to clustering UMIs with VSEARCH[40] at 93% identity. Based on this model, the
570 estimated maximal average UMI collision rate per template molecule is 2.3×10^{-10} for plasmid (19
571 individual TCRs in the largest VDJdb-10 TCR similarity bin × 50,000 copies) and 4.6×10^{-13} gDNA
572 (19 individual TCRs × 100 copies) templates.
573 Dual-UMI-tagging PCR reactions were performed by combining the calculated amount of
574 template DNA (50,000 plasmid copies corresponding to 0.178 pg or 100 gDNA copies
575 corresponding to 700 pg × the number of unique TCRs in the library) with 0.5 µM forward and
576 reverse primers, 1x NEBNext HiFi 2x PCR mastermix (NEB; M0541L) in a total volume of 15 µL
577 (plasmid) or 100 µL per 67 ng input gDNA, followed by incubation in a thermocycler (3 min at 98

578 °C, 2 cycles of 30 s, 90 s at 72°C and 90 s at 72°C, followed by 5 min at 72°C). After dual-UMI
579 tagging, tagging primers were removed using 0.6x AMPure XP beads (Beckman Coulter), and
580 reaction products were washed twice with 80% ethanol, and eluted in 20 µL (plasmid), or 20 µL
581 per 200 µL required dual-UMI-tagging reaction volume (gDNA), nuclease-free water.
582 Dual-UMI-tagged full-length TCR molecules were further amplified by two rounds of nested PCR
583 using the primers in Supplementary Table 5. In the first PCR, complete dual-UMI-tagged reaction
584 products were PCR-amplified using 0.5 µM nanop_UVP_umi_1_Fw and nanop_UVP_umi_1_Rev
585 primers and 1x NEBNext HiFi 2x PCR mastermix in a total volume of 96 µL. The reaction mixture
586 was incubated in a thermocycler (1 min at 98 °C, 5 cycles of 15 s at 98 °C, 25 s at 67 °C and 90 s
587 at 72 °C, immediately followed by 5 cycles of 15 s at 98 °C, 25 s at 72 °C and 90 s at 72 °C, and a
588 final 2 min at 72 °C). PCR products were purified using 0.6x AMPure XP beads, washed twice with
589 80% ethanol, eluted in 20 µL nuclease-free water. Amplification products were then further PCR-
590 amplified by combining complete round one reaction products with 0.5 µM nanop_UVP_2_Fw
591 and nanop_UVP_2_Rev primers and 1x NEBNext HiFi 2x PCR mastermix in a total volume of 96
592 µL, followed by incubation in a thermocycler (1 min at 98 °C, followed by a set number of cycles
593 of 15 s at 98 °C, 25 s at 70 °C and 90 s at 72 °C, and 2 min at 72 °C). Here, the number of PCR
594 cycles was based on the amount of initial plasmid or gDNA template that was required to tag all
595 TCR template molecules with UMIs at sufficient coverage. PCR products were purified using 0.6x
596 AMPure XP beads, washed twice with 80% ethanol, eluted in 13 µL nuclease-free water, and
597 product concentrations were measured using an Agilent Bioanalyzer 2100 (DNA 7500 kit).
598
599 *ONT long read TCR sequencing*
600 Full-length TCR amplicons were prepared for ONT sequencing using the Native Barcoding Kit 24
601 V14 protocol (Oxford Nanopore Technologies, SQK-NBD114.24). In brief, 200 fmol of amplicon
602 DNA was end-prepped and dA-tailed using NEBNext Ultra II End Repair/dA-Tailing Module (NEB;
603 E7546), followed by ligation of barcodes from the Native Barcoding Expansion kit using Blunt/TA
604 Ligase (NEB; M0367). Barcoded samples were pooled and subjected to native adapter ligation
605 with the NEBNext Quick Ligation Module (NEB; E6056). Resulting libraries were quantified using
606 a Qubit dsDNA HS Assay (Thermo Fisher) and sequenced at 2,000-2,500x coverage, using a
607 R10.4.1 flow cell on MinION or PromethION sequencing devices.
608
609 ***Expression of individual TCRs and TCR libraries in reporter T cells***
610 For expression in Jurkat T cells, TCRs were retrovirally transduced into TCRα$^-$/TCRβ$^-$ CD8αβ$^+$
611 Jurkat T cells. In brief, retroviral supernatant was produced by transfection of FLY-RD18

612 packaging cells with pMX plasmid DNA encoding indicated TCR libraries or individual TCRs using
613 Lipofectamine 3000 Reagent (Thermo Fisher Scientific). After 48 h, supernatants were collected,
614 filtered and centrifuged (2,000 g for 90 min) in 24-well plates precoated with RetroNectin
615 (Takara). $CD8^+$ $TCR\alpha^-/TCR\beta^-$ Jurkat cells were infected by incubation on virus-coated plates for
616 21 hours. For the production of TCR libraries, Jurkat cells were transduced at infection rates of
617 less than 5% to ensure single retroviral integrations in the majority of transduced cells. After 72
618 h, transduction efficiency was assessed by cell staining with anti-mouse TCRβ constant domain
619 antibody and flow cytometry. TCR-expressing Jurkat cells were selected by culture in the
620 presence of 2.5 µg/ml puromycin and then 5.0 µg/ml puromycin for 48 h each. Subsequently,
621 TCR-transduced cells were isolated by staining with anti-mouse TCRβ constant domain antibody
622 and cell sorting on a FACSAria Fusion (BD Biosciences) to achieve at least 80% purity. Sorted
623 Jurkat cells were cultured in RPMI 1640 medium supplemented with 10% FBS (Sigma-Aldrich)
624 until use in TCR discovery screens. TCR library coverage of at least 250x was maintained during
625 all steps.
626 For TCR expression in primary T cells, $CD8^+$ T cells were isolated from healthy donor PBMCs using
627 the $CD8^+$ T Cell Isolation Kit (Miltenyi Biotec) and stimulated with anti-CD3/anti-CD28 Dynabeads
628 (Life Technologies) in RPMI 1640 medium supplemented with 10% human AB serum (Life
629 Technologies), penicillin–streptomycin (Roche) and 150 U mL^{-1} IL-2 (Proleukin, Novartis) for 48
630 h. 24-well plates precoated with RetroNectin (Takara) were coated with retrovirus by
631 centrifugation (2,000 g for 90 min) of retroviral supernatants, followed by incubation of
632 stimulated $CD8^+$ PBMCs on these retrovirus-coated plates for 21 h. Transduction efficiency was
633 measured after 72 h by cell staining with anti-mouse TCRβ constant domain antibody and flow
634 cytometry. Transduced cells were selected by culture in the presence of 2.5 µg/ml puromycin for
635 48 h, with subsequent addition of new IL-2 and medium every 3–4 days.
636
637 ***Generation of antigen-expressing B cell lines***
638 The HLA-A*02:01-restricted `GLCTLVAML` (GLC) EBV-derived and `YLQPRTFLL` (YLQ) SARS-Cov2-
639 derived epitopes were introduced into an HLA-A*02:01$^+$ Bcl-6/Bcl-xL-immortalized B cell line by
640 lentiviral transduction. In brief, lentiviral supernatant was produced by transfection of 293T
641 LentiX HEK cells (Takara) with a lentiviral plasmid encoding mKelly1 and either the GLC or YLQ
642 epitope, and psPAX2 and pGALV[41] helper plasmids, using Lipofectamine™ 3000 Reagent (Thermo
643 Fisher Scientific). After 48 h, supernatants were collected, filtered and centrifuged (2,000 g for 90
644 min) in 24-well plates precoated with RetroNectin (Takara). Immortalized B cells were infected
645 by incubation on virus-coated plates for 21h. Transduction efficiency was measured after 72 h by

646 analysis of mKelly1 expression by flow cytometry. Transduced B cells were subsequently
647 selected with 2.5 – 5.0 µg/mL puromycin for 48 h and cultured in the presence of irradiated (55
648 Gy) CD40L^{+} mouse L cells in IMDM medium (Gibco) supplemented with 10% heat-inactivated
649 FBS (Sigma-Aldrich), penicillin-streptomycin and 50ng/mL IL-21 (BioLegend)) until further use. B
650 cells received new irradiated CD40L^{+} L cells every 10 days, and new medium and IL-21 every 3-4
651 days.
652
653 ***T cell activation assays***
654 Reactivity of T cells transduced with single TCRs was analyzed by flow cytometry after overnight
655 coculture with target cells in U-bottom 96-well plates at an effector-to-target ratio of 1:1. T cells
656 incubated in the absence of target cells with or without phorbol 12-myristate 13-acetate
657 (50 ng/mL; Sigma-Aldrich) and ionomycin (1 µg/mL; Sigma-Aldrich), served as positive and
658 negative controls, respectively. Cells were stained with IR-Dye and anti-mouse TCRβ constant
659 domain antibody, in combination with anti-CD8 and anti-CD137 (when using primary T cells) or
660 with anti-CD69 (when using Jurkat cells) antibodies, and subsequently analyzed by flow
661 cytometry. Coculture with irrelevant epitope-expressing B cells was used to determine
662 background T cell activation levels.
663
664 ***Functional TCR reactivity screens***
665 TCR discovery screens were performed as described previously for TIL-derived TCRs[26]. In brief,
666 functional genetic screens were performed by incubating VDJdb-10 TCR library-expressing CD8^{+}
667 Jurkat T cells with B cells expressing the indicated epitopes at an effector-to-target ratio of 1:1.
668 Subsequently, Jurkat cells with the highest 10% of CD69 expression were isolated using a
669 FACSAria Fusion cell sorter and TCR abundance was determined by nanopore sequencing, as
670 described above. Functional genetic screens were performed in triplicate, with a TCR library
671 coverage of 500x in all experiments. Reactive TCRs and non-reactive TCRs were identified from
672 nanopore sequencing data as described below.
673
674
675 **Data analysis and Statistical Analysis**
676 ***Creation of the VDJdb-10 dataset***
677 The VDJdb database was downloaded from the antigengenomics/vdjdb-db GitHub page on the
678 19th of May 2023 and was built directly from submissions using provided instructions. The
679 vdjdb_full.txt file was filtered to only contain human CD8^{+} T cell-derived TCRs and MHC-I binding

epitopes. In addition, the database was filtered to only contain unique and complete TCR-epitope annotations (filtering on TRAV, TRBV, TRAJ and TRBJ genes, CDR3α and CDR3β amino acid sequences, and epitope amino acid sequence). All entries derived from 10X Genomics data from the Application Note - A New Way of Exploring Immunity (v1 Chemistry) were removed (https://www.10xgenomics.com/resources/datasets), as prior work provided evidence for the low quality of this dataset[42]. This resulted in 6,450 paired TCR-pMHC entries. Fifty duplicate TCRs were removed because the annotated epitopes were at >2 edit distance of each other. One TCR entry was removed because of a missing HLA-gene annotation. 28 epitope entries were removed because their HLA-gene annotation did not match the HLA annotated in an experimentally verified dataset of peptide HLA allele binders[43]. 64 entries were removed because they contained an HLA annotation that does not occur in the IMGT HLA allele database[44] (i.e., "A*08:01", "A*02:256", and "B*51:193"). The final parsed VDJdb dataset, in which TCR-pMHC pairs that would be most likely to contain errors have been removed, contained 6,307 entries.

To create the VDJdb-10 TCR library from the final parsed VDJdb dataset, we extracted all 3,251 TCRs annotated as reactive to the top 10 most densely sampled epitopes in VDJdb: `GILGFVFTL`, `YLQPRTFLL`, `TTDPSFLGRY`, `NLVPMVATV`, `SPRWYFYYL`, `CINGVCWTV`, `LLWNGPMAV`, `GLCTLVAML`, `LTDEMIAQY`, and `ATDALMTGF`. Additionally, we included 549 negative control TCRs annotated as reactive to epitopes that have only one or two TCR entries in VDJdb, collectively spanning 466 unique epitopes.

