Generative Design of High-affinity T-cell Receptors by Progressive Learning with Structural Confidence

Xinyuan Zhu*[1], Qingshuo Jin*[2], Jiadong Lu[2], Jun Wu[2], Lixia Zhu[3,4,5], Yuhua Shang[4], Tengchuan Jin†[3,4,5], and Fuli Feng†[2]

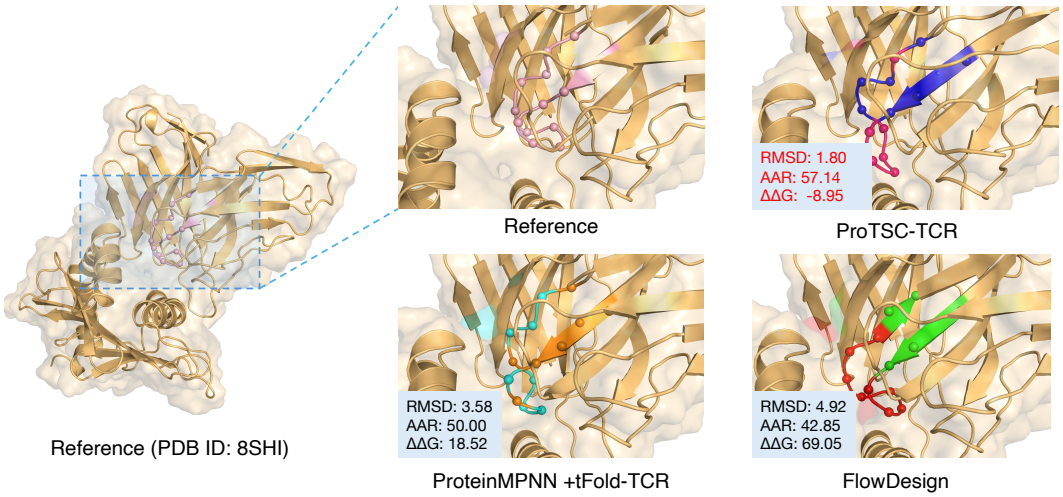
[1]School of Information Science and Technology, University of Science and Technology of China
[2]School of Artificial Intelligence and Data Science, University of Science and Technology of China
[3]Laboratory of Structural Immunology, State Key Laboratory of Immune Response and Immunotherapy, Division of Life Sciences and Medicine, University of Science and Technology of China
[4]Anhui Genebiol Biotech. Ltd
[5]Institute of Health and Medicine, Hefei Comprehensive National Science Center

