

An in silico pipeline prioritizes candidate viral and bacterial mimics of hybrid insulin peptide autoantigens in type 1 diabetes

Liam Gershony

liam.gershony@gmail.com

Benaroya Research Institute <https://orcid.org/0009-0005-2387-5154>

Research Article

Keywords: type 1 diabetes, hybrid insulin peptides, molecular mimicry, HLA-DQ8, T-cell receptor, AlphaFold 3, TCRModel 2, SARS-CoV-2

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11341818/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

An in silico pipeline prioritizes candidate viral and bacterial mimics of hybrid insulin peptide autoantigens in type 1 diabetes

Liam Gershony*

Research Intern at the Benaroya Research Institute, Seattle, WA, United States

ARTICLE INFO

Keywords:

type 1 diabetes
hybrid insulin peptides
molecular mimicry
HLA-DQ8
T-cell receptor
AlphaFold 3
TCRModel2
SARS-CoV-2

ABSTRACT

Hybrid insulin peptide (HIP) fusion junctions are not encoded in the germline proteome and may be absent during thymic negative selection, allowing microbial mimics to encounter T cells that escaped central tolerance. We present a structure-first screening pipeline for HIP mimics in type 1 diabetes, dataset-scoped negative results for motif yield and backbone RMSD, and ten peptides prioritized for experimental testing. Five reference HIP–TCR–MHC systems guided motif construction. The Swiss-Prot screen targeted ten taxa; a contemporary reconstruction contains reviewed sequences for nine. Of 1,092 motif hits, 55 passed predicted HLA-binding and AlphaFold 3 confidence filters and were modelled with TCRModel2 across five stochastic seeds. A composition-preserving shuffled-proteome control showed no aggregate motif-yield enrichment in the contemporary reconstruction (observed/null ratio 0.97). Candidate backbone RMSD had median 0.315 Å and interquartile range 0.274–0.389 Å; the three controls lay inside the candidate distribution at the 21.8th, 61.8th, and 81.8th percentiles. No single RMSD cutoff perfectly separated these sampled groups. A three-metric native-HIP gate using TCR–pMHC ipTM, interface–pLDDT, and ΔG retained 13 candidates; exclusion of three animal-associated Rotavirus A peptides left ten nominees. The nominated set was unchanged by dropping the redundant K_d criterion but was sensitive to the acceptance thresholds. Supporting comparisons found higher ipTM for 48 of 55 candidates, robust to nonparametric testing and multiple-comparison correction, but the candidate-only AlphaFold 3 pre-filter confounds that separation. RRNVATLQAENVTVG from SARS-CoV-2 and SDLLKSVDSEEVRE from human rhinovirus 3 are co-priorities undergoing experimental testing; confidence and directional agreement favor the former, whereas interface energetics favor the latter. Their relative ordering is not statistically established. The pipeline supplies testable peptide–TCR–MHC combinations and documents the limits of motif-yield and backbone-similarity filters in these analyses.

1. Introduction

Type 1 diabetes (T1D) is a chronic autoimmune disease in which CD4⁺ and CD8⁺ T cells destroy the insulin-producing β cells of the pancreatic islets, resulting in lifelong dependence on exogenous insulin. An estimated 43% of the 513,000 incident cases projected for 2025 occur in individuals younger than 20 years, and current estimates place global prevalence at 9.5 million (95% CI 9.0 to 10.0 million), with a projection of 14.7 million by 2040 [1]. Despite decades of investigation, the events that initiate islet autoimmunity remain unresolved.

Genetic susceptibility to T1D is strongly influenced by the HLA class II region, particularly the DR3-DQ2 and DR4-DQ8 haplotypes [2, 3]. Because class II molecules present peptides to CD4⁺ T cells, this association implicates specific presented epitopes in disease risk.

Two lines of evidence motivated this study. The first is the discovery of hybrid insulin peptides (HIPs).

HIPs form within β cells when proteolytic fragments of proinsulin become covalently joined to fragments of other islet proteins. Cathepsin D contributes to this reaction within the secretory granule [4]. The resulting junctional sequences are not encoded in the germline genome and may therefore be absent during thymic negative selection. This could allow HIP-reactive T cells to escape central tolerance. Baker and colleagues established that HIPs are autoantigens in human T1D by identifying HIP-reactive CD4⁺ T-cell clones from patient islets and characterizing their TCRs and restricting HLA alleles [5]. Related work identified proinsulin-reactive CD4⁺ T cells directly from human islets [18]. Later structural studies resolved how a T1D-associated TCR recognizes a HIP presented by DQ8 [6, 7].

The second line of evidence is the molecular-mimicry hypothesis. In this model, infection introduces a peptide that structurally resembles a self-epitope, priming T cells that may later cross-react with host tissue. Enteroviruses, particularly coxsackievirus B, have repeatedly been

¹Corresponding author

liam.gershony@gmail.com (Liam Gershony)

ORCID(s): 0009-0005-2387-5154 (Liam Gershony)

associated with islet autoimmunity [8, 9]. Reports of increased pediatric T1D during and after the COVID-19 pandemic further renewed interest in environmental triggers, although a mechanistic link has not been established.

These two lines of evidence have not been integrated systematically. Previous mimicry searches have focused on conventional islet autoantigens such as GAD65, IA-2, and native proinsulin. HIPs are especially relevant targets because their junctions may be absent from the thymic proteome. A microbial peptide resembling a HIP junction could therefore encounter reactive T cells that were not removed during central tolerance.

A second methodological challenge concerns how mimicry should be assessed. Conventional approaches often rely on linear sequence similarity, such as BLAST. A TCR instead recognizes the three-dimensional surface formed by solvent-exposed peptide side chains within the MHC groove. Peptides with little sequence identity can still form similar surfaces if their anchor and contact residues share physicochemical properties. Conversely, high sequence identity may be less informative when differences occur at TCR-facing positions.

This study therefore used a structure-first approach. We modelled reference TCR–pMHC–HIP complexes, identified residues contacting the TCR and MHC, and translated those contacts into physicochemical motifs for searching microbial proteomes. Candidates were filtered for predicted HLA binding, remodelled with the relevant TCR, and compared with the native HIP using structural-confidence and energetic criteria. The structural analysis was restricted to HIPs with characterized TCR sequences. Its output is a prioritized set of microbial peptides for experimental testing as potential mimics of T1D autoantigens.

The structural screening and nomination workflow is summarized in Figure 1.

2. Methods

All archive files and scripts cited in this manuscript refer to release v1.2 at commit 0ddf984 (see Data availability).

2.1. TCR sequence reconstruction

Published reports often identify T1D-reactive TCR clones by V gene, J gene, and CDR3 sequence rather than by complete sequence. Such receptor sequences may also support the development of TCR-based biomarkers for T1D [19]. TCR α and β chains were reconstructed with Stitchr [10]. Stitchr retrieves the specified human variable and joining alleles from IMGT/GENE-DB and assembles them around the supplied CDR3. Commands followed

the form: `stitchr -v '[V gene]' -j '[J gene]' -cdr3 '[CDR3 AA sequence]' -nl`. The `-nl` flag suppresses the leader sequence. The TCR β D gene is integral to the CDR3 and was not supplied separately. Eight runs generated the raw α/β sequences for the reference clones.

Constant regions were removed before modelling because they do not contribute to the docking interface. Trimming began at these identifiers: TRAC, DIQN; TRBC1, EDLN or EDLK. After trimming, α chains ranged from 105 to 111 residues and β chains from 111 to 113 residues. HLA class II α - and β -chain sequences were retrieved from UniProt and RCSB PDB, then trimmed to remove signal peptides and transmembrane tails.

2.2. Reference and candidate complex modelling

Trimmed TCR α , TCR β , MHC II α , MHC II β , and HIP sequences were assembled and submitted to AlphaFold 3 [14]. Model quality was assessed using pLDDT, pTM, and ipTM. The model with the highest ipTM was retained as the reference complex for each system.

Candidate complexes were modelled with AlphaFold 3 using the same chain composition and model-selection procedure as the reference complexes. Candidates were retained only when the resulting complex had ipTM > 0.75. This candidate-level confidence filter was applied after the NetMHCIIpan-4.2 binding filter and before TCRModel2 modelling.

2.3. Pharmacophore definition

Reference complexes were loaded into PyMOL [15]. For all five complexes, peptide atoms within 3.0 Å of TCR atoms were selected. Peptide residues with side chains within this contact envelope were labelled TCR-contacting. The equivalent procedure was used for MHC contacts. Assignments were mapped onto the standard nine-residue class II binding core (Table 3).

