GDEGAN: GAUSSIAN DYNAMIC EQUIVARIANT GRAPH ATTENTION NETWORK FOR LIGAND BINDING SITE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate prediction of binding sites of a given protein, to which ligands can bind, is a critical step in structure-based computational drug discovery. Recently, Equivariant Graph Neural Networks (GNNs) have emerged as a powerful paradigm for binding site identification methods due to the large-scale availability of 3D structures of proteins via protein databases and AlphaFold predictions. The state-of-the-art equivariant GNN methods implement dot product attention, disregarding the variation in the chemical and geometric properties of the neighboring residues. To capture the variation in properties, we propose GDEGAN (Gaussian Dynamic Equivariant Graph Attention Network), which replaces simple dot-product attention with adaptive kernels that recognize binding sites. The proposed attention mechanism captures variation in neighboring residues using statistics of their characteristic local feature distributions. Our mechanism dynamically computes neighborhood statistics at each layer, using local variance as an adaptive bandwidth parameter with learnable per-head temperatures, enabling each protein region to determine its own context-specific importance. Our model shows better predictive performance, outperforming existing methods with relative improvements of 37-66 % in DCC and 7-19 % DCA success rates across COACH420, HOLO4k, and PDBBind2020 datasets. These advances have direct application in accelerating protein-ligand docking by identifying potential binding sites for therapeutic target identification.

1 INTRODUCTION

The functional behavior of proteins is governed mainly by their interaction with other molecules to modulate their function, such as small-molecule ligands (Du et al., 2016). These interactions are precise, occurring at well-defined geometric and chemical regions on the protein surface known as binding sites or pockets. Precisely predicting potential binding sites based on a protein's 3D structure is a foundational challenge in rational drug design (Zheng et al., 2013) and structural biology (Schomburg et al., 2014). The success of models like AlphaFold (Jumper et al., 2021; Abramson et al., 2024) in accurately predicting protein 3D structures has significantly advanced the capabilities of structure-based drug design methodologies (Tunyasuvunakool et al., 2021; Sadybekov & Katritch, 2023). Within this field, it is crucial to differentiate between two complementary computational tasks. The first, *protein-ligand binding sites* identification from 3D structures of proteins, is the fundamental challenge of discovering surface pockets capable of binding novel or unknown ligands. This is particularly vital for the majority of proteins with no known ligands or binding partners. The second, *docking* (Zhang et al., 2023; Stärk et al., 2022; Lu et al., 2022) builds upon this by predicting the precise binding pose and orientation of a known ligand within a target site. While historically these have been formidable challenges, both have recently seen significant progress driven by the application of geometric deep learning (Stärk et al., 2022; Lu et al., 2022; Zhang et al., 2023; Hussein et al., 2015; Gainza et al., 2020; Méndez-Lucio et al., 2021; Ganea et al., 2022; Satorras et al., 2021; Zhang et al., 2024; Schütt et al., 2018), leading to powerful new tools for drug discovery.

Ligand Binding Site (LBS) Identification. Over the years, various computational methods from classical machine learning to deep learning have been successfully employed for LBS identification, combining proteins' physical, chemical and geometric information. Earlier approaches, including

P2Rank (Krivák & Hoksza, 2018), a random-forest based technique that utilized protein surface information, and Fpocket (Le Guilloux et al., 2009), which depended on Voronoi tessellation and alpha spheres (Liang et al., 1998) for efficacy, are constrained by the limited expressivity of protein representation. Early applications of deep learning to this problem, pioneered by Convolution Neural Networks (CNNs) (LeCun et al., 2002), led to various methods including DeepSite (Jiménez et al., 2017), DeepPocket (Aggarwal et al., 2021) and DeepSurf (Mylonas et al., 2021) treating proteins as 3D volumetric data and applying 3D CNNs to predict binding regions. However, these voxel-based approaches under-perform due to their fundamental misalignment with the irregular, sparse nature of protein structures. More importantly, they are sensitive to the protein's orientation in 3D space (Zhang et al., 2024). These limitations motivated representing proteins more naturally as graphs, where the atoms or residues serve as nodes and the interactions between them are the edges. This perspective is perfectly suited for Graph Neural Networks (GNNs), and in particular, equivariant GNNs (Satorras et al., 2021), which have become the standard for 3D geometric deep learning. By design, these models respect the rotational and translational symmetries of the physical world, directly addressing the key failure of CNNs. Consequently, modern methods like EquiPocket (Zhang et al., 2024), which utilize EGNN (Satorras et al., 2021) as backbone, have proven to be powerful for LBS identification.

Equivariant Graph Neural Networks. Equivariant graph neural networks have progressed in two directions: scalarization-based models (Satorras et al., 2021; Schütt et al., 2018; 2021; Du et al., 2023) and high-degree steerable models (Batzner et al., 2022; Batatia et al., 2022; Musaelian et al., 2023; Qiao et al., 2022; Liao & Smidt, 2023; Liao et al., 2024). The scalarization-based models function by transforming 3D data, such as coordinates, into scalar characteristics (e.g., distance), hence enhancing computational efficiency and scalability. Nonetheless, their expressivity is constrained in contexts such as protein data modeling, where the comprehension of symmetry and spatial relationships is essential for capturing geometric patterns. In contrast, high-degree steerable models operate directly on rich geometric features (irreducible representations) via the Clebsch-Gordan product, preserving essential spatial relationships. Despite their robust theoretical foundation and superior performance, they are computationally intensive, particularly for larger graphs such as proteins Cen et al. (2024). GotenNet (Aykent & Xia, 2025) presents a solution that balances expressiveness and computing efficiency by implementing a spherical-scalarization model. Building on this, we adopt GotenNet (Aykent & Xia, 2025) as the backbone for our model, applying its efficient framework for processing higher-degree features to the protein-ligand binding site identification task.

Yet even Equivariant GNNs applied to proteins exhibit a critical limitation. While recent E(3)-equivariant methods, such as EquiPocket, the current state-of-the-art (Zhang et al., 2024) method, have significantly advanced ligand binding site identification. Their message-passing frameworks are inherently based on a static, context-agnostic attention mechanism known as dot-product attention or self-attention, formally expressed as $\alpha_{ij} \propto exp(f(h_i, h_j))$, where f quantifies similarity. These methods employ a globally fixed similarity metric to assess inter-atomic importance, which is fundamentally misaligned with the nature of proteins, as they are characterized by extreme structural and chemical heterogeneity.

Our Approach. Our approach is motivated by the key insight that binding sites often appear as statistically distinct, tightly clustered regions compared to the rest of the protein surface. Exploiting this property, we overcome the aforementioned limitations by replacing dot-product attention with dynamic, context-aware statistical fitting. Inspired by recent work in probabilistic attention for non-geometric modalities (Ioannides et al., 2024), we introduce Gaussian Dynamic Attention mechanism, adapted for the first time to equivariant graph representations. We build upon GotenNet architecture (Aykent & Xia, 2025) as it already handles the high-degree steerable features efficiently. Our proposed model GDEGAN, computes attention scores by measuring how statistically probable a neighboring atom's features are, given a Gaussian distribution defined by the target atom's local neighborhood. By dynamically computing the mean and variance of each atom's local environment at every layer, our attention mechanism becomes inherently adaptive. This adaptation to the emergent local geometry of the protein graphs provides a more powerful and physically grounded inductive bias, enabling the model to learn more robust representations of complex protein structures.

