Chemical Science

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: Y. Zhang, C. Huang, Y. Liu, Q. Yuan, Z. Xunyong and S. Sun, *Chem. Sci.*, 2026, DOI: 10.1039/D5SC09548D.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemical-science

View Article Online
DOI: 10.1039/D5SC09548D

ARTICLE

HCGT-PL: A heterogeneous contrastive graph transformer unifying protein–ligand affinity prediction and structure-based virtual screening

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Yunjiang Zhang [a], Chenyu Huang [a], Yuetong Liu [b], Qing Yuan [c], Xunyong Zhou [d] and Shaorui Sun *[a]

Structure-based virtual screening and binding affinity prediction remain challenging due to solvation/entropy effects, protein flexibility, and induced fit. We present a Heterogeneous Contrastive Graph Transformer for Protein-Ligand (HCGT-PL) framework. Each complex is represented as a directed heterogeneous graph with multiple node and relation types; relation-specific multi-head attention enables message passing and aggregation. Unsupervised augmentations yield transferable interaction representations that are fine-tuned for affinity regression and virtual screening. Across diverse benchmarks and hold-out evaluations, the approach delivers robust accuracy, strong ranking capability, and pronounced early-enrichment, with consistent generalization over varied protein families and pocket conditions. Interpretability visualizations indicate that the model prioritizes ligand functional groups and contacting receptor side chains within the binding pocket. By unifying heterogeneous graph modeling, graph Transformers, and contrastive learning, this framework provides a general, transferable, and interpretable solution for protein–ligand modeling.

Introduction

In the lengthy and costly process of modern drug discovery, accurately predicting the binding affinity between ligands and their target proteins, as well as their binding conformations, is crucial. High binding affinity is not only essential for the efficacy of drugs but also critical for minimizing side effects.

Isothermal Titration Microcalorimetry[1] and Surface Plasmon Resonance,[2] as two widely applied techniques for analyzing molecular interactions and determining binding affinities, face challenges such as complexity, high cost, and time consumption in large-scale drug screening. There is an urgent need for innovative computational methods to accelerate the drug discovery process and reduce costs.

Virtual screening (VS), as a computational method, identifies molecules that may bind to drug targets by screening large chemical libraries, thereby accelerating the drug discovery process. Among these, structure-based virtual screening (SBVS) simulates ligand-target interactions through computational modeling to achieve optimal spatial and physicochemical complementarity.[3,4]

Classic affinity prediction calculation methods include force field-based scoring functions, empirical scoring functions, and knowledge-based scoring functions. However, these methods generally face problems such as difficulty in handling solvation effects, ignoring or simplifying entropy contributions, inability to capture non-linear relationships, challenges in simulating protein flexibility and induced fit, as well as inaccurate binding pose predictions. These issues result in poor correlation with experimental binding affinities, creating a disconnect between simplified models and the complex reality of biological interactions. This leads to significant inaccuracies and a lack of comprehensive physical representation, necessitating more sophisticated computational methods to model these complex factors for accurate protein-ligand affinity prediction.

Graph Neural Networks (GNNs) have overcome the limitations of early deep learning architectures (CNNs, RNNs) in processing molecular data's graph structure and three-dimensional spatial relationships, providing a powerful framework for analyzing graph-form data, particularly suitable for non-Euclidean data structures.[5] In drug discovery, they have advanced the graph representation of drug compounds and target proteins, enabling more comprehensive and accurate understanding of molecular interactions. The core of GNNs follows a message passing mechanism, including message passing and node updating, with specific architectures such as GCNs[6] and GATs,[7] each having their own characteristics. Compared to traditional deep learning models, GNNs show clear advantages in handling graph topologies, arbitrarily complex topologies, and structures without fixed node ordering, effectively capturing relationships and propagating information. For molecular representation, they are more targeted, providing more natural, powerful, and

[a] Department of Chemical Engineering and Technology, College of Materials Science and Engineering, Beijing University of Technology, Beijing 100124, P. R. China.
[b] Department of Biology, College of Chemistry and Life Sciences, Beijing University of Technology, Beijing 100124, P. R. China.
[c] Center of Excellence for Environmental Safety and Biological Effects, Department of Chemistry, Beijing University of Technology, Beijing 100124, P. R. China.
[d] Zhencui (Jiangsu) Enzyme Technology Development Co., Ltd. Shuyang County, Suqian 223600, P. R. China.
Supplementary Information available: Additional experimental details, methods, results, and analysis.

View Article Online
DOI: 10.1039/D5SC00548D

accurate representations for learning protein-ligand interactions.

Biological systems involve multiple types of entities (e.g., different types of protein atoms, different types of ligand atoms) and interactions (e.g., covalent bonds within proteins, covalent bonds within ligands). Isomorphic GNNs treat all nodes and edges equally. Furthermore, Heterogeneous Graph Neural Networks (HGNNs)[8] explicitly model this heterogeneity, allowing for different types of nodes and edges, each of which may have unique feature spaces and message passing/aggregation functions.

Protein-ligand complexes can be represented as heterogeneous graphs, where protein atoms and ligand atoms are different node types, while intramolecular covalent bonds and intermolecular non-covalent interactions are different edge types. For example, HGEF-Net models ligands as heterogeneous graphs, capturing inter-atomic correlations, subgraph topology, and chemical functional group information. The advantage of HGNNs for PLI lies in their ability to represent complex biological interactions more accurately and in greater detail, and to learn patterns and relationships specific to certain types. For instance, the HGScore model separates embeddings for each edge type (protein-protein, protein-ligand, etc.), allowing submodels to learn from specific contexts, thereby capturing richer semantic contexts and improving performance compared to homogeneous graphs.

Related models include HGScore[10], HGEF-Net[9], HNCL-DTI[11], and MVCL-DTI[12]. The transition from generic GNNs to HGNNs in the PLI field marks the maturation of this domain. This reflects a recognition that a singular approach to processing nodes/edges is insufficient to address the complexity of biological systems.

The Transformer architecture, with its attention mechanism, demonstrates excellent performance in capturing long-range dependencies and global context, while Graph Transformers apply the attention mechanism to graph-structured data, allowing nodes to attend to other nodes in the graph, thereby capturing global information and long-range dependencies both within and between molecules. Moreover, Contrastive Learning (CL)[13], as a self-supervised learning method, can leverage large unlabeled molecular databases to learn robust, generalizable, and intrinsic molecular representations, enhancing the performance of downstream tasks (such as PLA prediction, DTI prediction, and molecular property prediction). Our previous work, CL-GNN[14], has also demonstrated this.

Here, we present an end-to-end workflow, termed the Heterogeneous Contrastive Graph Transformer for Protein-Ligand Modeling (HCGT-PL) (Fig. 1). For each protein-ligand complex, we construct a directed heterogeneous graph with two node types (protein and ligand) and four relation (edge) types that distinguish intra-molecular covalent structure from inter-molecular contacts. During training, to support task-agnostic representation learning, we apply graph-view augmentations-including subgraph removal, atom masking, and bond deletion. Message passing and aggregation are performed through relation-specific Q-K-V projections under multi-head attention, with geometric context incorporated on inter-molecular edges to capture contact geometry efficiently. We first pretrain the model in a self-supervised contrastive manner to learn transferable interaction representations, and then fine-tune the pretrained encoder with lightweight task heads for downstream binding-affinity regression and structure-based virtual screening. Together, these designs explicitly reflect the intrinsic asymmetry of molecular recognition (ligand to protein and protein to ligand) while enabling relation-aware and geometry-sensitive interaction modeling with practical computational cost.

