Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

PHYSICS-INFORMED GRAPH NEURAL NETWORKS FOR BINDING FREE-ENERGY PREDICTION FROM DOCKING AND DYNAMICS DATA

Siti Rahman[1]*, Ahmad Zaki[1], Nurul Huda[2], Amir Faisal[1], Lim Wei[2]

1. ***Department of Pharmaceutical AI Engineering, Faculty of Pharmacy, University of Malaya, Kuala Lumpur, Malaysia.***
2. ***Department of Computational Drug Analytics, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru, Malaysia.***

ARTICLE INFO

Received:
20 November 2025
Received in revised form:
10 February 2026
Accepted:
16 February 2026
Available online:
28 February 2026

Keywords: Physics-informed learning, Graph neural networks, Binding free energy, Molecular docking, Molecular dynamics, Protein–ligand complexes

ABSTRACT

Reliable prediction of protein–ligand binding free energy is central to hit-to-lead optimization because affinity guides prioritization, analogue design, and resource allocation. Classical docking is fast and scalable, but its simplified scoring functions often struggle across chemically diverse ligand series. Purely data-driven models can learn dataset-specific chemical patterns that may not reflect realistic molecular recognition. Conversely, purely physics-based approaches can miss high-dimensional statistical regularities present in curated structural and affinity datasets. This article proposes a physics-informed graph neural network for predicting absolute binding free energy from protein–ligand complex structures. The model integrates docking scores, molecular-dynamics-derived interaction descriptors, and force-field-inspired energy terms within a unified graph representation. The proposed architecture uses three-dimensional, SE(3)-aware message passing over protein–ligand complex graphs. Atom and residue features are augmented with docking-derived pose information, molecular dynamics descriptors, and pre-computed electrostatic and van der Waals interaction terms.

Conceptually, such a model would be expected to provide more physically consistent affinity estimates than purely empirical scoring functions. It could also support interpretable atom-level and residue-level interaction analysis for medicinal chemistry decision-making. Physics-informed graph learning offers a principled bridge between rigorous molecular thermodynamics and flexible deep representation learning. This framework provides a scalable model-oriented route toward more reliable in silico affinity optimization.

*This is an **open-access** article distributed under the terms of the Creative Commons Attribution-Non Commercial-Share Alike 4.0 License, which allows others to remix, and build upon the work non commercially.*

To Cite This Article: Rahman S, Zaki A, Huda N, Faisal A, Wei L. Physics-Informed Graph Neural Networks for Binding Free-Energy Prediction from Docking and Dynamics Data. Pharmacophore. 2026;17(1):101-10. https://doi.org/10.51847/jGMqmzd7Dk

Introduction

Binding free energy is a central quantity in structure-based drug design because it connects molecular recognition to compound prioritization, lead optimization, and selectivity engineering. High-throughput docking provides an efficient first-pass strategy for ranking large chemical libraries, but classical scoring functions often simplify solvation, entropic effects, and induced fit in ways that limit transferability across targets and chemotypes. Deep structural models such as KDEEP demonstrated that protein–ligand complexes can be represented directly from three-dimensional grids for affinity prediction [1], while DeepDTA-style and complex-based neural approaches showed how supervised learning can exploit curated affinity data [2]. These developments motivate models that retain the speed of learned scoring while improving the physical realism missing from many docking workflows.

Physics-based free-energy methods occupy a different point on the accuracy–cost spectrum. Molecular-mechanics end-point approaches and free-energy perturbation methods can encode more detailed thermodynamic information than empirical docking scores, yet their computational burden restricts routine application to very large virtual screens. Hybrid strategies such as algebraic graph learning and molecular-mechanics-informed scoring indicate that physically motivated descriptors can improve learned affinity models when combined with statistical learning [3, 4]. Benchmarking resources such as CASF also

Corresponding Author: Siti Rahman; Department of Pharmaceutical AI Engineering, Faculty of Pharmacy, University of Malaya, Kuala Lumpur, Malaysia. E-mail: siti.rahman@outlook.com.

emphasize that scoring, ranking, docking, and screening should be evaluated as related but distinct tasks rather than treated as interchangeable measures of model quality [5, 6].

Current machine-learning scoring functions still face important limitations when used prospectively. Graph and convolutional models can learn strong correlations from PDBbind-like complexes, but they may also inherit dataset bias, pose bias, and protein-family leakage if training and evaluation protocols are not designed carefully [7, 8]. Interaction-based graph models such as GraphDTA, MONN, and contact-map neural networks show the value of molecular graph representations for compound–protein affinity prediction, yet many such models do not explicitly constrain learned interactions to be compatible with electrostatics, dispersion, or solvation physics [7-9]. Dynamic effects are also frequently underrepresented because many models consume a single static docked or crystallographic pose rather than descriptors derived from molecular fluctuations.

The thesis of this MDL article is that a physics-informed graph neural network can combine docking, molecular dynamics, and force-field-derived interaction features to predict binding free energy with improved physical interpretability. The proposed model treats the protein–ligand complex as a spatial graph, augments edges with physically meaningful pairwise terms, and uses global features from docking and MD simulations to support residual correction of approximate binding estimates. Attention-based and structure-aware graph models already suggest that learned interaction patterns can be mapped back to chemically meaningful contacts [10-14]. Building on that direction, the proposed framework is designed not as a replacement for thermodynamics, but as a differentiable approximation that should be evaluated against docking, MM-GBSA, and free-energy benchmarks.

