3D Interaction Geometric Pre-training for Molecular Relational Learning

Namkyeong Lee[1], **Yunhak Oh**[1], **Heewoong Noh**[1],
Gyoung S. Na[2], **Tianfan Fu**[3], **Chanyoung Park**[1]
[1] KAIST [2] KRICT [3] RPI
{namkyeong96,yunhak.oh,heewoongnoh,cy.park}@kaist.ac.kr
ngs0@krict.re.kr, futianfan@gmail.com

Abstract

Molecular Relational Learning (MRL) is a rapidly growing field that focuses on understanding the interaction dynamics between molecules, which is crucial for applications ranging from catalyst engineering to drug discovery. Despite recent progress, earlier MRL approaches are limited to using only the 2D topological structure of molecules, as obtaining the 3D interaction geometry remains prohibitively expensive. This paper introduces a novel 3D geometric pre-training strategy for MRL (3DMRL) that incorporates a 3D virtual interaction environment, overcoming the limitations of costly traditional quantum mechanical calculation methods. With the constructed 3D virtual interaction environment, 3DMRL trains 2D MRL model to learn the overall 3D geometric information of molecular interaction through contrastive learning. Moreover, fine-grained interaction between molecules is learned through force prediction loss, which is crucial in understanding the wide range of molecular interaction processes. Extensive experiments on various tasks using real-world datasets, including out-of-distribution and extrapolation scenarios, demonstrate the effectiveness of 3DMRL, showing up to a 24.93% improvement in performance across 40 tasks.

1 Introduction

Molecular relational learning (MRL) focuses on understanding the interaction dynamics between molecules and has gained significant attention from researchers thanks to its diverse applications (Lee et al., 2023a). For instance, understanding how a medication dissolves in different solvents (medication-solvent interaction) is vital in pharmacy (Pathak et al., 2020; Lu et al., 2024; Chen et al., 2024), while predicting the optical and photophysical properties of chromophores in various solvents (chromophore-solvent interaction) is essential for material discovery (Joung et al., 2021). Because of the expensive time and financial costs associated with conducting wet lab experiments to test the interaction behavior of all possible molecular pairs (Preuer et al., 2018), machine learning methods have been quickly embraced for MRL.

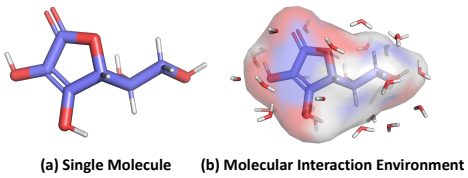

Figure 1: 3D geometry of (a) an individual molecule and (b) the molecular interaction environment.

Despite recent advancements in MRL, previous works tend to ignore molecules' 3D geometric information and instead focus solely on their 2D topological structures. However, in molecular science, the 3D geometric information of molecules (Figure 1 (a)) is crucial for understanding and predicting molecular behavior across various contexts, ranging from physical properties (Atkins et al., 2023) to biological functions (Fu et al., 2024; Zhang et al., 2021). This is particularly important in MRL,

AI for New Drug Modalities at NeurIPS 2024.

as geometric information plays a key role in molecular interactions by determining how molecules recognize, interact, and bind with one another in their interaction environment (Silverman & Holladay, 2014). This fact has been widely accepted in traditional molecular dyn In traditional molecular dynamics simulations, explicit solvent models, which directly consider the detailed environment of molecular interaction, have demonstrated superior performance compared to implicit solvent models, which simplify the solvent as a continuous medium, highlighting the significance of explicitly modeling the complex geometries of interaction environments (Zhang et al., 2017a).

However, acquiring stereochemical structures of molecules is often very costly, resulting in limited availability of such 3D geometric information for downstream tasks (Liu et al., 2021). Consequently, in the domain of molecular property prediction (MPP), there has been substantial progress in injecting 3D geometric information to 2D molecular graph encoders during the pre-training phase, while utilizing only the 2D molecular graph encoder for downstream tasks (Stärk et al., 2022; Liu et al., 2023). In contrast, compared to the MPP, pre-training and fine-tuning strategies for MRL have been surprisingly underexplored, primarily due to the following two distinct challenges associated with modeling complex molecular interaction environments.

Firstly, interactions between molecules occur through complex geometry as they are chaotically distributed in space as shown in Figure 1 (b). Therefore, it is essential to consider not only each molecule's independent geometry but also their relative positions and orientations in space. This requirement further complicates the acquisition of geometric information, making it more challenging to obtain detailed 3D geometry of molecular interaction environments. Consequently, it is essential to model an interaction environment that can simulate molecular interactions based solely on the 3D geometry of the individual molecules.

Secondly, in addition to the complexity of the interaction environment, the forces between molecules during interactions are vital in MRL, as they are key to understanding a wide range of physical, chemical, and biological processes. For instance, in solute-solvent interactions, polar solutes dissolve in polar solvents because of dipole-dipole interactions or hydrogen bonding. These forces allow solute molecules to form favorable interactions with solvent molecules, promoting solvation and enhancing solubility (Atkins et al., 2023). Thus, it is essential to develop pre-training strategies that effectively capture the forces between molecules within their interaction geometry.

To address these challenges, we introduce a novel 3D geometric pre-training strategy that is applicable to various MRL models by incorporating the 3D geometry of the interaction environment for molecules (3DMRL). Specifically, instead of relying on costly traditional quantum mechanical calculation methods to obtain interaction environments, we first propose a virtual interaction environment involving multiple molecules designed to simulate real molecular interactions. Then, during the pre-training stage, a 2D MRL model is trained to produce representations that are aligned with those of the 3D virtual interaction environment via contrastive learning. Additionally, the 2D MRL model is trained to predict the forces between molecules within this virtual interaction environment, allowing the model to effectively learn fine-grained atom-level interactions between molecules. These two pre-training strategies enable the 2D MRL model to be pre-trained to understand the nature of molecular interactions, facilitating positive transfer to a wide range of downstream MRL tasks. In this paper, we make the following contributions:

- Rather than relying on costly traditional quantum mechanical calculation methods to obtain interaction geometry, we propose a virtual interaction geometry made up of multiple molecules to mimic the molecular interaction environment observed in real-world conditions (Section 4.1).

- We propose pre-training strategies that enable the 2D MRL model to learn representations aligned with the 3D virtual interaction environment and capture the intermolecular forces between interacting molecules within the environment (Section 4.2).

- We conduct extensive experiments across various MRL models pre-trained with 3DMRL on a range of MRL tasks, including *out-of-distribution* and *extrapolation* scenarios. These experiments demonstrate improvements of up to 24.93% compared to MRL methods trained from scratch, underscoring the versatility of 3DMRL (Section 5).

To the best of our knowledge, this is the first paper proposing pre-training strategies specifically designed for molecular relational learning.

2 Related Works

2.1 Molecular Relational Learning (MRL)

Molecular Relational Learning (MRL) focuses on understanding the interaction dynamics between paired molecules. Delfos (Lim & Jung, 2019) employs recurrent neural networks combined with attention mechanisms to predict solvation-free energy, a key factor influencing the solubility of chemical substances, using SMILES string as input. Similarly, CIGIN (Pathak et al., 2020) utilizes message-passing neural networks (Gilmer et al., 2017) along with a cross-attention mechanism to capture atomic representations for solvation-free energy prediction. In a different context, Joung et al. (2021) use graph convolutional networks (Kipf & Welling, 2016) to generate representations of chromophores and solvents, which are then used to predict various optical and photophysical properties of chromophores, essential for developing new materials with vibrant colors. Meanwhile, MHCADDI (Deac et al., 2019) introduces a co-attentive message passing network (Veličković et al., 2017) designed for predicting drug-drug interactions (DDI), which aggregates information from all atoms within a pair of molecules, not just within individual molecules. Recently, CGIB (Lee et al., 2023a) and CMRL (Lee et al., 2023b) have introduced a comprehensive framework for MRL tasks, such as predicting solvation-free energy, chromophore-solute interactions, and drug-drug interactions. These models achieve this by identifying core functional groups involved in molecular interactions using information bottleneck and causal theory, respectively. However, prior studies have largely ignored molecules' 3D geometric information despite its well-established importance in comprehending various molecular properties.

2.2 3D Pre-training for Molecular Property Prediction (MPP)

Recently, the molecular science community has shown increasing interest in pre-training machine learning models with unlabeled data, primarily due to the scarcity of labeled data for downstream tasks (Lee et al., 2023b; Velez-Arce et al., 2024; Xu et al., 2024). A promising approach in this area leverages molecules' inherent nature, which can be effectively represented as both 2D topological graphs and 3D geometric graphs. For instance, 3D Infomax (Stärk et al., 2022) aims to enhance mutual information between 2D and 3D molecular representations using contrastive learning. GraphMVP (Liu et al., 2021) extends this concept by introducing a generative pre-training framework alongside contrastive learning. More recently, Noisy Nodes (Zaidi et al., 2022) and MoleculeSDE (Liu et al., 2023) have introduced methods to learn the 3D geometric distribution of molecules using a denoising framework, thereby uncovering the connection between the score function and the force field of molecules. Although the 3D structure of molecules has been effectively leveraged in pre-training for predicting single molecular properties, it remains surprisingly underexplored in the context of molecular relational learning (MRL).

3 Preliminaries

3.1 Problem Statement

Notations. Given a molecule g, we first consider a 2D molecular graph, denoted as $g_{\text{2D}} = (\mathbf{X}, \mathbf{A})$, where $\mathbf{X} \in \mathbb{R}^{N \times F}$ represents the atom attribute matrix, and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix, with $\mathbf{A}_{ij} = 1$ if a covalent bond exists between atoms i and j. Additionally, we define a 3D conformer as $g_{\text{3D}} = (\mathbf{X}, \mathbf{R})$, where $\mathbf{R} \in \mathbb{R}^{N \times 3}$ is the matrix of 3D coordinates, each row representing the spatial position of an individual atom.

Task Description. Given a 2D molecular graph pair $(g^1_{\text{2D}}, g^2_{\text{2D}})$ and 3D conformer pair $(g^1_{\text{3D}}, g^2_{\text{3D}})$, our goal is to pre-train the 2D molecular encoders f^1_{2D} and f^2_{2D} simultaneously with the virtual interaction geometry g_{vr}, derived from the 3D conformer pair. Then, the pre-trained 2D molecular encoders f^1_{2D} and f^2_{2D} are utilized for various MRL downstream tasks.