Methods

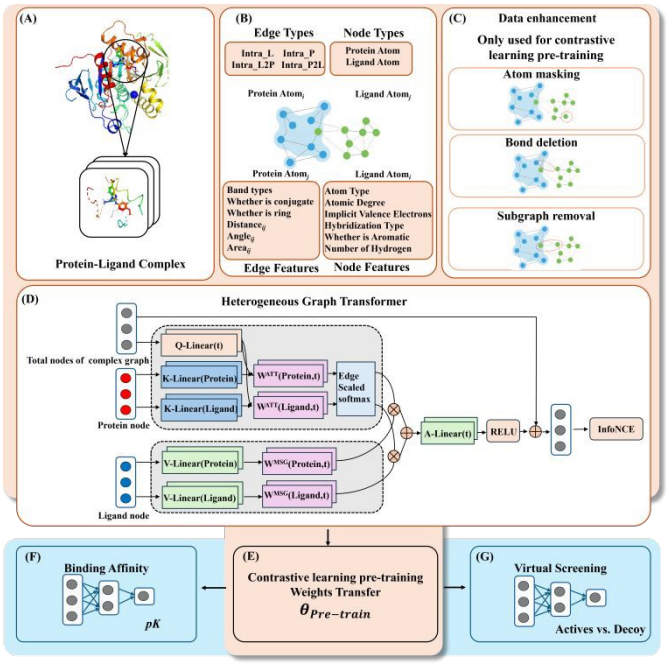

Figure 1. Overall workflow of the HCGT-PL. (A) Data preparation. Protein–ligand complexes are converted to pockets (B) Graph construction. A heterogeneous graph is built with two node types (protein atoms, ligand atoms) and four edge types; node/edge features include atom/bond types, degrees, valence, aromaticity, distances/angles, etc.(C) Data augmentation. Such as subgraph removal, atom masking, and bond deletion on the graph information. (D) Heterogeneous Graph Transformer. For a target node (Total nodes of the complex graph) of type t, relation-specific linear projections generate Q/K/V; multi-head attention with edge-scaled softmax computes message weights, followed by relation-specific message passing and target-specific aggregation. (E) Contrastive pre-training. Two augmented graph views are encoded and optimized with a contrastive loss to learn transferable interaction representations. (F)and (G) Downstream tasks. The pretrained encoder is fine-tuned for binding-affinity regression and structure-based virtual screening.

Graph representation in protein-ligand complex

We model each complex as a typed, directed heterogeneous graph.

$$\mathcal{G} = (V, E, \tau, \phi, X^V, X^E)$$

where $V = V^l \cup V^p$ is the union of ligand and protein atoms. The node–type mapping is $\tau: V \to A$ with $A = \{l, p\}$ (ligand, protein). The edge–type mapping is $\phi: E \to R$ with $R = \{11, \mathrm{pp}, 1\mathrm{p}, \mathrm{pl}\}$. Accordingly

$$E = E^{11} \cup E^{\mathrm{pp}} \cup E^{1\mathrm{p}} \cup E^{\mathrm{pl}}, E^{xy} \subseteq V^x \times V^y.$$

E^{11}: intra-ligand edges between ligand atoms (covalent bonds). E^{pp}: intra-protein edges between protein atoms in the pocket (covalent bonds).$E^{1\mathrm{p}}$ and E^{pl}: inter-molecular contact edges from ligand→protein and protein→ligand, respectively. These are constructed between atom pairs within a distance threshold d^{c}.

Edge features comprise bond attributes for intramolecular ligand–ligand and protein–protein edges and invariant geometric descriptors for intermolecular ligand–protein and protein–ligand edges. The 11-dimensional geometry-only edge feature is constructed as follows. For each directed edge i→j, all neighboring atoms k of atom j, excluding atom i, are enumerated. For each i–j–k triplet, three geometric quantities are calculated: the angle ∠ijk, the area of triangle △ijk, and the Euclidean distance d(i,k). Each quantity is independently aggregated over all neighboring atoms k using max, sum, and mean statistics, yielding 3×3=9 features. Two direct distance terms between atoms i and j, namely the L1 norm and the L2/Euclidean distance, are then appended, resulting in an 11-dimensional geometric feature vector. For intramolecular edges, a 6-dimensional chemical bond feature, including bond type, conjugation, and ring membership, is prepended, yielding a 17-dimensional edge feature. Intermolecular edges use only the 11-dimensional geometric feature. Full feature definitions are summarized in Table S1.

Heterogeneous graph transformer

In this section, we will introduce the detailed steps for constructing the Molecular Heterogeneous Graph Transformer (HGT). The HCGT-PL encoder consists of three stacked HGT convolutional layers. Each HGT layer uses two relation-specific attention heads, with a hidden dimension of 128 per head, giving a concatenated node embedding dimension of 256. Layer normalization is applied within each HGT layer, and a dropout rate of 0.05 is used in both the graph convolutional layers and the prediction head. For affinity regression, the graph-level representation is fed into a three-layer MLP prediction head with dimensions 256 → 256 → 1, where Dropout(0.05), LeakyReLU activation, and BatchNorm1d are applied after the hidden layer. The input node feature is a 35-dimensional one-hot atom descriptor, including element type, atom degree, implicit valence, hybridization, aromaticity, and total hydrogen count.

The entire HGT is divided into three modules: heterogeneous mutual attention, heterogeneous message passing, and target-specific aggregation.

Heterogeneous mutual attention. In contrast to the traditional transformer model, the HGT model assigns a distinct set of projection weights W to each meta-relationship. In a conventional transformer, however, all words share a common set of weights. With a more focused approach, we apply a linear projection technique, denoted as Q-Linear $^{i}_{\tau(t)}$, to transform the target node t into the i-th query vector.

$$\mathrm{Attention}_{HGT}(s,e,t) = \underset{\forall s \in N(t)}{\mathrm{Softmax}}\left(\|_{i\in[1,h]} ATT-\mathrm{head}^{i}(s,e,t)\right)$$

$$ATT-\mathrm{head}^{i}(s,e,t) = \left(K^{i}(s)W^{\mathrm{ATT}}_{\varphi(e)}Q^{i}(t)^{T}\right)\cdot\frac{\mu_{\langle\tau(s),\varphi(e),\tau(t)\rangle}}{\sqrt{d}}$$
$$K^{i}(s) = K-\mathrm{Linear}^{i}_{\tau(s)}\left(H^{(l-1)}[s]\right)$$
$$Q^{i}(t) = \text{Q-Linear}^{i}_{\tau(t)}\left(H^{(l-1)}[t]\right)$$

Here, ATT-head refers to the i-th attention head, K(s) represents the i-th key vector for source nodes, and Q(t) is the i-th query vector for target node t. The relationship e links source node s and target node t, while μ(·) scales each relational ternary adaptively. The attention operation involves combining the outputs from all h attention heads to generate the attention vector for each node pair (s,t). Essentially, it performs a softmax operation to create a probability distribution tailored to each target node t, derived from the accumulated attention vectors from neighboring nodes N(t).

Heterogeneous message passing. In this module, meta-relationships between edges are incorporated into the message-passing mechanism to address the discrepancies in how different node and edge types are distributed.

$$\mathrm{Message}_{HGT}(s,e,t) = \|_{i\in[1,h]}\mathrm{MSG}-\mathrm{head}^{i}(s,e,t)$$
$$\mathrm{MSG}-\mathrm{head}^{i}(s,e,t) = \text{M-Linear}^{i}_{\tau(s)}\left(H^{(l-1)}[s]\right)W^{\mathrm{MSG}}_{\varphi(e)}$$

Target-Specific aggregation. In this module, the results from the heterogeneous mutual attention and message passing are aggregated from the source node to the target node:

$$\tilde{H}^{(l)}[t] = \bigoplus_{\forall s\in N(t)}\left(\mathrm{Attention}_{HGT}(s,e,t)\cdot\mathrm{Message}_{HGT}(s,e,t)\right)$$

These results are then passed through an ELU activation layer, followed by linear transformation and residual connections:

$$H^{(l)}[t] = \sigma\left(\text{A-linear}_{\tau(t)}\tilde{H}^{(l)}[t]\right) + H^{(l-1)}[t]$$

Contrastive learning

We adopt unsupervised contrastive learning to obtain discriminative and transferable representations of protein–ligand interactions.

Given a complex G, we generate two augmented graph views $\tilde{G}^{(1)},\tilde{G}^{(2)}$via subgraph removal, atom masking, and bond deletion. Each view is encoded by a heterogeneous graph encoder $f_{\theta}(\cdot)$ and a projection head $g(\cdot)$; the resulting vectors are ℓ_2 — normalized:

$$\mathbf{z}^{(v)} = \frac{g\left(f_{\theta}\left(\tilde{G}^{(v)}\right)\right)}{\left\|g\left(f_{\theta}\left(\tilde{G}^{(v)}\right)\right)\right\|_2}\in\mathbb{R}^{d}, v\in\{1,2\}.$$

In a mini-batch of M complexes we obtain $2M$ views; the two views of the same complex form a positive pair, while all other views serve as in-batch negatives. We use cosine similarity$\mathrm{sim}(\mathrm{u},\mathrm{v}) = \mathrm{u}^{\top}\mathrm{v}$.