Supplementary Information

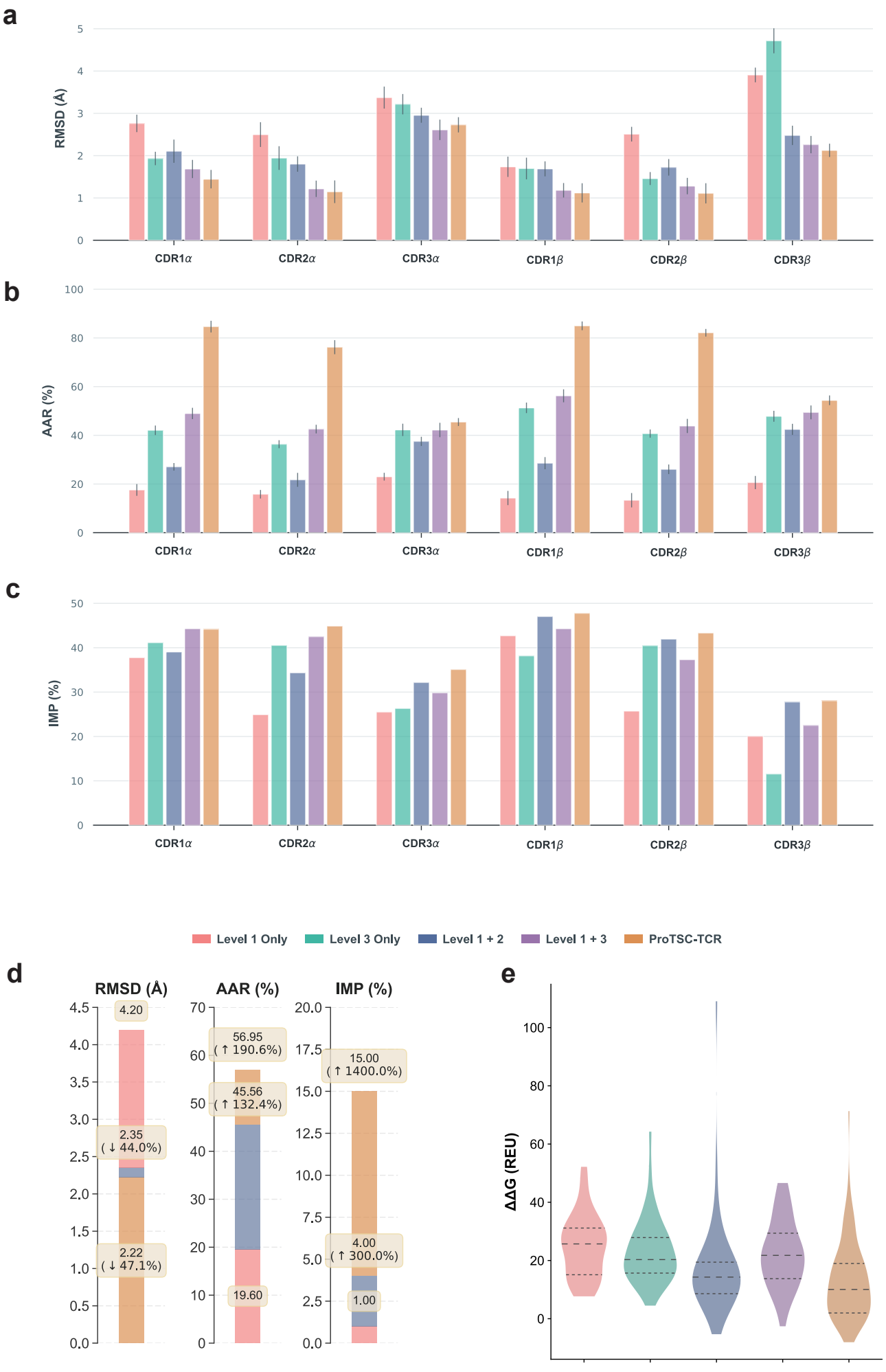
Contents

*These authors contributed equally to this work.
†Correspondence: jint@ustc.edu.cn; fengfl@ustc.edu.cn

