

TCREN2 — A METHODS REFERENCE:

CONTACT-POTENTIAL SCORING, RANKING, AND STRUCTURE OPERATIONS

Mikhail Shugay^{1*}

¹ISALGO laboratory

tcren technical appendix

Abstract

`tcren` scores a candidate epitope for a fixed TCR–peptide–MHC structure by summing a residue-level pairwise statistical potential over the observed contact map [2]. The score depends only on residue identity and the closest-atom contact geometry, so a candidate is scored *virtually*: its amino acids are threaded onto the native contacts with no atom moved. This appendix is the implementation-level companion to the method’s biophysical synopsis — it states, as the code computes them, the contact map, the statistical potentials and their derivation, the per-task readouts (epitope ranking, percentile rank, mutational $\Delta\Delta G$), the two coordinate-level operations (backbone-preserving substitution and knowledge-based pose refinement), and the interface mechanics. Every default parameter is collected in Table 1; the formulas are the exact quantities evaluated.

1. WORKFLOW

Figure 1 is the end-to-end pipeline. A single call (`tcren pipeline`) takes a structure through: **(1) import** (C-gene-trimmed parse of PDB / mmCIF); **(2) annotate** — TCR loci and CDR/framework regions by germline alignment (`ARDA`), MHC allele, class, chain role and groove positions; **(3) superimpose** (optional) onto the canonical database, placing the complex in a common MHC-groove C_α frame so downstream geometry is comparable across complexes; **(4) contacts** — the 5 Å heavy-atom contact map with a per-residue region markup; and **(5) score** — each interface summed under its potential. Candidate epitopes, percentile ranks and $\Delta\Delta G$ are read *virtually* off the resulting map; substitution, refinement and mechanics operate on coordinates. The same operations are exposed as subcommands (`annotate`, `orient`, `contacts`, `score`, `rank`, `ddg`, `refine`) and as a Python API.

2. CONTACT MAP

For a chain-typed complex `tcren` builds the inter-chain contact map: two residues on different chains are in contact when their closest heavy-atom pair lies within the cutoff $r_c = 5$ Å, reduced to *one* pair