3.2 2D MRL Model Architecture

In this paper, we mainly focus on 1) the construction of virtual interaction geometry, and 2) pre-training strategies for MRL. Therefore, we employ existing model architectures for 2D MRL, i.e., CIGIN (Pathak et al., 2020), which provides a straightforward yet effective framework for MRL as

depicted in Figure 2 (a). However, since our pre-training strategies are applicable to various architectures beyond CIGIN, we will explain how our approach has been integrated into other baseline models in Appendix B. For each pair of 2D molecular graphs, denoted as g^1_{2D} and g^2_{2D}, the graph neural networks (GNNs)-based molecular encoders f^1_{2D} and f^2_{2D} initially produce an atom embedding matrix for each molecule, formulated as:

$$\mathbf{E}^1 = f^1_{\text{2D}}\left(g^1_{\text{2D}}\right), \quad \mathbf{E}^2 = f^2_{\text{2D}}\left(g^2_{\text{2D}}\right), \tag{1}$$

where $\mathbf{E}^1 \in \mathbb{R}^{N^1 \times d}$ and $\mathbf{E}^2 \in \mathbb{R}^{N^2 \times d}$ are the atom embedding matrices for g^1_{2D} and g^2_{2D}, containing N^1 and N^2 atoms, respectively. Next, we capture the interactions between nodes in g^1_{2D} and g^2_{2D} using an interaction matrix $\mathbf{I} \in \mathbb{R}^{N^1 \times N^2}$, defined by $\mathbf{I}_{ij} = \mathsf{sim}(\mathbf{E}^1_i, \mathbf{E}^2_j)$, where $\mathsf{sim}(\cdot, \cdot)$ represents the cosine similarity measure. Subsequently, we derive new embedding matrices $\tilde{\mathbf{E}}^1 \in \mathbb{R}^{N^1 \times d}$ and $\tilde{\mathbf{E}}^2 \in \mathbb{R}^{N^2 \times d}$ for each graph, reflecting their respective interactions. This is computed using $\tilde{\mathbf{E}}^1 = \mathbf{I} \cdot \mathbf{E}^2$ and $\tilde{\mathbf{E}}^2 = \mathbf{I}^\top \cdot \mathbf{E}^1$, where $\cdot$ denotes matrix multiplication. Here, $\tilde{\mathbf{E}}^1$ represents the node embeddings of g^1_{2D} that incorporates the interaction information with nodes in g^2_{2D}, and similarly for $\tilde{\mathbf{E}}^2$. To obtain the final node embeddings, we concatenate the original and interaction-based embeddings for each graph, resulting in $\mathbf{H}^1 = (\mathbf{E}^1 || \tilde{\mathbf{E}}^1) \in \mathbb{R}^{N^1 \times 2d}$ and $\mathbf{H}^2 = (\mathbf{E}^2 || \tilde{\mathbf{E}}^2) \in \mathbb{R}^{N^2 \times 2d}$. Finally, we apply the Set2Set readout function (Vinyals et al., 2015) to compute the graph-level embeddings $\mathbf{z}^1_{\text{2D}}$ and $\mathbf{z}^2_{\text{2D}}$ for each graph g^1_{2D} and g^2_{2D}, respectively.

4 Methodology

In this section, we introduce our method, named 3DMRL, a novel pre-training framework for MRL utilizing 3D geometry information. In Section 4.1, we introduce how to construct the virtual interaction geometry that can be utilized instead of expensive calculation of real interaction geometry of molecules. Then, in Section 4.2, we present pre-training strategies for the 2D MRL model to acquire representations aligned with the constructed virtual interaction geometry and to learn the intermolecular forces between the molecules involved. The overall framework is depicted in Figure 2, and the pseudocode for the entire framework is provided in Appendix E.

4.1 Virtual Interaction Geometry Construction

While the 3D geometry of molecules plays a significant role in predicting molecular properties, acquiring this information involves a trade-off between cost and accuracy. For example, RDKit's ETKDG algorithm (Landrum, 2013) is fast but less accurate. In contrast, the widely adopted metadynamics method, CREST (Grimme, 2019), achieves a more balanced compromise between speed and accuracy, yet still requires around 6 hours to process a drug-like molecule. This challenge is even more pronounced in MRL, which necessitates not just the geometry of individual molecules but also the relative spatial arrangements between multiple molecules (Durrant & McCammon, 2011; Sosso et al., 2016). Therefore, this study aims to develop a virtual interaction geometry consisting of multiple molecules to mimic real-world molecular interactions utilizing the 3D geometry of individual molecules. However, it is not trivial to model the environment of real-world molecular interaction environments due to its chaotic nature as shown in Figure 1 (b).

Drawing inspiration from the explicit solvent models used in traditional molecular dynamics simulations (Frenkel & Smit, 2023), we propose a one-to-many geometric configuration that involves a relatively larger molecule g^1_{3D}, determined based on its radius, surrounded by multiple smaller molecules g^2_{3D} as shown in Figure 2 (b). Specifically, for a given conformer pair $(g^1_{\text{3D}} = (\mathbf{X}^1, \mathbf{R}^1), g^2_{\text{3D}} = (\mathbf{X}^2, \mathbf{R}^2))$, we create an environment by arranging the smaller molecules $(g^{2,1}_{\text{3D}}, \ldots, g^{2,i}_{\text{3D}}, \ldots, g^{2,n}_{\text{3D}})$ around a centrally placed larger molecule g^1_{3D} as follows:

- **[Step 1] Select Target Atoms in the Larger Molecule.** We start by randomly selecting n atoms from the larger molecule g^1_{3D} that are not part of any aromatic ring. This choice is based on the fact that aromatic rings are more stable and less likely to engage in chemical reactions.
- **[Step 2] Positioning the Smaller Molecules.** Each smaller molecule in $(g^{2,1}_{\text{3D}}, \ldots, g^{2,i}_{\text{3D}}, \ldots, g^{2,n}_{\text{3D}})$ is then placed close to one of the n selected atoms in the larger

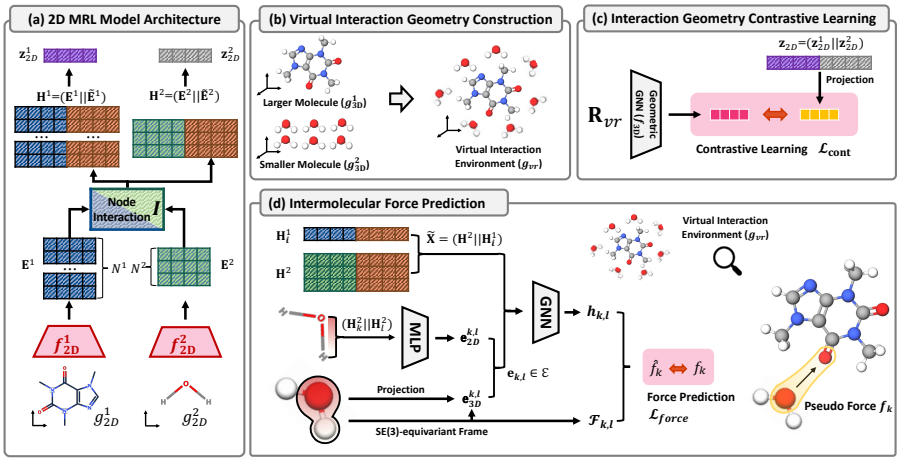

Figure 2: Framework: (a) 2D MRL model architecture (Section 3.2). (b) Virtual interaction geometry construction (Section 4.1). (c) Interaction geometry contrastive learning (Section 4.2.1). (d) Intermolecular force prediction (Section 4.2.2).

molecule g^1_{3D}. This positioning is achieved by transiting and rotating the original 3D coordinates $\mathbf{R}^2$ of the smaller molecule g^2_{3D}.

- **[Step 2-1] Determine Transition Direction and Distance.** We generate a normalized random Gaussian noise vector ε (with a norm of 1), which will be used to set the direction for the transition. We then scale this direction vector ε by the radius of the smaller molecule, r^2, to establish the transition distance.
- **[Step 2-2] Transit and Rotate to the New Position.** The new 3D coordinates for each smaller molecule are determined using the formula $\mathbf{R}^{2,i} = \mathbf{R}^2 + \varepsilon_i * r^2 + \mathbf{R}^1_i$, where $\mathbf{R}^1_i \in \mathbb{R}^3$ represents the 3D position of the i-th selected atom in the larger molecule g^1_{3D}. This operation is performed through broadcasting, meaning $\mathbf{R}^1_i$ and ε_i are added to each row of $\mathbf{R}^2$. Additionally, we apply a random rotation matrix to rotate the small molecule after its transition. This transition and rotation operations ensure that each smaller molecule is positioned close to its corresponding selected atom on the larger molecule, simulating a realistic interaction environment.

- **[Step 3] Constructing Virtual Interaction Geometry.** After positioning each smaller molecule $g^{2,i}_{3D}$ near the i-th selected atom in the larger molecule g^1_{3D}, we compile all the 3D coordinates to form a unified virtual environment g_{vr}. This process involves combining the coordinate matrix $\mathbf{R}^1$ of the larger molecule g^1_{3D}, with the transited coordinates $(\mathbf{R}^{2,1}, \ldots, \mathbf{R}^{2,i}, \ldots, \mathbf{R}^{2,n})$ of the smaller molecules $(g^{2,1}_{3D}, \ldots, g^{2,i}_{3D}, \ldots, g^{2,n}_{3D})$, resulting in $\mathbf{R}_{vr} = (\mathbf{R}^1 \| \mathbf{R}^{2,1} \| \ldots \| \mathbf{R}^{2,i} \| \ldots \| \mathbf{R}^{2,n}) \in \mathbb{R}^{(N^1 + n \cdot N^2) \times 3}$. Additionally, it involves concatenating all the atom attribute matrices to form $\mathbf{X}_{vr} = (\mathbf{X}^1 \| \mathbf{X}^2 \| \ldots \| \mathbf{X}^2) \in \mathbb{R}^{(N^1 + n \cdot N^2) \times F}$, thereby defining the virtual interaction geometry as $g_{vr} = (\mathbf{X}_{vr}, \mathbf{R}_{vr})$. Note that multiple small molecules share the same attribute matrix $\mathbf{X}^2$, since we use the atom attribute irrelevant to the atomic coordinates.

During the pre-training phase, we construct the virtual interaction geometry (**Step 1** to **Step 3**) at each epoch, allowing the 2D MRL model to learn the complex and diverse interaction geometries between paired molecules. It is important to note that, given each molecule's 3D geometry, the virtual environment can be generated in real time because transition and rotation are matrix operations. This ensures that the computational complexity of 3DMRL remains comparable to that of previous 3D pre-training approaches for single molecular property prediction (Stärk et al., 2022).

4.2 Pre-training Strategies

Once the virtual interaction geometry is established, we pre-train the 2D MRL model using two complementary strategies: interaction geometry contrastive learning (Section 4.2.1) and intermolecular force prediction (Section 4.2.2). Contrastive learning helps the model capture the overall interac-

tion geometry of the molecules, while intermolecular force prediction allows the model to learn the fine-grained atom-level interaction behavior between molecules.