For view i with its positive index j and temperature τ,

$$\ell(i,j) = -\log\frac{\exp\left(\frac{\mathrm{sim}\left(\mathbf{z}_i,\mathbf{z}_j\right)}{\tau}\right)}{\sum_{k=1,k\neq i}^{2M}\exp\left(\frac{\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_k)}{\tau}\right)}.$$

The symmetric batch objective is

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{2M}\sum_{c=1}^{M}\left[\ell(i_c,j_c) + \ell(j_c,i_c)\right].$$

This objective maximizes agreement between two augmentations of the same complex while pushing apart different complexes, yielding invariance to perturbations and sensitivity to cross-sample differences. Contrastive pre-training drives the encoder to pull together representations of interacting protein–ligand views and to repel mismatched pairs, providing a strong initialization for downstream binding-affinity regression and structure-based virtual screening. The temperature τ controls distribution sharpness, and $g(\cdot)$ is a small MLP used only in the contrastive space, while encoder outputs feed the downstream heads.

Architecture overview

Our framework consists of two stages: self-supervised contrastive pre-training and task-specific fine-tuning.

Pre-training. Each protein-ligand complex is converted into a directed, typed heterogeneous graph with two node types and four relation types. To construct the two views used in contrastive pre-training, we apply stochastic graph augmentations to the directed heterogeneous complex graph. Concretely, for each protein-ligand complex we create three candidate augmented graphs, each generated by one of the following operators (all applied with an augmentation ratio p= 0.2): (1) Atom masking: with probability p, we randomly select nodes and either remove them together with all incident edges. (2) Bond deletion: we randomly drop p fraction of edges while keeping all nodes. The operator is applied to both intra-molecular covalent edges and inter-molecular contact edges. (3) Subgraph removal: we sample a connected subgraph containing approximately p fraction of nodes using a random-walk expansion from a randomly chosen seed node. During pre-training, we randomly sample two views from the above candidates to form a positive pair for the NT-Xent (InfoNCE) objective. Each view is encoded by a Heterogeneous Graph Transformer[15] encoder f_θ; a small projection head $g(\cdot)$ maps the graph readout to a contrastive space. We optimize f_θ with the symmetric InfoNCE (NT-Xent) loss using in-batch negatives, optionally adding a protein–ligand pairing loss to explicitly pull together true pairs and push apart mismatched pairs. This stage yields a domain-specific encoder that captures interaction-relevant and augmentation-invariant representations.

Fine-tuning. For downstream tasks, we reuse the pretrained encoder as a feature extractor and attach lightweight heads: Binding affinity regression: an MLP on the graph readout with a regression objective (MSE). Structure-based virtual screening: an MLP classifier with a sigmoid output and class-balanced cross-entropy . Unless stated otherwise, we fine-tune the entire network end-to-end. This two-stage design provides strong initialization and consistently improves convergence speed and final accuracy on both affinity prediction and virtual screening.

Dataset preparation

Pre-training. Following CL-GNN, we compiled 368,896 unique unlabeled protein-ligand complexes from BioLiP for self-supervised pre-training. Structure preparation. Because raw PDB entries often contain missing or inconsistent records, we reprocessed all complexes with PDBFixer: (i) identify and rebuild missing residues and heavy atoms; (ii) replace non-standard residues with their standard counterparts; (iii) remove all heterogens (including water/ions/metals/cofactors), and add missing heavy atoms.; (iv) add missing hydrogens.

View Article Online DOI: 10.1039/D5SC09548D

For each complex we generate two graph views via subgraph removal, atom masking, and bond deletion, ensuring both views preserve pocket–ligand feasibility.

Each view is encoded by the heterogeneous graph encoder fθ and a projection head g(·). During self-supervised pre-training, the HCGT-PL encoder was optimized using only the NT-Xent contrastive loss, i.e., InfoNCE with symmetric cosine similarity, with a temperature parameter of τ=0.05. The model was pre-trained for 2,000 epochs with a batch size of 512 using the Adam optimizer, with a learning rate of 0.001, β1=β2=0.9, and a weight decay of 1×10−6. No gradient accumulation was applied. Although the model implementation supports an optional protein–ligand pairing loss, this auxiliary loss was not used in the final reported model. The downstream affinity regression model was fine-tuned on the labeled PDBbind dataset using MSE loss with a batch size of 64.

We randomly split the pre-training set into a training split and a validation split of 20,000 complexes. Checkpoints are selected by the best validation InfoNCE objective. No supervised labels are used at this stage.

Downstream test. We trained and validated our encoder-head model on PDBbind v2016 (general + refined), in line with prior work. From the PDBbind website we collected 13,285 complexes with experimentally determined affinities. After excluding entries that could not be parsed or repaired, 12,904 complexes remained and were randomly split into train (N = 11,904) and validation (N = 1,000). Affinity labels were converted to pK values, using reported Kd or Ki (larger is stronger binding). To assess generalization, we evaluated on PDBbind Core 2013 (N = 107) and Core 2016 (N = 285), [16] and additionally compared on CSAR-HiQ.[17] Pre-training, training and validation sets share no overlap with the test sets. Structures underwent the same preparation as in Pre-train.

Heads, loss, and metrics. A lightweight MLP head is attached to the graph readout for regression; we optimize with MSE (Huber as a robustness check). We report RMSE and Pearson's correlation coefficient r (Pearson's r), computed on complexes with valid labels.

Large-scale SBVS is cast as binary classification (actives vs. decoys) at the target level. We train on DUD-E[18](diverse actives paired with property-matched decoys across many targets) and evaluate generalization on Three independent benchmarks: DEKOIS2.0[19] (harder decoys to reduce artificial enrichment), TrueDecoy_set and TrueDecoy$_{gap}$ set.[20]

Training protocol and evaluation. We attach a classification MLP (sigmoid output) atop the pretrained encoder and optimize with class-balanced cross-entropy (focal loss as an option for class

imbalance). We evaluate early-enrichment metrics (EF@0.1/0.5/1/5%) per target, then report both per-target scores and cross-target averages. EF@k is computed per target and then summarized across targets (mean/median), consistent with the violin and heatmap presentations.Thresholds are selected on the validation targets and applied unchanged to test targets.

Results

Comparative evaluation of binding-affinity prediction

We first evaluated the proposed heterogeneous contrastive graph Transformer on the PDBbind v.2016 and PDBbind v.2013 core sets, comparing its performance with a wide range of state-of-the-art methods. As shown in Table 1, HCGT-PL achieves highly competitive results on both benchmarks. On the PDBbind v.2016 Core Set, our method reached an RMSE of 1.210 and Pearson's r of 0.845, which is the best Pearson's r among all compared models and comparable to the lowest RMSE (1.209 from CL-GNN). This indicates that the proposed framework not only preserves predictive accuracy but also captures the binding affinity trends more consistently, reflecting a strong ability to learn discriminative interaction patterns. On the PDBbind v.2013 Core Set, HCGT-PL obtained an RMSE of 1.512 and Pearson's r of 0.762, which are competitive with existing methods. While the RMSE is close to the best-performing baselines (e.g., IGN with 1.372), the correlation remains slightly lower compared to CL-GNN (0.815) and IGN (0.832). This observation suggests that although HCGT-PL generalizes well to older datasets, further optimization may be required to fully capture the variability present in earlier versions of PDBbind.

Table 1. Performance comparison of different models on the PDBbind v.2016 and PDBbind v.2013 Core Sets. Results are reported in terms of root mean-square error (RMSE, the lower the better) and Pearson correlation coefficient (Pearson's r , the higher the better). The best results in each column are highlighted in bold.