*Correspondence: mikhail.shugay@gmail.com

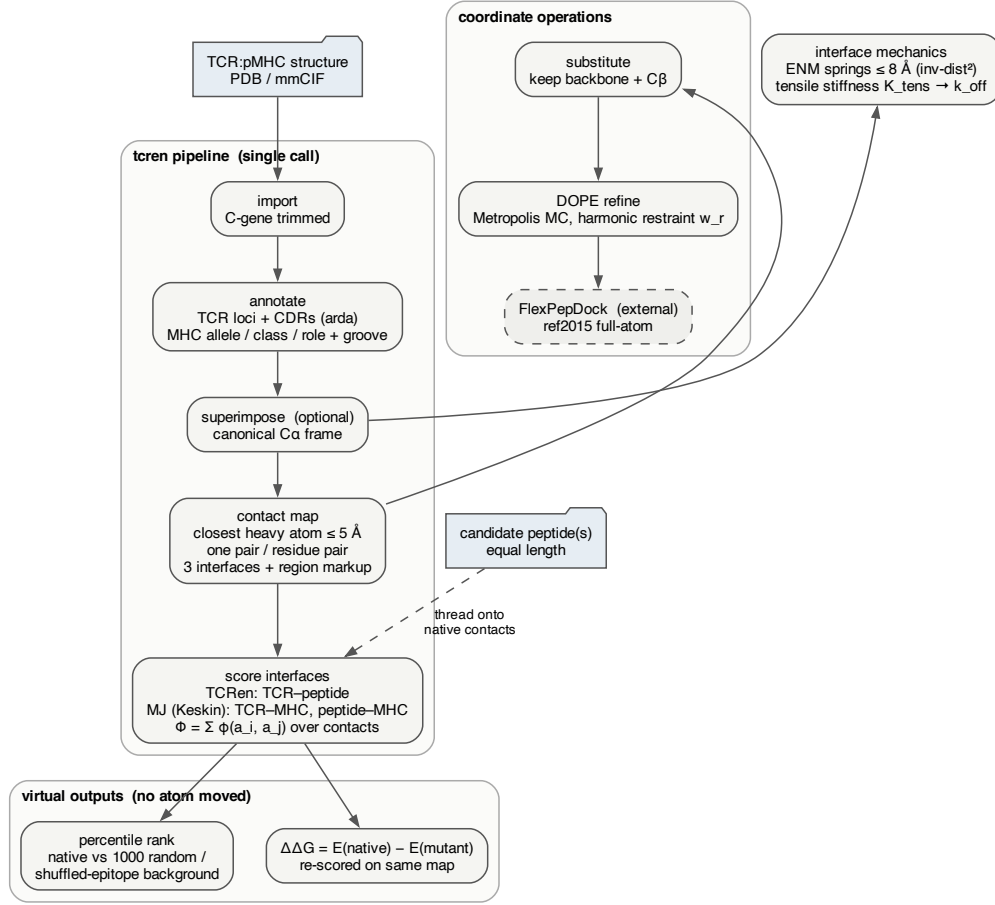


Figure 1: The `tcren2` workflow. The single-call pipeline parses, annotates, orients and builds the 5 Å contact map, then scores each interface (TCRen on TCR-peptide, Miyazawa-Jernigan on the presentation interfaces). Candidate peptides thread onto the native contacts for virtual scoring, percentile ranking and $\Delta\Delta G$ (no atom moved); the coordinate operations (substitute, DOPE refine, external FlexPepDock relaxation) and the interface-mechanics off-rate proxy consume the oriented structure. Node annotations are the code defaults of Table 1.

per residue pair (the closest). Partitioning by chain type gives the three interfaces of the ternary complex,

$$\mathcal{C} = \mathcal{C}_{\text{TCR:pep}} \cup \mathcal{C}_{\text{TCR:MHC}} \cup \mathcal{C}_{\text{pep:MHC}}.$$

Each contact carries the two residue identities and, optionally, its heavy-atom-pair count n_{ij}^{atoms} (built with `count_atoms=True`). The map is the single object every downstream step consumes; the contact energy is orientation-invariant, while the optional canonical frame (step 3) is what makes *geometry* comparable across complexes.

3. STATISTICAL POTENTIALS

TCRen. The TCRen potential is the Boltzmann inversion of observed TCR–peptide contact frequencies against an independent-pairing reference [2, 10]. From the contact table, with n_{ab} the count of directed contacts between partner residue a and peptide residue b (each incremented by a pseudo-count κ), row/column/grand totals $N_a = \sum_b n_{ab}$, $N_b = \sum_a n_{ab}$, $N = \sum_{ab} n_{ab}$,

$$\text{odds}_{ab} = \frac{n_{ab} N}{N_a N_b}, \quad \phi_{ab}^{\text{TCRen}} = -\ln \text{odds}_{ab} \quad (1)$$

(lower is more favourable). The grid enumerates the twenty amino acids; cysteine is dropped from the partner (“from”) axis *after* the log-odds, a data convention of the reference derivation. An alignment-matrix variant replaces $-\ln$ by $-\log_2(\cdot)/\beta$ over 21 symbols (a gap included), $\beta = 44$; the classic natural-log form above is the shipped default and reproduces `TCRen_potential.csv` exactly. Optional per-structure weights down-weight redundant complexes in the counts while keeping all data.

PRESENTATION INTERFACES. The peptide–MHC and TCR–MHC interfaces are scored with the Miyazawa–Jernigan quasi-chemical contact potential [6] (Keskin parameterisation), the same log-odds construction (1) applied to a generic protein-contact reference. Using an independent potential per interface — TCRen where thymic selection shaped the statistics, MJ where it did not — is the combination the benchmarks favour.

4. SCORING, RANKING, AND $\Delta\Delta G$

INTERFACE ENERGY. A candidate peptide s is scored on an interface by threading its residues onto the contacted peptide positions and summing the potential,

$$\Phi_{XY}(s) = \sum_{(i,j) \in \mathcal{C}_{XY}} w_{ij} \phi(a_i, s_{\pi(j)}), \quad (2)$$

where ϕ is the interface’s potential, a_i the fixed partner residue, $\pi(j)$ the contact’s peptide position, and the weight is $w_{ij} = 1$ (default, *residue* weighting) or $w_{ij} = n_{ij}^{\text{atoms}}$ (*atomic* weighting, which tracks the LJ+Coulomb atom-pair sum more closely). Pairs absent from the potential (e.g. a cysteine partner) contribute nothing. The total complex energy is $\Phi = \Phi_{\text{TCR:pep}}(\text{TCRen}) + \Phi_{\text{TCR:MHC}}(\text{MJ}) + \Phi_{\text{pep:MHC}}(\text{MJ})$. Scoring is fixed-backbone and fixed-contact-set — the native pose stands in for every candidate — so a whole library ranks from one structure in microseconds per peptide.

