From Fragments to Geometry: A Unified Graph Transformer for Molecular Representation from Conformer Ensembles

Duy Minh Ho Nguyen [1 2 3] **Trung Quoc Nguyen** [2] **Ha Thi Hong Le** [2] **Mai Thanh Nhat Truong** [2]
TrungTin Nguyen [4] **Nhat Ho** [5] **Khoa D Doan** [6] **Duy Duong-Tran** [7 8] **Li Shen** [8] **Daniel Sonntag** [2 9] **James Zou** [10]
Mathias Niepert [1 3] **Hyojin Kim** [11] **Jonathan E Allen** [11]

Abstract

Designing and understanding molecules for biological applications requires models that can integrate rich structural information from both 2D molecular graphs and diverse 3D conformer ensembles. We introduce a fragment-aware, structure-guided graph transformer that enables scalable and expressive molecular modeling by aggregating multiple 3D conformers while incorporating fragment-level inductive biases from the 2D topology. Our approach employs a trainable attention-based fusion mechanism within a graph transformer to dynamically combine 2D and 3D representations, moving beyond static solvers and rigid fusion heuristics. This architecture enables fine-grained reasoning over chemically diverse molecules, including organocatalysts and transition-metal complexes. While originally developed for molecular property prediction, the method's structure-aware and fragment-level modeling is readily applicable to other downstream applications in drug discovery, reaction modeling, and AI-driven biological research. The model scales to large datasets and achieves state-of-the-art results across molecular property benchmarks, demonstrating its potential as a foundational component for generative AI in molecular science.

1. Introduction

Machine learning has become a transformative tool in computational biology, chemistry, and drug discovery, enabling predictive and generative modeling of molecular systems (Butler et al., 2018; Vamathevan et al., 2019; Choudhary et al., 2022; Fedik et al., 2022; Batatia et al., 2023). Many existing molecular representation learning methods rely on either 2D molecular graphs, which capture topological connectivity efficiently (Kipf & Welling, 2017; Gilmer et al., 2017b; Xu et al., 2018; Veličković et al., 2018), or on 3D representations derived from a single conformer (Schütt et al., 2017; Schütt et al., 2021b; Batzner et al., 2022; Batatia et al., 2022). While 2D graphs are computationally efficient, they lack essential geometric context - critical for understanding molecular interactions, activity, and design. Incorporating 3D conformers introduces spatial features such as bond lengths and torsion angles, but relying on a single conformation fails to reflect the intrinsic flexibility and thermodynamic diversity of real molecules.

In biological systems, molecules dynamically sample a range of conformations due to bond rotations, vibrational modes, and environmental interactions (Ramsundar et al., 2019), and many functionally relevant properties — such as solubility, binding affinity, or reactivity — emerge from this conformational ensemble (Perola & Charifson, 2004). Yet, fully modeling the conformational distribution remains computationally intensive, as quantum mechanical methods for conformer generation are expensive (Rosa et al., 2016; Wankowicz & Bonomi, 2025; Medrano Sandonas et al., 2024). This challenge has motivated hybrid learning models that combine the scalability of 2D graphs with the geometric richness of a small but representative subset of 3D conformers. These approaches are currently opening a promising path toward generative molecular design, where both topological and spatial variations are essential for modeling bioactive compounds and synthesizable drug-like molecules.

To address this, structure-aware ensemble methods based on optimal transport - especially those using fused Gromov-Wasserstein (FGW) alignment - have shown promise (Brogat-Motte et al., 2022; Ma et al., 2023; Nguyen et al., 2024a). By aligning both feature and geometric spaces, these models better preserve spatial correspondences across conformers and enable expressive ensemble aggregation. However, such methods are computationally expensive and struggle to scale to large molecular datasets such as Drugs-75k (Zhu et al., 2023; Axelrod & Gomez-Bombarelli, 2022), limiting their utility for high-throughput applications in generative biology.

[1]Max Planck Research School for Intelligent Sytems (IMPRS-IS) [2]German Research Center for Artificial Intelligence (DFKI) [3]University of Stuttgart [4]University of Queensland [5]University of Texas at Austin [6]VinUniversity [7]United States Naval Academy [8]University of Pennsylvania [9]Oldenburg University [10]Stanford University [11]Lawrence Livermore National Labs. Correspondence to: Duy Minh Ho Nguyen <ho_minh_duy.nguyen@dfki.de>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

In this work, we propose a **scalable alternative: a trainable, geometry-aware graph transformer** that replaces costly FGW alignment with efficient attention-based conformer aggregation. By supervising the model with FGW distances during training, we learn a latent embedding space where conformer similarities reflect both topological and geometric structure. This enables fast, permutation-invariant conformer integration suitable for large-scale generative pipelines. Beyond efficiency, we further enrich our model with fragment-level structural priors from 2D molecular graphs, injecting chemically meaningful hierarchies into both message passing and 3D attention layers. This unified 2D–3D framework captures fine-grained spatial and topological interactions essential for applications such as molecular property prediction, virtual screening, and functional optimization.

In summary, our key contributions are:

- We propose a **scalable, geometry-aware conformer aggregation framework**, denoted as FACET, that replaces costly FGW alignment with a trainable Graph Transformer, enabling efficient, deterministic attention-based inference. We further provide theoretical bounds on the approximation error relative to FGW distances.
- We introduce a unified 2D–3D representation learning approach that embeds **fragment-level structural priors** into both 2D message passing and 3D spatial self-attention, capturing multi-scale interactions between molecular topology and geometry.
- Our method delivers over **6× faster aggregation** than prior geometry-aware baselines and achieves **state-of-the-art performance** across six benchmarks, including molecular property prediction and Boltzmann-weighted ensemble tasks, demonstrating robustness across diverse molecular scenarios and dataset scales.

2. Related Work

2.1. Conformer Ensemble Learning in Molecular Representations

Molecular representations range from fingerprints (Morgan, 1965) and 1D strings (Ahmad et al., 2022; Wang et al., 2019) to 2D graphs (Gilmer et al., 2017a; Rong et al., 2020) and 3D geometric models (Fang et al., 2021; Zhou et al., 2023). While 2D models are efficient, they lack spatial context; 3D models add geometric detail but often rely on a single conformer, overlooking structural flexibility. Recent hybrid approaches combine 2D graphs with conformer ensembles (Zhu et al., 2024b; Axelrod & Gómez-Bombarelli, 2023), using aggregation techniques like pooling or self-attention (Zaheer et al., 2017; Vaswani et al., 2017). Geometry-aware methods based on FGW alignment (Brogat-Motte et al., 2022; Nguyen et al., 2024a) better capture spatial similarity across conformers but are computationally costly and struggle to scale in generative or high-throughput settings (Zhou et al., 2023). Our method addresses this limitation by using graph transformer architectures to learn latent embeddings of 3D conformers, integrating both geometry-aware signals - akin to those used in FGW-based methods—and hierarchical features from molecular fragments. This yields a scalable and permutation-invariant framework that balances computational efficiency with high representational power, making it well-suited for accuracy-critical molecular tasks.

2.2. Scalable Optimal Transport for Graph Learning

Recent advances in learning-based Optimal Transport (OT) have introduced efficient alternatives to classical solvers. Early work leveraged differentiable Sinkhorn distances with entropic regularization to improve stability and scalability (Cuturi, 2013; Feydy et al., 2019; Genevay et al., 2018). Subsequent methods enhanced computational efficiency through structural simplifications, such as low-rank approximations (Scetbon et al., 2021; Cuturi et al., 2020) and spatially-aware geometry-based formulations (Bachmann et al., 2022; Solomon et al., 2015). Additionally, meta-learning techniques accelerated optimization by learning better initializations (Amos et al., 2023), while more recent approaches have trained neural OT models directly on data to bypass iterative solvers altogether (Courty et al., 2017; Tong et al., 2021; Haviv et al., 2024).

Despite these advances, most of this work is limited to standard OT and does not extend to structure-aware variants like Fused Gromov-Wasserstein (FGW), which account for both feature similarity and relational graph structure. To address this, we introduce the first learned approximation of FGW via a graph transformer architecture, enabling scalable and geometry-aware aggregation across conformer ensembles. By integrating fragment-level structural priors into both 2D and 3D encoders, our framework supports multi-scale reasoning that unifies topological connectivity with spatial conformational diversity—essential for rich molecular representation and downstream biological modeling.

2.3. Fragment-biases in Molecular GNN

Fragment-level molecular substructures, such as rings, functional groups, and pharmacophores, play a central role in property prediction and drug development (Merlot et al., 2003; Varnek et al., 2005). Recent studies have harnessed these motifs for scaffold-aware molecule generation (Lee et al., 2024; Li, 2020; Chan et al., 2024), fragment-centric self-supervised tasks like masking and contrastive learning (Rong et al., 2020; Zhang et al., 2021; Wen et al., 2024), and in graph neural networks that encode inductive biases at the fragment level (Fey et al., 2020; Wang et al., 2025; Wollschläger et al., 2024). These approaches consistently demonstrate improved generalization, interpretability, and

data efficiency. Motivated by these findings, we take a complementary approach: embedding fragment-level priors directly into hybrid 2D–3D molecular models. Our method encodes hierarchical substructures into both 2D message passing and 3D spatial attention, supporting multi-scale reasoning across topological and geometric domains. This design enhances conformer ensemble aggregation and produces richer, geometry-aware representations for tasks that depend on molecular flexibility and spatial precision.

3. Fragment-Aware Conformer Ensemble Transformer

Notations. Let $\Delta_N := \{\boldsymbol{\omega} \in \mathbb{R}_+^N : \boldsymbol{\omega}^\top \mathbf{1}_N = 1\}$ denote the probability simplex, where $\mathbf{1}_N$ is the all-ones vector in $\mathbb{R}^N$. For $x \in \boldsymbol{\Omega}$, δ_x is the Dirac measure at x. We write $[K] := \{1, \ldots, K\}$ for $K \in \mathbb{N}$, and use $\langle \cdot, \cdot \rangle$ to denote the Frobenius inner product. For a tensor $\mathbf{L} = (L_{ijkl})$ and matrix $\boldsymbol{B} = (B_{kl})$, define the contraction $\mathbf{L} \otimes \boldsymbol{B} := \left(\sum_{kl} L_{ijkl} B_{kl}\right)_{ij}$. A graph $G = (V, E)$ has $N := |V|$ nodes and edges $E \subseteq \{\{u, v\} \subseteq V : u \neq v\}$. An attributed graph is given by $\mathcal{G} := (\boldsymbol{H}, \boldsymbol{A}, \boldsymbol{\omega})$, where $\boldsymbol{H} \in \mathbb{R}^{N \times d}$ is the node feature matrix (with row $\boldsymbol{H}_v$ for node v), $\boldsymbol{A}$ encodes structure (e.g., adjacency or shortest-path), and $\boldsymbol{\omega} \in \Delta_N$ is a node weight distribution.

Given two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ with N_1 and N_2 nodes, the *Fused Gromov-Wasserstein (FGW)* distance (Peyré et al., 2016; Titouan et al., 2019; 2020) is: $\mathrm{FGW}_{p,\alpha}(\mathcal{G}_1, \mathcal{G}_2) := \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)} \langle (1-\alpha)\boldsymbol{M} + \alpha \mathbf{L}(\boldsymbol{A}_1, \boldsymbol{A}_2) \otimes \boldsymbol{\pi}, \boldsymbol{\pi} \rangle$, where $\Pi(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2) := \{\boldsymbol{\pi} \in \mathbb{R}_+^{N_1 \times N_2} : \boldsymbol{\pi} \mathbf{1}_{N_2} = \boldsymbol{\omega}_1,\ \boldsymbol{\pi}^\top \mathbf{1}_{N_1} = \boldsymbol{\omega}_2\}$ is the set of valid couplings, $\boldsymbol{M}[i, j] = d_f(\boldsymbol{H}_1[i], \boldsymbol{H}_2[j])^p$ is the node feature cost, $\mathbf{L}(\boldsymbol{A}_1, \boldsymbol{A}_2)[i, j, l, m] = |\boldsymbol{A}_1[i, j] - \boldsymbol{A}_2[l, m]|^p$ captures structural mismatch, and $\alpha \in [0, 1]$ balances feature and structure alignment.

3.1. Conformer Generation

Following prior work, we generate molecular conformers using distance geometry methods that convert interatomic constraints, such as bond lengths, angles, stereochemistry, and steric limits, into 3D coordinates (Hawkins, 2017). A lightweight force field refines the structures toward low-energy conformations. Compared to quantum methods like DFT, this approach is highly scalable and efficient for large datasets. As in prior studies (Raza et al., 2022; Nguyen et al., 2024b), we use RDKit (Landrum, 2016) for fast and reliable conformer generation.

3.2. Framework Overview

We propose a neural architecture as in Figure 1 composed of three components. First, a 2D MPNN captures topological features from the molecular graph, while another MPNN operates on a fragment-induced hypergraph to encode higher-order structural priors (Sec.3.3). The outputs from both are fused and passed through a lightweight adaptor module, which dynamically refines and calibrates the feature representations before feeding them into the pre-trained graph transformer (Sec.3.4). Given a set of 3D conformers sampled from an input molecule graph, we use a 3D-MPNN to extract their embedding features (Sec. 3.4.1), followed by another adaptor layer. These adaptors are crucial for handling the variability in 3D conformer and 2D molecule features extracted by the 3D-MPNN ($\Phi(.)$) and 2D-MPNN. Then a graph transformer is used to aggregate the conformer feature sets into a geometry-aware molecular embedding, guided by atom-level and fragment-level attention. Finally, a permutation- and E(3)-invariant fusion module combines the 2D and 3D representations into a unified embedding for downstream tasks (Sec. 3.4.5).

3.3. Fragment-Enhanced 2D Molecular Graph

Each molecule is represented as a 2D graph $G = (V, E)$, where nodes V correspond to atoms and edges E to covalent bonds. Atom features $\mathbf{h}_v^{(0)} \in \mathbb{R}^d$ encode properties like atom type and valence, while bonds (u, v) are annotated with features $\boldsymbol{e}(u, v)$ (Scarselli et al., 2008; Gilmer et al., 2017a). We adopt a 2D message-passing neural network (MPNN) that updates node embeddings layer-wise:

$$\mathbf{h}_v^\ell = \mathsf{UPD}^\ell(\mathbf{h}_v^{\ell-1}, \mathsf{AGG}^\ell(\mathbf{M}^\ell(\mathbf{h}_v^{\ell-1}, \mathbf{h}_u^{\ell-1}, \boldsymbol{e}_{v,u}) \mid u \in N(v))), \tag{1}$$

where $\mathbf{M}^\ell$ is a message function, AGG^ℓ is sum aggregation, and UPD^ℓ is identity or multilayer perception layers. We use Graph Attention Networks (GATs) (Veličković et al., 2017), where messages are computed as:

$$\begin{aligned} \mathbf{M}_{v,u}^\ell &= \alpha_{v,u}^\ell \mathbf{W}^\ell \mathbf{h}_u^{\ell-1}, \\ \alpha_{v,u}^\ell &= \mathrm{softmax}_u\left(\mathrm{LeakyReLU}\left({}^\top[\mathbf{W}^\ell \mathbf{h}_v^{\ell-1}, |, \mathbf{W}^\ell \mathbf{h}_u^{\ell-1}]\right)\right). \end{aligned} \tag{2}$$

After L layers, we obtain final atom-level features $\mathbf{h}_v^L$ for each atom v used for downstream tasks.

Fragment-Based Structural Augmentation. To enhance atomic representations with higher-order structural context, we construct a fragment hypergraph from the input molecular graph G using ring-path decomposition (Kong et al., 2022; Geng et al., 2023; Wollschläger et al., 2024) to identify key substructures such as aromatic rings and functional groups (Fig. 2). Each fragment is treated as a node in a new graph $G^{\text{frag}} = (V^{\text{frag}}, E^{\text{frag}})$, where nodes correspond to fragments and edges are induced from the connectivity in G, two fragments are connected if they share an atom or are directly bonded.

