Self-supervised Blending Structural Context of Visual Molecules for Robust Drug Interaction Prediction

Tengfei Ma[*] **Kun Chen**[*] **Yongsheng Zang** **Yujie Chen** **Xuanbai Ren** **Bosheng Song**
Hongxin Xiang **Yiping Liu** **Xiangxiang Zeng**[†]
State Key Laboratory of Chemo and Biosensing, College of Computer Science and
Electronic Engineering, Hunan University, Changsha 410082, China
The Ministry of Education Key Laboratory of Fusion Computing of Supercomputing
and Artificial Intelligence, Hunan University, Changsha 410082, China.
tfma@hnu.edu.cn, xzeng@hnu.edu.cn

Abstract

Identifying drug-drug interactions (DDIs) is critical for ensuring drug safety and advancing drug development, a topic that has garnered significant research interest. While existing methods have made considerable progress, approaches relying solely on known DDIs face a key challenge when applied to drugs with limited data (e.g., novel and few-shot drugs): *insufficient exploration of the space of unlabeled pairwise drugs*. To address these issues, we innovatively introduce S^{2}VM, a **S**elf-**s**upervised **V**isual pretraining framework for pair-wise **M**olecules, to fully fuse structural representations and explore the space of drug pairs for DDI prediction. S^{2}VM incorporates the explicit structure and correlations of visual molecules, such as the positional relationships and connectivity between functional substructures. Specifically, we blend the visual fragments of drug pairs into a unified input for joint encoding and then recover molecule-specific visual information for each drug individually. This approach integrates fine-grained structural representations from unlabeled drug pair data. By using visual fragments as anchors, S^{2}VM effectively captures the spatial information of local molecular components within visual molecules, resulting in more comprehensive embeddings of drug pairs. Experimental results show that S^{2}VM achieves state-of-the-art performance on widely used benchmarks, with Macro-F1 score improvements of 4.21% and 3.31%, respectively. Further extensive results and theoretical analysis demonstrate the effectiveness of S^{2}VM for both few-shot and novel drugs. The code and data are available at https://github.com/xiaomingaaa/S2VM.

1 Introduction

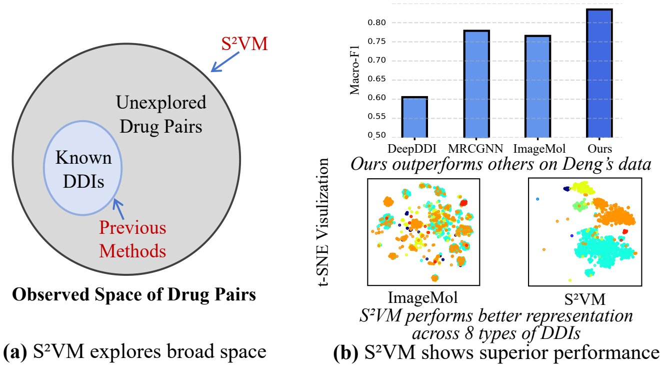
Combinatorial therapy, which involves the simultaneous use of multiple drugs, is a promising strategy for treating patients with complex diseases [1, 2]. However, this approach poses challenges due to potential drug-drug interactions (DDIs) that can alter the intended therapeutic outcomes. When patients take multiple drugs at the same time, these interactions can result in unexpected side effects or diminished clinical efficacy [3, 4]. Therefore, accurately predicting DDIs is essential to avoid potential adverse effects, making it a critical task in the common therapeutic field [5]. Despite ongoing efforts, predicting these interactions remains a significant challenge.

Numerous computational prediction methods have been developed to address these challenges to predict unknown drug-drug interaction (DDI) events [6, 4, 7]. Many of these methods use handcrafted

[*] Equal contribution
[†] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

features (e.g., molecule structure [8, 9, 1], side effects [10], and phenotypic similarity [11]) to represent each drug for predicting potential DDIs. However, these methods rely heavily on expert domain knowledge to design these features accurately. To address this, some approaches use deep learning models to extract low-dimensional features from molecular sequences, learning representations from SMILES in an end-to-end manner [12, 13]. Additionally, to represent drug structures from a functional perspective, several works [14, 15, 16] extract molecular substructures and employ graph neural networks to model the associations between drug pairs, resulting in promising predictive performance. However, they primarily focus on molecular features, neglecting other biological entities involved in drug interaction events, such as proteins, pathways, and diseases, which are crucial to identifying DDIs. Recent works [17, 18, 19, 20, 21, 22, 23] have taken advantage of the semantic relations and topological structures of biomedical knowledge graphs to improve the structural representation of molecules for accurate prediction of DDI. While these methods have achieved some improvements, they primarily predict unknown DDIs by learning drug representations from known DDIs, which are limited to novel and few-shot drugs due to the challenge: *limited exploration for the space of drug pairs from huge unlabeled data*. As illustrated in Figure 1, previous methods mainly represent drug pairs by concatenating the molecular embeddings from individual drug encoders, which were trained on existing DDIs (Figure 1a), resulting in weak structural fusion and exploration capabilities for broad unknown drug pairs. We provide a more detailed discussion in Appendix C.1.



(a) S^2VM explores broad space

(b) S^2VM shows superior performance

Figure 1: (a) S^2VM explores a comprehensive space of drug pairs for existing drugs. (b) The self-supervised S^2VM shows superior performance and representations.

To address these limitations, we propose a self-supervised pretraining framework (called S^2VM) to learning from over 200M drug pairs, designed to encode input drugs jointly by capturing both intrinsic structures and extrinsic interactions between molecules. Specifically, S^2VM first samples and blends input drugs based on their local visual fragments for joint encoding by the drug encoder. Then, S^2VM introduces a decoder to reconstruct the original visual structure of input molecules from the blended representation. This reconstruction process establishes structural correlations between input drugs using molecular visual information in a self-supervised manner. The pretrained encoder is subsequently adopted for DDI prediction. Empirical observations (Figure 1b) and theoretical (Section 4.2) analysis indicate that S^2VM is designed for effective structural representation of drug pairs, exhibiting superior exploration capabilities compared to the visually pretrained molecule representation model ImageMol [24]. Our contributions include: (1) To the best of our knowledge, we are the first to develop a self-supervised pretraining model based on large-scale unlabeled drug pairs that jointly encodes the visual structural relations of drug pairs for DDI prediction. (2) By representing the blended visual fragments of observed paired molecules and recovering their original visual structures, S^2VM effectively captures extrinsic relations and intrinsic structures between molecules from both experimental and theoretical perspectives. (3) Through theoretical analysis and empirical validation, we demonstrate that S^2VM effectively integrates visual structural relationships across diverse drug pairs, achieving state-of-the-art performance in DDI prediction under various scenarios.

2 Related Work

Drug Interaction Prediction. Identifying potential drug interaction events is crucial to drug discovery. Some works mainly adopt handcraft features of molecules to predict unknown DDIs [8, 9]. However, these handcraft features are limited by reliance on domain knowledge of drugs [25, 17], suffering from low expressive ability [26]. DeepDDI [12] and CASTER [13] utilize deep learning models to mine low-dimensional representations of drugs and predict the interaction associations between input drug pairs. Further, SSI-DDI [14], SA-DDI [27], and DSN-DDI [16] proposed substructure-based GNN and fused the representation of molecules based on substructures adaptively. However, these methods overlook the drug-related knowledge from biomedical networks [20, 17]. To model the structure of molecules and the interactive information of drugs, MUFFIN [18] and SumGNN [19] adopt GNN [28] to represent the molecular structure and the relational semantics of the biomedical knowledge graph. To further represent the interactive association between drugs, MRCGNN [21] and TIGER [22] utilized a shared encoder with a contrastive learning mechanism to integrate the structural information of molecules and the information of multi-relational DDI events. However, they learn separate drug inputs from known DDIs, limited by modeling the structural relations between them. We innovatively designed a self-supervised pretraining framework to introduce a unified model to represent huge unlabeled drug pairs jointly.

Representation of Visual Molecules. The molecular images are intuitive in representing spatial information such as the positional relations and connectivity between functional substructures. Some researchers consider representing molecules as images and adopting computer vision techniques to extract features for chemical properties prediction [29, 30]. To effectively understand the structural information of visual molecules, ImageMol [24] proposed a pretraining model based on molecular images to learn representation from 10 million molecules. To enhance the image representation of visual molecules, CGIP [31] is further proposed to model the molecular graphs and images in a contrastive manner. Although vision-based molecular representation has shown excellent performance, these methods are limited to modeling the paired molecules simultaneously. In this paper, we design a novel architecture to pretrain a unified encoder for representing the paired drugs.

3 Preliminaries

Background. In this paper, we use RDKit to convert molecular SMILES [32] into visual molecules (i.e., 2D images). The molecular images contain more spatial information (e.g., the positional relationships between functional groups and atoms), which are intuitive and informative for representing molecules. Therefore, we consider molecular images as our inputs.

Structure-level Visual Pretraining. In order to explore the structure-level representation fusion of drugs, we design a self-supervised pretraining framework based on visual fragments. Specifically, the framework is in an encoder-decoder architecture, which contains a transformer-based encoder $\mathcal{F}$ and decoder $\hat{\mathcal{F}}$. Given a pair of drugs (d_u, d_v), where $d_u \in \mathbb{R}^{H \times W \times C}$ and $d_v \in \mathbb{R}^{H \times W \times C}$ are the visual molecules converted by RDKit, our goal is to learn the structural fusion embedding $\hat{\mathrm{e}}_{uv}$ as follows:

$$\begin{aligned} \mathbf{z} &= \mathcal{F}(blend(d_u, d_v)); \hat{d}_u, \hat{d}_v = \hat{\mathcal{F}}(\mathbf{z}); \\ \hat{\mathbf{z}} &= \arg\min_{\mathbf{z}} \Delta(d_u, \hat{d}_u) + \Delta(d_v, \hat{d}_v), \end{aligned} \tag{1}$$

where (H, W, C) is the resolution of input images and the pretrained encoder $\mathcal{F}$ is adopted to embed inputs for downstream DDI prediction.

