Explainability of Molecular Graph Neural Network

Ataollah Kamal
INSA Lyon, LIRIS UMR 5205
F-69621 Villeurbanne, France

Matej Hladiš
Univ. Côte d'Azur, ICN, UMR 7272 CNRS
F-06108 Nice, France

Jérémie Topin
Univ. Côte d'Azur, ICN, UMR 7272 CNRS
F-06108 Nice, France

Marc Plantevit
EPITA Research Laboratory (LRE)
F-94276 Le Kremlin-Bicêtre, France

Sébastien Fiorucci
Univ. Côte d'Azur, ICN, UMR 7272 CNRS
F-06108 Nice, France

Céline Robardet
INSA Lyon, LIRIS UMR 5205
F-69621 Villeurbanne, France
celine.robardet@insa-lyon.fr

***Abstract*—Graph Neural Networks (GNNs) have demonstrated strong performance in molecular interaction prediction, but their interpretability remains limited, especially in domain-specific applications like ligand–receptor modeling. This paper presents a model-agnostic explainer for GNN-CLS, a specialized GNN model designed to predict interactions between molecules and olfactory receptor proteins. The proposed method uses cooperative game theory to identify influential molecular substructures and receptor sequence regions, offering faithful and theoretically grounded explanations of model predictions. This approach enhances transparency by revealing which features drive predictive outcomes, helping bridge the gap between model performance and chemical insight. The contributions include a formal framework for relevance attribution and interaction analysis, positioning this work at the intersection of explainable AI and computational chemistry.**

I. Introduction

As machine learning models grow in complexity, they are achieving unprecedented levels of performance across a wide range of domains. However, this increase in capability often comes at the cost of interpretability. The internal decision-making processes of these models become harder to understand, making it challenging for researchers and practitioners to trust their outputs or gain meaningful insights from their behavior. As a response, the field of Explainable Artificial Intelligence (XAI) has emerged, aiming to bridge this gap by providing tools to make model predictions more understandable and interpretable to humans.

The need for explainability is particularly critical in the domain of molecular interaction prediction, where understanding the rationale behind a model's output can generate valuable biological hypotheses and guide experimental design. In this context, domain-specific architectures have been developed to more accurately capture the complexity of biochemical systems, extending beyond general-purpose Graph Neural Networks (GNNs). However, the increased specificity and architectural sophistication of these models often obscure their inner workings, reinforcing the need for dedicated explainability methods capable of elucidating how predictions are made. One such model is GNN-CLS [1], a specialized GNN architecture designed to predict interactions between molecules and olfactory receptor (OR) proteins. GNN-CLS integrates molecular graph representations with protein sequence embeddings to model ligand–receptor interactions within a unified framework. While the model achieves strong predictive performance, its intricate design makes it challenging to trace which molecular or receptor features are most influential in the prediction.

To address this, we develop a model-agnostic explainer tailored to GNN-CLS model. Our method identifies the most important parts of the molecule and the receptor sequence for each prediction, as well as the key interactions between them. Importantly, the explainer is grounded in cooperative game theory, providing theoretical guarantees for the faithfulness and consistency of the generated explanations. By combining domain-aware relevance analysis with formal interpretability principles, our approach improves transparency and brings GNN-CLS closer to practical, insight-driven applications in molecular interaction modeling.

This work offers two primary contributions. First, we introduce a model-agnostic Graph Neural Network (GNN) explainer grounded in cooperative game theory, which provides theoretical guarantees regarding the faithfulness of the generated explanations. This framework allows for reliable interpretation across a wide range of GNN architectures. Second, we present a novel methodology to capture and explain the interactions between molecular substructures and specific regions of the receptor, offering deeper insights into structure–activity relationships that are critical for tasks such as drug discovery and molecular design.

The remainder of this paper is organized as follows. In Section II, we review related work on Explainable Artificial Intelligence (XAI), positioning our contribution within the broader research landscape. Section III introduces the GNN-CLS model and outlines the key principles of cooperative game theory that underpin our approach. In Section IV, we formally present our model-agnostic explainer, ESPAM, developed specifically for molecular machine learning model explanation, and detail its game-theoretic foundation. Section V describes the experimental setup and reports both quantitative and qualitative results. Finally, Section VI concludes the paper

This work was supported by the CNRS (MITI interdisciplinary program) and by the French government through the France 2030 investment plan managed by the ANR (grants PORTRAIT ANR-22-CE23-0006 and PANDORA ANR-24-CE23-0950), and as part of the Initiative of Excellence Université Côte d'Azur (ANR-15-IDEX-01). We also thank the Université Côte d'Azur's OPAL HPC center for resources and support.

by summarizing our main contributions and findings, and by outlining potential directions for future research.

II. Related Works

In recent years, self-explainable GNNs [2], [3] have emerged as a promising step toward interpretability. Yet, such models are not always practical to deploy, as many GNNs are already implemented without intrinsic explainability. As a result, most GNNs are still black boxes. Therefore, despite progress on inherently interpretable designs, post-hoc explanation methods remain essential for understanding the behavior of these widely used opaque models.

A variety of methods have been proposed to explain GNN predictions by identifying influential substructures within input graphs. Two prominent examples are GNNExplainer [4] and PGExplainer [5], which make clear a model's decision by learning edge masks over the input graph using perturbation-based techniques. In contrast, PGM-Explainer [5] constructs a probabilistic graphical model to approximate the causal relationships between subgraph components and predictions. Also, adaptations of Grad-CAM have been applied to GNNs to attribute importance using gradient-based saliency maps [6].

Unlike the aforementioned methods, which focus on instance-level explanations, GLGExplainer [7] offers model-level insights through logical propositions of concepts, leveraging subgraphs obtained from an existing instance-level explainer. Nevertheless, the quality of these model-level explanations remains highly dependent on the reliability of the underlying instance-level explanations [8]. Moreover, previous studies have shown that concept-based models either exhibit limited effectiveness when relying solely on concept truth-values or sacrifice interpretability when using concept embeddings, as the embedding dimensions lack clear semantic meaning and thus hinder human-understandable reasoning [9].

In general, while these methods provide valuable insights, they are primarily heuristic and lack formal theoretical guarantees regarding the faithfulness and consistency of their explanations. Furthermore, most of these approaches are not model-agnostic, but are tightly coupled to specific GNN architectures or require access to internal gradients or parameters. This limitation reduces their applicability, making them difficult or even infeasible to use with more complex or black-box models.

To address the need for theoretical guarantees concerning the faithfulness and consistency of explanations, game-theoretic approaches have emerged, offering a principled framework for attributing importance. SubgraphX [10] employs Monte Carlo tree search to approximate Shapley values [11], capturing the effects of subgraph interactions. GraphSVX [12] uses surrogate modeling on perturbed inputs to estimate Shapley values and provide fair attributions across features and nodes. While these methods rely on Shapley values, GStarX [13] instead leverages the Hamiache–Navarro (HN) value [14], which explicitly incorporates graph structure when forming coalitions, enhancing the structural coherence of the explanations. However, computing exact Shapley or Hamiache–Navarro values is computationally intractable in practice. Consequently, these methods resort to sampling or heuristic approximations [15], [16], which introduce a trade-off between explanation fidelity and computational efficiency. This compromise limits their scalability and applicability in real-time or large-scale settings.