We apply the same GAT formulation in Eq. (1) to the fragment graph to obtain fragment embeddings $\{\mathbf{h}_f^{\text{frag}}\}_{f \in V^{\text{frag}}}$. Then **for each atom v that belongs to its fragment** $f(v)$, we fuse their atom-level representations $\mathbf{h}_v^{(L)}$ with $\{\mathbf{h}_f^{\text{frag}}\}$

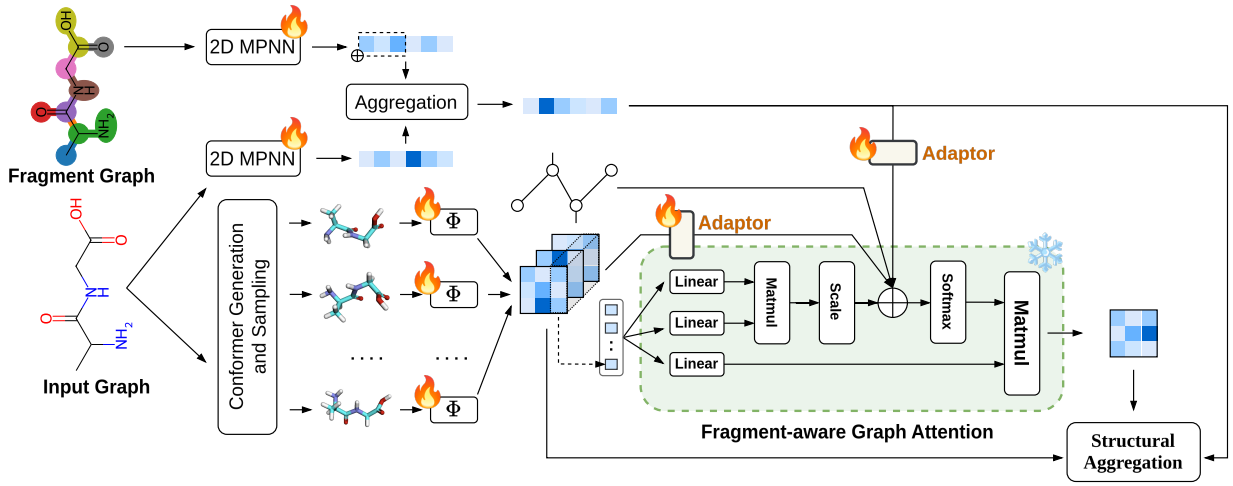


Figure 1. **FACET overview**. The model integrates 2D topological features from molecular and fragment graphs using message passing neural networks (MPNNs), combines them with 3D conformer features processed through a 3D-MPNN (Φ) and a lightweight adaptor, and feeds them into a fragment-aware graph attention for geometry-aware embedding guided by FGW distance. A final structural aggregation module fuses 2D and 3D representations into a unified embedding for downstream tasks.

by:

$$\widetilde{\mathbf{h}}_v^{(L)} = \mathbf{h}_v^{(L)} + \mathrm{FFN}\left(\mathbf{h}_{f(v)}^{\mathrm{frag}}\right), \tag{3}$$

where $\mathrm{FFN}(\cdot)$ is a learnable feedforward network that projects fragment-level context into the same space as atom features. If an atom belongs to multiple fragments, its atom-level features are aggregated with the corresponding fragment structures using the shared $\mathrm{FFN}(\cdot)$. Finally, we define a fragment-enhanced graph-level representation that is computed by applying a readout function $\mathbf{h}_{\mathrm{2D}} = \mathrm{READOUT}\left(\{\widetilde{\mathbf{h}}_v^{(L)} \mid v \in V\}\right) = \sum_{v\in V} \widetilde{\mathbf{h}}_v^{(L)}$. Intuitively, the **dual-level encoding** combining local atomic features and global fragment-level context as Eq.(3) allows the model to **reason over both fine-grained and coarse-grained structures**, enhancing the expressivity of the molecular representation.

3.4. Learning Graph Transformer for 3D Molecule Aggregations

A molecular conformer is represented as a set $S = \{\mathbf{r}_i, Z_i\}_{i=1}^N$, where N denotes the number of atoms, $\mathbf{r}_i \in \mathbb{R}^3$ corresponds to the 3D Cartesian coordinates of atom i, and $Z_i \in \mathbb{N}$ indicates its atomic number.

3.4.1. 3D CONFORMER FEATURE REPRESENTATION

For each conformer S, we can define its graph $\mathcal{G}_S$ and compute its 3D feature embedding by using a geometric message-passing network SchNet (Schütt et al., 2017), though other $E(3)$-invariant neural architectures can be readily substituted without modification. We represent the matrix of atom-level features from the final message-passing layer L of SchNet as $\mathbf{H}$, where each column $\mathbf{H}[v]$ corresponds to the feature vector $\mathbf{h}_{\mathrm{3d},v}^{(L)}$ of atom v. We then compute the vector representation for a conformer S as $\mathbf{h}_{\mathrm{3d,S}} = \sum_{v\in V}(\mathbf{W}_{\mathrm{3d}})\,\mathbf{h}_{\mathrm{3d},v}^{(L)} + \mathbf{b}_{\mathrm{3d}} \in \mathbb{R}^d$ with $\mathbf{W}_{\mathrm{3d}}$ and $\mathbf{b}_{\mathrm{3d}}$ are learnable vectors. Given a set of K conformers $\{\mathbb{S}_k\}_{k=1}^K$, we define $\mathbf{H}_{\mathrm{3d}}[k] = \mathbf{h}_{\mathrm{3D},\mathbb{S}_k}$ as the feature embedding for the k-th conformer. The matrix $\mathbf{H}_{\mathrm{3d}} \in \mathbb{R}^{K\times d}$ thus summarizes the feature representations of all conformers in the set.

3.4.2. FRAGMENT-AWARE GRAPH FORMER

Given the atom-wise feature matrix $\mathbf{H}$ for each conformer S, we aim to learn structure-encoded latent representations using Graph Transformer architectures (Zhang et al., 2020; Ying et al., 2021; Kreuzer et al., 2021; Luo et al., 2024). We adopt the architecture from (Ying et al., 2021) due to its strong expressiveness on small molecular graphs, and further *extend its attention mechanism with fragment substructures* (Fig .2). It is important to note that our framework is flexible and can incorporate alternative transformer-based models.

In particular, we compose N of transformer layers (Vaswani et al., 2017). Each Transformer layer consists of a self-attention mechanism followed by a position-wise feed-forward network. Given $\mathbf{H} = [\mathbf{h}_1^\top, \ldots, \mathbf{h}_n^\top]^\top \in \mathbb{R}^{n\times d}$ computed in Section 3.4.1 by a 3D-MPNN, where $\mathbf{h}_i = \mathbf{h}_{\mathrm{3d},v_i}^{(L)} \in \mathbb{R}^{1\times d}$ is the vector embedding of an atom v_i with d is the hidden size. We compute self-attention, by linearly projecting $\mathbf{H}$ into query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) matrices using learned weights $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$:

$$\mathbf{Q} = \mathbf{H}W_Q,\ \mathbf{K} = \mathbf{H}W_K,\ \mathbf{V} = \mathbf{H}W_V,$$

$$\widetilde{\mathbf{A}} = \mathbf{Q}\mathbf{K}^\top/\sqrt{d}, \quad \mathrm{Attention}(\mathbf{H}) = \mathrm{softmax}(\widetilde{\mathbf{A}})\mathbf{V}. \tag{4}$$

Here, $\widetilde{\mathbf{A}}$ denotes the attention score matrix representing pairwise similarities between tokens. For clarity, we present the single-head version; extending to multi-head attention is straightforward. Bias terms are omitted for brevity.

While the attention in Eq. (4) only uses feature nodes, leveraging the structural information of the 3D conformer graph is essential. Follow (Ying et al., 2021), we incorporate the (i) `centrality encoding`, measuring how important a node is in the graph using its degree, and (ii) `spatial encoding`, measuring spatial relation between two nodes v_i and v_j in a graph $\mathcal{G}_S$ by the distance of the shortest path distance (SPD) (Cormen et al., 2022; Balaban, 1985) together with a weighted learnabe value along edges of SPD between two nodes. Specifically, we incorporate (i) by:

$$\mathbf{h}_i = \mathbf{h}_i + z^-_{\deg^-(v_i)} + z^+_{\deg^+(v_i)}, \tag{5}$$

where $z^-, z^+ \in \mathbb{R}^d$ are learnable embedding vectors specified by the indegree $\deg^-(v_i)$ and outdegree $\deg^+(v_i)$ of atom v_i respectively. Assume $\widetilde{\mathbf{A}}_{ij}$ as the (i,j)-element of the Query-Key product matrix $\widetilde{\mathbf{A}}$, the condition (ii) extends $\widetilde{\mathbf{A}}_{ij}$ as:

$$\widetilde{\mathbf{A}}_{ij} = (\mathbf{h_i}\mathbf{W}_Q)(\mathbf{h}_j\mathbf{W}_K)^T/\sqrt{d} + s_{\phi(v_i,v_j)} + c_{ij}, \tag{6}$$

where $s_{\phi(v_i,v_j)}$ is a learnable scalar indexed by $\phi(v_i,v_j)$, which denotes for SPD distance between v_i and v_j, and shared across all layers; $c_{ij} = \mathbb{E}(x_{e_n}(w^E_n)^T)$ where $\mathbb{E}(.)$ is the expectation operation, x_{e_n} is the feature of the n-th edge e_n in SPD_{ij}, $w^E_n \in \mathbb{R}^{d_E}$ is the n-th weight embedding, and d_E is the dimensionality of edge feature compute as difference in feature embeddings of two nodes belong to it.

While the spatial encoding in Eq.(6) is implicated by the SPD, we argue that this might inadequately capture chemically meaningful substructures (ablation in Tab. 5). This motivates us to extend attention scores in Eq. (6) using values derived from (iii) `fragment-level node features` computed on 2D topology graph in Eq. (3), directly *guiding attention toward structurally and functionally relevant regions* such as rings, functional groups, or scaffolds. To this end, we compute an adjacency-like matrix $\mathbf{A}(G)$ using cosine distance over the final node embeddings $\widetilde{\mathbf{h}}^{(L)}_v$. Specifically, for each pair of atoms (v_i, v_j) in the 2D molecular graph, we define

$$\mathbf{A}(G)_{ij} = 1 - \frac{\langle \widetilde{\mathbf{h}}^{(L)}_i, \widetilde{\mathbf{h}}^{(L)}_j \rangle}{|\widetilde{\mathbf{h}}^{(L)}_i|_2 \cdot |\widetilde{\mathbf{h}}^{(L)}_j|_2}, \tag{7}$$

which quantifies their directional dissimilarity in the embedding space. Finally, we use the attention score as:

$$\widetilde{\mathbf{A}}_{ij} = (\mathbf{h_i}\mathbf{W}_Q)(\mathbf{h}_j\mathbf{W}_K)^T/\sqrt{d} + s_{\phi(v_i,v_j)} + c_{ij} + \mathbf{A}(G)_{ij}. \tag{8}$$

3.4.3. Learning to Approximate FGW Distance

We denote by $\mathcal{T}_\theta(.)$ be a graph transformer model that has its attention operation as Eq.(8), our goal is to train $\mathcal{T}_\theta(\cdot)$ to map the feature representation of each conformer S into a latent space where the L_2 distance between any pair S_i, S_j approximates their FGW distance - an *effective*, yet *computationally expensive*, geometry-aware metric (Brogat-Motte et al., 2022; Ma et al., 2023; Nguyen et al., 2024a). To this end, given a set of $\Omega = \{S_i\}_{i=1}^K$ of K generated conformers, we sample B conformers from Ω, then compute their encoding features by $\mathcal{T}_\theta(\mathbf{H}_i)$ for each $S_i \in B$. These outputs are compared with their pair-wise FGW distance to optimize the loss:

$$\mathcal{L}_{\text{enc}} = \sum_{ij} \left[||\mathcal{T}_\theta(\mathbf{H}_i) - \mathcal{T}_\theta(\mathbf{H}_j)||_2^2 - \text{FGW}_{p,\alpha}(\mathcal{G}(S_i), \mathcal{G}(S_j)) \right]. \tag{9}$$

By minimizing the loss $\mathcal{L}_{\text{enc}}$, we update the parameters of the transformation module $\mathcal{T}_\theta(\cdot)$ using gradient descent: $\theta \leftarrow \theta - \epsilon\nabla\mathcal{L}_{\text{enc}}$. Once trained, we freeze $\mathcal{T}_\theta$ and incorporate it back into the framework to compute a *geometry-aware representation* across K conformers $\{\mathbb{S}_k\}_{k=1}^K$ as follows: $\overline{\mathbf{H}} = \mathbb{E}\left(\{\mathcal{T}_\theta(\mathbf{H}_i)\}_{i=1}^K\right)$, where $\overline{\mathbf{H}}$ denotes the aggregated structural embedding. However, the 3D conformer feature distribution, extracted by 3D-MPNN, used to train $\mathcal{L}_{\text{enc}}$ (Eq. 9) may experience a *domain shift* when co-trained with other components in the full framework (Sec. 3.4.4) due to the continuous updating of 3D-MPNN. To address this, we design *adapter layers* as simple FFN layers to transform the input features in Eq. (9), aligning them to the seen distribution during training $\mathcal{T}_\theta$.

3.4.4. Invariant Aggregation of 2D and 3D Representation

We integrate representations from the 2D molecular graph and multiple 3D conformers using both average pooling and a GraphTransformer-based aggregation. The transformer captures rich spatial interactions while ensuring permutation invariance across conformers and E(3) equivariance, preserving robustness to 3D transformations. Given K conformers, using $\overline{\mathbf{H}}$ as the GraphTransformer (GT)-aggregated atom features. We compute the global GT representation as: $\mathbf{h}_{\text{GT}} = \sum_{v\in V}\left(\mathbf{W}_{\text{GT}}, \overline{\mathbf{h}}_v + \mathbf{b}_{\text{GT}}\right)$, where $\overline{\mathbf{h}}_v = \overline{\mathbf{H}}[v]$ and $\mathbf{W}_{\text{GT}}, \mathbf{b}_{\text{GT}}$ are learnable parameters. We then define $\mathbf{H}_{\text{2D}}$ and $\mathbf{H}_{\text{GT}}$ be the matrices whose columns are, respectively, K copies of the 2D feature $\mathbf{h}_{\text{2D}}$ (Sec.3.3) and $\mathbf{h}_{\text{GT}}$ representations from previous section. We fuse those representations with the 3D conformer features $\mathbf{H}_{\text{3D}}$ to produce the final atom-wise embedding: $\mathbf{H}_{\text{comb}} = \widetilde{\mathbf{W}}_{\text{2D}}\mathbf{H}_{\text{2D}} + \widetilde{\mathbf{W}}_{\text{3D}}\mathbf{H}_{\text{3D}} + \widetilde{\mathbf{W}}_{\text{GT}}\mathbf{H}_{\text{GT}}$, where each $\mathbf{W}_i$, $i \in \{\text{2D}, \text{3D}, \text{GT}\}$ are trainable projection matrix. The combined embedding $\mathbf{H}_{\text{comb}}$ is fed into a final FFN layer to predict the target property.

4. Theoretical Bounds for Embedding Non-Euclidean FGW Distance Matrices

Learning a Transformer $\mathcal{T}_\theta(.)$ to predict the FGW problem is closely related to multidimensional scaling (MDS) (Torgerson, 1952). Building on recent advances (Haviv et al., 2024; Sonthalia et al., 2021), we extend MDS theory to derive bounds on the error of embedding non-Euclidean distances,

specifically Wasserstein and FGW, into a Euclidean space suitable for graph transformer integration. While computing FGW barycenters is costly, our embedding enables efficient approximation via averaging and decoding in latent space. Prior work (Haviv et al., 2024) validated this approach for Wasserstein distances; we generalize it to FGW and provide theoretical justification, offering a scalable path for structure-aware graph alignment.

Cumulative Stress Optimization Problem via Pairwise FGW Distance Matrix. We define the **pairwise FGW distance matrix** $\boldsymbol{D}$ for a set of K distributions as $D_{ij} := \mathrm{FGW}_{p,\alpha}(\mathcal{G}(S_i), \mathcal{G}(S_j))$ for all $i, j \in [K]$, following Section 3.4.3. The **empirical FGW barycenter** is given by $\overline{\mathcal{G}}_K \in \arg\min_{\mathcal{G}\in\mathcal{P}_p(\boldsymbol{\Omega})} \frac{1}{K}\sum_{i=1}^{K} \mathrm{FGW}^p_{p,\alpha}(\mathcal{G}, \mathcal{G}(S_i))$, where $\mathcal{P}_p(\boldsymbol{\Omega})$ denotes the space of attributed graphs with finite p-th order FGW distance.