4.2.1 Interaction Geometry Contrastive Learning

Given a paired 2D molecular graphs $(g^1_{\text{2D}}, g^2_{\text{2D}})$ and its corresponding 3D virtual interaction geometry g_{vr}, we first encode them with a 2D MRL model, and a geometric deep learning model, respectively. For 2D molecular graphs, we compute the molecule-level representations, $\mathbf{z}^1_{\text{2D}}$ and $\mathbf{z}^2_{\text{2D}}$, for each molecule g^1_{2D} and g^2_{2D}, respectively, as outlined in the Section 3.2. Following this, we derive the 2D interaction representation $\mathbf{z}_{\text{2D}}$, by concatenating these two representations, i.e., $\mathbf{z}_{\text{2D}} = (\mathbf{z}^1_{\text{2D}} || \mathbf{z}^2_{\text{2D}})$. On the other hand, to encode the 3D virtual interaction geometry $g_{\text{vr}} = (\mathbf{X}_{\text{vr}}, \mathbf{R}_{\text{vr}})$, we use geometric GNNs f_{3D} that output $SE(3)$ invariant (Duval et al., 2023) representations $\mathbf{z}_{\text{3D}}$ given the coordinates of atoms $\mathbf{R}_{\text{vr}}$ in virtual interaction geometry (Schütt et al., 2017), i.e., $\mathbf{z}_{\text{3D}} = f_{\text{3D}}(\mathbf{R}_{\text{vr}})$. Then, as shown in Figure 2 (c), we align the 2D interaction representation $\mathbf{z}_{\text{2D}}$ and the 3D geometry representation $\mathbf{z}_{\text{3D}}$ via Normalized temperature-scaled cross entropy (NTXent) loss (Chen et al., 2020) as follows:

$$\mathcal{L}_{\text{cont}} = -\frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \left[\log \frac{e^{\texttt{sim}(\mathbf{z}_{\text{2D},i}, \mathbf{z}_{\text{3D},i})/\tau}}{\sum_{k=1}^{N_{\text{batch}}} e^{\texttt{sim}(\mathbf{z}_{\text{2D},i}, \mathbf{z}_{\text{3D},k})/\tau}} + \log \frac{e^{\texttt{sim}(\mathbf{z}_{\text{3D},i}, \mathbf{z}_{\text{2D},i})/\tau}}{\sum_{k=1}^{N_{\text{batch}}} e^{\texttt{sim}(\mathbf{z}_{\text{3D},i}, \mathbf{z}_{\text{2D},k})/\tau}} \right], \tag{2}$$

where $\texttt{sim}(\cdot, \cdot)$ represents cosine similarity, τ denotes the temperature hyperparameter, and N_{batch} refers to the number of pairs within a batch. By training the 2D MRL model to output interaction representations that align with the 3D interaction geometry, the model can effectively learn the overall geometry of molecular interactions during the pre-training phase.

4.2.2 Intermolecular Force Prediction

Beyond the overall geometry of interaction, it is essential to learn about the intermolecular forces between molecules during molecular interactions, as these forces govern how molecules behave, interact, and react in various environments. Inspired by scientific knowledge, we propose a pre-training strategy to predict the direction of forces acting between molecules based on the assumption that forces are exerted between molecules during their interactions (London, 1937). That is, we aim to pre-train the 2D MRL model to predict forces in the constructed virtual interaction geometry. However, predicting forces from a 2D representation is challenging because the prediction must adhere to the physical properties of forces, specifically being equivariant to rotations and transitions in 3D Euclidean space, also known as *SE(3)-equivariance* (Duval et al., 2023). To address this, we propose predicting the force between molecules by utilizing local frame (Du et al., 2022), which allows for flexible conversion between invariant and equivariant features.

More specifically, given the position $\mathbf{R}^{2,i}$ of the i-th small molecule $g^{2,i}_{\text{3D}}$ in the constructed virtual interaction geometry, we first define an orthogonal local frame $\mathcal{F}_{k,l}$ between atoms k and l within molecule $g^{2,i}_{\text{3D}}$ as follows:

$$\mathcal{F}_{k,l} = \left(\frac{\mathbf{r}_k - \mathbf{r}_l}{||\mathbf{r}_k - \mathbf{r}_l||}, \frac{\mathbf{r}_k \times \mathbf{r}_l}{||\mathbf{r}_k \times \mathbf{r}_l||}, \frac{\mathbf{r}_k - \mathbf{r}_l}{||\mathbf{r}_k - \mathbf{r}_l||} \times \frac{\mathbf{r}_k \times \mathbf{r}_l}{||\mathbf{r}_k \times \mathbf{r}_l||} \right), \tag{3}$$

where $\mathbf{r}_k \in \mathbb{R}^3$ and $\mathbf{r}_l \in \mathbb{R}^3$ indicate the position of atoms k and l in constructed virtual interaction geometry, respectively. For simplicity, please note that we will omit the molecule index i in the notation from here. With the established local frame, we derive the invariant 3D feature for the edge between atoms k and l by projecting their coordinates into the local frame, i.e., $\mathbf{e}^{k,l}_{\text{3D}} = \text{Projection}_{\mathcal{F}_{k,l}}(\mathbf{r}_k, \mathbf{r}_l) \in \mathbb{R}^d$. Additionally, we obtain the 2D invariant edge feature between atoms k and l by concatenating the respective features from the 2D molecular graph, i.e., $\mathbf{e}^{k,l}_{\text{2D}} = \text{MLP}(\mathbf{H}^2_k || \mathbf{H}^2_l) \in \mathbb{R}^d$. Now that we have both invariant 2D and 3D features, we can derive the final invariant edge feature $\mathbf{e}^{k,l}$ by combining these invariant edge features as follows:

$$\mathbf{e}_{k,l} = \mathbf{e}^{k,l}_{\text{2D}} + \mathbf{e}^{k,l}_{\text{3D}}. \tag{4}$$

We define the edge feature set $\mathcal{E}$, which includes $\mathbf{e}_{k,l}$ for every possible pair of atoms.

With the invariant final edge feature set $\mathcal{E}$, we can further process the small molecule information through GNNs to predict the interaction forces between the small molecule and the central larger

molecule. To achieve this, we first obtain the atom features specific to the i-th small molecule by concatenating the i-th atom representation of the larger molecule (to which the i-th small molecule is assigned) with each atom representation of the small molecule, i.e., $\tilde{\mathbf{X}} = (\mathbf{H}^2 || \mathbf{H}_i^1) \in \mathbb{R}^{N^2 \times 4d}$ using broadcasting. This approach allows the model to learn a more precise force direction by incorporating the features of the assigned atom in the larger molecule. Next, with the edge feature set $\mathcal{E}$ and the atom feature $\tilde{\mathbf{X}}$, we derive the final edge representation $\mathbf{h}_{k,l}$ through multiple GNN layers, represented as $\mathbf{h}_{k,l} = \text{GNN}(\tilde{\mathbf{X}}, \mathcal{E})$. Finally, we determine the force direction $\hat{f}_k$ between the atom k of the small molecule and the central larger molecule by combining the final invariant edge representation $\mathbf{h}_{k,l}$ with our $SE(3)$-equivariant frame $\mathcal{F}_{k,l}$ as follows:

$$\hat{f}_k = \sum_l \mathbf{h}_{k,l} \odot \mathcal{F}_{k,l}, \tag{5}$$

where $\odot$ indicates element-wise product. This approach guarantees our predicted force $\hat{f}_k$ to be $SE(3)$-equivariant. Then, we calculate the force prediction loss as follows:

$$\mathcal{L}_{\text{force}} = \frac{1}{n \cdot N^2} \sum_{i=1}^{n} \sum_{k=1}^{N^2} ||f_k^i - \hat{f}_k^i||_2^2, \tag{6}$$

where f_k^i represents the ground truth force direction between the larger molecule and the k-th atom of the i-th small molecule, whose precise calculation is both costly and sometimes impractical. Therefore, we propose using the direction between the k-th atom of the i-th small molecule and the i-th atom of the larger molecule to which the small molecule is attached, i.e., $f_k^i = \mathbf{R}_k^{2,i} - \mathbf{R}_i^1 / ||\mathbf{R}_k^{2,i} - \mathbf{R}_i^1||_2$, as the pseudo force between these atoms is the dominant force due to their close proximity. Note that $\mathcal{L}_{\text{force}}$ is calculated for every molecule pair in the batch, although we have omitted this notation for simplicity.

Finally, we pre-train the 2D MRL model by jointly optimizing two proposed losses, i.e., contrastive loss and force prediction loss, as follows:

$$\mathcal{L}_{\text{pre-train}} = \mathcal{L}_{\text{cont}} + \alpha \cdot \mathcal{L}_{\text{force}}, \tag{7}$$

where α is a hyperparameter that determines the trade-off between the contrastive loss and the force prediction loss. After task-agnostic pre-training, the 2D molecular encoders f_{2D}^1 and f_{2D}^2 are fine-tuned for specific downstream tasks where access to 3D geometric information is limited.

5 Experiments

5.1 Experimental Setup

Pre-training Datasets. We utilize three distinct datasets to pre-train 3DMRL for each downstream task. Specifically, we use the **Chromophore** (Joung et al., 2020) dataset for pre-training when downstream tasks involve the optical properties of chromophores, the **CombiSolv** (Vermeire & Green, 2021) dataset when downstream tasks related to the solvation free energy of solutes, and the **DDI** (drug-drug interaction) dataset, which we created for the drug-drug interaction downstream task. We provide further details on how to construct pre-training pairs in the dataset in Appendix A.1.

Downstream Task Datasets. Following a prior study (Lee et al., 2023a), we employ **ten** datasets to comprehensively evaluate the performance of 3DMRL on two tasks: 1) molecular interaction prediction, and 2) drug-drug interaction (DDI) prediction. For the molecular interaction prediction task, we utilize the **Chromophore** dataset (Joung et al., 2020), which pertains to three optical properties of chromophores, along with five other datasets related to the solvation free energy of solutes: **MNSol** (Marenich et al., 2020), **FreeSolv** (Mobley & Guthrie, 2014), **CompSol** (Moine et al., 2017), **Abraham** (Grubbs et al., 2010), and **CombiSolv** (Vermeire & Green, 2021). In the Chromophore dataset, we focus on the maximum absorption wavelength (**Absorption**), maximum emission wavelength (**Emission**), and excited state lifetime (**Lifetime**) properties. For the DDI prediction task, we use two datasets: **ZhangDDI** (Zhang et al., 2017b) and **ChChMiner** (Zitnik et al., 2018), both of which contain labeled DDI data. We provide further details on datasets in Appendix A.2.

Baseline methods. We validate the effectiveness of 3DMRL by using it to enhance various recent state-of-the-art molecular relational learning methods, including **MPNN** (Gilmer et al., 2017),

Table 1: Performance improvement in molecular interaction tasks across different models with our proposed pre-training strategy (RMSE) (↓). We conduct 15 independent runs for each model and report their mean along with the standard deviation (in parentheses). Colors indicate the performance improvement compared to the models trained from scratch.