Problem Definition. We focus on predicting the potential drug interaction events between drugs. The prediction is achieved by blending input drug pairs and fusing the substructure-based representation based on visual molecules. We formulate the visual-based DDI prediction as a multi-classification task, aiming to estimate the probability of corresponding interaction events. Specifically, given a pair of drugs (d_u, d_v), we propose a model to identify the interaction event denoted as $\hat{y}_{(d_u,d_v)} = \Gamma((d_u, d_v)|\Theta, \mathcal{F})$.

4 Method

4.1 Proposed S^{2}VM

Overview. S^{2}VM aims to learn the essential representations of input drugs (i.e., a pair of molecules) that possess inherent connections while appearing distinctive between different molecules. Specifically, S^{2}VM proposes an image-based self-supervised framework to pretrain a unified encoder for representing a pair of molecules. As illustrated in Figure 2, S^{2}VM mainly consists of four components: (a) *Structure-level encoding* module encodes the input drugs into a sequence of visual tokens;

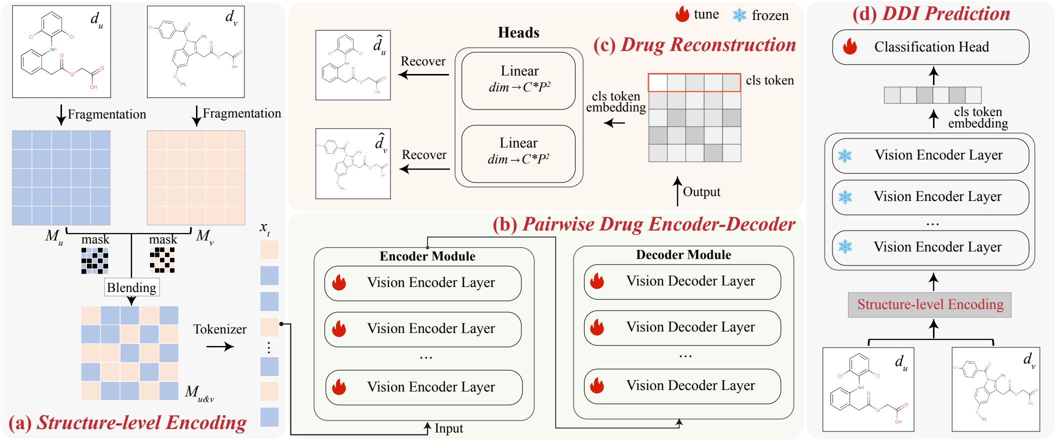


Figure 2: S^{2}VM consists of four components: (a) To fuse the drug pairs into unified input, we sample and blend them into structural tokens (i.e., fragments of visual molecules); (b) We feed the structural tokens into a vision-based Encoder-Decoder to model the semantic relations of molecular fragments; (c) To promote the structural fusion of drug pairs, we set a reconstruction operation to recover the input drugs; (d) The pretrained encoder is adopted to predict potential drug interactions.

(b) The *Pairwise Drug Encoder-decoder* architecture embeds the sequence of visual tokens; (c) We introduce a self-supervised objective to *reconstruct* original molecular images; (d) The pretrained encoder is used to represent a pair of drugs for *DDI prediction*.

Structure-level Encoding. In the context of molecules, the local substructures are the common intrinsic attributes across different molecules. Based on this, we leverage the substructures as anchors, to represent a pair of drugs in a fine-grained manner, blending molecular local structures in the early stage. Specifically, we propose a *Structure-level Encoding* module to blend molecules at the structure level (Figure 2a). Given the input drugs (d_u, d_v), to focus on the local structure of molecules, we split them into a matrix of visual fragments $M_u \in \mathbb{R}^{m \times n}$ and $M_v \in \mathbb{R}^{m \times n}$, where $m = H/P, n = W/P$, (P, P, C) is the resolution of each fragment, and $N = HW/P^2$ indicates the number of fragments. To deeply fuse the structures of input drugs, we design a sampling strategy to blend M_u and M_v into one fused matrix $M_{u\&v} \in \mathbb{R}^N$, which is then fed into a single image encoder for molecule representation. We define a binomial distribution S with a probability vector $p = (p_1, p_2)$. For each fragment of $M^{ij}_{u\&v}(0 \le i < m, 0 \le j < n)$ in the fused matrix, we sample $s^{ij} \in \{1, 2\}$ following the probability distribution p, determining the corresponding element of the blended matrix:

$$M^{ij}_{u\&v} = \begin{cases} M^{ij}_u & \text{if } s^{ij} = 1 \\ M^{ij}_v & \text{otherwise,} \end{cases} \tag{2}$$

where each element in position (i, j) are randomly selected from M^{ij}_u and M^{ij}_v. According to this process, the extrinsic and intrinsic relations of local structures across molecules are blended into a single structure-blended matrix $M_{u\&v}$. We then inject the blended matrix $M_{u\&v}$ into a sequence of visual tokens $x_t \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where C denotes the number of channels. The tokenized sequence x_t is represented by a transformer-based encoder that mines the semantics between the local structures of blended molecules.

Pairwise Drug Encoder-Decoder. We utilize an encoder-decoder architecture to embed the visual tokens x_t into hidden space and decode the latent embedding for the reconstruction of molecular images. To effectively model the semantic relations of local structures within x_t, we apply standard ViT [33] as our encoder $\mathcal{F}$ (i.e., 12 blocks of ViT). Following ViT, we prepend a learnable embedding $x_{cls} \in \mathbb{R}^{1 \times (P^2 \cdot C)}$ to the sequence of embedded tokens x_t, whose state at the output of the encoder $\mathcal{F}$ serves as the representation of input drugs. Specifically, the forward process of the encoder is as follows:

$$\begin{aligned} \mathbf{z}_0 &= [x_{cls}\mathbf{W}; x^1_t\mathbf{W}; x^2_t\mathbf{W}; \ldots; x^N_t\mathbf{W}] + \mathbf{W}^{enc}_{pos}, \\ \mathbf{z}'_l &= \mathrm{MSA}(\mathrm{LayerNorm}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \\ \mathbf{z}_l &= \mathrm{MLP}(\mathrm{LayerNorm}(\mathbf{z}'_l)) + \mathbf{z}'_l, \end{aligned} \tag{3}$$

where the $\mathbf{W} \in \mathbb{R}^{(P^2 \cdot C) \times dim}$ and $\mathbf{W}_{pos}^{enc} \in \mathbb{R}^{(N+1) \times dim}$ are the trainable parameters and positional embedding. MSA represents the multiheaded self-attention. After L layers of iterations, we obtain the fused latent embedding $\mathbf{z}_L$. To further the structural fusion of input drugs, we feed $\mathbf{z}_L$ into a lightweight decoder $\hat{\mathcal{F}}$ (i.e., 4 blocks of ViT) for molecule reconstruction. Similar to the encoder, the reasoning process of decoder $\hat{\mathcal{F}}$ is defined as follows:

$$\begin{aligned} \mathbf{e}_0 &= \mathbf{z}_L + \mathbf{W}_{pos}^{dec}, \\ \mathbf{e}'_l &= \text{MSA}(\text{LayerNorm}(\mathbf{e}_{l-1})) + \mathbf{e}_{l-1}, \\ \mathbf{e}_l &= \text{MLP}(\text{LayerNorm}(\mathbf{e}'_l)) + \mathbf{e}'_l, \end{aligned} \tag{4}$$

where $\mathbf{W}_{pos}^{dec} \in \mathbb{R}^{(N+1) \times dim}$ denotes the positional embedding of the decoder and $\mathbf{e}_l \in \mathbb{R}^{(N+1) \times dim}$ is the decoded representation by $\hat{\mathcal{F}}$. Then the decoded embedding $\mathbf{e}_l$ is input into two heads for image reconstruction, described in the next section. The decoder is only used during pretraining to reconstruct the original molecular images, and the encoder is adopted for the downstream DDI prediction task.

Drug Reconstruction. S^{2}VM introduces a reconstruction objective, recovering the original molecular images d_u and d_v from $\mathbf{e}_l$ by predicting the pixel value of each missing patch within the target molecule. Specifically, we introduce two linear projections to scale the latent embedding $\mathbf{e}_l$, defined as follows:

$$\begin{aligned} \mathbf{h}_u &= \mathbf{e}_l[1: ,]\mathbf{E}_1 + b_1, \\ \mathbf{h}_v &= \mathbf{e}_l[1: ,]\mathbf{E}_2 + b_2, \end{aligned} \tag{5}$$

where $\mathbf{E}_1 \in \mathbb{R}^{dim \times (P^2 \cdot C)}$ and $\mathbf{E}_2 \in \mathbb{R}^{dim \times (P^2 \cdot C)}$ represent learnable parameters. $\mathbf{h}_u \in \mathbb{R}^{N \times (P^2 \cdot C)}$ and $\mathbf{h}_v \in \mathbb{R}^{N \times (P^2 \cdot C)}$ denote the constructed latent embedding, which then is reshaped to form the reconstructed molecular images $\hat{d}_u$ and $\hat{d}_v$. As depicted in Eq. (5), to emphasize the visual information in molecules, we remove the global *[class]* token (i.e., $\mathbf{e}_l[0]$) as $\mathbf{e}_l[1: ,]$. Subsequently, we introduce a mean squared error (MSE) as our loss function to optimize the reconstruction process:

$$\ell_{rec} = \text{MSE}(d_u, \hat{d}_u) + \text{MSE}(d_v, \hat{d}_v). \tag{6}$$

By minimizing ℓ_{rec} during pretraining, we can obtain a unified encoder to represent the paired drugs.