To approximate this barycenter in embedding space, we require $\|\overline{\boldsymbol{e}}_K - \boldsymbol{e}_j\|_2^2 \approx \mathrm{FGW}_{p,\alpha}(\overline{\mathcal{G}}_K, \mathcal{G}(S_j)) := \overline{D}_{K,j}$ for all $j \in [K]$, where $\overline{\boldsymbol{e}}_K = \frac{1}{K}\sum_{i=1}^{K} \boldsymbol{e}_i$ is the mean embedding and $\boldsymbol{e}_i := \mathcal{T}_\theta(\mathbf{H}_i)$ is the learned representation. To assess how well the embeddings $\{\boldsymbol{e}_i\}_{i=1}^K \subset \mathbb{R}^d$ preserve both pairwise FGW distances and barycenter structure, we define the **cumulative stress**: $\mathcal{S} = \min_{\boldsymbol{e}_i\in\mathbb{R}^d} \sum_{i,j\in[K]} \left(\|\boldsymbol{e}_i - \boldsymbol{e}_j\|_2^2 - D_{ij}\right)^2 + \sum_{j\in[K]} \left(\|\overline{\boldsymbol{e}}_K - \boldsymbol{e}_j\|_2^2 - \overline{D}_{K,j}\right)^2$. This objective enforces faithful reconstruction of both the distance structure and the barycenter alignment in the learned embedding space, as formalized in Theorem 4.1 (see proof in Appendix D).

Theorem 4.1. *Let $\boldsymbol{D}$ denote the pairwise $\mathrm{FGW}_{p,\alpha}$ distance matrix, and let $\{\lambda_i, \boldsymbol{v}_i\}_{i=1}^K$ represent the eigendecomposition of the associated criterion matrix $\boldsymbol{F} = -\boldsymbol{CDC}$, where $\boldsymbol{C} = \boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$ is the centering matrix. The optimal stress value, denoted by $\mathcal{S}^*$, is bounded as follows: $\mathcal{L} \le \mathcal{S}^* \le \mathcal{U}$, where $\mathcal{L} := \sum_{i:\lambda_i<0} \lambda_i^2$, $\mathcal{U} := \sum_{ij}(\Delta g_i + \Delta g_j)^2 + \mathcal{L} + \mathcal{C}$, $\Delta g_i = \frac{1}{2}\sum_{j:\lambda_j<0} \lambda_j \cdot \boldsymbol{v}_{ij}^2$. Here, $\boldsymbol{v}_{ij}$ denotes the j-th component of the i-th eigenvector $\boldsymbol{v}_n$ of $\boldsymbol{F}$, and $\mathcal{C}$ quantifies the approximation error between the empirical barycenter in the Euclidean embedding space and that in the original space of undirected attributed graphs.*

5. Experiments

5.1. Implementation Details

General pipeline. Our training consists of three stages. **Stage 1**: We train 2D and 3D MPNNs to extract features from molecular graphs and conformers. These features are also used to supervise the Graph Transformer in the next stage. **Stage 2**: The Graph Transformer is trained independently to approximate FGW distances between conformers using the features from Stage 1. We use the architecture of Graphormer (Ying et al., 2021), with 12 attention layers, 8 heads, and a hidden size of 64 (372k parameters). It is trained for 1000 epochs with a learning rate of $1e^{-5}$. **Stage 3**: We integrate all components into an end-to-end model, where only the 2D and 3D MPNNs are updated (300 epochs, learning rate $5e^{-4}$). To address feature distribution shift caused by finetuning, we apply FFN-based adaptor layers to the 2D and 3D features before feeding them to the Graph Transformer.

We use Adam for all stages. Further experimental details are provided in the Appendix.

5.2. Approximation of FGW Distance via Graph Transformer

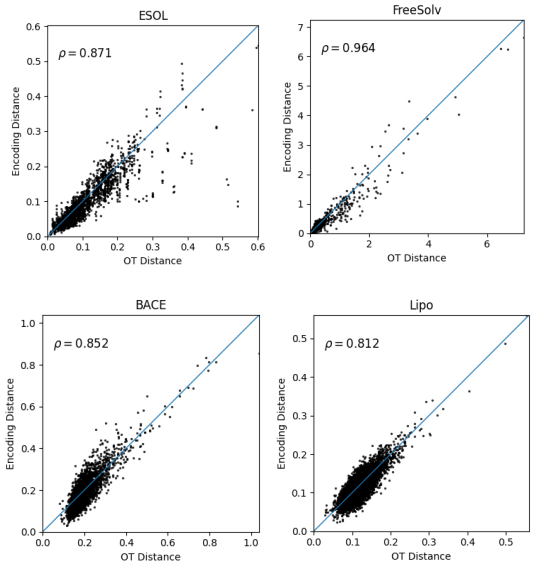


Figure 2. Correlations between FGW distance and trained Graph-Transformer on four datasets in **MoleculeNet** benchmark. For each test molecule, we compute pairwise FGW distances between conformers and compare them with Euclidean distances between their Graph Transformer embeddings. The correlation ρ is reported, with the reference line $y = x$ shown in blue.

Beyond theoretical estimation, we empirically evaluate how well the Graph Transformer approximates FGW distances between conformers in Euclidean space. As shown in Figure 2, results on the MoleculeNet benchmarks reveal a strong correlation between learned embeddings and true FGW distances, validating the transformer's effectiveness in simulating costly FGW computations. While correlation varies slightly across datasets, the results consistently highlight the model's reliability as a fast FGW surrogate, especially as the number of conformers in the aggregation increases

5.3. Scaling Fragment Geometry-Aware Aggregation

To validate the scalability of FACET model, based on a Graph Transformer for structure-aware aggregation, we compare it against Conan-FGW (Nguyen et al., 2024a), a method computing FGW distances on-the-fly during training and inference. We evaluate two key aspects: (i) ***inference-time efficiency with varying numbers of con-***

formers, and (ii) ***average training time per epoch at different dataset scales***. For inference, we measure the time required to generate output embeddings from K conformers ($K \in 5, 10, 15, 20$) using single and multi-GPU settings. Experiments are conducted on the BACE dataset and summarized in Figure 3.

It can be seen that (a) **FACET** exhibits strong scalability, maintaining a nearly constant runtime across varying numbers of conformers, in both single- and multi-GPU environments. In contrast, ConAN-FGW scales poorly, with runtime increasing sharply as the number of conformers grows. Although using multiple GPUs reduces the runtime compared to a single GPU, the upward trend persists, with runtimes surpassing 50 seconds for 20 conformers. (b) Secondly, FACET's similar runtime on single- and multi-GPU setups reflects its efficiency and the small workload in this experiment. In such cases, multi-GPU overhead can outweigh speedup gains. We expect multi-GPU acceleration to be beneficial mainly for large-scale tasks, like processing a large batch size of thousands of molecules or handling memory-intensive inputs.

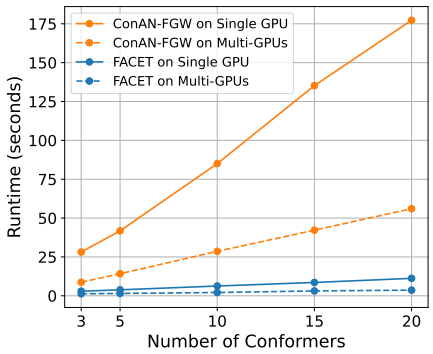


Figure 3. Total inference time on a test set of BACE to extract geometry-aware embedding aggregation using CoNan-FGW and our FACET.

In the second setting, we compare the average per-epoch training time of FACET and Conan-FGW on two datasets of different scales: Kraken (1,086 molecules) and Drugs-75k (52,569 molecules). As summarized in Figure 4, FACET exhibits linear scaling with the number of conformers and achieves 5–6× faster runtime on average than Conan-FGW. This efficiency is critical for scaling to large datasets and longer training schedules - for example, training Conan-FGW on Drugs-75k for 300 epochs requires 1,107.58 GPU hours, while FACET only takes 214 hours. This can be further reduced to 26.75 hours with 8 GPUs, compared to 138 hours for Conan-FGW under the same hardware setup.

5.4. State-of-the-Art Performance Comparison on Molecular Tasks

Datasets. We evaluate molecular property regression on the **MoleculeNet** (Wu et al., 2018) and **MARCEL** (Zhu et al., 2024a) benchmarks. **MoleculeNet** includes four

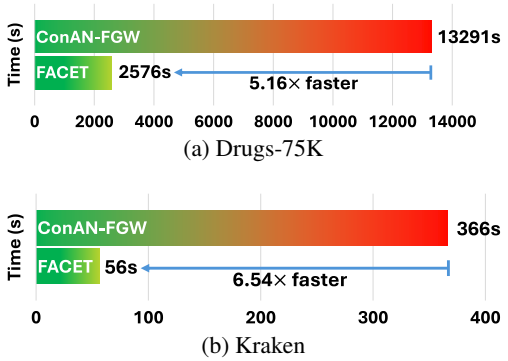


Figure 4. Comparison of the one-epoch training time of ConAN-FGW (Nguyen et al., 2024b) and the proposed FACET on the Drugs-75K and Kraken datasets from the **MARCEL** benchmark.

datasets, **ESOL**, **BACE**, **Lipo**, and **FreeSolv**, with targets covering solubility, inhibitory concentration (pIC_{50}), lipophilicity, and hydration free energy. **MARCEL** consists of **Drugs-75K** and **Kraken**, where the goal is to predict the Boltzmann-averaged property $\langle y \rangle_{k_B}$ from sampled conformers. **Drugs-75K** uses quantum descriptors (IP, EA, χ), while **Kraken** focuses on Sterimol features (B_5, L, and their buried forms). The Boltzmann average is computed as a weighted sum over conformer-specific values y_i with probabilities p_i. All datasets follow the original random split settings, using the provided sampled conformers.

Table 3. Number of samples for each split on tasks of **MoleculeNet** and the **MARCEL** benchmark.

Model	Lipo	ESOL	FreeSolv	BACE	Drugs-75k	Kraken
Train	2940	789	449	1059	52569	1086
Valid.	420	112	64	151	7509	155
Test	2940	227	129	303	15021	311
Total	4200	1128	642	1513	**75099**	1552

Baselines. For the **MoleculeNet** benchmark (Wu et al., 2018), we compare FACET with a wide range of baselines, including (i) `2D supervised` methods (e.g., GAT (Veličković et al., 2018), D-MPNN (Yang et al., 2019a), AttentiveFP (Xiong et al., 2019)), (ii) `pre-training approaches` (e.g., PretrainGNN (Hu et al., 2020b), GROVER (Rong et al., 2020), ChemBERTa-2* (Ahmad et al., 2022), ChemRL-GEM (Fang et al., 2022), MolFormer (Ross et al., 2022)), (iii) `3D-conformers` based models ConfNet (Liu et al., 2021)), UniMol (Zhou et al., 2023), SchNet (Schütt et al., 2017), ChemProp3D (Axelrod & Gómez-Bombarelli, 2023),ConAN-FGW (Nguyen et al., 2024b)). Training follows the setup in ConAN-FGW (Nguyen et al., 2024b).

For the **MARCEL** benchmark (Zhu et al., 2024a), we compare FACET against `2D models` (e.g., GIN (Xu et al., 2019), GIN+VN (Hu et al., 2020a), ChemProp (Yang et al., 2019b), GraphGPS (Rampášek et al., 2022)), `3D models` (e.g., SchNet (Schütt et al., 2017), DimeNet++ (Klicpera et al., 2020), GemNet (Gasteiger

Table 1. Comparison of molecular property regression performance on the **MoleculeNet** benchmark (MSE ↓). The results of competing methods are adapted from (Nguyen et al., 2024b). FACET uses a SchNet backbone.

Model	Lipo	ESOL	FreeSolv	BACE
2D-GAT (Veličković et al., 2018)	1.387 ± 0.206	2.288 ± 0.017	8.564 ± 1.345	1.844 ± 0.33
D-MPNN (Yang et al., 2019a)	0.534 ± 0.022	0.923 ± 0.045	4.213 ± 0.068	0.723 ± 0.021
Attentive FP (Xiong et al., 2019)	0.520 ± 0.001	0.771 ± 0.026	4.197 ± 0.193	-
PretrainGNN (Hu et al., 2020b)	0.545 ± 0.003	1.210 ± 0.005	6.392 ± 0.003	-
GROVER_large (Rong et al., 2020)	0.676 ± 0.012	0.798 ± 0.018	5.162 ± 0.047	-
ChemBERTa-2* (Ahmad et al., 2022)	0.639 ± 0.006	0.795 ± 0.033	-	1.858 ± 0.029
ChemRL-GEM (Fang et al., 2022)	0.486 ± 0.008	0.706 ± 0.061	3.924 ± 0.436	-
MolFormer (Ross et al., 2022)	0.492 ± 0.012	0.766 ± 0.026	5.485 ± 0.045	1.091 ± 0.021
ConfNet (Liu et al., 2021)	1.360 ± 0.038	2.115 ± 0.484	-	1.329 ± 0.042
UniMol (Zhou et al., 2023)	**0.374 ± 0.012**	0.741 ± 0.014	2.867 ± 0.186	-
SchNet-scalar (Schütt et al., 2017)	0.704 ± 0.032	0.672 ± 0.027	1.608 ± 0.158	0.723 ± 0.100
SchNet-emb (Schütt et al., 2017)	0.589 ± 0.022	0.635 ± 0.057	1.587 ± 0.136	0.692 ± 0.028
ChemProp3D (Axelrod & Gómez-Bombarelli, 2023)	0.602 ± 0.035	0.681 ± 0.023	2.014 ± 0.182	0.815 ± 0.170
CONAN (Nguyen et al., 2024b)	0.556 ± 0.013	0.571 ± 0.019	1.496 ± 0.158	0.635 ± 0.051
CONAN-FGW (Nguyen et al., 2024b)	0.422 ± 0.016	0.529 ± 0.022	1.068 ± 0.083	0.549 ± 0.016
FACET	0.424 ± 0.009	**0.516 ± 0.044**	**0.967 ± 0.082**	**0.495 ± 0.115**

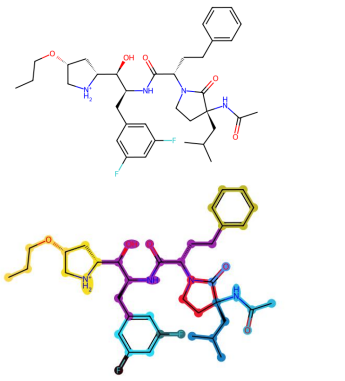


Table 2. RingsPaths decomposition on BACE, splitting molecules into rings, paths, and linkers. This reflects molecular topology and improves interpretability and generalization.

et al., 2021), PaiNN (Schütt et al., 2021a), ClofNet (Du et al., 2022), LEFTNet (Du et al., 2023)), and `ensemble strategies` such as DeepSets-based ensemble (Zaheer et al., 2017), self-attention (Vaswani et al., 2017), etc. All methods are evaluated under the same settings as described in the MARCEL benchmark.

5.4.1. RESULTS

MoleculeNet. FACET achieves state-of-the-art results on ESOL, FreeSolv, and BACE, reporting the lowest MSE scores across all tasks. Its consistent outperformance of CONAN-FGW underscores the value of incorporating fragment-level substructures into geometry-aware attention, which enhances the model's ability to capture localized chemical contexts and improves molecular property prediction.

MARCEL. FACET also demonstrates strong performance on the MARCEL benchmark, improving results across both SchNet and GemNet backbones. By combining structure-aware aggregation with hierarchical fragment information, FACET remains robust even at scale - unlike CONAN-FGW, which struggles with MARCEL's larger dataset. Together, these results highlight FACET's effectiveness and scalability across diverse molecular modeling tasks.