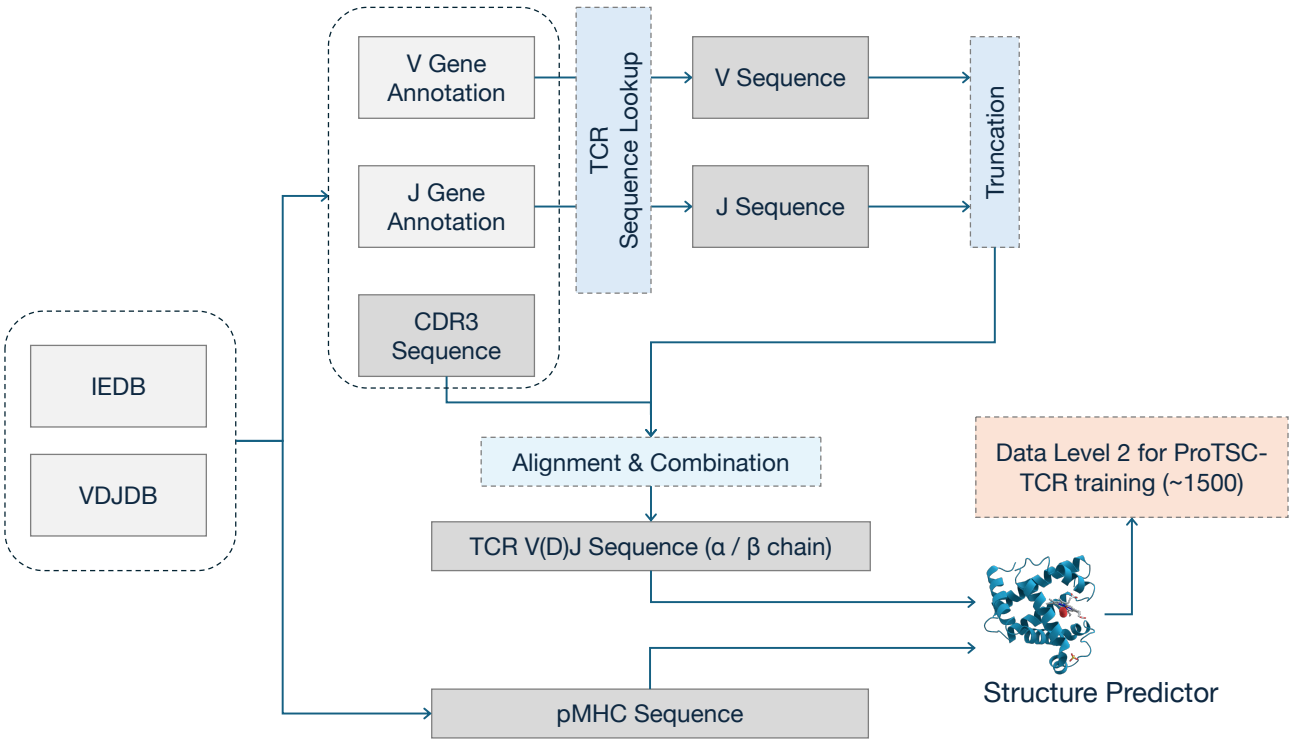
1 Supplementary Figures



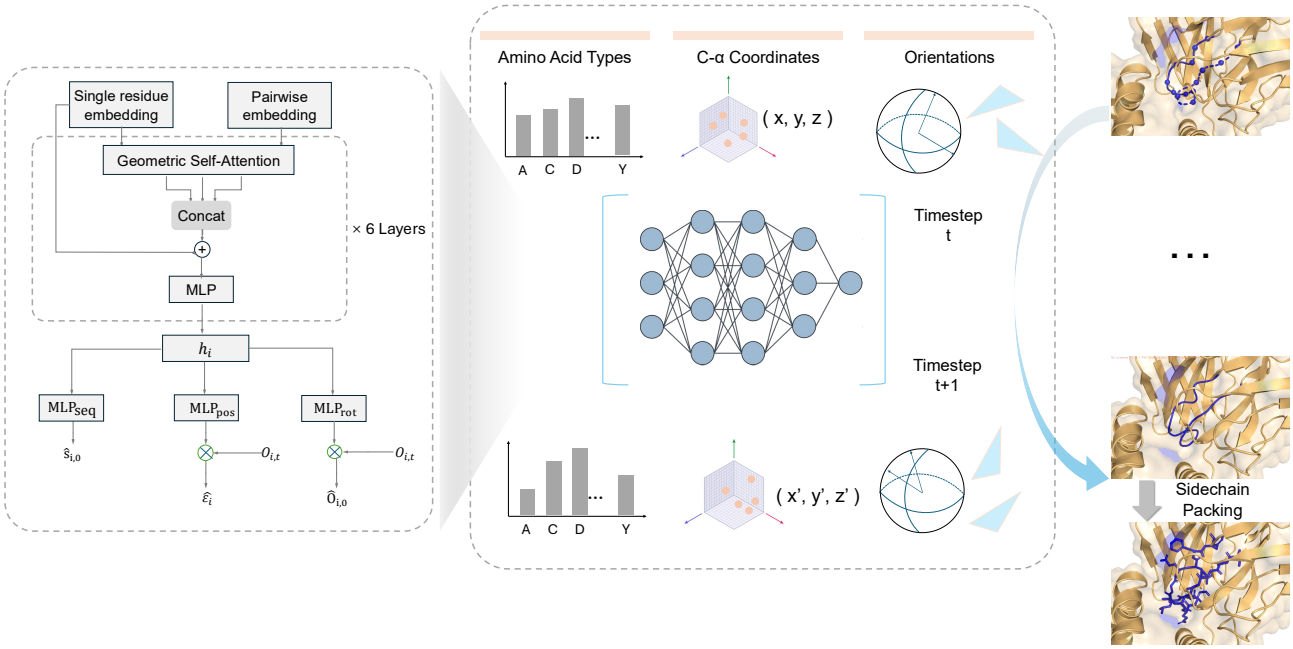
Supplementary Fig. 1 | Aligned zoom-in views of the AAR-ranked CDR3β designs for 8SHI in the native complex (gold). The reference panel shows the native loop and generated loops colored by method. ProTSC-TCR accurately recapitulates the native CDR3β backbone trajectory and interface placement. Baselines produce trade-off as their AAR-ranked designs exhibit larger deviations in loop placement and less favorable predicted binding energetics.


