# Computational Modeling of Protein–Ligand Interactions

## A Machine Learning Approach to Binding Affinity Prediction

**Marcus J. Chen**

A Dissertation

in

Bioengineering

Presented to the Faculties of the University of Pennsylvania

in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2025

**Supervisor of Dissertation**

Dr. Elena M. Vasquez

Associate Professor of Bioengineering

School of Engineering and Applied Science

**Graduate Group Chairperson**

Dr. Robert K. Nakamura

Professor of Bioengineering

# Abstract

The accurate prediction of protein–ligand binding affinities remains one of the central challenges in computational drug discovery. Despite decades of methodological advances in molecular docking and scoring functions, the quantitative reliability of computational predictions often falls short of experimental accuracy. This dissertation presents a novel framework that integrates physics-based molecular simulations with modern machine learning architectures to achieve state-of-the-art binding affinity prediction.

We introduce the Ensemble Graph Attention Network (EGAN), a deep learning architecture that operates directly on three-dimensional protein–ligand complex representations. Unlike conventional methods that rely on hand-crafted geometric features, EGAN learns interaction fingerprints through a hierarchical attention mechanism that captures both local atomic environments and global topological patterns. The model is trained on a curated dataset of 14,000 experimentally measured binding constants spanning seven major protein families.

Our results demonstrate a Pearson correlation coefficient of $R = 0.89$ on independent test sets, representing a 23% improvement over the best-performing classical scoring functions. We further show that EGAN generalizes effectively to challenging targets including intrinsically disordered proteins and metalloenzymes, domains where traditional docking approaches consistently underperform. Ablation studies confirm that the graph attention mechanism contributes significantly to model performance, with attention weights aligning with experimentally validated binding hot spots.

This work establishes a foundation for integrating deep learning into structure-based drug design pipelines and suggests that hybrid physics-informed neural architectures represent a promising direction for the next generation of computational drug discovery tools.

# Acknowledgments

I wish to express my deepest gratitude to my advisor, Dr. Elena Vasquez, whose unwavering support, intellectual rigor, and patience made this work possible. Her ability to see connections across disciplines — from statistical mechanics to deep learning — shaped the interdisciplinary character of this research.

I thank the members of my dissertation committee: Dr. Robert Nakamura, Dr. Sarah Okonkwo, and Dr. David Thornton, for their insightful feedback and generous mentorship throughout my doctoral studies. I am also grateful to the members of the Vasquez Laboratory, particularly Dr. Priya Sharma and James O'Brien, for countless discussions that refined the ideas presented here.

This research was supported by a Graduate Research Fellowship from the National Science Foundation and by computational resources provided by the Penn Institute for Computational Science. I thank my collaborators at Nimbus Therapeutics for providing access to proprietary binding affinity datasets that enriched the benchmarking studies in Chapter 4.

Finally, I owe an immeasurable debt to my family — my parents, Lin and Wei Chen, and my sister, Dr. Julia Chen — for their love, encouragement, and steadfast belief in the value of this pursuit. Any errors or omissions remain my own.

# Dedication

*To my parents, who taught me to ask why.*

# Contents

# List of Figures

# List of Tables

# CHAPTER 1

# Introduction

## 1.1 Motivation and Significance

The discovery and development of new pharmaceutical compounds remains an extraordinarily resource-intensive endeavor, with the average cost of bringing a single drug to market exceeding $2.6 billion and timelines stretching beyond a decade [1]. At the heart of this challenge lies a fundamental molecular recognition problem: predicting how strongly a small molecule ligand will bind to its target protein. Accurate binding affinity predictions can dramatically reduce the experimental burden of early-stage drug discovery by enabling computational prioritization of candidate molecules [2].