2.4. Motif construction and proteome screening

Contact assignments were generalized into physicochemical classes (Table 4) and expressed as ScanProsite patterns [11] (Table 5). ScanProsite searched UniProtKB/Swiss-Prot using the NCBI taxonomy identifiers in Table 10. The motifs were applied to the taxa in Table 6. Because screening was performed at the level of the listed taxonomic groups, some hits came from strains associated with non-human hosts. The human-association criterion was applied during candidate nomination rather than during the initial screen. Table 10 lists the searched taxa and their human-association classifications. ScanProsite was used because the

objective was motif homology at structurally defined positions rather than overall sequence similarity.

Ten taxa were intended search targets, not ten demonstrably represented sequence sets. The contemporary UniProt coverage reconstruction used for this revision contains reviewed entries for nine targets (six viral groups and three bacterial species); *B. dorei* (TaxID 357276) contributes none, although unreviewed TrEMBL entries exist. The retained methods specify Swiss-Prot, and no retained record establishes that TrEMBL was included. Because the original sequence snapshot and identifier-retrieval date were not retained, present-day coverage cannot establish the exact historical search contents. Table 10 distinguishes the recorded query identifiers from current taxonomy.

The motif-null control scanned current reviewed sequences for the Table 10 queries, substituting Orthoherpesviridae (3044472) for Herpesviridae (10292). Python regular-expression lookaheads counted overlapping matches of the three structural-screen motifs at every eligible start. Forty independent within-protein shuffles preserved each sequence's composition and length (pseudorandom seed 0; N_SHUFFLE=40). The empirical one-sided enrichment probability was $(b + 1)/41$, where b counts null totals at least as large as the observed total. A scrambled-motif check used 12 permutations per motif of its constrained-class and wildcard blocks, preserving motif length and block composition. This scan-only control required no structural modelling (scripts/motif_null_control.py; results in Section 3.12).

2.5. MHC-binding prediction

Motif hits in the structural screen were submitted to NetMHCIIpan-4.2 [20], which incorporates expanded HLA-DQ immunopeptidomics data. Predictions used the HLA allele restricting the source system. Peptides with EL rank below 5% were retained for complex modelling; binding affinity and affinity rank were recorded when available.

An additional motif and NetMHCIIpan-4.2 screen considered twelve HIPs lacking characterized TCR sequences, using standard anchor/contact assignments rather than observed complexes, and returned 474 peptides with EL rank below 1%; the peptide list was not retained. This screen stopped at predicted MHC binding: no structural modelling, control comparison, matched background analysis, or nominations followed. Recorded summaries noted human rhinovirus 3, *K. oxytoca*, and *P. distasonis* in both screens, a descriptive overlap

indicating that their organism lists were not entirely screen-specific.

2.6. TCR-pMHC complex modelling

Surviving candidates were modelled with TCRModel2 [12], followed by Amber relaxation. Five models were generated per candidate with different random seeds to assess variation between modelling runs. Recorded metrics were TCR-pMHC ipTM and interface-pLDDT, per seed. Overall model pLDDT, pTM, and model-confidence scores were not retained in the results workbook for the 55 candidates (Supplementary Table S2). TCR-pMHC ipTM was the primary metric because it isolates the predicted TCR-pMHC interface. Interface-pLDDT was retained as a secondary measure of local confidence at that interface.

Prioritization assumed that higher TCR-pMHC ipTM corresponds to a higher likelihood of genuine TCR recognition. This assumption was not tested here: no cognate/non-cognate recognition benchmark was run. ipTM measures predicted interface confidence, not a calibrated probability of recognition. AlphaFold-Multimer confidence scoring has shown limited binding-discrimination performance in an external TCR-pMHC benchmark [21]; those results motivate caution but do not provide a performance estimate for the present HIP systems.

2.7. Interface energetics

PDB files from the five TCRModel2 structures were submitted to PRODIGY [13] to estimate Gibbs free energy (ΔG) for the interface between the TCR α/β chains and the pMHC complex. A more negative ΔG was interpreted as a more stable predicted complex. Predicted K_d was retained descriptively: at fixed temperature, $K_d = \exp(\Delta G/RT)$, so it is not an independent energetic measurement or gate criterion.

2.8. Structural homology

Backbone RMSD between each candidate and its native-HIP complex was calculated in PyMOL with rms_cur after superposition of the conserved MHC domains. The model with the highest TCR-pMHC ipTM was selected for each candidate. RMSD below 1.0 Å was prespecified as indicating close reproduction of peptide backbone geometry. The retained RMSD records were also summarized by range, quartiles, and control percentiles within the pooled 55-candidate distribution. Percentiles denote the fraction of candidate RMSDs strictly below the value being evaluated; no group-comparison test was performed with three controls.

2.9. Statistical analysis

2.9.1. Supporting negative-control comparisons

Each candidate was compared only with the designated negative control for the same TCR–MHC system. Controls were not pooled across systems. The HIP6/A2.11/DQ8 control, RRDRKPPDRDRKRR, was a synthetic peptide rather than a microbial sequence. The HIP11/8.E3/DQ8 and HIP11/8.E3/DQ8-trans controls used the same 14-residue Epstein–Barr virus peptide, KDVDAAVDAEVVQF, from strains B95-8 and AG876, respectively. Each was tested against the corresponding MHC molecule. Thus, the three system controls comprised one synthetic decoy and one viral sequence evaluated in two distinct MHC contexts. They were not three independent microbial decoys.

For each candidate, five seed-specific TCR–pMHC ipTM values were compared with the five seed-specific values of the designated same-system negative control using an unpaired Welch’s t -test at $\alpha = 0.05$:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (1)$$

Candidates with $p < 0.05$ were counted as having higher ipTM than their control. This supporting comparison was not used as a prerequisite for nomination.

Sensitivity analyses used the existing five-seed TCR–pMHC ipTM records only, without new modelling. Candidate–control comparisons were also assessed by two-sided Mann–Whitney U tests. Benjamini–Hochberg false-discovery-rate correction was applied across all 55 comparisons, separately for the Welch and Mann–Whitney test families, with a 0.05 threshold. Cliff’s delta summarized rank separation, with positive values favoring the candidate; bootstrap intervals summarized resampling variation. These analyses concern separation from the same-system control, not improvement over the native HIP.

2.9.2. Comparison with the native HIP

For each native HIP, 95% confidence intervals across five stochastic seeds were constructed for TCR–pMHC ipTM, interface-pLDDT, and ΔG :

$$CI = \bar{x} \pm t^{crit} \frac{s}{\sqrt{n}} \quad (2)$$

The three-metric gate required each candidate’s mean ipTM and interface-pLDDT to be at or above the corresponding native-HIP lower interval bound and its mean ΔG to be at or below the upper bound. Confidence intervals used $t^{crit} = 2.776$ with $n = 5$ seeds. The gate was applied to all 55 modelled candidates without requiring

significance against a negative control; human association was assessed separately at nomination.

K_d is a deterministic transform of ΔG and is not an additional criterion. Existing-data sensitivity analysis compared this gate with the former four-metric gate and with each two-metric confidence/energy pair using `scripts/gate_dimensionality_and_audit.py`. The final ten nominees were unchanged on dropping K_d , although the number passing before the human-association filter increased (Section 3.7); this is a revision of the criterion, not mathematical equivalence of the two gates. Means and symmetric intervals calculated after a nonlinear transformation need not reproduce the gate in the original units.

Threshold sensitivity retained all three metrics and the same host-association filter but replaced the native-HIP intervals with fixed tolerances of 0.5, 1, 2, 3, or 5% of the absolute native mean. A confidence metric passed at or above $\mu_{\text{native}} - f|\mu_{\text{native}}|$, and ΔG passed at or below $\mu_{\text{native}} + f|\mu_{\text{native}}|$, allowing better values without restriction. The absolute value ensures the tolerance for negative ΔG is in the less favorable direction. These post hoc comparisons assess sensitivity of the prespecified CI-based acceptance rule, not replacement nomination rules; `scripts/rmsd_distribution_and_sensitivity.py` implements the analysis using system-specific records.

2.9.3. Metric dependence and descriptive prioritization

Existing workbook records were used to assess dependence among TCR–pMHC ipTM, interface-pLDDT, and ΔG without further modelling. Pearson and Spearman correlations were calculated across the 55 candidate means pooled across systems; Pearson correlations and effective dimensionality were also evaluated separately within each TCR–MHC system. Correlations between ipTM and interface-pLDDT across the five seeds within each candidate were calculated when both metrics varied. Effective dimensionality was summarized from the three-metric correlation matrix by its eigenvalues λ_j , the first-component variance fraction, and the participation ratio $(\sum_j \lambda_j)^2 / \sum_j \lambda_j^2$. These are descriptive measures of covariance, not tests of independence or a number of independent trials for a sign test. The archive script `scripts/metric_independence.py` implements these analyses.