Downstream DDI Prediction and Optimization. We consider the DDI prediction a multi-class classification task. Given a predicted drug pair (d_u, d_v), we use the structure-level encoding module to convert it into a sequence x_t of visual tokens. We then feed them into the pretrained encoder $\mathcal{F}$ and obtain their latent embedding $\mathbf{z}$. To focus on the global representation of the drug pair, we adopt the embedding of *[class]* token $\mathbf{z}[0, :]$ to predict the interaction probability of the given drug pair as follows:

$$\hat{y}_{(d_u, d_v)} = \sigma(\text{MLP}(\mathbf{z}[0, :])), \tag{7}$$

where $\sigma(\cdot)$ is the softmax activation function. We then utilize the cross-entropy loss:

$$\ell_{pre} = -\sum_{k \in \mathcal{K}} \log(\hat{y}^k_{(d_u, d_v)}) y^k_{(d_u, d_v)}, \tag{8}$$

where $\mathcal{K}$ is the number of DDI event types and $y_{(d_u, d_v)}$ represent the ground truth.

4.2 Theoretical Analysis

In this section, we present a theoretical perspective based on mutual information maximization [34, 35] to understand better the effectiveness of S^{2}VM. Given a pair of drugs (d_u, d_v), as described in *Structure-level Encoding* module, they are randomly partitioned into two parts, represented as $d_u = [A_1, A_2]$ and $d_v = [B_1, B_2]$. A_i shares identical indexes of visual fragments with B_i, $i \in \{1, 2\}$ and the blended matrix is denoted as $M_{u\&v} = [A_1, B_2]$.

Proposition 1 (Mutual Information Maximization) *S^{2}VM represents input molecules with structure-level encoding into latent space and then recovers them, maximizing the lower bound of the mutual information:* $\mathbb{E}_S I(A_2; A_1, B_2) + I(B_1; A_1, B_2)$. The proof is detailed in Appendix A.1.

Proposition 2 (Objectives of Pretraining Process) *The mutual information* $I(A_2; A_1, B_2) + I(B_1; A_1, B_2)$ *can be decomposed into* ***extrinsic*** *and* ***intrinsic*** *objectives: (1) contrastive and* *generative between input drugs, and (2) recover missing visual fragments for each molecule* (i.e., **Eq. (9)**). The proof refers to Appendix A.2.

$$\begin{aligned}&\frac{1}{2}[\underbrace{I(A_1;B_1)+I(A_2;B_2)}_{\textbf{contrastive and generative}}+\underbrace{I(A_1;B_1|B_2)+I(A_2;B_2|A_1)}_{\textbf{conditional contrastive and generative}}]\\&+\frac{1}{2}[\underbrace{I(A_1;A_2)+I(B_1;B_2)}_{\textbf{recovery}}+\underbrace{I(A_1;A_2|B_2)+I(B_1;B_2|A_1)}_{\textbf{conditional recovery}}]\end{aligned}\tag{9}$$

Based on the above propositions, we conclude that S^{2}VM has **two strengths** in embedding paired molecules: (i) S^{2}VM using *contrastive and generative* objectives can learn finer-grained associations (e.g., structural interactions between different molecules) from large-scale paired drugs, which improves the generability of molecular representations (i.e., Extrinsic Relations); (ii) S^{2}VM can effectively model the relationship between local structures within a molecule through the *recovery* of missing visual fragments, which helps to enhance the structural representation of the molecule (i.e., Intrinsic Structure). In conclusion, S^{2}VM effectively enhances downstream DDI prediction by modeling paired molecular representation from both external and internal perspectives.

5 Experiments

In this section, to evaluate the effectiveness of S^{2}VM, we carefully consider the following *key research* questions: **Q1**: Does S^{2}VM outperform SOTA baselines on DDI prediction across various scenarios? **Q2**: Are the designed self-supervised pretraining architecture and unified encoder effective? **Q3**: Can S^{2}VM achieve superior performance in new drugs and explore structural mechanisms for DDIs?

5.1 Experimental Settings

Datasets. To evaluate our S^{2}VM, we adopt widely-used datasets: (1) *Deng's dataset* [36] contains 65 types of DDI events with a total of 37,264 DDIs among 570 drugs, (2) *Ryu's dataset* (i.e., DrugBank) [12] includes 86 types of DDI events with a total of 191,570 DDIs between 1,700 drugs, and (3) *TWOSIDES* [37] has 604 drugs and 252,111 for 200 event types. Further, following MRCGNN [21], we count the number of DDI instances involving each DDI event as event frequency and split these DDI events into two groups for few-shot settings (*Few* and *Rare*). We present event types and corresponding proportions in each group in Appendix B.1. The TWOSIDES is adopted to evaluate the performance of S^{2}VM in emerging drugs. Specifically, we adopt two strategies: *S1* setting, determining the interaction type between an emerging drug and an existing drug, and *S2* setting, predicting the interaction type between two new drugs. **For pretraining**, we adopt 200,000 molecules from PubChem to construct $\sim$ 200M pairs of drugs. Refer to Appendix B.1 for details. Each molecule is transformed into a molecular image through a standardized and reproducible pipeline, which serves as the visual input to our model, as detailed in B.1.

Evaluation. Following MRCGNN [21], we split Deng's and Ryu's datasets into training, validation, and test sets with a ratio of 7:1:2, ensuring that each set contains DDI events from all interaction types. We treat the prediction on Deng's and Ryu's datasets as a multi-class classification task, employing Accuracy, Macro-F1, Macro-Recall, and Macro-Precision as our evaluation metrics in both common and few-shot scenarios. In the TWOSIDES dataset under the inductive setting, a drug pair may exhibit multiple interaction types. The task here is to predict whether a specific type of interaction would occur between the paired drugs using a binary classification setting. So we utilize the Accuracy and ROC-AUC metrics on the TWOSIDES datasets. In addition, we select the best model on the validation set based on Macro-F1 for the multi-class classification task and ROC-AUC for the multi-type classification task. Table 1 reports the average results from five runs on the test set.

Implementation Details. For the pretraining process, we set the learning rate $lr = 1.5 \times 10^{-4}$, the number of iterations as $2,000$, the size of the molecular image is $224 \times 224 \times 3$, the size of the visual fragment is $16 \times 16 \times 3$, and the numbers of transformer layers in the encoder and decoder are 12 and 4, respectively. For the downstream DDI prediction task, we set the learning rate $lr = 1 \times 10^{-3}$ and the number of iterations as 100. The image and fragment sizes remain consistent with the pretraining process. The encoder's weights are frozen and utilized to model the representation of DDI pairs. We provide **hyperparameter analysis** in Appendix C. All experiments are conducted on the Linux server with one RTX 3090 (24GB RAM) or RTX 2080Ti (12GB RAM) (refer to Appendix B.2).

Table 1: Results of S^{2}VM and baselines for drug interaction prediction on two datasets. We mark the best score with a bold font and the second best with an underline.

Method	**Deng's dataset**				**Ryu's dataset**			
	ACC.	**Macro-F1**	**Macro-Rec.**	**Macro-Pre.**	**ACC.**	**Macro-F1**	**Macro-Rec.**	**Macro-Pre.**
DeepDDI	78.07	60.55	58.39	66.11	93.23	86.43	85.12	89.28
SSI-DDI	78.66	42.16	38.96	51.39	90.08	66.63	62.87	75.07
MUFFIN	82.69	52.45	48.44	62.04	95.10	85.66	83.39	89.80
KGNN	85.57	72.62	69.87	77.14	92.31	83.77	83.91	89.81
GoGNN	87.66	69.38	68.41	73.16	94.24	85.89	84.51	89.49
MRCGNN	89.79	77.91	76.88	81.01	95.67	88.94	87.27	92.21
CGIP	87.57	76.33	76.41	81.72	93.35	85.72	87.65	88.47
ImageMol	88.75	77.83	76.13	82.72	91.74	87.57	86.62	89.93
CSSE-DDI	82.90	63.46	61.19	70.05	90.90	87.21	85.64	89.82
S^{2}VM	**91.05**	**82.12**	**79.31**	**85.42**	**95.86**	**92.07**	**91.48**	**94.31**
Impr. (%)	↑1.26	↑3.83	↑2.43	↑2.70	↑0.19	↑3.13	↑4.21	↑2.10

Table 2: Results comparison on the few-shot Deng's dataset.

Method	*Few Setting*		*Rare Setting*	
	ACC.	Macro-F1	ACC.	Macro-F1
DeepDDI	47.18	41.91	36.36	31.86
SSI-DDI	64.40	61.73	41.17	38.04
META-DDIE	76.85	74.12	55.13	51.01
MRCGNN	81.89	79.92	47.27	43.75
ImageMol	87.12	89.76	63.69	66.67
S^{2}VM	**91.53**	**91.76**	**68.54**	**73.33**
Impr. (%)	↑2.91	↑2.0	↑3.6	↑6.66

Table 3: Performance comparison on the few-shot Ryu's dataset.

Method	*Few Setting*		*Rare Setting*	
	ACC.	Macro-F1	ACC.	Macro-F1
DeepDDI	65.17	62.32	42.43	37.23
SSI-DDI	72.21	71.15	56.17	52.37
META-DDIE	84.06	79.25	69.58	64.21
MRCGNN	90.16	89.06	66.67	61.21
ImageMol	95.21	91.56	92.72	92.03
S^{2}VM	**99.51**	**95.37**	**99.23**	**98.44**
Impr. (%)	↑4.30	↑3.81	↑6.51	↑6.41

Baselines. To evaluate the performance of S^{2}VM, we compare it with several SOTA methods: the descriptor-based **DeepDDI** [12], the molecular structure-based **SSI-DDI** [14], the biomedical knowledge graph based **KGNN** [17], the molecular substructure together with DDI-related biomedical knowledge **MUFFIN** [18], **GoGNN** [38], **MRCGNN** [21], and **CSSE-DDI** [23], and the image-based molecule representation methods **ImageMol** [24] and **CGIP** [31]. Additionally, we include scenario-specific methods: **META-DDIE** [39] for the few-shot scenario and **STNN-DDI** [40] together with **CSMDDI** [25] for the inductive scenario. Refer to Appendix B.3 for more details.

5.2 Main Results (Q1)

In response to **Q1**, we design various experiments to evaluate S^{2}VM in different scenarios.

