SnapBind: Lightweight CNN for Protein Binding Pocket Prediction

David Z Barth, Alexander Haas, Elias M Bruss

Abstract:

We present *SnapBind*, a lightweight CNN (0.5M parameters) for predicting protein druggability and binding pockets from amino acid sequences. Using 100k protein-ligand pairs from BindingDB, our model enables local deployment for high-throughput screening without expensive computational infrastructure.

1 Methods

Dataset: 100k protein-ligand pairs from BindingDB plus 20k negative controls (Antibodies, structural proteins, nucleases). Binding sites defined by 5Å cutoff with binary residue annotation.

Architecture: FastCNNBindingPredictor with 64-dim embeddings, 128-dim hidden layers, 0.1 dropout, handling 20-300 residue sequences.

Training: Batch size 32, AdamW (lr=2e-4), focal loss (α=1, γ=2) for class imbalance, early stopping, Apple Silicon GPU acceleration.

2 Progressive Design

Three-pass architecture: (1) Sequence-based CNN for druggability, (2) ESM embeddings for evolutionary context, (3) SMILES integration for small-molecule ligand-specific predictions.

3 Results

Efficiency: 0.5M parameters enable consumer hardware deployment with rapid batch processing and local execution.

Performance: Stable training with effective class imbalance handling through focal loss and positive class weighting.

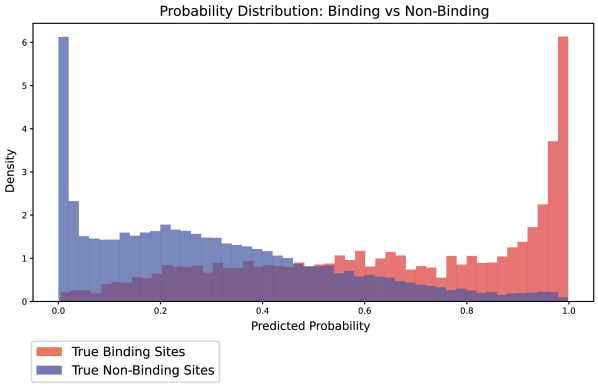

Figure 1: Score distribution for ground truth.

4 Applications

High-throughput protein screening, binding site localization, academic accessibility without proprietary platforms, local computational pipeline integration.

5 Conclusion

Our lightweight CNN democratizes binding pocket prediction by balancing performance with computational efficiency, enabling researchers to perform drug discovery screening without extensive infrastructure requirements.