Models	PDBbind v. 2016 Core Set		PDBbind v. 2013 Core Set	
	RMSE	Pearson's r	RMSE	Pearson's r
RF-Score[21]	1.488	0.785	1.616	0.776
PotentialNet[22]	1.503	0.772	1.607	0.773
GNN-DTI[23]	1.384	0.779	1.533	0.767
RosENet[24]	1.24	0.820	1.430	0.800
Pafnucy[25]	1.420	0.780	1.620	0.700
AGL-Score[26]	1.733	0.833	1.973	0.793
OnionNet[27]	1.278	0.816	1.503	0.782
AEScore[28]	1.300	0.800	1.460	0.760
PSH-ML[29]	1.280	0.835	1.482	0.783
PerSpect ML[30]	1.724	0.840	1.956	0.793
BAPA[31]	1.308	0.819	1.457	0.771
IGN[32]	1.220	0.837	1.372	0.832
PointTransformer[33]	1.260	0.833	1.613	0.781
CL-GNN[14]	1.200±0.009	0.838±0.004	1.345±0.004	0.812±0.003
HCGT-PL	1.238±0.041	0.842±0.003	1.441±0.004	0.783±0.003

Impact of the protein binding pocket on binding affinity

Without invoking pre-training, we systematically evaluated the impact of the pocket cropping radius r∈{5,6,7,8,9,10} Å on binding-affinity prediction performance (Table 2). On the PDBbind 2016 core set, the highest correlation was obtained at r=5 Å (Pearson's r=0.821). Increasing the radius from 7 to 9 Å consistently worsened performance, with RMSE increasing from 1.362 to 1.530 and correlation decreasing from 0.790 to 0.777. A consistent trend was observed on the PDBbind 2013 core set: r=5 Å yielded superior results (RMSE=1.498; Pearson's r=0.771), whereas expanding the radius to 9 Å markedly

View Article Online
DOI: 10.1039/D5SC09548D

degraded performance (RMSE = 1.883; Pearson's r=0.546). These findings indicate that, for a model trained from scratch, smaller pockets (5-6 Å) provide a higher signal-to-noise ratio and accentuate local interactions; in contrast, including many distal residues introduces noise and structural heterogeneity that exceed the model's ability to disentangle informative signals at initialization, leading to higher errors and lower correlations. Balancing robustness and computational cost, we therefore adopt r=5 Å as the default in subsequent experiments.

Table 2. Effect of pocket radius on model performance (no pretraining)

Å	PDBbind v. 2016 core set		PDBbind v. 2013 core set	
	RMSE	Pearson's r	RMSE	Pearson's r
5	1.333	0.821	1.498	0.771
6	1.300	0.808	1.504	0.750
7	1.362	0.790	1.543	0.740
8	1.431	0.791	1.629	0.717
9	1.530	0.777	1.883	0.546

Regression accuracy and ranking performance on comprehensive benchmarks

We plot the squared error as a function of experimental $p K = -logK_d/K_i$ and fit it with a three-segment piecewise linear model (PWL-3).[34] The curve exhibits a characteristic "U" shape: errors are minimal around the near-neutral/moderate affinity regime (pK≈6.5–8.0), but rise markedly at the extremes of very strong or very weak affinity. This suggests that (i) training samples are denser and have a higher signal-to-noise ratio in the middle range; and (ii) extreme regimes are more prone to measurement noise, pocket conformational heterogeneity, and systematic contributions from solvent/entropy, all of which inflate residuals. The segmented pattern motivates subsequent error calibration (e.g., interval-wise isotonic regression or temperature scaling) and indicates a need to augment data or inject physical priors in the extreme ranges to curb extrapolation error.

On the CASF-2016 ranking benchmark, our method achieves the highest ⟨p⟩ score, outperforming traditional scoring functions (e.g., X-Score[35], DrugScore[36]) and learning-based baselines (e.g., ChemPLP[37], iScore-Hybrid[38]). This indicates that the proposed heterogeneous graph Transformer captures relative affinity trends within ligand series more robustly, maintaining pairwise ranking consistency even under substantial conformational variation and broad chemical diversity. Together with the interval-wise error analysis above, this implies particularly strong discriminability in the moderate-affinity regime, which lifts overall ranking metrics.

On CASF-2016 (PDBbind v2016 core set), the model attains Pearson's r =0.845 and RMSE=1.210 (as annotated in the figure). The regression line is essentially parallel to the identity line in the scatter p lot, indicating not only reduced absolute error but also preservation of the global linear relationship both within and across targets. On the independent CSAR-HiQ dataset, our approach surpasses representative baselines—including RF-Score, PotentialNet, GNN-DTI, and IGN—on both RMSE and Pearson's r , demonstrating stronger cross-target generalization and robustness.

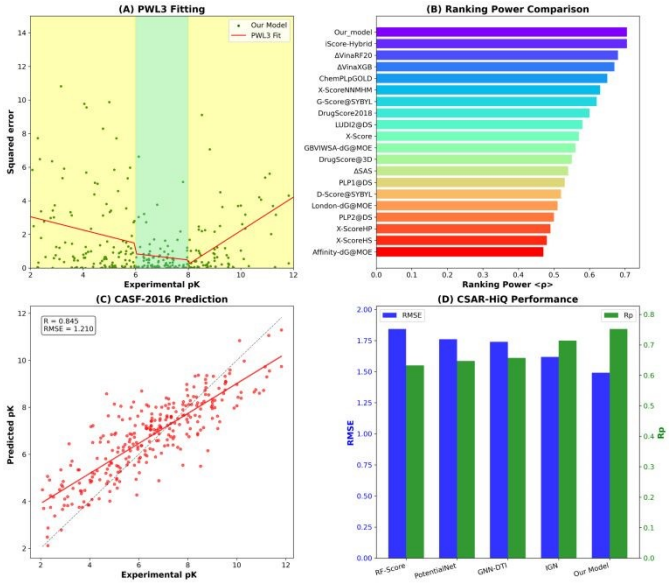

Figure 2. Comprehensive evaluation of the heterogeneous contrastive graph Transformer for protein–ligand binding affinity. (A) Residual structure captured by a three-segment piecewise linear fit (PWL-3). Squared prediction error is plotted against experimental pK (x-axis; labeled "Experimental pK" in the panel); the red line denotes the PWL-3 fit. (B) Ranking power comparison (average ⟨p⟩). (C) CASF-2016 regression: scatterplot of experimental vs. predicted pK. (D) Cross-dataset generalization on CSAR-HiQ: comparison of RMSE (left axis, blue) and Pearson's r (right axis, green).

Ablation analysis of contrastive learning pre-trained models

To evaluate the impact of contrastive pre-training on model performance, we conducted an ablation study using varying pre-training factors. Table 3 presents the results on the PDBbind v.2016 and PDBbind v.2013 core sets with models trained without pre-training and with pre-training factors (Temperature τ) of 0.01, 0.05, 0.1, 0.5, and 1. The baseline model without pre-training achieved an RMSE of 1.333 and an Pearson's r of 0.821 for the PDBbind v.2016 core set, indicating suboptimal performance. When pre-training was applied with a factor of 0.01, the model's performance improved slightly, with RMSE of 1.287 and Pearson's r of 0.832. The best performance was observed at a pre-training factor of 0.05, where RMSE reached 1.210 and Pearson's r increased to 0.845, indicating that contrastive pre-training significantly enhances the model's ability to capture transferable interaction representations. However, further increases in the pre-training factor (0.1 to 1) resulted in diminishing returns, with RMSE increasing to 1.282 and 1.257, and Pearson's r decreasing slightly to 0.818 and 0.819, respectively. These findings suggest that while contrastive pre-training improves model generalization, the optimal factor is around 0.05, beyond which the improvements plateau or decline. This ablation study highlights the critical role of contrastive pre-training in enhancing the performance of the HCGT-PL model for protein-ligand binding affinity prediction and virtual screening.

Table 3. Ablation Results of Contrastive Learning Pre-training on Model Performance

t	PDBbind v. 2016 core set		PDBbind v. 2013 core set	
	RMSE	Pearson's r	RMSE	Pearson's r
Without pre-train	1.333	0.821	1.498	0.771
0.01	1.287	0.832	1.522	0.754
0.05	1.210	0.845	1.437	0.786
0.1	1.282	0.818	1.501	0.760
0.5	1.257	0.819	1.487	0.765
1	1.276	0.814	1.482	0.758

Practical applications of HCGT-PL: Quantitative affinity vs MD-based MM/GBSA.

In binding free-energy estimation, MD-MM/GBSA[32] has long been used as a reference method for its clear physical intuition and manageable computational cost. Accurate MD workflows are time-consuming, which constrains their use in large-scale virtual screening. Deep learning scoring can be orders-of-magnitude faster, so a head-to-head comparison with MD-MM/GBSA clarifies when a learned model is a viable drop-in or complementary replacement. Zou *et al.*[39] systematically juxtaposed their Deep learning model with MD-MM/GBSA; adopting the same datasets and measurement conventions (including MAE) ensures that differences primarily reflect the scoring function rather than implementation.