Comparison with Baselines. We present the absolute performance gains of S^{2}VM and baselines for predicting DDIs in Table 1. As shown in Table 1, we observe that S^{2}VM achieves the best results in the DDI prediction task on both Deng's and Ryu's datasets. Specifically, S^{2}VM improves the Macro-F1 and Macro-Rec. by at least 4.21% and 2.43% respectively on Deng's dataset, and achieves the 3.13% and 4.21% absolute increase over the best baseline on Ryu's dataset. Furthermore, we have the following observations: (1) Compared with DeepDDI and SSI-DDI, which focus solely on modeling molecular structures, KGNN, which utilizes local semantic relations of drug entities, performs better. This suggests that DDI-related semantics are more effective than molecular structures alone in predicting potential DDIs. (2) Compared with KGNN, MRCGNN, which leverages both the semantic relations of drug interaction networks and molecular structures, achieves better performance. This indicates that integrating DDI-related semantics with molecular structures enhances the prediction task. (3) Compared with MRCGNN, CGIP, and ImageMol, which mine visual information (e.g., positional relations of functional substructures) from separate molecular images, show comparable results on both Deng's and Ryu's datasets. This demonstrates the potential of visual molecular information in predicting unknown DDIs. (4) S^{2}VM, which considers paired drugs as a unified input for joint encoding and explores a wide space of drug pairs using self-supervised pretraining, outperforms all other methods, especially the methods based on semantic relations together with molecular structure. This demonstrates that finer-grained structural fusion and exploration of broad drug pairs can effectively capture the extrinsic and intrinsic associations between drugs.

Few-shot Scenario. To investigate the effectiveness of S^{2}VM on the few-shot DDI prediction task, we design two subsets (See Table 1 in Appendix B.1) with *Few Setting* and *Rare Setting* from Deng's and Ryu's datasets, respectively. The results of S^{2}VM on these few-shot scenarios are presented in Table 2 and Table 3. S^{2}VM consistently outperforms the baselines, achieving a Macro-F1 improvement of 2% and 6.66% in the *Few* and *Rare* settings on Deng's dataset, respectively. Similarly, S^{2}VM increases the Accuracy score by 4.3% and 5.51% in the *Few* and *Rare* settings on Ryu's dataset. Besides, we observe that (1) MRCGNN, which incorporates drug-related semantic relations and molecular structures, outperforms the structure-based methods SSI-DDI and DeepDDI, indicating that the inclusion of biomedical information from knowledge graphs enhances the few-shot DDI prediction task; (2) Image-based methods CGIP and ImageMol perform comparably to methods like MRCGNN, demonstrating that visual information from molecular images is effective for predicting DDIs under limited supervision; (3) S^{2}VM, which unifies paired drugs through structural fusion, achieves the best performance, highlighting that finer-grained structural representations of visual molecules are crucial for identifying unknown DDIs, even with limited interaction information. These findings suggest that S^{2}VM unifying input drugs through structural fusion and self-supervised learning offers a novel and effective perspective for few-shot DDI prediction.

5.3 Ablation Study (Q2)

To investigate the impact of each module in S^{2}VM, we perform an ablation study on Deng's and Ryu's datasets by (1) removing the pretraining process (called **w/o pretrain**) and (2) considering the pretrained encoder shared for encoding input drugs separately (called **w/ shared**). We can observe that all variants perform worse than the S^{2}VM in Figure 3, verifying the effectiveness of S^{2}VM.

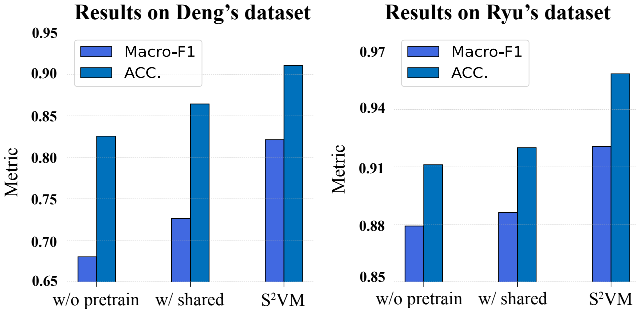


Figure 3: The results on different variants of S^{2}VM.

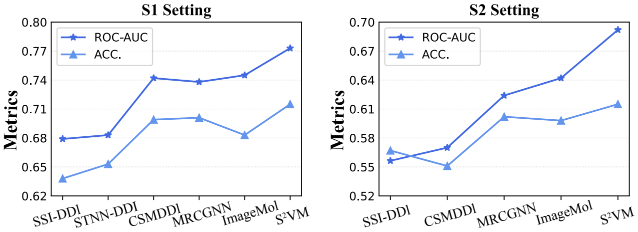


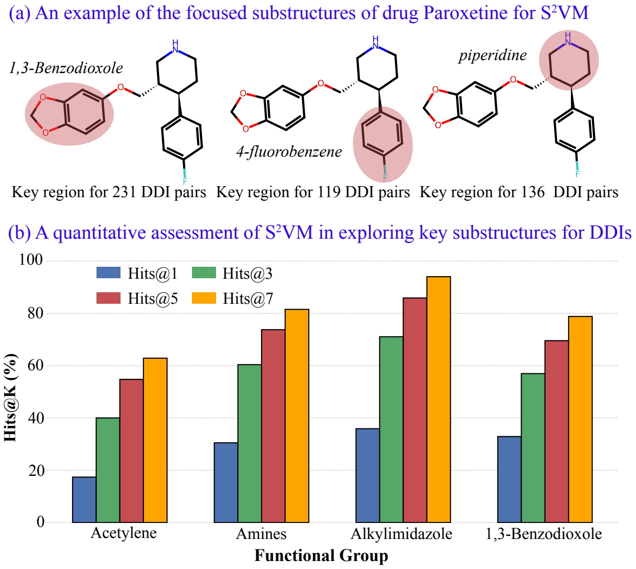
Figure 4: The performance of S^{2}VM based on TWOSIDES on inductive scenarios.

w/o pretrain. We observe a significant reduction in performance across all datasets for DDI prediction after removing the self-supervised learning process for paired drugs. Similarly, we can see that the performance of S^{2}VM without pretraining shows worse results than the variant **w/ shared**, which implies that self-supervised pretraining has positive impacts on representing DDIs for shared encoders. This is because S^{2}VM's use of a self-supervised objective for jointly encoding paired drugs effectively extracts both extrinsic and intrinsic structural correlations between them.

w/ shared. From the results reported in Figure 3, we notice a degradation in performance compared with S^{2}VM on both Deng's and Ryu's datasets. This observation demonstrates representing paired drugs jointly is beneficial in fusing structural visual information from molecular images. The performance reductions observed in **w/o pretrain** and **w/ shared** underscore the importance of mining structural relationships between drug pairs in a self-supervised manner and jointly encoding paired drugs for DDI prediction.

5.4 Effectiveness and Interpretability of S^{2}VM (Q3)

Inductive Scenario. Predicting potential DDIs for new drugs remains a significant challenge. S^{2}VM introduces a self-supervised framework that mines both extrinsic and intrinsic mechanisms of drug interactions, showing potential for predicting unknown DDIs for emerging drugs. To evaluate the prediction ability of S^{2}VM on new drugs, we design inductive experiments for two settings: *S1* and *S2*. As shown in Figure 4, we can see that S^{2}VM performs best on both *S1* and *S2* settings. Specifically, the results indicate that MRCGNN outperforms molecular structure-based methods like SSI-DDI and STNN-DDI, suggesting that incorporating semantic relations can enhance the inductive prediction ability for emerging drugs. Besides, ImageMol shows better performance than previous models, demonstrating that the representations of visual molecules are beneficial to DDI prediction in inductive scenarios. Furthermore, S^{2}VM, by introducing a self-supervised framework that effectively represents paired drugs jointly and explores structural correlations within the broad space of drug pairs, achieves notable improvements in both *S1* and *S2* settings. This suggests that S^{2}VM self-supervised exploration of large-scale observed drug pairs can effectively extract the structural relations between new drugs.


(a) An example of the focused substructures of drug Paroxetine for S^{2}VM

(b) A quantitative assessment of S^{2}VM in exploring key substructures for DDIs

Figure 5: The structure-based explainability of S^{2}VM.

Structural Interpretation for DDI Mechanisms. In DDI prediction, perpetrators can alter the pharmacokinetics (PK) of victim drugs by inducing or inhibiting metabolic enzymes [41]. To evaluate the interpretability of S^{2}VM, we focused on substructures of perpetrator drugs that are reported in the literature to inhibit metabolic enzymes. We utilize a manually curated dataset comprising multiple chemicals known to inhibit metabolic enzymes via specific substructures. Figure 5a demonstrates how S^{2}VM identifies the most salient structural motifs in the drug *Paroxetine* across multiple DDI pairs. Notably, the highlighted substructures in *Paroxetine* correspond to known inhibitors of CYP2D6, such as *1,3-Benzodioxole* [42]. These key fragments were consistently highlighted in DDIs involving *Paroxetine* [43], suggesting mechanistic relevance. To quantitatively assess the ability of S^{2}VM to focus on key substructures, we analyze the predictions that emphasized known inhibitory motifs across 4,543 DDIs involving nine distinct drugs. As shown in Figure 5b, we introduce four metrics to assess the hit rate of the top-weighted substructures by S^{2}VM. The results indicate that the model's top-attended substructures are well-aligned with domain knowledge, underscoring its strong interpretability in identifying biologically meaningful features for DDI prediction. Additional details on the evaluation metrics, cases of highlighted regions, and annotated data are provided in Appendix D.

6 Limitation and Conclusion

Limitation. While S²VM advances DDI prediction, three considerations warrant attention. Real-world DDI distributions are influenced by temporal emergence patterns, therapeutic classes, and toxicity profiles—factors not explicitly modeled into the pretraining stage, which could enhance adverse interaction detection. Second, 2D molecular representations ignore 3D conformational effects, which reflect inherent scalability-granularity trade-offs rather than critical flaws, suggesting future directions. Third, the current framework primarily focuses on structural learning to support new or under-annotated drugs, where biological context is often limited or unavailable, thereby potentially overlooking pharmacological or mechanistic factors underlying DDIs. Compared to knowledge-enhanced models such as MUFFIN , which rely on entity coverage, S^{2}VM achieves up to 32.5% higher accuracy on rare DDIs, demonstrating stronger generalization under low-resource scenarios.