PERCENTILE RANK. To turn a raw energy into a calibrated rank, the native peptide is scored against a background of $N_{\text{bg}} = 1000$ peptides of the same length — uniform-random over the twenty amino acids, or position-shuffled draws from a supplied epitope list — and its percentile reported,

$$\text{rank\%} = \frac{1}{N_{\text{bg}}} |\{b : \Phi(b) \leq \Phi(\text{native})\}|. \quad (3)$$

A small `rank_pct` marks a strong binder relative to chance; the background seed is fixed (default 0) for reproducibility.

MUTATIONAL $\Delta\Delta G$. A peptide point mutation is scored *virtually* — no atom moves, no re-docking — by re-scoring the mutant on the same contact map,

$$\Delta\Delta G = \Phi(\text{native}) - \Phi(\text{mutant}), \quad (4)$$

so a positive value flags a destabilising mutation. Only contacts touching the mutated position survive the difference; an interface without the peptide (TCR–MHC) contributes identically zero. This is the fast triage matrix (`alanine_scan`, `neoantigen_ddg`); a quantitative, sampling-dependent $\Delta\Delta G$ is left to the coordinate path below and, ultimately, to full relaxation.

POLY-ALANINE REFERENCE. Summing (4) over every peptide position gives the poly-alanine reference difference $\Delta\Phi = \Phi(s) - \Phi(A^L)$ (`reference_delta`) — the full-peptide alanine scan, which subtracts the pose’s identity-independent baseline $\Phi(A^L)$ (what the geometry scores when every peptide residue is alanine) and keeps the sequence-specific part. On a *fixed* contact map $\Delta\Phi = \Phi(s) - \text{const}$, so it re-ranks nothing; it matters only across candidates that each carry their *own* structure (e.g. AlphaFold peptide-swap models), where it normalises out per-pose interface geometry and rescues forced or mis-registered poses whose geometry corrupts the raw energy. Both Φ and $\Delta\Phi$ are *relative*, *within-receptor* peptide rankings, not binding constants: they order a peptide panel against one fixed TCR:pMHC (e.g. the graded B*27:05 EC₅₀ series, Spearman $\rho \approx 0.75$ –0.9) but predict equilibrium affinity *across* complexes only weakly — on the ATLAS SPR set raw Φ and $\Delta\Phi$ track ΔG , K_d , k_{off} and k_{on} at $|\rho| \leq 0.3$; the off-rate, when a structure predicts anything, comes from the interface mechanics of §7, not the contact sum.

5. BACKBONE-PRESERVING PEPTIDE SUBSTITUTION

DEFINITION 1 (Substitution). Let the native peptide chain have residues $r_1, \dots, r_L$ with heavy-atom coordinate sets $X_1, \dots, X_L$, and let $s \in \mathcal{A}^L$ be a new sequence over the twenty amino acids. The substitution $\text{sub}(\cdot, s)$ returns a chain whose i -th residue has identity s_i and retains exactly the backbone atoms $\{\text{N}, \text{C}_\alpha, \text{C}, \text{O}\}$ together with C_β when $s_i \neq \text{Gly}$, with all coordinates inherited from X_i ; side-chain atoms beyond C_β are discarded.

RATIONALE. A residue-level potential reads the contact map through residue *identity* and the closest-atom (or C_β) distance [6]; the backbone + C_β frame therefore carries everything the score and the refinement energy of §6 consume, while the native backbone—rigidly constrained between the MHC groove walls and the TCR loops—is the appropriate scaffold for an equal-length epitope. Rebuilding the discarded side chains (a backbone-dependent rotamer search [8]) and full physics relaxation are deliberately *not* performed here; substitution only fixes identity and C_β geometry, leaving heavier modelling to the refiner or to an external tool. The map is purely combinatorial—no alignment, no optimisation—so it is exact and order-preserving on the peptide’s sequential indexing.