Design of TCR library assemblies

We developed a Python package, called `TCRtoolbox`, to design and prepare TCR library assemblies, requiring only the TRAV, TRAJ, TRBV and TRBJ gene names and CDR3α and CDR3β amino acid sequences as input. As a first step, the package translates TRAV, TRAJ, TRBV, and TRBJ gene names into standardized IMGT format. `TCRtoolbox` subsequently trims the TRAJ and TRBJ gene amino acid sequences from their N-terminus to the first amino acid of the J gene end motif (either `FGXG`, `WGXG`, `XGXG`, or `FXXG`)[45]. The trimmed J genes are then joined to their corresponding CDR3α/β regions to generate CDR3α/β-Jα/β amino acid sequences. Resulting sequences are codon optimized using a custom codon optimizer within `TCRtoolbox` to eliminate type IIS BbsI recognition sites, four-base overhangs (Figure S1C), Kozak sequences, primer binding sites, repeats longer than four nucleotides, and terminator motifs (`TTTTT` or `AAAAA`). Codon optimization enforces a global GC content between 35–60%, with a relaxed range of 25–70% allowed only for sequences that cannot otherwise be optimized. Sequences failing these

713 criteria are filtered out. All VDJdb-10 TCR CDR3α/β-Jα/β sequences were successfully codon
714 optimized.
715 TCRtoolbox subsequently assigns individual TCRs to wells of 384-well plates, together with
716 unique primer pairs that were selected from a set of highly orthogonal primers. For each TCR that
717 is assigned to a defined well of a 384-well plate, the corresponding primer nucleotide sequences
718 are added to the 5' and 3' ends of the codon-optimized CDR3α/β-Jα/β nucleotide sequences by
719 TCRtoolbox. Resulting nucleotide sequences are written to an Excel worksheet that may directly
720 be used to obtain oligonucleotide pools from a commercial supplier. In parallel, TCRtoolbox
721 assigns TRAV and TRBV genes from each TCR to each well. These genes are stored as reusable
722 plasmid stocks in 2D-barcoded tubes, currently covering the *01 alleles of all TRAV and TRBV
723 genes (Supplementary Table 4). For seamless joining with CDR3 sequences, TRAV and TRBV
724 amino acid sequences are trimmed from their C-termini up to the cysteine residue at position
725 104 (Cys104, IMGT numbering). TCRtoolbox automatically keeps track of TRAV and TRBV stock
726 volumes. TCRtoolbox also writes a Hamilton automatic pipetting instruction Excel sheet to
727 generate an I.DOT dispenser source plate from the 2D-barcoded tubes, and I.DOT dispense
728 instruction CSV file is written to add the required TRAV and TRBV genes from the resulting I.DOT
729 source plate to the corresponding 384-wells for TCR assembly.
730 To check the constructed CDR3-J fragment sequences for design errors and TRAV and TRBV 384-
731 well assignment before start of physical TCR assembly, we simulate cloning of each TCR *in silico*
732 using pydna[46], modeling both restriction digestion, ligation, and PCR to yield the fully assembled
733 TCR-encoding retroviral plasmid sequences. The resulting TCR amino acid sequence is then
734 compared to the independently reconstructed amino acid sequence of the same TCR using
735 custom TCR reconstruction software implemented in TCRtoolbox, as previously described[19].
736
737 ***TCR UMI consensus creation and counting pipeline***
738 ONT sequencing reads were base-called with dorado using the sup@v5.0.0 model with a --
739 min-qscore of 10: https://github.com/nanoporetech/dorado. To create high-accuracy full-
740 length TCR UMI consensus reads that can be counted, we developed a Python package called
741 ONT-TCRconsensus, with steps from[29,45]. All exact parameter settings and program versions
742 used by this package are provided on its GitHub page. In this package, primers are first trimmed
743 from raw base-called ONT reads using dorado, and trimmed reads with an expected error rate [46]
744 (EE rate) >7%, as predicted from base-call qualities are removed using VSEARCH[40]. As it is not
745 possible to unambiguously assign filtered raw base-called reads to reference TCRs that have
746 highly similar sequences, we developed a pipeline that first assigns raw base-called reads to TCR

747 similarity bins that contain reference TCRs with highly related sequences. This allows the
748 subsequent creation of UMI-based consensus sequences with sufficient accuracy to
749 unambiguously assign the consensus reads to specific TCR reference sequences. In this
750 approach, the TCR library reference FASTA file is self-aligned using minimap2[47] and TCRs with
751 >93% sequence identity (as raw base-called reads maximally have a 7% EE rate) are clustered
752 into 'TCR similarity bins'.
753 Next, raw base-called reads are aligned to the TCR library reference with `minimap2`. Alignments
754 are assigned to their respective TCR similarity bins based on the reference TCR that each raw
755 base-called read aligns to. Within each TCR similarity bin, high-quality UMIs are detected on both
756 ends based on their designed sequence pattern (see above) and correct UMI lengths. Subreads
757 originating from the same unique TCR molecule are then identified by clustering the detected
758 UMIs into UMI bins with `VSEARCH`. PCR chimeras, which create low-frequency novel UMI
759 combinations by recombination, are removed by filtering out UMI bins with a subread coverage
760 <4. Of note, `VSEARCH` UMI clustering within TCR identity clusters minimizes computational
761 resource requirements. To validate TCR similarity binning, we confirmed that UMIs did not
762 overlap between TCR similarity bins. The consensus sequence for each UMI bin is then generated
763 from binned subreads using `medaka` polishing: https://github.com/nanoporetech/medaka. The
764 resulting high-accuracy consensus sequences are aligned to the TCR reference library using
765 `minimap2`. Alignments with a BLAST sequence identity (exact nucleotide matches at the same
766 position, counting each consecutive gap separately) below the minimum required to distinguish
767 all closest-identity TCR pairs in the reference are filtered, ensuring that only consensus
768 sequences accurate enough to distinguish all TCRs in the library are counted. To validate the
769 entire `ONT–TCRconsensus` pipeline, we confirmed that UMIs do not overlap between alignments
770 to unique reference TCRs.
771 To calculate precision as % sequence-perfect reads, we filtered alignments below 1.0 BLAST
772 identity (i.e., exact matching sequences). Finally, TCR UMI counts were generated by counting
773 the number of unique UMIs of which the consensus sequences aligned to each reference TCR.
774
775 ***Reactive TCR identification using DESeq2***
776 To identify reactive TCRs in pooled TCR screens, per-sample count tables were normalized and
777 compared with `PyDESeq2`[48,49] (version 0.5.0). Within the set of 3,428 VDJdb-10 TCRs (92.8% of all
778 VDJdb-10 TCRs) that was detected across all YLQ and GLC triplicates, reactive and non-reactive
779 TCRs were defined using the DESeq2 Wald test with an adjusted P < 1e-5. All default parameters
780 were used, except `refit_cooks`, which was set to False.

tcrdist3 network analysis

Pairwise tcrdist3 distances were computed between all query TCRs, resulting in an n × n distance matrix, where n is the number of TCRs compared. Edges between TCR nodes were defined for any TCR pair with a TCRdist3 distance less than 120. This threshold was consistently used to generate all TCR networks. Force-directed layouts of the resulting networks were visualized using the `spring_layout` function from `NetworkX` (v2.7), with the layout parameter `k=0.15`.

Visualization of Alphafold3 predicted TCR-pMHC complex structures

AlphaFold3-predicted TCR–pMHC structures were visualized using ChimeraX, with residues colored according to AlphaFold3's pLDDT confidence scores. Circus plots were generated using AlphaBridge. Protein interfaces, as defined by AlphaBridge, were colored green for TCRα-peptide interfaces, purple for TCRβ-peptide interfaces and grey for all other interfaces. The outer ring reflects the per residue model confidence (pLDDT) of the respective sequences.

TCR-pMHC reactivity modeling

tcrdist3

tcrdist3 distances were computed between each query TCR and all YLQ- or GLC-annotated TCRs, resulting in a 1 × (n−1) vector of distances. The nearest neighbors for each query TCR were identified by selecting the (n−1)/10 smallest distances (rounded to the nearest integer). A weighted average of these nearest neighbor distances was then computed, where smaller distances received higher weights. The final prediction score was the negative of this weighted average.

STAPLER

STAPLER was trained on the parsed VDJdb dataset of 6,307 TCR-pMHC entries. Training was performed using the training procedure outlined in[19]. Query TCRs were then predicted using the best model checkpoint from each fold, and the mean of the predicted scores across folds was used as the final prediction score for each query TCR.

pLDDT

AlphaFold3's pLDDT confidence scores for predicted TCR-pMHC complex structures were parsed from the CIF files outputted by the AlphaFold3 inference implementation. Final pLDDT scores were computed as the mean pLDDT of the CDR3α, CDR3β and epitope that was extracted

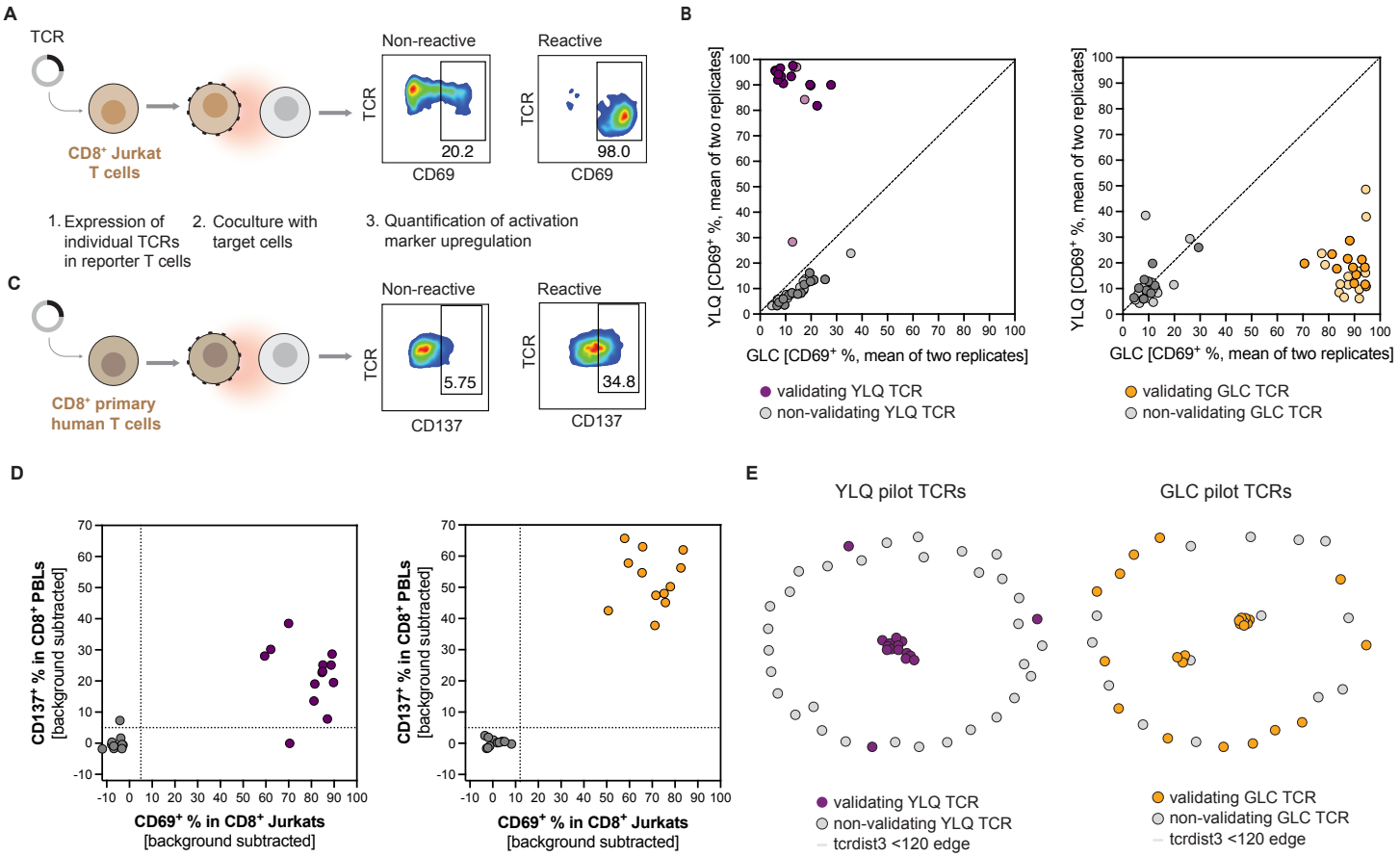
using the PDB parser of BioPython[50] by taking b-factor score (pLDDT) of the Cα atom of each amino acid.

