

A distribution-aware Bayesian logistic recognizer for real vs shuffled TCR–pMHC complexes

tcren supplementary note

1 Motivation

The Gaussian Bayesian network of the companion note treats *every* interface feature as Gaussian, which mis-specifies the many features that are not: 15 are non-negative integer **counts** (contact and interaction tallies, naturally Poisson) and the TCRdock `dock_torsion` is a **circular angle** (von Mises, wrapping $[0, 2\pi)$). We replace it with a **discriminative Bayesian logistic regression**, which handles these correctly: in a discriminative model the counts are covariates that enter linearly (their Poisson canonical form), so switching model already removes the Gaussian-on-counts mis-specification for free, and the one representation that genuinely matters — a circular predictor — is entered as its von Mises sufficient statistics $(\cos \theta, \sin \theta)$.

2 Feature encoding

Each feature enters the linear predictor through the canonical statistic of its natural family (Table 1); all columns are then standardised (train statistics only).

Table 1: Distribution-aware feature encoding.

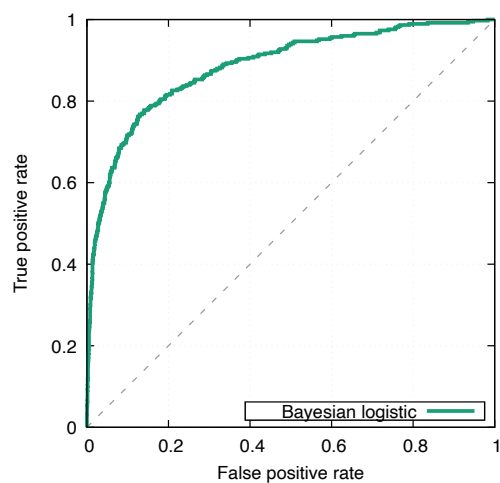
family	features	encoding
circular (von Mises)	<code>dock_torsion</code>	$\rightarrow (\cos \theta, \sin \theta)$
bounded ratio (Beta)	<code>chain_balance</code> $\in [0, \frac{1}{2}]$	$\rightarrow \text{logit}(2x)$
counts (Poisson)	<code>extent</code> , <code>n_contacts_*</code> , <code>ct_*</code>	linear (Poisson canonical)
unit-vector cos/sin	<code>dock_{tcr,mhc}_u{y,z}</code>	linear (already circular)
continuous	energies $F/\Delta F/e$, <code>burial</code> , <code>dock_d</code> , <code>pitch</code> , <code>crossing</code>	linear

3 Model

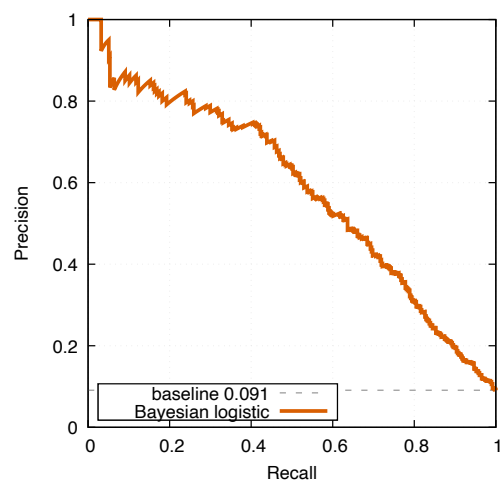
$y \sim \text{Bernoulli}(\sigma(\alpha + \mathbf{z}^\top \boldsymbol{\beta}))$ on the encoded, standardised design $\mathbf{z}$, with weakly-informative priors $\alpha \sim \mathcal{N}(0, 2.5)$, $\boldsymbol{\beta} \sim \mathcal{N}(0, 1)$ (a regularized-horseshoe variant is available for Bayesian sparsity). Fit with PyMC (NUTS, 2 chains). The posterior mean is frozen into `tcren` (`tcren.recognition.BayesianLogisticRecognizer`). Real (Oriented2026, $y=1$) vs Shuffled2026 $10\times$ decoys ($y=0$); 5-fold cross-validated out-of-fold predictions give the results below. The proper encoding lifts ROC–AUC to match the random forest with a fully interpretable linear model, above the Gaussian BN (0.865) and a plain logistic on raw features (0.870).

