MOLRIME: RECIPROCAL INNOVATION-DRIVEN MULTIMODAL EXCHANGE FOR MOLECULAR REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning transferable molecular representations benefits from integrating complementary information from 2D chemical topology and 3D molecular geometry. Existing molecular pretraining methods exploit this complementarity through cross-modal alignment, knowledge transfer, or representation fusion. However, these approaches typically operate on modality representations as a whole, without distinguishing between information already captured by the other modality and information that has not yet been captured. This highlights a central challenge in cross-modal interaction: determining what information should actually be exchanged between modalities. To address this challenge, we propose Reciprocal Innovation-Driven Multimodal Exchange (RIME), which makes cross-modal communication conditional on what the receiving modality can already explain. RIME uses reciprocal cross-modal prediction to identify prediction residuals as cross-modal innovations, which are dynamically re-estimated and exchanged as the modality representations evolve. Building on RIME, we develop MolRIME, a molecular pretraining framework that realizes innovation-driven 2D–3D exchange through symmetry-preserving updates and combines it with geometric denoising and topology reconstruction. We evaluate MolRIME across molecular dynamics, quantum-chemical, physicochemical, and biological tasks. Controlled frozen-encoder comparisons further demonstrate the benefit of exchanging cross-modal innovations over complete modality representations, with MolRIME reducing test MAE by 6.1-8.5% on four of six evaluated QM9 properties using only 10% of the training labels.

1 INTRODUCTION

Learning transferable molecular representations benefits from integrating complementary information from 2D chemical topology and 3D molecular geometry (Stärk et al., 2022; Zhu et al., 2022). 2D molecular graphs characterize atoms, bonds, and chemical connectivity, whereas 3D conformations provide complementary information about spatial arrangements and interactions relevant to molecular energetics and dynamics (Schütt et al., 2017; Gasteiger et al., 2020). This complementarity makes molecular pretraining a natural setting for jointly leveraging the two structural views from large collections of molecular structures and conformations without task-specific annotations (Stärk et al., 2022; Zhu et al., 2022).

Existing molecular pretraining methods have explored various ways of connecting 2D topology and 3D geometry (Li et al., 2022; Stärk et al., 2022; Zhu et al., 2022). Early approaches establish cross-modal correspondence through contrastive learning or bidirectional generation (Liu et al., 2022; 2023b), while subsequent methods enable richer interaction through relation-level modality blending, unified multi-view pretraining, cross-modal distillation, and progressive multimodal fusion (Yu et al., 2024; Feng et al., 2024; Cho et al., 2025; Jing et al., 2025). Despite these advances, existing methods largely focus on how the two modalities interact, while paying less attention to what information should actually be exchanged (Zang et al., 2026; Chen et al., 2025). In particular, they do not explicitly distinguish between information that is already predictable from the other modality and information that remains unexplained by it. This distinction is especially important for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2D topology and 3D geometry, which contain partially overlapping but non-identical information. Molecular geometry contains substantial information predictable from 2D topology while also providing spatial information unavailable from topology alone (Alhamoud et al., 2024; Hou et al., 2024; Li et al., 2022); conversely, 2D topology explicitly represents chemical connectivity that may not be fully captured by geometric representations (Zhu et al., 2022; Chen et al., 2023). Information already captured by the receiving modality and information that remains unexplained should therefore not necessarily receive equal emphasis in cross-modal communication (Tjandrasuwita et al., 2025; Yang et al., 2026). However, exchanging complete representations makes no such distinction, because the communicated content is determined by the source representation rather than by what the receiver has yet to capture (Barezi & Fung, 2019; Mo et al., 2023). Moreover, as each modality progressively incorporates information from the other, this unexplained component changes over time. Cross-modal exchange should therefore adapt its communicated content to the evolving information gap, rather than repeatedly communicating complete source representations.

To address this challenge, we introduce Reciprocal Innovation-Driven Multimodal Exchange (RIME), a cross-modal exchange mechanism that redefines the content of communication according to what the receiving modality can already explain. Instead of treating the complete source representation as the default unit of exchange, RIME uses reciprocal cross-modal prediction to estimate what is predictable across modalities, with the corresponding prediction residuals capturing what remains unexplained. These residuals define cross-modal innovations, which serve as the actual content exchanged between modalities. This turns cross-modal prediction from an auxiliary learning signal into a mechanism for determining what information should be communicated. Moreover, RIME treats innovations as dynamic rather than fixed, re-estimating them at successive interaction blocks as reciprocal interaction changes what each modality captures. RIME thereby changes the unit of cross-modal communication from complete representations to dynamically estimated innovations, allowing the exchanged content to adapt to what the receiving modality has yet to explain.

Building on RIME, we develop MolRIME, a molecular pretraining framework for innovation-driven exchange between 2D chemical topology and 3D molecular geometry. A further key challenge is that the two modalities have different transformation properties: topological representations are independent of spatial transformations, whereas geometric representations must preserve the appropriate invariance and equivariance. MolRIME addresses this challenge through symmetry-preserving reciprocal innovation exchange, allowing geometry- and topology-derived innovations to enrich the opposite modality without violating geometric equivariance. Beyond cross-modal interaction, MolRIME integrates molecular denoising, topology reconstruction, and cross-modal objectives into a unified pretraining framework. Together, these components enable topology and geometry to mutually enrich their representations while preserving their distinct structural properties.

Our main contributions are summarized as follows:

- We introduce RIME, a cross-modal interaction mechanism that shifts communication from complete source representations to receiver-conditioned innovations. These innovations are defined as reciprocal cross-modal prediction residuals and dynamically re-estimated as the modality representations evolve.
- We develop MolRIME, a molecular pretraining framework that adapts RIME to 2D chemical topology and 3D molecular geometry through symmetry-preserving reciprocal exchange, and combines it with molecular denoising, topology reconstruction, and cross-modal learning objectives.
- We evaluate MolRIME across molecular dynamics, quantum-chemical, physicochemical, and biological tasks. Controlled frozen-encoder comparisons isolate the effect of the exchanged signal and demonstrate the benefit of exchanging cross-modal innovations over complete modality representations.