Model	Chromophore			MNSol	FreeSolv	CompSol	Abraham	CombiSolv
	Absorption	Emission	Lifetime					
MPNN	22.00 (0.30)	26.34 (0.41)	0.789 (0.021)	0.643 (0.005)	1.127 (0.110)	0.420 (0.018)	0.640 (0.008)	0.614 (0.031)
+ 3DMRL	19.96 (0.12)	25.21 (0.31)	0.753 (0.018)	0.609 (0.008)	1.068 (0.087)	0.377 (0.020)	0.550 (0.051)	0.599 (0.025)
Improvement	9.27%	4.29%	4.56%	5.28%	5.24%	10.24%	14.06%	2.44%
AttentiveFP	22.86 (0.30)	28.70 (0.23)	0.871 (0.010)	0.570 (0.021)	1.019 (0.070)	0.350 (0.008)	0.426 (0.042)	0.471 (0.028)
+ 3DMRL	22.80 (0.61)	28.54 (1.97)	0.784 (0.013)	0.562 (0.031)	0.901 (0.059)	0.271 (0.009)	0.378 (0.027)	0.448 (0.011)
Improvement	0.26%	0.55%	9.99%	1.40%	11.57%	22.57%	11.26%	4.88%
CIGIN	19.66 (0.69)	25.84 (0.23)	0.821 (0.017)	0.582 (0.022)	0.958 (0.116)	0.369 (0.018)	0.421 (0.018)	0.464 (0.002)
+ 3DMRL	18.00 (0.17)	24.21 (0.09)	0.729 (0.014)	0.528 (0.019)	0.839 (0.105)	0.277 (0.006)	0.371 (0.031)	0.435 (0.006)
Improvement	8.44%	6.30%	11.20%	9.28%	12.42%	24.93%	11.87%	6.25%
CGIB	18.37 (0.35)	24.52 (0.25)	0.808 (0.015)	0.562 (0.008)	0.876 (0.037)	0.321 (0.002)	0.404 (0.037)	0.448 (0.008)
+ 3DMRL	17.93 (0.35)	23.92 (0.29)	0.733 (0.009)	0.538 (0.020)	0.842 (0.078)	0.274 (0.002)	0.370 (0.027)	0.442 (0.015)
Improvement	2.40%	5.90%	9.28%	4.27%	3.88%	14.64%	8.42%	1.33%
$\text{CGIB}_{\text{Cont}}$	18.59 (0.24)	24.68 (0.49)	0.803 (0.019)	0.561 (0.012)	0.897 (0.098)	0.333 (0.005)	0.404 (0.039)	0.452 (0.015)
+ 3DMRL	17.90 (0.17)**	23.94 (0.24)	0.720 (0.020)	0.524 (0.018)*	0.863 (0.075)	0.284 (0.007)	0.372 (0.021)	0.441 (0.022)
Improvement	3.71%	3.00%	10.33%	6.59%	3.79%	14.71%	7.92%	2.43%

AttentiveFP (Xiong et al., 2019), **CIGIN** (Pathak et al., 2020), **CGIB** (Lee et al., 2023a), and $\textbf{CGIB}_{\textbf{Cont}}$ (Lee et al., 2023a). Additionally, we compare our proposed pre-training framework, 3DMRL, with recent molecular pre-training approaches that aim to learn 3D structure of individual molecules, such as **3D Infomax** (Stärk et al., 2022), **GraphMVP** (Liu et al., 2021), and **MoleculeSDE** (Liu et al., 2023). It is important to note that these approaches involve pre-training a single encoder for molecular property prediction (**MPP Pre-training** in Table 2), whereas our work is pioneering in training two separate encoders simultaneously during pre-training for molecular relational learning (**MRL Pre-training** in Table 2). For the baseline methods, we use the original authors' code and conduct the experiments in the same environment as 3DMRL to ensure a fair comparison. We provide more details on the compared methods in Appendix B.

Evaluation metrics. For regression tasks, we use Root Mean Squared Error (RMSE) to measure the difference between the predicted and the ground truth values. For classification tasks, we measure the model performance using the Area Under the Receiver Operating Characteristic (AUROC).

Evaluation protocol. Following Pathak et al. (2020), for the molecular interaction prediction task, we evaluate the models under a 5-fold cross-validation scheme. The dataset is randomly split into 5 subsets and one of the subsets is used as the test set, while the remaining subsets are used to train the model. A subset of the test set is selected as the validation set for hyperparameter selection and early stopping. We repeat 5-fold cross-validation three times (i.e., 15 runs in total) and report the accuracy and standard deviation of the repeats.

For the DDI prediction task (Lee et al., 2023a), we conduct experiments on two different *out-of-distribution* scenarios, namely **molecule split** and **scaffold split**. For the **molecule split**, the performance is evaluated when the models are presented with new molecules not included in the training dataset. Specifically, let $\mathbb{G}$ denote the total set of molecules in the dataset. Given $\mathbb{G}$, we split $\mathbb{G}$ into $\mathbb{G}_{\text{old}}$ and $\mathbb{G}_{\text{new}}$, so that $\mathbb{G}_{\text{old}}$ contains the set of molecules that have been seen in the training phase, and $\mathbb{G}_{\text{new}}$ contains the set of molecules that have not been seen in the training phase. Then, the new split of dataset consists of $\mathcal{D}_{\text{train}} = \{(\mathcal{G}^1, \mathcal{G}^2) \in \mathcal{D} | \mathcal{G}^1 \in \mathbb{G}_{\text{old}} \wedge \mathcal{G}^2 \in \mathbb{G}_{\text{old}}\}$ and $\mathcal{D}_{\text{test}} = \{(\mathcal{G}^1, \mathcal{G}^2) \in \mathcal{D} | (\mathcal{G}^1 \in \mathbb{G}_{\text{new}} \wedge \mathcal{G}^2 \in \mathbb{G}_{\text{new}}) \vee (\mathcal{G}^1 \in \mathbb{G}_{\text{new}} \wedge \mathcal{G}^2 \in \mathbb{G}_{\text{old}}) \vee (\mathcal{G}^1 \in \mathbb{G}_{\text{old}} \wedge \mathcal{G}^2 \in \mathbb{G}_{\text{new}})\}$. We use a subset of $\mathcal{D}_{\text{test}}$ as the validation set in inductive setting. In the **scaffold split** setting (Huang et al., 2021), just like in the molecule split, molecules corresponding to scaffolds that were not seen during training will be used for testing. For both splits, we repeat 5 independent experiments with different random seeds on split data, and report the accuracy and the standard deviation of the repeats. In both scenarios, we split the data into training, validation, and test sets with a ratio of 60/20/20%. We provide details on model implementation and training in Section C.

5.2 Experimental Results

We begin by comparing each model architecture trained from scratch with the same architecture pre-trained using our proposed strategy, referred to as +3DMRL in Table 1. We have the following

Table 2: Performance of CIGIN model on molecular interaction tasks using different pre-training strategies (RMSE) (↓). We conduct 15 independent experiments and report their mean along with the standard deviation (in parentheses). For each dataset, we highlight the best method **in bold**.

Strategy	Chromophore			MNSol	FreeSolv	CompSol	Abraham	CombiSolv
	Absorption	Emission	Lifetime					
No Pre-training	19.66 (0.69)	25.84 (0.23)	0.821 (0.017)	0.567 (0.014)	0.884 (0.074)	0.331 (0.029)	0.412 (0.028)	0.458 (0.002)
MPP (molecular property prediction) Pre-training								
3D Infomax	18.71 (0.61)	24.59 (0.22)	0.790 (0.022)	0.585 (0.015)	0.873 (0.103)	0.321 (0.041)	0.426 (0.036)	0.464 (0.004)
GraphMVP	18.40 (0.62)	24.73 (0.14)	0.797 (0.022)	0.561 (0.025)	1.010 (0.115)	0.301 (0.025)	0.418 (0.020)	0.437 (0.015)
MoleculeSDE	18.56 (0.24)	24.91 (0.10)	0.836 (0.040)	0.564 (0.018)	0.971 (0.122)	0.308 (0.024)	0.426 (0.028)	0.454 (0.012)
MRL (molecular relational learning) Pre-training								
3DMRL	**18.00** (0.17)	**24.21** (0.09)	**0.729** (0.014)	**0.528** (0.019)	**0.839** (0.105)	**0.277** (0.006)	**0.371** (0.031)	**0.435** (0.006)

observations: **1)** 3DMRL obtains consistent improvements over the base graph neural networks in all 40 tasks (across various datasets and neural architectures), achieving up to 24.93% relative reduction in RMSE. While the paper is written based on CIGIN for better understanding in Section 3.2, we could observe performance improvements not only in CIGIN but also in various other model architectures, demonstrating the versatility of proposed pre-training strategies. We further demonstrate how our pre-training strategies are adopted to various model architectures in Appendix B. **2)** We observe comparatively less performance improvement of AttentiveFP in the Chromophore dataset, which can be attributed to its limited ability to predict dipole moments, which is highly related to the optical properties of molecules, as demonstrated in their own work (Kim & Fukuda, 2006). **3)** Furthermore, the comparison between CIGIN and CGIB showed that CIGIN, when pre-trained with 3DMRL, can match or even surpass the performance of CGIB. This demonstrates that 3DMRL allows the model to perform efficiently, without requiring a complex model design for improvement.

Additionally, we compare our pre-training strategies with recent molecular pre-training approaches proposed for molecular property prediction (MPP) of a single molecule. Table 2 and Table 3 show the results for the molecular interaction prediction task, and the drug-drug interaction (DDI) task, respectively. As these approaches are originally designed for single molecules, we first pre-train the GNNs using each strategy, then incorporate the pre-trained GNNs into the CIGIN architecture and fine-tune them for various MRL downstream tasks. We have the following observations: **4)** Although MPP pre-training methods have demonstrated success in molecular property prediction in prior studies, they did not yield satisfactory results in molecular relational learning tasks and, in some cases, even resulted in negative transfer. This highlights the need for creating specialized pre-training strategies tailored to MRL tasks. We further demonstrate the MPP pre-training strategy with a large-scale dataset still performs worse than 3DMRL in Appendix D.1. **5)** On the other hand, pre-training with 3DMRL consistently delivers significant performance improvements across downstream tasks. This validates the effectiveness of our approach, as it successfully integrates scientific knowledge into the pre-training strategy, enhancing the model's overall performance. **6)** Additionally, for the DDI task in Table 3, we observed that the performance improvement is more pronounced in challenging scenarios ((b) Scaffold split) compared to less difficult ones ((a) Molecule split). This highlights the enhanced generalization ability of 3DMRL in out-of-distribution scenarios, demonstrating its potential for real-world drug discovery applications where robust generalization across diverse molecular structures is essential. We further explore the *extrapolation* capability of 3DMRL in Appendix D.2.

Table 3: Performance of CIGIN model on out-of-distribution DDI tasks using different pre-training strategies (AUROC) (↑). We conduct 5 independent experiments and report their mean along with the standard deviation (in parentheses). For each dataset, we highlight the best method **in bold**.