Background

Binding Free-Energy Prediction in Drug Discovery

Binding free-energy prediction spans a tiered workflow in which docking provides fast pose generation and approximate scoring, MM-PBSA or MM-GBSA provides end-point energetic refinement, and free-energy perturbation provides a more rigorous but more costly thermodynamic estimate. CASF studies highlight that docking benchmarks must distinguish pose reproduction, scoring power, ranking power, and screening power because a method can perform well on one dimension while failing on another [5, 6]. Machine-learning scoring functions such as AGL-Score and hybrid MM/GBSA–ML models have shown that structural and physical descriptors can be learned jointly rather than treated as separate post-processing steps [4, 15]. For lead optimization, the unmet need is therefore a rapid predictor that can preserve the physical interpretability of energy-based methods while scaling more like a learned scoring function.

Graph Neural Networks for Protein–Ligand Complexes

Graph neural networks represent atoms, bonds, residues, and intermolecular contacts as nodes and edges, making them naturally suited to protein–ligand complexes. GraphDTA showed how molecular graph representations can support affinity prediction in drug–target modeling [9], while contact-map-based approaches incorporated structural proximity information between ligand and protein regions [9]. More complex frameworks such as InteractionGraphNet and structure-aware interactive graph neural networks explicitly model intermolecular communication, allowing messages to pass across the binding interface rather than only within the ligand graph [13, 14]. These architectures motivate a binding free-energy model in which covalent ligand structure, protein environment, and cross-interface physical interactions are learned in a unified graph.

Physics-Informed Machine Learning

Physics-informed machine learning introduces prior knowledge through model architecture, input features, regularization terms, or auxiliary objectives that encourage learned functions to respect known molecular behavior. In protein–ligand modeling, this can mean embedding Coulombic, van der Waals, hydrogen-bonding, hydrophobic, or solvation-inspired terms into edge features rather than requiring the network to infer them only from coordinates and atom types. PIGNet formalized this idea by incorporating physics-informed interaction terms into deep learning for drug–target interaction prediction [16], and physics-guided neural scoring approaches similarly suggest that energy-like decompositions can constrain learned affinity estimates [17]. Such inductive bias is especially important when models are expected to extrapolate beyond closely related ligands or protein families.

Docking Scores and MD Descriptors as Features

Docking scores summarize approximate intermolecular complementarity, steric fit, and empirical interaction terms, but they often omit dynamic stability and solvent-mediated effects. Interaction fingerprint models and contact-based neural networks show that residue-level contact patterns can provide richer information than a single scalar docking score [18, 19]. MD-derived descriptors, including per-residue MM-GBSA decompositions, hydrogen-bond persistence, solvent exposure, and ligand mobility, can supplement a static docked pose with information about whether predicted interactions remain stable under thermal motion [15, 20]. A physics-informed GNN can therefore treat docking as an initial structural hypothesis and MD descriptors as evidence about the persistence and energetic plausibility of that hypothesis.

Benchmarks and Challenges in Binding Affinity Prediction

Benchmarking binding affinity models commonly involves PDBbind-derived training and test sets, CASF scoring tasks, and curated free-energy datasets for more thermodynamically rigorous comparisons. PDBbind-like benchmarks enabled many

deep scoring functions, including convolutional, topological, and graph-based models, but they also raise concerns about protein overlap, ligand-series leakage, and dependence on crystallographic pose quality [1, 3, 4]. CASF-2013 and CASF-2016 provide standardized assessments of scoring and ranking behavior, helping clarify whether a model is useful for pose ranking, affinity prediction, or virtual screening [5, 6]. Free-energy perturbation benchmark resources are important complementary tests because they focus on relative thermodynamic trends that docking-trained models may not capture without explicit physical structure.

Table 1 shows key benchmarking resources for evaluating binding affinity models, which differ in focus, methodology, and the type of insights they provide.

Table 1. Benchmarking Resources for Binding Affinity Model Evaluation

Benchmark Type	Example Resources	Focus / Purpose	Advantages	Limitations
PDBbind-derived sets	PDBbind	Training and testing deep scoring functions	Supports convolutional, topological, and graph-based models	Protein overlap, ligand-series leakage, reliance on crystallographic pose quality [1, 3, 4]
Standardized scoring tasks	CASF-2013, CASF-2016	Assess scoring and ranking behavior	Clarifies model utility for pose ranking, affinity prediction, or virtual screening [5, 6]	Limited to benchmarked scoring metrics; may not reflect thermodynamic rigor
Free-energy datasets	Curated FEP datasets	Evaluate relative thermodynamic trends	Captures physical trends beyond docking-trained models	Requires accurate physical models and more computational cost

Model Development Overview

High-Level Architecture

The proposed model is a three-dimensional graph neural network that takes a protein–ligand complex as input and outputs a scalar estimate of binding free energy. Each complex graph contains ligand atoms, selected binding-site protein atoms or residues, covalent edges, spatial contact edges, and energy-annotated intermolecular edges. Prior work on 3D convolutional affinity prediction, OnionNet contact shells, and geometric interaction graph neural networks supports the idea that spatial organization around the ligand is essential for learned scoring [1, 18, 21]. In this framework, docking scores and MD-derived descriptors are not used as final answers, but as contextual signals that help the graph model learn a physically meaningful correction.