5.5. Ablation study

In this section, we analyze the key components of FACET through ablation studies. Specifically, we evaluate the impact of: (i) removing fragment structures from both the 2D MPNN and the self-attention mechanism in the graph transformer (**w/o Frag**); (ii) using fragments only in the 2D MPNN but not in the graph transformer (**w/o Frag in Trans.**); and (iii) omitting the trainable adaptor (**w/o Adap.**) that aligns 3D conformer features with the graph transformer, which can lead to performance degradation due to domain shift during training. As shown in Table 5, the absence of (i) significantly reduces performance, making FACET comparable to CONAN-FGW but with better scalability. Incorporating fragments into both components (ii) provides further gains, while (iii) proves essential for mitigating the domain shift introduced by changes in the 3D MPNN during training.

Table 4. Comparison of molecular property regression performance on the **MARCEL** benchmark (MAE ↓). The results of competing methods are adapted from (Zhu et al., 2024a).

Category	Model	Drugs-75K			Kraken			
		IP	EA	χ	B_5	L	$BurB_5$	BurL
2D models	GIN (Xu et al., 2019)	0.4354	0.4169	0.2260	0.3128	0.4003	0.1719	0.1200
	GIN+VN (Hu et al., 2020a)	0.4361	0.4169	0.2267	0.3567	0.4344	0.2422	0.1741
	ChemProp (Yang et al., 2019b)	0.4595	0.4417	0.2441	0.4850	0.5452	0.3002	0.1948
	GraphGPS (Rampášek et al., 2022)	0.4351	0.4085	0.2212	0.3450	0.4363	0.2066	0.1500
3D models	SchNet (Schütt et al., 2017)	0.4394	0.4207	0.2243	0.3293	0.5458	0.2295	0.1861
	DimeNet++ (Klicpera et al., 2020)	0.4441	0.4233	0.2436	0.3510	0.4174	0.2097	0.1526
	GemNet (Gasteiger et al., 2021)	0.4069	0.3922	**0.1970**	0.2789	0.3754	0.1782	0.1635
	PaiNN (Schütt et al., 2021a)	0.4505	0.4495	0.2324	0.3443	0.4471	0.2395	0.1673
	ClofNet (Du et al., 2022)	0.4393	0.4251	0.2378	0.4873	0.6417	0.2884	0.2529
	LEFTNet (Du et al., 2023)	0.4174	0.3964	0.2083	0.3072	0.4493	0.2176	0.1486
Ensemble Strategy with DeepSets	SchNet (Schütt et al., 2017)	0.4452	0.4232	0.2243	0.2704	0.4322	0.2024	0.1443
	DimeNet++ (Klicpera et al., 2020)	0.4126	0.3944	0.2267	0.2630	0.3468	0.1783	0.1185
	GemNet (Gasteiger et al., 2021)	0.4066	0.3910	0.2027	0.2313	**0.3386**	0.1589	**0.0947**
	PaiNN (Schütt et al., 2021a)	0.4466	0.4269	0.2294	**0.2225**	0.3619	0.1693	0.1324
	ClofNet (Du et al., 2022)	0.4280	0.4033	0.2199	0.3228	0.4485	0.2178	0.1548
	LEFTNet (Du et al., 2023)	0.4149	0.3953	0.2069	0.2644	0.3643	0.2017	0.1386
FACET	SchNet (Schütt et al., 2017)	0.4235	0.3971	0.2155	0.2508	0.3982	0.1803	0.1245
	GemNet (Gasteiger et al., 2021)	**0.3891**	**0.3852**	**0.1970**	**0.2225**	0.3402	**0.1503**	0.0952

Table 5. FACET ablation study.

Settings	FACET	w/o Frag.	w/o Frag. in Trans.	w/o Adap.
ESOL	**0.516**	0.531	0.525	0.546
FreeSolv	**0.967**	1.072	0.973	1.085

6. Conclusion

We present the FACET, a scalable method that integrates 3D conformer features with fragment-level 2D graph information. Using a trainable attention mechanism, it dynamically fuses 2D and 3D representations, outperforming FGW-based baselines across all MoleculeNet tasks. It also scales to 75,000 molecules and large conformer ensembles in the MARCEL benchmark, achieving state-of-the-art results in property and reaction prediction with efficient runtimes.

Acknowledgement

This work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016, the DARPA ANSR program under award FA8750-23-2-0004, the DARPA CODORD program under award HR00112590089. It was also partly funded by the Defense Threat Reduction Agency (DTRA), HDTRA1242044. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Duy M. H. Nguyen. Duy M. H. Nguyen and Daniel Sonntag are also supported by the No-IDLE project (BMBF, 01IW23002), the MASTER project (EU, 101093079), and the Endowed Chair of Applied Artificial Intelligence, Oldenburg University.

Disclaimer

The views expressed in this manuscript are those of the author(s) and do not reflect the official policy or position of the U.S. Naval Academy, Department of the Navy, the Department of Defense, or the U.S. Government.

References

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta-2: Towards chemical foundation models, 2022.

Amos, B., Cohen, S., Luise, G., and Redko, I. Meta optimal transport. *International Conference on Machine Learning*, 2023.

Axelrod, S. and Gomez-Bombarelli, R. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Axelrod, S. and Gómez-Bombarelli, R. Molecular machine learning with conformer ensembles. *Mach. Learn.: Sci. Technol.*, 4(3):035025, September 2023. ISSN 2632-2153. doi: 10.1088/2632-2153/acefa7.

Bachmann, F., Hennig, P., and Kobak, D. Wasserstein t-sne. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 104–120. Springer, 2022.

Balaban, A. T. Applications of graph theory in chemistry. *Journal of chemical information and computer sciences*, 25(3):334–343, 1985.

Batatia, I., Kovacs, D. P., Simm, G., Ortner, C., and Csanyi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436. Curran Associates, Inc., 2022.

Batatia, I., Benner, P., Chiang, Y., Elena, A. M., Kovács, D. P., Riebesell, J., Advincula, X. R., Asta, M., Baldwin, W. J., Bernstein, N., Bhowmik, A., Blau, S. M., Cărare, V., Darby, J. P., De, S., Pia, F. D., Deringer, V. L., Elijošius, R., El-Machachi, Z., Fako, E., Ferrari, A. C., Genreith-Schriever, A., George, J., Goodall, R. E. A., Grey, C. P., Han, S., Handley, W., Heenen, H. H., Hermansson, K., Holm, C., Jaafar, J., Hofmann, S., Jakob, K. S., Jung, H., Kapil, V., Kaplan, A. D., Karimitari, N., Kroupa, N., Kullgren, J., Kuner, M. C., Kuryla, D., Liepuoniute, G., Margraf, J. T., Magdău, I.-B., Michaelides, A., Moore, J. H., Naik, A. A., Niblett, S. P., Norwood, S. W., O'Neill, N., Ortner, C., Persson, K. A., Reuter, K., Rosen, A. S., Schaaf, L. L., Schran, C., Sivonxay, E., Stenczel, T. K., Svahn, V., Sutton, C., van der Oord, C., Varga-Umbrich, E., Vegge, T., Vondrák, M., Wang, Y., Witt, W. C., Zills, F., and Csányi, G. A foundation model for atomistic materials chemistry, 2023.

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.*, 13(1):2453, May 2022. ISSN 2041-1723.

Brogat-Motte, L., Flamary, R., Brouard, C., Rousu, J., and d'Alché Buc, F. Learning to predict graphs with fused gromov-wasserstein barycenters. In *International Conference on Machine Learning*, pp. 2321–2335. PMLR, 2022.

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., and Walsh, A. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, Jul 2018. ISSN 1476-4687.

Chan, B. W., Lynch, N. B., Tran, W., Joyce, J. M., Savage, G. P., Meutermans, W., Montgomery, A. P., and Kassiou, M. Fragment-based drug discovery for disorders of the central nervous system: designing better drugs piece by piece. *Frontiers in Chemistry*, 12:1379518, 2024.

Cheng, A. H., Sun, C., and Aspuru-Guzik, A. Scalable autoregressive 3d molecule generation. *arXiv preprint arXiv:2505.13791*, 2025.

Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., Park, C. W., Choudhary, A., Agrawal, A., Billinge, S. J. L., Holm, E., Ong, S. P., and Wolverton, C. Recent advances and applications of deep learning methods in materials science. *npj Comput. Mater.*, 8(1): 59, Apr 2022. ISSN 2057-3960.

Chung, Y. G., Haldoupis, E., Bucior, B. J., Haranczyk, M., Lee, S., Zhang, H., Vogiatzis, K. D., Milisavljevic, M., Ling, S., Camp, J. S., et al. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. *Journal of Chemical & Engineering Data*, 64(12):5985–5998, 2019.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to algorithms*. MIT press, 2022.

Courty, N., Flamary, R., and Ducoffe, M. Learning wasserstein embeddings. *International Conference on Learning Representations (ICLR)*, 2017.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Cuturi, M., Teboul, O., Niles-Weed, J., and Vert, J.-P. Supervised quantile normalization for low rank matrix factorization. In *International Conference on Machine Learning*, pp. 2269–2279. PMLR, 2020.

Du, W., Zhang, H., Du, Y., Meng, Q., Chen, W., Zheng, N., Shao, B., and Liu, T.-Y. Se (3) equivariant graph neural networks with complete local frames. In *International Conference on Machine Learning*, pp. 5583–5608. PMLR, 2022.

Du, Y., Wang, L., Feng, D., Wang, G., Ji, S., Gomes, C. P., Ma, Z.-M., 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.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Chemrl-gem: Geometry enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 2021. doi: 10.48550/ARXIV.2106.06130.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Fedik, N., Zubatyuk, R., Kulichenko, M., Lubbers, N., Smith, J. S., Nebgen, B., Messerly, R., Li, Y. W., Boldyrev, A. I., Barros, K., Isayev, O., and Tretiak, S. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nat. Rev. Chem.*, 6(9):653–672, Sep 2022. ISSN 2397-3358. doi: 10.1038/s41570-022-00416-3.

Fey, M., Yuen, J.-G., and Weichert, F. Hierarchical inter-message passing for learning on molecular graphs. *arXiv preprint arXiv:2006.12179*, 2020.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on aRtIfIcIaL InTeLlIgEnCe and Statistics*, pp. 2681–2690. PMLR, 2019.

Gasteiger, J., Becker, F., and Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Geng, Z., Xie, S., Xia, Y., Wu, L., Qin, T., Wang, J., Zhang, Y., Wu, F., and Liu, T.-Y. De novo molecular generation via connection-aware motif mining. *arXiv preprint arXiv:2302.01129*, 2023.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pp. 1263–1272, 2017a.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 06–11 Aug 2017b.

Gower, J. Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and its Applications*, 67:81–97, June 1985. ISSN 0024-3795. doi: 10.1016/0024-3795(85)90187-9. URL https://www.sciencedirect.com/science/article/pii/0024379585901879.

Haviv, D., Kunes, R. Z., Dougherty, T., Burdziak, C., Nawy, T., Gilbert, A., and Pe'Er, D. Wasserstein Wormhole: Scalable Optimal Transport Distance with Transformer. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17697–17718. PMLR, July 2024. URL https://proceedings.mlr.press/v235/haviv24a.html.

Hawkins, P. C. Conformation generation: the state of the art. *Journal of chemical information and modeling*, 57 (8):1747–1756, 2017.

Higham, N. J. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and*

its Applications, 103:103–118, May 1988. ISSN 0024-3795. doi: 10.1016/0024-3795(88)90223-6. URL https://www.sciencedirect.com/science/article/pii/0024379588902236.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020a.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020b.

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

Klicpera, J., Groß, J., Günnemann, S., et al. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, pp. 1–13, 2020.

Kong, X., Huang, W., Tan, Z., and Liu, Y. Molecule generation by principal subgraph mining and assembling. *Advances in Neural Information Processing Systems*, 35: 2550–2563, 2022.

Kreuzer, D., Beaini, D., Hamilton, W., Létourneau, V., and Tossou, P. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*, 34:21618–21629, 2021.

Landrum, G. Rdkit: open-source cheminformatics http://www.rdkit.org, 2016.

Lee, S., Lee, S., Kawaguchi, K., and Hwang, S. J. Drug discovery with dynamic goal-aware fragments. *International Conference on Machine Learning*, 2024.

Li, Q. Application of fragment-based drug discovery to versatile targets. *Frontiers in molecular biosciences*, 7: 180, 2020.

Liu, M., Fu, C., Zhang, X., Wang, L., Xie, Y., Yuan, H., Luo, Y., Xu, Z., Xu, S., and Ji, S. Fast quantum property prediction via deeper 2d and 3d graph networks. *arXiv preprint arXiv:2106.08551*, 2021.

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015.

Luo, Y., Li, H., Shi, L., and Wu, X.-M. Enhancing graph transformers with hierarchical distance structural encoding. *Advances in Neural Information Processing Systems*, 37:57150–57182, 2024.

Ma, X., Chu, X., Wang, Y., Lin, Y., Zhao, J., Ma, L., and Zhu, W. Fused gromov-wasserstein graph mixup for graph-level classifications. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Medrano Sandonas, L., Van Rompaey, D., Fallani, A., Hilfiker, M., Hahn, D., Perez-Benito, L., Verhoeven, J., Tresadern, G., Kurt Wegner, J., Ceulemans, H., et al. Dataset for quantum-mechanical exploration of conformers and solvent effects in large drug-like molecules. *Scientific Data*, 11(1):742, 2024.

Merlot, C., Domine, D., Cleva, C., and Church, D. J. Chemical substructures in drug discovery. *Drug Discovery Today*, 8(13):594–602, 2003.

Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, May 1965. ISSN 1541-5732. doi: 10.1021/c160017a018.

Nguyen, D. M., Lukashina, N., Nguyen, T., Le, A. T., Nguyen, T., Ho, N., Peters, J., Sonntag, D., Zaverkin, V., and Niepert, M. Structure-aware e (3)-invariant molecular conformer aggregation networks. *International Conference on Machine Learning*, 2024a.

Nguyen, D. M. H., Lukashina, N., Nguyen, T., Le, A. T., Nguyen, T., Ho, N., Peters, J., Sonntag, D., Zaverkin, V., and Niepert, M. Structure-aware E(3)-invariant molecular conformer aggregation networks. In *International Conference on Machine Learning*, pp. 37736–37760. PMLR, 2024b.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., and Yamazaki, M. Polyinfo: Polymer database for polymeric materials design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, pp. 22–29. IEEE, 2011.

Perola, E. and Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *Journal of medicinal chemistry*, 47(10):2499–2510, 2004.

Peyré, G., Cuturi, M., and Solomon, J. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2664–2672, New York, New York, USA, June 2016. PMLR.

Peyré, G., Cuturi, M., and others. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. Publisher: Now Publishers, Inc.

Rampášek, L., Galkin, M., Dwivedi, V. P., Luu, A. T., Wolf, G., and Beaini, D. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

Ramsundar, B., Eastman, P., Walters, P., and Pande, V. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. O'Reilly Media, 2019.

Raza, A., Henle, E. A., and Fern, X. Non-equilibrium molecular geometries in graph neural networks. *arXiv preprint arXiv:2203.04697*, 2022.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

Rosa, M., Micciarelli, M., Laio, A., and Baroni, S. Sampling molecular conformers in solution with quantum mechanical accuracy at a nearly molecular-mechanics cost. *Journal of Chemical Theory and Computation*, 12 (9):4385–4389, 2016.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Scetbon, M., Cuturi, M., and Peyré, G. Low-rank sinkhorn factorization. In *International Conference on Machine Learning*, pp. 9344–9354. PMLR, 2021.

Schütt, K., Kindermans, P., Felix, H. E. S., Chmiela, S., Tkatchenko, A., and Müller, K. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 991–1001, 2017.

Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Schütt, K., Unke, O., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pp. 9377–9388. PMLR, 2021a.

Schütt, K. T., Unke, O. T., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *ICML*, pp. 1–13, 2021b.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

Sonthalia, R., Van Buskirk, G., Raichel, B., and Gilbert, A. How can classical multidimensional scaling go wrong? In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12304–12315. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/66121d1f782d29b62a286909165517bc-Paper.pdf.

Titouan, V., Courty, N., Tavenard, R., Laetitia, C., and Flamary, R. Optimal Transport for structured data with application on graphs. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284. PMLR, June 2019.