For the secondary composite (Supplementary Table S1), each prioritization metric was expressed as percentage change from the native HIP, with positive values indicating improvement, and divided by the sample SD of

those changes among the nominated candidates in that system (four HIP6/A2.11/DQ8 and six HIP11/8.E3/DQ8-trans nominees, not the full pools of 28 and 20). The numerator retained the native-HIP baseline rather than subtracting the cohort mean. The three standardized changes were averaged with equal, a priori weights. Scores describe relative position only within each selected group and did not determine nomination or establish an ordering of the experimental co-priorities.

2.10. Software and databases

The analysis used AlphaFold 3 [14] and TCRModel2 [12] for structural modelling, PRODIGY [13] for interface energetics, PyMOL [15] for structural analysis, Stitchr [10] for TCR reconstruction, ScanProsite/PROSITE [11] for motif screening, and NetMHCIIpan-4.2 [20] for MHC-binding prediction. Sequence and annotation resources included UniProt [16], IMGT/GENE-DB, RCSB PDB, and IEDB [17].

3. Results and discussion

3.1. Reference complexes were reconstructed with high confidence

Five TCR–HIP–MHC systems with experimentally validated components were assembled from the published literature (Tables 1 and 2). The reports provided the V gene, J gene, and CDR3 sequence for each clone. Full-length α - and β -chain sequences were reconstructed with Stitchr [10]. Constant regions and leader sequences were removed before modelling (Section 2.1).

All five AlphaFold 3 complexes had ipTM values of 0.85–0.91 and pTM values of 0.87–0.91. Each exceeded the predefined ipTM threshold of 0.75 for a confident predicted interface. After constant regions were removed, reconstructed TCR α chains ranged from 105 to 111 residues and β chains from 111 to 113 residues.

The five reference-complex models are shown in Figure 2.

3.2. Pharmacophore definition from modelled contacts

For each reference complex, peptide atoms within 3.0 Å of TCR atoms were identified in PyMOL. Residues whose side chains fell within that contact envelope were labelled TCR-facing. The same procedure was applied to MHC contacts. Mapping these assignments onto the canonical nine-residue class II binding core produced the contact map in Table 3.

Because TCRs recognize shape and charge rather than residue identity alone, contact positions were generalized into physicochemical classes instead of fixed amino acids

(Table 4). The screen therefore searched for chemical equivalence at structurally defined positions rather than overall sequence identity.

3.3. Motif screening returns 1,092 candidates

Pharmacophores were translated into three structural-screen ScanProsite patterns (Table 5). The search targeted ten T1D-associated taxa and returned 1,092 peptides; the contemporary Swiss-Prot reconstruction has sequence coverage for nine (Tables 6 and 10). The two additional binding-screen motifs are listed for completeness and described in Section 2.5.

3.4. MHC binding and complex-confidence filtering yields 55 candidates

All 1,092 peptides were evaluated with NetMHCIIpan-4.2 using the HLA allele restricting their source system. Peptides with an eluted-ligand (EL) rank below 5% were retained as predicted binders; binding was not measured experimentally. Surviving peptides were also required to pass the AlphaFold 3 complex-confidence filter (ipTM > 0.75). Together, these filters retained 55 candidates: 28 for HIP6/A2.11/DQ8, 7 for HIP11/8.E3/DQ8, and 20 for HIP11/8.E3/DQ8-trans.

3.5. Negative controls lie within the candidate backbone-RMSD distribution

Each of the 55 candidates was modelled with TCRModel2 using five stochastic seeds. The model with the highest TCR–pMHC ipTM was used for backbone RMSD comparison with its corresponding native-HIP complex after alignment of the conserved MHC domains. Candidate RMSDs ranged from 0.198 to 1.896 Å, with first quartile 0.274 Å, median 0.315 Å, and third quartile 0.389 Å. Excluding the single 1.896 Å outlier gave median 0.311 Å and maximum 0.755 Å.

The KDVDAAVDAEVVQF controls in HIP11/8.E3/DQ8 and HIP11/8.E3/DQ8-trans had RMSDs of 0.270 and 0.356 Å, respectively, at the 21.8th and 61.8th percentiles of the pooled candidate distribution. The synthetic HIP6/A2.11/DQ8 control, RRDRKPPPPDRDRKRR, had RMSD 0.458 Å (81.8th percentile). Thus, controls were interspersed with candidates rather than confined to a distributional extreme: the cis control had lower RMSD than 78.2% of candidates. RRNVATLQAENVVG had RMSD equal to the candidate median (49.1% of candidate values were strictly lower). These distributions cannot be perfectly separated by a single RMSD cutoff; this descriptive result does not depend on choosing 1.0 Å.

At the prespecified 1.0 Å cutoff, 54 of 55 candidates and all three controls passed. The outlier was

GHGEINDSTTVEPIL from Rotavirus A strain RVA/Human/US/D/1974; it remained in subsequent analyses and failed the native-HIP gate independently of RMSD. No candidate was excluded by RMSD. Controls were not subjected to the ipTM pre-filter, so their close alignment cannot be attributed to selection for high ipTM. The analysis concerns these models and three control complexes built from two unique sequences, not a general recognition benchmark; no inferential test was performed. The complete values are in `results/rmsd_distribution.csv`.

3.6. Supporting candidate–control ipTM comparisons

For each candidate, five TCR–pMHC ipTM seed values were compared with the designated same-system negative control using an unpaired Welch’s t -test ($\alpha = 0.05$). Forty-eight of 55 candidates (87%) had significantly higher scores; this count summarizes a supporting comparison, with interpretation subject to Section 3.12. Nomination instead uses the three-metric native-HIP gate applied to all 55 candidates (Section 3.7). The candidate KHVANSVFAEIIKT (human herpesvirus 8 type P, isolate GK18; HIP11/8.E3/DQ8-trans) had an originally stored workbook p -value of 0.0658 and a recomputed two-sided Welch p -value of 0.0404. It is included in the 48 candidates passing the recomputed test. It did not meet the native-HIP gate (Section 3.7): its mean TCR–pMHC ipTM was 0.36, compared with 0.822 for native HIP11-trans, so the final nominated set was unchanged. RRNVATLQAENVGTG had mean ipTM 0.84 ± 0.02 , compared with 0.26 ± 0.03 for the EBV AG876 control (95% confidence intervals); interpretation of this comparison is discussed in Section 3.12. The synthetic HIP6/A2.11/DQ8 control had mean ipTM 0.682, illustrating that a designed decoy can still produce a moderately confident interface.

Figure 3 shows the candidate and negative-control seed distributions relative to each system’s native-HIP mean and 95% confidence interval.

The existing-data sensitivity analysis identified 48 of 55 candidates with significantly higher ipTM under Mann–Whitney U testing, matching the Welch count. After Benjamini–Hochberg correction across all 55 comparisons, 48 survived in each test family. Forty-five candidates showed complete rank separation from their control (Cliff’s delta = +1.0, bootstrap interval [1.0, 1.0]), and none scored below its control. The degenerate bootstrap intervals reflect complete separation in the observed samples; they do not establish a precisely known population effect. With five seeds per group, the minimum attainable two-sided exact permutation p -value

under perfect separation is $2/C(10, 5) = 0.0079$. This resolution limit applies to an exact rank test; asymptotic approximations and tie handling can yield different numerical values. At this sample size, significance largely reflects complete or near-complete rank separation rather than precise estimation of an effect.

This robustness applies to the candidate–control comparison only. For example, RRNVATLQAENVGTG and its trans-system control differ substantially in mean ipTM (approximately 0.84 versus 0.26). The much smaller margins over native HIP11, which drive follow-up prioritization, are not individually significant (Section 3.9). These sensitivity analyses do not strengthen the native-HIP margins; see Section 3.12 for their scope.

The motif screen also returned hits from animal-associated Rotavirus A strains. Three had higher mean TCR–pMHC ipTM than every nominated HIP6/A2.11/DQ8 candidate: RHGEVNDSTTVEPVL (RVA/Pig/US/Gottfried/1983, 0.876), GPGEVNDSTTVEPIL (RVA/Equine/UK/H2/1976, 0.874), and GPGEVNDSTTVEPVL (RVA/Equine/US/FI-14/1980, 0.870), compared with 0.850 for the highest-scoring nominee. These peptides were retained through modelling but excluded from nomination because of their host association, not their ipTM score. All three passed the revised three-metric native-HIP gate; two passed the former four-metric gate. Thus, a high ipTM score did not track the human-association criterion used for nomination. Table 8 displays these three gate-passers as excluded rows, alongside the four nominees; Table 10 gives their query classifications.