Strategy	(a) Molecule Split		(b) Scaffold Split	
	ZhangDDI	ChChMiner	ZhangDDI	ChChMiner
No Pre-training	71.75 (0.76)	76.21 (1.19)	70.96 (1.40)	75.81 (0.79)
MPP (molecular property prediction) Pre-training				
3D Infomax	71.01 (2.19)	76.05 (1.30)	70.90 (1.63)	74.87 (1.08)
GraphMVP	71.82 (1.44)	76.42 (1.68)	71.73 (0.95)	76.13 (1.01)
MoleculeSDE	70.07 (0.58)	76.37 (1.14)	69.46 (1.55)	76.03 (1.13)
MRL (molecular relational learning) Pre-training				
3DMRL	**74.00** (0.72)	**78.93** (0.59)	**74.85** (1.58)	**78.56** (1.03)

5.3 Model Analysis

Ablation Studies. To further understand our model, we conduct an ablation study to investigate the impact of two key components on the final performance. Specifically, as shown in Equation 7, the objective function contains two terms: (i) contrastive learning-based loss and (ii) intermolecular

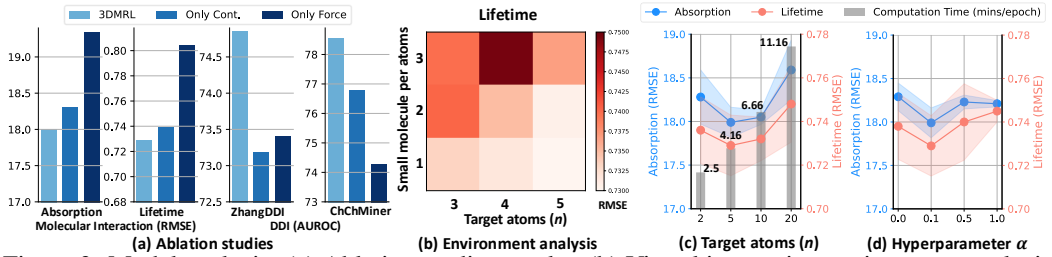

Figure 3: Model analysis: (a) Ablation studies results. (b) Virtual interaction environment analysis. Sensitivity analysis on (c) the number of target atoms n, and (d) hyperparameter α.

force prediction loss; we curate two variants that involve only (i) (denoted **only cont.**) and only (ii) (denoted **only force**) in Figure 3 (a). As shown in Figure 3 (a), the contrastive learning-based loss plays a particularly critical role. Removing it from 3DMRL results in a significant performance drop, even falling below MPP pre-training strategies such as 3D Infomax and GraphMVP. This is because the contrastive loss allows the model to capture the overall interaction geometry at the molecular level, while the force prediction loss focuses on learning more fine-grained, atom-level interactions. However, combining both losses, as in 3DMRL, yields the best results, demonstrating the importance of leveraging the strengths of both levels of granularity. We provide further detailed results of ablation studies in Appendix D.3.

Environment analysis. While we propose assigning a single small molecule to each target atom during molecular interaction in Section 4.1, we also investigate the impact of varying the number of assigned small molecules per atom in the larger molecule. As illustrated in Figure 3 (b), we observe a decline in model performance as the number of small molecules per atom increases, given a fixed number of target atoms n. This suggests that modeling interactions between multiple small molecules and a single atom in a larger molecule can degrade model performance. This is consistent with scientific understanding that, although hydrogen bonding can occasionally allow multiple molecules to interact with a single atom simultaneously, steric and electronic hindrances frequently impede such interactions. Thus, we contend that our proposed virtual interaction geometry appropriately reflects the real-world physics in molecular interactions.

Sensitivity analysis on n. Moreover, we conduct a sensitivity analysis to explore the empirical effect of the number of target atoms n, which determines the number of small molecules in a virtual interaction geometry. To do so, we examine the Chromophore dataset, where the larger molecules primarily consist of 34 atoms each. In Figure 3 (c), we observe that the model achieves optimal performance when using five small molecules to construct the virtual interaction geometry. More specifically, using too few small molecules ($n = 2$) results in poorer performance, as it fails to adequately simulate real-world interaction environments. On the other hand, the model performance also declines as the number of small molecules increases, likely due to the 3D geometry encoder overfitting to the small molecules with an excessive count. Furthermore, we observe that as the number of target atoms increases, more extensive computational resources are required to encode the 3D interaction geometry during pre-training. Hence, selecting an appropriate number of target atoms is crucial for both model performance and computational efficiency. We provide additional analyses on different datasets in Appendix D.4.

Sensitivity analysis on α. We also conduct sensitivity analysis on α, which controls the weight of force prediction loss, in Equation 7. In Figure 3 (d), the model's performance declines as α increases from 0.1, primarily because it overly emphasizes atom-level interactions between the molecules instead of considering the overall interaction geometry. Conversely, we also notice a drop in performance when force prediction loss is not utilized (i.e., $\alpha = 0.0$), as this causes the model to lose ability in learning fine-grained atom-level interactions. It is important to note that while we set $n = 5$ and $\alpha = 0.1$ across all datasets during pre-training, models pre-trained with varying n and α consistently outperform those trained from scratch, demonstrating the robustness of 3DMRL.

6 Conclusion

In this work, we propose 3DMRL, a novel pre-training framework that effectively integrates 3D geometric information into molecular relational learning (MRL). By constructing a virtual interaction geometry and employing contrastive learning and intermolecular force prediction, our approach

successfully injects complex 3D geometry information of molecular interactions into 2D MRL models. Experimental results demonstrate that 3DMRL significantly enhances the performance of 2D MRL models across various downstream tasks and neural architectures, validating the importance of incorporating 3D geometric data.

Future work will extend the current research to (1) drug-target binding affinity prediction, which is a fundamental task in drug discovery, where the larger molecule is a protein target, involving more complex protein structures, and (2) organic-inorganic interaction prediction tasks mimicking the dissociation process, focusing on accurately modeling the behavior of organic-inorganic complexes.

References

Peter William Atkins, Julio De Paula, and James Keeler. *Atkins' physical chemistry*. Oxford university press, 2023.

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

Tianyi Chen, Nan Hao, Yingzhou Lu, Capucine Van Rechem, and Tianfan Fu. Uncertainty quantification and interpretability for clinical trial approval prediction. *Health Data Science*, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Andreea Deac, Yu-Hsiang Huang, Petar Veličković, Pietro Liò, and Jian Tang. Drug-drug adverse effect prediction with graph co-attention. *arXiv preprint arXiv:1905.00534*, 2019.

Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Se (3) equivariant graph neural networks with complete local frames. In *International Conference on Machine Learning*, pp. 5583–5608. PMLR, 2022.

Jacob D Durrant and J Andrew McCammon. Molecular dynamics simulations and drug discovery. *BMC biology*, 9:1–9, 2011.

Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Lio, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023.

Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2023.

Yi Fu, Yingzhou Lu, Yizhi Wang, Bai Zhang, Zhen Zhang, Guoqiang Yu, Chunyu Liu, Robert Clarke, David M Herrington, and Yue Wang. Ddn3. 0: Determining significant rewiring of biological network structure with differential dependency networks. *Bioinformatics*, pp. btae376, 2024.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Stefan Grimme. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *Journal of chemical theory and computation*, 15(5):2847–2862, 2019.

Laura M Grubbs, Mariam Saifullah, E Nohelli, Shulin Ye, Sai S Achi, William E Acree Jr, and Michael H Abraham. Mathematical correlations for describing solute transfer into functionalized alkane solvents containing hydroxyl, ether, ester or ketone solvents. *Fluid phase equilibria*, 298 (1):48–53, 2010.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Joonyoung F Joung, Minhi Han, Minseok Jeong, and Sungnam Park. Experimental database of optical properties of organic compounds. *Scientific data*, 7(1):1–6, 2020.

Joonyoung F Joung, Minhi Han, Jinhyo Hwang, Minseok Jeong, Dong Hoon Choi, and Sungnam Park. Deep learning optical spectroscopy based on experimental database: potential applications to molecular design. *JACS Au*, 1(4):427–438, 2021.

Jun Young Kim and Takashi Fukuda. Relationship between dipole moment of chromophores and photoinduced birefringence. *Japanese journal of applied physics*, 45(1S):456, 2006.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Namkyeong Lee, Dongmin Hyun, Gyoung S Na, Sungwon Kim, Junseok Lee, and Chanyoung Park. Conditional graph information bottleneck for molecular relational learning. In *International Conference on Machine Learning*, pp. 18852–18871. PMLR, 2023a.

Namkyeong Lee, Kanghoon Yoon, Gyoung S Na, Sein Kim, and Chanyoung Park. Shift-robust molecular relational learning with causal substructure. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1200–1212, 2023b.

Hyuntae Lim and YounJoon Jung. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chemical science*, 10(36):8306–8315, 2019.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*, 2021.

Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *International Conference on Machine Learning*, pp. 21497–21526. PMLR, 2023.

Fritz London. The general theory of molecular forces. *Transactions of the Faraday Society*, 33: 8b–26, 1937.

Yingzhou Lu, Yaojun Hu, and Chenhao Li. Drugclip: Contrastive drug-disease interaction for drug repurposing. *arXiv preprint arXiv:2407.02265*, 2024.

Aleksandr V Marenich, Casey P Kelly, Jason D Thompson, Gregory D Hawkins, Candee C Chambers, David J Giesen, Paul Winget, Christopher J Cramer, and Donald G Truhlar. Minnesota solvation database (mnsol) version 2012. 2020.

David L Mobley and J Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28(7):711–720, 2014.

Edouard Moine, Romain Privat, Baptiste Sirjean, and Jean-Noël Jaubert. Estimation of solvation quantities from experimental thermodynamic data: Development of the comprehensive compsol databank for pure and mixed solutes. *Journal of Physical and Chemical Reference Data*, 46(3): 033102, 2017.

Yashaswi Pathak, Siddhartha Laghuvarapu, Sarvesh Mehta, and U Deva Priyakumar. Chemically interpretable graph interaction network for prediction of pharmacokinetic properties of drug-like molecules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 873–880, 2020.

Kristina Preuer, Richard PI Lewis, Sepp Hochreiter, Andreas Bender, Krishna C Bulusu, and Günter Klambauer. Deepsynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, 34(9):1538–1546, 2018.

Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the national academy of sciences*, 115(18):E4304–E4311, 2018.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Richard B Silverman and Mark W Holladay. *The organic chemistry of drug design and drug action*. Academic press, 2014.

Gabriele C Sosso, Ji Chen, Stephen J Cox, Martin Fitzner, Philipp Pedevilla, Andrea Zen, and Angelos Michaelides. Crystal nucleation in liquids: Open questions and future challenges in molecular dynamics simulations. *Chemical reviews*, 116(12):7078–7116, 2016.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pp. 20479–20502. PMLR, 2022.

Alejandro Velez-Arce, Kexin Huang, Michelle M Li, Xiang Lin, Wenhao Gao, Tianfan Fu, Manolis Kellis, Bradley L Pentelute, and Marinka Zitnik. Tdc-2: Multimodal foundation for therapeutic science. *bioRxiv*, pp. 2024–06, 2024.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Florence H Vermeire and William H Green. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 418:129307, 2021.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015.

Yingheng Wang, Yaosen Min, Xin Chen, and Ji Wu. Multi-view graph contrastive representation learning for drug-drug interaction prediction. In *Proceedings of the Web Conference 2021*, pp. 2921–2933, 2021.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, 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.

Bohao Xu, Yingzhou Lu, Chenhao Li, Ling Yue, Xiao Wang, Nan Hao, Tianfan Fu, and Jim Chen. Smiles-mamba: Chemical mamba foundation models for drug admet prediction. *arXiv preprint arXiv:2408.05696*, 2024.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. *arXiv preprint arXiv:2206.00133*, 2022.