To align with Zou *et al.*,[39] we evaluated the same (aligned) set of protein–ligand complexes and compared three binding free energies: the experimental reference (ΔG_{exp}), the MD–MM/GBSA[38] estimate ($\Delta G_{MM/GBSA}$), and HCGT-PL prediction (ΔG_{pred}). When the model outputs pK_d , we convert it to free energy using $\Delta G = -RT\ln K_d$ at T=300 K, (with $pK_d = -\log_{10} K_d$), ensuring consistent units and enabling direct comparison with the literature. For each complex, the absolute error is defined as $|\Delta\Delta G| = |\Delta G - \Delta G_{\mathrm{exp}}|$; across complexes, performance is summarized by the mean absolute error (MAE).

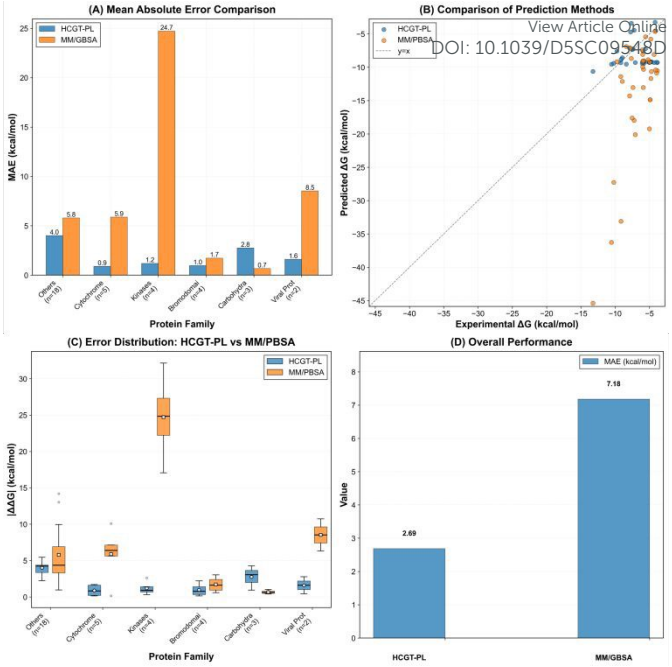

Figure 3. Comparison of Different Prediction Methods. (A) Shows the Mean Absolute Error (MAE) of the prediction model and MM/GBSA across different protein families. Blue bars represent the prediction model, and orange bars represent MM/GBSA. (B) A scatter plot comparing predicted ΔG with experimental ΔG, illustrating the correlation between predicted and experimental values. (C) Box plot comparison of the error distribution between the prediction model and MM/GBSA. (D) Overall performance comparison, where the blue bars represent the MAE of the prediction model, and the orange bars represent the MAE of MM/GBSA.

Figure 3 provides a benchmark comparison between the HCGT-PL and MM/GBSA workflow under a consistent evaluation protocol. We note that these two approaches represent fundamentally different paradigms: HCGT-PL is a data-driven statistical model trained to minimize prediction error on PDBbind-like supervision, whereas MM/GBSA is an approximate physics-based simulation whose accuracy depends on force-field parameterization, conformational sampling, and end-point free-energy approximations. Therefore, the observed differences should be interpreted as a practical trade-off on this benchmark rather than a universal superiority of one paradigm over the other. In Fig. 3A, HCGT-PL achieves lower MAE across multiple protein families than MM/GBSA in this setting, suggesting that the learned interaction representations can serve as an efficient surrogate scorer for affinity estimation and ranking. Fig. 3B shows that, despite some deviations, the model captures the overall trend of experimental ΔG variations, indicating that it encodes informative protein-ligand interaction patterns. Fig. 3C further summarizes the error distribution, where HCGT-PL exhibits a narrower spread for several systems, especially those that are challenging for the chosen MM/GBSA setup. Consistently, Fig. 3D reports a reduced error for HCGT-PL relative to MM/GBSA, including an average error reduction of 62.5% under the current protocol. For difficult cases such as GSK3β/CDK2 (e.g., 4ACG/4ACC/4ACM), Table S12 shows that MM/GBSA can yield large |ΔΔG| deviations (32.15/25.72/23.95 kcal/mol), whereas HCGT-PL remains within 0.8–3.75 kcal/mol error. This discrepancy may reflect the sensitivity of MM/GBSA

to sampling and parameter choices in highly polar and conformationally heterogeneous systems, while the learned model provides a stable estimate within its training distribution. In contrast, for systems such as 1OH4 dominated by aromatic stacking and hydrogen bonding, both approaches are closer to the experimental values, suggesting that when the dominant interactions are well captured, MM/GBSA and HCGT-PL can both provide reasonable estimates. Overall, these results support the complementary use of the two paradigms: HCGT-PL for high-throughput prescreening/ranking and MM/GBSA for physics-based refinement and mechanistic analysis on a smaller subset.

Performance comparison of virtual screening

We treat virtual screening as a binary classification ranking task: for each target, active molecules (actives) and decoy molecules (decoys) are scored and ranked separately. The receptor preprocessing, pocket cropping, and scoring/post-processing workflows are kept consistent across methods, with the following metrics used: enrichment factor EF{0.5%, 1%, 2%, 5%}, and compared against three major baselines: (i) traditional docking/scoring function-based methods (Glide_SP[40]); (ii) machine learning-based scorers (RF-Score[18]); (iii) deep learning methods (IGN[32]). All baselines are either reproduced under the public implementation or settings from the original papers, or directly cited from published results.

As shown in Fig. 4, Across EF@0.1–5%, HCGT-PL achieves the highest median and interquartile range, with the largest gains at EF@0.1% and EF@0.5%, indicating superior early enrichment. IGN ranks second, while RF-score and Glide_SP trail behind. As the threshold widens to 1% and 5%, differences narrow but HCGT-PL remains top or tied, suggesting better cross-target consistency. These gains likely stem from relation-specific attention over heterogeneous protein–ligand graphs and contrastive pretraining that enhances transferable interaction representations.

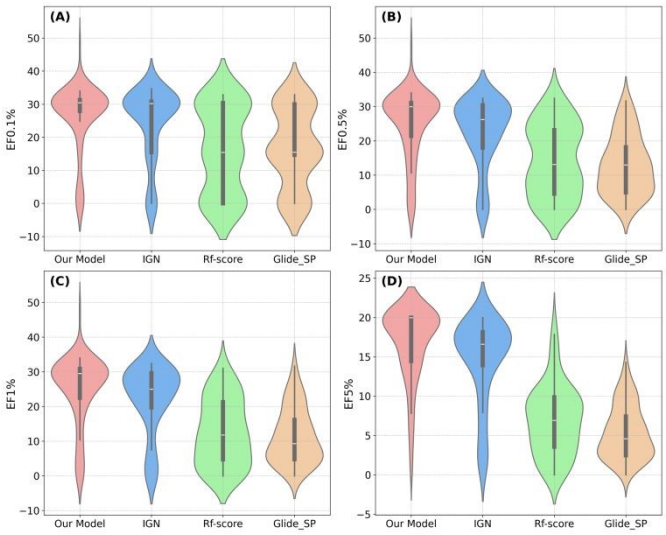

Figure 4. Early-enrichment comparison across methods (violin plots). (A–D) show EF@0.1%, 0.5%, 1%, and 5%, respectively. The x-axis lists the methods (HCGT-PL, IGN, RF-score, Glide_SP); the y-axis reports the enrichment factor (EF; higher is better). In each violin, the central line marks the median, the gray box denotes the interquartile range (IQR), and the violin width reflects the per-target density.

Fig. 5 shows a heatmap comparison of EF@0.1%, 0.5%, 1%, and 5% across different targets. Overall, HCGT-PL exhibits deeper color scales at most targets, with the most significant advantage in the early enrichment regions of EF@0.1%/0.5%. IGN is the strongest baseline, but still weaker than HCGT-PL on most targets. RF-score and Glide_SP generally show shallower results. As the threshold widens from 0.1% to 5%, the differences between methods narrow, but HCGT-PL's advantage remains highly consistent across all thresholds, demonstrating better robustness across both thresholds and targets. For a few targets, baseline methods are close to or slightly better than HCGT-PL, suggesting that these targets may have stronger regularity or be better suited for physics/empirical-based scoring. Overall, this heatmap validates the broad early enrichment improvement and ranking stability of our method across most targets.