2 Related Work

2.1 Geometric Molecular Representation Learning

Three-dimensional molecular representation learning incorporates spatial structure beyond chemical connectivity. A central line of work develops symmetry-aware architectures that preserve invari-

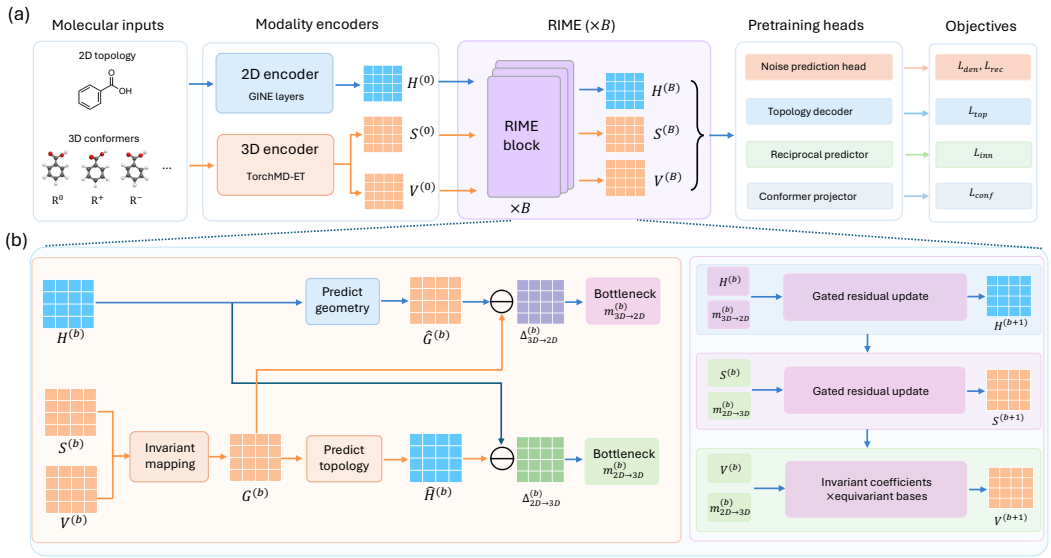

Figure 1: **Overview of MolRIME. (a)** MolRIME encodes molecular topology and geometry into invariant topology features H, invariant geometric features S, and equivariant vector features V, and processes each geometric view independently through a shared stack of B RIME blocks. **(b)** Each RIME block reciprocally predicts one modality from the other and defines the corresponding prediction residuals as cross-modal innovations. After bottleneck projection, the innovations drive sequential updates of topology, invariant geometric, and equivariant vector features, with invariant coefficients modulating equivariant bases to preserve geometric symmetry. Predictions and innovations are recomputed at every block.

ance or equivariance under geometric transformations, including directional and equivariant message passing, equivariant transformers, and joint invariant-equivariant representations (Schütt et al., 2021; Gasteiger et al., 2021; Thölke & Fabritiis, 2022; Liao & Smidt, 2023; Chen et al., 2024). Building on these advances, large-scale pretraining frameworks exploit molecular conformations to learn transferable geometric representations (Zhou et al., 2023), with geometric denoising emerging as an effective self-supervised objective that learns from perturbed conformations by recovering molecular structures or geometric signals (Feng et al., 2023; Cho et al., 2025; Liao et al., 2024). Together, these developments provide the architectural and learning foundations for symmetry-aware molecular representations from 3D geometry.

2.2 2D-3D Multimodal Molecular Pretraining

Recent molecular pretraining methods increasingly integrate 2D topology and 3D geometry to exploit their complementary structural information. Early approaches primarily establish cross-modal correspondence through alignment or generation. GraphMVP combines contrastive and generative objectives between molecular graphs and conformations (Liu et al., 2022), while MoleculeSDE models bidirectional conditional generation using group-symmetric stochastic differential equations (Liu et al., 2023b). Other approaches exploit 2D-3D correspondence through joint or multi-view pretraining objectives (Li et al., 2022; Zhu et al., 2022; Chen et al., 2023; Zhu et al., 2024).

More recent methods enable finer-grained interaction or knowledge transfer between modalities. MoleBLEND blends modality-specific atomic relations for relation-level multimodal learning (Yu et al., 2024), while MolGT promotes cross-modal knowledge sharing through a modality-shared graph Transformer and 2D-3D contrastive learning (Chen et al., 2025). Other approaches transfer geometric knowledge through cross-modal distillation (Cho et al., 2025), support flexible cross-modal integration under missing modalities (Song et al., 2025), or progressively fuse topological and geometric representations through structure-aware interaction (Jing et al., 2025).

These methods provide increasingly sophisticated forms of cross-modal alignment, knowledge transfer, and representation fusion. MolRIME instead focuses on what information should be communicated: reciprocal prediction identifies source information that remains unexplained by the receiving modality, and only this innovation is exchanged. Additionally, the innovations are recomputed after each interaction block, allowing the communicated content to evolve with the modality representations.

3 METHODS

To address cross-modal exchange between 2D chemical topology and 3D molecular geometry, we first formulate the underlying interaction mechanism as Reciprocal Innovation-Driven Multimodal Exchange (RIME), which determines what information should be communicated through reciprocal prediction and innovation-based exchange. We then instantiate RIME for molecular 2D–3D interaction, yielding the MolRIME pretraining framework. This molecular realization must account for the distinct transformation properties of topological and geometric representations, which we address through symmetry-preserving innovation exchange. Details of molecular-view construction, encoder architectures, auxiliary objectives, and optimization are provided in Appendix A.

3.1 RECIPROCAL INNOVATION-DRIVEN MULTIMODAL EXCHANGE

Consider two aligned modality representations $X^{(b)}, Y^{(b)} \in \mathbb{R}^{n\times d}$ at interaction block b, where the alignment may correspond to tokens, spatial locations, or entities such as atoms. Conventional bidirectional exchange directly communicates the source representations $X^{(b)}$ and $Y^{(b)}$. The exchanged content is therefore defined by the source alone without accounting for what the receiving modality can already explain.