III. GNN-CLS model and cooperative game

The GNN-CLS model extends the general principles of GNNs through a domain-specific architecture designed to capture the interactions between molecules and olfactory receptor (OR) proteins. To support our analysis of this model's predictions, we propose an explainability approach grounded in cooperative game theory. In what follows, we first describe the structure and functioning of GNN-CLS, then present the theoretical foundations of cooperative game-based explanation methods, which provide a principled way to attribute contributions to individual molecular and receptor features.

A. GNN-CLS model

Graph classification aims to assign labels to entire graph structures based on their topology and node attributes. This task is widely used in bioinformatics [17], cheminformatics [18], and social network analysis [19], where entities and their relationships are naturally represented as graphs. GNNs have become a standard approach for this problem, leveraging message-passing mechanisms to iteratively update node representations based on their local neighborhoods [20]. The final graph-level representation is typically obtained by aggregating node embeddings using permutation-invariant functions such as summation, mean pooling, or attention mechanisms.

The GNN-CLS model [1] is specifically designed to predict molecular interactions with olfactory receptors by jointly processing two distinct inputs: a molecular graph, where nodes correspond to atoms and edges to chemical bonds, and a receptor protein sequence, encoded using a pre-trained protein language model [21]. The model extends standard GNN-based graph classification by integrating multi-head attention mechanisms and adopting a proteo-chemometric modeling framework, enabling it to effectively capture the complex interplay between molecular structure and receptor sequence.

To effectively capture receptor-ligand interactions, GNN-CLS injects receptor sequence embeddings into the molecular graph as additional node features, allowing the message-passing process to account for both molecular topology and receptor-specific properties. The model alternates between graph-based message passing [20], fully connected layers, and attention mechanisms to facilitate information exchange beyond bonded interactions. Additionally, an edge-conditioned convolution (ECC) layer [22] and an attention-based readout function [23] generate the final graph-level representation used for classification. Figure 1 provides an overview of GNN-CLS, illustrating its architecture and how it integrates these components.

By combining molecular structure and sequence-based representations, GNN-CLS effectively models complex biochemical interactions. However, its interpretability remains a chal-

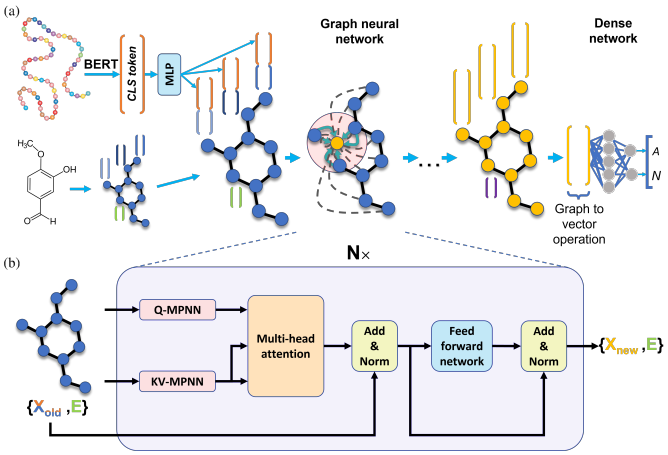

Fig. 1: Figure from [1]. The model takes as input a pair consisting of a protein sequence and a molecular graph. The protein sequence is first encoded using ProtBERT, and the resulting embedding is concatenated to the feature vector of each node in the molecular graph. Within the GNN architecture, node embeddings are initially updated through two successive and identical message passing neural networks. Subsequently, pairwise interactions between nodes are modeled using a multi-head attention mechanism.

lenge. Since GNNs inherently rely on multi-step message passing, identifying the most relevant molecular substructures or receptor-specific features is nontrivial. In this work, we develop explainability methods specifically designed for GNN-CLS, shedding light on its decision-making process and improving trust in its predictions.

B. Cooperative Games

To interpret the predictions of complex models such as GNN-CLS, we turn to cooperative game theory, which offers a rigorous framework for attributing the contribution of individual features to a model's output through concepts such as Shapley values.

A cooperative game is defined as a pair (N, g), where N is a finite set of players, and $g : 2^N \to \mathbb{R}$ is the *characteristic function*, satisfying $g(\emptyset) = 0$. A *coalition* is any subset $S \subseteq N$, while N itself is referred to as the *grand coalition*. The characteristic function assigns a real value to each coalition, representing its collective worth. A solution in cooperative game theory is a function that, for each game (N, g), allocates the total worth of the grand coalition, $g(N)$, among the players. Formally, a solution is a function defined over the set of all cooperative games, $\mathcal{G}$, such that:

$$\begin{aligned} &\phi : \mathcal{G} \to \mathbb{R}^{|N|}, \\ &(N, g) \mapsto \phi(N, g) = (\phi_1(g), \ldots, \phi_{|N|}(g)), \end{aligned} \tag{1}$$

where, for each player $i \in N$, the value $\phi_i(g)$ represents their allocated contribution to the total worth of the game.

Various solutions can be introduced for cooperative games based on the axioms imposed on ϕ. Among these, the Shapley values are a foundational solution concept that satisfy a specific set of desirable axioms: Efficiency, Symmetry, Null Player, and Additivity. These axioms ensure that contributions are distributed fairly among all players based on their marginal impact across all possible coalitions. In the context of explainable AI, Shapley values have been widely adopted to attribute the prediction of a model to its input features in a principled and interpretable way.

Building on this framework, [24] propose an alternative attribution method, Fair-Efficient-Symmetric-Perturbation (FESP) values, grounded in a modified set of axioms: Efficiency, Symmetry, and Fair Treatment. These axioms aim to preserve fairness in feature attribution while allowing for more flexibility in perturbation-based explanations. The definitions of these axioms are presented below.

- Efficiency: The total value of the grand coalition is fully distributed among all players: $\sum_{i \in N} \phi_i(g) = g(N)$.
- Symmetry: For any permutation π on the players, the following equations holds: $\phi_i(g) = \phi_{\pi(i)}(\pi \cdot g)$
- Fair Treatment: If adding a player i to every coalition adds more contribution than adding a player j to the same coalition, then the contribution value of i should be higher than j:

$$\begin{aligned} \forall i, j \in N, S \subseteq N \setminus \{i, j\}; g(\{i\} \cup S) \geq g(\{j\} \cup S) \\ \Rightarrow \phi_i(g) \geq \phi_j(g). \end{aligned}$$

The calculation of FESP values is defined in Equation (2):

$$\begin{aligned} \phi_i(g) &= w \times g(\{i\}) - (1 - w) \times g(N \setminus \{i\}) \\ \text{with } w &= \frac{g(N) + \sum_{j \in N} g(N \setminus \{j\})}{\sum_{j \in N} g(\{j\}) + \sum_{j \in N} g(N \setminus \{j\})} \end{aligned} \tag{2}$$

Shapley values are one of the most widely used attribution methods in cooperative game theory, offering a principled way to distribute the overall output of a model among its input features. However, their exact computation is known to be computationally expensive, requiring exponential time in the number of features [11]. To address this, several approximation methods have been proposed [15], [16], but these approaches often suffer from slow convergence and may yield only coarse estimates of feature contributions.