6. POTENTIAL-GUIDED REFINEMENT

Write $x \in \mathbb{R}^{3M}$ for the M peptide heavy-atom coordinates (the only degrees of freedom) and $y \in \mathbb{R}^{3M'}$ for the fixed partner (TCR + MHC) heavy atoms. Each heavy atom carries a `DOPE` atom class

$c(\cdot) \in \{1, \dots, 158\}$ —a MODELLER mean-force atom type, i.e. a (residue, atom-name) label—and let $\varphi_{c,c'} : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ be the DOPE distance-dependent potential of mean force for the class pair (c, c') [9], tabulated on 29 knots at 0.75–14.75 Å (0.5 Å spacing), linearly interpolated, and set to 0 beyond $d_{\max} = 14.75$ Å. Crucially φ is *not* TCRen or Miyazawa–Jernigan—those score the final epitope (§4), so reusing them to refine would optimise the pose against the very quantity reported. DOPE is a single, chain-agnostic, atom-resolved potential applied uniformly to every peptide↔partner pair.

DEFINITION 2 (Refinement energy).

$$E(x) = \underbrace{\sum_{p=1}^M \sum_{q=1}^{M'} \varphi_{c(p), c(q)}(\|x_p - y_q\|)}_{\text{DOPE}} + w_r \sum_{p=1}^M \|x_p - x_p^0\|^2.$$

DOPE’s short-range knots are strongly repulsive (capped near +10), so the potential supplies its own excluded-volume term: there is no separate clash penalty. The second term is a harmonic restraint to the input pose x^0 (default $w_r = \frac{1}{2}$).

BOUNDEDNESS AND THE ROLE OF THE RESTRAINT. Unlike a residue contact-count energy—whose trivial rigid-body minimum can be “no contacts” (the peptide ejected)—the DOPE term is bounded below and tends to 0 as the peptide is translated away, so a finite rigid-body minimiser always exists.

PROPOSITION 1. *Restrict x to the rigid-body family $x = Rx^0 + t$, $R \in SO(3)$, $t \in \mathbb{R}^3$, and let $E_D(x) = \sum_{p,q} \varphi_{c(p), c(q)}(\|x_p - y_q\|)$. Then E_D is bounded ($|E_D| \leq MM' \max |\varphi|$) and $E_D \rightarrow 0$ as $\|t\| \rightarrow \infty$. Its rigid-body minimiser is the globally best-packed pose, which need not coincide with the input x^0 : the peptide may slide within the groove. Adding $w_r \sum_p \|x_p - x_p^0\|^2$ makes E coercive and confines the minimiser to a neighbourhood of x^0 whose radius decreases with w_r , yielding a local refinement—a native pose (already near a DOPE optimum) barely moves, while a clashed or displaced input relaxes locally.*

SAMPLING. E is minimised by simulated-annealing Metropolis Monte Carlo [5, 3]. At step t a trial pose is proposed by a rigid move of the peptide: a Gaussian translation $\delta \sim \mathcal{N}(0, \sigma_t^2 I_3)$ and a rotation about the peptide centroid by a random axis and angle $\theta \sim \mathcal{N}(0, \sigma_r^2)$. The trial is accepted with probability $\min(1, \exp(-\Delta E/T_t))$, with the temperature cooled linearly $T_t : T_0 \downarrow T_1$; the lowest-energy pose visited is returned. The chain is deterministic given its pseudo-random seed.

IMPLEMENTATION AND COST. Each energy evaluation is $O(MM')$ over heavy-atom pairs; partner atoms are pre-filtered to those within a 12 Å shell of the peptide, so M' is the interface, not the whole complex, and several thousand steps complete in well under a second. The kernel is C++ (pybind11); the symmetric $158 \times 158 \times 29$ DOPE table is passed by buffer, an atom’s class is its (residue, atom) label, and φ is a single buffer lookup plus a linear interpolation. The bundled table is parsed once from the MODELLER mean-force libraries (`atmcls-mf.lib`, `dist-mf.lib`) as redistributed by the pymod/altmod project.¹ A native pose is a fixed point of the move set up to thermal noise (it barely moves); a clashed or displaced input relaxes locally.

¹<https://github.com/pymodproject/altmod> (DOPE libraries from MODELLER; [9]).

7. INTERFACE MECHANICS

The affinity-adjacent quantity a native structure actually predicts is not the equilibrium free energy but the dissociation *rate* k_{off} , which sets the bond lifetime a T cell reads through kinetic proofreading. `tcren` models it mechanically (`tcren.mechanics`): heavy-atom interface contacts within 8 Å become elastic-network springs with inverse-square-distance stiffness, and the interface is pulled along the TCR↔pMHC dissociation axis. The tensile stiffness K_{tens} (`stiffness_tensor`) tracks experimental $\log k_{\text{off}}$ — a stiffer interface ruptures more slowly (Bell–Evans) — while a breakable-spring variant records the peak resisting force and work past a strain threshold (`rupture`). These consume the oriented structure only; no potential or sequence enters.