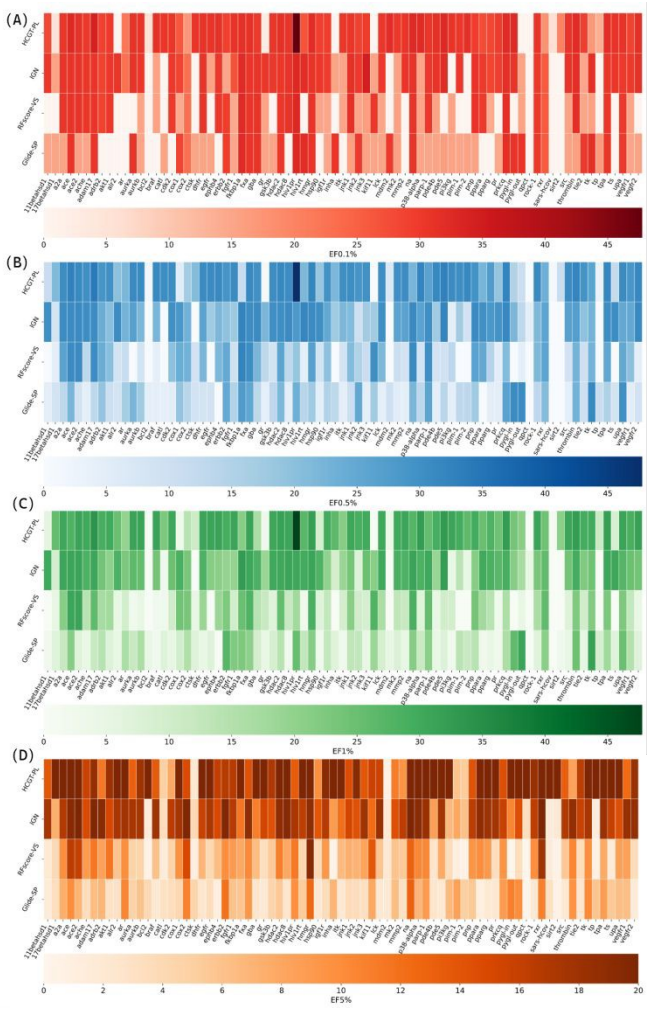

Figure 5. Target-wise early enrichment heatmaps. (A–D) show EF@0.1%, 0.5%, 1%, and 5%, respectively. Rows correspond to methods (HCGT-PL, IGN, RF-score, Glide_SP); columns correspond to individual targets. Cell color encodes the enrichment factor (EF), with the color bar indicating the value range—darker colors denote higher EF (better performance).

View Article Online
DOI: 10.1039/D5SC09548D

We evaluated HCGT-PL for SBVS on the TrueDecoy and TrueDecoygap datasets following Gu et al.[20], comparing against two families of baselines: (i) traditional docking/empirical scoring (Glide[40], rDock[41], Surflex[42], LeDock[43] and their combinations with RTMScore[44]; KarmaDock[45]) and (ii) learning-based/deep-learning scorers (RF-Score[46], IGN[32], DiffDock[47], FlexPose[48]). Performance was summarized by EF@0.5%, 1%, and 5%.

To avoid confounding factors, we distinguish pose-generation from pose-scoring in SBVS. For docking baselines (Glide/rDock/Surflex/LeDock), ligand poses are obtained from docking under the same receptor preparation and pocket definition; the Top-1 pose and the corresponding docking score are used for ranking. For docking+rescoring variants (e.g., RTMScore), we rescore the same docked pose using the specified rescoring function. For learning-based scorers (e.g., RF-Score, IGN) that require a pose as input, we use a shared set of docked poses generated once with a fixed docking protocol, and apply different scorers on exactly the same poses. For learning-based methods that generate poses (e.g., DiffDock, FlexPose), we run the official implementations with default pipelines/settings and use their Top-1 predicted pose for ranking.

As shown in Fig. 6, violin plots of cross-target EF distributions indicate that at the most stringent cutoffs (EF@0.5% and EF@1%), HCGT-PL achieves a higher median than all—or the vast majority of—baselines. The distribution is visibly shifted toward larger EF values with a "thick" interquartile range (IQR), implying that early-enrichment gains arise across many targets rather than being driven by a few outliers. In contrast, several docking/empirical methods exhibit violin bodies concentrated in the low-EF regime, with medians near zero, underscoring limited ability to prioritize the very top of the ranked lists. On the more challenging TrueDecoy$_{gap}$ benchmark (Fig. 6D-E), HCGT-PL retains the highest median and a similarly substantial IQR, demonstrating consistent, target-wide early enrichment even when actives and decoys are harder to distinguish. Collectively, these results suggest that the learned interaction representations in HCGT-PL capture transferable binding determinants that translate into robust, practically valuable early retrieval in SBVS.

Like other structure-based learning scoring functions, HCGT-PL depends on the quality of the input 3D pose. When crystal structures are used, the model receives experimentally supported protein-ligand geometries and can learn reliable contact patterns. When docked poses are used, inaccurate ligand orientations or distorted contacts may perturb intermolecular distances, angles, and edge construction, leading to reduced prediction accuracy. Nevertheless, HCGT-PL may retain a degree of robustness because it integrates atom-level chemical descriptors, intramolecular bond information, directed protein-ligand contact edges, relation-specific attention, and contrastive pre-training, rather than relying on a single geometric feature. These design choices help the model focus on chemically meaningful interaction motifs and transferable local patterns. Future work will further improve pose robustness by incorporating multiple docked poses or conformational ensembles, pose-noise augmentation, receptor flexibility, and solvent-aware descriptors.

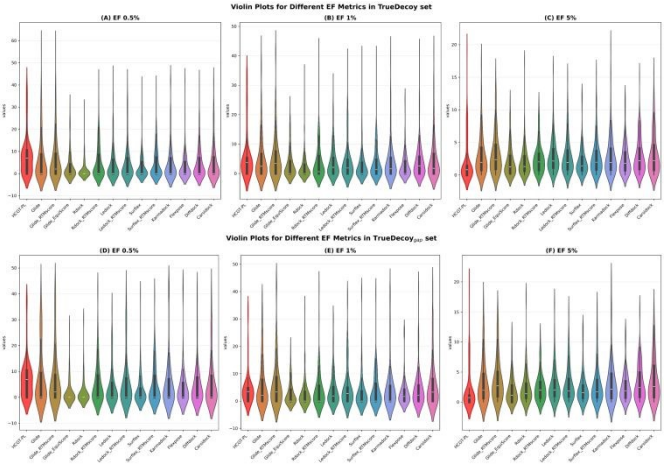

Figure 6. Early-enrichment distributions on the TrueDecoy and TrueDecoy$_{gap}$ benchmarks. Panels (A–C) report EF@0.5%, EF@1% and EF@5% on TrueDecoy_set; panels (D–F) show the corresponding results on TrueDecoy$_{gap}$_set.

Interpretability case study

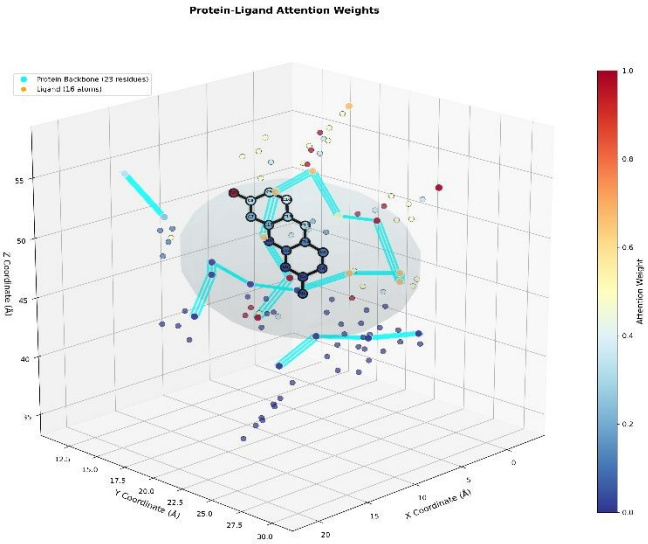

Figure 7. Interpretability visualization based on a Heterogeneous Graph Transformer (PDB ID: 1BCU). The protein backbone is shown as a cyan trace, the binding pocket as a gray envelope, and the ligand in a ball-and-stick representation. The displayed atom-wise importance weights indicate each atom's contribution to the model readout (prediction) in the current forward pass; higher weights denote greater contribution.