Structure-based drug design (SBDD) leverages the three-dimensional structures of protein targets to guide the optimization of lead compounds. Molecular docking — the computational prediction of ligand binding poses and energies — has served as the workhorse of SBDD for over three decades [3]. Classical scoring functions, which estimate binding free energy through simplified physical models incorporating terms for van der Waals interactions, electrostatics, hydrogen bonding, and desolvation, offer computational efficiency but often sacrifice quantitative accuracy.

**Key Insight**

The central limitation of classical scoring functions is their reliance on linear additive energy models. Protein–ligand binding is fundamentally governed by non-additive effects including conformational entropy, solvent reorganization, and cooperative hydrogen bond networks — phenomena that linear models cannot adequately capture.

## 1.2 The Binding Affinity Prediction Problem

Formally, the binding affinity prediction problem can be stated as follows: given a protein structure $\mathcal{P}$ and a ligand structure $\mathcal{L}$, predict the equilibrium dissociation constant $K_d$ or the binding free energy $\Delta G_{\text{bind}}$, related by the thermodynamic relationship:

$$\Delta G_{\text{bind}} = -RT \ln K_d = \Delta H_{\text{bind}} - T\Delta S_{\text{bind}} \tag{1.1}$$

where $R$ is the gas constant, $T$ is the absolute temperature, $\Delta H_{\text{bind}}$ is the binding enthalpy, and $\Delta S_{\text{bind}}$ is the binding entropy. The challenge lies in the fact that $\Delta G_{\text{bind}}$ is typically on the order of $-5$ to $-15$ kcal/mol, representing a small difference between large opposing energetic contributions [4].

## 1.3 Contributions of This Dissertation

This dissertation makes four primary contributions to the field of computational binding affinity prediction:

(1) The Ensemble Graph Attention Network (EGAN), a novel deep learning architecture that operates on protein–ligand complex graphs with physics-informed attention mechanisms.

(2) A comprehensive benchmarking framework for evaluating binding affinity predictors across diverse protein families, including intrinsically disordered regions and metalloproteins.

(3) An integrated computational pipeline that combines molecular dynamics sampling with graph neural network predictions, achieving state-of-the-art accuracy on standard benchmarks.

(4) A systematic analysis of attention weight distributions that reveals interpretable patterns corresponding to experimentally validated binding hot spots.

## 1.4 Organization of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 provides essential background on molecular recognition, classical scoring functions, and the application of graph neural networks to molecular property prediction. Chapter 3 describes the EGAN architecture in detail, including the ensemble sampling strategy and the physics-informed attention mechanism. Chapter 4 presents comprehensive benchmarking results across multiple protein families and compares EGAN performance against established methods. Chapter 5 discusses implications for drug discovery, limitations of the current approach, and directions for future research.

# CHAPTER 2

# Background and Related Work

## 2.1 Classical Approaches to Binding Affinity Prediction

Classical scoring functions can be broadly categorized into three families: force-field-based, empirical, and knowledge-based [5]. Force-field-based methods compute non-bonded interaction energies using molecular mechanics potentials, typically with an AMBER or CHARMM parameterization. Empirical scoring functions fit weighted sums of physically motivated terms to experimental binding data. Knowledge-based potentials derive statistical pair potentials from the frequency of interatomic contacts observed in structural databases.

Table 2.1 summarizes the key characteristics of representative scoring functions from each family.

Table 2.1: Representative scoring functions across three classical families.

| Function | Core Methodology | Type | Year |
|---|---|---|---|
| AutoDock Vina | Hybrid empirical/knowledge-based scoring | Empirical | 2010 |
| GlideScore SP | Force-field with empirical hydrophobic term | Force-field | 2004 |
| DrugScore | Distance-dependent pair potentials from PDB statistics | Knowledge-based | 2006 |
| X-Score | Atomic-level empirical scoring with desolvation | Empirical | 2002 |
| PMF04 | Statistical potential with protein-specific corrections | Knowledge-based | 2004 |
| GoldScore | Genetic algorithm-optimized hydrogen bonding terms | Empirical | 1997 |
| DSX | Distance-scaled finite ideal-gas reference state | Knowledge-based | 2011 |

## 2.2 Deep Learning in Molecular Property Prediction

Recent years have witnessed an explosion of interest in applying deep learning to molecular problems. Graph neural networks (GNNs) have emerged as a particularly natural framework for modeling molecular structures, where atoms represent nodes and bonds represent edges [6]. The message-passing paradigm, in which node representations are iteratively updated by aggregating information from neighboring nodes, provides a flexible inductive bias well-suited to chemical systems.