RIME instead makes cross-modal exchange conditional on the current state of the receiver. Each modality predicts the representation of the other through a cross-modal predictor p, and the discrepancy between the observed and predicted representations defines the corresponding cross-modal innovation:

$$\widehat{Y}^{(b)} = p^{(b)}_{X\to Y}(X^{(b)}),\ \Delta^{(b)}_{Y\to X} = Y^{(b)} - \widehat{Y}^{(b)}, \widehat{X}^{(b)} = p^{(b)}_{Y\to X}(Y^{(b)}),\ \Delta^{(b)}_{X\to Y} = X^{(b)} - \widehat{X}^{(b)}. \quad (1)$$

Here, $p^{(b)}_{X\to Y}$ predicts the source representation $Y^{(b)}$ from the current receiver representation $X^{(b)}$, thereby estimating what information in $Y^{(b)}$ is already predictable from $X^{(b)}$. The residual $\Delta^{(b)}_{Y\to X}$ then captures what remains unexplained by $X^{(b)}$, with the opposite direction defined symmetrically. These innovations rather than the complete source representations form the content of cross-modal communication.

RIME regulates the resulting innovations through bottlenecked and gated residual updates. Crucially, reciprocal predictions are recomputed after each interaction block as the modality representations evolve. This continual re-estimation allows the innovations to evolve with their receivers as previously unexplained information becomes progressively predictable through cross-modal interaction. RIME thus replaces repeated exchange of complete modality representations with dynamic exchange of receiver-conditioned innovations.

3.2 MOLRIME FRAMEWORK

MolRIME instantiates RIME between atom-aligned 2D topology and 3D geometry while preserving their distinct transformation properties. As illustrated in Figure 1, the framework consists of a topology encoder, an equivariant geometry encoder, a stack of molecular RIME blocks, and objective-specific readouts. We represent each molecule by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N = |\mathcal{V}|$ atoms and atomic identities $Z \in \mathbb{N}^N$, together with a reference conformation $R^0 \in \mathbb{R}^{N\times 3}$. During pretraining, we augment the reference conformation R^0 to obtain alternative conformations R^+, with both serving as parent conformations from which corrupted conformations R^- are generated for geometric denoising. Details of molecular-view construction are provided in Appendix A.1.

Let $\mathcal{R} = \{R_q\}_{q=1}^{J}$ denote the set of J geometric views associated with the same molecular graph. These views include the reference conformation as well as topology-preserving and corrupted conformations. Each view preserves the atom identities and ordering of $\mathcal{G}$ while differing only in its

coordinates. This shared correspondence enables cross-modal interaction at the atom level. To preserve view-specific geometric information, each geometric view interacts with topology independently and cross-view aggregation is deferred until after the RIME stack.

The topology encoder $f_{2\mathrm{D}}$ maps $\mathcal{G}$ to invariant atom representations H, while the geometry encoder $f_{3\mathrm{D}}$ maps each R_q to invariant scalar features S_q and equivariant vector features V_q:

$$H^{(0)} = f_{2\mathrm{D}}(\mathcal{G}), \qquad \left(S_q^{(0)}, V_q^{(0)}\right) = f_{3\mathrm{D}}(Z, R_q), \tag{2}$$

where the superscript (0) denotes representations before cross-modal interaction, with $H^{(0)}, S_q^{(0)} \in \mathbb{R}^{N\times d}$ and $V_q^{(0)} \in \mathbb{R}^{N\times 3\times d}$. Under rigid transformations of the input coordinates, the encoded representations satisfy

$$H'^{(0)} = H^{(0)}, \qquad S_q'^{(0)} = S_q^{(0)}, \qquad V_q'^{(0)} = QV_q^{(0)}, \quad Q \in O(3). \tag{3}$$

Thus, $H^{(0)}$ and $S_q^{(0)}$ are invariant, while $V_q^{(0)}$ is equivariant. For each geometric view, $H^{(0)}$ is paired with $\left(S_q^{(0)}, V_q^{(0)}\right)$ through their shared atom correspondence. Each triplet is processed independently by the shared RIME stack before cross-view aggregation. We omit the view index q below when describing a single view.

3.3 Symmetry-Preserving Molecular RIME Block

Equation 1 assumes that paired representations can be directly compared. In MolRIME, however, $H^{(b)}$ and $S^{(b)}$ are invariant, whereas $V^{(b)}$ is equivariant. Cross-modal innovations must therefore be exchanged while preserving these distinct transformation properties.

Invariant geometry and reciprocal prediction. For $b = 0, \ldots, B-1$, let $\left(H^{(b)}, S^{(b)}, V^{(b)}\right)$ denote the representations entering the current RIME block. To obtain a geometric representation compatible with the invariant topology stream, we combine the scalar features with the channel-wise norms of the vector features, and apply a learnable invariant projection $\phi_{\mathrm{inv}}^{(b)}$ followed by layer normalization LN:

$$U_{i,c}^{(b)} = \left\| V_{i,:,c}^{(b)} \right\|_2, \qquad G^{(b)} = \mathrm{LN}\left(\phi_{\mathrm{inv}}^{(b)}\left([S^{(b)} \parallel U^{(b)}]\right)\right). \tag{4}$$

Here, $i \in \{1, \ldots, N\}$ indexes atoms and $c \in \{1, \ldots, d\}$ indexes feature channels. Since both $S^{(b)}$ and $U^{(b)}$ are invariant, $G^{(b)} \in \mathbb{R}^{N\times d}$ serves as the invariant geometric representation for cross-modal prediction.