TCRbridge

TCR-pMHC complex structures were predicted using an implementation of the inference pipeline of AlphaFold3 (version 3.0.0): https://github.com/google-deepmind/alphafold3. AlphaBridge interface scores (ABi) were calculated for TCRα-peptide and TCRβ-peptide interfaces at an interface threshold of 0.4. If no interface was identified at this interface threshold, the TCRbridge prediction score was set to 0. The final TCRbridge prediction score was computed as the mean of the two TCRα-peptide and TCRβ-peptide ABi scores. Next to ABi scores, AlphaBridge also provides PDE, PMC and iptm_i scores for each interface.

Model performance evaluation using PRC and ROC-AUC

The scikit-learn[51] implementation of `precision_recall_curve` was used to calculate the precision and recall for every threshold using prediction scores and labels. The `average_precision_score` was used to calculate average precision (AP), defined as the area under the precision-recall curve. Random average precision is defined as the number of positive samples divided by the total number of samples. The scikit-learn `roc_curve` implementation was used to calculate true positive and false positive rates for every threshold using prediction scores and labels. The `roc_auc_score` was used to calculate the area under the receiver operating characteristic curve (AUC). The random line is defined as 0.5 for all AUC curves. Labels can either be validating or non-validating TCRs (Figure 4D), or reactive or non-reactive TCR-pMHC pairs (Figure 4E, 4F, and S4G).


A
TCR
CD8+ Jurkat T cells
Non-reactive
Reactive
TCR
20.2
98.0
CD69
1. Expression of individual TCRs in reporter T cells
2. Coculture with target cells
3. Quantification of activation marker upregulation
B
YLQ [CD69+ %, mean of two replicates]
GLC [CD69+ %, mean of two replicates]
validating YLQ TCR
non-validating YLQ TCR
validating GLC TCR
non-validating GLC TCR
C
CD8+ primary human T cells
Non-reactive
Reactive
TCR
5.75
34.8
CD137
D
CD137+ % in CD8+ PBLs [background subtracted]
CD69+ % in CD8+ Jurkats [background subtracted]
E
YLQ pilot TCRs
GLC pilot TCRs
validating YLQ TCR
non-validating YLQ TCR
tcrdist3 <120 edge
validating GLC TCR
non-validating GLC TCR
tcrdist3 <120 edge

849 **Figure 1 Exploration of data quality in VDJdb**

850 (A) Assay setup and representative flow cytometry data depicting CD69 expression on TCR-
851 transduced CD8αβ$^+$ TCR-null Jurkat cells following coculture with either YLQ- or GLC-expressing
852 HLA-A*02:01$^+$ B cells.

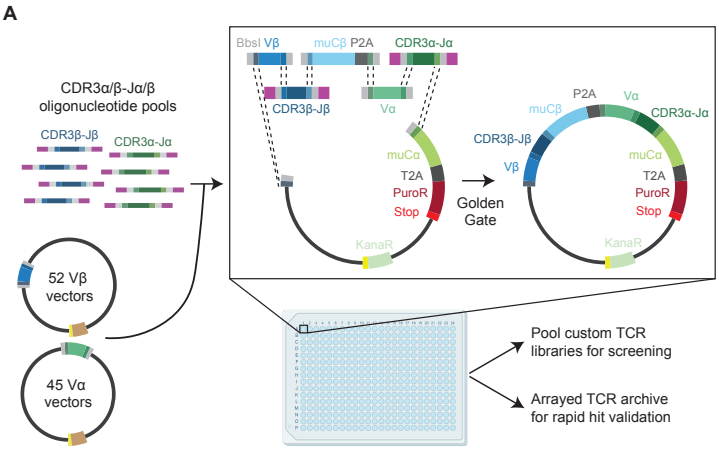
853 (B) Scatterplot depicting CD69 expression on TCR-transduced CD8$^+$ TCR-null Jurkat cells upon
854 exposure to either YLQ- or GLC-expressing HLA-A*02:01$^+$ B cells. Dots represent individual TCRs.
855 Dots with darker shading indicate TCRs that were also evaluated in primary CD8$^+$ T cells, as
856 shown in (C and D).

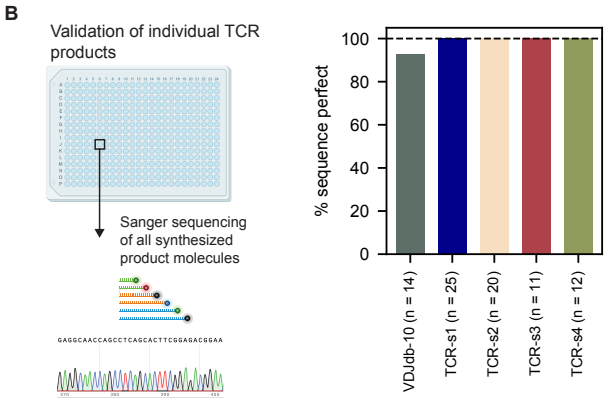
857 (C). Assay setup and representative flow cytometry data depicting CD137 expression on TCR-
858 transduced primary CD8$^+$ T cells following coculture with either YLQ- or GLC-expressing HLA-
859 A*02:01$^+$ B cells.

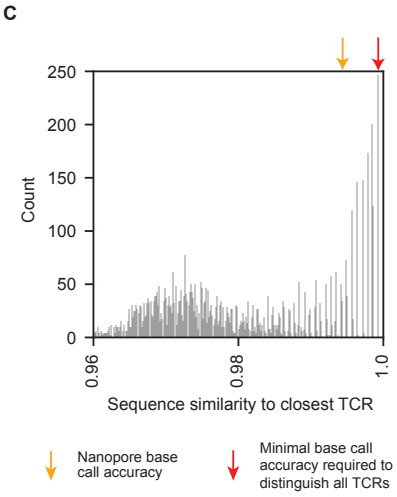
860 (D). Scatterplot comparing CD69 expression on TCR-transduced CD8$^+$ TCR-null Jurkat cells and
861 CD137 expression on TCR-transduced primary CD8$^+$ T cells, following coculture with HLA-
862 A*02:01$^+$ B cells expressing the indicated epitopes. Dots represent individual TCRs. Background
863 CD69/CD137 signal observed upon co-culture with HLA-A*02:01$^+$ B cells expressing irrelevant
864 epitopes was subtracted.

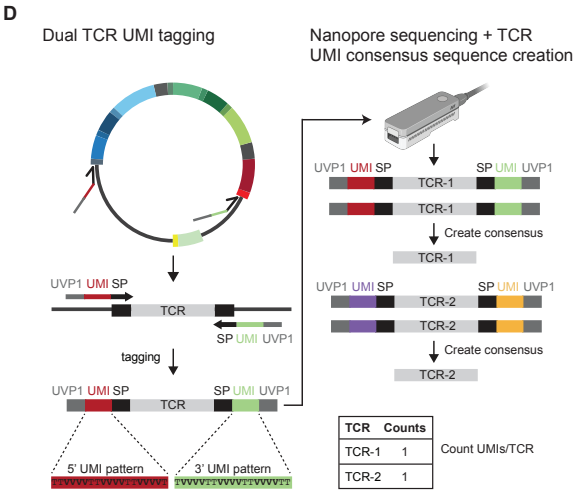
865 (E) tcrdist3 distance-based network formed by YLQ- and GLC-annotated TCRs evaluated in (A-
866 D). Grey nodes represent non-validating TCRs, colored nodes represent validating TCRs in (A,B).
867 Nodes are connected by edges when TCRs represented by those nodes have a tcrdist3 distance
868 below 120.

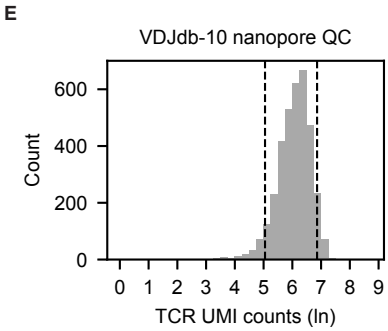
869
870
871
872
873
874
875
876
877
878
879
880
881
882











Sensitivity (% TCRs successfully assembled)	96.2% (3551/3693)
Precision (% sequence-perfect TCR UMIs)	79.0%
95th/5th percentile ratio of TCR abundance	6.12

883 **Figure 2 Scalable synthetic TCR library assembly and characterization**
884 (A) Schematic overview of synthetic TCR library assembly platform (see Figure S1A for details).
885 (B) Percentage of full-length sequence-perfect (i.e., exact matching sequences) synthetic TCRs,
886 as evaluated using Sanger-sequencing of bulk TCR assembly products. Indicated numbers of
887 assembly products of the VDJdb-10 (3,693 TCRs), TCR-s1 (3,510 TCRs), TCR-s2 (983 TCRs), TCR-
888 s3 (1,000 TCRs), and TCR-s4 (2,892 TCRs) libraries were evaluated.
889 (C) Sequence identity between each TCR in the VDJdb-10 library and its closest match, based on
890 pairwise comparisons. Identities were calculated by self-aligning all TCRs using minimap2[47],
891 defined as BLAST identity: exact nucleotide matches at the same position, counting each
892 consecutive gap separately. The orange arrow marks the average raw base-calling accuracy of
893 Oxford Nanopore Technologies (ONT), while the red arrow indicates the minimum sequence
894 accuracy threshold above which all TCRs in the VDJdb-10 library are distinguishable.
895 (D) Schematic of ONT sequence unique molecular identifier (UMI)-based error correction
896 method to sequence unique full-length TCR molecules at the accuracy required to distinguish all
897 TCRs. (see Figure S2A-B for details of ONT-sequencing DNA library preparation and error
898 correction bioinformatics). Unique synthetic TCR molecules are tagged with UMIs at both ends.
899 ONT sequencing-reads originating from the same unique TCR molecule are identified by UMI
900 binning and used to create an accurate consensus sequence by medaka polishing. Resulting
901 consensus sequences are aligned to the reference sequence list to allow counting of unique full-
902 length TCR molecules.
903 (E) Histogram of TCR UMI counts of the VDJdb-10 library and corresponding count QC summary
904 statistics.
905
906
907
908
909
910
911
912
913
914
915
916