### 2.2.1 Graph Neural Network Architectures

The fundamental operation in a message-passing neural network can be expressed as:

$$\mathbf{h}_v^{(k+1)} = U\left(\mathbf{h}_v^{(k)}, \bigoplus_{u \in \mathcal{N}(v)} M\left(\mathbf{h}_v^{(k)}, \mathbf{h}_u^{(k)}, \mathbf{e}_{vu}\right)\right) \tag{2.1}$$

where $\mathbf{h}_v^{(k)}$ is the feature vector of node $v$ at layer $k$, $\mathcal{N}(v)$ denotes the neighbors of $v$, $\mathbf{e}_{vu}$ is the edge feature between nodes $v$ and $u$, $M$ is the message function, $\bigoplus$ is a permutation-invariant aggregation operator (typically sum, mean, or max), and $U$ is the update function. Both $M$ and $U$ are implemented as learnable neural network layers.

### 2.2.2 Application to Binding Affinity Prediction

Several groups have applied GNN architectures to binding affinity prediction with promising results. The PDBbind benchmark dataset [7] provides a standardized evaluation framework, containing approximately 19,000 protein–ligand complexes with experimentally measured binding constants. State-of-the-art GNN-based methods achieve Pearson correlation coefficients in the range of $R = 0.80$ to 0.84 on PDBbind core sets, representing substantial improvements over classical scoring functions.

**Limitations of Current Approaches**

Most existing deep learning methods for binding affinity prediction operate exclusively on ligand structures, treating the protein environment implicitly or through pre-computed descriptors. This simplification discards critical information about protein conformational dynamics, solvent effects, and long-range electrostatic interactions that are known to modulate binding thermodynamics.

## 2.3 Ensemble Methods in Computational Chemistry

Molecular recognition occurs on rugged free energy landscapes characterized by multiple local minima separated by appreciable kinetic barriers. Single-structure representations — whether from X-ray crystallography, NMR spectroscopy, or single docking poses — capture only one snapshot of this dynamic equilibrium. Ensemble methods

address this limitation by sampling multiple conformations and weighting predictions by their thermodynamic populations [8].

The EGAN architecture introduced in this work extends the ensemble paradigm to graph neural networks. Rather than operating on a single protein–ligand complex graph, EGAN processes an ensemble of conformations sampled from molecular dynamics trajectories, with an attention mechanism that learns to weight conformations according to their relevance for binding affinity prediction. This approach bridges the gap between the statistical mechanical foundations of binding thermodynamics and the representational power of modern deep learning.

# Bibliography

[1] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.*, 47:20–33, 2016.

[2] J. Lyu, S. Wang, T. E. Balius, et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

[3] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.*, 3(11):935–949, 2004.

[4] D. L. Mobley and K. A. Dill. Binding of small-molecule ligands to proteins: "What you see" is not always "what you get". *Structure*, 17(4):489–498, 2010.

[5] J. Liu and R. Wang. Classification of current scoring functions. *J. Chem. Inf. Model.*, 55(3):475–482, 2018.

[6] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1263–1272, 2017.

[7] R. Wang, X. Fang, Y. Lu, and S. Wang. The PDBbind database: Collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.*, 47(12):2977–2980, 2004.

[8] R. E. Amaro, J. Baudry, J. Chodera, et al. Ensemble docking in drug discovery. *Biophys. J.*, 114(10):2271–2278, 2018.