a
RMSD (Å)
CDR1α
CDR2α
CDR3α
CDR1β
CDR2β
CDR3β
b
AAR (%)
c
IMP (%)
Level 1 Only
Level 3 Only
Level 1 + 2
Level 1 + 3
ProTSC-TCR
d
RMSD (Å)
AAR (%)
IMP (%)
4.20
2.35
(↓ 44.0%)
2.22
(↓ 47.1%)
56.95
(↑ 190.6%)
45.56
(↑ 132.4%)
19.60
15.00
(↑ 1400.0%)
4.00
(↑ 300.0%)
1.00
e
ΔΔG (REU)

Supplementary Fig. 2 ***(previous page)*** **| Ablation across CDR regions and an additional case study under the** ***Top-100*** **protocol. a–c**, Performance across the six CDR regions (CDR1α, CDR2α, CDR3α, CDR1β, CDR2β, CDR3β) on the 35-target benchmark. Bars report backbone RMSD (**a**), AAR (**b**), and IMP (**c**) for variants: Level 1 Only, Level 3 Only, Level 1+2, Level 1+3, and the ProTSC-TCR. **d**, Stepwise improvements on an additional target (PDB ID 8SHI), highlighting how incorporating Data Level 2 and 3 contributes to superior RMSD, AAR, and IMP than Level 1 Only model. **e**, Distribution of Rosetta $\Delta\Delta G$ over 100 independent designs for 8SHI, showing a progressive shift toward more favorable predicted binding energetics with ProTSC-TCR.



Supplementary Fig. 3 | Workflow for full TCR sequence reconstruction and construction of the Level 2 data. To construct the dataset from binding records that typically lack full variable-domain sequences, a reconstruction pipeline was implemented. Germline V and J gene segments are retrieved based on annotated gene identifiers and truncated at conserved recombination motifs. These segments are then aligned and combined with the specific CDR3 sequence to reconstruct complete TCR α and β V(D)J sequences. Finally, the reconstructed TCR sequences are paired with pMHC sequences and processed by a structure predictor (ColabFold-AF2) to generate 1,500 predicted TCR–pMHC complexes for ProTSC-TCR training.



Supplementary Fig. 4 | Model architecture and generative inference workflow of ProTSC-TCR. The model utilizes an SE(3)-equivariant architecture (left) that processes single-residue and pairwise geometric embeddings through invariant self-attention layers to produce context-aware hidden representations (h_i). These representations are decoded by three modality-specific heads (MLP_{seq}, MLP_{pos}, and MLP_{rot}). The geometric predictions are computed in local frames and projected globally to ensure equivariance. The inference pipeline (right) demonstrates the iterative denoising of the joint state (sequence, C_α coordinates, and orientations) from random noise, followed by side-chain packing to generate the final full-atom structure.