RIME then performs reciprocal prediction between the topology and invariant geometry representations:

$$\widehat{G}^{(b)} = \mathrm{LN}\left(p_{2\mathrm{D}\to 3\mathrm{D}}^{(b)}(H^{(b)})\right), \qquad \widehat{H}^{(b)} = \mathrm{LN}\left(p_{3\mathrm{D}\to 2\mathrm{D}}^{(b)}(G^{(b)})\right). \tag{5}$$

The predictors $p_{2\mathrm{D}\to 3\mathrm{D}}^{(b)}$ and $p_{3\mathrm{D}\to 2\mathrm{D}}^{(b)}$ are multilayer perceptrons. Each predictor estimates the component of its target representation that is recoverable from the other modality. The corresponding prediction residuals define the directional cross-modal innovations:

$$\Delta_{3\mathrm{D}\to 2\mathrm{D}}^{(b)} = G^{(b)} - \widehat{G}^{(b)}, \qquad \Delta_{2\mathrm{D}\to 3\mathrm{D}}^{(b)} = H^{(b)} - \widehat{H}^{(b)}. \tag{6}$$

Here, the subscripts of the innovations denote their exchange directions, which are opposite to the corresponding cross-modal prediction directions. The first innovation captures geometric information not explained by the current topology representation, while the second captures topological information not explained by the current geometry representation. Because the predictions are recomputed from the evolving representations at each block, the resulting innovations are dynamically re-estimated throughout the interaction stack.

Bottlenecked reciprocal updates. Each innovation is projected through a low-dimensional bottleneck before being exchanged:

$$M_{u\to v}^{(b)} = \phi_{u\to v,\uparrow}^{(b)}\left(\phi_{u\to v,\downarrow}^{(b)}\left(\Delta_{u\to v}^{(b)}\right)\right), \tag{7}$$

where $(u, v) \in \{(\mathrm{3D}, \mathrm{2D}), (\mathrm{2D}, \mathrm{3D})\}$. The resulting messages are incorporated through adaptive gates:

$$A_H^{(b)} = \sigma\Big(\phi_H^{(b)}\big([H^{(b)} \parallel G^{(b)} \parallel \Delta_{\mathrm{3D}\to\mathrm{2D}}^{(b)}]\big)\Big), \tag{8}$$

$$H^{(b+1)} = \mathrm{LN}\Big(H^{(b)} + \mathrm{Dropout}\Big(A_H^{(b)} \odot M_{\mathrm{3D}\to\mathrm{2D}}^{(b)}\Big)\Big), \tag{9}$$

$$A_S^{(b)} = \sigma\Big(\phi_S^{(b)}\big([S^{(b)} \parallel H^{(b+1)} \parallel G^{(b)}]\big)\Big), \tag{10}$$

$$S^{(b+1)} = \mathrm{LN}\Big(S^{(b)} + \mathrm{Dropout}\Big(A_S^{(b)} \odot M_{\mathrm{2D}\to\mathrm{3D}}^{(b)}\Big)\Big). \tag{11}$$

Here, σ denotes the sigmoid function and $\odot$ denotes element-wise multiplication. The gates control the atom- and channel-wise contribution of each innovation message to the receiving representation. Within each block, the topology representation is updated first with the gated geometry-to-topology message. The resulting $H^{(b+1)}$ then conditions the geometry update, allowing the reverse exchange to depend on the updated topology state.

Equivariant vector update. The topology-to-geometry message is invariant and therefore cannot directly introduce a spatial direction into the geometry stream. Instead, it predicts invariant coefficients that modulate equivariant vector bases:

$$[A_V^{(b)} \parallel A_R^{(b)}] = \sigma\Big(\phi_V^{(b)}\big([S^{(b+1)} \parallel H^{(b+1)} \parallel M_{\mathrm{2D}\to\mathrm{3D}}^{(b)}]\big)\Big). \tag{12}$$

The vector features of atom i are then updated as

$$V_i^{(b+1)} = V_i^{(b)} + V_i^{(b)} W_V^{(b)} \,\mathrm{Diag}\Big(a_{V,i}^{(b)}\Big) + \widetilde{r}_i \Big(a_{R,i}^{(b)}\Big)^{\top}, \tag{13}$$

where $W_V^{(b)} \in \mathbb{R}^{d\times d}$ is a bias-free channel projection, $a_{V,i}^{(b)}$ and $a_{R,i}^{(b)}$ are the i-th rows of $A_V^{(b)}$ and $A_R^{(b)}$, respectively, and $\widetilde{r}_i = r_i - \bar{r}_{g(i)} \in \mathbb{R}^3$ denotes the position of atom i relative to the center of its molecule. The latter two terms in Equation 13 are equivariant because the coefficients are invariant, while $V_i^{(b)} W_V^{(b)}$ and $\widetilde{r}_i$ transform as vectors. In particular, under $r_i' = Qr_i + t$ with $Q \in O(3)$, centering gives $\widetilde{r}_i' = Q\widetilde{r}_i$, and the bias-free channel projection satisfies $(QV_i^{(b)})W_V^{(b)} = Q(V_i^{(b)} W_V^{(b)})$. Hence $V_i'^{(b+1)} = QV_i^{(b+1)}$, so the update preserves translation invariance and $O(3)$-equivariance. A complete derivation is provided in Appendix A.3.

Together, Eqs. (4)-(13) define the symmetry-preserving implementation of a RIME block in MolRIME, which can be summarized as

$$\big(H^{(b+1)}, S^{(b+1)}, V^{(b+1)}\big) = \mathrm{RIME}^{(b)}\big(H^{(b)}, S^{(b)}, V^{(b)}, R\big), \tag{14}$$

where R denotes the coordinates of the current geometric view. Stacking B such blocks re-estimates cross-modal predictions and innovations from the updated representations, so information that becomes predictable contributes less to subsequent innovations while unexplained information can continue to be communicated.

3.4 Representation Readout and Pretraining Objectives

After the final RIME block, each geometric view yields interacted representations $\big(H_q^{(B)}, S_q^{(B)}, V_q^{(B)}\big)$. Rather than collapsing them into a single representation, MolRIME retains their distinct structural forms and applies readouts appropriate to different pretraining signals. The atom-level representations are preserved for local supervision, while an attention-based readout produces invariant scalar and equivariant vector graph features for molecular-level learning. The norm of the vector graph feature further provides an invariant geometric summary.