In contrast, the Fair-Efficient-Symmetric-Perturbation (FESP) attribution method [24] provides an exact solution with a much lower computational cost. FESP achieves linear-time computation, specifically $O(|N|)$. This makes it a practical and efficient alternative to Shapley values, particularly for complex machine learning models.

IV. Explaining GNN-CLS Model

We propose a methodology to explain the predictions of the GNN-CLS model by identifying both the key molecular substructures and the relevant regions of the receptor that contribute to the model's output. Our approach combines two complementary attribution strategies. In Section IV-A, we introduce ESPAM (Efficient Symmetric Perturbation Attribution Method), a FESP-inspired technique designed to highlight the most influential parts of the molecular graph. In Section IV-B, we present a corresponding method for analyzing the receptor

sequence, enabling a joint interpretation of the ligand-receptor interaction captured by GNN-CLS.

A. ESPAM for explaining the role of the molecular graph in predictions

A distinctive feature of Graph Neural Networks (GNNs) is that node embeddings evolve through iterative interactions with their neighbors. At each layer, a node aggregates information from its local neighborhood, gradually expanding its receptive field. This process naturally leads to the notion of an ego network, that is the subgraph that directly influences a node's representation at a given layer.

Formally, the ego network of a node v at layer ℓ, denoted $Ego_\ell(v) = (\mathcal{V}_\ell(v), \mathcal{E}_\ell(v))$, is recursively defined as:

$$Ego_\ell(v) = \begin{cases} (\{v\}, \emptyset) \text{if } \ell = 0, \\ (\mathcal{V}_{\ell-1}(v) \cup \{w \mid (u,w) \in E, u \in \mathcal{V}_{\ell-1}(v)\}, \\ \mathcal{E}_{\ell-1}(v) \cup \{(u,w) \mid u \in \mathcal{V}_{\ell-1}(v), (u,w) \in E\}) \\ \text{if } \ell > 0. \end{cases} \tag{3}$$

This recursive formulation captures how information propagates in the graph through local interactions, ultimately encoding both neighborhood structure and broader topological context.

In the framework of cooperative game theory, consider a game (N, g) where the set of players N corresponds to the nodes of a molecular graph G, and each coalition corresponds to an induced subgraph. The Fair Treatment axiom in this setting posits that if, for any coalition S disjoint from both v and u, we have $g(S \cup \{v\}) \geq g(S \cup \{u\})$, then it should follow that $\phi_v(g) \geq \phi_u(g)$. However, this classical formulation fails to consider the graph's structure: if v is adjacent to S while u is not, their contributions are not being fairly compared. In practice, such asymmetries can distort attribution, especially in GNNs where node relevance depends strongly on graph connectivity.

One potential remedy is to restrict comparisons to subgraphs that are connected to both u and v. However, this approach suffers when u and v are far apart in the graph, as the number of such subgraphs becomes vanishingly small, introducing a new source of bias.

To overcome this limitation, we propose redefining the Fair Treatment axiom in terms of ego networks. Since ego networks directly reflect how GNNs propagate information and build node representations, they offer a more appropriate basis for comparing node importance. Specifically, we introduce a GNN-aware Fair Treatment axiom: given two nodes v and u, if the following conditions hold for all layers ℓ, then $\phi_v(g) \geq \phi_u(g)$:

$$\forall \ell : \begin{cases} g(Ego_\ell(v)) \geq g(Ego_\ell(u)), \\ g(N \setminus Ego_\ell(v)) \leq g(N \setminus Ego_\ell(u)). \end{cases} \tag{4}$$

This formulation leverages the structure of ego networks to assess the relative influence of nodes in a way that is consistent with how GNNs process input data, offering a principled and efficient foundation for attribution in molecular graphs.

Indeed, Equation (4) assesses the relative importance of vertices v and u by considering two factors: (1) the influence each node exerts on the GNN's prediction when its corresponding ego-network is removed, and (2) the similarity between the GNN's prediction on ego-network-based subgraphs and the original full-graph prediction. This dual criterion ensures a robust comparison grounded in both structural and functional perspectives.

Building on this principle, we introduce ESPAM, a feature attribution method tailored for GNNs, which quantifies node-level contributions based on ego-networks and the FESP framework. Given a graph $G = (V, E)$ and a trained GNN model f, we define the cooperative game (N, g) and its solution ϕ as follows:

- The set of players: The set of players N corresponds to the collection of ego-networks $\{Ego_\ell(v) | v \in V\}$ for a given layer ℓ.
- The characteristic function: Let the predicted class of the input graph G by model f be c. The characteristic function g is defined as: $g = f_c$, where $f_c(G')$ denotes the predicted probability for class c when the model is evaluated on subgraph $G' \subseteq G$.
- The solution ϕ: The attribution score $\phi_i(g)$ for node i is given by:

$$\phi_i(g) = \frac{1}{L} \sum_{\ell=1}^{L} \phi_i^{(\ell)}(g), \tag{5}$$

where $\phi^{(\ell)}$ represents the FESP attribution computed for radius ℓ over the cooperative game defined by the ego-networks at that radius, i.e., $(\{Ego_\ell(v) : v \in V\}, f_c)$. The parameter L controls the aggregation depth, i.e., the number of GNN layers considered when computing attributions. Since our goal is a model-agnostic explainer with no assumptions on the GNN architecture, L is treated as a hyperparameter, chosen empirically based on the relevance and stability of the explanations. More precisely, the optimal L is chosen by maximizing H-Fidelity [13], an objective function that measures the quality of the explanations.

The following result guarantees that our attribution method ESPAM adheres to a principled foundation by satisfying key axiomatic properties, including our proposed adaptation of Fair Treatment.

Theorem 1. *The solution ϕ satisfies the three axioms of Efficiency, Symmetry, and the proposed GNN-aware Fair Treatment (as defined in Equation (4))*[1].