Core Inputs

The core inputs consist of molecular graphs with atom and bond features, docked pose information, and molecular-dynamics-derived energetic descriptors. Docking contributes pose coordinates, empirical scoring terms, and interaction fingerprints that describe contacts between the ligand and binding-site residues, while MD contributes stability-oriented descriptors such as residue energy decompositions and persistent interaction patterns. Hybrid approaches combining molecular mechanics with machine learning show that end-point physical descriptors can be informative when learned together with structural features [15, 22]. The resulting input representation is intended to retain the speed and accessibility of docking while selectively incorporating dynamic and energetic information from short MD refinement.

Design Principles

The model is designed around four principles: physical awareness, geometric consistency, multimodal fusion, and interpretability. Physical awareness comes from energy-based edge features and optional regularization inspired by electrostatics and dispersion, while geometric consistency is supported by equivariant or geometry-aware message passing that respects rotations and translations of the complex. Equivariant line-graph and geometric interaction graph approaches indicate that preserving spatial symmetries can be valuable for affinity prediction from three-dimensional protein–ligand structures [21, 23]. Interpretability is built into the architecture through atom-level attribution, residue-level aggregation, and feature-gating mechanisms that reveal whether a prediction is driven primarily by docking, MD stability, or pairwise interaction physics.

Data Sources and Feature Engineering

Training and Benchmark Datasets

The conceptual training and evaluation strategy would use curated protein–ligand structural affinity data, CASF-style decoy and scoring benchmarks, and supplementary free-energy datasets when thermodynamic labels are available. PDBbind-derived datasets have supported many structure-based deep learning models, including KDEEP, OnionNet, and graph-based scoring functions [1, 4, 18]. CASF benchmarks provide standardized comparisons against classical scoring functions and help test whether a model can distinguish pose quality, rank related compounds, and prioritize likely binders [5, 6]. To test generalizability rather than memorization, splits should be designed by protein sequence, binding-site similarity, ligand scaffold, or target family rather than by random complex assignment.

Pre-Computation of Docking Scores and Interaction Fingerprints

Docking-derived features would be generated by producing one or more plausible binding poses and extracting scalar scores, per-pose energy components, contact maps, and residue-level interaction fingerprints. OnionNet demonstrated that layered intermolecular contacts around the ligand can be transformed into a learnable representation of binding affinity [18], while attention-based interaction models suggest that contact patterns can provide interpretable signals about which residues contribute to predicted affinity [19]. These features can be encoded as global graph attributes, intermolecular edge annotations, or residue-node descriptors depending on whether the information describes the whole pose or a local contact. Multiple poses should be represented explicitly so that the model can learn pose sensitivity instead of assuming that the highest-ranked docking pose is necessarily the most physically plausible.

MD-Derived Energy Features

MD-derived features would describe how the docked complex behaves after local relaxation and thermal sampling, without treating the simulation as a full free-energy calculation. Per-residue MM-GBSA or MM-PBSA decompositions can be converted into residue-node features or residue–ligand edge features, while hydrogen-bond persistence, ligand mobility, solvent exposure, and binding-site flexibility provide complementary descriptors of dynamic stability. Machine-learning models based on molecular mechanics and generalized Born surface area terms indicate that decomposed physical energy estimates can support affinity prediction when integrated with learned correction functions [15]. Persistent-homology and curvature-based descriptors further suggest that geometric and dynamic summaries can capture structural organization beyond static contact counts [3, 20].

Table 2 defines the proposed multimodal input representation and clarifies how docking, molecular dynamics, and force-field-derived descriptors should be encoded as graph-level, node-level, or edge-level evidence for binding free-energy prediction.

Table 2. Physics-Informed Input Representation for Protein–Ligand Binding Free-Energy Prediction