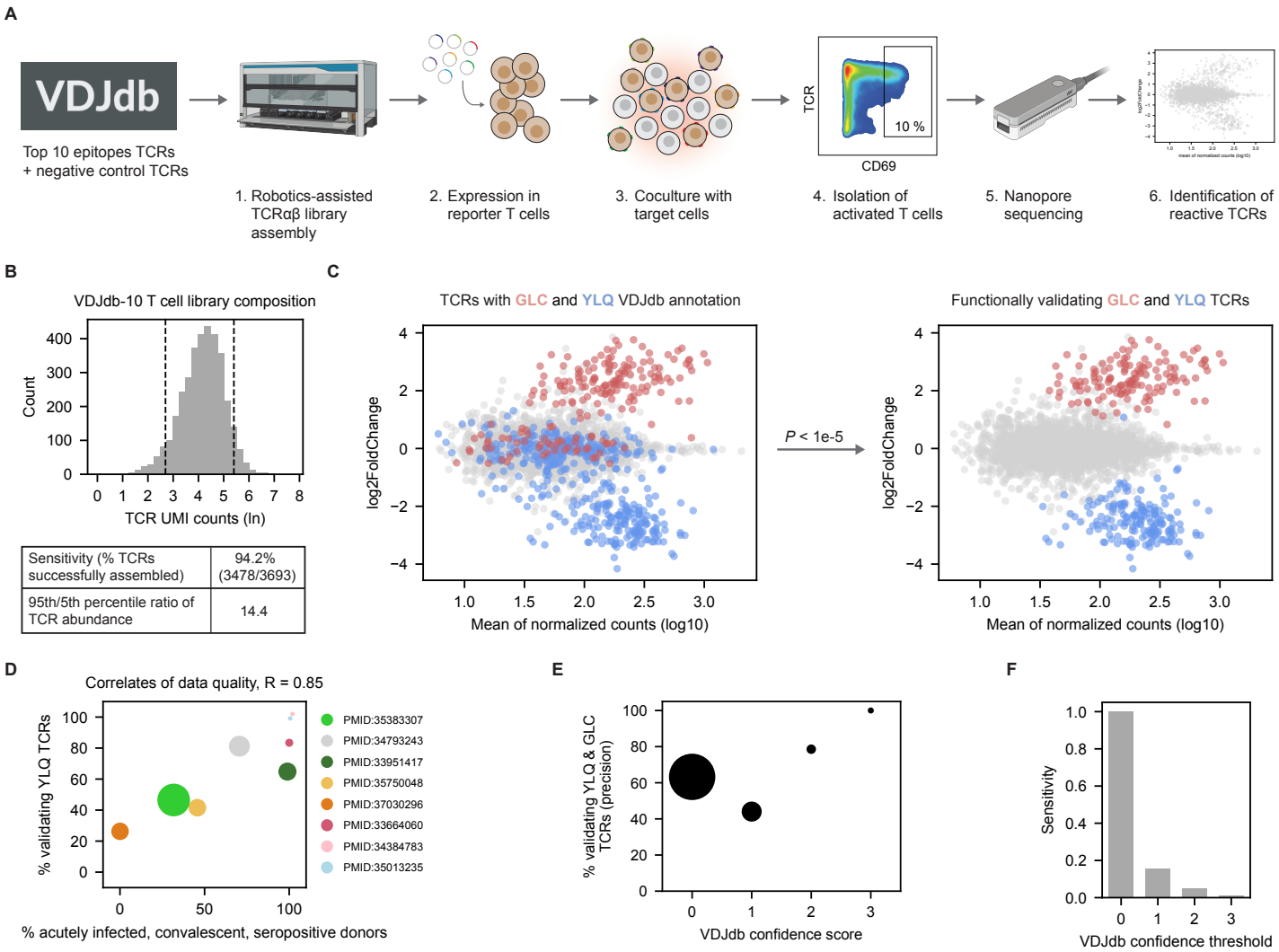

A
VDJdb
Top 10 epitopes TCRs
+ negative control TCRs
1. Robotics-assisted TCRαβ library assembly
2. Expression in reporter T cells
3. Coculture with target cells
TCR
10 %
CD69
4. Isolation of activated T cells
5. Nanopore sequencing
6. Identification of reactive TCRs
B
VDJdb-10 T cell library composition
Count
TCR UMI counts (ln)
Sensitivity (% TCRs successfully assembled) 94.2% (3478/3693)
95th/5th percentile ratio of TCR abundance 14.4
C
TCRs with GLC and YLQ VDJdb annotation
log2FoldChange
Mean of normalized counts (log10)
P < 1e-5
Functionally validating GLC and YLQ TCRs
log2FoldChange
Mean of normalized counts (log10)
D
Correlates of data quality, R = 0.85
% validating YLQ TCRs
% acutely infected, convalescent, seropositive donors
PMID:35383307
PMID:34793243
PMID:33951417
PMID:35750048
PMID:37030296
PMID:33664060
PMID:34384783
PMID:35013235
E
% validating YLQ & GLC TCRs (precision)
VDJdb confidence score
F
Sensitivity
VDJdb confidence threshold

Figure 3 Pooled functional genetic screening of the VDJdb-10 TCR library

(A) Schematic overview of pooled functional genetic screening approach. The VDJdb-10 TCR library was introduced into CD8αβ$^+$ TCR-null Jurkat cells. VDJdb-10-expressing Jurkat cells were exposed to HLA-A*02:01$^+$ B cells expressing either the YLQ or GLC epitope and reactive T cells were isolated using CD69$^+$ cell sorting. Abundance of individual VDJdb-10 TCRs in the CD69$^+$ cell fractions was determined by full-length TCR UMI ONT-sequencing. Reactive TCRs were then identified by their relative abundance in the different CD69$^+$ cell fractions.

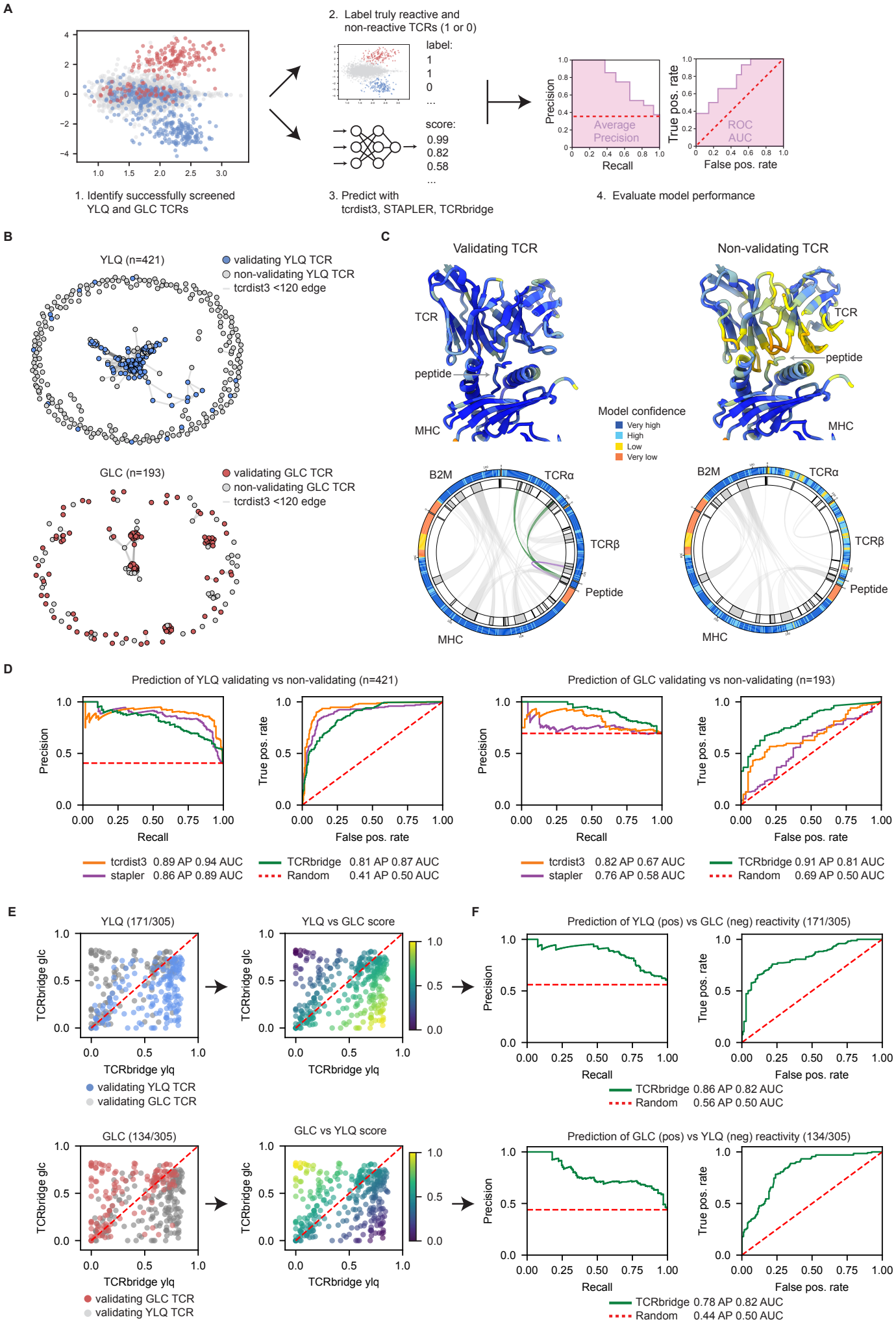
(B) Histogram of TCR UMI counts of the VDJdb-10 library in Jurkat T cells and corresponding count QC summary statistics.

(C) MA plot of fold changes, representing the relative abundance of TCR UMI counts in the YLQ epitope exposed and GLC epitope-exposed CD69$^+$ cell fractions over the mean normalized TCR UMI count across CD69$^+$ cell fractions. Dots represent individual TCRs. Left MA plot: dots are colored by VDJdb annotation for reactivity against GLC (red) or YLQ (blue). Right MA plot: significantly enriched (*P* value < 1e-5) dots are colored. *P* values were calculated using the DESeq2[48,49] Wald test and adjusted for multiple comparisons.

(D) Scatterplot depicting the percentage of validating YLQ-reactive TCRs in individual studies as a function of the percentage of samples from acutely infected, convalescent, or SARS-CoV-2 seropositive donors. In all studies, TCRs were identified based on pMHC multimer binding. In those cases where clonotype-level donor information was available (the majority of studies) the fraction of TCR clonotypes originating from acutely infected, convalescent, and seropositive donors was determined. Dot size indicates the number of TCRs contributed by each study (range 1-191 TCRs).

(E) Plot depicting the percentage of validating YLQ- and GLC-annotated TCRs (i.e., precision) at the indicated VDJdb confidence scores. Dot size indicates the total number of annotated TCRs at each confidence score threshold (range 4-512 TCRs).

(F) Barplot depicting the percentage of all validating YLQ- and GLC-annotated TCRs that is retained (i.e., sensitivity) when using the indicated thresholds of VDJdb confidence scores as a minimum quality threshold.


A
1. Identify successfully screened YLQ and GLC TCRs
2. Label truly reactive and non-reactive TCRs (1 or 0)
label:
1
1
0
...
score:
0.99
0.82
0.58
...
3. Predict with tcrdist3, STAPLER, TCRbridge
Precision
Recall
Average Precision
True pos. rate
False pos. rate
ROC AUC
4. Evaluate model performance
B
YLQ (n=421)
validating YLQ TCR
non-validating YLQ TCR
tcrdist3 <120 edge
GLC (n=193)
validating GLC TCR
non-validating GLC TCR
tcrdist3 <120 edge
C
Validating TCR
Non-validating TCR
TCR
peptide
MHC
Model confidence
Very high
High
Low
Very low
B2M
TCRα
TCRβ
Peptide
MHC
D
Prediction of YLQ validating vs non-validating (n=421)
Prediction of GLC validating vs non-validating (n=193)
Precision
Recall
True pos. rate
False pos. rate
tcrdist3 0.89 AP 0.94 AUC
stapler 0.86 AP 0.89 AUC
TCRbridge 0.81 AP 0.87 AUC
Random 0.41 AP 0.50 AUC
tcrdist3 0.82 AP 0.67 AUC
stapler 0.76 AP 0.58 AUC
TCRbridge 0.91 AP 0.81 AUC
Random 0.69 AP 0.50 AUC
E
YLQ (171/305)
YLQ vs GLC score
GLC (134/305)
GLC vs YLQ score
TCRbridge glc
TCRbridge ylq
validating YLQ TCR
validating GLC TCR
F
Prediction of YLQ (pos) vs GLC (neg) reactivity (171/305)
TCRbridge 0.86 AP 0.82 AUC
Random 0.56 AP 0.50 AUC
Prediction of GLC (pos) vs YLQ (neg) reactivity (134/305)
TCRbridge 0.78 AP 0.82 AUC
Random 0.44 AP 0.50 AUC

951 **Figure 4 TCRvdb improves TCR-pMHC reactivity model evaluation**

952 (A) Schematic overview of prediction model evaluation pipeline using validated VDJdb (TCRvdb).
953 Validating and non-validating TCRs are identified using the functional genetic screening results
954 as depicted in Figure 3C. tcrdist3, STAPLER, and TCRbridge are then evaluated on their ability to
955 distinguish identified validating and non-validating TCRs.