Model interpretability is essential for practical deployment. To visualize how ligand–protein atomic-pair interactions contribute to the final prediction, we analyzed the complex PDB ID: 1BCU. For each atom, we computed an importance score by aggregating its incoming attention mass across all relation types (edge categories), weighting by head and layer importance, and summing the contributions. As shown in Fig. 7, high-weight (orange/red) atoms cluster along the ligand surface within the pocket envelope; these hot spots cover the ligand's key

functional groups as well as receptor side chains that form hydrogen bonds, hydrophobic contacts, and π-mediated interactions with them. This pattern indicates that the model concentrates representational capacity on regions where physical interactions actually occur, rather than distributing weight uniformly inside and outside the pocket.

Within the ligand, heteroatom sites (donors/acceptors, charged/polar groups) and the edges of aromatic rings receive prominently higher weights, whereas the inert carbon scaffold is typically assigned medium-to-low weights. Such intramolecular selectivity suggests the model discerns not only whether contact occurs but also which chemical fragments are functionally responsible. Point clouds far from the ligand are mostly low-weight (blue), with only weak signals at a few remote locations, consistent with a "near-field–dominant, far-field–decaying" physical intuition. The resulting distance–attention trend implies that the model primarily relies on short-range interactions (hydrogen bonding, hydrophobic packing, π–π/π–cation) for discrimination, while retaining limited sensitivity to longer-range electrostatic or shape constraints.

Conclusions

This study introduces a HCGT-PL that unifies, in a single framework, heterogeneous graph representation (multiple node and relation types), relation-specific multi-head attention, contrastive pre-training, and downstream fine-tuning, thereby overcoming the limitations of homogeneous graphs in capturing the diverse semantics and cross-scale dependencies of biological systems. Systematic experiments show that HCGT-PL achieves both high accuracy and strong ranking power on binding-affinity regression and structure-based virtual screening, with stable generalization and robustness across multiple datasets, protein families, and pocket conditions; interpretability visualizations further reveal that the model concentrates within the pocket on ligand key functional groups and receptor side chains where interactions truly occur. Analysis of pocket-cropping radii and residual structure indicates that smaller radii improve the signal-to-noise ratio, whereas very strong/very weak binding regimes are more susceptible to measurement noise, solvent/entropy effects, and conformational heterogeneity; ablations of contrastive pre-training confirm that cross-sample invariant representations facilitate downstream convergence and enhance final performance. Taken together, HCGT-PL offers a unified route for protein–ligand modeling that is scalable, transferable, and interpretable, and it is poised to serve as a general component for affinity prediction and early-enrichment screening in drug discovery.

Despite these advances, there remains room for improvement: performance in extreme affinity regimes and in strongly polar/strong induced-fit scenarios is still constrained by upper-bound errors, which may be alleviated by injecting physical priors (explicit/implicit solvent terms, polarization and long-range electrostatics, many-body interactions, or energy regularization) and, going forward, by incorporating conformational ensembles and pocket dynamics to improve extrapolability. In parallel, dataset shift and target-label imbalance persist, calling for stronger robustness via multi-benchmark joint training, hard-example mining, and imbalance-aware learning.

Looking forward, several promising directions warrant further investigation. First, replacing a single static complex structure with conformational ensembles generated by molecular dynamics or enhanced sampling could better capture receptor flexibility and induced-fit effects, although this would increase computational cost. Second, incorporating explicit or implicit solvent effects, such as GB/SA-type solvation descriptors or learned water-placement representations, may improve affinity prediction for polar and solvent-exposed binding sites where electrostatic desolvation and water-mediated interactions strongly influence binding thermodynamics. Third, extending the self-supervised pre-training framework to larger structural and bioactivity databases, such as the full PDB and ChEMBL, may provide richer supervisory signals and further improve generalization to novel protein families and chemotypes. Finally, adapting HCGT-PL to covalent docking scenarios and allosteric binding sites represents an important direction toward broader applications in computational drug discovery.

Author contributions

Yunjiang Zhang (first author): Conceptualization; Methodology; Software; Investigation; Data curation; Writing – original draft. Chenyu Huang: Software; Methodology; Investigation. Xunyong Zhou: Formal analysis; Validation; Visualization. Yuetong Liu: Validation; Writing – review & editing. Qing Yuan: Validation; Writing – review & editing. Shaorui Sun (corresponding author): Conceptualization; Supervision; Project administration; Funding acquisition; Writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

The code generated in this study is available at https://github.com/Shaoruisun/HCGT-PL. All datasets used in this study are publicly available from their respective sources: BioLiP:https://www.aideepmed.com/BioLiP/; PDBbindv2016: https://www.pdbbind-plus.org.cn/; CASF-2016 Core Set and CASF-2013 Core Set: https://www.pdbbind-plus.org.cn/casf; CSAR-HiQ:http://www.csardock.org/; DUD-E: http://dude.docking.org/;DEKOIS2.0: https://www.dekois.com/; Original SBVS benchmark could be found at https://doi.org/ 10.5281/zenodo.13684010.

Acknowledgements

This work was supported by the National Key Research & Development Program of China (grant no. 2021YFA1201000).

References

1 R. J. Falconer, *J. Mol. Recognit.*, 2016, **29**, 504–515. https://doi.org/10.1002/jmr.2550.

2 W. Hou and S. B. Cronin, *Adv. Funct. Mater.*, 2013, **23**, 1612–1619. https://doi.org/10.1002/adfm.201202148.

3 E. Lionta, G. Spyrou, D. Vassilatis and Z. Cournia, *Curr. Top. Med. Chem.*, 2014, **14**, 1923–1938. https://doi.org/10.2174/1568026614666140929124445.

4 H. Li, K. Sze, G. Lu and P. J. Ballester, *WIREs Comput. Mol. Sci.*, https://doi.org/10.1002/wcms.1478.

5 Z. Zhang, L. Chen, F. Zhong, D. Wang, J. Jiang, S. Zhang, H. Jiang, M. Zheng and X. Li, *Curr. Opin. Struct. Biol.*, 2022, **73**, 102327. https://doi.org/10.1016/j.sbi.2021.102327.

6 S. Zhang, H. Tong, J. Xu and R. Maciejewski, *Comput. Soc. Netw.*, 2019, **6**, 11. https://doi.org/10.1186/s40649-019-0069-y.

7 P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y. Bengio, *arXiv*, 2018, preprint, arXiv:arXiv:1710.10903, https://doi.org/10.48550/arXiv.1710.10903.

8 C. Zhang, D. Song, C. Huang, A. Swami and N. V. Chawla, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, Anchorage AK USA, 2019, pp. 793–803. https://doi.org/10.1145/3292500.3330961.

9 W. Li, X. Li, M. Wang, F. Liu, Y. Luo, R. Guo and Q. Pan, *J. Biomol. Struct. Dyn.*, 2025, 1–13. https://doi.org/10.1080/07391102.2025.2475229.

10 K. Crampon, A. Giorkallos, X. Vigouroux, S. Baud and L. A. Steffenel, *Explor. Drug Sci.*, 2023, 126–139. https://doi.org/10.37349/eds.2023.00010.

11 J. Hu, M. Bewong, S. Kwashie, W. Zhang, V. M. Nofong, G. Wu and Z. Feng, in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Lisbon, Portugal, 2024, pp. 294–299. https://doi.org/10.1109/BIBM62325.2024.10822850.

12 B. Zhang, L. Quan, Z. Zhang, L. Cao, Q. Chen, L. Peng, J. Wang, Y. Jiang, L. Nie, G. Li, T. Wu and Q. Lyu, *J. Chem. Inf. Model.*, 2025, **65**, 1009–1026. https://doi.org/10.1021/acs.jcim.4c02073.

13 Y. Wang, J. Wang, Z. Cao and A. Barati Farimani, *Nat. Mach. Intell.*, 2022, **4**, 279–287. https://doi.org/10.1038/s42256-022-00447-x.

14 Y. Zhang, C. Huang, Y. Wang, S. Li and S. Sun, *J. Chem. Inf. Model.*, 2025, **65**, 1724–1735. https://doi.org/10.1021/acs.jcim.4c01290.