Input Layer	Specific Information Captured	Graph Encoding Strategy	Binding-Free-Energy Relevance	Main Modeling Risk	Design Safeguard
Docked pose geometry	Ligand coordinates, binding-site orientation, pose alternatives, steric fit	Atom coordinates; ligand–residue spatial edges; pose-level global attributes	Provides the initial structural hypothesis for binding and enables post-docking re-scoring	Model may overtrust the top-ranked docking pose even when it is physically implausible	Encode multiple poses and evaluate pose sensitivity rather than assuming one pose is correct
Docking score components	Empirical affinity score, steric terms, hydrogen-bond terms, hydrophobic contacts, interaction fingerprints	Global graph features; residue-node descriptors; contact-edge annotations	Supplies a fast approximate prior that the GNN can correct using structural and physical evidence	Docking scores may reproduce empirical scoring-function bias	Treat docking as a prior, not as the target answer; learn residual corrections
Protein–ligand contact topology	Atom-pair contacts, residue–ligand proximity, binding-site interaction patterns	Intermolecular edges connecting ligand atoms to protein atoms or residues	Captures the local recognition environment that drives affinity and selectivity	Static contact maps may confuse transient proximity with durable binding	Combine contact topology with MD persistence and energy decomposition
Force-field-inspired interaction terms	Coulombic interactions, van der Waals terms, steric clashes, hydrophobic complementarity	Energy-annotated intermolecular edge features	Embeds physical inductive bias directly into message passing	Force-field parameters may be unreliable for metals, covalent binders, unusual protonation states, or water-mediated systems	Add domain flags and reliability checks for difficult chemical environments
MD-derived stability descriptors	Hydrogen-bond persistence, ligand mobility, residue flexibility, solvent exposure, interaction persistence	Residue-node features; ligand-node descriptors; dynamic edge attributes	Distinguishes stable binding interactions from pose artifacts	MD features introduce computational cost and protocol dependence	Use tiered deployment: docking-only screening first, MD-augmented scoring for prioritized compounds
Per-residue energy decomposition	Residue-level MM-GBSA/MM-PBSA-like energetic contributions	Residue-node features; residue–ligand edge features	Enables residue-level interpretation of favorable and unfavorable interaction regions	Decomposition values may be noisy or sensitive to simulation setup	Compare energy decompositions with neural attributions and docking contacts
Ligand chemical graph	Atom type, formal charge, aromaticity, hybridization, bond type, ring structure	Ligand atom nodes and covalent bond edges	Preserves medicinal chemistry information needed for analogue design	Ligand patterns may be memorized when scaffold overlap is present	Use scaffold-based splits and external chemotype testing
Protein binding-site context	Residue identity, local geometry, charge environment, secondary-structure neighborhood, flexibility	Protein atom or residue nodes; local spatial edges	Encodes target-specific recognition features and selectivity-relevant context	Protein-family leakage can inflate benchmark accuracy	Use protein-sequence, family, and binding-site similarity splits

Global simulation and preparation metadata	Protonation state, pose source, MD protocol, force-field choice, simulation length	Global graph attributes or metadata embeddings	Helps identify when predictions depend on preparation assumptions	Hidden preprocessing variation may reduce reproducibility	Report preparation protocol and include metadata-aware sensitivity analysis

Physics-Informed Graph Neural Network Architecture

Equivariant Message Passing with Energy Edges

The GNN operates on three-dimensional coordinates, where messages are passed along covalent bonds, intramolecular spatial contacts, and protein–ligand interaction edges. Edge features include distances, atom-pair types, contact categories, and pre-computed force-field-inspired terms such as Coulombic and Lennard-Jones-like contributions. Physics-informed models such as PIGNet show that explicitly encoding interaction physics can guide neural prediction of drug–target interactions [16], while geometric and equivariant GNNs suggest that spatial symmetry should be respected when learning from protein–ligand structures [21, 23]. This design allows the model to learn both local chemical compatibility and global binding-site geometry without confusing coordinate orientation with molecular meaning.

Figure 1 illustrates the proposed physics-informed graph neural network architecture, showing how docking priors, molecular-dynamics descriptors, and force-field-inspired interaction terms are fused into an interpretable spatial graph model for binding free-energy prediction.

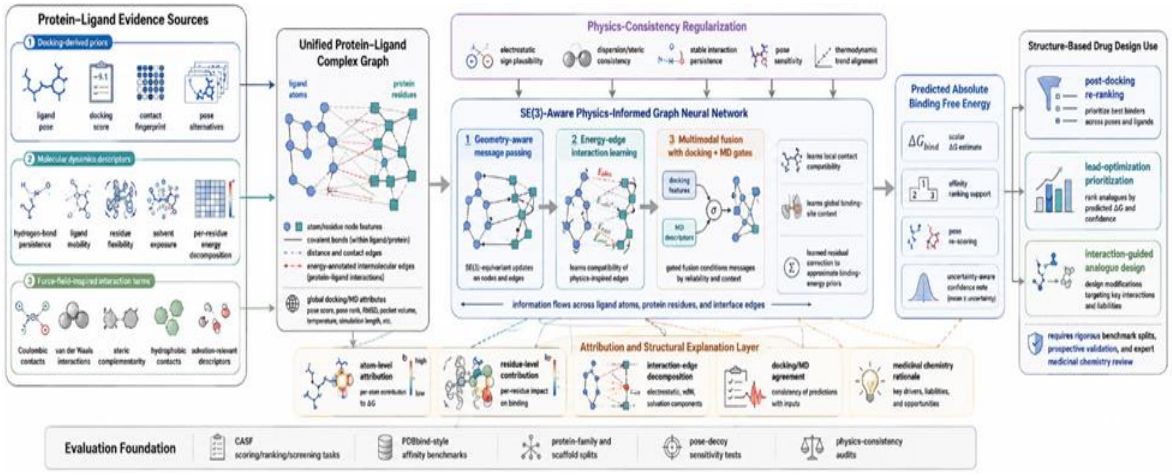

Figure 1. Physics-Informed Graph Neural Network Architecture for Binding Free-Energy Prediction from Docking and Dynamics Data

Multi-Modal Feature Fusion

The architecture fuses graph-derived embeddings with docking and MD descriptors through a multimodal module that projects each information source into a shared latent space. Docking features can act as a fast empirical prior, MD descriptors can describe dynamic stability, and graph embeddings can encode atomistic complementarity across the binding interface. Multi-objective and structure-aware neural models such as MONN, PLANET, and SS-GNN support the use of integrated representations that combine molecular topology, structural context, and interaction-level information [7, 24, 25]. A gating mechanism would allow the model to adjust how strongly it relies on each modality for a given complex, making the prediction less dependent on any single imperfect source of evidence.