956 (B) tcrdist3 distance-based network formed by YLQ- and GLC-annotated TCRs. Grey nodes
957 represent non-validating TCRs, colored nodes represent validating TCRs. Nodes are connected
958 by edges when TCRs represented by those nodes have a tcrdist3 distance below 120.

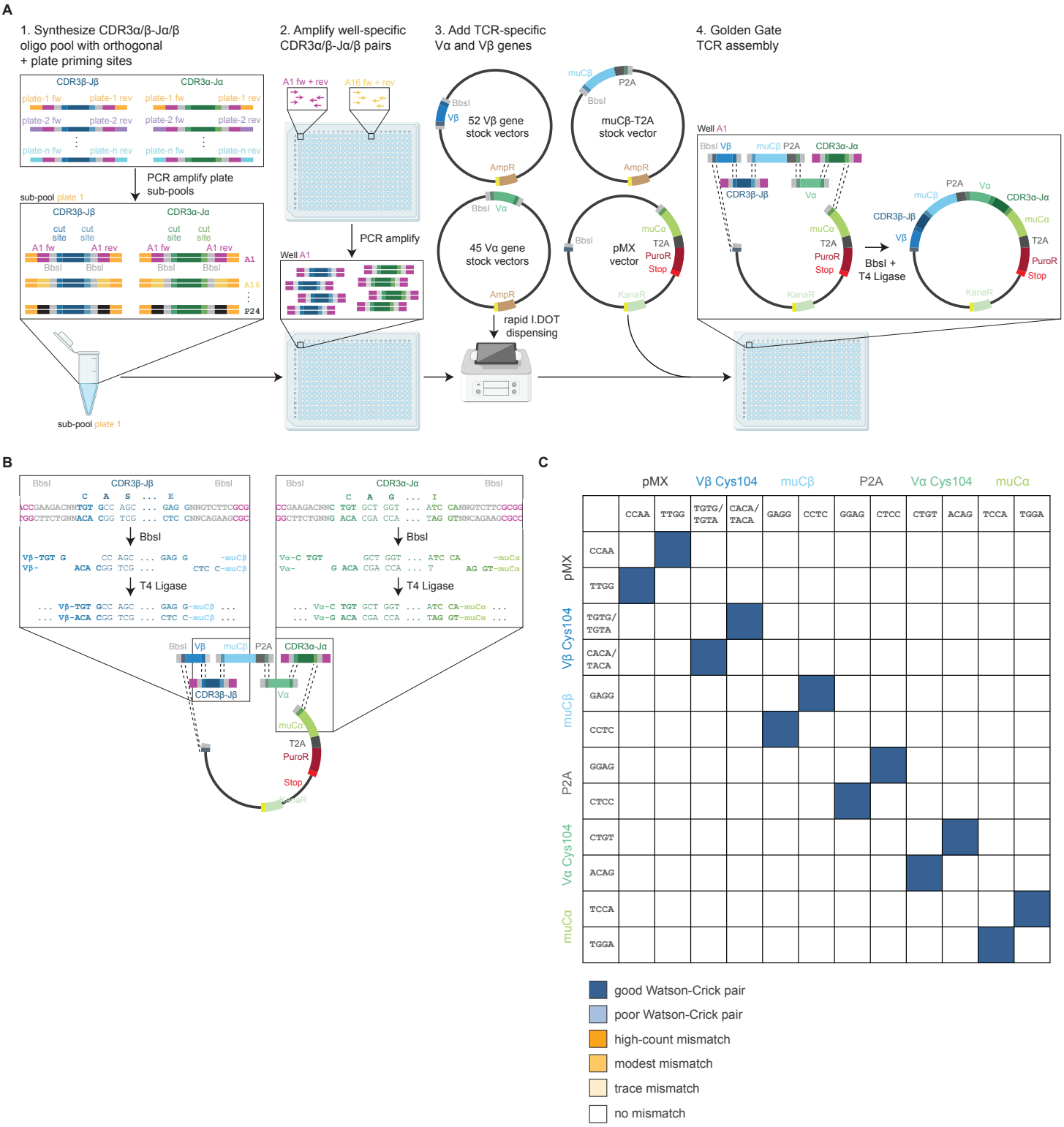
959 (C) AlphaFold3 (AF3) predicted TCR-pMHC structure complexes of a validating (top-left) and non-
960 validating (top-right) YLQ-annotated TCR. Structures are colored by AF3's confidence in the
961 placement of each atom within its local environment (pLDDT; predicted Local Distance
962 Difference Test). The corresponding AlphaBridge[31] circos plots of the validating (bottom-left) and
963 non-validating (bottom-right) predicted structure complexes show pLDDT for TCRα, TCRβ,
964 peptide, MHC, and B2M (outer ring), along with confident interfaces identified by AlphaBridge
965 (inner ring and ribbons). Ribbons representing confident interfaces between TCRα and peptide
966 are colored green, between TCRβ and peptide are colored purple, and all other interfaces are
967 shown in grey. Note the presence of confident interactions between TCR and peptide and TCR
968 and MHC in the left model but not the right model.

969 (D) Comparison of tcrdist3 (orange), STAPLER (purple), and TCRbridge (green) performance in
970 predicting whether YLQ-annotated (left) and GLC-annotated (right) TCRs do or do not validate in
971 functional genetic screening. Note that the labels identified in Fig. 3 were not included in the
972 development of either model. Left: Precision-Recall curve showing the fraction of validating pairs
973 of all pairs (i.e. precision) at the indicated recall (sensitivity) levels. Right: Receiver Operating
974 Characteristic (ROC) curve showing ability of the indicated models to rank validating pairs above
975 non-validating pairs, as measured as the true positive rate (sensitivity) versus the false positive
976 rate (1 − specificity) across prediction thresholds. Average Precision (AP; area under the
977 Precision-Recall curve) summarizes overall model performance on the Precision-Recall curve,
978 with values closer to 1.0 indicating better precision. Area under the ROC curve (AUC)
979 summarizes overall model performance in ranking validating pairs relative to non-validating pairs
980 across prediction thresholds, with values closer to 1.0 indicating a higher ability to distinguish
981 validating from non-validating pairs. Dashed red lines show random performance.

982 (E) Left: Scatterplots comparing TCRbridge-predicted confidence scores of validating YLQ and
983 GLC TCRs for structure predictions in combination with either their cognate or non-cognate
984 epitopes. Top-left plot depicts scores for YLQ TCRs in blue. Bottom-left plot depicts scores for

GLC TCRs in red. Top-right and bottom-right plots depict data for the same TCRs but colored by relative confidence score for both predictions. Top-right: higher scores indicate preference for YLQ over GLC. Bottom-right higher scores indicate preference for GLC over YLQ. Relative scores were computed by calculating the scalar projections of each TCR prediction onto a vector contrasting the two epitopes (either (1, –1) for YLQ or (–1, 1) for GLC), followed by min-max normalization to scale values from 0 (favoring the non-target epitope) to 1 (perfectly aligned with the target epitope).

(F) Performance of TCRbridge in distinguishing cognate (positive) from swapped (negative) validating TCR-epitope pairs using the predicted relative score from (E). Precision-recall and ROC curves show model performance for validating YLQ (top) and GLC (bottom) TCRs. Precision-recall curves show the fraction of cognate epitope pairs of total pairs (precision) at the indicated levels of recall (sensitivity). ROC curves reflect the model's ability to rank cognate epitope pairs relative to swapped pairs. As explained in (D), AP and AUC summarize overall model performance. Dashed red lines represent random performance.


A
1. Synthesize CDR3α/β-Jα/β oligo pool with orthogonal + plate priming sites
CDR3β-Jβ
CDR3α-Jα
plate-1 fw
plate-1 rev
plate-2 fw
plate-2 rev
plate-n fw
plate-n rev
PCR amplify plate sub-pools
sub-pool plate 1
cut site
A1 fw
A1 rev
BbsI
A1
A16
P24
sub-pool plate 1
2. Amplify well-specific CDR3α/β-Jα/β pairs
A1 fw + rev
A16 fw + rev
PCR amplify
Well A1
3. Add TCR-specific Vα and Vβ genes
BbsI
Vβ
52 Vβ gene stock vectors
AmpR
muCβ
P2A
muCβ-T2A stock vector
Vα
45 Vα gene stock vectors
muCα
T2A
PuroR
Stop
pMX vector
KanaR
rapid I.DOT dispensing
4. Golden Gate TCR assembly
Well A1
BbsI + T4 Ligase
B
CDR3β-Jβ
CDR3α-Jα
BbsI
T4 Ligase
C
pMX
Vβ Cys104
muCβ
P2A
Vα Cys104
muCα
CCAA
TTGG
TGTG/TGTA
CACA/TACA
GAGG
CCTC
GGAG
CTCC
CTGT
ACAG
TCCA
TGGA
good Watson-Crick pair
poor Watson-Crick pair
high-count mismatch
modest mismatch
trace mismatch
no mismatch

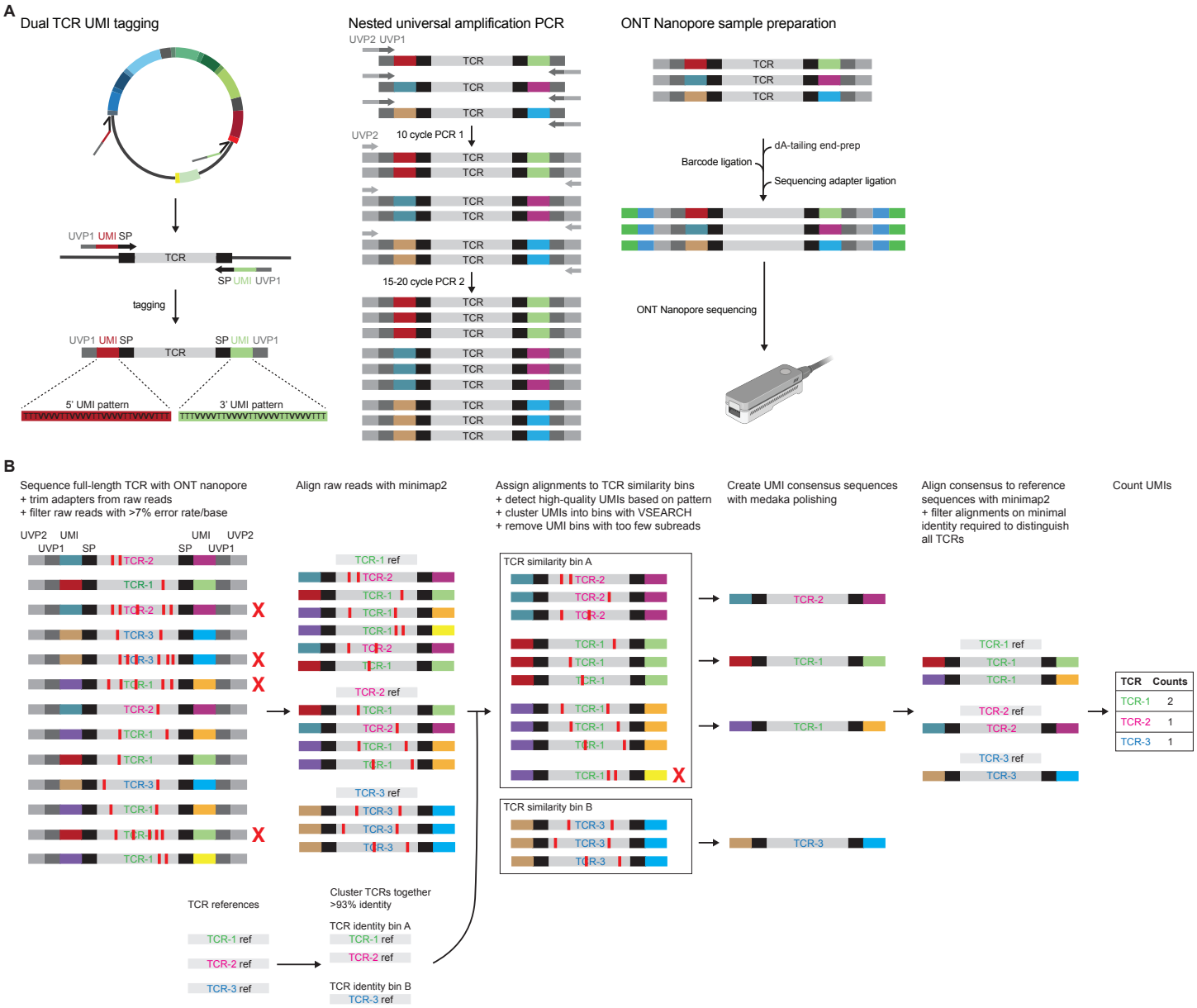
1019 **Supplementary Figure 1 Scalable synthetic TCR library assembly platform**