3.7. Thirteen candidates pass the three-metric native-HIP gate; ten are nominated

Nomination used the native-HIP intervals for TCR–pMHC ipTM, interface-pLDDT, and ΔG (Section 2.9.2). Applied to all 55 candidates, this gate retained 13; the separate human-association criterion excluded the three animal-associated Rotavirus A peptides, leaving four HIP6/A2.11/DQ8 and six HIP11/8.E3/DQ8-trans nominees, with none in HIP11/8.E3/DQ8. Their organism-level distribution is summarized in Table 7. Passing the gate describes compatibility with the reference intervals, not a statistical equivalence test.

The former four-metric gate retained 12 candidates and the identical ten nominees. Removing the redundant K_d criterion therefore changes the intermediate count but not the final nominated set. Reducing further to ipTM plus ΔG yields 17 passers and 14 nominees, while interface-pLDDT plus ΔG yields 19 passers and 16 nominees; neither two-metric gate preserves the ten-peptide set. The

three-metric gate is retained to preserve both confidence measures without treating transformed energetics as an additional measurement. Its nominated set also remains unchanged if the supporting Welch comparison is imposed first: the seven Welch-negative candidates have ipTM means of 0.370–0.676, below their corresponding native-HIP lower bounds of 0.773–0.835.

The nominated set is sensitive to the acceptance region. Relative ipTM CI half-widths were 1.31% for HIP6/A2.11/DQ8, 4.76% for HIP11/8.E3/DQ8, and 5.89% for HIP11/8.E3/DQ8-trans, a 4.5-fold range. Replacing all three metric intervals with fixed relative tolerances of 0.5, 1, 2, 3, and 5% yielded respectively 5, 5, 6, 11, and 20 passers, all human-associated, versus 13 passers and ten nominees under the CI rule. None of these tested fixed tolerances reproduced the ten-peptide set: only five of the ten survived the tightest two, whereas the 5% set contained all ten plus ten others. The trans system's wider reference intervals therefore admit nominees excluded by tighter fixed tolerances. These sensitivity results do not replace the CI-derived nominees or establish another optimal threshold.

Among the ten nominees, RRNVATLQAENVGTG was the only candidate in the HIP11/8.E3/DQ8-trans system to exceed its native HIP on all three prioritization metrics. SESEVESDTALESEV was the only candidate to do so in HIP6/A2.11/DQ8. These directional comparisons describe the reported means, not statistically established superiority.

3.8. Metric dependence and prioritization

Within-system metric dependence, rather than pooled estimates, informs prioritization (Table 11). In HIP6/A2.11/DQ8, HIP11/8.E3/DQ8-trans, and HIP11/8.E3/DQ8, respectively, ipTM/interface-pLDDT correlations were 0.648, 0.905, and 0.972, versus 0.560 pooled; ipTM/ ΔG correlations were +0.086, -0.240, and -0.457, versus -0.580 pooled; interface-pLDDT/ ΔG correlations were -0.063, -0.251, and -0.452, versus +0.076 pooled. Pooling therefore masks stronger within-system confidence-metric associations while overstating the consistency of the ipTM/ ΔG association across systems. The trans-system effective dimensionality was 1.84 of three: directional agreement is not three independent confirmations. Full analysis output is available through `scripts/metric_independence.py` in the public archive. The descriptive composite is reported in Supplementary Table S1 and did not determine nomination.

3.9. Co-priorities for experimental follow-up: RRNVATLQAENVGTG and SDLLKSVDSEEVRE

RRNVATLQAENVGTG (SARS-CoV-2) and SDLLKSVDSEEVRE (human rhinovirus 3) are co-priorities among the six HIP11/8.E3/DQ8-trans nominees, and both are undergoing testing at Benaroya (Section 3.13). RRNVATLQAENVGTG has higher mean TCR-pMHC ipTM (0.842 versus 0.826) and favorable mean changes relative to native HIP11 on all three metrics. SDLLKSVDSEEVRE has more favorable interface energetics (ΔG -9.92 versus -9.46 kcal mol⁻¹), with a +9.251% change relative to native HIP11 (Table 9). These are complementary descriptive reasons for testing both, not a statistically established ordering. RRNVATLQAENVGTG's native-HIP margins remain individually nonsignificant ($p = 0.33$ for ipTM, 0.71 for interface-pLDDT, and 0.067 for ΔG).

If three independent metric signs are equally likely to be favorable or unfavorable, agreement in the favorable direction has probability $0.5^3 = 0.125$; a two-independent-axis illustration gives $0.5^2 = 0.25$. Across six nominees, the corresponding probabilities of at least one full agreement are $1 - (1 - p)^6 = 0.55$ and 0.82, assuming candidates are also independent. The ten nominees' favorable-sign rates are 0.30, 0.40, and 0.70 for ipTM, interface-pLDDT, and ΔG ; using their mean (0.467) as a common independent sign probability gives 0.102 per candidate and 0.47 across six. None is a calibrated selection-adjusted p -value: a pooled rate near one-half establishes neither independence nor absence of gate-induced bias. Directional agreement alone does not justify preferring one peptide.

RRNVATLQAENVGTG's five ipTM values were 0.84, 0.85, 0.85, 0.82, and 0.85 (SD 0.0130), compared with 0.80, 0.86, 0.82, 0.85, and 0.80 (SD 0.0279) for trans-system SDLLKSVDSEEVRE. Using unrounded seed SDs, the variance ratio is 4.588, below the upper critical value of 9.60 for a two-sided 5% $F(4, 4)$ test ($p = 0.169$ under independent normal sampling). The difference is unresolved at five seeds and motivates the expanded sampling in Section 3.13. The present comparison uses recomputed per-seed SDs: the stored trans-system SD for SDLLKSVDSEEVRE was 0.826 (its mean), which would imply an approximately 64-fold SD ratio rather than the corrected 2.1-fold ratio.

The secondary descriptive composite favors SDLLKSVDSEEVRE over RRNVATLQAENVGTG by 0.024 standardized units, less than the effect of changing the standardization cohort by one member (Supplementary Table S1). It did not determine

nomination and is not used to order these co-priorities; both peptides are already being tested.

3.10. Interpretation

The contribution is a quantified test of a plausible but uninformative screening feature: the controls lie inside the candidate backbone-RMSD distribution, and RRNVATLQAENVGTG lies at its median (Section 3.5). Class II MHC binding constrains peptides toward an extended, polyproline-II-like backbone through conserved interactions [22], so shared backbone geometry is structurally expected rather than surprising. Documenting its lack of discrimination in this dataset helps prevent low RMSD from being treated as a stand-alone mimicry filter.

The completed composition-preserving control provides a second empirical negative result: aggregate motif yield did not exceed the shuffled background within the reconstruction (Section 3.12). Contact-derived motifs are linear class-based filters, not complete TCR-facing surface comparisons. The control supports their role as a tractability device rather than a demonstrated enrichment step. Alongside the RMSD overlap, it documents a distinct screening limitation, without assessing downstream-filter performance.

The practical output remains ten microbial peptides for experimental follow-up, nominated through downstream HLA-binding, complex-confidence, native-HIP, and human-association filters. In the sequence-based reconstruction, Orthoherpesviridae and Enterovirus contribute 1,093,964 residues (72.6% of the total) and account for three nominees (HHV-1, HHV-5, and HRV-3; Table 7); organism counts must be read against unequal search volumes.

3.11. The SARS-CoV-2 candidate in context

Reports of increased pediatric T1D during and after the COVID-19 pandemic have prompted interest in a possible role for SARS-CoV-2. Proposed mechanisms include bystander inflammation, direct β -cell infection, and immune dysregulation. The present result offers a narrower, testable hypothesis: one viral peptide is predicted to bind a high-risk HLA-DQ molecule with an EL rank of 0.03% and to reproduce key structural and energetic features of a validated HIP autoantigen when modelled with a patient-derived TCR.

RRNVATLQAENVGTG is a priority for follow-up, not an experimentally established mimic or evidence that the peptide causes disease. Its three margins over native HIP11 were not individually significant (Section 3.9). A causal interpretation would require a link between infection and seroconversion, T-cell reactivity in patient

samples, and evidence that the peptide is naturally processed and presented in vivo. At present, the result provides a specific target for experimental testing, without establishing superiority over the native HIP or enrichment of mimics in SARS-CoV-2.

