GNN-Diffusion Framework for Structure-Based Drug Discovery of Undruggable NR4A Transcription Factors

Yanting Xing and Bocheng Wu

Department of Chemistry & Biochemistry, Texas State University

Introduction

Transcription factors regulate gene expression and are implicated in virtually every cancer type, yet many lack well-defined binding pockets accessible to conventional small-molecule drugs, classifying them as "undruggable" by traditional approaches. NR4A nuclear receptors [1], comprising NR4A1 (Nur77), NR4A2 (Nurr1), and NR4A3 (NOR1), exemplify this challenge. These orphan nuclear receptors act as critical oncogenic drivers in leukemia, lung cancer, and colorectal cancer, promoting tumor cell survival and suppressing apoptosis. Despite their therapeutic relevance, no approved small-molecule drugs target NR4A, because their ligand-binding domains are filled with bulky hydrophobic residues that leave no accessible cavity for traditional drug binding.

Classical structure-based drug discovery fails for such targets because the protein-protein interaction surfaces that NR4A uses are large, flat, and dynamic, incompatible with the rigid pocket geometry assumed by conventional docking. A fundamentally different computational strategy is needed: one that can learn the complex 3D geometry of NR4A's interaction surface and generate novel molecules shaped to complement it.

We address this by integrating two complementary approaches. First, high-throughput AutoDock virtual screening of a 64,000-compound library identifies scaffolds with demonstrated binding affinity to NR4A, yielding four priority scaffolds with over 100 high-affinity binders each. Second, UniMoMo [2], a state-of-the-art SE(3)-equivariant graph neural network diffusion model, is fine-tuned on these scaffold-specific hits to learn the chemical and geometric profile of validated NR4A binders, then used to generate novel candidate molecules directly conditioned on the NR4A pocket geometry. This pipeline leverages TACC Lonestar6 high-performance computing and provides a reusable framework extensible to other undruggable transcription factor targets.

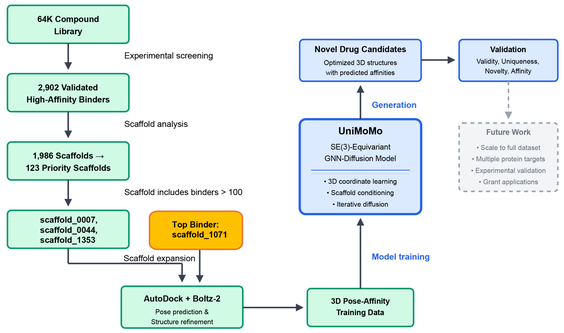
References:

[1] Safe S and Karki K. The Paradoxical Roles of Orphan Nuclear Receptor 4A (NR4A) in Cancer. *Mol Cancer Res.* 2021;19(2):180-191.

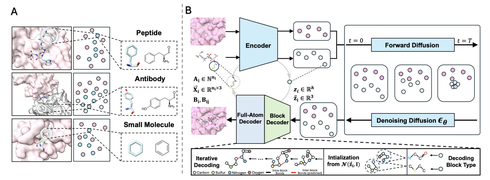
[2]Kong et al. UniMoMo: Unified Generative Modeling of 3D Molecules for De Novo Binder Design. ICML 2025.

Approach & Pipeline

SE(3)-Equivariant GNN-Diffusion Framework



Unified Generative Modeling of 3D Molecules (UniMoMo)

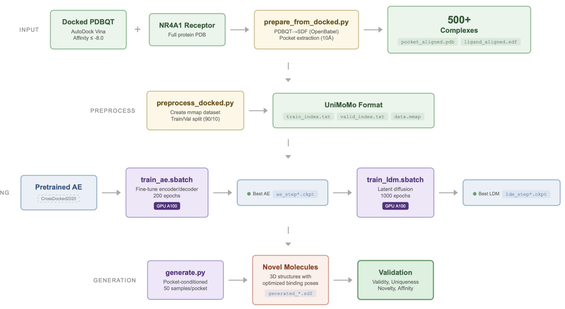


Unified Block Representation: Atoms grouped into fragments with a shared vocabulary across small molecules, peptides, and antibodies

SE(3)-Equivariant Architecture: Handles 3D molecular geometry regardless of orientation-critical for pocket-ligand modeling