Contributions. In this work, we aim to improve on finding the most probable binding candidates for LBS identification task. We argue that a more powerful approach is to make the attention mechanism dynamic and context-aware. To this end, we make the following contributions:

- We introduce Gaussian Dynamic Attention mechanism that characterizes each atom's neighborhood using learnable Gaussian parameters. This design preserves E(3)-equivariance by computing attention from invariant local statistics.
- We investigate the use of high-degree steerable E(3)-Equivariant GNNs to the critical task of protein-ligand binding site identification, demonstrating their effectiveness in capturing complex geometries.
- We demonstrate through extensive experiments that our proposed GDEGAN surpasses state-of-the-art methods on multiple benchmarks and achieves a significant improvement in inference speed, validating the efficacy and efficiency of our adaptive attention design.

2 PRELIMINARIES

Protein Graph Representation. We represent a protein structure as a geometric graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$, where $\mathcal{V}$ denotes the set of N residues, $\mathcal{E}$ denotes edges between spatially proximate residues, and $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^N$ represents the 3D coordinates of C_α atoms. Each node $v_i \in \mathcal{V}$ is characterized by it's scaler features $h_i \in \mathbb{R}^{n_d}$ and equivariant features $\tilde{\mathbf{X}}_i^{(l)}$ of degree l. Edges connect residues within a spatial cutoff: $\mathcal{E} = \{(i,j) : \|\mathbf{p}_i - \mathbf{p}_j\| < r_c, i \neq j\}$, where r_c is set to 10 Å for capturing relevant interactions and n_d is initial node feature dimension.

To initialize nodes with dimension n_d, we use pre-trained ESM-2 embeddings $h_i \in \mathbb{R}^{n_d}$, which are then projected to hidden-dimensional features h_d using learned transformations. These embeddings capture sequence context and evolutionary patterns needed to identify binding sites. Nodes are labeled binding ($y_i = 1$) or non-binding ($y_i = 0$) based on closeness to ligand atoms during training. See Appendices A.1 and A.2 for more information on geometric representations and binding site definitions.

Equivariance and Invariance. In 3D geometric learning, the symmetry of physical laws necessitates that models adhere to the Euclidean group E(3), which includes rotations, reflections, and translations. A function f is E(3)-equivariant if for rotation/reflection $\mathbf{R} \in O(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$, it satisfies $f(g \cdot \mathbf{P}, h) = g \cdot f(\mathbf{P}, h)$, where $g \cdot \mathbf{P} = \mathbf{R}\mathbf{P} + \mathbf{t}$ denotes the group action on positions $\mathbf{P}$, and invariant features h. Invariant functions satisfy: $f(g \cdot \mathbf{P}, h) = f(\mathbf{P}, h)$, producing unchanged outputs for scalar quantities.

Task Formulation. Given a geometric protein graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$, we formulate the task as learning an equivariant model $f(\mathcal{G}, y)$ to predict the binding probabilities $\hat{y}_i \in [0, 1]$ for each residue $v_i \in \mathcal{V}$.

3 GDEGAN: GAUSSIAN DYNAMIC EQUIVARIANT GRAPH ATTENTION NETWORK

We enhance GotenNet (Aykent & Xia, 2025) by incorporating a Gaussian dynamic attention module. Although GotenNet attains superior performance in molecular property prediction via equivariant tensor attention, its uniform attention approach to atomic neighborhoods constrains its efficacy in predicting protein-ligand binding sites, where geometric heterogeneity is essential. To address this limitation, we leverage the observation of binding sites having different geometric patterns, with varying curvature and chemical properties. This motivates three key modifications: statistical attention that adapts to local variance, protein-specific embeddings, and directed supervision. The complete architecture of GDEGAN is illustrated in Figure 1.

3.1 EQUIVARIANT GEOMETRIC TENSORS

Here, we describe different representations for both scalars and tensors. Tensor representations are initialized using spherical harmonics to capture spatial information from rank 0 to L_{max}. Edge geometry is encoded via spherical harmonics: $\tilde{\mathbf{r}}_{ij}^{(l)} = Y^{(l)}(\hat{\mathbf{r}}_{ij})$ where $\hat{\mathbf{r}}_{ij} = (\mathbf{p}_i - \mathbf{p}_j)/\|\mathbf{p}_i - \mathbf{p}_j\|$ is the unit vector and $Y^{(l)} : S^2 \to \mathbb{R}^{2l+1}$ denotes the degree-l spherical harmonics that map the unit sphere to a $(2l+1)$ dimensional vector. These basis functions enable the network to process geometric information while preserving equivariance. For $l = 0$, $\tilde{\mathbf{r}}_{ij}^{(0)}$: scalar invariants; $l = 1$,

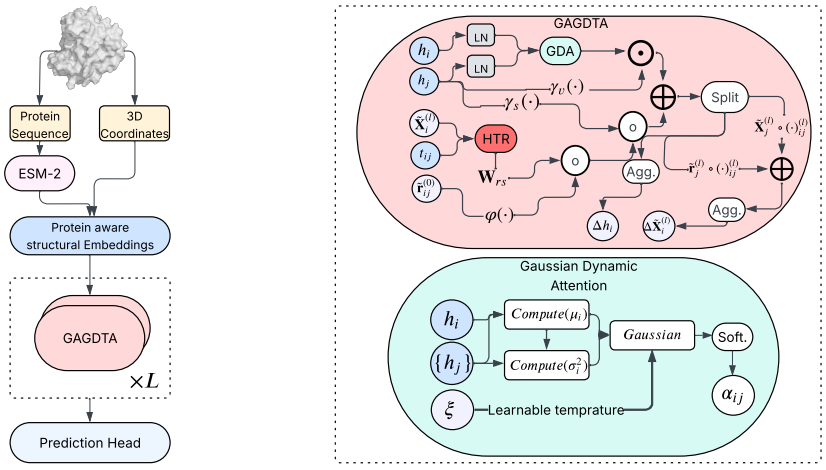

Figure 1: **GDEGAN architecture for protein ligand binding site identification. Left:** Overview of the GDEGAN framework showing the integration of protein-specific ESM-2 embeddings with geometric processing through L layers of Gaussian Dynamic Attention. **Right:** Detailed view of the Gaussian Dynamic Attention (GDA) module. In this $\oplus$, $\cdot$ and $\circ$ denotes addition, dot product and element-wise product respectively. HTR is inherited from GotenNet (Aykent & Xia, 2025). Soft. stands for Softmax, Agg. for Aggregation.

$\tilde{\mathbf{r}}_{ij}^{(1)}$: directional vectors; $l = 2$, $\tilde{\mathbf{r}}_{ij}^{(2)}$: quadrupole moments. Where, $\tilde{\mathbf{r}}_{ij}^{l}$ is initialized based on their relative positions $\mathbf{p}_i$ and $\mathbf{p}_j$ of nodes i and j in increasing order of geometric complexity, specifically $\tilde{\mathbf{r}}_{ij} = \{\tilde{\mathbf{r}}^{(0)}, \tilde{\mathbf{r}}^{(1)}, ..., \tilde{\mathbf{r}}^{(L_{max})}\}$ and $\tilde{\cdot}$ represents the steerable features. Node features comprise of invariant scalars $h \in \mathbb{R}^{n_d}$ and equivariant high degree steerable features $\tilde{\mathbf{X}}^{(l)} \in \mathbb{R}^{(2l+1)\times h_d}$ of degree $l \in \{1, ..., L_{\max}\}$. These features transform predictably under E(3) operations, with scalars remaining invariant and steerable features transforming according to their degree l where h_d denotes node embedding hidden dimensions.

