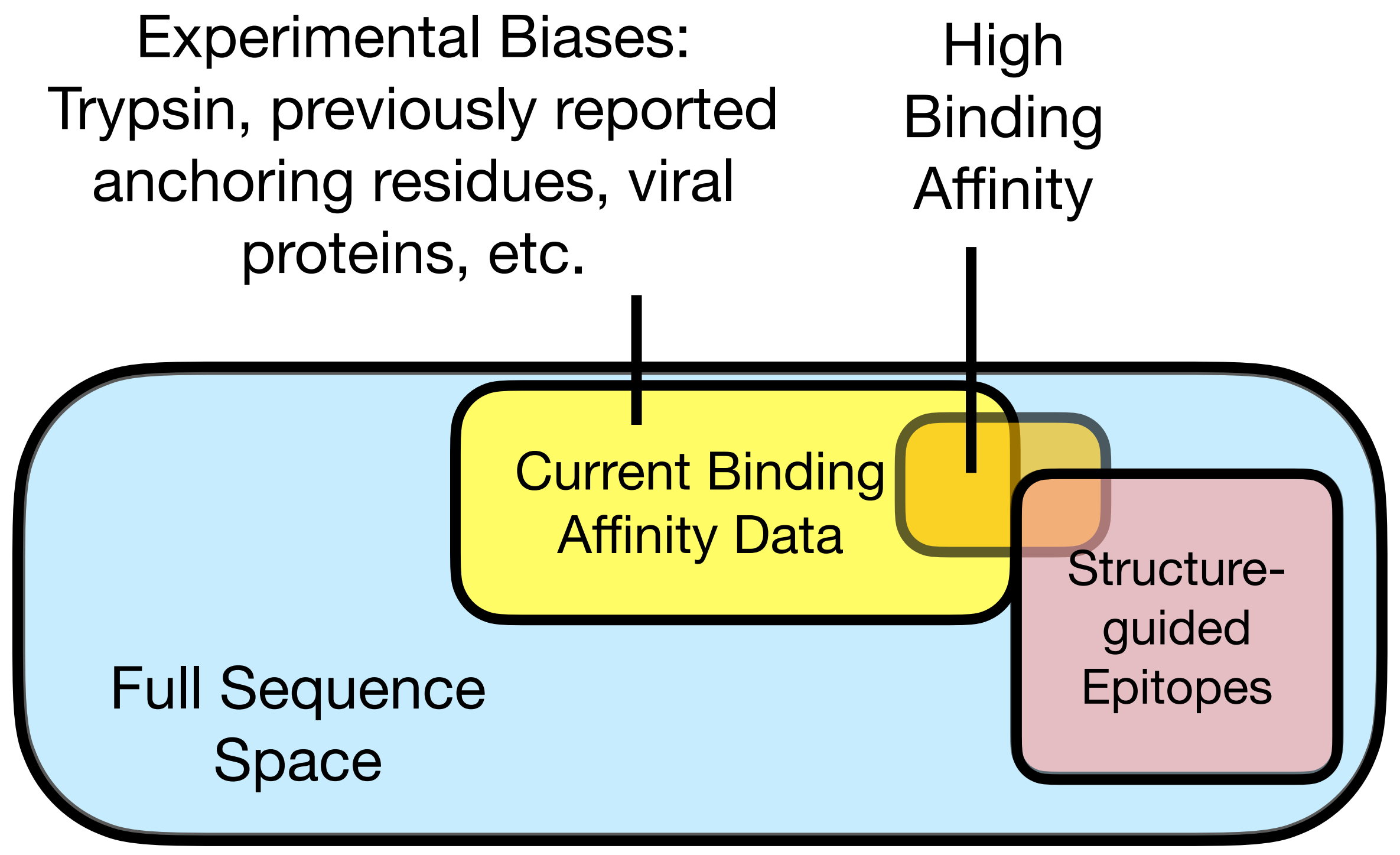
UC Berkeley
Center for
Computational Biology

Generation of structure-guided pMHC libraries with Diffusion Models

ICML
International Conference
On Machine Learning

Sergio Emilio Mares, Ariel Espinoza-Weinberger, Nilah M. Ioannidis

Abstract


Experimental Biases: Trypsin, previously reported anchoring residues, viral proteins, etc.