We pretrain MolRIME with complementary objectives over molecular geometry, topology, cross-modal prediction, and conformational variation. Geometric denoising provides the primary geometric supervision. For atoms in corrupted conformations, the model predicts the normalized coordinate perturbation from $S_q^{(B)}$ and $V_q^{(B)}$:

$$\mathcal{L}_{\mathrm{den}} = \frac{1}{3|\mathcal{A}^-|} \sum_{i\in\mathcal{A}^-} \|\widehat{\epsilon}_{i,\mathrm{norm}} - \epsilon_{i,\mathrm{norm}}\|_2^2, \tag{15}$$

where $\mathcal{A}^-$ denotes the set of atom instances across corrupted views, $\epsilon_{i,\mathrm{norm}}$ denotes the normalized coordinate perturbation applied to atom i, and $\widehat{\epsilon}_{i,\mathrm{norm}}$ is its prediction. We additionally employ coordinate recovery $\mathcal{L}_{\mathrm{rec}}$ and masked atom and bond reconstruction $\mathcal{L}_{\mathrm{top}}$ to supervise geometric structure and local chemical information, respectively.

To train the reciprocal predictors that define the cross-modal innovations, we apply a stop-gradient cosine objective at every interaction block:

$$\mathcal{L}_{\mathrm{inn}} = \frac{1}{2B} \sum_{b=0}^{B-1} \left[2 - \cos\Big(\widehat{G}^{(b)}, \mathrm{sg}(G^{(b)})\Big) - \cos\Big(\widehat{H}^{(b)}, \mathrm{sg}(H^{(b)})\Big) \right], \tag{16}$$

where sg denotes stop-gradient and $\cos(\cdot,\cdot)$ denotes atom-wise cosine similarity averaged over the aligned atoms. Stop-gradient blocks gradients through the prediction targets in this objective, encouraging each predictor to estimate what is recoverable from the other modality without directly adapting the target representation to reduce the prediction loss.

We further apply multi-conformer regularization $\mathcal{L}_{\mathrm{conf}}$ to separate information shared across conformations from conformation-specific variation. A margin constraint prevents the conformation-specific representations from collapsing, while an orthogonality constraint reduces their overlap with the shared representation. Corrupted conformations are excluded from this regularization but remain involved in the other pretraining objectives. Together, the pretraining objectives are combined as

$$\mathcal{L} = \lambda_{\mathrm{den}}\mathcal{L}_{\mathrm{den}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{top}}\mathcal{L}_{\mathrm{top}} + \lambda_{\mathrm{inn}}\mathcal{L}_{\mathrm{inn}} + \lambda_{\mathrm{conf}}\mathcal{L}_{\mathrm{conf}}, \tag{17}$$

where the corresponding λ coefficients denote the loss weights. Complete formulations of the auxiliary objectives and the corresponding weighting coefficients are provided in Appendix A.4.

4 Experiments

4.1 Experimental Setup

Pretraining. We pretrain MolRIME on PCQM4Mv2 (Hu et al., 2021) using the objectives defined in Section 3.4. The checkpoint with the lowest validation pretraining loss is used to initialize all downstream models. The pretraining data, model configuration, objective weights, and optimization settings are detailed in Appendix A.5.

Benchmarks and evaluation. We evaluate MolRIME in three downstream settings. MD22 contains molecular-dynamics trajectories for seven systems ranging from 42 to 370 atoms, providing a test of atomic force prediction across molecular conformations (Chmiela et al., 2023). QM9 comprises approximately 134,000 small organic molecules with quantum-chemical properties computed using density functional theory (Ramakrishnan et al., 2014). For molecular property prediction, we use eight MoleculeNet classification datasets covering bioactivity, toxicity, and drug-related endpoints. We fine-tune a separate model for each MD22 system, QM9 property, and MoleculeNet dataset, reporting force MAE, property MAE, and ROC-AUC, respectively. Dataset splits, fine-tuning hyperparameters, and implementation details are provided in Appendix B.

4.2 Molecular Dynamics and Force Prediction

For MD22, we fine-tune MolRIME separately on each molecular system and evaluate atomic force predictions using MAE. We compare against representative force prediction methods, including MACE (Batatia et al., 2022), ViSNet (Wang et al., 2024), FreeCG (Shao et al., 2025), LiTEN (Su et al., 2026), and the denoising-pretrained Frad (Ni et al., 2024). We report their published force MAEs. Dataset splits and fine-tuning configurations are provided in Appendix B.

As shown in Table 1, MolRIME achieves the lowest force MAE on three of the seven systems: Ac-Ala3-NHMe, DHA, and AT-AT, with errors of 0.035, 0.040, and 0.060, respectively. Relative to the best competing result on each system, these correspond to error reductions of 34.0%, 21.6%, and 4.8%, respectively. MolRIME does not achieve the lowest error on the remaining four systems, indicating that its gains on MD22 are system dependent rather than uniform.

4.3 QUANTUM-CHEMICAL PROPERTY PREDICTION

For a broad comparison, we include representative molecular representation learning methods spanning multimodal pretraining and geometry-based self-supervision, including GraphMVP (Liu et al., 2022), MoleculeSDE (Liu et al., 2023a), Frad (Feng et al., 2023), SubGDiff (Zhang et al., 2024), MolCL-SP (Luo & Deng, 2025), and MolSpectra (Liu et al., 2025). Performance is evaluated using mean absolute error (MAE), with the unit of each target reported in Table 2. Further experimental details are provided in Appendix B.

As shown in Table 2, MolRIME achieves the lowest MAE on U_0 and H, with errors of 5.20 and 5.43 meV, respectively, and ties MolSpectra for the lowest LUMO error at 13.10 meV. Relative to the best competing result, MolRIME reduces the error by 2.4% on U_0 and 2.2% on H. It also achieves a HOMO MAE of 15.50 meV, close to the lowest error of 15.30 meV obtained by Frad. On the remaining properties, MolRIME remains competitive but does not consistently achieve the lowest error, indicating that its gains on QM9 are property dependent rather than uniform.

