

BACKBONE-PRESERVING PEPTIDE SUBSTITUTION AND POTENTIAL-GUIDED REFINEMENT OF TCR–PEPTIDE–MHC STRUCTURES

Mikhail Shugay^{1*}

¹ISALGO laboratory

tcren technical appendix

Abstract

tcren ranks candidate epitopes for a fixed TCR–peptide–MHC structure by summing a residue-level pairwise statistical potential over the observed contact map [2]; because the potential depends only on residue identity and the closest-atom contact geometry, a candidate is scored *virtually*—its amino acids are threaded onto the native contacts with no atom moved. This appendix documents two coordinate-level extensions that the virtual score does not need but downstream re-docking does: (i) a backbone-preserving *peptide substitution* that produces the coordinates of a new equal-length epitope on the native scaffold, and (ii) a knowledge-based *potential-guided refinement* of the peptide pose—the DOPE atom-level distance-dependent statistical potential [9] plus a harmonic restraint to the input pose, sampled by rigid-body Metropolis Monte Carlo [5, 3] and implemented as a pybind11/C++ kernel. The refinement potential is deliberately *distinct* from the TCReN / Miyazawa–Jernigan potentials **tcren** *scores* epitopes with: refining a pose against the same quantity one then reports would be circular, so the pose is optimised under an independent, more physical energy. We give the energy function, prove that the restraint is necessary (without it the rigid-body minimiser drifts the peptide out of its pocket), and state the sampler and its cost. The method is knowledge-based, in the spirit of distance-dependent potentials of mean force [10, 6, 9]; it is *not* a physics force field. Physics-grade relaxation is delegated to Rosetta FlexPepDock [7, 4].

1. SCOPE

For a chain-typed complex **tcren** builds the inter-chain contact map—closest heavy-atom pairs at a 5 Å cutoff, reduced to one pair per residue pair—and scores a candidate peptide s on the TCR↔peptide interface as $\sum_{(i,j)} \phi(a_i, s_{\pi(j)})$, where ϕ is the TCReN statistical potential, a_i the fixed partner residue and π the contact’s peptide position [2]. This is a fixed-backbone, fixed-contact-set approximation: the native pose stands in for every candidate. It is fast (microseconds per peptide) and is the basis of all ranking in **tcren**. The two operations below relax that approximation when a candidate’s *geometry*—not just its score on the native contacts—is required, e.g. to seed re-docking, to relieve a clash introduced by substitution, or to compare poses across candidates.

*Correspondence: mikhail.shugay@gmail.com

2. 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 §3 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.

3. 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 (§1), 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.

4. 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. Shcherbinin, 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.

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

- [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 Flex-PepDock 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.