SnapBind: Lightweight CNN for Protein Binding Pocket Prediction

David Z Barth, Alexander Haas, Elias M Bruss

1 Challenge Tackled

Accurate protein AI models are costly in time and computational resources. *SnapBind* addresses these challenges by providing a lightweight architecture that maintains high performance while being efficient enough for local deployment. We provide high-throughput protein screening, binding site localization, academic accessibility without proprietary platforms and a local computational pipeline integration.

2 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.

3 What Worked Well

Data Acquisition and the Training Pipeline

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

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.

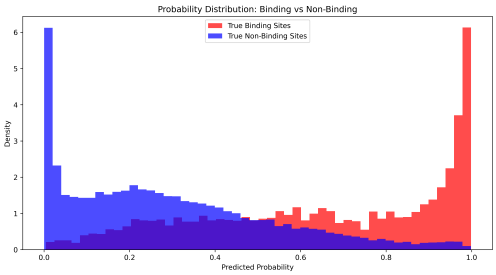

Figure 1: Score distribution for ground truth.

4 What Was Challenging

Balancing model complexity with performance was a key challenge. While we aimed for a lightweight architecture, ensuring sufficient representational power to capture intricate protein-ligand interactions required careful design choices.

Front-end integration paired with a meaningful design philosophy required a lot of iterative refinement to achieve a user-friendly experience.

5 Time Spent

- 0 - 1 h: Deciding on specific problem and first approach
- 1 - 5 h: Shared responsibilities - Data processing, different model architectures and benchmarking
- 5 - 10 h: Shared responsibilities - Model training, evaluation, and hyperparameter tuning, first steps towards integration
- 10 - 15 h: Shared responsibilities - Front-end integration, user interface design, and presentation

6 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.