Output Head and Training

The output head maps the pooled graph and multimodal embedding to a scalar predicted binding free energy, while optional auxiliary heads can predict pose preference or interaction consistency. Training would use a regression loss for absolute affinity and could include a ranking-oriented objective to encourage chemically sensible ordering of related ligands without reporting or assuming any numerical performance. Cascade graph convolutional networks, deep fusion inference, and optimized graph scoring models illustrate how supervised affinity learning can be structured around complex-level embeddings and ranking-sensitive objectives [11, 12, 26]. In the proposed physics-informed formulation, auxiliary losses should regularize the model toward physically plausible interaction decompositions rather than merely improving benchmark fit.

Integrating Docking, Dynamics, and Physics-Based Energy Terms

Docking Scores as Priors

Docking scores should be treated as approximate priors rather than final estimates of binding free energy. In the proposed model, the docking score provides a baseline energetic hypothesis, while the GNN learns a residual correction from atomistic contacts, pose geometry, and physics-derived interaction features. This design is consistent with interaction-based inductive-bias models that use structural information to refine affinity prediction beyond raw geometric proximity [27]. It also aligns

with generic scoring frameworks that integrate physical prior knowledge into data-driven protein–ligand interaction modeling [28].

MD-Derived Stability and Interaction Descriptors

MD-derived descriptors provide information about whether the interactions observed in a docked pose remain stable under molecular motion. Hydrogen-bond persistence, ligand mobility, residue flexibility, and per-residue energy decomposition can help the GNN distinguish transient steric complementarity from durable binding interactions. Physics-guided neural scoring models suggest that energy-like descriptors can improve interpretability when incorporated into affinity prediction [17], while augmented free-energy data strategies indicate that learned scoring functions may benefit from closer alignment with thermodynamic simulation outputs. In this framework, MD descriptors serve as dynamic context rather than as substitutes for rigorous free-energy calculations.

Physics-Based Constraints

Physics-based constraints can be introduced by requiring learned interaction contributions to remain qualitatively consistent with electrostatic, dispersion, and steric expectations. For example, an auxiliary regularization term could encourage edge-level electrostatic attributions to follow the sign and relative magnitude implied by charge-based interactions, while still allowing the neural network to learn corrections for solvation, entropy, and environment. Local-global geometric equivariant graph representation learning supports the idea that affinity prediction should combine local atomic interactions with broader structural context. Such constraints would be expected to improve extrapolation when the model encounters unfamiliar scaffolds, targets, or binding-site environments.

Model Interpretability and Structural Insights

Atom-Level Attribution

Atom-level attribution can expose which ligand atoms, protein residues, and intermolecular contacts drive the predicted free energy. Attention weights, gradient-based saliency, and edge-contribution decomposition can be mapped onto the three-dimensional complex to create a visual rationale for medicinal chemists. Attention-based protein–ligand affinity models show that interaction-focused neural mechanisms can connect predictions to specific binding-site contacts [19], and structure-aware graph models further support residue- and atom-level interpretation of learned affinity signals [14]. In a physics-informed architecture, these attribution maps should be compared with docking interactions and MD-derived energy decompositions to assess whether the model is learning chemically plausible explanations.

From Prediction to Design

The interpretability layer can translate a scalar free-energy estimate into design hypotheses for lead optimization. If a solvent-exposed hydrophobic group contributes weakly to the predicted interaction pattern, the model could motivate replacement with a polar substituent that stabilizes a persistent hydrogen bond observed during MD analysis. Multi-objective affinity models such as MONN and PLANET suggest that learned representations can support broader optimization logic beyond a single affinity value [7, 25]. When coupled with physics-informed attributions, the proposed GNN could guide analogue design by identifying which molecular changes are most consistent with stable and energetically favorable binding. **Figure 2** illustrates how atom-level and residue-level attribution can transform a predicted protein–ligand free-energy value into chemically interpretable design hypotheses by linking model saliency, binding-site contacts, docking evidence, and MD-derived energy decomposition.