Conclusion. Predicting drug-drug interactions (DDIs) is essential for ensuring patient safety and optimizing therapeutic strategies. However, existing models are often limited by insufficient representation of structural correlations between paired drugs and inadequate exploration of the vast space of potential drug pairs. To address these issues, we propose S^2VM, a self-supervised pretraining framework with a pre-fusion strategy that enhances structural modeling and generalization using over 200 million drug pairs. While S^2VM demonstrates effectiveness, challenges such as computational costs, data diversity, and limited interpretability remain, presenting opportunities for further improvement. Moving forward, we aim to refine the pretrained encoder as a backbone for drug representation and extend its applications to broad drug discovery.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (625B2067; U22A2037; 62425204; 62122025; 62450002; and 62432011). This work was supported by the Beijing Natural Science Foundation (L248013)

References

[1] Santiago Vilar, Eugenio Uriarte, Lourdes Santana, Tal Lorberbaum, George Hripcsak, Carol Friedman, and Nicholas P Tatonetti. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nature Protocols*, 9(9):2147–2163, 2014.

[2] Mukesh Bansal, Jichen Yang, Charles Karan, Michael P Menden, James C Costello, Hao Tang, Guanghua Xiao, Yajuan Li, Jeffrey Allen, Rui Zhong, et al. A community computational challenge to predict the activity of pairs of compounds. *Nature Biotechnology*, 32(12):1213–1222, 2014.

[3] Bo Jin, Haoyu Yang, Cao Xiao, Ping Zhang, Xiaopeng Wei, and Fei Wang. Multitask dyadic prediction and its application in prediction of adverse drug-drug interaction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[4] Yaqing Wang, Zaifei Yang, and Quanming Yao. Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning. *Communications Medicine*, 4(1):59, 2024.

[5] Xuan Lin, Lichang Dai, Yafang Zhou, Zu-Guo Yu, Wen Zhang, Jian-Yu Shi, Dong-Sheng Cao, Li Zeng, Haowen Chen, Bosheng Song, et al. Comprehensive evaluation of deep and graph learning on drug–drug interactions prediction. *Briefings in Bioinformatics*, 24(4):bbad235, 2023.

[6] Yongqi Zhang, Quanming Yao, Ling Yue, Xian Wu, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. Emerging drug interaction prediction enabled by a flow-based graph neural network with biomedical network. *Nature Computational Science*, 3(12):1023–1033, 2023.

[7] Xinyue Li, Zhankun Xiong, Wen Zhang, and Shichao Liu. Deep learning for drug-drug interaction prediction: A comprehensive review. *Quantitative Biology*, 12(1):30–52, 2024.

[8] Takako Takeda, Ming Hao, Tiejun Cheng, Stephen H Bryant, and Yanli Wang. Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. *Journal of Cheminformatics*, 9:1–9, 2017.

[9] Santiago Vilar, Rave Harpaz, Eugenio Uriarte, Lourdes Santana, Raul Rabadan, and Carol Friedman. Drug—drug interaction through molecular structure similarity analysis. *Journal of the American Medical Informatics Association*, 19(6):1066–1074, 2012.

[10] Assaf Gottlieb, Gideon Y Stein, Yoram Oron, Eytan Ruppin, and Roded Sharan. Indi: a computational framework for inferring drug interactions and their associated recommendations. *Molecular Systems Biology*, 8(1):592, 2012.

[11] Peng Li, Chao Huang, Yingxue Fu, Jinan Wang, Ziyin Wu, Jinlong Ru, Chunli Zheng, Zihu Guo, Xuetong Chen, Wei Zhou, et al. Large-scale exploration and analysis of drug combinations. *Bioinformatics*, 31(12):2007–2016, 2015.

[12] 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.

[13] Kexin Huang, Cao Xiao, Trong Hoang, Lucas Glass, and Jimeng Sun. Caster: Predicting drug interactions with chemical substructure representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 702–709, 2020.

[14] Arnold K Nyamabo, Hui Yu, and Jian-Yu Shi. Ssi–ddi: substructure–substructure interactions for drug–drug interaction prediction. *Briefings in Bioinformatics*, 22(6):bbab133, 2021.

[15] Shenggeng Lin, Yanjing Wang, Lingfeng Zhang, Yanyi Chu, Yatong Liu, Yitian Fang, Mingming Jiang, Qiankun Wang, Bowen Zhao, Yi Xiong, et al. Mdf-sa-ddi: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. *Briefings in Bioinformatics*, 23(1):bbab421, 2022.

[16] Zimeng Li, Shichao Zhu, Bin Shao, Xiangxiang Zeng, Tong Wang, and Tie-Yan Liu. Dsn-ddi: an accurate and generalized framework for drug–drug interaction prediction by dual-view representation learning. *Briefings in Bioinformatics*, 24(1):bbac597, 2023.

[17] Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In *IJCAI*, volume 380, pages 2739–2745, 2020.

[18] Yujie Chen, Tengfei Ma, Xixi Yang, Jianmin Wang, Bosheng Song, and Xiangxiang Zeng. Muffin: multi-scale feature fusion for drug–drug interaction prediction. *Bioinformatics*, 37(17): 2651–2658, 2021.

[19] Yue Yu, Kexin Huang, Chao Zhang, Lucas M Glass, Jimeng Sun, and Cao Xiao. Sumgnn: multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics*, 37(18):2988–2995, 2021.

[20] Tengfei Lyu, Jianliang Gao, Ling Tian, Zhao Li, Peng Zhang, and Ji Zhang. Mdnn: A multimodal deep neural network for predicting drug-drug interaction events. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 2021, pages 3536–3542, 2021.

[21] Zhankun Xiong, Shichao Liu, Feng Huang, Ziyan Wang, Xuan Liu, Zhongfei Zhang, and Wen Zhang. Multi-relational contrastive learning graph neural network for drug-drug interaction event prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5339–5347, 2023.

[22] Xiaorui Su, Pengwei Hu, Zhu-Hong You, S Yu Philip, and Lun Hu. Dual-channel learning framework for drug-drug interaction prediction via relation-aware heterogeneous graph transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 249–256, 2024.

[23] Haotong Du, Quanming Yao, Juzheng Zhang, Yang Liu, and Zhen Wang. Customized subgraph selection and encoding for drug-drug interaction prediction. *Advances in Neural Information Processing Systems*, 37:109582–109608, 2024.

[24] Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11): 1004–1016, 2022.

[25] Zun Liu, Xing-Nan Wang, Hui Yu, Jian-Yu Shi, and Wen-Min Dong. Predict multi-type drug–drug interactions in cold start scenario. *BMC Bioinformatics*, 23(1):75, 2022.

[26] Xiaorui Su, Zhuhong You, Deshuang Huang, Lei Wang, Leon Wong, Boya Ji, and Bowei Zhao. Biomedical knowledge graph embedding with capsule network for multi-label drug-drug interaction prediction. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 5640–5651, 2022.

[27] Ziduo Yang, Weihe Zhong, Qiujie Lv, and Calvin Yu-Chian Chen. Learning size-adaptive molecular substructures for explainable drug–drug interaction prediction by substructure-aware graph neural network. *Chemical Science*, 13(29):8693–8703, 2022.

[28] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, pages 1–14, 2016.

[29] Michael Fernandez, Fuqiang Ban, Godwin Woo, Michael Hsing, Takeshi Yamazaki, Eric LeBlanc, Paul S Rennie, William J Welch, and Artem Cherkasov. Toxic colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. *Journal of Chemical Information and Modeling*, 58(8):1533–1543, 2018.

[30] Shifa Zhong, Jiajie Hu, Xiong Yu, and Huichun Zhang. Molecular image-convolutional neural network (cnn) assisted qsar models for predicting contaminant reactivity toward oh radicals: Transfer learning, data augmentation and model interpretation. *Chemical Engineering Journal*, 408:127998, 2021.

[31] Hongxin Xiang, Shuting Jin, Xiangrong Liu, Xiangxiang Zeng, and Li Zeng. Chemical structure-aware molecular image representation learning. *Briefings in Bioinformatics*, 24(6):bbad404, 2023.

[32] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28 (1):31–36, 1988.

[33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, pages 1–21, 2021.

[34] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.

[35] Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. Multimodal molecular pretraining via modality blending. In *International Conference on Learning Representations*, 2024.

[36] Yifan Deng, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, and Shichao Liu. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics*, 36(15): 4316–4322, 2020.

[37] Nicholas P Tatonetti, Patrick P Ye, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125):125ra31–125ra31, 2012.

[38] Hanchen Wang, Defu Lian, Ying Zhang, Lu Qin, and Xuemin Lin. Gognn: Graph of graphs neural network for predicting structured entity interactions. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1317–1323, 2020.

[39] Yifan Deng, Yang Qiu, Xinran Xu, Shichao Liu, Zhongfei Zhang, Shanfeng Zhu, and Wen Zhang. Meta-ddie: predicting drug–drug interaction events with few-shot learning. *Briefings in Bioinformatics*, 23(1):bbab514, 2022.

[40] Hui Yu, ShiYu Zhao, and JianYu Shi. Stnn-ddi: a substructure-aware tensor neural network to predict drug–drug interactions. *Briefings in Bioinformatics*, 23(4):bbac209, 2022.

[41] Jukka Hakkola, Janne Hukkanen, Miia Turpeinen, and Olavi Pelkonen. Inhibition and induction of cyp enzymes in humans: an update. *Archives of Toxicology*, 94(11):3671–3722, 2020.

[42] Zhao Wang, Alisa R Kosheleff, Lilian W Adeojo, Oyinkansola Odebo, Toyin Adewole, Peibing Qin, Vladimir Maletic, Stefan Schwabe, and Azmi Nasser. Impact of paroxetine, a strong cyp2d6 inhibitor, on spn-812 (viloxazine extended-release) pharmacokinetics in healthy adults. *Clinical Pharmacology in Drug Development*, 10(11):1365–1374, 2021.