1020 (A) Schematic overview of robotics-assisted arrayed synthetic TCR library assembly. All TCRs in
1021 a library are assembled in 384-well plate(s), with one unique TCR per well. CDR3α/β-Jα/β
1022 oligonucleotide pools are demultiplexed using highly orthogonal (non-interacting) primers,
1023 indicated in the oligonucleotide sequence in purple[27]. Each well is assigned a unique pair of
1024 orthogonal primers, enabling PCR amplification of DNA encoding the CDR3α-Jα and CDR3β-β
1025 segments of a unique TCR. Importantly, by amplifying all CDR3α-Jα and CDR3β-β segments
1026 individually until the PCR plateau phase, we can overcome the non-uniformity of the supplier-
1027 manufactured oligonucleotide pools, thereby generating highly uniform libraries. Robotics-
1028 assisted dispensing is then used to add the corresponding TRAV and TRBV fragments to each
1029 well. Golden Gate assembly is used to join the CDR3α-Jα and CDR3β-β, TRAV, and TRBV
1030 fragments with the murine TRBC-P2A (muTRBC-P2A), yielding a full-length TCRβ-P2A-TCRα gene
1031 segment encoded in the pMX retroviral vector.

1032 (B) Schematic overview of Golden Gate-based assembly of TRBV, CDR3β-Jβ, muTRBC, and TRAV,
1033 CDR3α-Jα, and muTRAC fragments. BbsI is used to generate user-defined four-base sticky
1034 overhangs that can be end-joined by T4 DNA Ligase. The four-base overhang that joins TRBV and
1035 CDR3β-Jβ is formed by the conserved Cys104 `TGT` codon at the start of the CDR3β-Jβ, together
1036 with a conserved `G` in all TRBV genes except TRBV20-1 and TRBV29-1, which instead contain a
1037 conserved `A` (conservation identified using sequence analysis of ~89,000 TCRs[52]). The four-base
1038 overhang that joins TRAV and CDR3α-Jα is formed by the conserved Cys104 `TGT` codon at the
1039 start of the CDR3α-Jα, together with a conserved `C` nucleotide (conservation identified using the
1040 same sequence analysis of ~89,000 TCRs). CDR3β-Jβ is joined to muTRBC using the `GAGG`
1041 overhang, and CDR3α-Jα is joined to muTRAC using the `TCCA` overhang. All other used four-base
1042 overhangs are shown in (C).

1043 (C) NEBridge Ligase Fidelity Viewer output for all four-base overhangs used to join fragments
1044 together forming a synthetic TCR. NEBridge Ligase Fidelity Viewer is based on a comprehensive
1045 T4 DNA ligase end-joining fidelity dataset[28]. Empty squares indicate no predicted mismatches.

1046
1047
1048
1049
1050
1051
1052



Sensitivity (% TCRs successfully assembled)	97.0% (3406/3510)
Precision (% sequence-perfect TCR UMIs)	76.7%
95th/5th percentile ratio of TCR abundance	6.85

Sensitivity (% TCRs successfully assembled)	98.6% (969/983)
Precision (% sequence-perfect TCR UMIs)	69.6%
95th/5th percentile ratio of TCR abundance	5.02

Sensitivity (% TCRs successfully assembled)	96.8% (968/1000)
Precision (% sequence-perfect TCR UMIs)	63.8%
95th/5th percentile ratio of TCR abundance	2.52

Sensitivity (% TCRs successfully assembled)	97.2% (2811/2892)
Precision (% sequence-perfect TCR UMIs)	70.2%
95th/5th percentile ratio of TCR abundance	4.60

1053 **Supplementary Figure 2 Nanopore-sequencing TCR library QC analysis**

1054 (A) Schematic overview of Oxford Nanopore Technologies (ONT) full-length TCR DNA sequencing
1055 library preparation. Unique TCR molecules are tagged on both ends with unique molecular
1056 identifiers (UMIs) using a 2-cycle PCR. UMI-tagged TCRs are then amplified by two nested PCRs.
1057 ONT barcodes and native sequence adapters are ligated, and resulting libraries are sequenced.

1058 (B) Schematic overview of the full-length TCR UMI sequence error correction pipeline. Due to the
1059 inherent base-calling errors of raw nanopore reads, TCRs with high sequence identity are difficult
1060 to distinguish. To address this, reference TCR sequences from the library are first self-aligned
1061 using minimap2, and sequences sharing greater than 93% identity are clustered into TCR
1062 similarity bins. Next, raw TCR nanopore reads are filtered using VSEARCH[40] to exclude reads with
1063 more than a 7% expected error rate[46] (note that this falls within the 93% identity threshold used
1064 for TCR reference sequence binning), and non-filtered raw reads are aligned to the reference TCR
1065 library using minimap2[47]. Each read is then assigned to a TCR similarity bin based on the
1066 reference sequence it aligns to. Within each TCR similarity bin, high-quality UMIs are identified
1067 and clustered using VSEARCH to group subreads derived from the same unique molecule into
1068 UMI bins. Importantly, UMI clustering is performed after alignment and binning by TCR identity,
1069 thereby reducing the chance of UMI collisions and computational load, as fewer UMI
1070 comparisons are required. Use of paired UMIs on both ends of the full-length TCR facilitates the
1071 detection and removal of PCR chimeras by filtering out UMI bins with insufficient subread
1072 support. A consensus sequence for each UMI bin is then created using the medaka polisher.
1073 Finally, high-accuracy consensus sequences are aligned to the reference TCR library using
1074 minimap2, and UMIs are counted for each unique TCR reference.

1075 (C) Histogram of TCR UMI count density for the VDJdb-10 plasmid library, when either
1076 synthesized with one plate omitted (left, plate 1 omitted) or synthesized in full (right). TCR UMI
1077 counts were normalized such that the area under each histogram integrates to 1, thereby
1078 representing the probability density of TCR UMI counts. The histogram for plate 1 TCRs is shown
1079 in purple, and all VDJdb-10 TCRs (including plate 1) are shown in grey. The orange vertical line
1080 marks the minimum TCR UMI count threshold used to filter out low UMI count noise.

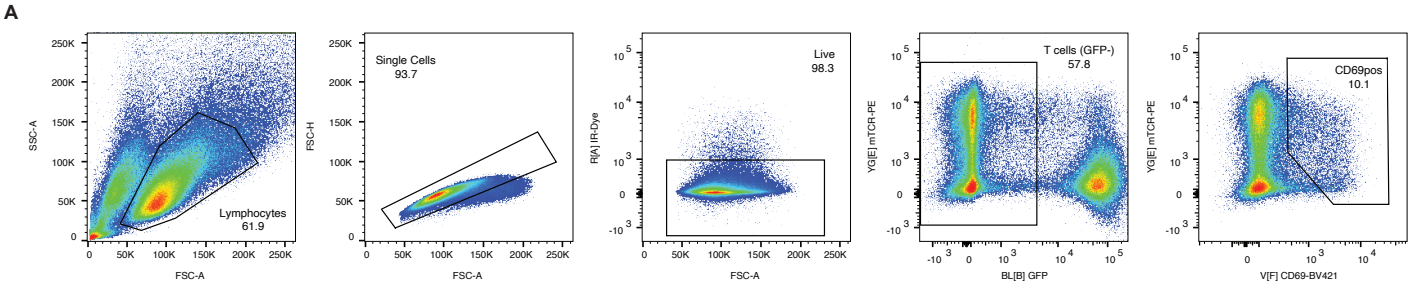
1081 (D) Histogram of TCR UMI counts of the TCR-s1, TCR-s2, TCR-s3, and TCR-s4 libraries and their
1082 corresponding count QC summary statistics.