3.2 Protein-Aware Structural Embeddings Diffused with Geometry

Unlike standard molecular GNNs that rely solely on atomic properties, our approach leverages pretrained protein representations while incorporating spatial relationships through geometric information, allowing efficient message passing for both nodes and edges. Given pre-computed ESM-2 (Lin et al., 2022) embeddings for each residue $h_i \in \mathbb{R}^{n_d}$, we construct initial node features through a two-stage process that combines evolutionary information with local structural context.

Neighborhood-Aware Message Aggregation. For each residue i, we aggregate information from spatial neighbors:

$$\mathbf{m}_i = \sum_{j \in \mathcal{N}(i)} \mathbf{W}_{\text{a}}(h_j) \circ \left(\phi(\tilde{\mathbf{r}}_{ij}^{(0)})\mathbf{W}_{\text{rbf}}\right) \circ \varphi(\tilde{\mathbf{r}}_{ij}^{(0)}) \tag{1}$$

where $\mathcal{N}(i) = \{j : \|\mathbf{p}_i - \mathbf{p}_j\| < r_c, j \neq i\}$ defines the spatial neighborhood, $\phi : \mathbb{R} \to \mathbb{R}^K$ represents K radial basis functions encoding distances, $\mathbf{W}_{\text{rbf}} \in \mathbb{R}^{K\times h_d}$ projects RBF features, $\mathbf{W}_{\text{a}} \in \mathbb{R}^{n_d\times h_d}$ projects node evolutionary features, $\circ$ denotes element-wise product and $\varphi(\tilde{\mathbf{r}}_{ij}^{(0)})$ is a smooth cutoff function ensuring differentiability.

Context-Enriched Feature Construction. The final node node features combining self and neighborhood information:

$$h_i^{(0)} = \mathbf{W}_u \left(\sigma \left(\text{LN}\left[\mathbf{W}_h(h_i \| \mathbf{m}_i\right] \mathbf{W}_d\right)\right) \tag{2}$$

where $\|$ denotes concatenation, LN is layer normalization for training stability, σ is the activation function, and $\mathbf{W}_h, \mathbf{W}_d, \mathbf{W}_u$ are learned projections. This formulation enables each residue to incorporate both its evolutionary signature and local structural environment.

Geometry-Aware Edge Scaler Features. Edge scaler features are initialized to capture pairwise relationships enhanced by spatial information:

$$t_{ij}^{(0)} = (h_i^{(0)} + h_j^{(0)}) \circ \left(\phi(\tilde{\mathbf{r}}_{ij}^{(0)})\mathbf{W}_{\text{e}}\right) \tag{3}$$

This symmetric formulation guarantees that $\mathbf{t}_{ij} = \mathbf{t}_{ji}$, preserving consistency in undirected protein graphs while integrating distance-dependent modulation via RBF-encoded spatial characteristics.

Here, $\mathbf{t}_{ij}^{(0)} \in \mathbb{R}^{e_d}$ represents the edge of dimension e_d, and $\mathbf{W}_\text{e}$ denotes a learned transformation matrix.

High Degree Equivariant Steerable features Initialization. These high degree steerable features that capture complex geometric information are initialized as $\mathbf{0}$ initially and are updated during attention aware feature update module, which will be discussed in the later section.

$$\tilde{\mathbf{X}}_i^{(l),(0)} = \mathbf{0} \in \mathbb{R}^{(2l+1)\times h_d}, \quad \forall l \in \{1, ..., L_\text{max}\} \tag{4}$$

3.3 Geometry Aware Gaussian Dynamic Tensor Attention

Unlike standard dot-product attention that treats all nodes uniformly, protein binding sites exhibit distinct geometric and chemical patterns: binding pockets are characterized by clustered residues with specific spatial arrangements, while surface regions show more dispersed distributions (Krivák & Hoksza, 2018). We hypothesize that high local chemical diversity translates to a high variance in a learned feature space of the surrounding residue. Therefore, local feature variance acts as a reliable signal for identifying functionally significant transition areas. Our Gaussian Dynamic Attention exploits this inherent heterogeneity by computing local neighborhood statistics (μ_i, σ_i^2) from the ESM-2 features h_i, enabling adaptive attention that responds to the local geometric context. The variance σ_i^2 modulates attention weights: high variance amplifies attention to capture complex binding site boundaries, while lower variance reduces unnecessary focus. Further, it is enhanced by incorporating spatial information. Specifically, for each residue i, with neighborhood $\mathcal{N}(i)$ we compute:

$$\mu_i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} h_j \tag{5}$$

$$(\sigma_i)^2 = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} (h_j - \mu_i)^2 \tag{6}$$

These statistics provide a distributional summary of the local neighborhood, capturing both the central tendency and spread of features.

Learnable Gaussian Parameters. We introduce H learnable variance parameter ξ that controls the temperature of attention for each of the H attention heads. This parameter adaptively modulates the sensitivity of attention to feature differences, allowing each head to specialize in different scales of molecular interactions. For molecular graphs, we compute the Gaussian attention score directly from pairwise node differences:

$$\alpha_{ij} = \frac{\exp\left(-\frac{||h_j - h_i||^2}{2\xi \cdot (\sigma_i^2 + \epsilon)}\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(-\frac{||h_k - h_i||^2}{2\xi \cdot (\sigma_i^2 + \epsilon)}\right)} \tag{7}$$

where σ_i^2 is the neighborhood variance providing context-aware scaling.

Integrating with Equivariant Features. To combine gaussian dynamic attention with equivariant features we follow GotenNet (Aykent & Xia, 2025) framework for updating both scalar and steerable features. Given attention scores α_{ij} from equation 7 we compute attention-weighted messages and combine them with geometric encoding:

$$\mathbf{o}_{ij} = \alpha_{ij} \cdot \gamma_v(h_j) + (t_{ij}\mathbf{W}_{rs}) \circ \gamma_s(h_j) \circ \varphi(\tilde{\mathbf{r}}_{ij}^{(0)}) \tag{8}$$

$$\{o_{ij}^s, \{o_{ij}^{d,(l)}\}_{l=1}^{L_\text{max}}, \{o_{ij}^{t,(l)}\}_{l=1}^{L_\text{max}}\} = \text{split}(\mathbf{o}_{ij}, h_d) \tag{9}$$

Here $\gamma_v, \gamma_s : \mathbb{R}^{h_d} \to \mathbb{R}^{S \cdot h_d}$ are MLPs and $\mathbf{W}_{rs} \in \mathbb{R}^{e_d \times (S \cdot h_d)}$ is a learnable weight matrix. S is multiplying factor to generate different coefficients for different l values calculated as $1 + 2 \times L_{max}$. Finally, features are updated maintaining equivariance:

$$\Delta h_i = \oplus_{j \in \mathcal{N}(i)}(o_{ij}^s) \tag{10}$$

$$\Delta \tilde{\mathbf{X}}_i^{(l)} = \oplus_{j \in \mathcal{N}(i)} \left[o_{ij}^{d,(l)} \circ \tilde{\mathbf{r}}_{ij}^{(l)} + o_{ij}^{t,(l)} \circ \tilde{\mathbf{X}}_j^{(l)} \right] \tag{11}$$

Here, each degree $l \in [1, L_{max}]$ contributes its own component and representations of residues are updated as follows:

$$h_i \leftarrow h_i + \Delta h_i, \quad \tilde{\mathbf{X}}_i^{(l)} \leftarrow \tilde{\mathbf{X}}_i^{(l)} + \Delta \tilde{\mathbf{X}}_i^{(l)} \tag{12}$$

By replacing uniform dot-product attention with geometry-aware Gaussian kernels that adapt to local feature distributions, GAGDTA enables precise discrimination between binding pockets and surface regions. This adaptive mechanism proves particularly effective for ligand binding site prediction, where the inherent heterogeneity of protein surfaces from tightly clustered binding pockets to dispersed surface residues demands context-aware attention patterns, as demonstrated in our experiments Section 4.