[43] Suvi TM Orr, Sharon L Ripp, T Eric Ballard, Jaclyn L Henderson, Dennis O Scott, R Scott Obach, Hao Sun, and Amit S Kalgutkar. Mechanism-based inactivation (mbi) of cytochrome p450 enzymes: structure–activity relationships and discovery strategies to mitigate drug–drug interaction risks. *Journal of Medicinal Chemistry*, 55(11):4896–4933, 2012.

[44] MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006.

[45] Tengfei Ma, Cao Xiao, Jiayu Zhou, and Fei Wang. Drug similarity integration through attentive multi-view graph auto-encoders. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 3477–3483, 2018.

[46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[47] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, June 2023.

[48] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[49] Yi Zhong, Gaozheng Li, Ji Yang, Houbing Zheng, Yongqiang Yu, Jiheng Zhang, Heng Luo, Biao Wang, and Zuquan Weng. Learning motif-based graphs for drug–drug interaction prediction via local–global self-attention. *Nature Machine Intelligence*, 6(9):1094–1105, 2024.

Technical Appendices and Supplementary Material

A Theoretical Analysis

In this section, we use uppercase to denote the random variables and lowercase to represent samples of the random variables, followed by the common notations from [44, 35].

A.1 Lemma 1 (Chain rule of mutual information)

Mutual information with conditions follows the law below, i.e.,

$$I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1) \tag{10}$$

Proof.

$$\begin{aligned} I(X_1;Y) + I(X_2;Y|X_1) = E_{p(x_1,y)}\big[\log \frac{p(x_1,y)}{p(x_1)p(y)}\big] + \\ E_{p(x_1,x_2,y)}\big[\log \frac{p(x_2,y|x_1)}{p(x_2|x_1)p(y|x_1)}\big] \\ = E_{p(x_1,x_2,y)}\big[\log \frac{p(x_1,y)}{p(x_1)p(y)} \frac{p(x_2,y|x_1)}{p(x_2|x_1)p(y|x_1)}\big] \\ = E_{p(x_1,x_2,y)}\big[\log \frac{p(x_1,y)p(x_2,y,x_1)}{p(y)p(x_2,x_1)p(y,x_1)}\big] \\ = E_{p(x_1,x_2,y)}\big[\log \frac{p(x_2,y,x_1)}{p(y)p(x_2,x_1)}\big] = I(X_1,X_2;Y) \end{aligned}$$

End Proof.

Based on the above lemma, we can decompose our mutual information $\mathbb{E}_S I(A_2; A_1, B_2) + I(B_1; A_1, B_2)$, described next.

A.2 Proposition 1 (Objectives of Pretrain Process)

As described in the main paper, the $I(A_2; A_1, B_2)+I(B_1; A_1, B_2)$ can be decomposed into following parts:

$$\begin{aligned}\frac{1}{2}\big[I(A_1;B_1) + I(A_2;B_2) + I(A_1;B_1|B_2) + I(A_2;B_2|A_1)\big]+ \\ \frac{1}{2}\big[I(A_1;A_2) + I(B_1;B_2) + I(A_1;A_2|B_2) + I(B_1;B_2|A_1)\big]\end{aligned} \tag{11}$$

Proof. We provide the first term in $I(A_2; A_1, B_2) + I(B_1; A_1, B_2)$, i.e. $I(A_2; A_1, B_2)$. Based on the *Lemma 1*, and let $X_1 = A_1, X_2 = B_2, Y = A_2$, we have:

$$I(A_2; A_1, B_2) = I(A_1; A_2) + I(A_2; B_2|A_1). \tag{12}$$

Also use *Lemma 1* and let $X_1 = B_2, X_2 = A_1, Y = A_2$, then we have:

$$I(A_2; A_1, B_2) = I(B_2; A_2) + I(A_2; A_1|B_2). \tag{13}$$

Based on Eq. (12,13), the $I(A_2; A_1, B_2)$ can be divided into:

$$\frac{1}{2}\Big[I(A_1;A_2) + I(A_2;B_2|A_1) + I(B_2;A_2) + I(A_2;A_1|B_2)\Big]. \tag{14}$$

Similarly, we adopt *Lemma 1* to decompose the second term $I(B_1; A_1, B_2)$:

$$\frac{1}{2}\big[I(B_1;A_1) + I(B_2;B_2|A_1) + I(B_1;B_2) + I(B_1;A_1|B_2)\big]. \tag{15}$$

End Proof.

Table 4: The DDIs division of Deng's and Ryu's datasets.

	Deng's dataset **Few Setting**	 **Rare Setting**
Event range	#39 - #49	#50 - #65
Event Frequency	> 15 and ≤ 50	≤ 15
Event Proportion	1.11%	0.35%
	Ryu's dataset **Few Setting**	 **Rare Setting**
Event range	#64 - #75	#76 - #86
Event Frequency	>15 and <50	≤ 15
Event Proportion	7.42%	4.71%

Table 5: The statistics of TWOSIDES for inductive settings.

	S1 Setting	**S2 Setting**
# Drugs in Train	514	514
# Drugs in Valid	30	30
# Drugs in Test	60	60
# training set	185,673	185,673
# valid set	16,113	467
# test set	45,365	2,466

B Experimental Details

All experiments of S^{2}VM and baseline methods were implemented on a Linux Server with 12 vCPU Intel(R) Xeon(R) Platinum 8255C and one RTX 3090/RTX 2080Ti.

B.1 Datasets.

For few-shot settings, we split the source data into two groups according to their event frequency. As shown in Table 4, we reported detailed event types and their proportions for *few* and *rare* settings. We can observe from Table 4 together with performance in the main paper that S^{2}VM can also achieve superior performance under a few supervised signals. For the inductive scenario, we follow [6] and the detailed drugs and DDIs are reported in Table 5.

In the self-supervised pretraining stage, we build large-scale drug pairs from a set of base drugs. Specifically, we randomly select $200k$ molecules from PubChem[3] as a base set of molecules. Then we randomly sample $2,000$ molecules from PubChem for each molecule of the base set. Based on this, we construct $\sim$ 200M pairs of drugs for pretraining. The detailed data is provided in the anonymous repository: https://anonymous.4open.science/r/S^{2}VM. We conduct more experiments to verify S^{2}VM on different scales of the base set (Appendix C.2).

We generate molecular images through a standardized and reproducible pipeline designed to ensure visual consistency and structural fidelity. All molecules are first canonicalized using RDKit to obtain a unique and deterministic SMILES representation, eliminating variations due to atom ordering or tautomers. The 2D molecular structures are then rendered using RDKit's MolsToGridImage function, explicitly depicting atoms and bonds, with each molecule represented as a 224×224 pixel image without stochastic augmentation to guarantee deterministic and consistent visual representation across runs. Finally, all layout-related parameters, including sub-image spacing, drawing style, and molecule alignment, are fixed to ensure that chemically identical molecules yield identical image representations.

[3] https://drive.google.com/file/d/1t1Ws-wPYPeeuc8f_SGgnfUCVCzlM_jUJ/view?usp=sharing

Table 6: The hyperparameters of S^{2}VM.

	Pretraining	**DDI prediction**
learning rate	1.5×10^{-4}	1×10^{-3}
patch size	16	16
#layers of encoder	12	12
#layers of decoder	4	-
scale	$200k$	-
batch size	512	64
p	$p = (0.5, 0.5)$	-
embedding dim	192	192

Table 7: Performance (Macro-F1 (%)) of predicting DDIs on Deng's dataset for different fusion strategies.

Fusion Strategy	**Feature Operation**	
	Concat	Sum
Post Fusion	56.37	58.94
Pre Fusion	59.73	60.25

B.2 Implementation details of S^{2}VM.

In the pretraining stage, we tune the learning rate among $\{1.5 \times 10^{-1}, 1.5 \times 10^{-2}, 1.5 \times 10^{-3}, 1.5 \times 10^{-4}, 1.5 \times 10^{-5}, 1.5 \times 10^{-6}\}$, the size of visual fragment/patch in $\{8, 16, 32, 48\}$, the number of transformer layers in encoder among $\{4, 8, 12, 16\}$, the number of transformer layers in decoder in $\{2, 4, 6\}$, the scale of training data in $\{50k, 100k, 200k, 300k\}$. Furthermore, we vary the blending probability vector $p = (p_1, p_2)$ into $p = (0.3, 0.7)$, $p = (0.5, 0.5)$, and $p = (0.7, 0.3)$. The $p = (0.5, 0.5)$ is selected finally. For the DDI prediction task, we tune the learning rate in $\{1 \times 10^{-1}, 1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}\}$, the batch size among $\{32, 64, 128, 256\}$. The final hyperparameters are shown in Table 6.

B.3 Implementation details of baselines.

In the common prediction scenario, we implemented KGNN[4], CGIP[5], CSSE-DDI[6], and ImageMol[7] using their official code. The results of other methods MRCGNN, DeepDDI, SSI-DDI, MUFFIN, and GoGNN are from MRCGNN [21]. In the few-shot scenario, we implemented DeepDDI[8], SSI-DDI[9], META-DDIE[10], MRCGNN[11], ImageMol using their source code. In the inductive settings, we implemented SSI-DDI, MRCGNN, and ImageMol based on their available sources. The results of STNN-DDI and CSMDDI were from the method [6]. Note that, for the image-based molecular representation model CGIP and ImageMol, we concat the embeddings of paired drugs and feed it into a 3-layer MLPs for classification. The parameters of CGIP or ImageMol are jointly trained with the classifier.

Table 8: The Macro-F1 (%) performance of S^{2}VM and its variant under inductive scenario on TWOSIDES.