2 Supplementary Tables

Supplementary Table 3 | Experimental SPR affinities and interface features for top-ranked KRAS-G12V TCR designs. The wild-type reference complex (PDB ID 8I5D) and the top-10 energy-ranked ProTSC-TCR CDR3β designs. For each unique CDR3β sequence, we report the FoldX-predicted binding energy change ($\Delta\Delta G$), the number of interfacial hydrogen bonds (#HB) and salt bridges (#SB), and the buried interface area (Å^2) computed from the structure generated by ProTSC-TCR, together with the SPR-measured dissociation constant (K_D, μM). Duplicate-ranked entries that share an identical CDR3β sequence are highlighted with matching colors, share the same measured K_D, and omit repeated structural statistics.

ID	$\Delta\Delta G$ (FoldX)	#HB	#SB	Interface Area	KD (μM)	CDR3β
Reference (PDB ID 8I5D)	–	*2*	*3*	*349.2*	*37.8*	***CASSLEGTVEETLYF***
Design-1	-33.60	4	3	474.7	22.9	**CASSLLASAFGELYF**
Design-2	-30.57	6	3	560.7	8.17	**CASSLRGGYFQPLHF**
Design-3	-28.79	–	–	–	22.9	**CASSLLASAFGELYF**
Design-4	-27.23	5	3	538.9	14.4	**CASSWWGDGGYYLYF**
Design-5	-24.21	6	3	548.6	23.2	**CASSFSGLDGQPQTF**
Design-6	-23.86	3	3	545.8	17.1	**CASSILGEPGNYVYF**
Design-7	-23.19	–	–	–	37.8	**CASSLEGTVEETLYF**
Design-8	-21.91	7	4	555.1	3.54	**CASSLWLDMFYTLYF**
Design-9	-20.69	3	2	413.4	8.67	**CASSDWFGSLDTLFF**
Design-10	-19.73	–	–	–	3.54	**CASSLWLDMFYTLYF**

Supplementary Table 4 | Experimental SPR affinities and interface features for top-ranked SARS-CoV-2 Spike-Y453F TCR designs. The wild-type reference complex (PDB ID 8YE4) and the top-10 energy-ranked ProTSC-TCR CDR3β designs. For each unique CDR3β sequence, we report the FoldX-predicted binding energy change ($\Delta\Delta G$), the number of interfacial hydrogen bonds (#HB) and salt bridges (#SB), and the buried interface area (Å^2) computed from the structure generated by ProTSC-TCR, together with the SPR-measured dissociation constant (K_D, μM). Duplicate-ranked entries that share an identical CDR3β sequence are highlighted with matching colors, share the same measured K_D, and omit repeated structural statistics.

ID	$\Delta\Delta G$ (FoldX)	#HB	#SB	Interface Area	KD (μM)	CDR3β
Reference (PDB ID 8YE4)	\	*0*	*0*	*155*	*6.37*	*CASSETGGYEQYF*
Design-1	-7.55	3	7	252.7	16.9	CASSDEPFDEQYF
Design-2	-3.79	4	0	376	20.3	CASSLLGPGEQYF
Design-3	-3.19	2	3	269.9	12.1	CASSDDASNEQYF
Design-4	-2.73	3	4	251.7	9.65	CASSDGAGDEQYF
Design-5	**-1.43**	**4**	**5**	**229.9**	**4.71**	**CASSDDAGDEQFF**
Design-6	**0.32**	**1**	**1**	**244.4**	**3.97**	**CASSDGRGNEQFF**
Design-7	0.78	0	0	277	8.86	CASSNGPYGEQYF
Design-8	0.88	4	2	262	13.1	CASSYRLGDEQYF
Design-9	1.18	5	2	220.2	11.1	CASSDGWPDEQFF
Design-10	1.26	3	2	214.9	10.3	CASSYRLGDEQYF

Supplementary Table 5 | Distribution of MHC Class and Alleles in the 35-target benchmark.