High Binding Affinity

Current Binding Affinity Data

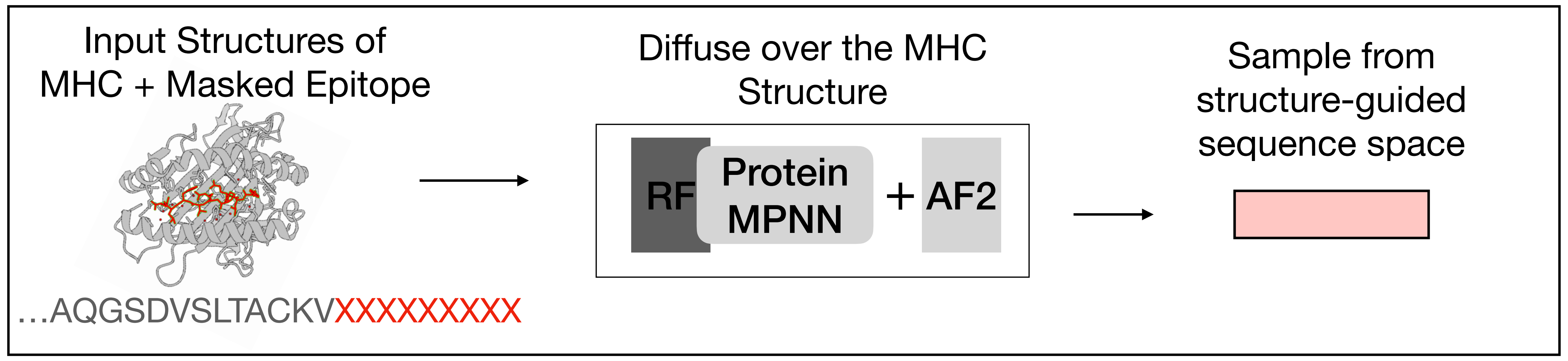
Structure-guided Epitopes

Full Sequence Space

- **Motivation**: Personalized vaccines and T-cell immunotherapies hinge on finding new peptide-MHC-I binders, yet MS and binding-assay datasets carry strong biochemical and protocol biases that hide large parts of sequence space.
- **Experimental biases called out**: Trypsin cleavage motifs, previously reported anchor residues, viral proteins, and high-affinity selection cutoffs.
- **Approach**: Build an unbiased benchmark by generating pMHC-I peptides with a diffusion pipeline conditioned on crystal-structure contact maps.
- **Results**: Library spans 20 clinically important HLA alleles, reproduces canonical anchor motifs, but is otherwise independent of known peptides, showing true structural generalization.
- **Take-away**: State-of-the-art sequence-based predictors miss many of these structurally valid designs, revealing allele-specific blind spots and providing a resource for fair model training and evaluation.

Methods

- Start with high-resolution pMHC-I structures and mask the peptide.
- **RFdiffusion** generates compatible peptide backbones while keeping MHC fixed.
- **ProteinMPNN** samples sequences that fit the backbone.
- **AlphaFold2** validates 3-D stability; only designs with peptide pLDDT > 0.8 are kept.
- Result: large, structure-guided peptide sequence space for each allele.