	S1 setting	**S2 setting**
S^{2}VM w/o pretrain	62.33	57.85
S^{2}VM	78.19	69.34

C Additional Experiments

C.1 Motivation Discussion

- **Limited representation of structural correlations between paired drugs.** A major mechanism of drug interactions results from a few local functional substructures instead of the whole chemical substructure [13, 14]. While the remaining substructures are less relevant. Therefore, the structural correlations between drugs are crucial to predict DDIs. To deeply model the structural representation of the whole drug interactions, we adopt a pre-fusion strategy to encode the input drugs jointly. In table 7, we conduct a simple experiment to validate the effectiveness of *pre-fusion*. We introduce two fusion strategies based on molecular morgan fingerprints: (1) *Post Fusion*, concatenating or summing the latent embeddings of a pair of drugs from a 3-layer DNN encoder based on their fingerprint features (2048-dimensional vectors); (2) *Pre Fusion*, previously concatenating or summing the molecular fingerprints of paired drugs as a unified input and then encode the input into a latent embedding using a 3-layer DNN. The experimental settings of the two strategies are the same. As shown in Table 7, we observe that the *pre-fusion* strategy performs better. This phenomenon suggests that direct joint encoding of inputs helps to model drug interactions.
- **Limited exploration for the space of drug pairs.** Previous methods mainly learned the representations of drug pairs from known DDIs, which are limited by the labeled data and generalizability, especially for the new drugs [45, 13]. To address this limitation, we propose a self-supervised pretraining framework learning from over 200M drug pairs to extract comprehensive structural correlations between molecules. To validate the effectiveness of the self-supervised objective, we design a simple experiment on S^{2}VM for the inductive scenario. Specifically, we perform **S^{2}VM** and its variant **S^{2}VM w/o pretrain** on TWOSIDES with S1 and S2 settings. As shown in Table 8, we observe that the S^{2}VM has a significant improvement in predicting DDIs compared with **S^{2}VM w/o pretrain**. This shows that S^{2}VM using self-supervised learning on a broad range of drug pairs has the potential to predict unknown DDIs over emerging drugs. Similarly, in the common scenario depicted in Figure 6, S^{2}VM shows better interaction distribution than others, indicating S^{2}VM is efficient in embedding latent space.

Table 9: The Macro-F1 (%) of S^{2}VM under different probability vector P on Deng's and Ryu's datasets.

$p_1 : p_2$	**Deng's dataset**	**Ryu's dataset**
7:3	81.59	91.76
5:5	**82.97**	**92.53**
3:7	80.87	92.08

[4] https://github.com/xzenglab/KGNN

[5] https://github.com/HongxinXiang/CGIP

[6] https://github.com/LARS-research/CSSE-DDI

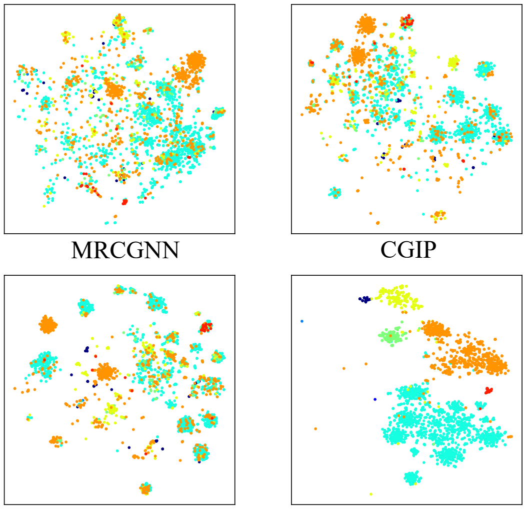
[7] https://github.com/HongxinXiang/ImageMol

[8] https://github.com/deepddi-transfer-learning/deepddi

[9] https://github.com/kanz76/SSI-DDI

[10] https://github.com/YifanDengWHU/META-DDIE

[11] https://github.com/Zhankun-Xiong/MRCGNN



MRCGNN

CGIP

ImageMol

VisualDDI

Figure 6: Distributions of DDI representations from S^{2}VM, ImageMol, CGIP, and MRCGNN across 8 event types.

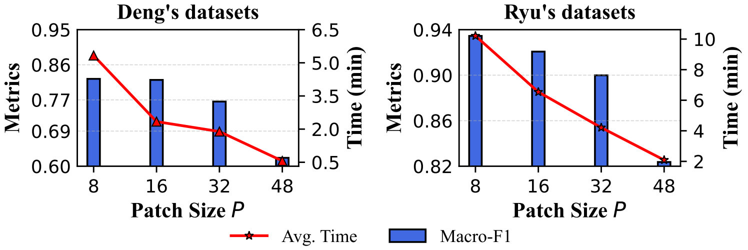


Figure 7: The performance of S^{2}VM and corresponding running time for one step on various patch sizes.

C.2 Representational Distribution of drug pairs.

To explore the representation distributions of drug pairs under different methods, we visualize the embeddings of the paired drugs using the T-SNE [46] tool. Specifically, we randomly select 2890 pair of drugs across 8 event types and then extract their embeddings from MRCGNN, CGIP, ImageMol, and S^{2}VM for visualization. As shown in Figure 6, we can find that S^{2}VM can effectively divide the space for different DDI types, performing best representation distributions.

C.3 Performance on various patch sizes.

We investigate the performance of S^{2}VM on different patch sizes (i.e., the value of P and (P, P, C) is the resolution of each patch/fragment). We vary P across $\{8, 16, 32, 48\}$. The results and corresponding average running time (for on step) are reported in Figure 7. We observe that S^{2}VM achieves best under $P = 8$ but suffers expensive time costs. In contrast, S^{2}VM performs a better balance between the predictive capabilities and time costs when $P = 16$. Meanwhile, the effect decreases as P increases and is accompanied by a low time cost. This is because a larger P reduces the number of tokens and brings about inefficient structural fusion, thus exhibiting high time efficiency and low prediction accuracy. Therefore, we finally select $P = 16$ as our patch size.

C.4 Performance of S^{2}VM on various scales of pretraining data.

To study the performance of S^{2}VM on various scales of pretraining data, we select different numbers ($\{50k, 100k, 200k, 300k\}$) of molecules from PubChem as our base set of drugs. We then randomly construct drug pairs from the base set by sampling 2000 molecules for each drug. The performance

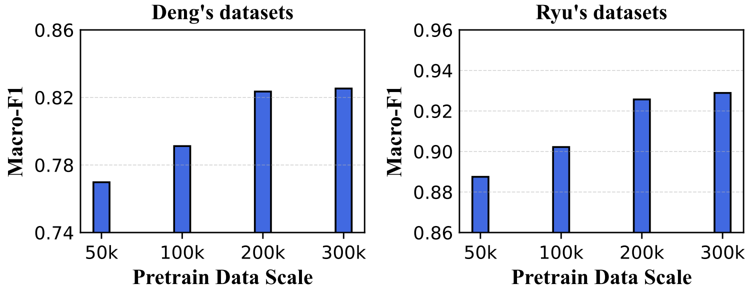
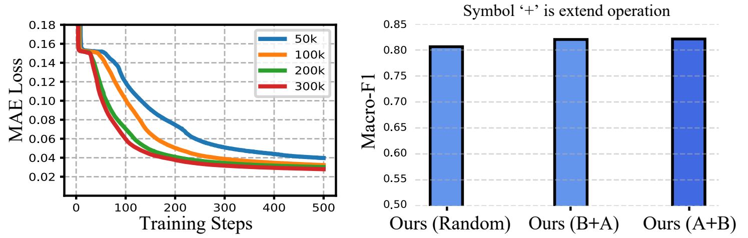


Figure 8: The performance of S^{2}VM under various scales of pretraining data.



(a) The loss curve on various scale **(b)** Performance on various position arrangements

Figure 9: Performance on reconstruction and DDI prediction (Deng's dataset) under different input strategies.

of S^{2}VM under different scales is depicted in Figure 8. As the data size increases, S^{2}VM shows a linear growth trend, while the growth slows down when it reaches $300k$. Meanwhile, the increase in data size can bring more time costs, so we finally consider $200k$ molecules as our base set of drugs.

C.5 Performance of S^{2}VM on different blending ratios.

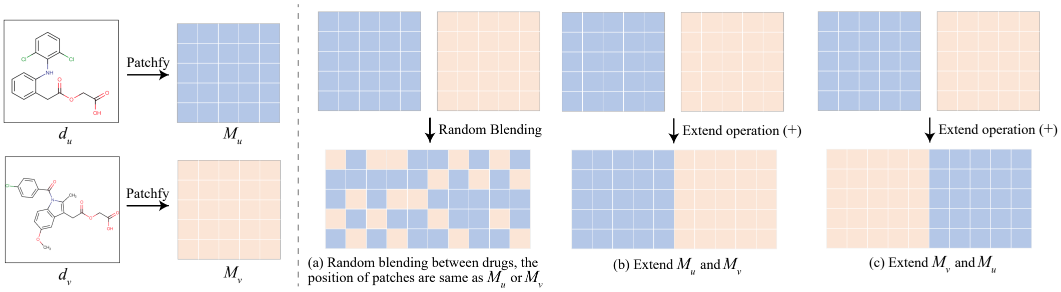
We study the impact of blending ratios for input drugs by varying the probability vector $p = (p_1, p_2)$ for the binomial distribution S. In table 9, we show the performance of S^{2}VM by varying p into $p = (0.3, 0.7)$, $p = (0.5, 0.5)$, and $p = (0.7, 0.3)$. S^{2}VM performs best when $p = (0.5, 0.5)$ and we finally select $p = (0.5, 0.5)$ as our sampling ratio.

C.6 Performance of Reconstruction.

To investigate the performance of molecular image reconstruction, we illustrate the loss curve on various scales of drugs and show two cases of recovery effect. As depicted in Figure 9a, we observe that the curve is smooth and the loss is coverage gradually on various scales of training data, indicating S^{2}VM performs well in this pretraining setting. Further, from the reconstruction cases, we see that S^{2}VM can recover most missing regions of the molecular images. These observations demonstrate that S^{2}VM is effective in capturing the local structural correlations between molecules, enhancing its ability to predict missing regions from the visible ones.

C.7 Position-independent Structural Fusion.