3.12. Limitations

Candidates and controls underwent different selection. Candidates passed an AlphaFold 3 ipTM > 0.75 pre-filter that the negative controls never faced. The 48/55 comparison therefore contrasts a pool selected on a quantity correlated with the tested score against unscreened controls. Its effect size is confounded and cannot establish discrimination. This comparison is not part of the nomination path: applying the three-metric gate to all 55 candidates yields the same 13 passers and ten nominees as applying it only to the 48 Welch-positive candidates (Section 3.7). Mann–Whitney testing and Benjamini–Hochberg correction address test choice and multiplicity, not this selection mechanism. Identical pre-filtering and matched decoys are required before interpreting the separation as a discrimination result (Section 3.13).

Recognition and biological activity remain untested. Higher ipTM is assumed to imply recognition, but no cognate/non-cognate benchmark was run here; an external AlphaFold-Multimer benchmark reported limited discrimination [21]. Without experimental validation, predicted binding, confidence, and energy do not establish natural processing, presentation, or T-cell activation. The trans-encoded DQ8 restriction element also requires validation, because trans-only variants may contribute little to measured HLA-DQ immunopeptidomes [20].

The motif yield does not exceed a composition-matched null. The reconstruction contained 3,039 reviewed sequences and 1,507,130 residues; *B. dorei* contributed none. Motifs 1, 2, and 5 returned 225, 619, and 174 overlapping matches (total 1,018). Across 40 shuffles, the null total had mean 1,054, SD 38, and range 982–1,136; observed/null was 0.97 ($z = -0.95$; one-sided empirical $p = 0.85$). The comparison is real versus shuffled sequences within this reconstruction: its yield is of similar scale to the original 1,092 hits but does not recover the original snapshot. Scrambled-block motifs showed no consistent yield advantage for the original ordering. This empirical control supersedes the approximate background estimate and detects no aggregate enrichment beyond per-protein composition; it does not test downstream filters. Aggregate results are in `results/motif_null_control.json`; the script also prints the scrambled-motif results.

Statistical scope is limited by seed count, metric dependence, and system differences. Five seeds measure model variation, not biological replication, and normality of bounded ipTM was not established. Candidate–control sensitivity tests do not validate nonsignificant native-HIP margins. Confidence metrics are correlated (Section 3.8), and K_d adds no independent evidence to ΔG . Threshold sensitivity (Section 3.7) and unequal candidate-pool sizes prevent interpreting the 4/0/6 nomination split as comparable success rates. Three control complexes represent only two unique sequences, including a highly charged synthetic decoy with mean ipTM 0.682, so the false-positive rate cannot be estimated.

Per-seed score provenance remains incomplete. Original model structures have been partially recovered, but per-seed TCR–pMHC ipTM score files and PRODIGY outputs remain unavailable. Structural checks can corroborate interface-pLDDT, not exactly reproduce its curated column; incomplete model coverage does not restore the per-seed series (Data availability; Supplementary Table S2).

3.13. Future directions

Three modelling tasks remain before journal submission: a cognate/non-cognate recognition benchmark; applying the identical AlphaFold 3 ipTM > 0.75 pre-filter to controls and adding 30–50 decoys from the same taxa matched on ipTM or NetMHCIIpan EL rank; and rerunning the ten nominees with 20–25 seeds each and all per-job outputs retained. Partial structural recovery does not substitute for these retained-output reruns.

Experimental testing of RRNVATLQAENV TG and SDLLKSVDSEEVRE is underway at the Benaroya Research Institute using the 8.E3 TCR and DQ8-expressing lentiviral cell lines. No results from that experimental work are included here.

4. Conclusion

This structure-informed pipeline provides ten microbial peptides for experimental follow-up and two executed negative controls: the reconstructed motif yield did not exceed a composition-matched null (Section 3.12), and candidate and control backbone geometries overlapped. Controls lie within the candidate RMSD distribution, with no single cutoff perfectly separating the sampled groups (Section 3.5). The three-metric native-HIP gate retained 13 candidates; the separate human-association filter left the same ten nominees as the former four-metric gate, without counting K_d as an independent criterion. The nominated set does not depend on the supporting candidate–control test.

RRNVATLQAENV TG and SDLLKSVDSEEVRE are co-priorities undergoing experimental testing at Benaroya, without a statistically established ordering. Their prioritization supplies specific peptide–TCR–MHC combinations to test, while the motif-null and RMSD results document distinct limits of the upstream screening features. The limitations in Section 3.12 and deferred work in Section 3.13 define the next steps from computational nomination to assessment of recognition and function.

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Declaration of competing interest

The author has no relevant financial or non-financial interests to disclose.

CRedit authorship contribution statement

Liam Gershony is the sole author and performed conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, and writing.

Ethical approval

Not applicable. This study involved no human participants, animal subjects, or patient-derived material.

Acknowledgements

The author thanks James M. Heather (Massachusetts General Hospital Cancer Center and Department of Medicine, Harvard Medical School) for supervision and methodological guidance.

Data availability

The results workbook, provenance-tagged per-seed records, per-peptide summaries, reference-complex metrics, analysis scripts, and regenerated results are publicly available in release v1.2 of the reproducibility archive. Supplementary Table S2 summarizes the provenance inventory. The archived `run_all.sh` orchestrates the core recomputations; metric dependence, gate dimensionality/workbook auditing, and RMSD distribution/threshold sensitivity are provided as separate analysis scripts, together with the completed motif-null control (`scripts/motif_null_control.py`; results in `results/motif_null_control.json`). The archive version underlying this preprint is tagged as release v1.2 (commit

Oddf9843591926c9dca4801d5e32f228fdab80a7), identifying the fixed snapshot of the data and code reported here. A DOI-registered deposit will accompany journal submission; for this preprint, the tagged commit serves as the citable reference.

Structural output from the original TCRModel2 runs has been partially recovered: 175 ranked_*.pdb downloads yielded 113 unique structures after removing 58 exact duplicates. Structures were mapped by modelled peptide core and disambiguated by MHC α allele; recovered_candidate_models/manifest.csv records the assignments and file hashes. Coverage spans 51 of 56 distinct workbook peptide sequences, including native-HIP and control records. Six sequences have five recovered models and 27 have one, so the collection does not reconstitute the per-seed series. Per-seed score files and PRODIGY outputs were not recovered; reported TCR–pMHC ipTM, ΔG , and K_d remain curated records. Interface-pLDDT recomputed from the recovered structures' pLDDT-bearing B-factors agrees with the workbook to approximately 2–3 pLDDT units, corroborating but not reproducing the curated column. No reported metric or nomination was changed by this recovery.

The audit compared stored TCR–pMHC ipTM SDs (column AE), candidate–control p -values (AA), and CI-text means and half-widths (AC) against per-seed recomputation across the 61-peptide workbook, identifying 44 discrepancies (22 SDs, 12 p -values, and 10 CI texts), including 17 SD entries that instead contained the corresponding mean, chiefly in the trans-system block (results/workbook_audit.csv). Reported seed-derived statistics use recomputed values rather than those stored summary columns; only the already-disclosed KHVANSVFAEIIKT case changed significance classification (Section 3.6). The interface-pLDDT, ΔG , and K_d CI fields were empty, so no stored intervals for those metrics were available to audit; their means and intervals used in the analyses were calculated from the per-seed records, not taken from stored CI fields.