Input Structures of MHC + Masked Epitope

...AQGSDVSLTACKVXXXXXXXXX

Diffuse over the MHC Structure

RF

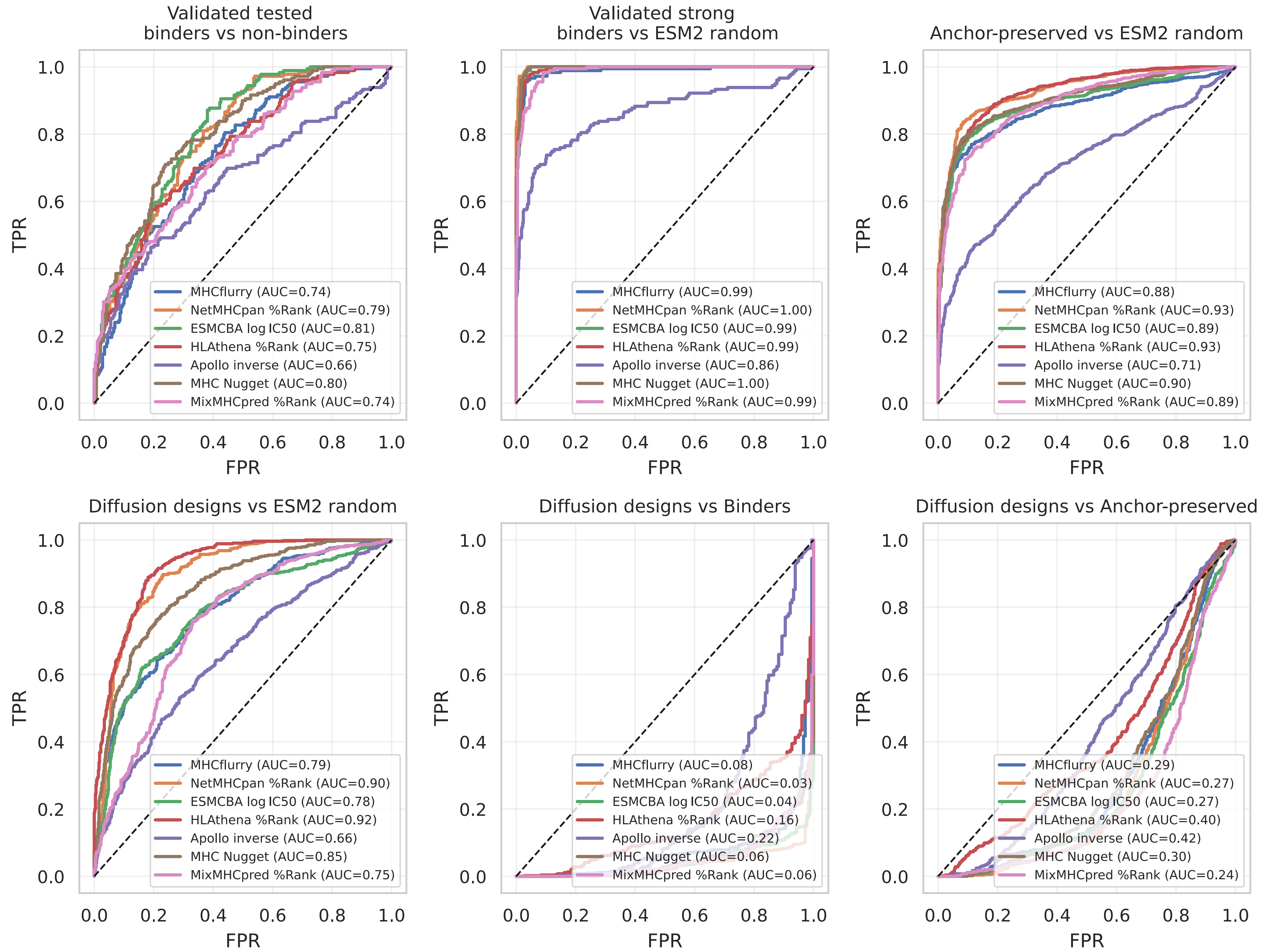
Protein MPNN

+ AF2

Sample from structure-guided sequence space

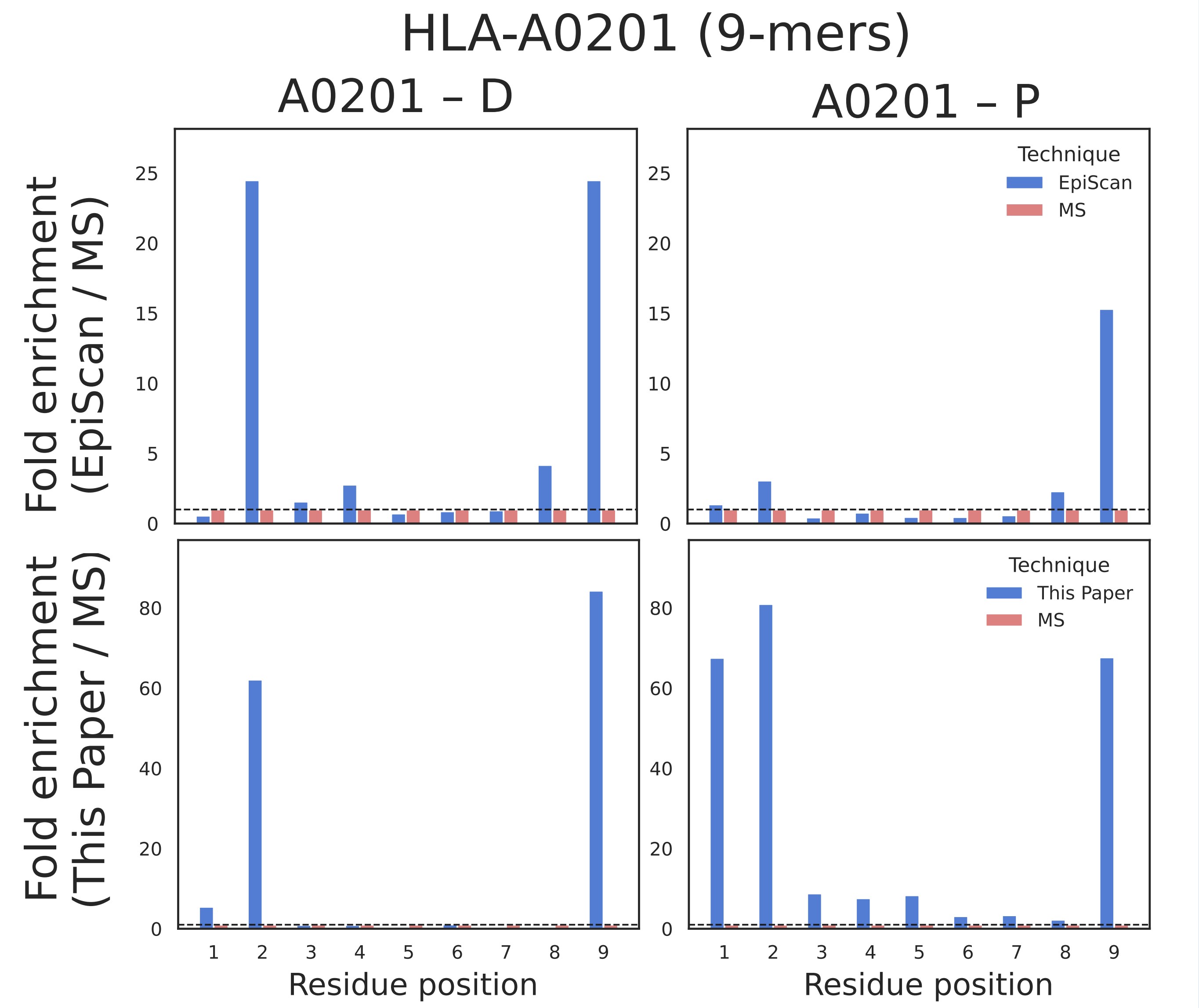
Current Binding Affinity Models overfit to Anchor residues

- **Current binding affinity models overfit to anchor residues**
- Benchmarked seven popular predictors on two stress tests.
  - **Anchor-preserved permutations**: keep P2 and PΩ residues, randomize the rest.
  - **Diffusion designs**: structurally validated peptides from the new library.
- Models give high scores to anchor-only controls (AUROC 0.71–0.93), showing they largely rely on anchors and ignore global context.
- When ranking diffusion designs versus validated strong binders, performance collapses (AUROC 0.06–0.22).
- **Conclusion: existing sequence-only models cannot recognize many structurally plausible binders that lie outside their training distribution.**



Consistency with experimental work

- Compared library motifs with **EpiScan**, a cell-based display assay that avoids MS biases.
- EpiScan tested >100 000 peptides but still sampled a narrow sequence slice (<400 A*02:01 binders, <1500 B*57:01).
- Diffusion library fills that gap by producing tens of thousands of anchor-compatible peptides that explore poorly sampled regions, broadening diversity and reducing bias.
- Our work follows closely the distribution captured by coupled biological work.



Conclusion and Future Work

- Structure-guided diffusion generation exposes critical weaknesses in current pMHC-I prediction models and offers an unbiased benchmark of ~10^5 peptides.
- Workflow can serve both as challenging test data and as supplemental training material to improve generalization.
- Next steps:
  - Expand experimental validation beyond the four alleles covered by EpiScan.
  - Integrate TCR-peptide interactions to assess immunogenicity, not just binding.
  - Use the benchmark to retrain or fine-tune sequence-and-structure-aware models, closing the observed blind spots.