B. From molecule-level to receptor-molecule interaction explanation

We propose a method to analyze how specific atoms within a molecule interact with distinct regions of a receptor's amino

[1]Proof available at https://github.com/atakml/ESPAM

acid sequence. This approach extends traditional molecular graph attribution by incorporating receptor–ligand interactions, thereby enhancing the interpretability of molecular binding mechanisms. G protein-coupled receptors (GPCRs), including olfactory receptors, are composed of well-defined structural segments, each contributing to different aspects of the receptor's function and activation. These segments can be broadly categorized as follows:

- **Transmembrane Helices:** GPCRs typically span the cell membrane through seven α-helical segments. These transmembrane helices form a barrel-like architecture, enclosing a ligand-binding pocket. They are critical for detecting extracellular molecules and transmitting conformational changes to the receptor's intracellular side.
- **Extracellular Loops (ECLs):** Positioned between the transmembrane helices on the extracellular side, these loops contribute to ligand specificity and binding affinity. They may undergo structural rearrangements during ligand interaction, influencing the receptor's activation state.
- **Intracellular Loops (ICLs):** Located on the cytoplasmic side, these loops connect the helices and mediate signal transduction by interacting with intracellular partners such as G proteins. Their configuration directly affects the initiation of downstream signaling cascades.

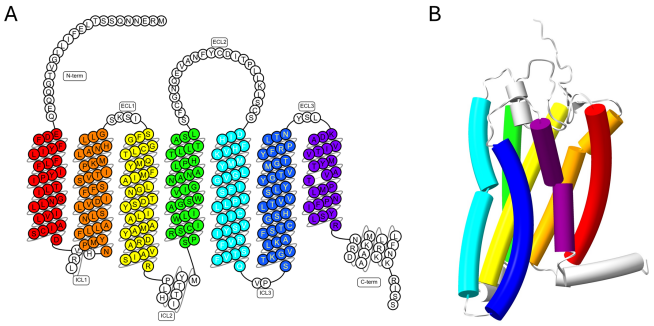

Fig. 2: Snake plot representation of the OR1A1 sequence (A) and three-dimensional structure (B). The colored regions indicate the transmembrane helices (TMH). The loops between TMH2-3, TMH4-5, and TMH6-7 are the extracellular loops 1, 2, and 3 (named ECL1, ECL2 and ECL3) while the loops between TMH1-2, TMH3-4, and TMH5-6 are the intracellular loops 1, 2, and 3 (ICL1, ICL2, ICL3).

Figure 2 presents a visualization of the segmental structure of the OR1A1 receptor. Since the transmembrane helices are primarily responsible for ligand binding and activation, we focus on assessing how each helical segment influences the model's interpretation of molecular atom importance. Specifically, we aim to understand how masking different helical regions of the receptor affects the contribution of individual atoms in the ligand to the model's prediction.

To formalize this, we construct a matrix $C \in \mathbb{R}^{n \times m}$, where each row corresponds to one of the n atoms in the molecule and each column to one of the m helical segments in the receptor. The entry C_{ij} quantifies the influence of helical segment j on the contribution of atom i to the model's decision.

Each row of the matrix is derived from a separate cooperative game, defined as follows:

Players: The set of players consists of all helical segments of the receptor.

Characteristic Function: For a given atom i, the characteristic function $g_i(S)$ measures the change in its contribution to the prediction when a coalition $S \subseteq$ helical segments of the receptor is masked. Formally:

$$g_i(S) = \text{sigmoid}\left(\phi_i(g) - \phi_i^S(g)\right), \tag{6}$$

where $\phi_i(g)$ is the original attribution score of atom i (computed using Equation (5)) and $\phi_i^S(g)$ is the attribution after masking the amino acids in segments S. Masking is performed by substituting the corresponding residues with the `[MASK]` token in the ProtBert-encoded [21] sequence, effectively removing their contribution from the receptor representation.

Given this setup, each entry C_{ij} is defined as the FESP attribution score of helical segment j in the game associated with atom i. That is:

$$C_{i,j} = \varphi_j(g_i), \tag{7}$$

where φ_j denotes the FESP solution for player j, and g_i is the characteristic function for atom i as defined in Equation (6).

Proposition 2. *The sum of columns at row i in the matrix C corresponds to the sigmoid of the original contribution of atom i to the model's prediction: $\sum_j C_{ij} = \text{sigmoid}\left(\phi_i(g)\right)$.*

By aggregating row-wise, we define the overall contribution of helical segment j as the sum of its corresponding row in the matrix C. We can then show the following:

Theorem 3. *Let the cooperative game be defined with helical segments as players and the characteristic function $g(S) = \sum_{i \in atoms} sigmoid(\phi_i - \phi_i^S)$. Then, the resulting attribution of helical segments, obtained by aggregating the matrix C satisfies the axioms of Efficiency, Symmetry, and the Fair Treatment.*

V. Experiments and Results

In this section, we address the following research questions related to the explainability and effectiveness of ESPAM:

- Q1 How does ESPAM compare to state-of-the-art explanation methods in terms of performance and fidelity?
- Q2 Can ESPAM capture changes in key molecular substructures when molecular properties are modified in the context of protein–molecule interaction tasks?
- Q3 Is ESPAM capable of revealing how specific atoms interact with the helical regions of the receptor, as interpreted by the underlying predictive model? To what extent can ESPAM help uncover the sources of the model's prediction errors, thereby improving trust and transparency?

To address Q1, we conduct two types of quantitative experiments. First, we evaluate the effectiveness of ESPAM using

a conventional GNN and benchmark it against state-of-the-art explanation methods on the synthetic BA2-Motifs dataset, as well as two real-world molecular datasets: AIDS and BBBP. This allows us to assess the faithfulness and relevance of the explanations in both controlled and practical settings. Second, we extend our evaluation to a more complex, large-scale scenario by applying ESPAM to five distinct sets of protein–molecule pairs, using the GNN-CLS model specifically designed for protein–ligand interaction prediction.

To address Q2 and Q3 research questions, we still use GNN-CLS along with known protein-molecule pairs, leveraging chemical background knowledge regarding their interactions.

The remainder of this section is organized as follows. In Section V-A, we present the evaluation metrics used to assess explanation quality. Section V-B addresses Q1 by outlining the evaluation protocol and comparing ESPAM against state-of-the-art methods on benchmark datasets. Section V-C focuses on Q2, evaluating the ability of ESPAM to identify relevant molecular substructures in protein–molecule interactions. Finally, in Section V-D, we address Q3 by providing qualitative results that illustrate ESPAM's behavior.

A. Evaluation Method

A good explanation should be both concise and faithful to the model's decision. In other words, it must not only highlight a minimal subset of input features but also capture the essential elements responsible for the model's prediction. Ideally, removing the explanation should significantly impact the prediction, while the explanation alone should approximate the original decision.

To quantitatively assess these properties, we employ three core metrics: fidelity, infidelity, and sparsity. Let G denote the input graph, m the explanation mask (a subgraph or node subset), and f the model being explained.

Fidelity evaluates how much the model's prediction drops when the explanatory part of the input is removed. A high fidelity score indicates that the explanation captures crucial elements of the model's decision: $\text{Fidelity}(G, m) = f_{c^\star}(G) - f_{c^\star}(G \setminus m)$, where $c^\star$ is the predicted class label with the highest confidence assigned by f to G.

Infidelity measures how closely the model's prediction on the explanation alone matches the original output. A lower infidelity indicates that the masked part alone is sufficient to support the model's prediction: $\text{Infidelity}(G, m) = f_{c^\star}(G) - f_{c^\star}(m)$.

Sparsity quantifies how compact the explanation is by measuring the relative size of the mask compared to the full input: $\text{Sparsity}(G, m) = 1 - \frac{|m|}{|G|}$, where $|.|$ denotes the number of nodes in the corresponding graph or subgraph.

These metrics are often in tension: for instance, using the entire input trivially maximizes fidelity and minimizes infidelity, but yields no sparsity. Conversely, an overly small mask may be sparse but provide poor fidelity and infidelity scores. Hence, evaluating them jointly is essential for fair and meaningful comparisons. To address this trade-off, we adopt the H-Fidelity metric [13] which harmonizes fidelity, infidelity, and sparsity into a single scalar measure. The idea is to reward explanations that are both effective (high fidelity, low infidelity) and concise (high sparsity). It does so by first normalizing fidelity and infidelity with respect to sparsity:

$\text{N-Fidelity} = \text{Fidelity}(G, m) \cdot (1 - \text{Sparsity}(G, m))$

$\text{N-Infidelity} = \text{Infidelity}(G, m) \cdot \text{Sparsity}(G, m)$

so that large masks are penalized when measuring fidelity and small masks are penalized when measuring infidelity. These normalized quantities are then combined via a harmonic function:

$$\text{H-Fidelity}(G, m) = \frac{(1 + \text{N-Fidelity})(1 - \text{N-Infidelity})}{(2 + \text{N-Fidelity} - \text{N-Infidelity})}$$

This formulation ensures that explanations which are compact, preserve the model's confidence, and faithfully reflect its reasoning receive higher scores. Throughout our experiments, we report H-Fidelity as a primary evaluation metric for comparing the overall quality of different explanation methods.

While the aforementioned metrics are designed to evaluate explanations in the form of *hard masks* (that is, discrete subgraphs extracted from the original input) many explainability methods, including ESPAM, GStarX, GraphSVX, and GradCAM, produce *soft masks*, which assign continuous importance scores to individual elements of the graph (e.g., nodes or edges). These soft masks are not directly compatible with the defined metrics, which assume a binary inclusion of input elements. To enable a fair and consistent evaluation across both soft- and hard-mask methods, we adopt the following standardization procedure. Given a soft mask generated for a graph G, we convert it into a corresponding hard mask by selecting the top-k most important elements such that the resulting mask satisfies a sparsity constraint of at least 0.5. In other words, the selected subset includes no more than half the original graph elements. Among all possible values of k, we choose the one that yields the highest H-Fidelity while satisfying this sparsity threshold. If multiple values of k produce similar results, we select the one that achieves the best trade-off between fidelity and sparsity. For methods that inherently produce hard masks, we directly use their outputs to compute H-Fidelity without any transformation.

B. Answering Q1: Comparison of ESPAM with state-of-the-art methods

To assess the performance of ESPAM relative to state-of-the-art explainability techniques, we conduct experiments using a standard Graph Neural Network (GNN) architecture. Specifically, we employ a model composed of three Graph Convolutional Network (GCN) layers [25], each with a hidden dimension of 20. The node-level embeddings are aggregated using both max and average pooling to construct a graph-level representation, which is then passed through a linear layer to produce the final prediction.

We evaluate this model across three datasets: BA2-Motifs, AIDS, and BBBP. The BA2-Motifs dataset [5] is a synthetic benchmark tailored for evaluating graph explanation methods. It comprises graphs labeled positive if they contain a house-shaped motif, and negative if they contain a 5-cycle, providing

a controlled environment to assess an explainer's ability to recover ground-truth motifs. The AIDS dataset [26], derived from molecular data, focuses on predicting anti-HIV activity and is widely used in the evaluation of molecular explainability. The BBBP dataset [27] assesses whether molecules can cross the blood-brain barrier, making it particularly pertinent for applications in pharmacokinetics and drug design.

For each dataset, we split the data into training (80%), validation (10%), and test (10%) sets. Model selection is performed based on the highest validation accuracy during training. Table I summarizes the key characteristics of these datasets and reports the corresponding classification performance of the GNN model used in our experiments.

Table II presents a comparison of ESPAM against seven state-of-the-art explainability methods: PGExplainer [5], GNNExplainer [4], GradCAM [6], PGMExplainer [5], GraphSVX [12], GStarX [13], and SubgraphX [10]. ESPAM consistently achieves the highest H-Fidelity scores across all three benchmark datasets, highlighting its superior capability in generating faithful and compact explanations. To evaluate the statistical significance of these results, we performed paired t-tests comparing ESPAM to each baseline method. The findings show that ESPAM significantly outperforms all competitors, with p-values consistently below 0.01, confirming strong statistical significance. The only exception occurs in the comparison with GraphSVX on the AIDS dataset, where the p-value is 0.06. Although this result still favors ESPAM, the difference does not meet the conventional significance threshold of 0.05.

C. Answering Q2: Evaluating ESPAM's ability to capture important molecular parts in protein–molecule interaction

We evaluate the ability of ESPAM to explain predictions made by the GNN-CLS model proposed in [1], focusing on its capacity to detect variations in crucial molecular substructures as molecular properties are altered. Our analysis is conducted on protein–molecule pairs involving four olfactory receptors with distinct response profiles: broad (OR1A1), specific (OR5K1 and OR51E2), and narrow (OR7D4). Additionally, we include a mixed test dataset comprising diverse protein–molecule interactions. These receptors were selected because their functional mechanisms are well-documented in the chemical and biological literature [28], [29], [30], [31], providing a basis for qualitative validation of the generated explanations.

Among the seven explainability methods evaluated in the previous section, only GraphSVX, GStarX, and SubgraphX are fully model-agnostic, relying solely on the model's output without requiring access to internal parameters or gradients. Therefore, we focus our comparison on these three approaches. However, due to SubgraphX's prohibitive computational cost, it is not practically applicable to the GNN-CLS model. As a result, GStarX and GraphSVX are retained as the primary baselines in this setting.

It is worth noting that GraphSVX does not natively support edge features, which are crucial in the GNN-CLS input representation. Although we attempted to modify GraphSVX to accommodate edge attributes, these efforts were unsuccessful. Given this limitation, we instead employ KernelSHAP [15] on the nodes of the graph for our protein–molecule interaction experiments. This approach is closely aligned with GraphSVX, differing primarily in that it does not attempt to reconnect disconnected components within coalitions. However, this distinction is not problematic in the context of GNN-CLS, which relies on transformer-based global attention rather than local connectivity. In fact, reconnecting disconnected components may introduce artifacts when explaining such architectures, and we argue that preserving the original graph structure yields more realistic attributions.

It is important to emphasize that the GNN-CLS model used in these experiments is pre-trained on a large and diverse set of protein–molecule pairs and remains fixed across all receptor categories. This ensures that any observed differences in explanation quality are attributable to the explanation methods themselves rather than to variations in model behavior. Table III provides detailed statistics for each receptor dataset along with the corresponding prediction performance of the model.

Table IV presents the H-Fidelity scores for three explanation methods across five receptor-molecule pair datasets. ESPAM consistently outperforms the baselines, achieving the highest H-Fidelity score on all datasets, indicating superior ability to generate concise and faithful explanations. While KernelSHAP generally ranks second, especially on OR51E2 and Mix, its performance shows higher variance. GStarX demonstrates stable but lower fidelity across the board. Paired t-tests reveal that ESPAM's improvements over both baselines are statistically significant, with p-values below 0.01 for all comparisons. These results highlight the effectiveness and robustness of ESPAM across a variety of molecular prediction tasks.

D. Answering Q3: Qualitative ESPAM's results

In this section, we focus on verifying the alignment of explanations generated by ESPAM with established background knowledge. Our experiments specifically analyze the explanations provided for molecular interactions with the olfactory receptors OR1A1, OR5K1, and OR51E2. Predictions for the OR7D4 receptor were not further analyzed due to the limited number of active pairs.

The qualitative experiment is divided into two complementary parts, each focusing on a different aspect of the molecule-receptor interaction. On the molecule side, we investigate whether the molecular substructures identified by ESPAM as important are indeed responsible for the binding interaction. Conversely, on the receptor side, we examine whether the predictions made by ESPAM can provide insights into the key features of the receptor that drive the binding interaction. Together, these two analyses aim to address the following question: Can ESPAM accurately identify the crucial factors on both the molecule and receptor sides that contribute to the binding interaction? We will also attempt to diagnose the causes of the model's prediction errors. Specifically, we will

TABLE I: Summary statistics of classical benchmark datasets and corresponding conventional GNN performance metrics.

Dataset	# Graphs	(#Positive, #Negative)	Avg Nodes	Avg Edges	Model Acc (Test)	F1	ROC-AUC	PR-AUC
BA2-Motifs	1000	(500, 500)	25	50.93	0.970	0.99	1.00	1.00
AIDS	2000	(400, 1600)	15.69	32.39	0.990	1.00	1.00	1.00
BBBP	1640	(389, 1251)	24.08	51.96	0.787	0.81	0.76	0.89

TABLE II: Comparison of H-Fidelity scores (mean ± standard deviation) achieved by different explanation methods across multiple datasets, each evaluated over five random seeds. For each dataset, the highest score is shown in bold. If the second-best score is not statistically different from the best, it is underlined.

Method	BA2-Motifs	AIDS	BBBP
PGExplainer	0.544 ± 0.0029	0.526 ± 0.0051	0.542 ± 0.0008
GNNExplainer	0.481 ± 0.0005	0.515 ± 0.0007	0.522 ± 0.0003
GradCAM	0.539 ± 0.0000	0.506 ± 0.0003	0.511 ± 0.0015
PGMExplainer	0.482 ± 0.0000	0.492 ± 0.0000	0.501 ± 0.0000
GStarX	0.549 ± 0.0013	0.527 ± 0.0002	0.532 ± 0.0001
GraphSVX	0.575 ± 0.0009	0.529 ± 0.0016	0.550 ± 0.0017
SubGraphX	0.541 ± 0.0121	0.502 ± 0.0028	0.521 ± 0.0096
ESPAM	**0.580 ± 0.0000**	**0.531 ± 0.0000**	**0.557 ± 0.0000**

examine whether the errors arise from the quality of the dataset used to train the GNN-CLS model.

1) Identifying the Binding Atoms: In this section we examine the capability of ESPAM in finding the atoms of the molecule responsible for the binding.

The results separated by the receptor are as follows[2]:

- **OR51E2:** This receptor is mainly sensitive to short-chain fatty acids [30]. In chemistry, a fatty acid is a carboxylic acid (R-COOH functional group) with an aliphatic chain (chain of carbon atoms). Figure 3 **(c–d)** shows that ESPAM effectively captures these functional groups and makes a different decision between long-chain and short-chain fatty acids. The model indeed captures these chemical functions as being important for the decision. For octanoic acid, it is interesting to note that although the carboxylic acid function is detected in the structure of this odorant, the decision to predict the molecule as inactive appears to come from the length of the aliphatic chain.
- **OR5K1:** [31] investigated how various ligands interact with the 5K1 odorant receptor. Their findings indicate that the aliphatic (non-aromatic) side chains attached to the pyrazine ring of these ligands play a crucial role in binding to OR5K1. The explanations in Figure 3 **(e–g)** demonstrate the same results as [31]. Interestingly, a comparison between Diethylpyrazine and 2-isopentyl-3-methoxypyrazine shows that ESPAM well destinguishes between the aliphatic chains and ether functional group (C-O-C).
- **OR1A1:** We evaluate the expressivity of ESPAM on two molecules (-)-carvone and (+)-carvone (Figure 3 **(a–b)**). Notably, the model identifies the carbonyl group (C=O) and the isopropenyl tail as key features distinguishing the two isomers. Functional assays by [28] found that (+)-carvone has reduced activity compared to (-)-carvone, due to the carbonyl group's hydrogen bonding within the binding pocket and the isopropenyl group's orientation. Our model successfully captures these subtle differences, highlighting its ability to discern nuanced structural features influencing binding affinity. As shown in Figure 3 **(a–b)**, ESPAM indicates that the oxygen atom of the ketone group and the isopropenyl group in (+)-carvone have lower activity than their counterparts in (-)-carvone, which are primarily responsible for binding. This difference arises from variations in chirality, affecting the relative positioning of these functional groups within the binding pocket, as reported by [28].

[2]The code to reproduce the result for all examined molecules is available at https://github.com/atakml/ESPAM

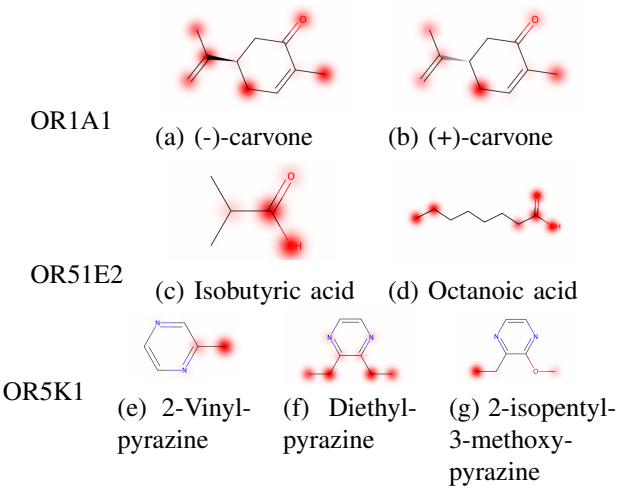

Fig. 3: Heat map visualizations highlighting substructural relevance for molecules binding to three olfactory receptors. **(a–b)** Stereoisomers of carvone interacting with OR1A1, both correctly predicted as positives. **(c–d)** Short- and long-chain fatty acids binding to OR51E2 with correct model predictions. **(e–g)** Molecules sharing a pyrazine scaffold binding to OR5K1, where the model consistently identifies aliphatic chains as important regions. These examples demonstrate ESPAM's effectiveness in identifying chemically meaningful binding patterns.