References

- [1] Ogle GD, Wang F, Haynes A, Gregory GA, King TW, Deng K, et al. (2025). Global type 1 diabetes prevalence, incidence, and mortality estimates 2025: results from the International Diabetes Federation Atlas, 11th Edition, and the T1D Index Version 3.0. *Diabetes Research and Clinical Practice* 225:112277. doi:10.1016/j.diabres.2025.112277.
- [2] Erlich H, Valdes AM, Noble J, Carlson JA, Varney M, Concannon P, et al.; Type 1 Diabetes Genetics Consortium (2008). HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the Type 1 Diabetes Genetics Consortium families. *Diabetes* 57:1084–1092.
- [3] Zubkiewicz-Kucharska A, Jamer T, Chrzanowska J, Akutko K, Pytrus T, Stawarski A, et al. (2022). Prevalence of haplotype DQ2/DQ8 and celiac disease in children with type 1 diabetes. *Diabetology & Metabolic Syndrome* 14:128. doi:10.1186/s13098-022-00897-8.
- [4] Crawford SA, Wiles TA, Wenzlau JM, Powell RL, Barbour G, Dang M, et al. (2022). Cathepsin D drives the formation of hybrid insulin peptides relevant to the pathogenesis of type 1 diabetes. *Diabetes* 71:2793–2803. doi:10.2337/db22-0303.
- [5] Baker RL, Rihanek M, Hohenstein AC, Nakayama M, Michels AW (2019). Hybrid insulin peptides are autoantigens in type 1 diabetes. *Diabetes* 68:1830–1840. doi:10.2337/db19-0128.
- [6] Wiles TA, Powell RL, Michels AW (2021). Structural basis of the recognition of a hybrid insulin peptide by a type 1 diabetes-associated T-cell receptor. *Frontiers in Immunology* 12:668680. doi:10.3389/fimmu.2021.668680.
- [7] Wiles TA, Hohenstein A, Landry LG, Dang M, Powell R, Guyer P, et al. (2021). The structural basis of hybrid insulin peptide recognition by the immune system. *Nature Communications* 12:5404. doi:10.1038/s41467-021-25404-x.
- [8] Isaacs SR, Foskett DB, Maxwell AJ, Ward EJ, Faulkner CL, Luo JYX, et al. (2021). Viruses and type 1 diabetes: from enteroviruses to the virome. *Microorganisms* 9:1519. doi:10.3390/microorganisms9071519.
- [9] Richardson SJ, Morgan NG (2018). Enteroviral infections in the pathogenesis of type 1 diabetes: new insights for therapeutic intervention. *Current Opinion in Pharmacology* 43:11–19.
- [10] Heather JM, Spindler MJ, Alonso MH, Shui YI, Millar DG, Johnson DS, et al. (2022). Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information. *Nucleic Acids Research* 50:e68. doi:10.1093/nar/gkac190.
- [11] de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research* 34:W362–W365. doi:10.1093/nar/gkl124.
- [12] Yin R, et al. (2023). TCRmodel2: high-resolution modeling of T-cell receptor recognition using deep learning. *Nucleic Acids Research* 51:W569–W576. doi:10.1093/nar/gkad369.

- [13] Xue LC, Rodrigues JPGLM, Kastritis PL, Bonvin AMJJ, Vangone A (2016). PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics* 32:3676–3678. doi:10.1093/bioinformatics/btw514.
- [14] Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630:493–500. doi:10.1038/s41586-024-07487-w.
- [15] Schrödinger LLC, DeLano W (2020). PyMOL molecular graphics system. <https://www.pymol.org/>.
- [16] UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* 51:D523–D531.
- [17] Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research* 47:D339–D343.
- [18] Michels AW, Landry LG, Nakayama M, et al. (2017). Islet-derived CD4 T cells targeting proinsulin in human autoimmune diabetes. *Diabetes* 66:722–734.
- [19] Nakayama M, Michels AW (2021). Using the T cell receptor as a biomarker in type 1 diabetes. *Frontiers in Immunology* 12:777788.
- [20] Nilsson JB, Kaabinejadian S, Yari H, Peters B, Barra C, Gragert L, et al. (2023). Machine learning reveals limited contribution of trans-only encoded variants to the HLA-DQ immunopeptidome. *Communications Biology* 6:442.
- [21] Deleuran SN, Nielsen M (2025). NetTCR-struc, a structure driven approach for prediction of TCR-pMHC interactions. *Frontiers in Immunology* 16:1616328. doi:10.3389/fimmu.2025.1616328.
- [22] Jardetzky TS, Brown JH, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC (1996). Crystallographic analysis of endogenous peptides associated with HLA-DR1 suggests a common, polyproline II-like conformation for bound peptides. *Proceedings of the National Academy of Sciences USA* 93:734–738. doi:10.1073/pnas.93.2.734.
- [23] Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020:baaa062. doi:10.1093/database/baaa062.

Table 1

Reference TCR clone composition.

Clone	TRAV	TRAJ	CDR3 α	TRBV	TRBJ	CDR3 β
8.E3	TRAV2	TRAJ37	CAVDGSGNTGKLIF	TRBV4-1	TRBJ2-7	CASSQDLAGVREQYF
E2	TRAV8-4	TRAJ43	CAVGATNNNDMRF	TRBV5-4	TRBJ2-1	CASSPIGASGGNEQFF
A3.10	TRAV38-1	TRAJ54	CAFFGQGAGKLVF	TRBV5-1	TRBJ2-3	CASSLSASGGATDTQYF
A2.11	TRAV38-1	TRAJ54	CAFMGAGAQLVF	TRBV4-3	TRBJ2-3	CASSQILRGPPDQYF

Clone 8.E3 was modelled in both *cis* and *trans* DQ8 contexts, as detailed in Table 2.**Table 2**

Antigen, restriction element, and AlphaFold 3 model confidence for each reference system.

#	Clone	HIP	MHC II α	MHC II β	Restriction	ipTM	pTM
1	8.E3	HIP11	DQA1*03:01	DQB1*03:02	DQ8	0.88	0.90
2	8.E3	HIP11	DQA1*05:01	DQB1*03:02	DQ8-trans	0.87	0.89
3	E2	HIP11	DQA1*05:01	DQB1*02:01	DQ2	0.91	0.91
4	A3.10	HIP6	DQA1*03:01	DQB1*03:02	DQ8	0.86	0.89
5	A2.11	HIP6	DQA1*03:01	DQB1*03:02	DQ8	0.85	0.87

HIP6, GQVELGGG-NAVEVLK; HIP11, SLQPLAL-EAEDLQV. The double dash marks the hybrid junction.

Table 3

Consensus contact assignment across the canonical nine-residue class II core used to guide motif construction. This schematic is not a per-system binding register.

Position	Contact	Position	Contact	Position	Contact
P1	MHC anchor	P4	MHC	P7	MHC/TCR
P2	TCR face	P5	TCR	P8	TCR
P3	MHC/TCR	P6	MHC	P9	MHC

Table 4

Physicochemical substitution classes permitted at contact positions.

Class	Residues	Proposed structural role
Hydrophobic	L, I, V	Create hydrophobic contact protrusions
Small/polar	A, G, S, T	Provide steric spacers without disrupting the interface
Acidic	E, D	Provide negative charge

Table 5

Search motifs derived from HIP pharmacophores.

Screen	ID	ScanProsite motif
Structural	1	x(5) - [AGST] - [LIV] - [ED] - [AGST] - [ED] - x(4)
Structural	2	x(5) - [AGST] - [LIV] - x - [AGST] - [ED] - x - [LIV] - x(2)
Structural	5	x(3) - [ED] - [LIV] - x(3) - x - [AGST] - [LIV] - [ED] - x(3)
Additional binding	3	x(6) - [LIV] - [ED] - [AGST] - [ED] - [ED] - [LIV] - x(2)
Additional binding	4	x(4) - [LIV] - x(4) - [AGST] - [LIV] - [ED] - [LIV] - x(2)

Table 6

Intended T1D-associated targets of the Swiss-Prot screen; nine have reviewed entries in the contemporary reconstruction, while *B. dorei* has none. Enterovirus and Herpesviridae are viral groups rather than individual organisms; the NCBI taxonomy identifiers used to query UniProtKB/Swiss-Prot are listed in Table 10.

Viral group	Bacterial species
Mumps virus	<i>Bacteroides dorei</i>
Rubella virus	<i>Clostridium sporogenes</i>
SARS-CoV-2	<i>Klebsiella oxytoca</i>
Enterovirus	<i>Parabacteroides distasonis</i>
Rotavirus	
Herpesviridae	

Table 7

Organism-level distribution of the 10 candidates meeting the native-HIP structural and energetic criteria.

Organism	Count
<i>Parabacteroides distasonis</i> strain ATCC 8503	3
<i>Klebsiella oxytoca</i>	3
Human herpesvirus 1 strain 17	1
Human cytomegalovirus strain AD169 (HHV-5)	1
Human rhinovirus 3	1
SARS-CoV-2	1
Total	10

Table 8

HIP6/A2.11/DQ8 three-metric-gate passers: four nominees and three animal-associated exclusions, relative to native HIP6.