Table 1: Force prediction on MD22 across seven molecular systems. Force MAE is reported in $\mathrm{kcal\,mol^{-1}\,\AA^{-1}}$. Baseline results are from the corresponding publications; the best result for each system is highlighted in **bold**.

Method	Ac-Ala3-NHMe ↓	DHA ↓	AT-AT ↓	Stachyose ↓	AT-AT-CG-CG ↓	Buckyball ↓	Double-walled ↓
MACE	0.088	0.065	0.099	0.088	**0.115**	**0.085**	0.277
ViSNet	0.083	0.060	0.081	0.088	0.148	0.184	0.362
Frad	0.073	0.078	0.233	0.231	0.308	0.430	0.729
FreeCG	0.053	0.051	0.063	0.612	0.125	0.178	0.245
LiTEN	0.076	0.055	0.092	**0.062**	0.149	0.114	**0.214**
MolRIME	**0.035**	**0.040**	**0.060**	0.125	0.145	0.214	0.463

Table 2: Quantum-chemical property prediction on QM9 across 12 molecular properties. Performance is measured by MAE, with units specified in the column headers. Baseline results are from the corresponding publications, and the best result for each property is highlighted in **bold**.

Method	μ (D) ↓	α (a_0^3) ↓	HOMO (meV) ↓	LUMO (meV) ↓	Gap (meV) ↓	R^2 (a_0^2) ↓	ZPVE (meV) ↓	U_0 (meV) ↓	U (meV) ↓	H (meV) ↓	G (meV) ↓	C_v ($\mathrm{cal\,mol^{-1}\,K^{-1}}$) ↓
GraphMVP	0.0270	0.056	25.75	21.58	41.99	0.136	1.609	13.07	13.03	13.31	13.43	0.029
MoleculeSDE	0.0260	0.054	25.74	21.41	41.77	0.151	1.587	12.04	12.54	12.05	13.07	0.028
Frad	**0.0100**	**0.037**	**15.30**	13.70	27.80	0.342	**1.420**	5.33	5.62	5.55	6.19	**0.020**
SubgDiff	0.0270	0.054	25.45	23.75	44.88	**0.110**	1.568	11.25	11.32	11.46	12.03	0.028
MolCL-SP	0.0250	0.052	20.09	20.77	35.68	0.146	1.507	11.32	10.49	11.37	12.76	0.028
MolSpectra	0.0110	0.048	15.50	**13.10**	**26.80**	0.410	1.710	5.67	**5.45**	5.87	**6.18**	0.021
MolRIME	0.0117	0.044	15.50	**13.10**	28.60	0.212	1.680	**5.20**	5.59	**5.43**	6.79	0.023

Table 3: Molecular property classification on eight MoleculeNet datasets using scaffold splits. Performance is reported as test ROC-AUC (%, mean ± standard deviation) over three seeds. Baseline results are from the corresponding publications. The best result is highlighted in **bold**.

Method ↑	BBBP ↑	Tox21 ↑	MUV ↑	BACE ↑	ToxCast ↑	SIDER ↑	ClinTox ↑	HIV ↑
MoleculeSDE (VE)	$\mathbf{73.2 \pm 0.5}$	76.5 ± 0.3	79.9 ± 0.2	80.4 ± 0.9	65.2 ± 0.3	59.6 ± 0.8	86.6 ± 3.7	78.5 ± 0.3
MoleculeSTM (Graph)	70.0 ± 0.5	76.9 ± 0.5	73.4 ± 3.0	80.8 ± 1.3	65.1 ± 0.4	61.0 ± 1.1	$\mathbf{92.5 \pm 1.1}$	76.9 ± 1.8
M2UMol	68.2 ± 0.6	76.2 ± 0.1	$\mathbf{80.5 \pm 0.8}$	84.3 ± 1.4	66.3 ± 0.5	$\mathbf{66.0 \pm 0.9}$	92.2 ± 0.5	$\mathbf{79.0 \pm 1.0}$
MolRIME	69.3 ± 0.4	$\mathbf{77.0 \pm 0.3}$	74.9 ± 1.5	$\mathbf{84.7 \pm 1.8}$	$\mathbf{66.5 \pm 0.7}$	61.2 ± 1.4	88.3 ± 0.8	78.1 ± 0.8

4.4 MOLECULAR PROPERTY PREDICTION

We evaluate MolRIME on eight MoleculeNet classification datasets using scaffold splits generated following the public M2UMol (Xiong et al., 2026) procedure. We report test ROC-AUC over three training seeds; implementation details are provided in Appendix B.

Table 3 compares MolRIME with published results from representative molecular pretraining methods. MoleculeSDE (Liu et al., 2023b) jointly models topology and geometry, MoleculeSTM combines graphs with text (Liu et al., 2023c), and M2UMol incorporates topology, geometry, text, and

biochemical information during pretraining. MolRIME achieves the highest reported mean ROC-AUC on Tox21, BACE, and ToxCast. Compared with M2UMol, MolRIME performs better on four of the eight datasets (BBBP, Tox21, BACE, and ToxCast) and trails on the other four (MUV, SIDER, ClinTox, and HIV). Because the external baselines were not rerun on our exact partitions, these results constitute a comparison with published scores rather than a matched-split evaluation.

4.5 Ablation Study

We evaluate frozen-encoder transfer under limited labels by comparing MolRIME with Direct Exchange, a separately pretrained variant that exchanges modality representations directly rather than their prediction residuals. The implementation details are provided in Appendix B.2. We select six QM9 targets, R^2, ZPVE, HOMO, α, μ, and G, to cover the different types of molecular properties considered in the benchmark, including structural, electronic, energetic, and thermodynamic properties. For each target, we freeze the pretrained encoder and train the same prediction-head architecture using 1%, 5%, or 10% of the original training split. Both models use the same labeled molecules for each seed, with results averaged over three seeds.