3.4 Hierarchical Processing and Equivariant Refinement

Following the GAGDTA layer, we adopt GotenNet (Aykent & Xia, 2025) hierarchical tensor refinement (HTR) and equivariant feed-forward (EQFF) modules with minimal modifications to maintain architectural consistency. To summarize:

Edge refinement via HTR. Edge features are refined using inner products of high-degree steerable features:

$$\mathbf{w}_{ij} = \text{Agg}_{l=1}^{L_{\max}} \langle \tilde{\mathbf{X}}_i^{(l)} \mathbf{W}_q, \tilde{\mathbf{X}}_j^{(l)} \mathbf{W}_k^{(l)} \rangle, \quad t_{ij} \leftarrow t_{ij} + \gamma_w(\mathbf{w}_{ij}) \circ \gamma_t(t_{ij}) \tag{13}$$

where $\mathbf{w}_{ij}$ is aggregated similarity between node i and j, $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{e_d \times e_d}$ are tensor projection matrices where, $\mathbf{W}_q$ is shared across degree $l \in [1, ..L_{max}]$ and $\mathbf{W}_k^{(l)}$ is degree specific. $\gamma_w : \mathbb{R}^{e_d} \to \mathbb{R}^{e_d}$ and $\gamma_t : \mathbb{R}^{e_d} \to \mathbb{R}^{e_d}$ are MLPs.

This refinement enriches edge representations with geometric information extracted from steerable features, enhancing the model's ability to capture spatial relationships between binding and non-binding residues.

Channel Mixing via EQFF. This is employed after GADGTA for efficient channel wise interaction while preserving equivariance:

$$\text{EQFF}(h, \tilde{\mathbf{X}}^{(l)}) = \Big((h + m_1) || (\tilde{\mathbf{X}}^{(l)} + m_2 \circ \tilde{\mathbf{X}}^{(l)} \mathbf{W}_v) \Big) \tag{14}$$

where $(m_1, m_2) = \text{split}(\gamma_m((\|\tilde{\mathbf{X}}^{(l)} \mathbf{W}_v\|_2) || h))$ are modulation factors computed from feature norms. $\mathbf{W}_v$ is learnable weight matrices, $\|\cdot\|_2$ denotes L_2 norm and γ_m is and MLP.

3.5 Theoretical Properties of GDEGAN

GotenNet (Aykent & Xia, 2025) proves in Appendices A and B that the end-to-end architecture is $E(3)$-equivariant. By introducing ESM-embeddings (Lin et al., 2022) as node features, we establish the following:

Proposition 3.1 (From $E(3)$ to $SE(3)$ Equivariance with Invariant Node Features)**.** *We break the $E(3)$ equivariance by introducing ESM-embeddings because, now node features do not encode chirality information. Therefore, the network maintains $SE(3)$ equivariance but loses reflection equivariance.*

Proof in Appendix B.1.

Proposition 3.2 (Gaussian Attention Preserves $SE(3)$ Equivariance)**.** *The Gaussian Dynamic Attention mechanism, when applied to invariant scalar features from ESM-embeddings, preserves the $SE(3)$ equivariance of the message-passing framework.*

Proof in Appendix B.2.

Remark 1. The loss of reflection equivariance $E(3) \to SE(3)$ occurs at the input level through ESM embeddings, not through the attention mechanism itself.

Remark 2 (Parameter and Computational Efficiency). The Gaussian Dynamic Attention requires only $O(H)$ learnable parameters complexity, in contrast to the $O(d^2)$ for standard dot-product attention's key-query projection. Calculating neighborhood statistics incurs merely $O(|N(i)|d)$ operations per node, which is insignificant relative to the $O(|N(i)|d^2)$ for conventional dot-product attention. This results in $O(d)$-fold reduction in computational complexity with negligible parameter overhead and a significant decrease in inference time, as shown in Appendix Table 3.

Table 1: Experimental results of baseline models and our framework measured by DCC and DCA success rates[a]. The table presents comparative results across three benchmark datasets. Bold values indicate the best performance in each metric.

Methods	Param (M)	Failure rate ↓	COACH420		HOLO4K[k]		PDBbind2020	
			DCC↑	DCA↑	DCC↑	DCA↑	DCC↑	DCA↑
Fpocket[b]	\	**0.000**	0.228	0.444	0.192	0.457	0.253	0.371
P2rank[b]	\	**0.000**	0.366	0.628	0.314	0.621	0.503	0.677
DeepSite[b]	1.00	\	\	0.564	\	0.456	\	\
Kalasanty[b]	70.6	0.120	0.335	0.636	0.244	0.515	0.416	0.625
DeepSurf[b]	33.1	0.054	0.386	0.658	0.289	0.635	0.510	0.708
RecurPocket[b]	21.2	0.075	0.354	0.593	0.277	0.616	0.492	0.663
GAT[b]	**0.03**	0.110	0.039(0.005)	0.130(0.009)	0.036(0.003)	0.110(0.010)	0.018(0.001)	0.088(0.011)
GCN[b]	0.06	0.163	0.049(0.001)	0.139(0.010)	0.044(0.003)	0.174(0.003)	0.018(0.001)	0.070(0.002)
GCN2[b]	0.11	0.466	0.042(0.098)	0.131(0.017)	0.051(0.004)	0.163(0.008)	0.023(0.007)	0.089(0.013)
SchNet[b]	0.49	0.140	0.168(0.019)	0.444(0.020)	0.192(0.005)	0.501(0.004)	0.263(0.003)	0.457(0.004)
Egnn[b]	0.41	0.270	0.156(0.017)	0.361(0.020)	0.127(0.005)	0.406(0.004)	0.143(0.007)	0.302(0.006)
EquiPocket[b]	1.70	0.051	0.423(0.014)	0.656(0.007)	0.337(0.006)	0.662(0.007)	0.545(0.010)	0.721(0.004)
GotenNet	2.20	0.049	0.464(0.007)	0.624(0.014)	0.454(0.001)	0.691(0.005)	0.553(0.008)	0.705(0.007)
GDEGAN	1.90	0.032	**0.580(0.008)**	**0.707(0.009)**	**0.560(0.013)**	**0.788(0.011)**	**0.675(0.010)**	**0.826(0.011)**

[1] [a]The standard deviation is indicated in brackets. [b] Results from the EquiPocket (Zhang et al., 2024) paper. [k] holo4k contains multi chains and complex with multiple copies, presenting a strong distribution shift.

3.6 Training Objective

We formulate binding site prediction as a multi-task learning problem that jointly optimizes localization accuracy and geometric understanding of protein-ligand interactions.