Candidate	Source	ipTM	$\Delta\%$	i-pLDDT	$\Delta\%$	ΔG	$\Delta\%$	Exceeds native on all three
VGVEVSAANTLEECW	<i>P. distasonis</i>	0.842	-0.473	83.48	+2.775	-12.14	+4.836	No
SESEVESDTALESEV	HHV-1	0.850	+0.473	81.39	+0.202	-12.14	+4.836	Yes
NDDELAGNLGLDVS	<i>K. oxytoca</i>	0.846	0.000	81.04	-0.229	-11.12	-3.972	No
RLCDLDTADAVEEMG	<i>K. oxytoca</i>	0.836	-1.182	81.21	-0.017	-11.20	-3.282	No
<i>Excluded at human-association filter; all three passed the metric gate</i>								
RHGEVNDSTTVEPVL	RVA/Pig/Gottfried	0.876	+3.546	88.01	+8.357	-10.92	-5.699	No
GPGEVNDSTTVEPIL	RVA/Equine/H2	0.874	+3.310	84.84	+4.457	-10.94	-5.527	No
GPGEVNDSTTVEPVL	RVA/Equine/FI-14	0.870	+2.837	86.28	+6.232	-10.66	-7.945	No

ΔG is reported in kcal mol⁻¹. Native HIP6: ipTM 0.846, i-pLDDT 81.22, and ΔG -11.58 kcal mol⁻¹. The identical ΔG means for VGVEVSAANTLEECW and SESEVESDTALESEV arise from different per-seed values, not duplicated records. The three flagged rows were excluded for host association, not score, and are not part of the ten nominees. Full Rotavirus A strain identifiers appear in Table 10.

Table 9

HIP11/8.E3/DQ8-trans three-metric-gate nominees: metrics and percentage changes relative to native HIP11.

Candidate	Source	ipTM	$\Delta\%$	i-pLDDT	$\Delta\%$	ΔG	$\Delta\%$	Exceeds native on all three
RRNVATLQAENVTG	SARS-CoV-2	0.842	+2.433	90.53	+0.714	-9.46	+4.185	Yes
SDLLKSVDSEEVRE	HRV-3	0.826	+0.487	89.67	-0.247	-9.92	+9.251	No
LQEILTITKSDDVVG	<i>P. distasonis</i>	0.822	0.000	89.96	+0.082	-9.10	+0.220	No
RDISSTIATEKIPF	HHV-5	0.820	-0.243	86.48	-3.791	-8.86	-2.423	No
DYVVNTINSDPME	<i>P. distasonis</i>	0.810	-1.460	85.75	-4.608	-9.34	+2.863	No
GKDIEGVGSEDLVN	<i>K. oxytoca</i>	0.782	-4.866	86.12	-4.196	-9.16	+0.881	No

ΔG is reported in kcal mol⁻¹. Native HIP11: ipTM 0.822, i-pLDDT 89.89, and ΔG -9.08 kcal mol⁻¹.

Table 10

Screened proteomes and application of the human-association nomination criterion.

Taxon (Table 6)	Species or strain screened	Query TaxID	Database searched	Human-associated (Yes/No)
Mumps virus	Mumps virus genotype A (Jeryl-Lynn)	2560602	UniProtKB/Swiss-Prot	Yes
Rubella virus	Rubella virus	11041	UniProtKB/Swiss-Prot	Yes
SARS-CoV-2	SARS-CoV-2	2697049	UniProtKB/Swiss-Prot	Yes
Enterovirus	Enterovirus (genus)	12059	UniProtKB/Swiss-Prot	Yes
Rotavirus	Rotavirus (genus)	10912	UniProtKB/Swiss-Prot	Mixed, see note
Rotavirus	RVA/Pig/US/Gottfried/1983	—	UniProtKB/Swiss-Prot	No
Rotavirus	RVA/Equine/UK/H2/1976	—	UniProtKB/Swiss-Prot	No
Rotavirus	RVA/Equine/US/FI-14/1980	—	UniProtKB/Swiss-Prot	No
Herpesviridae	Herpesviridae (family)	10292	UniProtKB/Swiss-Prot	Yes
Bacteroides dorei	B. dorei (no reviewed entries in reconstruction)	357276	UniProtKB/Swiss-Prot	Yes
Clostridium sporogenes	C. sporogenes	1505	UniProtKB/Swiss-Prot	Yes
Klebsiella oxytoca	K. oxytoca	571	UniProtKB/Swiss-Prot	Yes
Parabacteroides distasonis	strain ATCC 8503	823	UniProtKB/Swiss-Prot	Yes

Query TaxID denotes the taxonomy filter actually submitted, not a strain identifier. Rotavirus was queried once at genus level (10912), returning both human- and animal-associated strains. The three strain rows list hits from that query; dashes mean no separate strain-level query was used, not that a strain lacks a TaxID. Animal-associated candidates were excluded at nomination. The original identifier-retrieval date was not retained; a taxonomy-name check on 8 September 2026 identifies Orthoherpesviridae (TaxID 3044472) as the current family corresponding to the recorded Herpesviridae query (10292) [23]. *B. dorei* is currently named *Phocaeicola dorei*; its zero reviewed entries in the contemporary reconstruction mean that the ten intended targets represent nine available sequence sets, not demonstrated historical coverage of all ten. Current taxonomy and coverage do not reconstruct the original search snapshot.

Table 11

Pooled and within-system dependence of the three prioritization metrics across candidate means.

System	<i>n</i>	ipTM / i-pLDDT	ipTM / ΔG	i-pLDDT / ΔG	Effective dimension
Pooled	55	0.560	-0.580	0.076	2.09
HIP6/A2.11/DQ8	28	0.648	+0.086	-0.063	2.33
HIP11/8.E3/DQ8-trans	20	0.905	-0.240	-0.251	1.84
HIP11/8.E3/DQ8	7	0.972	-0.457	-0.452	1.57

Pairwise entries are Pearson correlations across each candidate's five-seed mean, not correlations obtained by pooling individual seed observations. Effective dimension is the participation ratio of the three-metric correlation matrix. The within-system estimates inform prioritization; these covariance summaries neither prove independence nor calibrate a sign-test probability. The cis-system estimates rest on only seven candidates; its correlations and effective dimension (1.57) are unstable and indicative only. Native HIP11/8.E3/DQ8 (cis): ipTM 0.812, interface-pLDDT 84.52, and ΔG -8.72 kcal mol⁻¹.