15 Z. Hu, Y. Dong, K. Wang and Y. Sun, *arXiv*, 2020, preprint, arXiv:arXiv:2003.01332, https://doi.org/10.48550/arXiv.2003.01332.

16 M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li and R. Wang, *J. Chem. Inf. Model.*, 2019, **59**, 895–913. https://doi.org/10.1021/acs.jcim.8b00545.

17 H. A. Carlson, R. D. Smith, K. L. Damm-Ganamet, J. A. Stuckey, A. Ahmed, M. A. Convery, D. O. Somers, M. Kranz, P. A. Elkins, G. Cui, C. E. Peishoff, M. H. Lambert and J. B. Dunbar, *J. Chem. Inf. Model.*, 2016, **56**, 1063–1077. https://doi.org/10.1021/acs.jcim.5b00523.

18 M. M. Mysinger, M. Carchia, John. J. Irwin and B. K. Shoichet, *J. Med. Chem.*, 2012, **55**, 6582–6594. https://doi.org/10.1021/jm300687e.

19 M. R. Bauer, T. M. Ibrahim, S. M. Vogel and F. M. Boeckler, *J. Chem. Inf. Model.*, 2013, **53**, 1447–1462. https://doi.org/10.1021/ci400115b.

20 S. Gu, C. Shen, X. Zhang, H. Sun, H. Cai, H. Luo, H. Zhao, B. Liu, H. Du, Y. Zhao, C. Fu, S. Zhai, Y. Deng, H. Liu, T. Hou and Y. Kang, *Nat. Mach. Intell.*, 2025, **7**, 509–520. https://doi.org/10.1038/s42256-025-00993-0.

21 P. J. Ballester and J. B. O. Mitchell, *Bioinformatics*, 2010, **26**, 1169–1175. https://doi.org/10.1093/bioinformatics/btq112.

22 E. N. Feinberg, D. Sur, Z. Wu, B. E. Husic, H. Mai, Y. Li, S. Sun, J. Yang, B. Ramsundar and V. S. Pande, *ACS Cent. Sci.*, 2018, **4**, 1520–1530. https://doi.org/10.1021/acscentsci.8b00507.

23 J. Lim, S. Ryu, K. Park, Y. J. Choe, J. Ham and W. Y. Kim, *J. Chem. Inf. Model.*, 2019, **59**, 3981–3988. https://doi.org/10.1021/acs.jcim.9b00387.

24 H. Hassan-Harrirou, C. Zhang and T. Lemmin, *J. Chem. Inf. Model.*, 2020, **60**, 2791–2802. https://doi.org/10.1021/acs.jcim.0c00075.

25 M. M. Stepniewska-Dziubinska, P. Zielenkiewicz and P. Siedlecki, *Bioinformatics*, 2018, **34**, 3666–3674. https://doi.org/10.1093/bioinformatics/bty374.

26 D. D. Nguyen and G.-W. Wei, *J. Chem. Inf. Model.*, 2019, **59**, 3291–3304. https://doi.org/10.1021/acs.jcim.9b00334.

27 Zheng L., Fan J. and Mu Y., *ACS Omega*, 2019, **4**, 15956–15965. https://doi.org/10.1021/acsomega.9b01997.

28 R. Meli, A. Anighoro, M. J. Bodkin, G. M. Morris and P. C. Biggin, *J. Cheminformatics*, 2021, **13**, 59. https://doi.org/10.1186/s13321-021-00536-w.

29 X. Liu, H. Feng, J. Wu and K. Xia, *Brief. Bioinform.*, 2021, **22**, bbab127. https://doi.org/10.1093/bib/bbab127.

30 Z. Meng and K. Xia, *Sci. Adv.*, 2021, **7**, eabc5329. https://doi.org/10.1126/sciadv.abc5329.

31 S. Seo, *BMC Bioinformatics*, 2021, **22**, 15. https://doi.org/10.1186/s12859-021-04466-0.

32 D. Jiang, C.-Y. Hsieh, Z. Wu, Y. Kang, J. Wang, E. Wang, B. Liao, C. Shen, L. Xu, J. Wu, D. Cao and T. Hou, *J. Med. Chem.*, 2021, **64**, 18209–18232. https://doi.org/10.1021/acs.jmedchem.1c01830.

33 Y. Wang, S. Wu, Y. Duan and Y. Huang, *Brief. Bioinform.*, 2022, **23**, bbab474. https://doi.org/10.1093/bib/bbab474.

34 S. J. Mahdizadeh and L. A. Eriksson, *J. Chem. Inf. Model.*, 2025, **65**, 2759–2772. https://doi.org/10.1021/acs.jcim.4c02192.

35 X. Zhang, X. Li and R. Wang, *J. Chem. Inf. Model.*, 2009, **49**, 1033–1048. https://doi.org/10.1021/ci8004429.

36 C. A. Sotriffer, H. Gohlke and G. Klebe, *J. Med. Chem.*, 2002, **45**, 1967–1970. https://doi.org/10.1021/jm025507u.

37 O. Korb, T. Stützle and T. E. Exner, *J. Chem. Inf. Model.*, 2009, **49**, 84–96. https://doi.org/10.1021/ci800298z.

38 T. Hou, J. Wang, Y. Li and W. Wang, *J. Chem. Inf. Model.*, 2011, **51**, 69–82. https://doi.org/10.1021/ci100275a.

39 Y. Zou, R. Wang, M. Du, X. Wang and D. Xu, *J. Phys. Chem. B*, 2023, **127**, 899–911. https://doi.org/10.1021/acs.jpcb.2c07592.

40 R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis and P. S. Shenkin, *J. Med. Chem.*, 2004, **47**, 1739–1749. https://doi.org/10.1021/jm0306430.

41 S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard and S. D. Morley, *PLoS Comput. Biol.*, 2014, **10**, e1003571. https://doi.org/10.1371/journal.pcbi.1003571.

42 A. N. Jain, *J. Med. Chem.*, 2003, **46**, 499–511. https://doi.org/10.1021/jm020406h.

43 N. Liu and Z. Xu, *IOP Conf. Ser. Earth Environ. Sci.*, 2019, **218**, 012143. https://doi.org/10.1088/1755-1315/218/1/012143.

44 C. Shen, X. Zhang, Y. Deng, J. Gao, D. Wang, L. Xu, P. Pan, T. Hou and Y. Kang, *J. Med. Chem.*, 2022, **65**, 10691–10706. https://doi.org/10.1021/acs.jmedchem.2c00991.

45 X. Zhang, O. Zhang, C. Shen, W. Qu, S. Chen, H. Cao, Y. Kang, Z. Wang, E. Wang, J. Zhang, Y. Deng, F. Liu, T. Wang, H. Du, L. Wang, P. Pan, G. Chen, C.-Y. Hsieh and T. Hou, *Nat. Comput. Sci.*, 2023, **3**, 789–804. https://doi.org/10.1038/s43588-023-00511-5.

46 H. Li, J. Peng, P. Sidorov, Y. Leung, K.-S. Leung, M.-H. Wong, G. Lu and P. J. Ballester, *Bioinformatics*, 2019, **35**, 3989–3995. https://doi.org/10.1093/bioinformatics/btz183.

47 G. Corso, H. Stärk, B. Jing, R. Barzilay and T. Jaakkola, *arXiv*, 2023, preprint, arXiv:arXiv:2210.01776, http://arxiv.org/abs/2210.01776.

48 T. Dong, Z. Yang, J. Zhou and C. Y.-C. Chen, *J. Chem. Theory Comput.*, 2023, **19**, 8446–8459. https://doi.org/10.1021/acs.jctc.3c00273.

View Article Online DOI: 10.1039/D5SC09548D

DOI: 10.1039/D5SC09548D

The code generated in this study is available at https://github.com/Shaoruisun/HGGT-PL. All datasets used in this study are publicly available from their respective sources:

BioLiP:https://www.aideepmed.com/BioLiP/; PDBbindv2016: https://www.pdbbind-plus.org.cn/; CASF-2016 Core Set and CASF-2013 Core Set: https://www.pdbbind-plus.org.cn/casf; CSAR-HiQ:http://www.csardock.org/; DUD-E: http://dude.docking.org/;DEKOIS2.0: https://www.dekois.com/; Original SBVS benchmark could be found at https://doi.org/ 10.5281/zenodo.13684010.