8. PARAMETER REFERENCE

Table 1 collects every default. Values are the code defaults (function signatures / CLI options); the last column points to the formula or section that uses each.

Table 1: Default parameters of the `tcren2` computations. One value per cell; parenthesised names are the CLI options.

Quantity	Default	Where
Contact cutoff r_c	5 Å (<code>--cutoff</code>)	§ 2
Contact reduction	closest heavy-atom pair, one per residue pair	§ 2
Contact weight	<code>residue</code> ($w=1$), or <code>atomic</code> ($w=n^{\text{atoms}}$)	(2)
TCRen potential	—ln odds, 20 aa (classic)	(1)
Pseudocount κ	1 (<code>--pseudocount</code>)	(1)
Cysteine	dropped from “from” axis (classic)	(1)
am-variant temperature β	44	§ 3
Presentation potential	Miyazawa–Jernigan (Keskin)	§ 3
Background size N_{bg}	1000 (<code>--background</code>)	(3)
Background seed	0	(3)
$\Delta\Delta G$	$\Phi(\text{nat}) - \Phi(\text{mut})$, peptide interfaces	(4)
Substitution frame	backbone {N, C $_{\alpha}$, C, O} + C $_{\beta}$	Def. 1
DOPE atom classes	158 MODELLER mean-force types	§ 6
DOPE knots	29 at 0.75–14.75 Å, 0.5 Å spacing	§ 6
DOPE cutoff d_{max}	14.75 Å ($q=0$ beyond)	§ 6
Restraint weight w_r	0.5 (<code>--restraint</code>)	Def. 2
Monte-Carlo steps	2000 (<code>--steps</code>)	§ 6
Partner shell	12 Å around the peptide	§ 6
Mechanics cutoff	8 Å	§ 7
Spring stiffness	inverse-distance ²	§ 7
Pull direction	tensile (dissociation axis)	§ 7

9. ASSUMPTIONS AND LIMITATIONS

The peptide is treated as a *rigid* body: there is no backbone flexibility and no side-chain repacking, so substitution leaves heavy atoms beyond C_β unmodelled and the refiner cannot resolve rotameric clashes (a backbone-dependent rotamer pass [8] is the natural next step; until then a substituted side chain contributes no DOPE terms beyond C_β). DOPE is an isotropic, distance-only potential—an orientation-dependent potential would resolve packing better at the cost of per-atom frames—and the restraint weight w_r is a tunable that sets the locality radius. For physically realistic, full-atom relaxation—side-chain optimisation and backbone minimisation under an all-atom force field—`tcren` defers to Rosetta FlexPepDock [7, 4] with the ref2015 energy function [1], invoked as an external process; the C++ kernel here is a fast knowledge-based pre-refinement, not a substitute for it.

REFERENCES

- [1] Rebecca F. Alford, Andrew Leaver-Fay, Jeliasko R. Jeliaskov, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13:3031–3048, 2017.
- [2] Vadim K. Karnaukhov, Dmitry S. Sheherbinin, Anton O. Chugunov, Dmitry M. Chudakov, Roman G. Efremov, Ivan V. Zvyagin, and Mikhail Shugay. Structure-based prediction of T cell receptor recognition of unseen epitopes using TCReN. *Nature Computational Science*, 4:510–521, 2024.
- [3] Scott Kirkpatrick, C. Daniel Gelatt, and Mario P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
- [4] Nir London, Barak Raveh, Eyal Cohen, Guy Fathi, and Ora Schueler-Furman. Rosetta FlexPepDock web server—high resolution modeling of peptide–protein interactions. *Nucleic Acids Research*, 39:W249–W253, 2011.
- [5] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.
- [6] Sanzo Miyazawa and Robert L. Jernigan. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256:623–644, 1996.
- [7] Barak Raveh, Nir London, and Ora Schueler-Furman. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins: Structure, Function, and Bioinformatics*, 78:2029–2040, 2010.
- [8] Maxim V. Shapovalov and Roland L. Dunbrack. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19:844–858, 2011.

- [9] Min-yi Shen and Andrej Sali. Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15:2507–2524, 2006.
- [10] Manfred J. Sippl. Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*, 213:859–883, 1990.