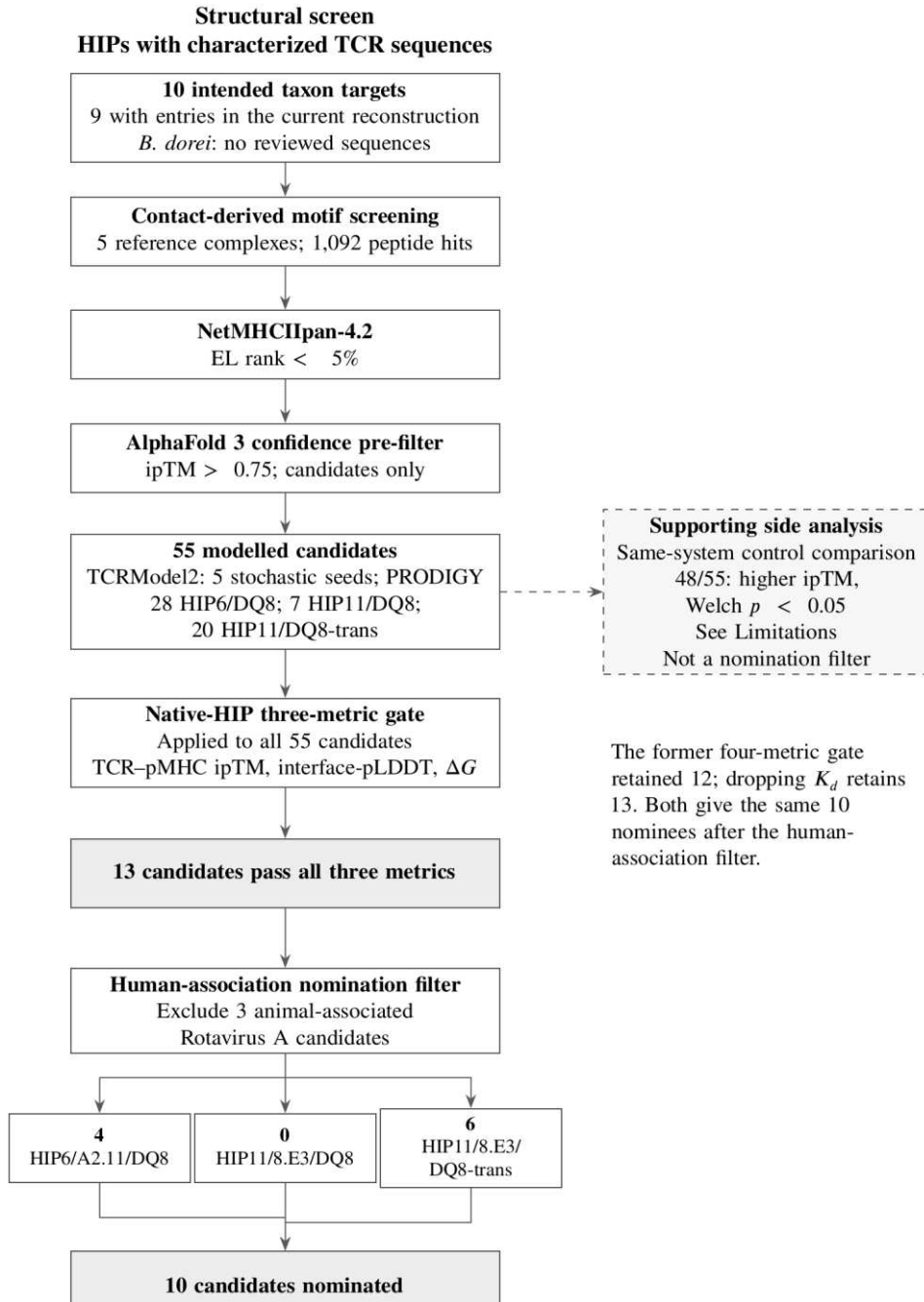


Figure 1: Structural screening and nomination workflow. The three-metric native-HIP gate takes all 55 modelled candidates as input; 13 pass, and the separate human-association filter excludes three animal-associated Rotavirus A peptides, leaving ten nominees in a 4/0/6 split. The former four-metric gate gave the same ten nominees; K_d is descriptive, not a gate criterion. The candidate-control comparison is a supporting side analysis, not part of nomination (Section 3.12). Taxon coverage refers to the contemporary Swiss-Prot reconstruction, not a recovered historical snapshot.

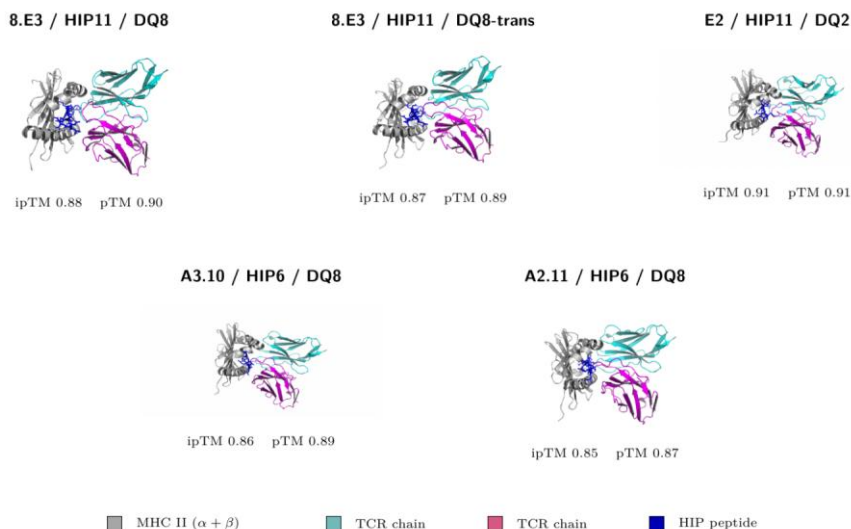


Figure 2: Reference TCR–pMHC–HIP complexes modelled with AlphaFold 3 and rendered in PyMOL. The model with the highest ipTM of five stochastic seeds was retained for each system. MHC class II is shown in gray, the TCR chains in teal and magenta, and the HIP peptide in blue. Model ipTM and pTM values are shown below each complex.

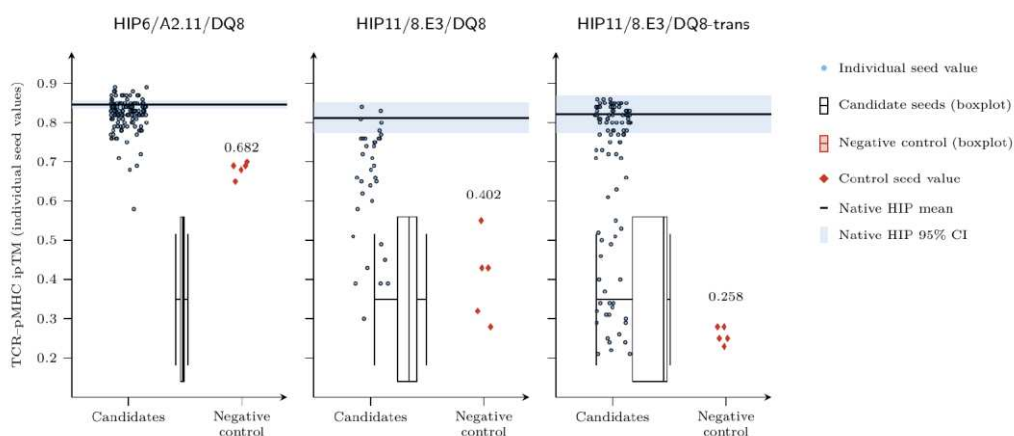


Figure 3: Per-system distributions of TCR–pMHC ipTM values across five stochastic seeds. Blue points show candidate seed values and red diamonds show negative-control seed values. Boxplots summarize each group; solid lines mark the native-HIP mean and shaded bands its 95% confidence interval. This supporting comparison is interpreted in Section 3.12.

Supplementary Material

Table S1

Within-system composite position (secondary descriptive analysis).

Rank	Candidate	Source	Standardized $\Delta\%$ (ipTM)	Standardized $\Delta\%$ (i-pLDDT)	Standardized $\Delta\%$ (ΔG)	Composite
<i>HIP6/A2.11/DQ8</i>						
1	VGVEVSAANTLEECW	P. distasonis	-0.670	+1.974	+0.988	+0.764
2	SESEVESDTALESEV	HHV-1	+0.670	+0.144	+0.988	+0.600
3	NDDELAGNLGLDVSN	K. oxytoca	0.000	-0.163	-0.812	-0.325
4	RLCDLDTADAVEEMG	K. oxytoca	-1.674	-0.012	-0.670	-0.786
<i>HIP11/8.E3/DQ8-trans</i>						
1	SDLLKSVDSSEEVRE	HRV-3	+0.199	-0.101	+2.304	+0.801
2	RRNVATLQAENV TG	SARS-CoV-2	+0.997	+0.293	+1.042	+0.777
3	LQEILTIKSDDVVG	P. distasonis	0.000	+0.034	+0.055	+0.030
4	DYVVNTINSDPVME	P. distasonis	-0.598	-1.894	+0.713	-0.593
5	RDISSTIATEKIPF	HHV-5	-0.100	-1.558	-0.603	-0.754
6	GKDIEGVGSEDLVN	K. oxytoca	-1.993	-1.724	+0.219	-1.166

Composite score is the equally weighted mean of the three standardized values. Scores are comparable only within a system. The standardization cohorts are the four HIP6 and six HIP11/DQ8-trans nominees, not the full candidate pools. See Section 2.9.3 for the calculation. Composite ranks did not determine nomination and do not establish an ordering of the two experimental co-priorities (Section 3.9).

Table S2

Metric-provenance inventory and separate structural recovery in the public archive.

Provenance tag	Entries	Interpretation
RAW	25	Values traced to retained original model output. The five reference-complex AlphaFold 3 jobs have retained raw outputs.
WORKBOOK	1,220	Curated spreadsheet metrics. Structures are partially recovered; original per-seed score files and PRODIGY outputs remain unavailable.
MISSING	610	Unavailable entries in the reported inventory, not measured zeros or imputed values.
Total inventory entries	1,855	Includes unavailable entries; the count is not a count of independent observations.

Counts describe metric entries, not structure files. Separately, 113 recovered original TCRModel2 structures cover 51 of 56 distinct workbook peptide sequences, including native-HIP and control records (*recovered_candidate_models/manifest.csv*). This partial, structures-only recovery does not reclassify any of the 1,220 WORKBOOK entries as RAW: original per-seed score files and PRODIGY outputs were not recovered, and structural interface-pLDDT checks do not exactly reproduce the curated values. The row-level table *data/per_seed_metrics.csv* identifies system, peptide, seed, metric, provenance, and the originating workbook cell in *source_file*. Raw AF3 reference metrics are listed in *data/reference_complex_af3_metrics.csv*. Overall pLDDT, pTM, and model-confidence scores were not retained in the workbook for the 55 candidates. See Data availability for recovery coverage and archive access.