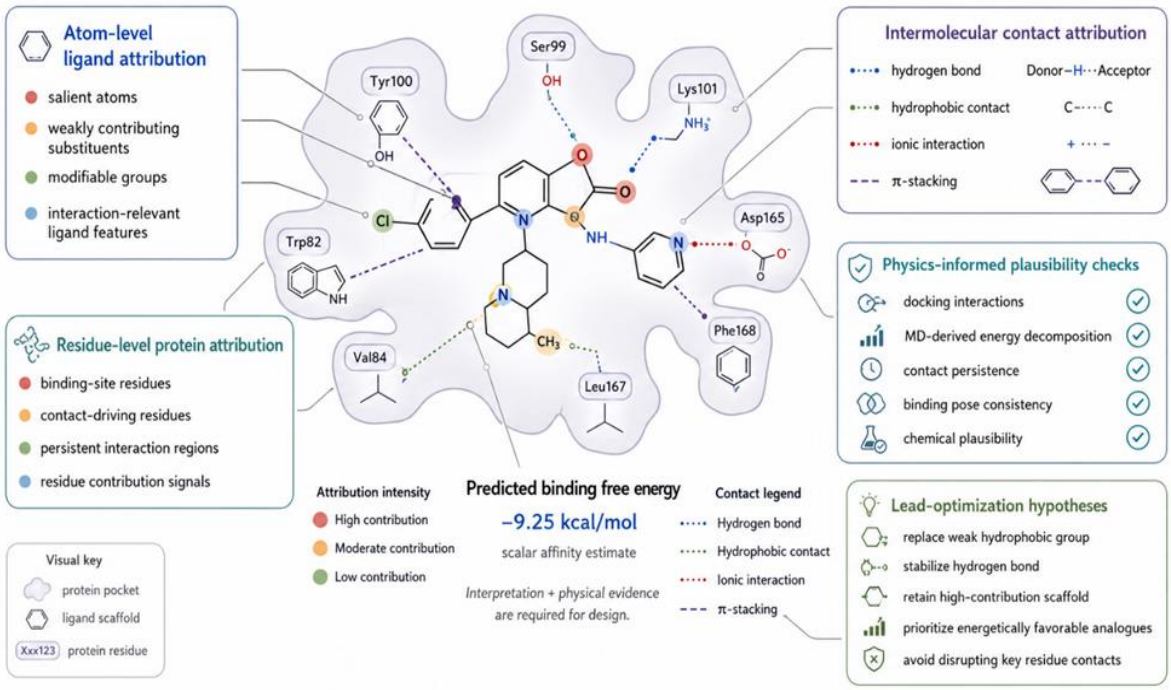

Figure 2. Atom-Level Attribution as a Bridge Between Binding Free-Energy Prediction and Ligand Design

Integration Into Structure-Based Drug Design Pipelines

Post-Docking Scoring and Re-Ranking

The most direct deployment of the proposed model is post-docking re-scoring, where docked compounds are first generated by a standard docking engine and then re-ranked by the physics-informed GNN. This approach preserves the throughput advantage of docking while replacing the native empirical score with a learned free-energy estimate informed by geometry, interaction physics, and optional MD descriptors. XLPFE and related machine-learning scoring functions illustrate how learned models can be positioned as scoring and ranking refinements after structure-based pose generation [22]. In practical pipelines, a lighter docking-only version could serve as an initial filter, while the full docking-plus-MD model could be applied to prioritized compounds.

Lead Optimization Guidance

During lead optimization, the model should provide both predicted affinity trends and interaction-level feedback. This dual output is useful because medicinal chemistry decisions often depend on whether an affinity change is supported by interpretable contacts, stable hydrogen bonding, reduced steric strain, or improved local complementarity. GraphscoreDTA and other optimized GNN affinity models demonstrate how graph representations can support compound prioritization [26], while studies on data bias emphasize that prospective usefulness depends on generalization beyond memorized structural patterns. A physics-informed design therefore should report not only a predicted free energy but also a confidence-oriented explanation of the structural evidence behind that prediction.

Evaluation Strategy

Predictive Accuracy

Predictive accuracy should be assessed on standardized structure-based affinity benchmarks using regression and ranking metrics, while avoiding overinterpretation of any single benchmark outcome. Comparisons should include native docking scores, MM-GBSA-style estimates, conventional machine-learning scoring functions, and graph neural network baselines. CASF-based studies provide a framework for separating scoring power from ranking and screening behavior [5, 6], while PDBbind-trained deep models such as KDEEP, OnionNet, and AGL-Score provide relevant learned-scoring baselines [1, 4, 18]. The evaluation should describe whether the physics-informed model improves conceptual reliability across tasks rather than presenting unsupported numerical results.

Generalization and Pose-Sensitivity

Generalization should be evaluated by splitting data according to protein families, ligand scaffolds, and binding-site similarity rather than relying only on random partitions. This is essential because apparent affinity prediction accuracy can reflect hidden overlap between training and test complexes, especially when structurally related ligands or homologous proteins appear in both sets. Bias-focused analysis of binding affinity prediction shows that careful dataset design is necessary for estimating real-world model transferability. Pose-sensitivity tests should additionally compare predictions across native-like poses, docking decoys, and perturbed conformations to determine whether the model rewards physically plausible binding modes.

Physics-Consistency Checks

Physics-consistency checks should examine whether learned atom-level and residue-level contributions align qualitatively with molecular interaction expectations. For example, favorable charged contacts, stable hydrogen bonds, and persistent hydrophobic packing should produce interpretable attribution patterns that agree with docking interactions and MD-derived energy decompositions. Physics-informed scoring models and generic prior-knowledge integration frameworks provide precedent for linking neural predictions to physically meaningful interaction components [16, 28]. Free-energy-augmented scoring approaches further suggest that learned models should be compared with thermodynamic trends, including cases where enthalpic gains may be offset by conformational or solvent-related penalties.

Table 3 provides an evaluation and deployment-readiness framework that separates benchmark accuracy from ranking value, pose sensitivity, physics consistency, generalization, uncertainty, and prospective medicinal chemistry utility.