MHC Category	**Allele/Locus**	**Count**
HLA Class I	A	15
	B	9
	C	1
	E	2
HLA Class II	DP	2
	DQ	2
Mouse MHC	H-2 (Class I)	2
	I-A (Class II)	2
Total		**35**

3 Supplementary Note

3.1 Details of numbering adaptation

Algorithm 1: Unified Receptor Numbering Adaptation (Chothia Scheme)

Input: Raw amino acid sequence S

Input: Topology Constant Table $\mathscr{T}_{topo}$ (contains boundary definitions for Heavy/Light topologies)

Output: Standardized Chain object $\mathscr{C}_{final}$ with unified numbering

$(\mathscr{A}, \tau) \leftarrow \text{RunHMMAlignment}(S, \text{'chothia'})$;

if $\tau \in \{IgH, TCR\beta, TCR\delta\}$ **then**

  $\mathscr{B} \leftarrow \text{Lookup}(\mathscr{T}_{topo}, \text{Topology}_{heavy})$; ; `// Retrieves {CDR1, CDR2, CDR3} bounds for Heavy-like chains`

else

  $\mathscr{B} \leftarrow \text{Lookup}(\mathscr{T}_{topo}, \text{Topology}_{light})$; ; `// Retrieves {CDR1, CDR2, CDR3} bounds for Light-like chains`

end

$\mathscr{C}_{final} \leftarrow \emptyset$;

foreach *residue tuple* $(n, i, aa) \in \mathscr{A}$ **do**

  $\rho \leftarrow \text{DetermineRegion}(n, \mathscr{B})$;

  $p \leftarrow \text{CreatePosition}(n, i, \tau, \rho)$;

  $\mathscr{C}_{final}.\text{append}((p, aa))$;

end

return $\mathscr{C}_{final}$;

3.2 TCR full sequence reconstruction

To generate full T cell receptor (TCR) variable region sequences from available V(D)J gene annotations, we employed a standardized assembly pipeline. Germline amino acid sequences corresponding to V and J gene annotations were retrieved from the IMGT/Gene-DB repository. These sequences were precisely truncated at conserved recombination signal motifs and subsequently concatenated with the experimentally determined CDR3 hypervariable junctions to assemble the complete variable domains of both TCR α and β chains. This reconstruction strategy maintains high data fidelity by directly querying up-to-date reference databases and explicitly omits the constant regions, thereby focusing the input features on the variable domains critical for discriminating specific binding interactions.

3.3 Surface plasmon resonance protocol

Instrumentation and reagents SPR assays were conducted on a Biacore 8K system (Cytiva) using Series S Sensor Chip CM5. The running buffer used throughout the experiments for immobilization, binding, and dissociation contained 20 mM HEPES (pH 7.5), 150 mM NaCl, and 0.05% Tween-20.

Ligand Immobilization The pMHC ligands (specifically pHLA-A*11:01 for KRAS-G12V targets and pHLA-A*24:02 for SARS-CoV-2 targets) were immobilized using standard amine coupling chemistry. The CM5 chip surface was first activated with a 1:1 mixture of 0.4 M 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC) and 0.1 M N-hydroxysuccinimide (NHS). The ligands were diluted to a final concentration of 20 μg/mL in 10 mM sodium acetate buffer (pH 4.5) and injected over the activated surface at a flow rate of 10 μL/min until a response level of approximately 2000 RU was achieved. Residual activated groups were blocked by injecting 1 M ethanolamine-HCl (pH 8.5).

Kinetic analysis Binding kinetics were measured using a multi-cycle kinetics approach. TCR analytes were diluted in the running buffer to generate a series of five concentrations using a 2-fold serial dilution factor. Samples were injected from low to high concentration. For each cycle, the association phase was monitored for 120 s, followed by a dissociation phase of 600 s. Between cycles, the sensor surface was regenerated with a brief injection of 50 mM NaOH to remove any bound TCRs.

Data evaluation The resulting sensorgrams were analyzed using the Biacore Insight Evaluation Software. Equilibrium dissociation constants (K_D) were derived by fitting the data to a 1:1 Langmuir binding model based on the multi-cycle kinetics data.