Bai Zhang, Yi Fu, Yingzhou Lu, Zhen Zhang, Robert Clarke, Jennifer E Van Eyk, David M Herrington, and Yue Wang. Ddn2. 0: R and python packages for differential dependency network analysis of biological systems. *bioRxiv*, pp. 2021–04, 2021.

Jin Zhang, Haiyang Zhang, Tao Wu, Qi Wang, and David Van Der Spoel. Comparison of implicit and explicit solvent models for the calculation of solvation free energy in organic solvents. *Journal of chemical theory and computation*, 13(3):1034–1043, 2017a.

Wen Zhang, Yanlin Chen, Feng Liu, Fei Luo, Gang Tian, and Xiaohong Li. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC bioinformatics*, 18(1):1–12, 2017b.

Marinka Zitnik, Rok Sosic, and Jure Leskovec. Biosnap datasets: Stanford biomedical network dataset collection. *Note: http://snap. stanford. edu/biodata Cited by*, 5(1), 2018.

Supplementary Material for 3D Interaction Geoemetric Pre-training for Molecular Relational Learning

A Datasets **15**

A.1 Pre-Training Datasets 15

A.2 Downstream Task Datasets 15

B Baselines Setup **16**

C Implementation Details **18**

C.1 Model architecture 18

C.2 Model training 18

D Additional Experimental Results **18**

D.1 Molecular Property Prediction Pre-training with Large-Scale Datasets 18

D.2 Extrapolation in Molecular Interaction Task 18

D.3 Ablation Studies 20

D.4 Environment Analysis 20

E Pseudocode **21**

A Datasets

A.1 Pre-Training Datasets

We utilize three distinct datasets, i.e., **Chromophore**, **CombiSolv**, and **DDI**, to pre-train 3DMRL for each downstream task as described in Section 5. Specifically, we use the **Chromophore** dataset for downstream tasks involving the optical properties of chromophores, the **CombiSolv** dataset for tasks related to the solvation free energy of solutes, and the **DDI** dataset, which we created for the drug-drug interaction task.

- The **Chromophore** dataset (Joung et al., 2020) consists of 20,236 combinations derived from 6,815 chromophores and 1,336 solvents, provided in SMILES string format. For pre-training, we initially convert chromophores and solvents into their respective 3D structures via rdkit, resulting in 6,524 3D structures for chromophores and 1,255 for solvents. These 6,524 unique chromophores are then randomly paired with the 1,255 solvents to generate a sufficient number of pairs. Out of the possible 8,187,620 chromophore-solvent combinations, we randomly sample 1%, which corresponds to 81,876 pairs, for pre-training.
- The **CombiSolv** dataset (Vermeire & Green, 2021) contains 10,145 combinations derived from 1,368 solutes and 291 solvents, provided in SMILES string format. Similar to our approach with the Chromophore dataset, we first convert solutes and solvents into their corresponding 3D structures, yielding 1,368 3D structures for solutes and 290 for solvents. From the potential random combinations, we select 79,344 solute-solvent pairs, representing 20% of all possible pairs.
- For the **DDI** dataset, we compile drug-drug pairs from the ZhangDDI (Zhang et al., 2017b), ChChMiner (Zitnik et al., 2018), and DeepDDI (Ryu et al., 2018) datasets. From a total of 235,547 positive pairs, we randomly sample 40% (i.e., 94,218 pairs) for use as the pre-training dataset. While chromophores and solutes act as the larger molecule g^1 in molecular interaction tasks, in the DDI dataset, we designate the drug with the larger radius as the larger molecule.

A.2 Downstream Task Datasets

Molecular Interaction Prediction. For the molecular interaction prediction task, we transform the SMILES strings into graph structures using the CIGIN implementation available on GitHub [1](Pathak et al., 2020). Regarding the datasets related to solvation free energies, such as MNSol, FreeSolv, CompSol, Abraham, and CombiSolv, we utilize SMILES-based datasets from previous studies (Vermeire & Green, 2021). Following previous work (Lee et al., 2023a), we specifically filter the data to include only solvation free energies measured at temperatures of 298 K ($\pm$ 2) and exclude any data involving ionic liquids and ionic solutes (Vermeire & Green, 2021).

- The **Chromophore** dataset (Joung et al., 2020) consists of 20,236 combinations derived from 6,815 chromophores and 1,336 solvents, provided in SMILES string format. This dataset includes optical properties sourced from scientific publications, with unreliable experimental results being excluded after thorough examination of absorption and emission spectra. In our work, we assess model performance by predicting three key properties: **maximum absorption wavelength (Absorption)**, **maximum emission wavelength (Emission)**, and **excited state lifetime (Lifetime)**, which are crucial for designing chromophores for specific applications. To ensure the integrity of each dataset, we remove any NaN values that were not reported in the original publications. Additionally, following previous work (Lee et al., 2023a), for the Lifetime data, we apply log normalization to the target values to mitigate skewness in the dataset, thereby enhancing training stability.
- The **MNSol** dataset (Marenich et al., 2020) features 3,037 experimentally measured free energies of solvation or transfer for 790 distinct solutes and 92 solvents. For our study, we focus on 2,275 pairs comprising 372 unique solutes and 86 solvents, in alignment with prior research (Vermeire & Green, 2021).
- The **FreeSolv** dataset (Mobley & Guthrie, 2014) offers 643 hydration free energy values, both experimental and calculated, for small molecules in water. In our research, we utilize 560 experimental measurements, consistent with the dataset selection criteria from previous studies (Vermeire & Green, 2021).

[1] `https://github.com/devalab/CIGIN`

Table 4: Statistics of datasets. $\mathcal{G}^1$ and $\mathcal{G}^2$ are defined in Section 5.1.

Task	Dataset		$\mathcal{G}^1$	$\mathcal{G}^2$	# $\mathcal{G}^1$	# $\mathcal{G}^2$	# Pairs
Molecular Interaction	Chromophore [3]	Absorption	Chromophore	Solvent	6,416	725	17,276
		Emission	Chromophore	Solvent	6,412	1,021	18,141
		Lifetime	Chromophore	Solvent	2,755	247	6,960
	MNSol [4]		Solute	Solvent	372	86	2,275
	FreeSolv [5]		Solute	Solvent	560	1	560
	CompSol [6]		Solute	Solvent	442	259	3,548
	Abraham [7]		Solute	Solvent	1,038	122	6,091
	CombiSolv [8]		Solute	Solvent	1,495	326	10,145
Drug-Drug Interaction	ZhangDDI [9]		Small-molecule Drug	Small-molecule Drug	544	544	40,255
	ChChMiner [10]		Small-molecule Drug	Small-molecule Drug	949	949	21,082

- The **CompSol** dataset (Moine et al., 2017) has been designed to illustrate the impact of hydrogen-bonding association effects on solvation energies. For our study, we analyze 3,548 solute-solvent pairs, encompassing 442 distinct solutes and 259 solvents, in accordance with prior research parameters (Vermeire & Green, 2021).
- The **Abraham** dataset (Grubbs et al., 2010), curated by the Abraham research group at University College London, provides extensive data on solvation. For this study, we focus on 6,091 solute-solvent combinations, comprising 1,038 distinct solutes and 122 solvents, as outlined in previous research (Vermeire & Green, 2021).
- The **CombiSolv** dataset (Vermeire & Green, 2021) integrates the data from MNSol, FreeSolv, CompSol, and Abraham, encompassing a total of 10,145 solute-solvent combinations. This dataset features 1,368 unique solutes and 291 distinct solvents.

Drug-Drug Interaction (DDI) Prediction. In the drug-drug interaction prediction task, we utilize the positive drug pairs provided in the MIRACLE GitHub repository[2], which excludes data instances that cannot be represented as graphs from SMILES strings. To create negative samples, we generate a corresponding set by sampling from the complement of the positive drug pairs. This approach is applied to both datasets. Additionally, for the classification task, we adhere to the graph conversion process outlined by MIRACLE (Wang et al., 2021).

- The **ZhangDDI** dataset (Zhang et al., 2017b) includes data on 548 drugs and 48,548 pairwise interactions, along with various types of similarity information pertaining to these drug pairs.
- The **ChChMiner** dataset (Zitnik et al., 2018) comprises 1,322 drugs and 48,514 annotated DDIs, sourced from drug labels and scientific literature.

Despite the **ChChMiner** dataset containing a significantly higher number of drug instances compared to the **ZhangDDI** dataset, the number of labeled DDIs is nearly equivalent. This suggests that the **ChChMiner** dataset exhibits a much sparser network of relationships between drugs.

B Baselines Setup

To validate the effectiveness of 3DMRL, we primarily evaluate molecular relational learning model architectures trained from scratch for downstream tasks, as well as the same models that are first pre-trained with 3DMRL and then fine-tuned for various downstream tasks. We include the following molecular relational learning model architectures:

[2] https://github.com/isjakewong/MIRACLE/tree/main/MIRACLE/datachem

[3] https://figshare.com/articles/dataset/DB_for_chromophore/12045567/2

[4] https://conservancy.umn.edu/bitstream/handle/11299/213300/MNSolDatabase_v2012.zip?sequence=12&isAllowed=y

[5] https://escholarship.org/uc/item/6sd403pz

[6] https://aip.scitation.org/doi/suppl/10.1063/1.5000910

[7] https://www.sciencedirect.com/science/article/pii/S0378381210003675

[8] https://ars.els-cdn.com/content/image/1-s2.0-S1385894721008925-mmc2.xlsx

[9] https://github.com/zw9977129/drug-drug-interaction/tree/master/dataset

[10] http://snap.stanford.edu/biodata/datasets/10001/10001-ChCh-Miner.html

- **MPNN** (Message Passing Neural Networks) (Gilmer et al., 2017) was originally proposed to predict the various chemical properties of a single molecule. For molecular relational learning tasks, we independently encode each molecule in a pair using MPNN and then concatenate their representations.

  To apply 3DMRL for **MPNN**, we first obtain the atom representation matrices $\mathbf{E}^1$ and $\mathbf{E}^2$ using f^1_{2D} and f^1_{2D}, which are MPNNs. Then, we directly use $\mathbf{E}^1$ and $\mathbf{E}^2$ instead of the $\mathbf{H}^1$ and $\mathbf{H}^2$, which considers the interaction between two molecules in Section 3.2. That is, we obtain graph-level embeddings $\mathbf{z}^1_{\text{2D}}$ and $\mathbf{z}^2_{\text{2D}}$ via $\mathbf{E}^1$ and $\mathbf{E}^2$ with Set2set readout function. Following contrastive learning is done with $\mathbf{z}^1_{\text{2D}}$ and $\mathbf{z}^2_{\text{2D}}$, and the edge representations $\mathbf{e}^{k,l}_{\text{2D}}$ and and initial atom representations for force prediction $\hat{\mathbf{X}}$ is obtained through $\mathbf{E}^1$ and $\mathbf{E}^2$. One can simply alternate $\mathbf{H}^1$ and $\mathbf{H}^2$ in Section 4 to $\mathbf{E}^1$ and $\mathbf{E}^2$.

- **AttentiveFP** (Xiong et al., 2019) was also initially proposed to predict various chemical properties of individual molecules by employing a graph attention mechanism to gather more information from relevant molecular datasets. For molecular relational learning tasks, we independently encode each molecule in a pair using MPNN and then concatenate their representations.