Table 3. Evaluation and Deployment Readiness Framework for Physics-Informed Binding Free-Energy GNNs

Evaluation Domain	Core Question	Recommended Test Design	What the Test Reveals	Failure Pattern to Watch For	Deployment Implication
Absolute affinity prediction	Does the model estimate binding free energy accurately across diverse protein–ligand complexes?	Compare against docking scores, MM-GBSA-style estimates, conventional ML scoring functions, and graph neural baselines on curated affinity benchmarks	Whether physics-informed fusion improves regression performance beyond empirical scoring	High benchmark correlation driven by dataset overlap rather than true generalization	Useful only as a re-scoring tool if validated beyond random splits
Ranking within ligand series	Can the model prioritize related analogues in a chemically meaningful order?	Evaluate congeneric or target-specific ligand series using ranking metrics	Whether predictions support lead optimization rather than only broad binder/non-binder separation	Correct global trends but poor analogue-level ordering	Limited value for medicinal chemistry decision-making
Screening enrichment	Can the model separate likely binders from decoys or weak binders?	CASF-style screening tasks and decoy-based enrichment analysis	Whether the model can assist virtual screening triage	Good scoring performance but poor early enrichment	Should not be used for large-library prioritization without enrichment validation
Pose sensitivity	Does the model reward physically plausible binding modes?	Compare native-like poses, docking decoys, perturbed conformations, and multiple docked poses	Whether the model recognizes pose quality rather than memorizing ligand identity	Similar affinity predictions for unrealistic and plausible poses	Requires pose-quality filtering before deployment
Scaffold generalization	Does the model transfer to unfamiliar chemotypes?	Scaffold-based train/test splits and external chemotype evaluation	Whether learned features generalize beyond known ligand families	Performance collapse on new scaffolds	Should be restricted to interpolation within known chemical space
Protein-family generalization	Does the model transfer across related or unrelated targets?	Protein-sequence, binding-site similarity, and target-family splits	Whether predictions reflect molecular recognition rather than protein-family leakage	Strong random-split results but weak target-split results	Not ready for prospective target expansion
Physics consistency	Do learned attributions align with plausible molecular interaction behavior?	Compare atom/residue attributions with electrostatics, van der Waals contacts, hydrogen-bond persistence, and energy decomposition	Whether explanations are chemically interpretable and physically coherent	Saliency highlights irrelevant atoms or unstable contacts	Predictions require expert review before design action
MD contribution value	Do dynamic descriptors improve prediction enough to justify added cost?	Compare docking-only, docking-plus-graph, and docking-plus-MD model variants	Whether MD-derived stability information adds practical signal	MD features add cost but little predictive or interpretive gain	Use MD-augmented model only for prioritized lead series
Uncertainty and reliability	Can the model identify low-confidence or out-of-domain cases?	Use uncertainty calibration, ensemble disagreement, distance-to-training-domain analysis, and failure-case review	Whether the model can flag cases requiring simulation or experimental confirmation	Overconfident predictions for unusual chemistry or poor poses	Necessary for responsible prioritization in discovery pipelines
Prospective utility	Does the model improve real design decisions?	Retrospective temporal splits, blinded prospective compound ranking, and experimental follow-up	Whether the model changes compound prioritization in a useful way	Benchmark success without experimental decision value	Required before claiming discovery-readiness

Limitations

Reliance on MD Simulations

The main practical limitation of the proposed model is its reliance on MD-derived descriptors for the full version of the pipeline. Although dynamic features can capture stability and interaction persistence that are absent from static docking poses, they introduce computational cost and protocol dependence. Hybrid MM/GBSA–machine-learning methods show that end-point energetic descriptors can be useful [15], but their value depends on simulation setup, conformational sampling, and the stability of the protein–ligand complex. A tiered deployment strategy would therefore use docking-only features for large early screens and reserve MD-augmented prediction for smaller lead-optimization sets.

Force-Field Limitations

The physics-derived edge features are constrained by the quality of the underlying force-field parameters and implicit assumptions used to compute them. This limitation is especially important for systems involving metal coordination, covalent binding, highly flexible loops, disordered regions, unusual protonation states, or strong water-mediated effects. Physics-guided neural models can reduce purely empirical bias, but they cannot fully correct physically inaccurate input descriptors if the molecular mechanics representation is poor [17]. The model should therefore be treated as a physics-informed approximation whose reliability depends on both learned representation quality and the validity of the molecular model used to generate its features.

Conclusion

A physics-informed graph neural network provides a coherent framework for binding free-energy prediction by integrating docking, molecular dynamics, and force-field-inspired interaction features. In this design, the protein–ligand complex is represented as a spatial graph whose messages encode both learned structural patterns and physically meaningful interaction terms. The model is therefore positioned between fast empirical docking and computationally intensive free-energy simulation. The principal strength of this framework is its ability to combine predictive flexibility with physical consistency. Docking scores provide rapid pose-level priors, MD descriptors contribute dynamic stability information, and energy-annotated graph edges support chemically interpretable learning. Together, these components could make the predicted free energy more useful for lead optimization than a purely black-box affinity score.