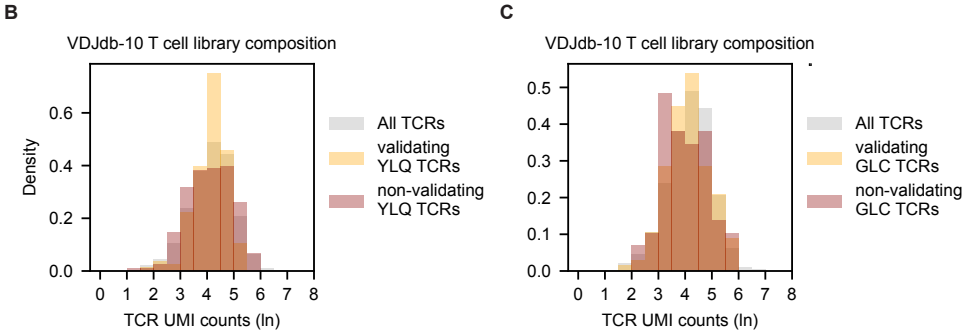
1083

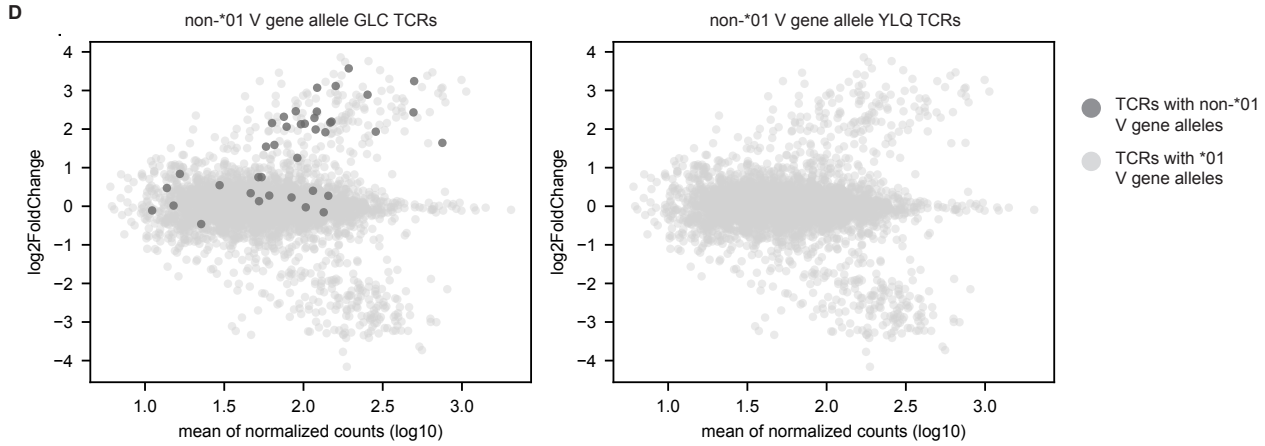
1084

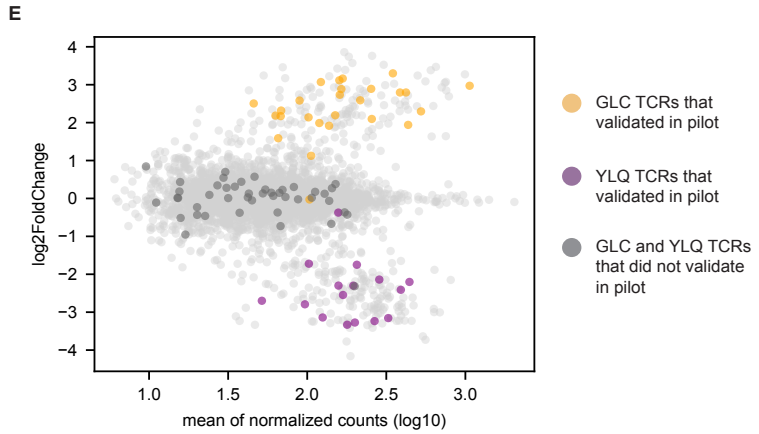
1085

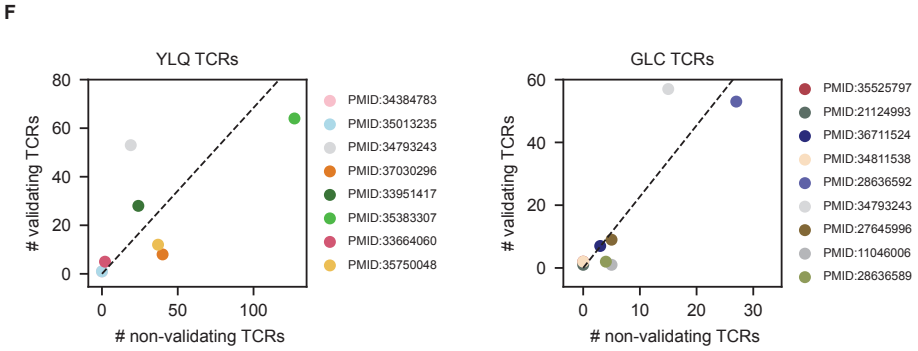
1086


A
SSC-A
FSC-A
Lymphocytes
61.9
FSC-H
Single Cells
93.7
R[A] IR-Dye
Live
98.3
YG[E] mTCR-PE
BL[B] GFP
T cells (GFP-)
57.8
V[F] CD69-BV421
CD69pos
10.1

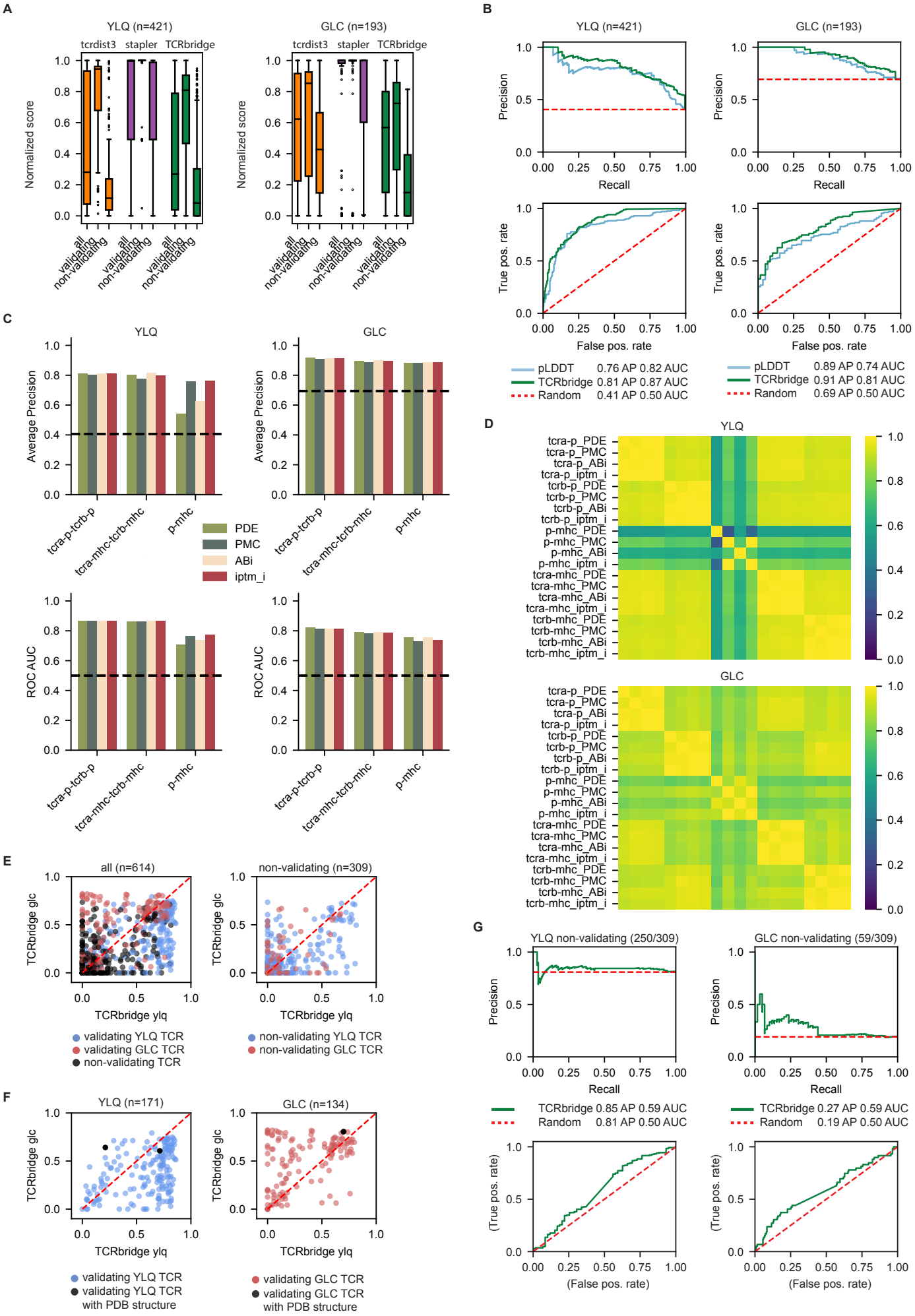

B
VDJdb-10 T cell library composition
Density
TCR UMI counts (ln)
All TCRs
validating YLQ TCRs
non-validating YLQ TCRs
C
VDJdb-10 T cell library composition
TCR UMI counts (ln)
All TCRs
validating GLC TCRs
non-validating GLC TCRs


D
non-*01 V gene allele GLC TCRs
non-*01 V gene allele YLQ TCRs
log2FoldChange
mean of normalized counts (log10)
TCRs with non-*01 V gene alleles
TCRs with *01 V gene alleles


E
log2FoldChange
mean of normalized counts (log10)
GLC TCRs that validated in pilot
YLQ TCRs that validated in pilot
GLC and YLQ TCRs that did not validate in pilot


F
YLQ TCRs
validating TCRs
non-validating TCRs
PMID:34384783
PMID:35013235
PMID:34793243
PMID:37030296
PMID:33951417
PMID:35383307
PMID:33664060
PMID:35750048
GLC TCRs
PMID:35525797
PMID:21124993
PMID:36711524
PMID:34811538
PMID:28636592
PMID:34793243
PMID:27645996
PMID:11046006
PMID:28636589

1087 **Supplementary Figure 3 Pooled functional genetic screening of the VDJdb-10 TCR library**
1088 (A) Gating strategy for the isolation of CD69^{+} Jurkat T cells in the VDJdb-10 TCR library screen
1089 (relating to Figure 3A and 3C).
1090 (B) Histogram of TCR UMI count density of all VDJdb-10 TCRs (grey), and of validating (red) and
1091 non-validating (purple) YLQ TCRs in the screen of Figure 3C. TCR UMI counts were normalized
1092 such that the area under each histogram integrates to 1.
1093 (C) Histogram of TCR UMI count density of all VDJdb-10 TCRs (grey), and of validating (red) and
1094 non-validating (purple) GLC TCRs in the screen of Figure 3C. TCR UMI counts were normalized
1095 such that the area under each histogram integrates to 1.
1096 (D) MA plot from Figure 3C, with colors representing TCRs originally described with non-*01 (dark
1097 grey) or *01 (light grey) TRAV or TRBV alleles. Non-*01-alleles are collapsed to *01 in our synthetic
1098 TCR assembly method. Note that 10X Genomics single-cell sequencing workflow does not
1099 annotate non-*01 alleles, explaining the lack of non-*01 TRAV or TRBV alleles for all YLQ-
1100 annotated TCRs.
1101 (E) MA plot from Figure 3C, with dots colored by validation outcome in Figure 1A. Colors depict
1102 validating YLQ-annotated (purple) and GLC-annotated TCRs (yellow), and non-validating YLQ-
1103 annotated and GLC-annotated TCRs (both dark grey).
1104 (F) Scatterplots depicting the number of non-validating TCRs and validating TCRs contributed by
1105 individual studies. Studies are shown by PubMed unique study identification number.
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120


A
YLQ (n=421)
GLC (n=193)
tcrdist3
stapler
TCRbridge
Normalized score
all
validating
non-validating
B
YLQ (n=421)
GLC (n=193)
Precision
Recall
True pos. rate
False pos. rate
pLDDT 0.76 AP 0.82 AUC
TCRbridge 0.81 AP 0.87 AUC
Random 0.41 AP 0.50 AUC
pLDDT 0.89 AP 0.74 AUC
TCRbridge 0.91 AP 0.81 AUC
Random 0.69 AP 0.50 AUC
C
YLQ
GLC
Average Precision
ROC AUC
tcra-p-tcrb-p
tcra-mhc-tcrb-mhc
p-mhc
PDE
PMC
ABi
iptm_i
D
YLQ
GLC
tcra-p_PDE
tcra-p_PMC
tcra-p_ABi
tcra-p_iptm_i
tcrb-p_PDE
tcrb-p_PMC
tcrb-p_ABi
tcrb-p_iptm_i
p-mhc_PDE
p-mhc_PMC
p-mhc_ABi
p-mhc_iptm_i
tcra-mhc_PDE
tcra-mhc_PMC
tcra-mhc_ABi
tcra-mhc_iptm_i
tcrb-mhc_PDE
tcrb-mhc_PMC
tcrb-mhc_ABi
tcrb-mhc_iptm_i
E
all (n=614)
non-validating (n=309)
TCRbridge glc
TCRbridge ylq
validating YLQ TCR
validating GLC TCR
non-validating TCR
non-validating YLQ TCR
non-validating GLC TCR
F
YLQ (n=171)
GLC (n=134)
TCRbridge glc
TCRbridge ylq
validating YLQ TCR
validating YLQ TCR with PDB structure
validating GLC TCR
validating GLC TCR with PDB structure
G
YLQ non-validating (250/309)
GLC non-validating (59/309)
Precision
Recall
TCRbridge 0.85 AP 0.59 AUC
Random 0.81 AP 0.50 AUC
TCRbridge 0.27 AP 0.59 AUC
Random 0.19 AP 0.50 AUC
(True pos. rate)
(False pos. rate)

1121 **Supplementary Figure 4 Analysis of model prediction scores**

1122 (A) Boxplot of tcrdist3, STAPLER, and TCRbridge normalized predicted confidence in TCRs being
1123 reactive. Confidence scores were normalized by dividing by the maximal prediction score for
1124 each model. Boxes indicate model median and 25th/75th percentiles of the normalized scores,
1125 whiskers show min/max except for outliers determined using the inter quartile range (IQR).

1126 (B) Comparison of the performance of TCRbridge and AlphaFold3 (AF3) mean pLDDT (predicted
1127 Local Distance Difference Test) confidence in the CDR3αβ and peptide regions of predicted TCR-
1128 pMHC complex structure in distinguishing validating and non-validating YLQ and GLC TCRs.
1129 Interpretation of Precision-Recall and ROC curves is described in Figure 4D.