Titouan, V., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. Fused Gromov-Wasserstein Distance for Structured Objects. *Algorithms*, 13(9):212, August 2020. ISSN 1999-4893. doi: 10.3390/a13090212. URL https://www.mdpi.com/1999-4893/13/9/212.

Tong, A. Y., Huguet, G., Natik, A., MacDonald, K., Kuchroo, M., Coifman, R., Wolf, G., and Krishnaswamy, S. Diffusion earth mover's distance and distribution embeddings. In *International Conference on Machine Learning*, pp. 10336–10346. PMLR, 2021.

Tong, V., Trung-Dung, H., Liu, A., Broeck, G. V. d., and Niepert, M. Learning to discretize denoising diffusion odes. *International Conference on Learning Representations (ICLR))*, 2025.

Torgerson, W. S. Multidimensional Scaling: I. Theory and Method. *Psychometrika*, 17(4):401–419, 1952. ISSN 0033-3123. doi: 10.1007/BF02288916. URL https://www.cambridge.org/core/product/5A026426380B7639E7EA6D92D5DF19AB. Edition: 2025/01/01 Publisher: Cambridge University Press & Assessment.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.*, 18(6):463–477, Jun 2019. ISSN 1474-1784.

Varnek, A., Fourches, D., Hoonakker, F., and Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *Journal of computer-aided molecular design*, 19:693–703, 2005.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *ICLR*, 2018.

Wang, J., Min, Y., Li, M., and Wu, J. Fragformer: A fragment-based representation learning framework for molecular property prediction. *Transactions on Machine Learning Research*, 2025.

Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. Smilesbert: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, pp. 429–436, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366663. doi: 10.1145/3307339.3342186.

Wankowicz, S. A. and Bonomi, M. From possibility to precision in macromolecular ensemble prediction. *arXiv preprint arXiv:2505.01919*, 2025.

Wen, Q., Ju, M., Ouyang, Z., Zhang, C., and Ye, Y. From coarse to fine: enable comprehensive graph self-supervised learning with multi-granular semantic ensemble. In *Forty-first International Conference on Machine Learning*, 2024.

Wollschläger, T., Kemper, N., Hetzel, L., Sommer, J., and Günnemann, S. Expressivity and generalization: Fragment-biases for molecular gnns. *International Conference on Machine Learning*, 2024.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, pp. 513–530, 2018.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., and Jegelka, S. Representation learning on graphs with jumping knowledge networks. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5453–5462. PMLR, 10–15 Jul 2018.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *International Conference on Learning Representations*, 2019.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, July 2019a. ISSN 1549-960X. doi: 10.1021/acs.jcim.9b00237.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019b.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Zhang, J., Zhang, H., Xia, C., and Sun, L. Graph-bert: Only attention is needed for learning graph representations. *arXiv preprint arXiv:2001.05140*, 2020.

Zhang, Z., Liu, Q., Wang, H., Lu, C., and Lee, C.-K. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882, 2021.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhu, Y., Hwang, J., Adams, K., Liu, Z., Nan, B., Stenfors, B., Du, Y., Chauhan, J., Wiest, O., Isayev, O., Coley, C. W., Sun, Y., and Wang, W. Learning over molecular conformer ensembles: Datasets and benchmarks, 2023.

Zhu, Y., Hwang, J., Adams, K., Liu, Z., Nan, B., Stenfors, B., Du, Y., Chauhan, J., Wiest, O., Isayev, O., et al. Learning over molecular conformer ensembles: Datasets and benchmarks. In *International Conference on Learning Representations*, 2024a.

Zhu, Y., Hwang, J., Adams, K., Liu, Z., Nan, B., Stenfors, B., Du, Y., Chauhan, J., Wiest, O., Isayev, O., et al. Learning over molecular conformer ensembles: Datasets and benchmarks. *International Conference on Learning Representations (ICLR)*, 2024b.

Supplementary Materials for "From Fragments to Geometry: A Unified Graph Transformer for Molecular Representation from Conformer Ensembles"

Contents

A. Limitations of FACET

A.1. FACET Operates on a Predefined Set of 3D Conformers.

Our method enables efficient geometry-aware aggregation without requiring expensive alignment procedures at inference time. While FACET demonstrates improved performance even with a small subset of conformers, *the quality and representativeness of this subset can still influence downstream predictions*. In particular, if the selected conformers are heavily biased or fail to capture key structural variations, some aspects of molecular flexibility may be underrepresented. Addressing this challenge through better conformer sampling strategies or task-aware selection mechanisms could further enhance model robustness, especially for highly flexible molecules.

Future direction: A promising extension would be to develop end-to-end models that can learn to generate conformers dynamically during training, using gradient feedback from downstream prediction losses. Such a differentiable conformer generation module could enable task-aware structural modeling, ensuring that the generated conformers are optimized not just for physical plausibility, but also for relevance to the predictive task at hand.

A.2. Limitations in Scope

A.2.1. Focusing on Small Molecules

FACET has primarily been evaluated on standard molecular property prediction benchmarks such as those in MoleculeNet, which consist mostly of small, drug-like molecules. While this setup is well-suited for many pharmacological applications, it limits the assessment of FACET's generalizability to more complex molecular systems. For example, **biomacromolecules** (e.g., peptides, proteins, nucleic acids) exhibit high flexibility, long-range dependencies, and hierarchical organization that are not present in small molecules. **Polymers and materials** often involve much larger structures without well-defined conformers, challenging FACET's reliance on discrete 3D inputs. Additionally, FACET currently models only single-molecule properties and has not been extended to multi-molecular interactions, such as protein-ligand binding.

Future direction: To broaden FACET's applicability, several promising future directions can be explored. First, incorporating efficient attention to capture both local fragment-level information and long-range structural dependencies is essential for handling large biomolecules. Second, adapting FACET to support flexible input formats, such as voxel grids or material-specific graphs, would allow it to process polymers and crystalline materials that lack stable conformers. Third, extending FACET to jointly model molecular interactions through cross-graph attention or co-embedding mechanisms could open applications in drug docking and molecular complex prediction. Finally, applying and evaluating FACET on broader

datasets, such as PDBbind (Liu et al., 2015), PolyInfo (Otsuka et al., 2011), or CoRE MOF 2019 (Chung et al., 2019), would provide a more comprehensive understanding of its strengths and limitations across molecular domains.

A.2.2. Limitation in Generative Capabilities.

While FACET demonstrates strong performance on discriminative tasks such as molecular property prediction across MoleculeNet and MARCEL benchmarks, its current formulation and evaluation are limited to regression settings where the goal is to predict properties from given molecular structures. As a result, the model's potential for generative applications such as de novo molecule generation, scaffold decoration, or fragment-based drug design remains unexplored. This limits our understanding of how well FACET can serve as a foundational model for inverse molecular problems, where structural creativity and diversity are critical. Future work should explore extensions of FACET with generative decoding mechanisms, such as auto-regressive sampling (Cheng et al., 2025), diffusion models (Tong et al., 2025), or variational objectives, to fully leverage its design for structure-conditioned generation.

B. Implementation Details

Our training pipeline includes three stages: In the first stage, we train only the 2D and 3D MPNNs to learn corresponding features from 2D molecular graph and 3D conformers. The features in this stage also serve as a dataset for approximating Graph Transformer to the FGW distance. In the next stage, the Graph Transformer is trained separately to simulate the costly computation of FGW distance between two conformers using learned features from stage 1. In the last stage, Graph Transformer is integrated in a single end-to-end training with 2D and 3D MPNNs. At this stage, only 2D and 3D MPNNs are trained. As a result of changing MPNNs during the last stage, a shift in the distribution of the Graph Transformer input might occur. We solve this problem by adding an adaptor layer using an MLP on both 3D and 2D features before feeding them to the GraphTransformer. For all experiments on the **MoleculeNet** and **MARCEL** benchmarks, we use the same number of conformers as specified in their original settings.

In all stages, we use Adam as our optimizer. We train our model on an 8 V100-GPUs cluster.

Stage 1. Learning 2D and 3D features. For each molecule, we define by $\mathbf{H}_{\text{2d-3d}} = \widetilde{\mathbf{W}}_{\text{2D}}\mathbf{H}_{\text{2D}} + \widetilde{\mathbf{W}}_{\text{3D}}\mathbf{H}_{\text{3D}}$, we then train for 150 epochs and set the learning rate to $1e^{-3}$. to optimize target property tasks $\mathcal{L}_{\text{pred}} = ||\widehat{\boldsymbol{y}}_{\text{2d-3d}} - \widetilde{\mathbf{y}}||_2^2$ where $\widetilde{\mathbf{y}}$ be the ground-truth value and $\widehat{\boldsymbol{y}}$ be our predicted value defined by:

$$\widehat{\boldsymbol{y}}_{\text{2d-3d}} = \mathbf{W}^{\mathcal{G}}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{\text{2d-3d}}[k]\right) + \mathbf{b}^{\mathcal{G}}, \tag{10}$$

with $\mathbf{W}^{\mathcal{G}}$ and $\mathbf{b}^{\mathcal{G}}$ are learnable parameters.

Stage 2. Training Graph Transformer to approximate FGW distance. The Graph Transformer is trained separately in the second stage to approximate the FGW distance by Euclidean embedding space. For the Graph Transformer architecture, we employ the same setting as Graphormer from (Ying et al., 2021). Specifically, a number of attention layers, a number of attention heads, and the hidden dimension of the transformer are set to 12, 8, and 64, respectively, which makes the total number of parameters of the Graph Transformer 372k. In our attention, we use the shortest-path distance (SPD) between a pair of nodes. Following practical implementation in (Ying et al., 2021), we pre-compute SPD distance for each 3D molecule graph and load these values during training and inference. We set a learning rate of $1e^{-5}$ and train for 1000 epochs with the following loss function:

$$\mathcal{L}_{\text{enc}} = \sum_{ij}\left[||\mathcal{T}_\theta(\mathbf{H}_i) - \mathcal{T}_\theta(\mathbf{H}_j)||_2^2 - \text{FGW}_{p,\alpha}(\mathcal{G}(S_i), \mathcal{G}(S_j))\right]. \tag{11}$$

Stage 3. Training Fragment-aware Graph Transformer. In the final stage, we freeze the trained GraphTransformer $\mathcal{T}\theta(\cdot)$ and use it to compute aggregated features from 3D conformer embeddings generated by the 3D-MPNN. To accommodate potential distribution shifts, we add lightweight FFN adaptor layers on top of both the 2D- and 3D-MPNNs used in $\mathcal{T}\theta(\cdot)$, while continuing to update the MPNNs during training. The full model is trained for 300 epochs with a reduced learning rate to optimize the training loss $\mathcal{L}_{\text{pred}} = ||\widehat{\boldsymbol{y}} - \widetilde{\mathbf{y}}||_2^2$ where

$$\widehat{\boldsymbol{y}} = \mathbf{W}^{\mathcal{G}}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{H}_{\text{comb}}[k]\right) + \mathbf{b}^{\mathcal{G}}. \tag{12}$$

C. Additional Analysis of FACET's Scalability and Performance with More 3D Conformers

In this section, we further analyze FACET's scalability on the following two factors:

C.1. Inference Time when Increasing the Number of 3D Conformers for Each Molecule.

We compare FACET against two versions of CONAN-FGW in running time to extract structure-aware embedding aggregation with different input of 3D conformers. We use two variations of CONAN-FGW, including a single GPU version and another relaxed solver that permits running Sinkhorn iterations on GPUs by matrix multiplication, thus supporting distributed multi-GPUs acceleration. The experiments are conducted on a **single GPU** using a batch size of 32 molecules, each with different conformers ranging from 3, 5, 10, 15, and 20, and another experiment with **four GPUs** on the same batch size, i.e., 8 molecules per GPU.

Figure 5 indicates our observations across four datasets of **MoleculeNet** benchmark, where we report the required time to extract embedding aggregations for all molecules in the test set. We see that (i) **FACET** demonstrates excellent scalability where its runtime remains nearly constant regardless of the number of conformers, both in single-GPU and multi-GPU settings. In contrast, ConAN-FGW shows poor scalability where runtime increases steeply with the number of conformers. While the multi-GPU usage improves runtime over single-GPU, the growth trend remains significant, with runtimes still exceeding 30 seconds at 20 conformers (e.g., with ESOL dataset).

Secondly, the nearly identical runtime of FACET across single- and multi-GPU settings, as shown in the plot, can be attributed to its computational efficiency and the relatively small workload in this experiment. In such cases, the overhead introduced by multi-GPU parallelization - such as inter-GPU communication and data synchronization - can outweigh its potential speedup benefits. Therefore, we argue that multi-GPU acceleration for FACET becomes advantageous only under substantially larger workloads, such as batch processing of thousands to millions of molecules or handling complex input representations that exceed the memory capacity of a single GPU.

C.2. Average Training Time per Epoch as a Function of Dataset Size.

We analyze the scalability of FACET with respect to the number of training molecules. To this end, we report the average training time per epoch across four datasets from the MoleculeNet benchmark. Figure 6 compares the training time of FACET and ConAN-FGW on a single GPU, using a batch size of 256 and 5 conformers per molecule. As shown in the figure, FACET achieves a 2.28× to 3.17× speedup over ConAN-FGW. Notably, this speedup is roughly proportional to the number of training molecules in each dataset, as reported in Table 3.

D. Proof of Theorem 4.1

Recall that we aim to establish the following novel theoretical bounds: Let $\boldsymbol{D}$ denote the pairwise $\text{FGW}_{p,\alpha}$ distance matrix, and let $\{\lambda_k, \boldsymbol{v}_k\}_{k=1}^K$ represent the eigendecomposition of the associated criterion matrix $\boldsymbol{F} = -\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}$, where $\boldsymbol{C} = \boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$ is the centering matrix. The optimal stress value, denoted by $\mathcal{S}^*$, is bounded as follows: $\mathcal{L} \leq \mathcal{S}^* \leq \mathcal{U}$, where

$$\mathcal{L} := \sum_{k:\lambda_k<0} \lambda_k^2, \quad \mathcal{U} := \sum_{kl} (\Delta g_k + \Delta g_l)^2 + \mathcal{L} + \mathcal{C}, \quad \Delta g_k = \frac{1}{2} \sum_{l:\lambda_l<0} \lambda_l \cdot \boldsymbol{v}_{kl}^2, \ \forall k \in [K].$$

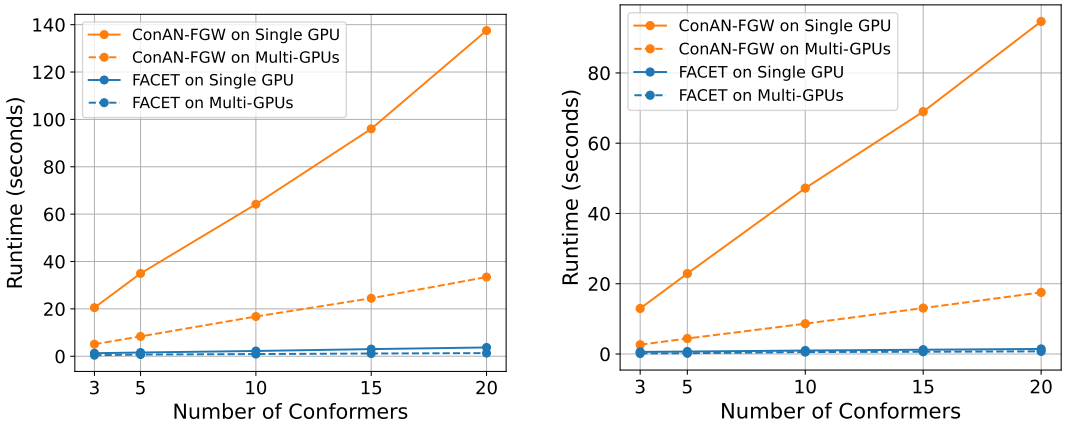
Here, $\boldsymbol{v}_{kl}$ denotes the l-th component of the k-th eigenvector $\boldsymbol{v}_k$ of $\boldsymbol{F}$, and $\mathcal{C}$ quantifies the approximation error between the empirical barycenter in the Euclidean embedding space and its counterpart in the original space of undirected attributed graphs. This is equivalent to that given $\boldsymbol{e} := \{\boldsymbol{e}_k\}_{k\in[K]} \in \mathbb{R}^{d\times K}$, our objective is to derive lower and upper bounds for the following cumulative stress:

$$\mathcal{S}^* = \min_{\boldsymbol{e}\in\mathbb{R}^{d\times K}} \mathcal{S}(\boldsymbol{e}), \quad \mathcal{S}(\boldsymbol{e}) = \mathcal{S}_1(\boldsymbol{e}) + \mathcal{S}_2(\boldsymbol{e}), \tag{13}$$