Table 2: Balanced cross-validated metrics (threshold at the Youden point for the point metrics).

metric	value
ROC-AUC	0.885
PR-AUC (baseline 0.091)	0.582
balanced accuracy	0.819
F_1	0.500
MCC	0.469



(a) ROC



(b) Precision-recall

Figure 1: Real-vs-shuffled discrimination, 5-fold cross-validated out-of-fold predictions.

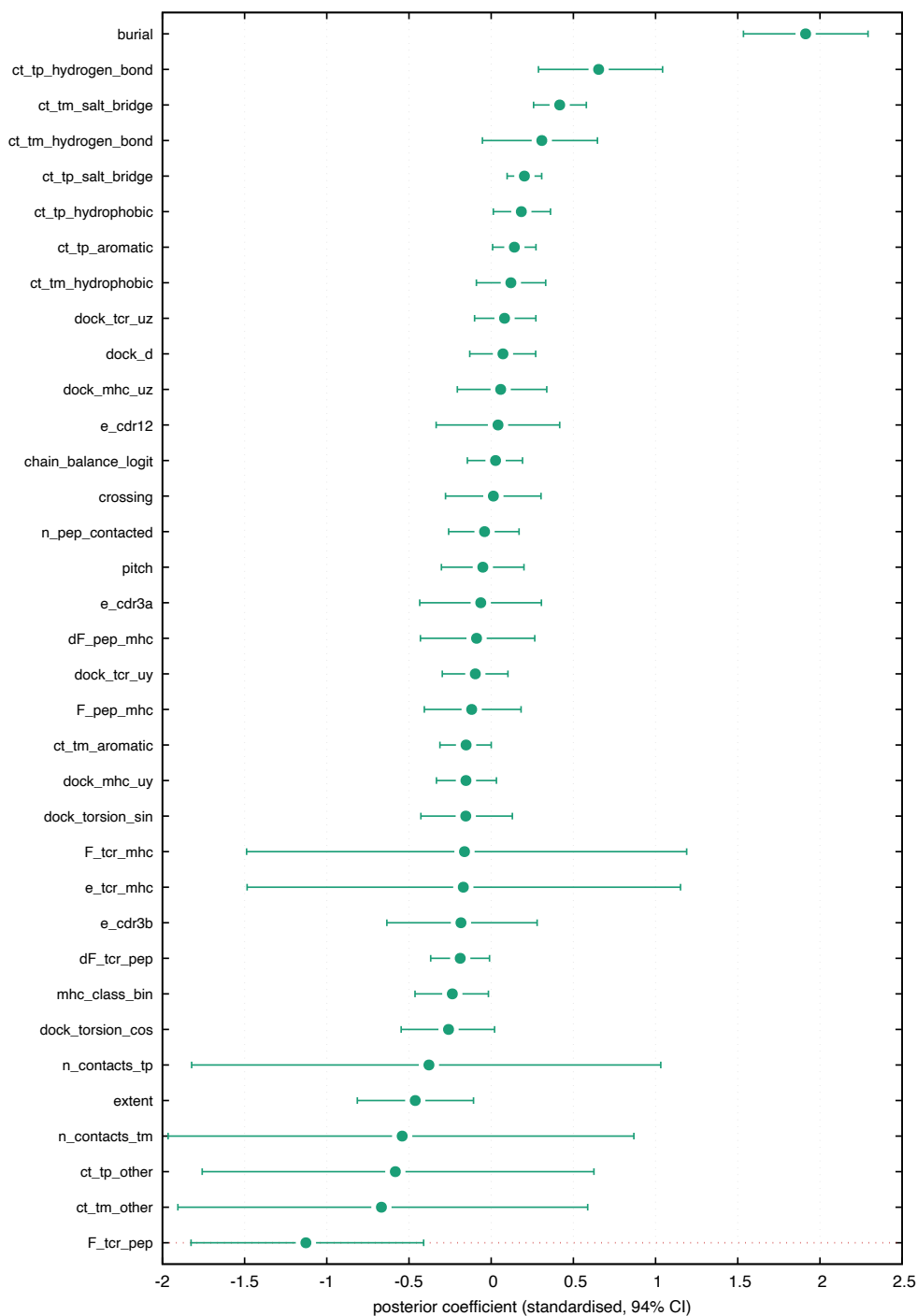


Figure 2: Posterior coefficient forest (standardised logistic weights, 94% credible intervals; sorted). The two **dock_torsion** terms are its cos/sin (von Mises) components; **chain_balance** enters as a logit. Coefficients whose interval excludes zero are the reliably-discriminative interface descriptors.