HLA-ProtBERT: AI-Powered HLA Analysis

NMDP Research Team

National Marrow Donor Program

March 2025

In Brief

- AI model (ProtBERT) for HLA protein analysis
- Captures functional relationships traditional methods miss
- Analyzed 17,109 Class I alleles in 81.57 seconds
- Reveals biological patterns matching HLA classification

Benefits

- **Mgmt:** Efficient analysis pipeline
- **Clinical:** Better matching insights
- **Tech:** Novel functional view

Overview

1. HLA Analysis Challenge
2. AI-Based Approach
3. Key Findings
4. Applications

Content for:

- Management
- Clinical Scientists
- Bioinformaticians

HLA Complexity

- Thousands of alleles with critical immune function
- Sequence similarity is not equivalent to functional similarity
- Growing database makes manual analysis impossible

Clinical: Better transplant matching
Mgmt: Efficiency gains
Tech: Novel computational approach

Key Takeaway

HLA complexity requires advanced methods beyond simple sequence comparison.

ProtBERT: AI for Proteins

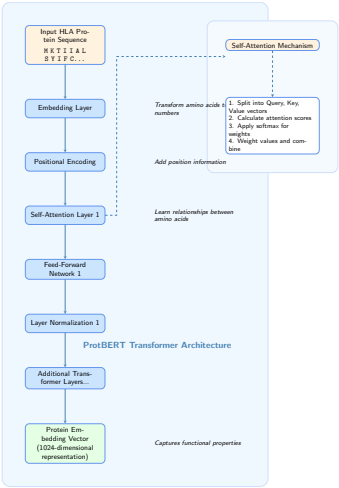
- AI model trained on 106+ million proteins
- Pre-trained for protein language understanding
- Captures functional/structural properties
- 16M parameters vs. billions in larger models

Clinical: Finds functionally similar HLAs
Mgmt: Leverages existing AI advances
Tech: Transfer learning from vast protein data

Key Takeaway

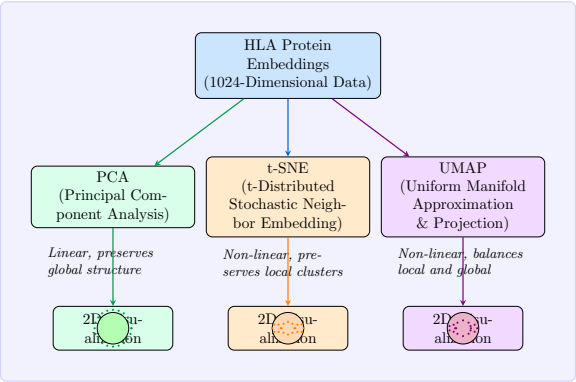
ProtBERT understands protein "language" beyond simple sequence similarity.

How It Works



- Processes HLA sequence while considering all amino acid relationships
- Multiple layers extract increasingly complex patterns

From Complex to Simple



Dimensionality Reduction Techniques

PCA
Global overview

t-SNE
Local clusters

UMAP
Balanced view

Analysis Overview

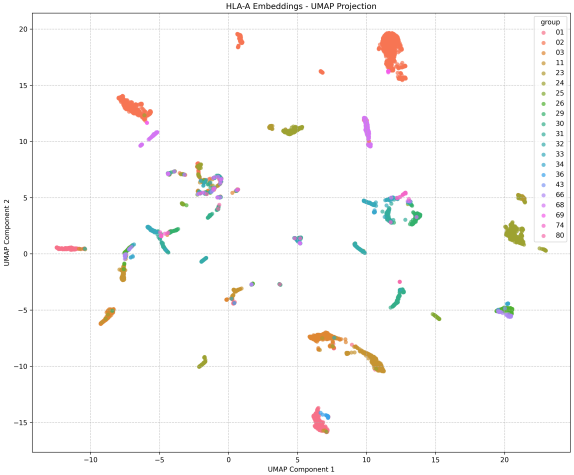
- 17,109 Class I alleles analyzed:
  - 5,432 HLA-A
  - 6,526 HLA-B
  - 5,151 HLA-C
- Total processing: 81.57 sec (0.005 sec/allele)
- Clear clustering by functional groups