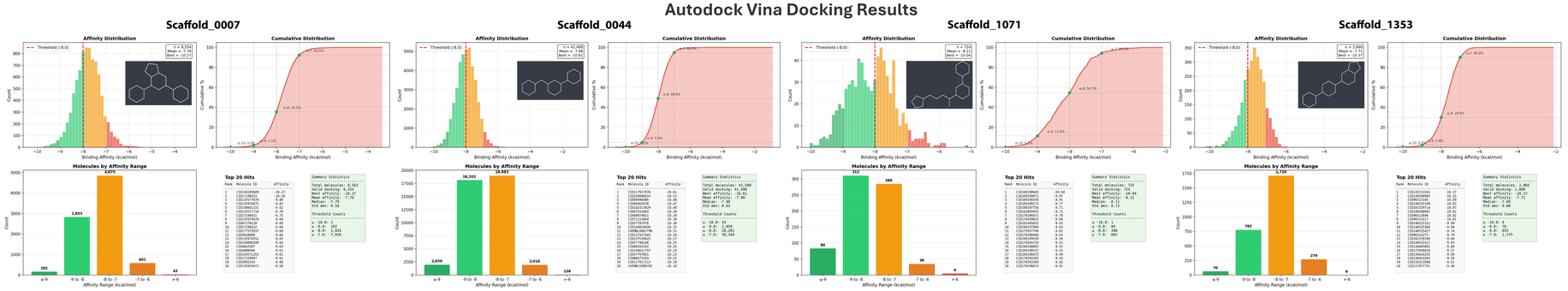
Fine-Tuned on NR4A Data: Adapted from a pretrained model on scaffold-specific AutoDock hits to generate novel NR4A binders

UniMoMo Fine-tuning Pipeline

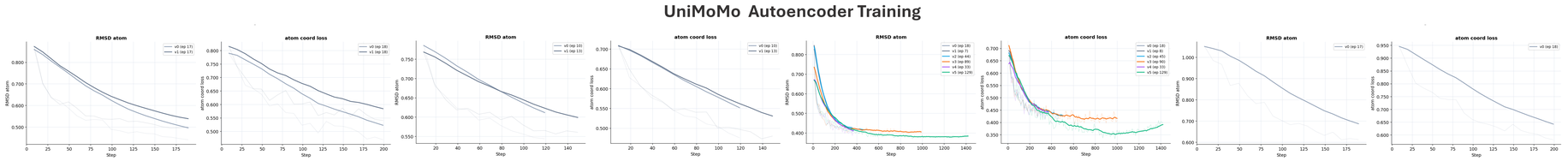


Results

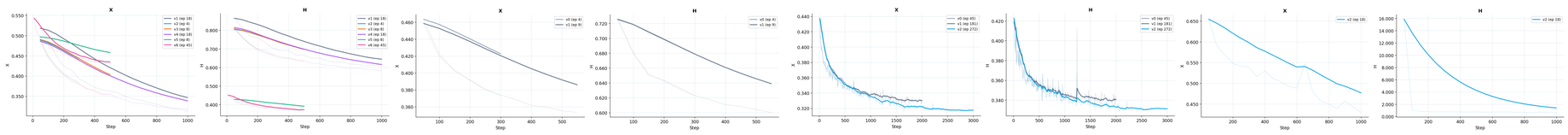
Autodock Vina Docking Results



UniMoMo Autoencoder Training



UniMoMo Latent Diffusion Model Training



- **AutoDock Vina Docking Results:** Scaffold expansion and AutoDock Vina docking against NR4A. Molecules with scores below -8 kcal/mol selected as UniMoMo training data.
- **UniMoMo Autoencoder Training:** CosineAnnealingLR decay (1e-4 to 1e-6) achieved best RMSD_atom of 0.378 Å for scaffold_1071, a 15% improvement over baseline.
- **UniMoMo Latent Diffusion Model Training:** Scaffold_1071 achieved coordinate loss of 0.318 after 3000 epochs with cosine decay, best across all runs.

Conclusion

We developed a computational pipeline integrating AutoDock virtual screening with UniMoMo fine-tuning for de novo binder design targeting undruggable NR4A transcription factors. Screening a 64,000-compound library identified four priority scaffolds (0007, 0044, 1071, 1353) with over 100 high-affinity binders each. Fine-tuning UniMoMo with an optimized CosineAnnealingLR schedule achieved autoencoder RMSD_atom of 0.378 Å, a 15% improvement over baseline, with the latent diffusion model converging to coordinate loss of 0.318. Novel NR4A binder candidates are currently being generated for experimental validation. This pipeline establishes a reusable framework for structure-based drug discovery extensible to other undruggable transcription factor targets.

Acknowledgements

This work used the Lonestar6 supercomputer at the Texas Advanced Computing Center (TACC) and the LEAP2 High Performance Computing cluster at Texas State University. The authors thank the Translational Health Research Center at Texas State University for supporting this research.

TEXAS STATE UNIVERSITY®

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM