SGCL-DPI: Structure-Guided Curriculum Learning for Drug-Protein Interaction Prediction

Othman Soufan[1]

[1] Sulaiman AlRajhi School of Business, Sulaiman AlRajhi University, Al Bukairiyah, Al-Qassim, Saudi Arabia

Corresponding Author:
Othman Soufan
Al Bukairiyah, Kingdom of Saudi Arabia
Email address: o.soufan@sr.edu.sa

Supplementary Material: Interpretability of SGCL-DPI via Integrated Gradients

Understanding the reasoning behind predictions made by deep learning models in drug–protein interaction (DPI) tasks is vital, particularly in biomedical domains where interpretability fosters trust, accountability, and scientific insight. In this study, we employed Integrated Gradients (IG), a gradient-based attribution method, to analyze the internal mechanisms of SGCL-DPI, a model that combines structural graph learning with guidance from a Random Forest (RF) teacher. The model processes molecular and protein graph structures and fuses their embeddings with coarse-grained predictions from an RF model, allowing it to leverage both fine-grained and global information during prediction.

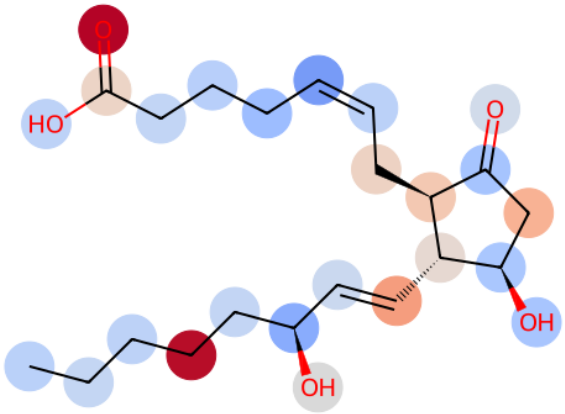
To assess the interpretability of SGCL-DPI, we visualized atom-level IG attributions on two representative molecules: one corresponding to a true positive (TP) prediction and the other to a false positive (FP) prediction. In these visualizations, red tones indicate atoms that positively contribute to the predicted interaction probability, while blue tones denote suppressive or negative contributions. Gray or neutral tones highlight atoms that had minimal influence on the decision.

The true positive sample (Supplementary Figure S2) reveals a clear concentration of high attribution scores around biologically meaningful substructures. Notably, strong attributions are observed on functional moieties such as a phosphate group, substituted aromatic rings, and polyhydroxy chains. These patterns suggest that SGCL-DPI has learned to focus on chemically significant regions known to facilitate molecular binding.

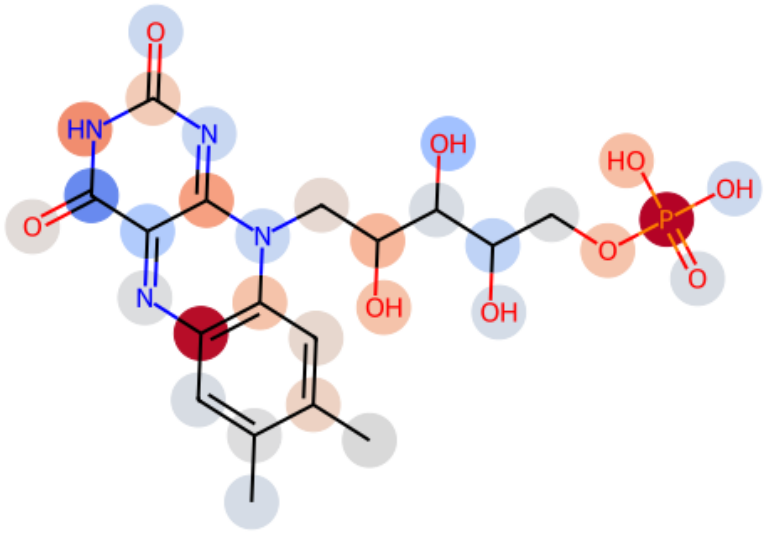
The attribution map is both sharp and localized, reinforcing the conclusion that the model's confident prediction was grounded in relevant chemical features.

In contrast, the false positive sample (Supplementary Figure S1) demonstrates a more diffuse and ambiguous attribution pattern. While some atoms are weakly highlighted, the lack of dominant red regions indicates that no specific substructure drove the model's high-confidence prediction. Instead, the attention appears scattered across non-specific hydroxyl chains and other motifs that, while common in active compounds, are insufficient to determine true binding. This behavior points to a tendency for overgeneralization or reliance on dataset artifacts, which may warrant further regularization or interpretability-aware training strategies.

Collectively, these interpretability analyses underscore the value of incorporating attribution methods like Integrated Gradients into model validation pipelines. In the case of SGCL-DPI, IG not only corroborates that the model attends to biochemically meaningful features in correct predictions but also provides insight into failure cases where diffuse or noisy attribution may underlie incorrect outputs. Such transparency is critical for building confidence in model predictions and guiding future improvements in data curation, model design, and deployment for real-world biomedical applications. Ask ChatGPT



Supplementary Figure S1: Integrated Gradients on False Positive Sample. This figure visualizes atom-level attributions on a false positive prediction. The model confidently predicted a binding interaction, but the ground-truth label indicates no binding. The attributions are widely dispersed with several moderately contributing atoms but no sharply dominant substructure. This diffuse attribution suggests the model may have overgeneralized from substructural motifs (e.g., hydroxyl chains) that are common in active molecules but not specific enough to indicate true interaction.



Supplementary Figure S2: Integrated Gradients on True Positive Sample. This figure presents attributions for a true positive prediction where the model correctly identified a drug–protein interaction. A strong concentration of red and orange attributions is evident around functional groups such as the phosphate moiety and substituted aromatic ring. These substructures are known to enhance molecular binding in biochemical literature, indicating that the model has likely learned meaningful chemical semantics. Blue regions denote suppressive or uninformative atomic contributions.

Implementation Details for Integrated Gradients-Based Interpretability

To enable post hoc interpretability of the SGCL-DPI model, we integrated a gradient-based attribution method, **Integrated Gradients (IG)**, to analyze the contribution of input substructures to final binding predictions. The following subsections describe the technical modifications and pipeline components introduced for this purpose.

Model Wrapping and Attribution Interface

We implemented a lightweight wrapper class FusionWrapper, encapsulating the model's fusion layer. This design enables compatibility with the Captum interpretability library, which requires a callable function mapping inputs (i.e., embeddings) to scalar predictions. The wrapper receives the GNN-derived drug and protein embeddings alongside the RF-based auxiliary prediction and returns the final fused score. This abstraction isolates the interpretability target while maintaining the original SGCL-DPI fusion behavior.

Extraction of Embeddings and Attention

A function run_ig_on_sample() was implemented to extract node-level representations from the GNN modules and aggregate them using the model's built-in attention mechanisms. Specifically:

- The drug_encoder and protein_encoder modules were executed on their respective graph structures.
- Attention weights were obtained via the learned attention heads and applied to the node embeddings to produce sample-specific graph-level embeddings.
- These aggregated embeddings, along with the RF prediction vector, were reshaped and marked with requires_grad_() to support gradient tracking.

These embeddings were then passed through Captum's IntegratedGradients class, where gradients are computed along a linear interpolation path between a baseline (e.g., zero vector) and the actual inputs.

Sample Selection for Case Analysis

To isolate meaningful case studies, we developed select_tp_fp_samples() to automatically extract one true positive (TP) and one false positive (FP) example. This function iteratively scans the dataset and returns two samples that:

- Are predicted with high confidence (based on a sigmoid threshold),
- Match the ground-truth label for TP, or diverge from it in the case of FP.

This automated sampling ensured that the visualizations reflected realistic and diagnostically relevant scenarios, while avoiding user bias or cherry-picking.

Molecule Visualization with Attribution

We introduced a new function visualize_molecule_with_ig() that converts SMILES strings to molecular graphs using RDKit and overlays atom-level IG scores. Key steps include:

- Aligning IG attribution values to the number of atoms in the molecule.
- Normalizing scores to the [0, 1] range.
- Mapping IG values to a diverging colormap (coolwarm), where:
    - **Red** represents atoms that positively contribute to the prediction,
    - **Blue** denotes atoms that suppress the prediction,
    - **Gray** indicates neutral or uninformative atoms.

The visualization was rendered using RDKit's MolDraw2DCairo engine and exported as high-resolution PNGs. Importantly, the color scale is automatically inferred from IG values without any manual annotation of TP/FP class, ensuring the visual outcome is attribution-driven rather than label-informed.

Robustness and Alignment

To ensure stability and accuracy, safeguards were implemented:

- Invalid SMILES strings or mismatched input dimensions between IG scores and atom counts are filtered.
- Per-atom attribution arrays are truncated or padded as necessary to align with molecular structure lengths.
- The embedding extraction and IG processes are wrapped in a no_grad() context except for the immediate attribution computation, reducing memory overhead and improving inference speed.