Protein Ligand Binding Site Prediction. For biding site identification we use Dice Loss $\mathcal{L}_{\text{Dice}}$ following (Aggarwal et al., 2021; Zhang et al., 2024) to address inherent class imbalance, then compute $\hat{y}_i = \mathbf{Sigmoid}(MLP(h_i^{(L)})$ are predicted binding probabilities after L layers, $y_i \in \{0, 1\}$ are ground truth labels, and $\epsilon = 1$ prevents division by zero. The Dice coefficient naturally handles imbalanced classes by focusing on the overlap between predictions and ground truth rather than individual classification accuracy.

Auxiliary Directional Loss. To enhance geometric understanding, we extract directional information from the learned equivariant features and supervise it with ground truth ligand directions. The $l = 1$ steerable features $\tilde{\mathbf{X}}_i^{(1)} \in \mathbb{R}^{3\times h_d}$ inherently encode directional information. We extract predicted directions as: $\hat{\mathbf{d}}_i = \frac{\bar{\mathbf{X}}_{ichannel}^{(1)}}{\| \bar{\mathbf{X}}_{ichannel}^{(1)} \|_2 + \epsilon}$. Here, $\bar{\mathbf{X}}_{ichannel}^{(1)} = \frac{1}{h_d} \sum_{k=1}^{h_d} \tilde{\mathbf{X}}_{i,k}^{(1)} \in \mathbb{R}^3$ averages across feature channels to obtain a single direction vector. The ground truth direction $\mathbf{d}_i^{\text{true}}$ points from residue i to the nearest ligand heavy atom: $\mathbf{d}_i^{\text{true}} = \frac{\mathbf{p}_{\text{lig}}^* - \mathbf{p}_i}{\|\mathbf{p}_{\text{lig}}^* - \mathbf{p}_i\|_2}$, where $\mathbf{p}_{\text{lig}}^* = \arg\min_{\mathbf{p}\in\mathcal{L}} \|\mathbf{p} - \mathbf{p}_i\|_2$. Here, $\mathcal{L}$ denotes the set of ligand atom positions. We compute directional loss $\mathcal{L}_{\text{Dir}}$ using cosine similarity between true and predicted directions. Hence, the training objective of our GDEGAN becomes $\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{Dir}}$.

4 Experiments

4.1 Datasets and Baseline Methods Compared

We utilize the datasets benchmark settings of EquiPocket (Zhang et al., 2024) for LBS identification. **scPDB** (Desaphy et al., 2015) is the widely utilized dataset used for training and validation, which contains the proteins' 3D structures and ligands. **PDBbind2020** (Wang et al., 2004), **COACH420** and **HOLO4K** Krivák & Hoksza (2018) are three diverse datasets used to evaluate our method. The detailed discussion on datasets can be found in Appendix C.1. We compare GDEGAN with several categories of methods, as seen in Table 1, with further information provided in Appendix C.2.

Table 2: Ablation Study.

Methods	EQ[c]	ADL[d]	ESM[e]	COACH420		HOLO4K		PDBbind2020	
				DCC↑	DCA↑	DCC↑	DCA↑	DCC↑	DCA↑
GotenNet	$E(3)$	No	No	0.454(0.007)	0.624(0.014)	0.464(0.001)	0.691(0.005)	0.553(0.008)	0.705(0.007)
GotenNet+ADL	$E(3)$	Yes	No	0.485(0.004)	0.642(0.011)	0.468(0.004)	0.732(0.004)	0.592(0.010)	0.748(0.003)
GotenNet+ESM	$SE(3)$	No	Yes	0.543(0.008)	0.693(0.006)	0.520(0.011)	0.753(0.004)	0.637(0.005)	0.760(0.004)
GotenNet(full)	$SE(3)$	Yes	Yes	0.556(0.002)	0.703(0.005)	0.529(0.011)	0.749(0.006)	0.649(0.005)	0.801(0.014)
GDEGAN+ESM	$SE(3)$	No	Yes	0.572(0.001)	0.702(0.002)	0.532(0.010)	0.769(0.006)	0.652(0.010)	0.810(0.011)
GDEGAN(full)	$SE(3)$	Yes	Yes	**0.580(0.008)**	**0.707(0.009)**	**0.560(0.013)**	**0.788(0.011)**	**0.675(0.010)**	**0.826(0.011)**

[1] [c]Equivariance of the Model. [d] Auxiliary Directional Loss. [e] Node Features generated through ESM-2 (Lin et al., 2022). GotenNet computes dot-product attention while GDEGAN computes gaussian dynamic attention.

4.2 Evaluation Metrics

We used well established metrics **DCC**, **DCA** and **Failure rate** (Chen et al., 2011) for LBS identification, detailed in Appendix C.3. We assess localization accuracy through **DCC** (Distance from Center of Center), measuring the Euclidean distance between predicted and true binding site centers, **DCA** (Distance to Closest Atom), measuring the minimum distance from the predicted center to any ligand atom, and **Failure rate** as percentage of proteins without any predicted binding site. Predictions are considered successful when DCC or DCA falls below the standard 4Å threshold, which captures typical protein-ligand interaction distances. During inference on novel proteins where the number of binding pockets are unknown, we employ mean-shift clustering (Comaniciu & Meer, 2002) on high-scoring residues ($\hat{y}_i > \tau$) following (Krivák & Hoksza, 2018) to automatically identify multiple binding pockets.

4.3 Implementation Details

We used 4 layers of GDEGAN with hidden dimensions as 128 throughout the model, $L_{max} = 2$ for steerable features, attention heads as 8, and edge spatial cutoff r_c is set to 10Å. We trained our model using the AdamW Optimizer (Loshchilov & Hutter, 2019) for 100 epochs, selecting the best checkpoints based on the validation set. LayerNorm, TensorLayerNorm and Dropouts (Aykent & Xia, 2025) were applied in each layer with SiLU activation. The learning rate was initially set to 0.0005 with Cosine Scheduler and weight decay with value 0.05. We trained our model on NVIDIA H100 NVL GPU with a batch size of 16. All the hyperparameters were selected based on the validation dataset, which is the 10% of the training dataset. More details are provided in Appendix D.

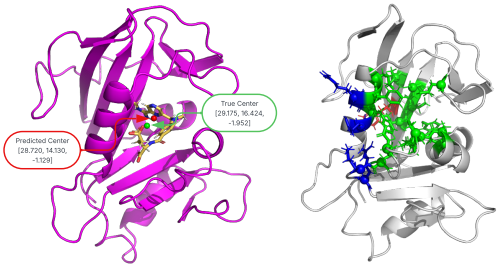

Figure 2: **Visualization of Protein 'PDB:1u72(A)'. Left:** Model prediction (red) vs true center (green) with coordinates. **Right:** Predicted residues: True Positive (green), False Positive (red), False Negative (blue).

4.4 Results and Attention Patters Analysis

The experimental results shown in Table 1 indicate substantial improvements across all benchmarks. GDEGAN attains the highest DCC success rate across all test datasets, with significant enhancements of 37.1%, 66.17%, and 23.8% over EquiPocket's DCC success rate for COACH420, HOLO4k, and PDBbind2020, respectively. For DCA metrics, we achieve enhancements of 7.7%, 19.0%, and 14.5% compared Equipocket's DCA. GDEGAN decreases the failure rate to 3.2%, in contrast to 5.1% for EquiPocket and 4.9% for GotenNet. Figure 2 illustrates the visualization of both predictions: the centroid and probable binding site candidates from our model. The comparison of inference speeds for our method is presented in Appendix C.4.