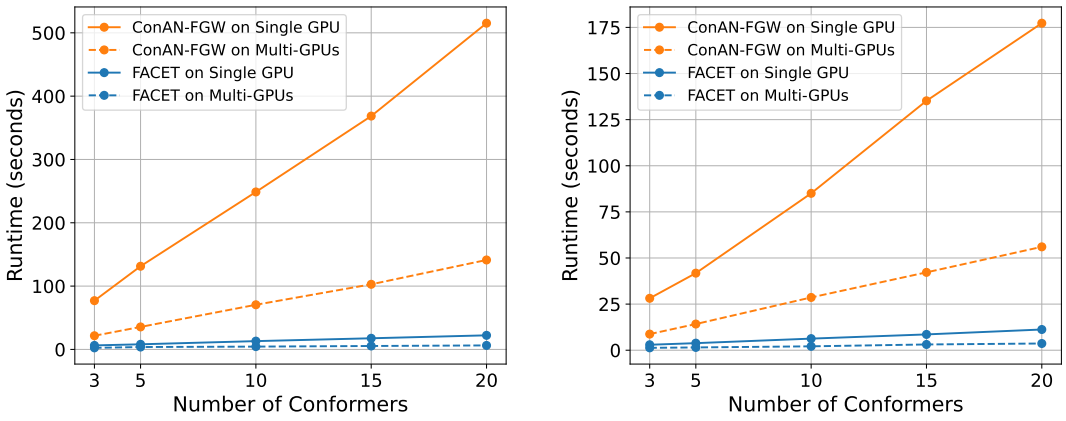
$$\mathcal{S}_1^* := \min_{\boldsymbol{e}\in\mathbb{R}^{d\times K}} \mathcal{S}_1(\boldsymbol{e}), \quad \mathcal{S}_1(\boldsymbol{e}) := \sum_{k,l\in[K]} \left(\|\boldsymbol{e}_k - \boldsymbol{e}_l\|_2^2 - D_{kl}\right)^2, \tag{14}$$

$$\mathcal{S}_2^* := \min_{\boldsymbol{e}\in\mathbb{R}^{d\times K}} \mathcal{S}_2(\boldsymbol{e}), \quad \mathcal{S}_2(\boldsymbol{e}) := \sum_{l\in[K]} \left(\|\overline{\boldsymbol{e}}_K - \boldsymbol{e}_l\|_2^2 - \overline{D}_{K,l}\right)^2. \tag{15}$$

To this end, we begin by specifying and formally defining the following important concepts in Appendix D.1.

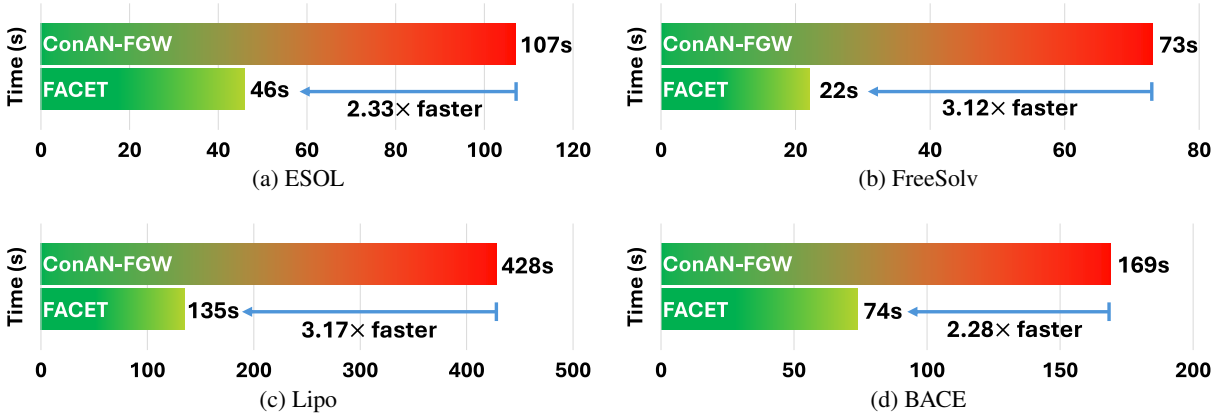


(a) ESOL (b) FreeSolv



(c) Lipo (d) BACE

Figure 5. Runtime comparison of structure-aware embedding aggregation between ConAN-FGW (Nguyen et al., 2024b) and the proposed FACET on four datasets from the MoleculeNet benchmark. Results are shown for both single-GPU and 4-GPU configurations. Reported runtimes represent the total time required to extract structural embeddings for all molecules in the test set of each dataset.



(a) ESOL (b) FreeSolv

(c) Lipo (d) BACE

Figure 6. Comparison of the one-epoch training time of ConAN-FGW (Nguyen et al., 2024b) and the proposed FACET on four datasets from the **MoleculeNet** benchmark.

D.1. Non-Euclidean Nature of Pairwise FGW Distance Matrix

Definition D.1 (Euclidean Distance Matrix)**.** A $K \times K$ distance matrix $\boldsymbol{D}$ is said to be *Euclidean* if there exists a set of points $\boldsymbol{e} = \{\boldsymbol{e}_k\}_{k=1}^K$ in some Euclidean space $\mathbb{R}^d$ such that

$$\forall k, l \in [K], \quad D_{kl} = \|\boldsymbol{e}_k - \boldsymbol{e}_l\|_2^2.$$

The space of all Euclidean distance matrices (EDM) is denoted by $\mathcal{E}$.

Fact 1 (Conditions for Euclidean Distance Matrix, see, *e.g.,* (Gower, 1985)). A matrix $\boldsymbol{D}$ is an EDM if and only if it satisfies the following three conditions:

(i) Non-negativity: $D_{kl} \geq 0$ for all $k, l \in [K]$,

(ii) Hollow diagonal: $D_{kk} = 0$ for all $k \in [K]$,

(iii) Positive semidefiniteness: the associated double-centered matrix $\boldsymbol{F} := -\boldsymbol{C}\boldsymbol{D}\boldsymbol{C}$ is positive semidefinite (PSD), where $\boldsymbol{C} = \boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$ is the centering matrix, and $\mathbf{1}_K$ denotes the K-dimensional vector of ones.

Recall that the pairwise FGW distance matrix $\boldsymbol{D}$ for a collection of K distributions is defined entry-wise by $D_{kl} := \mathrm{FGW}_{p,\alpha}(\mathcal{G}(\mathbb{S}_k), \mathcal{G}(S_l))$ for all $k, l \in [K]$, as introduced in Section 3. The following result establishes that this matrix does not correspond to a Euclidean distance matrix:

Lemma D.2 (Non-Euclidean Nature of Pairwise FGW Distance Matrix)**.** *Consider the case where $d_f = \|\cdot\|_2$. Then the FGW distance matrix $\boldsymbol{D}$, whose entries are given by*

$$FGW_{p,\alpha}(\mathcal{G}_1, \mathcal{G}_2) := \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)} \langle (1-\alpha)\boldsymbol{M} + \alpha\, \mathbf{L}(\boldsymbol{A}_1, \boldsymbol{A}_2) \otimes \boldsymbol{\pi}, \boldsymbol{\pi} \rangle\,,$$

with $\alpha \in [0, 1]$, does not define a Euclidean distance matrix.

As established in Lemma D.2, which is proved in Appendix D.4, the distance $\mathrm{FGW}_{p,\alpha}$ is not a Euclidean distance. Therefore, we are interested in quantifying how accurately non-Euclidean distance matrices can be approximated by pairwise distances between learned embeddings. To this end, we analyze the lower and upper bound of the set $\mathcal{S}$ in Appendices D.2 and D.3, respectively.

D.2. Lower Bounds on Embedding non-Euclidean FGW Distances

We would like to find the lower bound of $\mathcal{S}$. We note that the original formulation is non-convex, making it analytically intractable. Nonetheless, by reparameterizing the objective as a function of the pairwise squared distances $\widehat{D}_{kl} := \|\boldsymbol{e}_k - \boldsymbol{e}_l\|_2^2$ and $\overline{\widehat{D}}_{Kl} := \|\overline{\boldsymbol{e}}_K - \boldsymbol{e}_l\|_2^2$ induced by the embedding, and by incorporating the necessary conditions to ensure that $\widehat{\boldsymbol{D}}$ corresponds to a valid Euclidean distance matrix, the reformulated problem becomes convex for $\mathcal{S}_1$. Note that we can prove that $\mathcal{S}$ has a lower bound at $\widehat{\boldsymbol{L}}^*$, where $\widehat{\boldsymbol{L}}^*$ is a minimizer of $\mathcal{S}_1$, that is,

$$\mathcal{S}^* = \min_{\widehat{\boldsymbol{D}} \in \mathcal{E}} \left[\mathcal{S}_1(\widehat{\boldsymbol{D}}) + \mathcal{S}_2(\widehat{\boldsymbol{D}})\right], \quad \mathcal{S}_2(\widehat{\boldsymbol{D}}) := \sum_{l \in [K]} \left(\overline{\widehat{D}}_{Kl} - \overline{D}_{K,l}\right)^2, \tag{16}$$

$$\mathcal{S}_1(\widehat{\boldsymbol{L}}^*) = \min_{\widehat{\boldsymbol{D}} \in \mathcal{E}} \mathcal{S}_1(\widehat{\boldsymbol{D}}), \quad \mathcal{S}_1(\widehat{\boldsymbol{D}}) := \sum_{k,l \in [K]} \left(\widehat{D}_{kl} - D_{kl}\right)^2. \tag{17}$$

Indeed, given the previous reformulation of $\mathcal{S}$, we can establish the following lower bound via Proposition D.3. Notably, to simplify the problem, in Proposition D.3, we relax the EDM constraint by considering $\mathcal{E}_{\mathcal{L}}$, containing $\mathcal{E}$ by keeping only the PSD property from the EDM definition in Fact 1. We will reintroduce the missing constraints in $\mathcal{E}_{\mathcal{L}}$ and use the solution for the simplified problem to construct an upper bound in Appendix D.3.

Proposition D.3 (Error Lower Bound of $\mathcal{S}^*$)**.** *The lower bound of $\mathcal{S}$ is provided as follows:*

$$\mathcal{S}^* = \min_{\widehat{\boldsymbol{D}} \in \mathcal{E}} \left[\mathcal{S}_1(\widehat{\boldsymbol{D}}) + \mathcal{S}_2(\widehat{\boldsymbol{D}})\right] \geq \mathcal{S}_1(\widehat{\boldsymbol{L}}^*) + \mathcal{S}_2(\widehat{\boldsymbol{L}}^*) \geq \mathcal{L}_1 + \mathcal{L}_2 =: \mathcal{L}, \tag{18}$$

$$\mathcal{S}_1(\widehat{\boldsymbol{L}}^*) = \min_{\widehat{\boldsymbol{D}} \in \mathcal{E}_{\mathcal{L}}} \mathcal{S}_1(\widehat{\boldsymbol{D}}) \geq \sum_{k:\lambda_k < 0} \lambda_k^2 =: \mathcal{L}_1, \tag{19}$$

$$\mathcal{S}_2(\widehat{\boldsymbol{L}}^*) = \min_{\widehat{\boldsymbol{D}} \in \mathcal{E}_{\mathcal{L}}} \mathcal{S}_2(\widehat{\boldsymbol{D}}) = 0 =: \mathcal{L}_2. \tag{20}$$

Here $\mathcal{E}_{\mathcal{L}}$ contains $\mathcal{E}$ by keeping only the PSD property from the EDM definition in Fact 1.

Proof of Proposition D.3. Note that if $\mathcal{S}_1$ is minimized at $\widehat{\boldsymbol{L}}^*$, that is,

$$\mathcal{S}_1(\widehat{\boldsymbol{L}}^*) = \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}} \mathcal{S}_1(\widehat{\boldsymbol{D}}), \quad \mathcal{S}_1(\widehat{\boldsymbol{D}}) := \sum_{k,l\in[K]} \left(\widehat{D}_{kl} - D_{kl}\right)^2. \tag{21}$$

We then can find the lower bound of $\mathcal{S}^* = \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}} \left[\mathcal{S}_1(\widehat{\boldsymbol{D}}) + \mathcal{S}_2(\widehat{\boldsymbol{D}})\right]$ via the minimizer $\widehat{\boldsymbol{L}}^*$.

Using the definition of Frobenius norm and $\mathcal{E}_{\mathcal{L}}$, we can obtain:

$$\mathcal{S}_1(\widehat{\boldsymbol{L}}^*) := \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}} \mathcal{S}_1(\widehat{\boldsymbol{D}}) \geq \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \mathcal{S}_1(\widehat{\boldsymbol{D}}), \quad \mathcal{S}_1(\widehat{\boldsymbol{D}}) = \|\widehat{\boldsymbol{D}} - \boldsymbol{D}\|_F^2,$$

We then obtain the following decomposition:

$$\|\widehat{\boldsymbol{D}} - \boldsymbol{D}\|_F^2 = \|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2, \quad \boldsymbol{A} := \boldsymbol{C}\widehat{\boldsymbol{D}}\boldsymbol{C} - \boldsymbol{C}\boldsymbol{D}\boldsymbol{C},$$

$$\boldsymbol{B} := \frac{1}{K}\boldsymbol{O}\widehat{\boldsymbol{D}}\boldsymbol{C} + \frac{1}{K}\boldsymbol{C}\widehat{\boldsymbol{D}}\boldsymbol{O} + \frac{1}{K^2}\boldsymbol{O}\widehat{\boldsymbol{D}}\boldsymbol{O} - \left(\frac{1}{K}\boldsymbol{O}\boldsymbol{D}\boldsymbol{C} + \frac{1}{K}\boldsymbol{C}\boldsymbol{D}\boldsymbol{O} + \frac{1}{K^2}\boldsymbol{O}\boldsymbol{D}\boldsymbol{O}\right),$$

where $\boldsymbol{C} = \boldsymbol{I}_K - \frac{1}{K}\boldsymbol{O}$ is the centering matrix and $\boldsymbol{O} = \boldsymbol{1}_K\boldsymbol{1}_K^\top$ is the all-ones matrix. Indeed, using the definition of the centering matrix $\boldsymbol{C} = \boldsymbol{I}_K - \frac{1}{K}\boldsymbol{O}$, we have $\boldsymbol{I}_K = \boldsymbol{C} + \frac{1}{K}\boldsymbol{O}$.

$$\|\widehat{\boldsymbol{D}} - \boldsymbol{D}\|_F^2 = \|\boldsymbol{I}_K\widehat{\boldsymbol{D}}\boldsymbol{I}_K - \boldsymbol{I}_K\boldsymbol{D}\boldsymbol{I}_K\|_F^2 = \|\boldsymbol{A} + \boldsymbol{B}\|_F^2 = \|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2 + 2\,\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}) = \|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2,$$

Here we used the fact that the matrix product is invariant under cyclic permutation:

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{Tr}\left(\boldsymbol{C}(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{C}(\widehat{\boldsymbol{D}} - \boldsymbol{D})\frac{1}{K}\boldsymbol{O}\right) = \mathrm{Tr}\left(\frac{1}{K}\boldsymbol{O}\boldsymbol{C}(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{C}(\widehat{\boldsymbol{D}} - \boldsymbol{D})\right) = 0,$$

and

$$\frac{1}{K}\boldsymbol{O}\boldsymbol{C} = \frac{1}{K}\boldsymbol{O}\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{O}\right) = \frac{1}{K}\boldsymbol{O} - \frac{1}{K^2}\boldsymbol{O}\boldsymbol{O} = 0.$$

Under only the PSD constraint, the optimal solution $\widehat{\boldsymbol{L}}^*$ that minimizes $\mathcal{S}_1(\widehat{\boldsymbol{D}})$ can be decomposed as:

$$\widehat{\boldsymbol{L}}^* = \widehat{\boldsymbol{L}}^*_{\boldsymbol{A}} + \widehat{\boldsymbol{L}}^*_{\boldsymbol{B}},$$

where $\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}}$ and $\widehat{\boldsymbol{L}}^*_{\boldsymbol{B}}$ respectively minimize the terms $\|\boldsymbol{A}\|_F^2$ and $\|\boldsymbol{B}\|_F^2$ independently.