Important challenges remain before such a model could be used confidently in prospective discovery campaigns. MD-derived descriptors add computational cost, force-field terms may be unreliable for difficult chemical environments, and benchmark success does not guarantee experimental transferability. Prospective validation, careful uncertainty assessment, and rigorous split design would be essential for responsible deployment.

Open-source implementation would help make the framework transparent, reproducible, and extensible across academic and industrial settings. Integration into common docking and molecular simulation workflows would allow the model to function as a practical post-docking re-scoring and lead-optimization tool. A successful implementation should not replace physical simulation or experimental measurement, but should help prioritize where those resources are most valuable.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Jiménez J, Skalic M, Martinez-Rosell G, De Fabritiis G. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. J Chem Inf Model. 2018;58(2):287-96.
2. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. Bioinformatics. 2018;34(21):3666-74.
3. Cang Z, Wei GW. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. Int J Numer Method Biomed Eng. 2018;34(2):e2914.
4. Nguyen DD, Wei GW. AGL-score: algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. J Chem Inf Model. 2019;59(7):3291-304.
5. Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, et al. Comparative assessment of scoring functions: the CASF-2016 update. J Chem Inf Model. 2018;59(2):895-913.
6. Li Y, Su M, Liu Z, Li J, Liu J, Han L, et al. Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark. Nat Protoc. 2018;13(4):666-80.

7. Li S, Wan F, Shu H, Jiang T, Zhao D, Zeng J. MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. Cell Syst. 2020;10(4):308-22.
8. Jiang M, Li Z, Zhang S, Wang S, Wang X, Yuan Q, et al. Drug–target affinity prediction using graph neural network and contact maps. RSC Adv. 2020;10(35):20701-12.
9. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics. 2021;37(8):1140-7.
10. Son J, Kim D. Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities. PLoS One. 2021;16(4):e0249404.
11. Shen H, Zhang Y, Zheng C, Wang B, Chen P. A cascade graph convolutional network for predicting protein–ligand binding affinity. Int J Mol Sci. 2021;22(8):4023.
12. Jones D, Kim H, Zhang X, Zemla A, Stevenson G, Bennett WD, et al. Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. J Chem Inf Model. 2021;61(4):1583-92.
13. Jiang D, Hsieh CY, Wu Z, Kang Y, Wang J, Wang E, et al. Interactiongraphnet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. J Med Chem. 2021;64(24):18209-32.
14. Li S, Zhou J, Xu T, Huang L, Wang F, Xiong H, et al. Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021; p. 975-85.
15. Dong L, Qu X, Zhao Y, Wang B. Prediction of binding free energy of protein–ligand complexes with a hybrid molecular mechanics/generalized born surface area and machine learning method. ACS Omega. 2021;6(48):32938-47.
16. Moon S, Zhung W, Yang S, Lim J, Kim WY. PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. Chem Sci. 2022;13(13):3661-73.
17. Cain S, Risheh A, Forouzesh N. A physics-guided neural network for predicting protein–ligand binding free energy: from host–guest systems to the PDBbind database. Biomolecules. 2022;12(7):919.
18. Zheng L, Fan J, Mu Y. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. ACS Omega. 2019;4(14):15956-65.
19. Seo S, Choi J, Park S, Ahn J. Binding affinity prediction for protein–ligand complex using deep attention mechanism based on intermolecular interactions. BMC Bioinformatics. 2021;22(1):542.
20. Wee J, Xia K. Ollivier persistent Ricci curvature-based machine learning for the protein–ligand binding affinity prediction. J Chem Inf Model. 2021;61(4):1617-26.
21. Yang Z, Zhong W, Lv Q, Dong T, Chen CYC. Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures (gign). J Phys Chem Lett. 2023;14(8):2020-33.
22. Dong L, Qu X, Wang B. XLPFE: A simple and effective machine learning scoring function for protein–ligand scoring and ranking. ACS Omega. 2022;7(25):21727-35.
23. Yi Y, Wan X, Zhao K, Ou-Yang L, Zhao P. Equivariant line graph neural network for protein-ligand binding affinity prediction. IEEE J Biomed Health Inform. 2024;28(7):4336-47.
24. Zhang S, Jin Y, Liu T, Wang Q, Zhang Z, Zhao S, et al. SS-GNN: a simple-structured graph neural network for affinity prediction. ACS Omega. 2023;8(25):22496-507.
25. Zhang X, Gao H, Wang H, Chen Z, Zhang Z, Chen X, et al. Planet: a multi-objective graph neural network model for protein–ligand binding affinity prediction. J Chem Inf Model. 2023;64(7):2205-20.
26. Wang K, Zhou R, Tang J, Li M. GraphscoreDTA: optimized graph neural network for protein–ligand binding affinity prediction. Bioinformatics. 2023;39(6):btad340.
27. Yang Z, Zhong W, Lv Q, Dong T, Chen G, Chen CY. Interaction-based inductive bias in graph neural networks: enhancing protein-ligand binding affinity predictions from 3d structures. IEEE Trans Pattern Anal Mach Intell. 2024;46(12):8191-208.
28. Cao D, Chen G, Jiang J, Yu J, Zhang R, Chen M, et al. Generic protein–ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. Nat Mach Intell. 2024;6(6):688-700.