2) Identifying the Key Receptor Parts: In this section, we study the interaction between the atoms and the helical segments of the proteins. To this end, we use the molecules (-)-carvone and (+)-carvone as the molecular examples and the receptor OR1A1. The reason for selecting these molecules and this receptor is based on the existing background knowledge about the interaction between these two molecules and the

TABLE III: Datasets of receptor-molecule pairs and corresponding GNN-CLS performance metrics.

Proteins	#Pairs	(#Positive, #Negative)	Avg Nodes	Avg Edges	Model Acc (Test)	F1	ROC-AUC	PR-AUC
OR1A1	561	(136, 425)	10.33	20.02	0.81	0.70	0.90	0.75
OR5K1	240	(33, 207)	9.78	18.83	0.96	0.87	0.97	0.93
OR7D4	176	(14, 162)	11.68	23.50	0.96	0.82	0.98	0.81
OR51E2	196	(28, 168)	11.58	22.67	0.87	0.28	0.57	0.37
Mix	1565	(348, 1217)	10.05	19.39	0.86	0.70	0.86	0.69

TABLE IV: Comparison of H-Fidelity scores (mean ± standard deviation) across protein datasets for different explanation methods. Each row corresponds to a protein dataset, and the highest H-Fidelity score in each row is highlighted in bold.

Dataset	GStarX	KernelSHAP	ESPAM
OR5K1	0.508±0.0000	0.521±0.0002	**0.528±0.0000**
OR51E2	0.517±0.0000	0.531±0.0033	**0.549±0.0000**
OR1A1	0.546±0.0000	0.552±0.0011	**0.571±0.0000**
OR7D4	0.515±0.0000	0.524±0.0012	**0.536±0.0000**
Mix	0.523±0.0000	0.537±0.0004	**0.546±0.0000**

protein [28]. However, it should be noted that our goal is to investigate whether the interactions that the model identifies align with the knowledge of the activity of the receptor.

Our experiments involve a two-step process. First, we compute the matrix C from Equation (7). Then, for each segment, we mask it and measure the resulting change in the model's decision. By comparing the matrix, decision changes, and heatmaps from SectionV-D1, we can verify whether the matrix reflects the model's behavior. This comparison also enables us to determine if the model has learned the known interactions by aligning the matrix with background knowledge.

Figures 4 and 5 confirm that the total contribution of the segments is correlated with their ability to alter the decision when masked. Furthermore, when 3 **(a–b)** is compared with the heatmaps, it becomes clear that atoms with higher importance in the molecule tend to exhibit stronger interactions. These two observations collectively support the conclusion that the heatmap interactions are consistent with the model's behavior. Figures 4 and 5 also reveal that transmembrane helices 3 and 6 play a dominant role in the model's prediction for both molecules. It is well established that in GPCRs, transmembrane helices 3 and 6 harbor the most critical functional motifs for ligand recognition and receptor activation [32]. Notably, the contribution of these helices is more significant for (-)-carvone than for (+)-carvone, which is in line with the established activity profile of the receptor, where (-)-carvone exhibits higher activity than its (+) enantiomer.

3) Explanation of the Model Errors: Prediction errors are often related to the data used to train a model [33], particularly when the data is imbalanced. Therefore, we sought to analyze the role of local information density in the chemical space of the tested odorants for each OR. We counted the number of neighbors for each molecule and compared the results between two categories of molecules: those with correct predictions and those with incorrect predictions (Figure 6). Two molecules are considered neighbors if they have a Tanimoto coefficient less

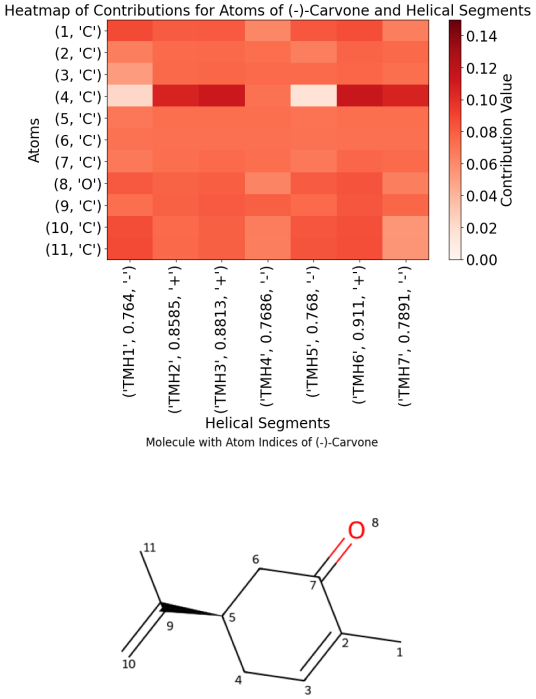

Fig. 4: Interaction Heat Map and Atom Indices of (-)-carvone. Numbers written next to each segment indicate its total contribution C_{*i} value. Plus indicates the decision change, and minus indicates no change in the decision by masking the segment.

than 0.3 using a Morgan2 fingerprint (2048 bits). Notably, for OR51E2 and OR5K1, the local information density is lower for molecules that are poorly predicted, suggesting that this may be a contributing factor to the model's errors. In contrast, the results for OR1A1 are more nuanced. While the average local information density is higher for correctly predicted molecules than for poorly predicted molecules, the difference is not statistically significant. This may be attributed to the greater chemical diversity of the OR1A1 dataset, given that it is a receptor with a very broad response spectrum.

VI. Conclusion

This study introduces the theoretical foundations of a novel method, called ESPAM, designed to explain the decision-making process of molecular graph neural networks. Our approach substantially outperforms existing methods in terms of accuracy, efficiency, and interpretability. Moreover, we have successfully applied ESPAM to a complex GNN-CLS model, which predicts the intricate interactions between a protein (here an olfactory receptor) and a small molecule (an odorant),

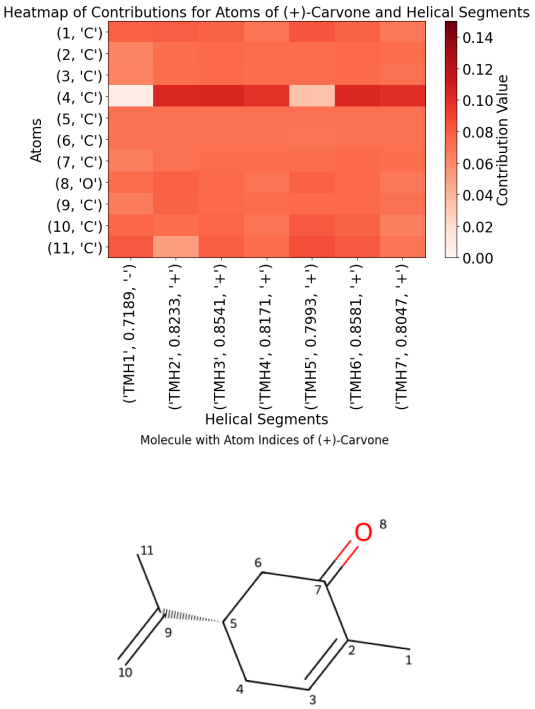

Fig. 5: Interaction Heat Map and Atom Indices of (+)-Carvone. Numbers written next to each segment indicate its total contribution C_{*i} value. Plus indicates the decision change, and minus indicates no change in the decision by masking the segment.