We visualize learned attention patterns for both Gaussian dynamic attention (GDEGAN) and dot product attention (GotenNet) as shown in Figure 3. The binding site attention sub-matrix on the left of GDEGAN shows higher values on binding site residues (darker red) and relevant neighbors

(dark yellow), suggesting strong local clustering, which demonstrates the adaptive nature of the method. This aligns with the claim that GDEGAN's Gaussian kernels impose localization by automatically nullifying distant interactions, concentrating the model's representational capability on statistically relevant neighborhoods. Meanwhile, the attention sub-matrix on the left of GotenNet shows a typical pattern of dot product attention capturing binding sites, but with less concentration on relevant neighbors. The attention values with reduced peak magnitudes in GDEGAN indicate optimal allocation of attention instead of indiscriminate distribution.

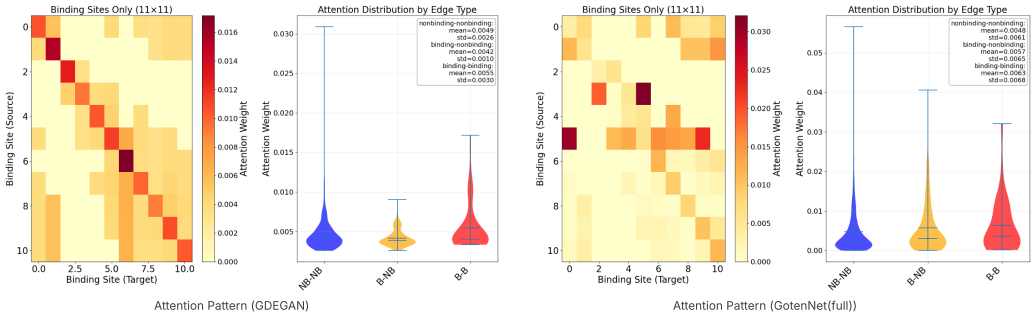

Figure 3: **Attention Patters Visualizations of Protein 'PDB:3c2f(A)'. Left:** On the left we show the attention patterns of GDEGAN, and on the **Right:** attention patterns of GotenNet(full).

4.5 Ablation Study

Table 2 presents the component-wise ablation results. Using ESM embeddings (GotenNet+ESM) instead of atomic numbers embeddings (GotenNet) improves the results by 15.61% DCC and 9.27% DCA averaged across datasets, indicating evolutionary information helps capture better binding regions than a purely geometric approach. Gaussian attention (GDEGAN+ESM) improves the results over dot-product attention (GotenNet+ESM) by 3.33% DCC and 3.32% DCA and both the techniques (full) with auxiliary directional loss as directional supervision further improves the results by an average of 2% DCC and 3.5% DCA respectively averaged across datasets. Critically, on the more structurally diverse HOLO4K dataset, GDEGAN(full) achieves a 2.4% DCC improvement over GotenNet(full), compared to only 1.5% on the more homogeneous COACH420 dataset, demonstrating that statistical adaptation provides greater benefits as structural heterogeneity increases. Notably, ESM features provide the largest enhancement, Gaussian attention's contribution grows with data complexity, validating our hypothesis that adaptive statistical kernels better handle protein heterogeneity than uniform similarity measures.

5 Conclusion and Limitations

This study presents GDEGAN, an enhancement of GotenNet (Aykent & Xia, 2025), which substitutes dot-product attention with Gaussian Dynamic Attention, specifically developed for the detection of protein-ligand interaction sites. This simple yet efficacious change demonstrates enhancements of 42.36% and 13.73% in DCC and DCA, respectively, averaged across all datasets in comparison to the prevailing state-of-the-art approach, EquiPocket (Zhang et al., 2024). Unlike previous methods that utilize atom-level information (Jiménez et al., 2017; Zhang et al., 2024; Aggarwal et al., 2021), our approach depends on residue-level information, hence improving efficiency in both training and inference due to a reduction in input graph size.

Our method is proposed to find the binding center by predicting probable binding site candidates instead of performing docking, a logical progression is to utilize our predictions to limit the docking search space. Even though the training data is limited in size and data points have well-defined pockets with a single ligand, our approach can generalize better. Our method would benefit more from training on more diverse data points with multi-ligand interacting pockets. Given that GDEGAN predicts binding sites Future research should methodically assess whether our statistical attention scores indicate ligand binding locations and predict binding affinity based on local feature coherence.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Rishal Aggarwal, Akash Gupta, Vineeth Chelur, CV Jawahar, and U Deva Priyakumar. Deeppocket: ligand binding site detection and segmentation using 3d convolutional neural networks. *Journal of Chemical Information and Modeling*, 62(21):5069–5079, 2021.

Sarp Aykent and Tian Xia. Gotennet: Rethinking efficient 3d equivariant graph neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35:11423–11436, 2022.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Jiacheng Cen, Anyi Li, Ning Lin, Yuxiang Ren, Zihe Wang, and Wenbing Huang. Are high-degree representations really unnecessary in equivariant graph neural networks? *Advances in Neural Information Processing Systems*, 37:26238–26266, 2024.

Ke Chen, Marcin J Mizianty, Jianzhao Gao, and Lukasz Kurgan. A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure*, 19 (5):613–621, 2011.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International conference on machine learning*, pp. 1725–1735. PMLR, 2020.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

Jérémy Desaphy, Guillaume Bret, Didier Rognan, and Esther Kellenberger. sc-pdb: a 3d-database of ligandable binding sites—10 years on. *Nucleic acids research*, 43(D1):D399–D404, 2015.

Xing Du, Yi Li, Yuan-Ling Xia, Shi-Meng Ai, Jing Liang, Peng Sang, Xing-Lai Ji, and Shu-Qun Liu. Insights into protein-ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, 17(2):144, 2016.

Yuanqi Du, Limei Wang, Dieqiao Feng, Guifeng Wang, Shuiwang Ji, Carla P Gomes, Zhi-Ming Ma, et al. A new perspective on building efficient and expressive 3d equivariant graph neural networks. *Advances in neural information processing systems*, 36:66647–66674, 2023.

Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, Davide Boscaini, Michael M Bronstein, and Bruno E Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature methods*, 17(2):184–192, 2020.

Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi S. Jaakkola, and Andreas Krause. Independent SE(3)-equivariant models for end-to-end rigid protein docking. In *International Conference on Learning Representations*, 2022.

Hiba Abi Hussein, Alexandre Borrel, Colette Geneix, Michel Petitjean, Leslie Regad, and Anne-Claude Camproux. Pockdrug-server: a new web server for predicting pocket druggability on holo and apo proteins. *Nucleic acids research*, 43(W1):W436–W442, 2015.

Georgios Ioannides, Aman Chadha, and Aaron Elkins. Gaussian adaptive attention is all you need: Robust contextual representations across multiple modalities, 2024.

Kandel Jeevan, Shrestha Palistha, Hilal Tayara, and Kil T Chong. Puresnetv2. 0: a deep learning model leveraging sparse representation for improved ligand binding site prediction. *Journal of Cheminformatics*, 16(1):66, 2024.

José Jiménez, Stefan Doerr, Gerard Martínez-Rosell, Alexander S Rose, and Gianni De Fabritiis. Deepsite: protein-binding site predictor using 3d-convolutional neural networks. *Bioinformatics*, 33(19):3036–3042, 2017.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10(1):39, 2018.

Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10(1):168, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Peiying Li, Boheng Cao, Shikui Tu, and Lei Xu. Recurpocket: recurrent lmser network with gating mechanism for protein binding site detection. In *2022 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 334–339. IEEE, 2022.

Jie Liang, Clare Woodward, and Herbert Edelsbrunner. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein science*, 7(9): 1884–1897, 1998.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2023.

Yi-Lun Liao, Brandon M Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems*, 35:7236–7249, 2022.

Oscar Méndez-Lucio, Mazen Ahmad, Ehecatl Antonio del Rio-Chanona, and Jörg Kurt Wegner. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3(12):1033–1039, 2021.

Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023.

Stelios K Mylonas, Apostolos Axenopoulos, and Petros Daras. Deepsurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics*, 37(12): 1681–1690, 2021.

Zhuoran Qiao, Anders S Christensen, Matthew Welborn, Frederick R Manby, Anima Anandkumar, and Thomas F Miller III. Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proceedings of the National Academy of Sciences*, 119(31): e2205221119, 2022.

Anastasiia V Sadybekov and Vsevolod Katritch. Computational approaches streamlining drug discovery. *Nature*, 616(7958):673–685, 2023.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Karen T Schomburg, Stefan Bietz, Hans Briem, Angela M Henzler, Sascha Urbaczek, and Matthias Rarey. Facing the challenges of structure-based target prediction by inverse virtual screening. *Journal of chemical information and modeling*, 54(6):1676–1686, 2014.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International conference on machine learning*, pp. 9377–9388. PMLR, 2021.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet-a deep learning architecture for molecules and materials. *The Journal of chemical physics*, 148(24), 2018.

Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, pp. 20503–20521. PMLR, 2022.

Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Improving detection of protein-ligand binding sites with 3d segmentation. *Scientific reports*, 10(1):5035, 2020.

Paolo Tosco, Nikolaus Stiefl, and Gregory Landrum. Bringing the mmff force field to the rdkit: implementation and validation. *Journal of cheminformatics*, 6(1):37, 2014.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

Yang Zhang, Zhewei Wei, Ye Yuan, Chongxuan Li, and Wenbing Huang. Equipocket: an e(3)-equivariant geometric graph neural network for ligand binding site prediction. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yangtian Zhang, Huiyu Cai, Chence Shi, and Jian Tang. E3bind: An end-to-end equivariant network for protein-ligand docking. In *The Eleventh International Conference on Learning Representations*, 2023.

Xiliang Zheng, LinFeng Gan, Erkang Wang, and Jin Wang. Pocket-based drug design: exploring pocket space. *The AAPS journal*, 15(1):228–241, 2013.

A Geometric Graph Formulation and Binding Site Representations

A.1 Representation of Proteins

The 3D structure of a protein is defined by the spatial coordinates of atoms associated with every amino acid, organized according to its amino acid sequence. For computational efficiency and biological relevance in detecting probable binding site candidates, we represent each amino acid residue by its C_α atom position $\mathbf{p}_i \in \mathbb{R}^3$, which provides a stable backbone reference point consistent across

all amino acid types. We formalize a protein structure as a geometric graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that captures both topological and spatial information necessary for binding site identification. Nodes in the geometric graph $\mathcal{V} = \{(\mathbf{p}_i, h_i)\}_{i=1}^N$ represent residue-feature pairs. The edges are established from each residue i to all residues j within a cutoff radius: $\mathcal{E} = \{(i,j) : \|\mathbf{p}_i - \mathbf{p}_j\| < r_c, i \neq j\}$, where we set $r_c = 10$ Å based on typical interaction distances in protein binding sites following (Zhang et al., 2024). Rather than encoding raw physicochemical properties, we leverage learned representations that capture both evolutionary conservation and structural context. Specifically, for each residue i, we extract a feature vector $h_i \in \mathbb{R}^d$ from the pretrained ESM-2 (Lin et al., 2022) model, which encodes evolutionary information and sequence context.

A.2 Binding Site Definition and Representation.

Protein binding sites are the critical regions where ligands, or other molecules interact with proteins in the biological process. We define binding sites based on proximity to known ligand positions from crystal structures, surrounded by the atoms of the protein. A residue is classified as belonging to the binding site when any of its constituent atoms lies within a threshold distance d_{bind} from any ligand atom. Formally, a residue i is considered part of a binding site as:

$$y_i = \begin{cases} 1 & \text{if } \min_{a \in \mathcal{A}_i, b \in \mathcal{L}} \|\mathbf{p}_a - \mathbf{p}_b\| < d_{bind} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where $\mathcal{A}_i$ represents all atoms of residue i, $\mathcal{L}$ represents all ligand atoms. d_{bind} represents the binding distance threshold taken as 4Å following (Zhang et al., 2024; Krivák & Hoksza, 2018; Jiménez et al., 2017).

B Proofs

B.1 Proof of Proposition 3.1

Proof. Considering the GDEGAN architecture with ESM embeddings as node features. We show that the network maintains SE(3) but not E(3) equivariance.

ESM embeddings are invariant scalars. ESM embeddings $h_i \in \mathbb{R}^{n_d}$ encode amino acid sequence information and evolutionary patterns. These are scalar features that do not transform under any spatial transformation. For any transformation $g = (\mathbf{R}, \mathbf{t})$ where $\mathbf{R} \in SO(3)$ (rotation) and $\mathbf{t} \in \mathbb{R}^3$ (translation):

$$h_i' = h_i \quad \forall g \in SE(3) \tag{16}$$

Loss of reflection equivariance. Consider a chiral molecule $\mathcal{M}$ and its mirror image $\mathcal{M}'$ obtained by reflection P. Both have identical ESM embeddings: $h_i = h_i'$ (same amino acid sequence) and identical pairwise distances: $\|\mathbf{x}_i - \mathbf{x}_j\| = \|\mathbf{x}_i' - \mathbf{x}_j'\|$, but different chirality. Since the network cannot distinguish between $\mathcal{M}$ and $\mathcal{M}'$ based on invariant features alone, it cannot be equivariant under reflections.

□

B.2 Proof of Gaussian Dynamic Attention Equivariance 3.2

Here we prove that our Gaussian Dynamic Attention mechanism maintains SE(3) equivariance:

B.2.1 Gaussian Attention Mechanism

Our Gaussian attention computes attention weights as:

$$\alpha_{ij} = \frac{\exp\left(-\frac{||h_j - h_i||^2}{2\xi \cdot (\sigma_i^2 + \epsilon)}\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(-\frac{||h_k - h_i||^2}{2\xi \cdot (\sigma_i^2 + \epsilon)}\right)} \tag{17}$$

where $h_i, h_j \in \mathbb{R}^{n_d}$ are invariant scalar features from ESM embeddings, σ_i^2 is the neighborhood variance, and ξ is the learnable temperature.

B.2.2 PROOF OF SE(3) EQUIVARIANCE

Proof. **Invariance of scalar features.** The ESM embeddings h_i are invariant under SE(3) transformations, because they encode sequential and evolutionary information, not geometric co-ordinates:

$$g(h_i) = h_i \tag{18}$$

Invariance of attention weights. Since h_i and h_j are invariant:

$$\|g(h_j) - g(h_i)\|^2 = \|h_j - h_i\|^2 \tag{19}$$

The neighborhood variance σ_i^2 computed from invariant features is also invariant:

$$\sigma_i^2 = \frac{1}{|\mathcal{N}(i)|} \sum_{k \in \mathcal{N}(i)} \|h_k - \mu_i\|^2 \tag{20}$$

where μ_i is the neighborhood mean. Under $g \in SE(3)$, both remain unchanged.