In particular, using the definition of the centering matrix $\boldsymbol{C} = \boldsymbol{I}_K - \frac{1}{K}\boldsymbol{O}$, the entries of $\widehat{\boldsymbol{L}}^*_{\boldsymbol{B}}$ are given by:

$$\widehat{\boldsymbol{L}}^*_{\boldsymbol{B},kl} := \left[\frac{1}{K}\boldsymbol{O}\boldsymbol{D}\boldsymbol{C} + \frac{1}{K}\boldsymbol{C}\boldsymbol{D}\boldsymbol{O} + \frac{1}{K^2}\boldsymbol{O}\boldsymbol{D}\boldsymbol{O}\right]_{kl}$$

$$= \left[\frac{1}{K}\boldsymbol{O}\boldsymbol{D} + \frac{1}{K}(\boldsymbol{O}\boldsymbol{D})^\top - \frac{1}{K^2}\boldsymbol{O}\boldsymbol{D}\boldsymbol{O}\right]_{kl} = \overline{\boldsymbol{D}}_k + \overline{\boldsymbol{D}}_l - \overline{\boldsymbol{D}},$$

where $\overline{\boldsymbol{D}}_k$ denotes the mean of the k-th row (or column) of $\boldsymbol{D}$, and $\overline{\boldsymbol{D}}$ is the global mean of all elements in $\boldsymbol{D}$. Therefore, the rows/columns mean of $\widehat{\boldsymbol{L}}^*_{\boldsymbol{B}}$ equal those of $\boldsymbol{D}$ itself, and hence

$$\widehat{\boldsymbol{L}}^*_{\boldsymbol{B}} = \arg\min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \|\boldsymbol{B}\|_F^2, \quad \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \|\boldsymbol{B}\|_F^2 = 0.$$

Therefore,

$$\min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \mathcal{S}_2(\widehat{\boldsymbol{D}}) = \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \sum_{l\in[K]} \left(\overline{\widehat{D}}_{Kl} - \overline{D}_{K,l}\right)^2 = 0.$$

Here we used the fact that the matrix $\boldsymbol{D}$ is given by $D_{kl} := \mathrm{FGW}_{p,\alpha}(\mathcal{G}(\mathbb{S}_k), \mathcal{G}(S_l))$ for all $k, l \in [K]$ and the empirical FGW barycenter is given by

$$\overline{\mathcal{G}}_K \in \arg\min_{\mathcal{G}\in\mathcal{P}_p(\boldsymbol{\Omega})} \frac{1}{K}\sum_{l=1}^K \mathrm{FGW}^p_{p,\alpha}(\mathcal{G}, \mathcal{G}(S_l)) = \arg\min_{\mathcal{G}\in\mathcal{P}_p(\boldsymbol{\Omega})} \frac{1}{K}\sum_{l=1}^K \mathrm{FGW}_{p,\alpha}(\mathcal{G}, \mathcal{G}(S_l)),$$

$$\overline{D}_{K,l} := \mathrm{FGW}_{p,\alpha}(\overline{\mathcal{G}}_K, \mathcal{G}(S_l)) = \min_{\mathcal{G}\in\mathcal{P}_p(\boldsymbol{\Omega})} \frac{1}{K}\sum_{l=1}^K \mathrm{FGW}_{p,\alpha}(\mathcal{G}, \mathcal{G}(S_l)) \text{ (=: column } l\text{-th means of } \boldsymbol{D}),$$

where $\mathcal{P}_p(\boldsymbol{\Omega})$ denotes the space of attributed graphs with finite p-th order FGW distance. To approximate this barycenter in embedding space, we require

$$\|\overline{\boldsymbol{e}}_K - \boldsymbol{e}_l\|_2^2 \approx \mathrm{FGW}_{p,\alpha}(\overline{\mathcal{G}}_K, \mathcal{G}(S_l)) := \overline{D}_{K,l} \text{ for all } l \in [K],$$

where $\overline{\boldsymbol{e}}_K = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{e}_k$ is the mean embedding and $\boldsymbol{e}_k := \mathcal{T}_\theta(\mathbf{H}_k)$ is the learned representation.

Now we would like to find a local analytic solution $\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^*$ minimizing $\|\boldsymbol{A}\|_F^2$ such that the global solution $\widehat{\boldsymbol{L}}^* = \widehat{\boldsymbol{L}}_{\boldsymbol{A}}^* + \widehat{\boldsymbol{L}}_{\boldsymbol{B}}^*$ minimizes both terms $\|\boldsymbol{A}\|_F^2$ and $\|\boldsymbol{B}\|_F^2$ simultaneously. That is,

$$\begin{aligned}\min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \|\boldsymbol{A}\|_F^2 &= \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \|\boldsymbol{C}(\widehat{\boldsymbol{L}}_{\boldsymbol{A}} + \widehat{\boldsymbol{L}}_{\boldsymbol{B}})\boldsymbol{C} - \boldsymbol{CDC}\|_F^2 \\ &= \|\boldsymbol{C}(\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^* + \widehat{\boldsymbol{L}}_{\boldsymbol{B}}^*)\boldsymbol{C} - \boldsymbol{CDC}\|_F^2 = \|\boldsymbol{C}\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^*\boldsymbol{C} - \boldsymbol{CDC}\|_F^2.\end{aligned}$$

Here we used the fact that by definition of $\widehat{\boldsymbol{L}}_{\boldsymbol{B}}^*$, it holds that $\boldsymbol{C}\widehat{\boldsymbol{L}}_{\boldsymbol{B}}^*\boldsymbol{C} = 0$. Hence, the optimization becomes:

$$\min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \|\boldsymbol{C}\widehat{\boldsymbol{L}}_{\boldsymbol{A}}\boldsymbol{C} - \boldsymbol{CDC}\|_F^2.$$

This is in fact the problem of computing the nearest PSD approximation $\boldsymbol{C}\widehat{\boldsymbol{L}}_{\boldsymbol{A}}\boldsymbol{C}$ to a symmetric matrix $\boldsymbol{CDC}$. Using the result from (Higham, 1988), we find the analytic solution as follows:

$$\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^* = -\sum_{k:\lambda_k>0} \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k^\top. \tag{22}$$

Here $\{\lambda_k, \boldsymbol{v}_k\}_{k\in[K]}$ are the eigenvalues and eigenvectors of $\boldsymbol{F} = -\boldsymbol{CDC}$. Because $\boldsymbol{CDC}$ has rows/columns means 0, the ones vector $\mathbf{1}_K$ is an eigenvector of $\boldsymbol{CDC}$ with eigenvalue 0. This leads to $\mathbf{1}_K$ is also in the null space $\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^*$ and:

$$\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^* = \boldsymbol{C}\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^*\boldsymbol{C}, \quad \frac{1}{K}\boldsymbol{O}\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^* = \frac{1}{K}\left(\boldsymbol{O}\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^*\right)^\top = 0.$$

Therefore,

$$\|\widehat{\boldsymbol{L}}^* - \boldsymbol{D}\|_F^2 = \|\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^* + \widehat{\boldsymbol{L}}_{\boldsymbol{B}}^* - \boldsymbol{D}\|_F^2 = \sum_{k:\lambda_k<0} \lambda_k^2.$$

Combining all together, Proposition D.3 is derived as follows:

$$\mathcal{S}^* \geq \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \|\boldsymbol{A}\|_F^2 + \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \|\boldsymbol{B}\|_F^2 + \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \mathcal{S}_2(\widehat{\boldsymbol{D}}) = \sum_{k:\lambda_k<0} \lambda_k^2 + 0 + 0 = \sum_{k:\lambda_k<0} \lambda_k^2 =: \mathcal{L}.$$

□

D.3. Upper Bounds on Embedding of Pairwise Empirical FGW Barycenter Distances

As discussed in Appendix D.2, the lower bound stated in Proposition D.3 is derived by simplifying the problem and relaxing the EDM constraint. Specifically, this relaxation involves considering the set $\mathcal{E}_{\mathcal{L}}$, which contains $\mathcal{E}$ but retains only the PSD requirement from the EDM characterization given in Fact 1. In Proposition D.4, we reintroduce the missing constraints excluded in $\mathcal{E}_{\mathcal{L}}$ and leverage the closed-form solution obtained from the relaxed problem to construct an upper bound under the original EDM constraint set $\mathcal{E}$.

Proposition D.4 (Error Upper Bound of $\mathcal{S}^*$)**.** *There exists a matrix $\widehat{\boldsymbol{U}}^* \in \mathcal{E}$ such that the following upper bounds hold:*

$$\mathcal{S}^* = \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}} \left[\mathcal{S}_1(\widehat{\boldsymbol{D}}) + \mathcal{S}_2(\widehat{\boldsymbol{D}})\right] \leq \mathcal{S}_1(\widehat{\boldsymbol{U}}^*) + \mathcal{S}_2(\widehat{\boldsymbol{U}}^*) \leq \mathcal{U}_1 + \mathcal{U}_2 =: \mathcal{U}, \tag{23}$$

$$\mathcal{S}_1(\widehat{\boldsymbol{U}}^*) = \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}} \mathcal{S}_1(\widehat{\boldsymbol{D}}) \leq \mathcal{U}_1 := \sum_{k:\lambda_k<0} \lambda_k^2 + \sum_{kl} (\Delta p_k + \Delta p_l)^2,$$

$$\Delta p_k = \frac{1}{2}\sum_{l:\lambda_l<0} \lambda_l \cdot \boldsymbol{v}_{kl}^2, \ \forall k \in [K] \tag{24}$$

$$\mathcal{S}_2(\widehat{\boldsymbol{U}}^*) = \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}} \mathcal{S}_2(\widehat{\boldsymbol{D}}) \leq \sum_{l} (\Delta \overline{\boldsymbol{p}}_l)^2 =: \mathcal{U}_2, \tag{25}$$

where the aggregated error term is defined as:

$$\Delta \overline{\boldsymbol{p}}_l := \frac{1}{2K}\sum_{k=1}^{K}\sum_{l:\lambda_l<0} \lambda_l \cdot \boldsymbol{v}_{kl}^2.$$

We aim to exploit the information derived from the truncation of the negative eigenspace of the matrix $\boldsymbol{CDC}$, specifically the matrix introduced in Equation (22), defined as:

$$\widehat{\boldsymbol{L}}_{\boldsymbol{A}}^* = -\sum_{k:\lambda_k>0} \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k^\top,$$

where $\{\lambda_k, \boldsymbol{v}_k\}_{k\in[K]}$ denote the eigenvalues and corresponding eigenvectors of the matrix $\boldsymbol{F} = -\boldsymbol{CDC}$.

Recall that the entries of $\widehat{\boldsymbol{L}}^*_{\boldsymbol{B}}$ are given by:

$$\widehat{\boldsymbol{L}}^*_{\boldsymbol{B},kl} = \left[\frac{1}{K}\boldsymbol{OD} + \frac{1}{K}(\boldsymbol{OD})^\top - \frac{1}{K^2}\boldsymbol{ODO}\right]_{kl} = \overline{\boldsymbol{D}}_k + \overline{\boldsymbol{D}}_l - \overline{\boldsymbol{D}}.$$

As a consequence, the sum $\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}} + \widehat{\boldsymbol{L}}^*_{\boldsymbol{B}}$ may not be strictly hollow or PSD when $\boldsymbol{D}$ is not an EDM. To address this, we seek to construct a symmetric matrix $\boldsymbol{P}$ to be added to $\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}}$, resulting in the matrix $\widehat{\boldsymbol{U}}^* := \widehat{\boldsymbol{L}}^*_{\boldsymbol{A}} + \boldsymbol{P}$, which is both hollow and PSD. This adjustment is designed to avoid any additional penalty on the term $\|\boldsymbol{A}\|_F^2$, though it may introduce some approximation errors in $\|\boldsymbol{B}\|_F^2$ and in the quantity $\mathcal{S}_2$. These induced errors contribute to the upper bound $\mathcal{U}$ for the optimal score $\mathcal{S}^*$.

We begin with the requirement that the matrix $\boldsymbol{P}$ does not contribute any additional penalty to the term $\|\boldsymbol{A}\|_F^2$. This can be ensured by imposing the constraint $\boldsymbol{CPC} = 0$. Under this condition, the matrix $\widehat{\boldsymbol{U}}^*$ remains a minimizer of $\|\boldsymbol{A}\|_F^2$, as demonstrated below:

$$\begin{aligned}
\min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \|\boldsymbol{A}\|_F^2 &= \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}_{\mathcal{L}}} \|\boldsymbol{C}(\widehat{\boldsymbol{L}}_{\boldsymbol{A}} + \widehat{\boldsymbol{L}}_{\boldsymbol{B}})\boldsymbol{C} - \boldsymbol{CDC}\|_F^2 \\
&= \|\boldsymbol{C}(\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}} + \boldsymbol{P} + \widehat{\boldsymbol{L}}^*_{\boldsymbol{B}})\boldsymbol{C} - \boldsymbol{CDC}\|_F^2 \\
&= \|\boldsymbol{C}\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}}\boldsymbol{C} - \boldsymbol{CDC}\|_F^2,
\end{aligned}$$

where the final equality holds due to the constraint $\boldsymbol{CPC} = 0$.

This leads to the condition $(\boldsymbol{CP})\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{PC}) = 0$, implying that $\boldsymbol{CP}$ lies in the left null space of $\boldsymbol{C}$, and $\boldsymbol{PC}$ lies in its right null space. As a result, all rows of $\boldsymbol{PC}$ must be constant, and this expression can be written as:

$$\mathbf{1}_K\boldsymbol{c}^\top = \boldsymbol{PC} = \boldsymbol{P}\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{O}\right) \text{ or } \boldsymbol{P} = \mathbf{1}_K\boldsymbol{c}^\top + \boldsymbol{P}\frac{1}{K}\boldsymbol{O},$$

where $\boldsymbol{c}$ is a column vector to be defined subsequently. Here, we have used the fact that $\boldsymbol{C}$ is the centering matrix defined by $\boldsymbol{C} = \boldsymbol{I}_K - \frac{1}{K}\boldsymbol{O}$.