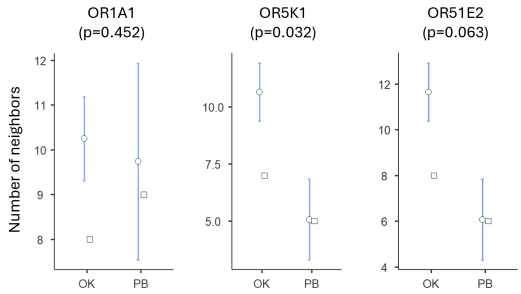

Fig. 6: Number of neighbors for correctly predicted (OK class) and poorly predicted (PB class) molecules in the datasets associated with the OR1A1, OR5K1, and OR51E2 receptors. A Mann-Whitney U test was performed to compare the distributions of neighbor counts between the two classes (circle and square are respectively average and median values).

yielding insights that are consistent with the known functions of these (macro)molecular entities.

References

[1] M. Hladis, M. Lalis, S. Fiorucci, and J. Topin, "Matching receptor to odorant with protein language and GNNs," in *ICLR*, 2023.

[2] Y. Chen, Y. Bian, B. Han, and J. Cheng, "How interpretable are interpretable graph neural networks?," in *ICML*, 2024.

[3] A. Ragno, B. L. Rosa, and R. Capobianco, "Prototype-based interpretable graph neural networks," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 4, pp. 1486–1495, 2024.

[4] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer," in *NeurIPS*, pp. 9240–9251, 2019.

[5] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," in *NeurIPS*, 2020.

[6] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *CVPR*, pp. 10764–10773, 2019.

[7] S. Azzolin, A. Longa, P. Barbiero, P. Lio, and A. Passerini, "Global explainability of GNNs via logic combination of learned concepts," in *The First Learning on Graphs Conference*, 2022.

[8] B. Armgaan, M. Dalmia, S. Medya, and S. Ranu, "Graphtrail: Translating GNN predictions into human-interpretable logical rules," in *Neurips*, 2024.

[9] P. Barbiero, G. Ciravegna, F. Giannini, M. E. Zarlenga, L. C. Magister, A. Tonda, P. Lió, F. Precioso, M. Jamnik, and G. Marra, "Interpretable neural-symbolic concept reasoning," in *ICML*, JMLR.org, 2023.

[10] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji, "On explainability of gnns via subgraph explorations," in *ICML*, pp. 12241–12252, 2021.

[11] L. S. Shapley, *A Value for n-Person Games*, pp. 307–318. Princeton University Press, 1953.

[12] A. Duval and F. D. Malliaros, "Graphsvx: Shapley value explanations for graph neural networks," in *ECML PKDD 2021*, p. 302–318, 2021.

[13] S. Zhang, Y. Liu, N. Shah, and Y. Sun, "Gstarx: Explaining graph neural networks with structure-aware cooperative games," in *NeurIPS*, 2022.

[14] G. Hamiache and F. Navarro, "Associated consistency, value and graphs," *Int. J. Game Theory*, vol. 49, no. 1, p. 227–249, 2020.

[15] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Neurips*, pp. 4765–4774, 2017.

[16] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the shapley value based on sampling," *Comp. Op. Res.*, vol. 36, no. 5, pp. 1726–1730, 2009.

[17] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.

[18] Y. Wang, Z. Li, and A. Barati Farimani, *Graph Neural Networks for Molecules*, pp. 21–66. 2023.

[19] T. Derr, Y. Ma, and J. Tang, " Signed Graph Convolutional Networks ," in *ICDM*, pp. 929–934, 2018.

[20] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *ICML*, vol. 70, pp. 1263–1272, PMLR, 2017.

[21] A. Elnaggar, M. Heinzinger, C. Dallago, and et al., "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE PAMI*, vol. 44, no. 10, pp. 7112–7127, 2022.

[22] Z. Wu and I. Savidis, "Circuit-gnn: A gnn for transistor-level modeling of analog circuit hierarchies," in *IEEE ISCAS*, 2023.

[23] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," 2017.

[24] C. Condevaux, S. Harispe, and S. Mussard, "Fair and efficient alternatives to shapley-based attribution methods," in *ECMLPKDD*, 2022.

[25] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[26] B. Wu, Y. Liu, B. Lang, and L. Huang, "DGCNN: Disordered GCN based on the gaussian mixture model," *Neurocomputing*, vol. 321, pp. 346–356, 2018.

[27] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, and Z. Wu, *Deep Learning for the Life Sciences*. 2019.

[28] C. Geithe, J. Protze, F. Kreuchwig, G. Krause, and D. Krautwurst, "Structural determinants of a conserved enantiomer-selective carvone binding pocket in the human odorant receptor or1a1," *CMLS*, vol. 74, no. 22, pp. 4209–4229, 2017.

[29] T. Shim, J. Pacalon, W.-C. Kim, X. Cong, J. Topin, J. Golebiowski, and C. Moon, "The third extracellular loop of mammalian odorant receptors is involved in ligand binding," *IJMS*, vol. 23, no. 20, 2022.

[30] C. B. Billesbølle, C. A. de March, W. van der Velden, and et al., "Structural basis of odorant recognition by a human odorant receptor," *Nature*, vol. 615, no. 7953, pp. 742–749, 2023.

[31] A. Nicoli, F. Haag, P. Marcinek, R. He, and et al., "Modeling the orthosteric binding site of the g protein-coupled odorant receptor or5k1," *JCIM*, vol. 63, no. 7, pp. 2014–2029, 2023.

[32] J. Topin, C. Bouysset, J. Pacalon, Y. Kim, M. Rhyu, S. Fiorucci, and J. Golebiowski, "Functional molecular switches of mammalian g protein-coupled bitter-taste receptors," *CMLS*, vol. 78, no. 23, 2021.

[33] W. P. Walters and R. Barzilay, "Applications of deep learning in molecule generation and molecular property prediction," *Accounts of Chemical Research*, vol. 54, no. 2, pp. 263–270, 2021.