1130 (C) Comparison of the performance of the indicated AF3 per-residue confidence scores and their
1131 combinations in predicted interfaces between the TCR and pMHC in distinguishing validating and
1132 non-validating YLQ and GLC TCRs. Performance of the following AF3 confidence metrics in its
1133 predicted interface were analyzed: interface predicted TM-score (ipTM), predicted distance error
1134 (PDE), and predicted merged confidence (PMC)—a combination of predicted aligned error (PAE)
1135 and pLDDT as performed by AlphaBridge[31]. TCRbridge uses ABi, which is formed by the
1136 combination of PDE and PMC scores. We evaluated performance using the sum of confidence in
1137 the TCRα-peptide and TCRβ-peptide interfaces (tcra-p-tcrb-p), the sum of confidence in the
1138 TCRα-MHC and TCRβ-MHC interfaces (tcra-mhc-tcrb-mhc), or the confidence in the peptide -
1139 MHC (p-mhc) interface.

1140 (D) Spearman correlation of confidence scores for the indicated AF3 and AlphaBridge metrics for
1141 predicted TCR-pMHC interfaces. Used confidence metrics and interfaces are described in (C).

1142 (E) Scatterplots depicting TCRbridge-predicted confidence scores in TCR-pMHC interfaces of all
1143 YLQ- and GLC-annotated (i.e., validating and non-validating, left) and of non-validating YLQ and
1144 GLC (right) TCRs in predicted structures with either their cognate or non-cognate epitopes.

1145 (F) Scatterplots depicting TCRbridge-predicted confidence scores in TCR-pMHC interfaces of
1146 validating YLQ (left) and validating GLC (right) TCRs in predicted structures with either their
1147 cognate or non-cognate epitopes. Dots representing TCRs deposited in PDB before the
1148 AlphaFold3 training cutoff date are highlighted in black.

1149 (G) Performance of TCRbridge in scoring predicted structures with presumed cognate peptides
1150 and non-cognate peptides for non-validating TCRs. Precision-recall and ROC curves show
1151 performance for non-validating YLQ (top) and GLC (bottom) TCRs. Note the limited performance
1152 for non-validating TCRs as compared to that observed for validating TCRs, as described in Figure
1153 4E. Dashed red lines represent random performance.

1154

1155 **Supplemental Information**

1156 Table S1 Parsed VDJdb database

1157 Table S2 Well-specific orthogonal primers

1158 Table S3 Plate-specific orthogonal primers

1159 Table S4 TRAV and TRBV sequences

1160 Table S5 Nanopore primers

1161

1162

1163 **References**

1164 1. Mora, T. & Walczak, A. M. How many different clonotypes do immune repertoires

1165 contain? *Curr. Opin. Syst. Biol.* **18**, 104–110 (2019).

1166 2. Schumacher, T. N., Scheper, W. & Kvistborg, P. Cancer Neoantigens. *Annu. Rev.*

1167 *Immunol.* **37**, 173–200 (2019).

1168 3. Zinkernagel, R. M. & Doherty, P. C. Immunological surveillance against altered self

1169 components by sensitised T lymphocytes in lymphocytes choriomeningitis. *Nature* **251**,

1170 547–548 (1974).

1171 4. Hedrick, S. M., Cohen, D. I., Nielsen, E. A. & Davis, M. M. Isolation of cDNA clones

1172 encoding T cell-specific membrane-associated proteins. *Nature* **308**, 149–153 (1984).

1173 5. Yanagi, Y. *et al.* A human T cell-specific cDNA clone encodes a protein having extensive

1174 homology to immunoglobulin chains. *Nature* **308**, 145–149 (1984).

1175 6. Hedrick, S. M., Nielsen, E. A., Kavaler, J., Cohen, D. I. & Davis, M. M. Sequence

1176 relationships between putative T-cell receptor polypeptides and immunoglobulins. *Nature*

1177 **308**, 153–158 (1984).

1178 7. Yang, X. *et al.* Autoimmunity-associated T cell receptors recognize HLA-B*27-bound

1179 peptides. *Nature* **612**, 771–777 (2022).

1180 8. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science*

1181 **348**, 69–74 (2015).

1182 9. Barouch, D. H. *et al.* Evaluation of a mosaic HIV-1 vaccine in a multicentre, randomised,
1183 double-blind, placebo-controlled, phase 1/2a clinical trial (APPROACH) and in rhesus
1184 monkeys (NHP 13-19). *The Lancet* **392**, 232–243 (2018).
1185 10. Li, X. *et al.* Molecular basis of differential HLA class I-restricted T cell recognition of
1186 a highly networked HIV peptide. *Nat. Commun.* **14**, 2929 (2023).
1187 11. Barouch, D. H. *et al.* Mosaic HIV-1 Vaccines Expand the Breadth and Depth of
1188 Cellular Immune Responses in Rhesus Monkeys. *Nat. Med.* **16**, 319–323 (2010).
1189 12. Fischer, W. *et al.* Polyvalent vaccines for optimal coverage of potential T-cell
1190 epitopes in global HIV-1 variants. *Nat. Med.* **13**, 100–106 (2007).
1191 13. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a
1192 three-track neural network. *Science* **373**, 871–876 (2021).
1193 14. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with
1194 AlphaFold 3. *Nature* **630**, 493–500 (2024).
1195 15. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*
1196 **596**, 583–589 (2021).
1197 16. wwPDB consortium. Protein Data Bank: the single global archive for 3D
1198 macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).
1199 17. Young, J. Y. *et al.* OneDep: Unified wwPDB System for Deposition, Biocuration, and
1200 Validation of Macromolecular Structures in the PDB Archive. *Structure* **25**, 536–545
1201 (2017).
1202 18. Gore, S. *et al.* Validation of Structures in the Protein Data Bank. *Struct. England1993*
1203 **25**, 1916–1927 (2017).
1204 19. Kwee, B. P. Y. *et al.* STAPLER: Efficient learning of TCR-peptide specificity
1205 prediction from full-length TCR-peptide data. 2023.04.25.538237 Preprint at
1206 https://doi.org/10.1101/2023.04.25.538237 (2023).

1207 20. Hudson, D., Fernandes, R. A., Basham, M., Ogg, G. & Koohy, H. Can we predict T
1208 cell specificity with digital biology and machine learning? *Nat. Rev. Immunol.* **23**, 511–
1209 521 (2023).
1210 21. Meysman, P. *et al.* Benchmarking solutions to the T-cell receptor epitope prediction
1211 problem: IMMREP22 workshop report. *ImmunoInformatics* **9**, 100024 (2023).
1212 22. Goncharov, M. *et al.* VDJdb in the pandemic era: a compendium of T cell receptors
1213 specific for SARS-CoV-2. *Nat. Methods* **19**, 1017–1019 (2022).
1214 23. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2024 update. *Nucleic Acids*
1215 *Res.* **53**, D436–D443 (2025).
1216 24. Zong, S. *et al.* Very rapid cloning, expression and identifying specificity of T-cell
1217 receptors for T-cell engineering. *PLOS ONE* **15**, e0228112 (2020).
1218 25. Hu, Z. *et al.* A cloning and expression system to probe T-cell receptor specificity and
1219 assess functional avidity to neoantigens. *Blood* **132**, 1911–1921 (2018).
1220 26. Moravec, Z. *et al.* Discovery of tumor-reactive T cell receptors by massively parallel
1221 library synthesis and screening. *Nat. Biotechnol.* **43**, 214–222 (2025).
1222 27. Subramanian, S. K., Russ, W. P. & Ranganathan, R. A set of experimentally
1223 validated, mutually orthogonal primers for combinatorially specifying genetic components.
1224 *Synth. Biol.* **3**, ysx008 (2018).
1225 28. Potapov, V. *et al.* Comprehensive Profiling of Four Base Overhang Ligation Fidelity
1226 by T4 DNA Ligase and Application to DNA Assembly. *ACS Synth. Biol.* **7**, 2665–2674
1227 (2018).
1228 29. Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique
1229 molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods* **18**, 165–169
1230 (2021).

1231 30. Mayer-Blackwell, K., Fiore-Gartland, A. & Thomas, P. G. Flexible Distance-Based
1232 TCR Analysis in Python with tcrdist3. *Methods Mol. Biol. Clifton NJ* **2574**, 309–366
1233 (2022).
1234 31. Álvarez-Salmoral, D. *et al.* AlphaBridge: tools for the analysis of predicted
1235 macromolecular complexes. 2024.10.23.619601 Preprint at
1236 https://doi.org/10.1101/2024.10.23.619601 (2024).
1237 32. Bradley, P. Structure-based prediction of T cell receptor:peptide-MHC interactions.
1238 *eLife* **12**, e82813 (2023).
1239 33. Culka, M. *et al.* Predicting specificity of TCR-pMHC interactions using machine
1240 learning and biophysical models. 2025.04.04.647165 Preprint at
1241 https://doi.org/10.1101/2025.04.04.647165 (2025).
1242 34. Lawson, C. L. *et al.* Cryo-EM model validation recommendations based on outcomes
1243 of the 2019 EMDataResource challenge. *Nat. Methods* **18**, 156–164 (2021).
1244 35. Read, R. J. *et al.* A new generation of crystallographic validation tools for the protein
1245 data bank. *Struct. Lond. Engl. 1993* **19**, 1395–1412 (2011).
1246 36. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically
1247 disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
1248 37. Motmaen, A. *et al.* Peptide-binding specificity prediction using fine-tuned protein
1249 structure prediction networks. *Proc. Natl. Acad. Sci.* **120**, e2216697120 (2023).
1250 38. Scheper, W. *et al.* Low and variable tumor reactivity of the intratumoral TCR
1251 repertoire in human cancers. *Nat. Med.* **25**, 89–94 (2019).
1252 39. *Introduction to Algorithms*. (MIT Press, Cambridge, Mass, 2009).
1253 40. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile
1254 open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

1255 41. Mock, U., Thiele, R., Uhde, A., Fehse, B. & Horn, S. Efficient Lentiviral
1256 Transduction and Transgene Expression in Primary Human B Cells. *Hum. Gene Ther.*
1257 *Methods* **23**, 408–415 (2012).
1258 42. Povlsen, H. R. *et al.* Improved T cell receptor antigen pairing through data-driven
1259 filtering of sequencing information from single cells. *eLife* **12**, e81810 (2023).
1260 43. Mei, S. *et al.* Anthem: a user customised tool for fast and accurate prediction of
1261 binding between peptides and HLA class I molecules. *Brief. Bioinform.* **22**, bbaa415
1262 (2021).
1263 44. Robinson, J. *et al.* IPD-IMGT/HLA Database. *Nucleic Acids Res.* **48**, D948–D955
1264 (2020).
1265 45. Amstler, S. *et al.* Nanopore sequencing with unique molecular identifiers enables
1266 accurate mutation analysis and haplotyping in the complex lipoprotein(a) KIV-2 VNTR.
1267 *Genome Med.* **16**, 117 (2024).
1268 46. Edgar, R. C. & Flyvbjerg, H. Error filtering, pair assembly and error correction for
1269 next-generation sequencing reads. *Bioinformatics* **31**, 3476–3482 (2015).
1270 47. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf.*
1271 *Engl.* **34**, 3094–3100 (2018).
1272 48. Muzellec, B., Teleńczuk, M., Cabeli, V. & Andreux, M. PyDESeq2: a python package
1273 for bulk RNA-seq differential expression analysis. *Bioinformatics* **39**, btad547 (2023).
1274 49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and
1275 dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
1276 50. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational
1277 molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
1278 51. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. Preprint at
1279 https://doi.org/10.48550/arXiv.1201.0490 (2018).

1280 52. Wu, T. D. *et al.* Peripheral T cell expansion predicts tumour infiltration and clinical
1281 response. *Nature* **579**, 274–278 (2020).
1282