Table 4 reports the relative reduction in mean test MAE of MolRIME over Direct Exchange. MolRIME yields lower mean MAE on R^2, ZPVE, μ, and G across all three label fractions. At 10% labels, the corresponding reductions are 8.5%, 6.1%, 7.9%, and 6.4%. For α, the reduction increases from 0.2% at 1% labels to 4.9% at 10% labels, while the two models perform comparably on HOMO. Since MolRIME and Direct Exchange differ only in the exchanged cross-modal signal, these results suggest that innovation-driven exchange can improve the utility of frozen representations across different QM9 properties, although the magnitude of the improvement varies by target. Further details of the Direct Exchange baseline and the frozen-encoder evaluation setup, together with the absolute MAEs and standard deviations, are provided in Appendix B.2.

Table 4: Relative reduction in test MAE of MolRIME over Direct Exchange for frozen-encoder transfer on QM9. Reductions are computed from the mean MAE over three seeds; positive values indicate better performance of MolRIME.

Property	1% labels	5% labels	10% labels
R^2	$+8.0\%$	$+7.2\%$	$+8.5\%$
ZPVE	$+3.2\%$	$+4.5\%$	$+6.1\%$
HOMO	$+1.7\%$	$+0.2\%$	-0.3%
α	$+0.2\%$	$+3.2\%$	$+4.9\%$
μ	$+2.5\%$	$+6.9\%$	$+7.9\%$
G	$+9.9\%$	$+8.8\%$	$+6.4\%$

4.6 Analysis of Innovation Dynamics

We further examine the internal dynamics of RIME on approximately 2,000 held-out pretraining molecules by tracking cross-modal predictability and prediction residuals across the three interaction blocks. Cross-modal predictability increases substantially from Block 1 to Block 2 and remains high in Block 3, while the normalized prediction residuals decrease in Block 2 and partially increase in Block 3. Atom-level analysis further shows that these changes are not uniform across atoms. Together, these observations are consistent with RIME's block-wise re-estimation of cross-modal predictions and innovations. Detailed definitions, statistics, and visualizations are provided in Appendix B.3.

5 Conclusion

We presented MolRIME, a multimodal molecular representation learning framework that uses cross-modal prediction residuals for reciprocal communication between molecular topology and geometry. Rather than exchanging complete representations, MolRIME exchanges prediction residuals that are re-estimated at each interaction block and incorporated through symmetry-preserving updates. MolRIME achieves competitive performance across molecular property and dynamics benchmarks, while controlled frozen-encoder experiments show that innovation-driven exchange can yield more effective downstream representations than direct representation exchange. These results highlight the exchanged information as an important design choice for multimodal molecular representation learning.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and improving the clarity and presentation of the manuscript. All AI-assisted text was carefully reviewed and revised by the authors. Generative AI was not used to generate experimental results or data. The authors take full responsibility for the final content of this work, including all claims, analyses, and AI-assisted text.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, pretraining objectives, and experimental protocols in the main text and appendices. Implementation and training details, including hyperparameter settings and dataset processing procedures, are provided in the appendix to facilitate reproduction of our results.

REFERENCES

Kumail Alhamoud, Yasir Ghunaim, Abdulelah S Alshehri, Guohao Li, Bernard Ghanem, and Fengqi You. Leveraging 2d molecular graph pretraining for improved 3d conformer generation with graph neural networks. *Computers & Chemical Engineering*, 183:108622, 2024.

Elham J Barezi and Pascale Fung. Modality-based factorization for multimodal fusion. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 260–269, 2019.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gabor Csanyi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436. Curran Associates, Inc., 2022. doi: 10.52202/068431-0830. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4a36c3c51af11ed9f34615b81edb5bbc-Paper-Conference.pdf.

Jiujiu Chen, Xuecang Zhang, Zhi-Ming Ma, Shengchao Liu, et al. Molecule joint auto-encoding: Trajectory pretraining with 2d and 3d diffusion. *Advances in Neural Information Processing Systems*, 36:55077–55096, 2023.

Ruizhe Chen, Chunyan Li, Longyue Wang, Mingquan Liu, Shugao Chen, Jiahao Yang, and Xiangxiang Zeng. Pretraining graph transformer for molecular representation with fusion of multimodal information. *Information Fusion*, 115:102784, 2025.

Tianlang Chen, Shengjie Luo, Di He, Shuxin Zheng, Tie-Yan Liu, and Liwei Wang. Geomformer: A general architecture for geometric molecular representation learning. *arXiv preprint arXiv:2406.16853*, 2024.

Stefan Chmiela, Valentin Vassilev-Galindo, Oliver T Unke, Adil Kabylda, Huziel E Sauceda, Alexandre Tkatchenko, and Klaus-Robert Müller. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9(2):eadf0873, 2023.

Sungjun Cho, Dae-Woong Jeong, Sung Moon Ko, Jinwoo Kim, Sehui Han, Seunghoon Hong, Honglak Lee, and Moontae Lee. 3d denoisers are good 2d teachers: molecular pretraining via denoising and cross-modal distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 110–118, 2025.

Shikun Feng, Yuyan Ni, Yanyan Lan, Zhi-Ming Ma, and Wei-Ying Ma. Fractional denoising for 3d molecular pre-training. In *International Conference on Machine Learning*, pp. 9938–9961. PMLR, 2023.

Shikun Feng, Yuyan Ni, Minghao Li, Yanwen Huang, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. UniCorn: A unified contrastive learning approach for multi-view molecular representation learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13256–13277. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/feng24f.html.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1eWbxStPH.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in neural information processing systems*, 34:6790–6802, 2021.

Xiaoyang Hou, Tian Zhu, Milong Ren, Bo Duan, Chunming Zhang, Dongbo Bu, and Shiwei Sun. Gtam: a molecular pretraining model with geometric triangle awareness. *Bioinformatics*, 40(9): btae524, 2024.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. OGB-LSC: A large-scale challenge for machine learning on graphs. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=qkcLxoC52kL.