  More specifically, **AttentiveFP** first obtain atom representation matrices $\mathbf{H}^1$ and $\mathbf{H}^2$ using f^1_{2D} and f^1_{2D}, which consist of GAT and GRU layers. Then, the model obtain initial molecule representation $\tilde{\mathbf{z}}^1_{\text{2D}}$ and $\tilde{\mathbf{z}}^2_{\text{2D}}$ which are further enhanced by considering other molecules in a batch through GAT layers. After passing multiple GAT layers, the model obtain final molecule representations $\tilde{\mathbf{z}}^1_{\text{2D}}$ and $\tilde{\mathbf{z}}^2_{\text{2D}}$. In our framework, contrastive learning is done with $\mathbf{z}^1_{\text{2D}}$ and $\mathbf{z}^2_{\text{2D}}$, and the edge representations $\mathbf{e}^{k,l}_{\text{2D}}$ and and initial atom representations for force prediction $\hat{\mathbf{X}}$ is obtained through $\mathbf{H}^1$ and $\mathbf{H}^2$.

- **CIGIN** (Chemically Interpretable Graph Interaction Network) (Pathak et al., 2020) proposes to model the interaction between the molecules through a dot product between atoms in paired molecules. By doing so, they successfully predict the solubility of drug molecules. We provide detailed descriptions on how to apply 3DMRL for CIGIN in Section 4.

- **CGIB** (Conditional Graph Information Bottleneck) and **CGIB**$_{\text{cont}}$ (Conditional Graph Information Bottleneck with Contrastive Learning)(Lee et al., 2023a) aim to enhance generalization in molecular relational learning by identifying the core substructure of molecules during chemical reactions, based on the information bottleneck theory. While CIGIN is limited to predicting drug solubility, **CGIB** and **CGIB**$_{\text{cont}}$ extend molecular relational learning to predict the optical properties of chromophores in various solvents, molecule solubility in various solvents, and drug-drug interactions.

  CGIB and **CGIB**$_{\text{cont}}$ model architectures are highly similar to CIGIN, but they have another branch named *compress module*, which aims to inject noise to the atoms that are not important during the model. Specifically, they obtain $\mathbf{T}^1$ that is node representation matrix with noise, and obtain $\mathbf{z}_{\mathcal{G}^1_{\text{CIB}}}$ from the noise injected matrix along with $\mathbf{z}_{\mathcal{G}^1}$ and $\mathbf{z}_{\mathcal{G}^2}$ which are obtained from $\mathbf{H}^1$ and $\mathbf{H}^2$, respectively. To apply 3DMRL for **CGIB**, we pre-train the model without noise injection module, thereby using $\mathbf{H}^1$, $\mathbf{H}^2$, $\mathbf{z}_{\mathcal{G}^1}$, and $\mathbf{z}_{\mathcal{G}^2}$ in **CGIB** as $\mathbf{H}^1$, $\mathbf{H}^2$, $\mathbf{z}^1_{\text{2D}}$, and $\mathbf{z}^2_{\text{2D}}$ in Section 4. After pre-training staget, all the modules including noise injection module is trained for the downstream tasks.

In addition to the model architectures, we also compare the recent state-of-the-art molecular pre-training methods based on CIGIN architecture. Since molecular pre-training methods are specifically designed for a single molecule, we pre-train each molecule encoder in CIGIN architecture and adopted the pre-trained weights for molecular relational learning downstream tasks. In Section 5, we include following molecular pre-training approaches:

- **No pre-training** does not involve pertaining process and fine-tune the model using labeled data
- **3D Infomax** (Stärk et al., 2022) increase the mutual information between 2D and 3D molecular representations using contrastive learning
- **GraphMVP** (Liu et al., 2021) incorporates a generative pre-training framework in addition to contrastive learning
- **MoleculeSDE** (Liu et al., 2023) designs a denoising framework to capture the 3D geometric distribution of molecules, thereby revealing the relationship between the score function and the molecular force field.

To apply these approaches for MRL, we first pre-train the each encoder f^1_{2D} and f^2_{2D} in Section 3.2 with the above approaches. Then, the pre-trained encoders f^1_{2D} and f^2_{2D} are utilized to output the representations $\mathbf{E}^1$ and $\mathbf{E}^2$, following the remaining pipeline of the model outlined in Section 3.2. That is, each molecule encoder f^1_{2D} and f^2_{2D} implicitly possesses knowledge about the 3D structure of individual molecules, but not the complex interaction geometry between multiple molecules.

C Implementation Details

C.1 Model architecture

For the 2D MRL model, following a previous work (Pathak et al., 2020), we use 3-layer MPNNs (Gilmer et al., 2017) as our backbone molecule encoder to learn the representation of solute and solvent for the molecular interaction prediction, while we use a GIN (Xu et al., 2018) to encode both drugs for the drug-drug interaction prediction task (Lee et al., 2023a). We utilize a hidden dimension of 56 for molecular interaction tasks and 300 for drug-drug interaction tasks, employing the ReLU activation function for both. For the 3D virtual environment encoder f_{3D}, we utilize SchNet (Schütt et al., 2017), which guarantees an *SE(3)-invariant* representation of the environment. For both molecular interaction and drug-drug interaction tasks, we configure SchNet with 128 hidden channels, 128 filters, 6 interaction layers, and a cutoff distance of 5.0.

C.2 Model training

For model optimization during **Pre-training** stage, we employ the Adam optimizer with an initial learning rate of 0.0005 for the chromophore task, 0.0001 for the solvation free energy task, and 0.0005 for the DDI tasks. The model is optimized over 100 epochs during pre-training.

In the **downstream tasks**, the learning rate was reduced by a factor of 10^{-1} after 20 epochs of no improvement in model performance in validation set, following the approach in a previous work (Pathak et al., 2020), with the initial learning rate of 0.005 for the chromophore task, 0.001 for the solvation free energy task, and 0.0005 for the DDI tasks.

Computational resources. We perform all pre-training on a 40GB NVIDIA A6000 GPU, whereas all downstream tasks are executed on a 24GB NVIDIA GeForce RTX 3090 GPU.

Software configuration. Our model is implemented using Python 3.7, PyTorch 1.9.1, RDKit 2020.09.1, and Pytorch-geometric 2.0.3.

D Additional Experimental Results

D.1 Molecular Property Prediction Pre-training with Large-Scale Datasets

Although MPP pre-training approaches demonstrate unsatisfactory performance in Section 5, a positive aspect is their ability to leverage large-scale datasets containing both 2D and 3D molecular information. Consequently, we further explore whether utilizing a large-scale pre-training dataset can enhance MPP pre-training strategies in MRL tasks. To do so, we pre-train the encoders with each strategy with randomly sampled 50K molecules in GEOM dataset (Axelrod & Gomez-Bombarelli, 2022), which consists of 2D topological information and 3D geometric information, following the previous work (Liu et al., 2021). In Table 5, we observe that a large-scale pre-training dataset does not consistently result in performance improvements for MRL downstream tasks and can still cause negative transfer in various tasks. On the other hand, we note that MoleculeSTM benefits the most from the large-scale dataset among the strategies, likely due to the complexity of its denoising framework, which necessitates a large-scale dataset to learn the data distribution effectively. Nevertheless, it still exhibits negative transfer in the FreeSolve dataset and performs worse than 3DMRL, highlighting the need for a pre-training strategy specifically tailored to molecular relational learning.

D.2 Extrapolation in Molecular Interaction Task

The model's generalization ability in out-of-distribution (OOD) datasets is crucial for its application in real-world scientific discovery processes. To this end, we further conduct experiments on molecular interaction tasks by assuming out-of-distribution scenarios, as shown in Table 6. Specifically,

Table 5: Performance comparison of CIGIN model on molecular interaction tasks using different pre-training strategies and pre-training dataset (RMSE) (↓). The blue color signifies a positive transfer between the pre-training task and the downstream task, whereas the orange color denotes a negative transfer between the pre-training task and the downstream task. **Pre-training Dataset** indicates the pre-training datasets used during pre-training.

Strategy	Pre-training Dataset	Chromophore			MNSol	FreeSolv	CompSol	Abraham	CombiSolv
		Absorption	Emission	Lifetime					
No Pre-training	-	19.66 (0.69)	25.84 (0.23)	0.821 (0.017)	0.567 (0.014)	0.884 (0.074)	0.331 (0.029)	0.412 (0.028)	0.458 (0.002)
MPP (molecular property prediction) Pre-training									
3D Infomax	MRL	18.71 (0.61)	24.59 (0.22)	0.790 (0.022)	0.585 (0.015)	0.873 (0.103)	0.321 (0.041)	0.426 (0.036)	0.464 (0.004)
	GEOM	18.82 (0.24)	25.14 (0.18)	0.795 (0.021)	0.589 (0.027)	0.899 (0.080)	0.319 (0.019)	0.418 (0.023)	0.466 (0.017)
GraphMVP	MRL	18.40 (0.62)	24.73 (0.14)	0.797 (0.022)	0.561 (0.025)	1.010 (0.115)	0.301 (0.025)	0.418 (0.020)	0.437 (0.015)
	GEOM	18.85 (0.74)	24.87 (0.54)	0.784 (0.014)	0.551 (0.013)	0.900 (0.059)	0.325 (0.007)	0.410 (0.036)	0.437 (0.007)
MoleculeSDE	MRL	18.56 (0.24)	24.91 (0.10)	0.836 (0.040)	0.564 (0.018)	0.971 (0.122)	0.308 (0.024)	0.426 (0.028)	0.454 (0.012)
	GEOM	18.72 (0.16)	24.77 (0.48)	0.773 (0.023)	0.560 (0.086)	0.909 (0.142)	0.290 (0.008)	0.399 (0.034)	0.449 (0.007)
MRL (molecular relational learning) Pre-training									
3DMRL	MRL	**18.00** (0.17)	**24.21** (0.09)	**0.729** (0.014)	**0.528** (0.019)	**0.839** (0.105)	**0.277** (0.006)	**0.371** (0.031)	**0.435** (0.006)

we split the dataset based on molecular structure, i.e., molecule split and scaffold split, similar to the approach used in the DDI task in Section 5. It is important to note that this scenario is significantly more challenging than the out-of-distribution DDI task in Section 5 because it involves a regression task, which can also be viewed as an **extrapolation** task. As shown in Table 6, we observe that pre-training approaches generally benefit model performance in extrapolation tasks, with the exception of one case, namely 3D Infomax for the Lifetime dataset. Among the pre-training approaches, 3DMRL performs the best, underscoring the extrapolation capability of 3DMRL.

Table 6: Performance comparison of the CIGIN model on extrapolation in molecular interaction tasks using different pre-training strategies (RMSE) (↓).