Therefore:

$$\alpha_{ij}^g = \exp\left(-\frac{||h_j - h_i||^2}{2\xi \cdot (\sigma_i^2 + \epsilon)}\right) = \alpha_{ij} \tag{21}$$

Hence, Gaussian Dynamic Attention mechanism preserves SE(3) equivariance by maintaining invariant attention weights while allowing equivariant features to transform appropriately under rotations. The scalar nature of ESM embeddings ensures that reflection equivariance is not required, reducing $E(3)$ to $SE(3)$ as stated in Proposition 3.1. □

C EXPERIMENTS

C.1 DATASETS

scPDB (Desaphy et al., 2015) comprises 17,594 protein-ligand complex structures from the 2017 release, representing 4,782 unique proteins and 6,326 ligands. We used the most frequently used dataset for LBS identification (Stepniewska-Dziubinska et al., 2020; Jeevan et al., 2024) for training and validation, with a split of $90:10$. Final dataset was preprocessed using the steps described in EquiPocket (Zhang et al., 2024). **PDBbind2020** Wang et al. (2004) contains experimentally determined binding affinity data paired with structural information. We utilize the refined subset as per (Zhang et al., 2024), consisting of 5,316 complexes selected for structural quality from the larger general set of 14,127 complexes. The refined set enforces strict quality criteria, including resolution better than 2.5Å and complete ligand electron density. **COACH420 and HOLO4K** serve as independent test sets following Krivák & Hoksza (2018). COACH420 contains 420 protein-ligand complexes with diverse binding site architectures, while HOLO4K comprises 4,288 structures. Both datasets use the $MLIG$ subsets following (Aggarwal et al., 2021; Jiménez et al., 2017) containing biologically relevant ligands as defined by the original curation. Notably, HOLO4K presents significant distribution shift challenges as it contains numerous multi-chain assemblies and oligomeric proteins absent from typical training sets. We have split the HOLO4K dataset into per-chain components and aggregated the predictions in our evaluated results. For all datasets, we exclude solvent molecules and apply standard preprocessing, such as removing hydrogen atoms. Structures with missing coordinates or ambiguous ligand positions are filtered during preprocessing using rDkit (Tosco et al., 2014).

C.2 BASELINE METHODS COMPARED

We compare GDEGAN with different categories of methods proposed for LBS identification. *Traditional Machine Learning-based:* Fpocket (Le Guilloux et al., 2009), and P2rank (Krivák & Hoksza, 2018). *CNN-based:* DeepSite (Aggarwal et al., 2021), Kalasanty (Stepniewska-Dziubinska et al., 2020), and RecurPocket (Li et al., 2022). *Topological Graph-based:* GAT Veličković et al. (2018), GCN Kipf & Welling (2017), and GCN2 (Chen et al., 2020). *Spatial Graph-based:* SchNet (Schütt et al., 2018), EGNN (Satorras et al., 2021), and Equipocket (Zhang et al., 2024). *High-degree steerable method:* GotenNet (Aykent & Xia, 2025).

C.3 Evaluation Metrics

DCC (Distance from Center to Center). For each predicted binding site center $\hat{\mathbf{p}}_i$ and true binding site center $\mathbf{p}_j$, DCC measures the Euclidean distance between centers:

$$\text{DCC} = \|\hat{\mathbf{p}}_i - \mathbf{p}_{ligand}\|_2 \tag{22}$$

where $\hat{\mathbf{p}}_i \in \mathbb{R}^3$ represents the i-th predicted center and $\mathbf{p}_j \in \mathbb{R}^3$ the j-th ground truth center.

DCA(Distance to Closest Atom). This metric evaluates whether predictions are within the actual binding region by measuring the minimum Euclidean distance from a predicted center to any ligand atom:

$$\text{DCA} = \min_{b \in \mathcal{L}} \|\hat{\mathbf{p}}_i - \mathbf{p}_b\|_2 \tag{23}$$

where $\mathcal{L}$ represents all ligand atoms.

For both metrics predictions are considered successful if they are within a standard threshold τ , which in this case we have taken as 4Å following Aggarwal et al. (2021); Le Guilloux et al. (2009); Mylonas et al. (2021); Zhang et al. (2024).

$$\text{Success Rate}_{\text{DCC/DCA}} = \frac{|\{\text{Predicted sites} \mid \text{DCC/DCA} < \tau\}|}{|\{\text{True sites}\}|} \tag{24}$$

$$\text{Failure Rate} = \frac{|\{\text{Proteins} \mid |\text{predicted centers}| = 0\}|}{|\{\text{Proteins}\}|} \tag{25}$$

where $|\cdot|$ denotes set cardinality and $\tau = 4$Å is the standard threshold for successful prediction.

We have used **DCC/DCA success rate** and **Failure rate** as the evaluation metrics to compare from the sate-of-the-art methods.

C.4 Computational Efficiency Analysis

We evaluate the inference speed on 100 proteins for each model and present the duration in seconds (s) in Table 3. GDEGAN demonstrates substantial enhancements, with inference time about 1.9 seconds, in contrast to GotenNet's 4.12 seconds and Equipocket's 37 seconds.

Table 3: Inference time comparison across methods.

Method	Time (s/100 proteins)	Speedup	Type
GDEGAN (Ours)	**1.90**	**1.00×**	Reside Level Nodes
GotenNet	4.12	0.46×	Reside Level Nodes
EquiPocket^{p}	37.00	0.05×	Atom Level Nodes
Fpocket^{p}	23.00	0.08×	Geometry Based
Kalasanty^{p}	86.00	0.02×	3D-CNN Based
DeepSurf^{p}	641.00	0.003×	3D-CNN Based

^{p} Results from the EquiPocket (Zhang et al., 2024) paper.

D Training Hyper-Parameters Selection

This section presents the hyperparameters for training as outlined in Table 4, selected based on the validation data, which comprises 10% of the training dataset. These hyperparameters can be employed to ensure reproducibility. **We will release the full code based on the acceptance of the work.**

E Declaration on the use of Large Language Models

In this work, we have utilised tools like **Grammarly** to check any grammatical oversight, and these tools are powered by LLMs.

Table 4: Hyperparameter selection and reproducibility details.

Hyperparameter	**Search Space**
Learning Rate	{0.003, 0.0003, **0.0005**}
Minimum Learning Rate	**1e-6**
Batch Size	{8, **16**, 32}
Optimizer	{Adam, **AdamW**}
Learning rate scheduler	**Cosine Annealing Warm Restarts**
Warmup Epochs	**10**
Maximum Epochs	**150**
Early Stopping Patience	**30**
Gradient clipping	{10, **15** }
Weight Decay	{0.01, **0.05**}
Dropout Rate	{0.1, 0.2, **0.5**}
Node hidden dimension	**128**
Edge dimension (e_d)	**128**
Edge refinement dimension	**128**
L_{max}	**2**
Number of Layers	{2, **4**, 6}
Number of RBFs	**32**
Maximum number of neighbors	**32**
Number of attention heads	{4, **8**}
Activation Function	{ReLU, **SiLU**}
τ	**0.5**