Multiplying both sides on the left by $\frac{1}{K}\boldsymbol{O}$ yields:

$$\frac{1}{K}\boldsymbol{OP} = \frac{1}{K}\boldsymbol{O}\mathbf{1}_K\boldsymbol{c}^\top + \frac{1}{K}\boldsymbol{O}\left(\frac{1}{K}\boldsymbol{PO}\right) = \mathbf{1}_K\boldsymbol{c}^\top + \frac{1}{K^2}\boldsymbol{OPO}.$$

This leads to

$$\boldsymbol{c}^\top = \frac{1}{K}\mathbf{1}_K^\top\boldsymbol{P} - \frac{1}{K^2}\mathbf{1}_K^\top\boldsymbol{OPO}.$$

Indeed, via the definition of $\boldsymbol{O} = \mathbf{1}_K\mathbf{1}_K^\top$, we can verify this as follows:

$$\begin{aligned}
\mathbf{1}_K\boldsymbol{c}^\top + \frac{1}{K^2}\boldsymbol{OPO} &= \mathbf{1}_K\left(\frac{1}{K}\mathbf{1}_K^\top\boldsymbol{P} - \frac{1}{K^2}\mathbf{1}_K^\top\boldsymbol{OPO}\right) + \frac{1}{K^2}\boldsymbol{OPO} \\
&= \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\boldsymbol{P} - \frac{1}{K^2}\mathbf{1}_K\mathbf{1}_K^\top\boldsymbol{OPO} + \frac{1}{K^2}\boldsymbol{OPO} \\
&= \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\boldsymbol{P} - \frac{1}{K^2}\boldsymbol{OOPO} + \frac{1}{K^2}\boldsymbol{OPO} \\
&= \frac{1}{K}\boldsymbol{OP}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\boldsymbol{P} &= \mathbf{1}_K\left(\frac{1}{K}\mathbf{1}_K^\top\boldsymbol{P} - \frac{1}{K^2}\mathbf{1}_K^\top\boldsymbol{OPO}\right) + \boldsymbol{P}\frac{1}{K}\boldsymbol{O} \\
&= \frac{1}{K}\mathbf{1}_K(\mathbf{1}_K^\top\boldsymbol{P}) + \frac{1}{K}(\boldsymbol{P}\mathbf{1}_K)\mathbf{1}_K^\top - \frac{1}{K^2}\mathbf{1}_K\mathbf{1}_K^\top\boldsymbol{OPO}
\end{aligned}$$

Since $\boldsymbol{P}\mathbf{1}_K$ is a column vector, to satisfy this constraint, $\boldsymbol{P}$ must be of the form:

$$\boldsymbol{P} = \mathbf{1}_K\frac{\boldsymbol{p}^\top}{K} + \frac{\boldsymbol{p}}{K}\mathbf{1}_K^\top - \widehat{\boldsymbol{p}}\frac{\mathbf{1}_K\mathbf{1}_K^\top}{K},$$

where $\boldsymbol{p} \in \mathbb{R}^K$ is a vector of free parameters, and $\widehat{\boldsymbol{p}}$ denotes its average. This construction implies that $\boldsymbol{P}$ has only K degrees of freedom. However, to ensure that $\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}} + \boldsymbol{P}$ has zero diagonal (i.e., the resulting matrix is hollow), the diagonal

entries of $\boldsymbol{P}$ must satisfy the following K linear constraints:

$$\boldsymbol{p}_k - \frac{1}{2}\widehat{\boldsymbol{p}} = -\frac{1}{2}[\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}}]_{kk}, \quad \forall k \in [K].$$

Solving this linear system yields:

$$\boldsymbol{p}_k = \frac{1}{2}\left(\sum_{l:\lambda_l>0} \lambda_l \cdot \boldsymbol{v}_{kl}^2 + \frac{1}{K}\widehat{\boldsymbol{p}}\right),$$

$$\widehat{\boldsymbol{p}} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{p}_k = \frac{1}{K}\sum_{k=1}^{K}\sum_{l:\lambda_l>0} \lambda_l \cdot \boldsymbol{v}_{kl}^2,$$

where we have used the fact that $\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}} = -\sum_{l:\lambda_l>0} \lambda_l \boldsymbol{v}_l \boldsymbol{v}_l^\top$, and hence its diagonal entries are given by $[\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}}]_{kk} = -\sum_{l:\lambda_l>0} \lambda_l \cdot \boldsymbol{v}_{kl}^2$.

Consequently, the resulting matrix $\boldsymbol{P}$ can be expressed element-wise as:

$$\boldsymbol{P}_{k,l} = -\frac{[\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}}]_{kk} + [\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}}]_{ll}}{2} \geq 0,$$

where the inequality follows from the fact that $\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}}$ is negative semi-definite.

In summary, the matrix $\widehat{\boldsymbol{U}}^* := \widehat{\boldsymbol{L}}^*_{\boldsymbol{A}} + \boldsymbol{P}$ satisfies all three constraints specified in Definition D.1.

Although $\widehat{\boldsymbol{U}}^*$ preserves the value of $\|\boldsymbol{A}\|_F^2$, it differs from $\widehat{\boldsymbol{L}}^*_{\boldsymbol{A}}$ and introduces approximation errors in the $\|\boldsymbol{B}\|_F^2$ term and the $\mathcal{S}_2$ term. Note that the sum of the untruncated version of $\boldsymbol{CDC}$ and the matrix

$$\frac{1}{K}\boldsymbol{ODC} + \frac{1}{K}\boldsymbol{CDO} + \frac{1}{K^2}\boldsymbol{ODO}$$

is equal to $\boldsymbol{D}$ and remains hollow. Recall the decomposition:

$$\|\widehat{\boldsymbol{D}} - \boldsymbol{D}\|_F^2 = \|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2, \quad \boldsymbol{A} := \boldsymbol{C}\widehat{\boldsymbol{D}}\boldsymbol{C} - \boldsymbol{CDC},$$

$$\boldsymbol{B} := \frac{1}{K}\boldsymbol{O}\widehat{\boldsymbol{D}}\boldsymbol{C} + \frac{1}{K}\boldsymbol{C}\widehat{\boldsymbol{D}}\boldsymbol{O} + \frac{1}{K^2}\boldsymbol{O}\widehat{\boldsymbol{D}}\boldsymbol{O} - \left(\frac{1}{K}\boldsymbol{ODC} + \frac{1}{K}\boldsymbol{CDO} + \frac{1}{K^2}\boldsymbol{ODO}\right),$$

where $\boldsymbol{C} = \boldsymbol{I}_K - \frac{1}{K}\boldsymbol{O}$ is the centering matrix and $\boldsymbol{O} = \boldsymbol{1}_K\boldsymbol{1}_K^\top$ is the all-ones matrix.

The matrix

$$\frac{1}{K}\boldsymbol{ODC} + \frac{1}{K}\boldsymbol{CDO} + \frac{1}{K^2}\boldsymbol{ODO}$$

can be written similarly to $\boldsymbol{P}$ by including the contributions from the negative eigenvalues, resulting in the matrix $\widetilde{\boldsymbol{P}}$, parameterized by:

$$\widetilde{\boldsymbol{p}}_k = \frac{1}{2}\left(\sum_{l} \lambda_l \cdot \boldsymbol{v}_{kl}^2 + \frac{1}{K}\widetilde{\widehat{\boldsymbol{p}}}\right),$$

$$\widetilde{\widehat{\boldsymbol{p}}} = \frac{1}{K}\sum_{k=1}^{K} \widetilde{\boldsymbol{p}}_k = \frac{1}{K}\sum_{k=1}^{K}\sum_{l} \lambda_l \cdot \boldsymbol{v}_{kl}^2.$$

Define the correction due to negative eigenvalues as:

$$\Delta\boldsymbol{p}_k := \frac{1}{2}\sum_{l:\lambda_l<0} \lambda_l \cdot \boldsymbol{v}_{kl}^2, \quad \forall k \in [K].$$

The resulting error in the $\|\boldsymbol{B}\|_F^2$ term is given by:

$$\|\boldsymbol{B}\|_F^2 = \|\widetilde{\boldsymbol{P}} - \boldsymbol{P}\|_F^2 = \sum_{k,l} \left(\Delta\boldsymbol{p}_k + \Delta\boldsymbol{p}_l\right)^2.$$

Furthermore, the contribution to $\mathcal{S}_2$ is bounded as:

$$\mathcal{S}_2 = \min_{\widehat{\boldsymbol{D}}\in\mathcal{E}} \mathcal{S}_2(\widehat{\boldsymbol{D}}) = \sum_{l\in[K]} \left(\overline{\widehat{\boldsymbol{D}}}_{K,l} - \overline{\boldsymbol{D}}_{K,l}\right)^2 \leq \sum_{l} \left(\Delta\overline{\boldsymbol{p}}_l\right)^2 =: \mathcal{U}_2,$$

where the aggregated error term is defined as:

$$\Delta\overline{\boldsymbol{p}}_l := \frac{1}{2K}\sum_{k=1}^{K}\sum_{l:\lambda_l<0}\lambda_l\cdot\boldsymbol{v}_{kl}^2.$$

D.4. Proof of Lemma D.2

The proof is proved via leveraging Proposition 8.2 from (Peyré et al., 2019), applied to the specific case $\alpha = 0$, and relies on the relationships among FGW, Wasserstein (W), and Gromov-Wasserstein (GW) distances.

The FGW cost $\mathrm{FGW}_{p,\alpha}(\mathcal{G}_1, \mathcal{G}_2)$ is defined via two components: the node feature cost matrix $\boldsymbol{M}[i,j] = d_f(\boldsymbol{H}_1[i], \boldsymbol{H}_2[j])^p$, and the structural discrepancy tensor $\mathbf{L}(\boldsymbol{A}_1, \boldsymbol{A}_2)[i,j,l,m] = |\boldsymbol{A}_1[i,j] - \boldsymbol{A}_2[l,m]|^p$.

Let $\mathcal{G}_1 = (\boldsymbol{H}_1, \boldsymbol{A}_1, \boldsymbol{\omega}_1)$ and $\mathcal{G}_2 = (\boldsymbol{H}_2, \boldsymbol{A}_2, \boldsymbol{\omega}_2)$ be two attributed graphs with N_1 and N_2 nodes, respectively. Their associated probability measures are

$$\mu_1 = \sum_k \omega_{1k}\delta_{(\boldsymbol{x}_{1k}, \boldsymbol{a}_{1k})}, \quad \mu_2 = \sum_l \omega_{2l}\delta_{(\boldsymbol{x}_{2l}, \boldsymbol{a}_{2l})}.$$

We define the marginals $\mu_{\boldsymbol{H}_1} = \sum_k \omega_k \delta_{\boldsymbol{x}_k}$ and $\mu_{\boldsymbol{A}_1} = \sum_k \omega_k \delta_{\boldsymbol{a}_k}$ (and analogously for $\mu_{\boldsymbol{H}_2}$ and $\mu_{\boldsymbol{A}_2}$) as projections of μ_1 and μ_2 onto the feature and structural spaces, respectively.

Using these definitions, we introduce the following notation:

$$J_p(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{\pi}) = \sum_{ijkl} L_{ijkl}(\boldsymbol{A}_1, \boldsymbol{A}_2)^p \boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{kl}, \tag{26}$$

$$\mathrm{GW}_p(\mu_{\boldsymbol{H}_1}, \mu_{\boldsymbol{H}_2})^p = \min_{\boldsymbol{\pi}\in\boldsymbol{\Pi}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)} J_p(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{\pi}), \tag{27}$$

$$H_p(\boldsymbol{M}, \boldsymbol{\pi}) = \sum_{kl} d_f(\boldsymbol{x}_{1k}, \boldsymbol{x}_{2l})^p \boldsymbol{\pi}_{kl}, \tag{28}$$

$$\mathrm{W}_p(\mu_{\boldsymbol{A}_1}, \mu_{\boldsymbol{A}_2})^p = \min_{\boldsymbol{\pi}\in\boldsymbol{\Pi}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)} H_p(\boldsymbol{M}, \boldsymbol{\pi}). \tag{29}$$

Let $\boldsymbol{\pi} \in \boldsymbol{\Pi}(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ be any admissible coupling. If both μ_1 and μ_2 are defined over a common metric space $(\boldsymbol{\Omega}, \boldsymbol{A}, \mu)$, then the FGW distance is given by:

$$\mathrm{FGW}_{p,\alpha}(\mathcal{G}_1, \mathcal{G}_2) := \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)} \langle (1-\alpha)\boldsymbol{M} + \alpha\,\mathbf{L}(\boldsymbol{A}_1, \boldsymbol{A}_2) \otimes \boldsymbol{\pi}, \boldsymbol{\pi}\rangle\,. \tag{30}$$

We now derive the following fundamental identity:

$$\begin{aligned}\mathbb{E}_{p,\alpha}\left(\boldsymbol{M}, \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{\pi}\right) &:= \sum_{ijkl}\left[(1-\alpha)d_f(\boldsymbol{x}_{1k}, \boldsymbol{x}_{2l})^p + \alpha\,|\boldsymbol{A}_1(i,k) - \boldsymbol{A}_2(j,l)|^p\right]\boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{kl} \\ &= (1-\alpha)H_p(\boldsymbol{M}, \boldsymbol{\pi}) + \alpha J_p(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{\pi}).\end{aligned} \tag{31}$$

Moreover, let $\boldsymbol{\pi}_\alpha$ denote the optimal coupling that minimizes the FGW objective $\mathbb{E}_{p,\alpha}\left(\boldsymbol{M}, \boldsymbol{A}_1, \boldsymbol{A}_2, \cdot\right)$. Then the FGW distance admits the following decomposition:

$$\begin{aligned}\mathrm{FGW}_{p,\alpha}^p(\mu_1, \mu_2) &= \min_{\boldsymbol{\pi}\in\boldsymbol{\Pi}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)} \mathbb{E}_{p,\alpha}\left(\boldsymbol{M}, \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{\pi}\right) = \mathbb{E}_{p,\alpha}\left(\boldsymbol{M}, \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{\pi}_\alpha\right) \\ &= (1-\alpha)H_p(\boldsymbol{M}, \boldsymbol{\pi}_\alpha) + \alpha J_p(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{\pi}_\alpha) \\ &\geq (1-\alpha)\mathrm{W}_p^p(\mu_{\boldsymbol{A}_1}, \mu_{\boldsymbol{A}_2}) + \alpha\,\mathrm{GW}_p^p(\mu_{\boldsymbol{H}_1}, \mu_{\boldsymbol{H}_2}).\end{aligned} \tag{32}$$

This inequality follows from the optimality of the W and GW distances with respect to the cost functions H_p and J_p, respectively, and highlights the interpolation nature of the FGW distance between these two metrics as governed by the parameter α.

The generalized FGW cost $\mathbb{E}_{p,\alpha}\left(\boldsymbol{M}, \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{\pi}\right)$ admits the following explicit formulation:

$$\begin{aligned}\mathbb{E}_{p,\alpha}\left(\boldsymbol{M}, \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{\pi}\right) &= \langle (1-\alpha)\,\boldsymbol{M}^p + \alpha\,\mathbf{L}(\boldsymbol{A}_1, \boldsymbol{A}_2)^p \otimes \boldsymbol{\pi}, \boldsymbol{\pi}\rangle \\ &= \sum_{i,j,k,l}\left[(1-\alpha)d_f(\boldsymbol{x}_{1k}, \boldsymbol{x}_{2l})^p + \alpha\,|\boldsymbol{A}_1(i,k) - \boldsymbol{A}_2(j,l)|^p\right]\boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{kl}.\end{aligned}$$

Based on the formulation above, we obtain the following upper bound on the FGW distance:

$$\begin{aligned}\mathrm{FGW}_{p,\alpha}(G_1, G_2) &\leq \langle (1-\alpha)\, \boldsymbol{M} + \alpha\, \mathbf{L}(\boldsymbol{A}_1, \boldsymbol{A}_2) \otimes \boldsymbol{\pi}, \boldsymbol{\pi} \rangle \\ &\leq \sum_{k,l} \left[(1-\alpha)\, d_f(\boldsymbol{x}_{1k}, \boldsymbol{x}_{2l}) + 2^{p-1} \alpha\, \boldsymbol{A}[k,l] \right]^p \boldsymbol{\pi}_{kl}, \end{aligned} \tag{33}$$

where the second inequality follows from the convexity of the function $x \mapsto x^p$ for $p \geq 1$ and an application of Minkowski-type bounds on the structural term. Importantly, inequality in equation (33) holds for any admissible coupling $\boldsymbol{\pi} \in \boldsymbol{\Pi}(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$, and in particular, it remains valid when $\boldsymbol{\pi} = \overline{\boldsymbol{\pi}}$, the optimal coupling associated with the Wasserstein distance $\mathrm{W}_p(\mu_1, \mu_2)$ over the product metric space $(\boldsymbol{\Omega}, \overline{d})$. Here, the effective distance $\overline{d}$ between structured nodes $(\boldsymbol{x}_1, \boldsymbol{a}_1)$ and $(\boldsymbol{x}_2, \boldsymbol{a}_2)$ is defined as:

$$\overline{d}((\boldsymbol{x}_1, \boldsymbol{a}_1), (\boldsymbol{x}_2, \boldsymbol{a}_2)) = (1-\alpha)\, d_f(\boldsymbol{x}_1, \boldsymbol{x}_2) + 2^{p-1} \alpha\, \boldsymbol{A}(\boldsymbol{a}_1, \boldsymbol{a}_2).$$

Combining this with the Wasserstein formulation in equation (29), we observe the following inequality:

$$\mathrm{FGW}_{p,\alpha}(\mathcal{G}_1, \mathcal{G}_2) \leq \mathrm{W}_p(\mu_{\boldsymbol{A}_1}, \mu_{\boldsymbol{A}_2}), \quad \text{and} \quad \mathrm{FGW}_{p,\alpha}(\mathcal{G}_1, \mathcal{G}_2) = \mathrm{W}_p(\mu_{\boldsymbol{A}_1}, \mu_{\boldsymbol{A}_2}) \text{ when } \alpha = 0. \tag{34}$$