Clinical: Reveals functional similarities

Mgmt: Demonstrates efficiency at scale

Tech: Effective protein embedding

HLA-A Visualization



- Clear separation between major allele families
- A*01, A*02, A*03 groups form discrete clusters

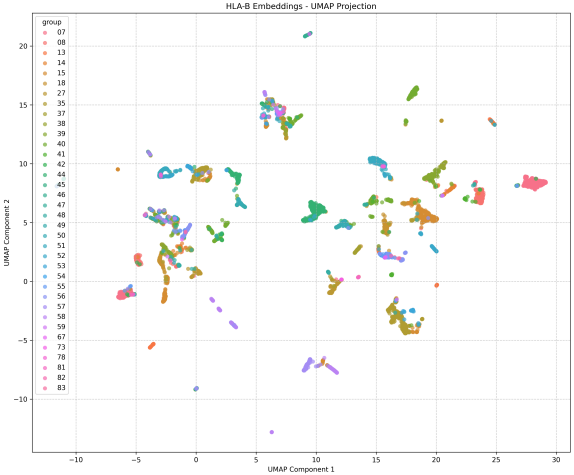
HLA-A: Multiple Perspectives

Figure: PCA (left), t-SNE (middle), UMAP (right)

Clinical: Helps identify potential cross-reactive epitopes

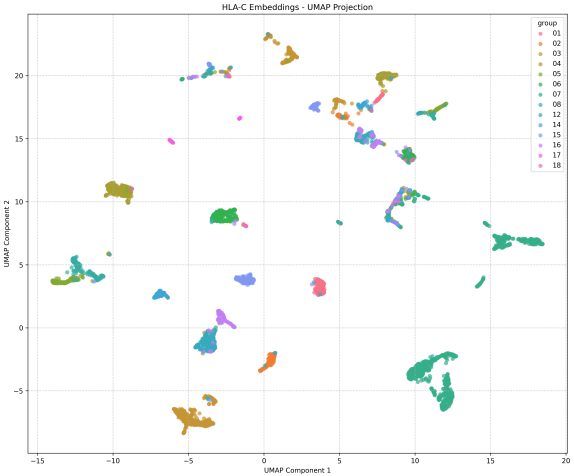
Tech: Different methods reveal different relationship aspects

HLA-B: Most Diverse



- Most diverse locus (6,526 alleles)
- Complex clustering reflects greater polymorphism

HLA-C: Distinct Patterns



- More compact clusters than A and B loci
- Clearer separation between major groups

Current Applications

Clinical

- **Clinical:** Better matching prediction
- **Clinical:** Novel allele classification
- **Clinical:** Cross-reactivity assessment

Research & Technical

- **Tech:** Structure-function modeling
- **Mgmt:** Efficient analysis workflow
- **Tech:** Evolutionary analysis

Understanding functional relationships between HLA alleles has direct applications in transplantation, disease association, and drug response prediction.

Future Directions

Technical Advancements

- HLA-specific training
- Structural information integration
- Multi-modal models

Data Integration

- Clinical outcome correlation
- Population genetics
- Cross-species analysis

Potential Impact

- **Clinical:** Improved virtual crossmatching
- **Mgmt:** Better matching algorithms
- **Tech:** Novel evolutionary insights

Key Takeaways

1. AI embeddings reveal functional HLA relationships
2. 17,109 alleles analyzed in 81.57 seconds
3. Clear functional clustering observed
4. Multiple visualization techniques provide complementary views
5. Applications span clinical, research, and data domains

Clinical: Explore matching prediction applications

Mgmt: Consider workflow integration

Tech: Explore additional models

References I

Technical Details

Model Specifications

- 30 transformer layers
- 16 attention heads
- 1024-dimensional embeddings
- 16M parameters (vs. billions)
- Pre-trained on 106M proteins

Dimensionality Reduction Parameters

- **UMAP:** n_neighbors=15, min_dist=0.1
- **t-SNE:** perplexity=30, learning_rate=200
- **PCA:** n_components=2

Implementation

- Hugging Face Transformers
- CUDA GPU acceleration
- Embedding caching
- Batch processing (batch_size=8)