Strategy	Molecule Split			Scaffold Split		
	Absorption	Emission	Lifetime	Absorption	Emission	Lifetime
No Pre-training	27.51 (0.74)	37.04 (1.07)	1.205 (0.033)	59.55 (1.35)	60.11 (1.98)	1.221 (0.033)
MPP (molecular property prediction) Pre-training						
3D Infomax	27.38 (1.19)	36.98 (1.24)	1.257 (0.050)	58.34 (1.89)	58.67 (1.00)	1.207 (0.041)
GraphMVP	26.93 (1.89)	36.51 (0.92)	1.201 (0.034)	59.27 (1.57)	57.67 (1.14)	1.199 (0.024)
MoleculeSDE	27.26 (1.19)	36.48 (1.12)	1.135 (0.077)	57.75 (0.74)	58.74 (1.02)	1.214 (0.010)
MRL (molecular relational learning) Pre-training						
3DMRL	**25.01** (1.51)	**34.66** (0.89)	**1.033** (0.027)	**57.58** (1.62)	**57.53** (1.13)	**1.178** (0.010)

D.3 Ablation Studies

We provide further ablation studies on molecular interaction task and drug-drug interaction task in Table 7 and 8, respectively.

Table 7: Further results from ablation studies on molecular interaction tasks.

Strategy	Chromophore			MNSol	FreeSolv	CompSol	Abraham	CombiSolv
	Absorption	Emission	Lifetime					
Only Cont.	18.30 (0.16)	24.70 (0.16)	0.739 (0.015)	0.531 (0.022)	0.874 (0.060)	0.301 (0.018)	0.376 (0.029)	0.458 (0.014)
Only Force	19.34 (0.50)	24.80 (0.05)	0.804 (0.011)	0.587 (0.019)	1.184 (0.173)	0.330 (0.028)	0.391 (0.020)	0.466 (0.021)
3DMRL	**18.00** (0.17)	**24.21** (0.09)	**0.729** (0.014)	**0.528** (0.019)	**0.839** (0.105)	**0.277** (0.006)	**0.371** (0.031)	**0.435** (0.006)

Table 8: Further results from ablation studies on drug-drug interaction tasks.

Strategy	(a) Molecule Split		(b) Scaffold Split	
	ZhangDDI	ChChMiner	ZhangDDI	ChChMiner
Only Cont.	73.09 (0.83)	77.68 (0.55)	73.18 (0.59)	76.79 (1.13)
Only Force	73.45 (1.29)	75.93 (1.14)	73.41 (2.28)	74.29 (1.79)
3DMRL	**74.00** (0.72)	**78.93** (0.59)	**74.85** (1.58)	**78.56** (1.03)

D.4 Environment Analysis

We provide further environment analysis in Figure 4. Once again, we observe that modeling a one-to-one relationship between target atoms and small molecules generally yields the best performance when the number of target atoms is fixed.

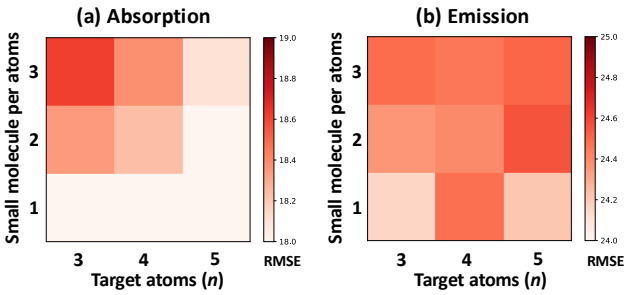

Figure 4: Environment analysis on (a) Absorption and (b) Emission properties in Chromophore Dataset

E Pseudocode

In this section, we provide pseudocode of 3DMRL in Algorithm 1.

Algorithm 1 Overall framework of 3DMRL.

Input:

- 2D molecular topology graphs $g^1_{\text{2D}}, g^2_{\text{2D}}$
- 3D molecular geometric graphs $g^1_{\text{3D}}, g^2_{\text{3D}}$
- 2D graph encoders $f^1_{\text{2D}}, f^2_{\text{2D}}$
- 3D Virtual Interaction Geometry Encoder f_{3D}

Pre-Training Stage:

For epoch **in** epochs:

$\mathbf{z}^1_{\text{2D}}, \mathbf{z}^2_{\text{2D}}, \mathbf{H}^1, \mathbf{H}^2 =$ 2D MRL ENCODER $(g^1_{\text{2D}}, g^2_{\text{2D}})$

$\mathbf{z}_{\text{2D}} = (\mathbf{z}^1_{\text{2D}} || \mathbf{z}^2_{\text{2D}})$

$g_{\text{vr}} =$ VIRTUAL INTERACTION GEOMETRY CONSTRUCTION $(g^1_{\text{3D}}, g^2_{\text{3D}})$

$\mathbf{z}_{\text{3D}} = f_{\text{3D}}(g_{\text{vr}})$ /* Virtual Geometry Encoding via SchNet */

$\mathcal{L}_{\text{cont}} =$ INTERACTION GEOMETRY CONTRASTIVE LOSS $(\mathbf{z}_{\text{2D}}, \mathbf{z}_{\text{3D}})$

$\mathcal{L}_{\text{force}} = \frac{1}{n}\sum_{i=1}^{n}$ INTERMOLECULAR FORCE PREDICTION LOSS $(g^{2,i}_{\text{3D}}, \mathbf{H}^1, \mathbf{H}^2)$

$\mathcal{L}_{\text{pre-train}} = \mathcal{L}_{\text{cont}} + \alpha \cdot \mathcal{L}_{\text{force}}$

Update $f^1_{\text{2D}}, f^2_{\text{2D}}$, and f_{3D}

Function 2D MRL ENCODER $(g^1_{\text{2D}}, g^2_{\text{2D}})$

$\mathbf{E}^1 = f^1_{\text{2D}}(g^1_{\text{2D}}), \quad \mathbf{E}^1 = f^2_{\text{2D}}(g^2_{\text{2D}})$

$\mathbf{I}_{ij} = \mathsf{sim}(\mathbf{E}^1_i, \mathbf{E}^2_j)$

where $\mathsf{sim}(\cdot,\cdot)$ is cosine similarity

$\tilde{\mathbf{E}}^1 = \mathbf{I} \cdot \mathbf{E}^2, \quad \tilde{\mathbf{E}}^2 = \mathbf{I}^\top \cdot \mathbf{E}^1$

$\mathbf{H}^1 = (\mathbf{E}^1 || \tilde{\mathbf{E}}^1), \quad \mathbf{H}^2 = (\mathbf{E}^2 || \tilde{\mathbf{E}}^2)$

$\mathbf{z}^1_{\text{2D}} = \text{Set2set}(\mathbf{H}^1), \quad \mathbf{z}^2_{\text{2D}} = \text{Set2set}(\mathbf{H}^2)$

return $\mathbf{z}^1_{\text{2D}}, \mathbf{z}^2_{\text{2D}}, \mathbf{H}^1, \mathbf{H}^2$

Function VIRTUAL INTERACTION GEOMETRY CONSTRUCTION $(g^1_{\text{3D}}, g^2_{\text{3D}})$

Randomly select n atoms in larger molecule g^1_{3D}

Copy small molecule g^2_{3D} to n small molecules $g^{2,1}_{\text{3D}}, \ldots, g^{2,i}_{\text{3D}}, \ldots, g^{2,n}_{\text{3D}}$

Generate a normalized random Gaussian noise vector ε

Create new 3D coordinates for each smaller molecule $g^{2,i}_{\text{3D}}$

$\mathbf{R}^{2,i} = \mathbf{R}^2 + \varepsilon_i * r^2 + \mathbf{R}^1_i$ /* Broadcasting operation */

Create virtual interaction geometry g_{vr}

$\mathbf{R}_{\text{vr}} = (\mathbf{R}^1 \| \mathbf{R}^{2,1} \| \ldots \| \mathbf{R}^{2,i} \| \ldots \| \mathbf{R}^{2,n})$

$\mathbf{X}_{\text{vr}} = (\mathbf{X}^1 \| \mathbf{X}^2 \| \ldots \| \mathbf{X}^2)$

$g_{\text{vr}} = (\mathbf{X}_{\text{vr}}, \mathbf{R}_{\text{vr}})$

return g_{vr}

Function INTERACTION GEOMETRY CONTRASTIVE LOSS $(\mathbf{z}_{\text{2D}}, \mathbf{z}_{\text{3D}})$

return $\mathcal{L}_{\text{cont}} = -\frac{1}{N_{\text{batch}}}\sum_{i=1}^{N_{\text{batch}}}\left[\log\frac{e^{\mathtt{sim}(\mathbf{z}_{\text{2D},i},\mathbf{z}_{\text{3D},i})/\tau}}{\sum_{k=1}^{N_{\text{batch}}} e^{\mathtt{sim}(\mathbf{z}_{\text{2D},i},\mathbf{z}_{\text{3D},k})/\tau}} + \log\frac{e^{\mathtt{sim}(\mathbf{z}_{\text{3D},i},\mathbf{z}_{\text{2D},i})/\tau}}{\sum_{k=1}^{N_{\text{batch}}} e^{\mathtt{sim}(\mathbf{z}_{\text{3D},i},\mathbf{z}_{\text{2D},k})/\tau}}\right]$

Function INTERMOLECULAR FORCE PREDICTION LOSS $(g^{2,i}_{\text{3D}}, \mathbf{H}^1, \mathbf{H}^2)$

For all edges (k,l) **in** $g^{2,i}_{\text{3D}}$:

$\mathcal{F}_{k,l} = \left(\frac{\mathbf{r}_k - \mathbf{r}_l}{||\mathbf{r}_k - \mathbf{r}_l||}, \frac{\mathbf{r}_k \times \mathbf{r}_l}{||\mathbf{r}_k \times \mathbf{r}_l||}, \frac{\mathbf{r}_k - \mathbf{r}_l}{||\mathbf{r}_k - \mathbf{r}_l||} \times \frac{\mathbf{r}_k \times \mathbf{r}_l}{||\mathbf{r}_k \times \mathbf{r}_l||}\right)$, /* Construct Orthogonal Frame */

where $\mathbf{r}_k \in \mathbb{R}^3$ indicates the position of atoms k.

$\mathbf{e}^{k,l}_{\text{3D}} = \text{Projection}_{\mathcal{F}_{k,l}}(\mathbf{r}_k, \mathbf{r}_l)$ /* Convert to $SE(3)$-Invariant Feature */

$\mathbf{e}^{k,l}_{\text{2D}} = \text{MLP}(\mathbf{H}^2_k || \mathbf{H}^2_l)$

$\mathbf{e}_{k,l} = \mathbf{e}^{k,l}_{\text{2D}} + \mathbf{e}^{k,l}_{\text{3D}}$.

$\tilde{\mathbf{X}} = (\mathbf{H}^2 || \mathbf{H}^1_i)$ /* Broadcasting operation */

$\mathbf{h}_{k,l} = \text{GNN}(\tilde{\mathbf{X}}, \mathcal{E})$, where $\mathcal{E}$ indicates all edges in $g^{2,i}_{\text{3D}}$ /* Obtain Edge Features */

$\hat{f}_k = \sum_l \mathbf{h}_{k,l} \odot \mathcal{F}_{k,l}$ /* Convert to $SE(3)$-equivariant Feature */

return $\mathcal{L}_{\text{force}} = \frac{1}{N^2}\sum_{k=1}^{N^2} ||f^i_k - \hat{f}^i_k||^2_2$