To verify the effectiveness of S^{2}VM in extracting robust structural relations from visual molecules, we experiment with adjusting the arrangements of blending paired drugs (input). Specifically, we design three strategies: (1) randomly blend the tokens (i.e., visual fragments) of molecular images (called **Ours (Random)**), (2) the tokens of the first drug extend tokens of the second one (called **Ours (A+B)**), and (3) the tokens of the second drug extend tokens of the first one (called **Ours (B+A)**). Note that the *Random* indicates that the relative positions of tokens in a molecule are kept the same as the original molecular structure, and the relative positions of tokens between two molecules are random. Refer to Appendix C.8 for more details. We observe that S^{2}VM with the three types of inputs achieves similar results, which demonstrates that S^{2}VM is insensitive to the absolute position of the input fragments. This phenomenon indicates that the positional relations between local fragments



(a) Random blending between drugs, the position of patches are same as M_u or M_v

(b) Extend M_u and M_v

(c) Extend M_v and M_u

Figure 10: Further explanations for different positional strategies.

in visual molecules can be efficiently modeled, which is crucial to representing drug pairs for DDI prediction.

C.8 Details of Position-independent Experiments.

To investigate the impact of different positional arrangements on DDI prediction, we experiment to verify the S^{2}VM. Specifically, we design three strategies: (a) **Ours (Random)**, (b) **Ours (A+B)**, and (c) **Ours (B+A)**. For ease of understanding, we visualize each strategy in Figure 10a, Figure 10b, and Figure 10c, respectively. **Ours (Random)** randomly blends visual fragments (i.e., tokens) of input drugs into a sequence. **Ours (A+B)** splices the fragments of drug A before the fragments of drug B. Similarly, **Ours (B+A)** concatenates the fragments of drug B before the fragments of drug A. This processed matrix will be tokenized into a sequence by column.

C.9 Performance of S^{2}VM under I-JEPA Pretraining

To further validate the robustness and generality of the S^{2}VM framework, we implemented an image-level I-JEPA [47] variant of S^{2}VM, trained under the same settings and scale (50K base molecules, 50M drug pairs) for fair comparison. For the I-JEPA variant, we replaced the ViT backbone in S^{2}VM with the I-JEPA architecture and adapted it to handle paired molecular images. Each drug image is masked and encoded separately to obtain its context features, which are then combined to reconstruct the missing regions of each image following the standard I-JEPA procedure. All default I-JEPA hyperparameters were retained, except for those shared with ViT (e.g., embed_dim=192). As summarized in Table 10, the I-JEPA variant achieves comparable performance on the Ryu dataset and only slightly underperforms on the Deng dataset. These consistent results across two distinct pretraining paradigms demonstrate that S^{2}VM effectively captures transferable molecular representations regardless of the underlying encoder design, while also highlighting the potential of JEPA-style models as a viable alternative for future extensions.

Table 10: Performance Macro-F1 (%) comparison between ViT-based and JEPA-based S^{2}VM models under small-scale pretraining

Method	**Deng's dataset**	**Ryu's dataset**
I-JPEA	71.54	88.78
S^{2}VM	**77.12**	**88.83**

Table 11: Performance comparison between S^{2}VM and single-molecule representation baselines.

Method	**Deng's dataset**		**Ryu's dataset**	
	ACC	**Macro-F1**	**ACC**	**Macro-F1**
MAE	82.76	69.61	92.15	89.20
S^{2}VM	**91.05**	**82.12**	**95.86**	**92.07**

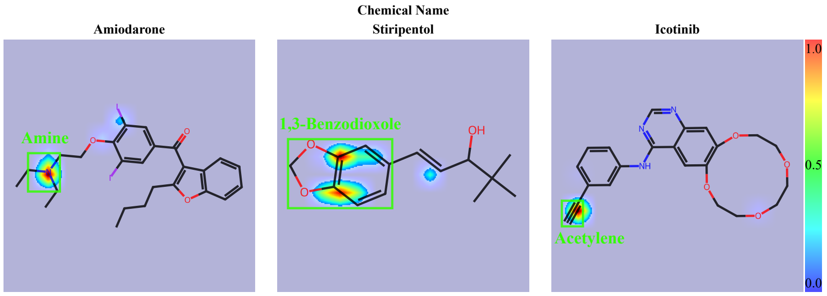


Figure 11: Examples of visualization by using Grad-CAM. The focused regions of S^{2}VM are mainly located in key substructures.

C.10 Performance of S^{2}VM under sing-molecule Pretraining

We conducted the control experiment by pretraining a ViT-based masked autoencoder (MAE) using single-molecule reconstruction only. This baseline uses the same architecture, the same 200K PubChem molecules, and identical training epochs as S^{2}VM, ensuring a fair comparison. During downstream DDI prediction, we adopted a post-fusion strategy where each drug is encoded individually and the resulting representations are concatenated for classification. As shown in Table 11, the single-molecule MAE baseline consistently underperforms S^{2}VM across all metrics and datasets. This demonstrates that joint reconstruction pretraining yields more expressive and interaction-aware representations, beyond what is achievable through standard single-drug encoding.

D Implementation details for interpretable analysis

D.1 Evaluation metrics

To quantifiably evaluate the ability of S^{2}VM in focusing key molecular substructures, we introduce the Hits@K metric. Specifically, the Hits@K is computed by filtering Grad-CAM [48] heatmaps using predefined anchor boxes (Figure 11). Let h_K denote the set of top-$K\%$ hot pixels within the entire molecular image (sorted by the pixel value), B indicate the whole pixels in the box, and T is the set of all pixels within the molecular image. The Hits@K is defined as follows:

$$
\begin{aligned}
Hits &= \sum_{i=1}^{N} f(h_K, B, T), \\
\textbf{Hits@K} &= \frac{Hits}{N}
\end{aligned}
\tag{16}
$$

where $Hits$ denotes the number of hit samples (i.e., S^{2}VM successfully focuses the key functional groups), $K \in \{1, 3, 5, 7\}$ and $f(h_K, B, T)$ is:

$$
f(h_K, B, T) = \begin{cases} 1 & \dfrac{|h_K \cap B|}{|B|} > \dfrac{|B|}{|T|} \times 3 \\ 0 & \text{others} \end{cases}
\tag{17}
$$

The core idea is to assess whether the S^{2}VM's region of interest is concentrated in key substructures.

D.2 Case

As shown in Figure 11, we used Grad-CAM to generate the attention map of S^{2}VM, with the high-attention regions that predominantly influence its predictions. These high-attention regions are mostly concentrated in critical substructures within the molecules, such as Amine, $1, 3$-Benzodioxole, and Acetylene groups, which are key contributors to the occurrence of adverse DDIs.

Table 12: The labeled key structures for DDI mechanisms. The evidence is available at [49]

Chemical Name	DrugBank ID	Functional Group	Enzyme inhibited	Chemical Formula
Paroxetine	DB00715	1,3-Benzdioxole	CYP2D6	$C_{19}H_{20}FNO_3$
Stiripentol	DB09118	1,3-Benzdioxole	CYP2C19||CYP2D6	$C_{14}H_{18}O_3$
Ethinylestradiol	DB00977	Acetylene	CYP3A4	$C_{20}H_{24}O_2$
Gestodene	DB06730	Acetylene	CYP3A4	$C_{21}H_{26}O_2$
Icotinib	DB11737	Acetylene	CYP3A4||CYP3A5	$C_{22}H_{21}N_3O_4$
Midazolam	DB00683	Alkylimidazole	CYP3A4	$C_{18}H_{13}ClFN_3$
Verapamil	DB00661	Amine	CYP3A4	$C_{27}H_{38}N_2O_4$
Troleandomycin	DB13179	Amine	CYP3A4	$C_{41}H_{67}NO_{15}$
Erythromycin	DB00199	Amine	CYP3A4	$C_{37}H_{67}NO_{13}$
Amiodarone	DB01118	Amine	CYP1A2||CYP2C9	$C_{25}H_{29}I_2NO_3$

D.3 Labeled Key Structures for DDI Mechanisms

We show the evaluation data used for the structured interpretation of the DDI mechanism in Table 12. For example, for drug Paroxetine and other drugs that produce adverse DDIs, the main reason is that the functional group 1,3-Benzdioxole in Paroxetine inhibits the drug metabolizing enzyme CYP2D6. Therefore, if S^{2}VM can effectively model these key substructures it will greatly improve the detection efficiency and discover new structuring mechanisms. We adopt a subset of all the labeled data [49] and are available at this link.

E Broader impacts

S^{2}VM's improved DDI prediction could enhance patient safety by reducing adverse drug interactions and accelerate therapeutic development through efficient preclinical screening. However, overreliance on AI predictions without clinical validation might risk misdiagnosis, while data biases could amplify healthcare disparities for underrepresented populations. Additionally, misuse of the model to design harmful drug combinations poses ethical concerns. To mitigate these risks, rigorous validation with pharmacologists, bias-aware data curation, and ethical governance frameworks are critical to ensure transparent and responsible deployment in real-world healthcare systems.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state the claims in the abstract and introduction. The claims match theoretical and experimental results and reflect how much the results can be expected to generalize to other settings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include a limitations paragraph in the Limitation and Conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper includes a theoretical analysis of the self-supervised blending of drug pairs in Section 4.2 and its proofs in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 5.1 and Appendix B detail training settings, datasets, implementation of baselines, architecture parameters, and optimization details. All necessary elements for reproducibility are described. The code and datasets are available.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A reproducible implementation and preprocessing scripts are made available at https://anonymous.4open.science/r/S2VM.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Hyperparameters, data splits, and implementation details are provided in Section 5.1 and Appendices B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report standard deviation across 5 runs for key metrics (Table 1).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are reported in Section 5.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with NeurIPS ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix E discusses the broader impacts of improved DDI detection tools in clinical patients.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The model and data pose no foreseeable risks requiring special safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets used are properly cited and license-compliant.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released codebase includes documentation on the training pipeline and dataset preprocessing.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were used only for editing and formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.