Zihao Jing, Yan Sun, Yan Yi Li, Sugitha Janarthanan, Alana Deng, and Pingzhao Hu. Structure-aware fusion with progressive injection for multimodal molecular representation learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Shuangli Li, Jingbo Zhou, Tong Xu, Dejing Dou, and Hui Xiong. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 4541–4549, 2022.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KwmPfARgOTD.

Yi-Lun Liao, Tess Smidt, Muhammed Shuaibi, and Abhishek Das. Generalizing denoising to non-equilibrium structures improves equivariant force fields. *arXiv preprint arXiv:2403.09549*, 2024.

Shaozhen Liu, Yu Rong, Deli Zhao, Qiang Liu, Liang Wang, et al. Molspectra: Pre-training 3d molecular representation with multi-modal energy spectra. In *International Conference on Learning Representations*, volume 2025, pp. 86733–86751, 2025.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xQUe1pOKPam.

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

Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21497–21526. PMLR, 23–29 Jul 2023b.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023c.

Yue Luo and Lei Deng. Molcl-sp: a multimodal contrastive learning framework with non-overlapping substructure perturbations for molecular property prediction. *Bioinformatics*, 41(10): btaf507, 2025.

Yujie Mo, Yajie Lei, Jialie Shen, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. Disentangled multiplex graph representation learning. In *International conference on machine learning*, pp. 24983–25005. PMLR, 2023.

Yuyan Ni, Shikun Feng, Xin Hong, Yuancheng Sun, Wei-Ying Ma, Zhi-Ming Ma, Qiwei Ye, and Yanyan Lan. Pre-training with fractional denoising to enhance molecular property prediction. *Nature Machine Intelligence*, 6(10):1169–1178, 2024.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):140022, 2014.

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

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International conference on machine learning*, pp. 9377–9388. PMLR, 2021.

Shihao Shao, Haoran Geng, Zun Wang, and Qinghua Cui. Freecg: Free the design space of clebsch-gordan transform for machine learning force fields. In *International Conference on Learning Representations*, volume 2025, pp. 5627–5654, 2025.

Tengwei Song, Min Wu, and Yuan Fang. Unified molecule pre-training with flexible 2d and 3d modalities: Single and paired modality integration. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 2750–2760, 2025.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 3D infomax improves GNNs for molecular property prediction. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20479–20502. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/stark22a.html.

Qun Su, Kai Zhu, Qiaolin Gou, Jintu Zhang, Renling Hu, Yurong Li, Yongze Wang, Hui Zhang, Ziyi You, Linlong Jiang, et al. A scalable and quantum-accurate foundation model for biomolecular force fields via linearly tensorized quadrangle attention. *Nature Communications*, 17(1):3639, 2026.

Philipp Thölke and Gianni De Fabritiis. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zNHzqZ9wrRB.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4NJCI4Q3Za.

Yusong Wang, Tong Wang, Shaoning Li, Xinheng He, Mingyu Li, Zun Wang, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15(1):313, 2024.

Zhankun Xiong, Ziyan Wang, Feng Huang, Minyao Qiu, Shuyan Fang, Liuqing Yang, Xionghui Zhou, Shichao Liu, Ping Zhang, and Wen Zhang. Multi-to-uni modal knowledge transfer pretraining for molecular representation learning. *Nature Communications*, 17(1):3797, 2026.

Zequn Yang, Yake Wei, Haotian Ni, Zhihao Xu, and Di Hu. Information-theoretic decomposition for multimodal interaction learning. *arXiv preprint arXiv:2606.11614*, 2026.

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*, volume 2024, pp. 13314–13332, 2024.

Xuan Zang, Junjie Zhang, and Buzhou Tang. Molecular representation learning via multimodal fusion and decoupling. *Information Fusion*, 125:103493, 2026.

Jiying Zhang, Zijing Liu, Yu Wang, Bin Feng, and Yu Li. Subgdiff: A subgraph diffusion model to improve molecular representation learning. *Advances in Neural Information Processing Systems*, 37:29620–29656, 2024.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The eleventh international conference on learning representations*, 2023.

Jinhua Zhu, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Unified 2d and 3d pre-training of molecular representations. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2626–2636, 2022.

Yanqiao Zhu, Dingshuo Chen, Yuanqi Du, Yingze Wang, Qiang Liu, and Shu Wu. Molecular contrastive pretraining with collaborative featurizations. *Journal of Chemical Information and Modeling*, 64(4):1112–1122, 2024.

APPENDIX

A Method Details **15**

A.1 Molecular View Construction 15

A.2 Encoders, View Processing, and Readout 15

A.3 Symmetry Verification 16

A.4 Pretraining Objectives 16

A.5 Pretraining Setup 17

B Downstream Experimental Details **18**

B.1 Data Splits and Evaluation Protocols 18

B.2 Frozen-Encoder Transfer under Limited Labels 19

B.3 Cross-Modal Predictability and Innovation Dynamics 19

A METHOD DETAILS

This appendix provides implementation details of MolRIME, including molecular-view construction, encoder and readout architectures, auxiliary objectives, symmetry analysis, and pretraining configuration.

A.1 MOLECULAR VIEW CONSTRUCTION

Starting from the reference conformation R^0, we construct K alternative conformations by perturbing molecular torsions followed by weak coordinate noise. For the k-th alternative conformation, we select at most two rotatable torsions and independently perturb their angles by values sampled uniformly from $[-10^\circ, 10^\circ]$:

$$\widetilde{R}^{+,k} = \mathcal{T}(R^0; \theta_k), \qquad R^{+,k} = \widetilde{R}^{+,k} + E^{+,k}, \qquad k = 1, \dots, K, \tag{18}$$

where $\mathcal{T}$ denotes the torsional transformation parameterized by θ_k, and $E^{+,k} \in \mathbb{R}^{N\times 3}$ denotes the weak coordinate perturbation. If no rotatable torsion is available, only the coordinate perturbation is applied. The reference and alternative conformations together form the parent set

$$\mathcal{R}_{\mathrm{par}} = \{R^{\mathrm{par},k}\}_{k=0}^{K} = \{R^0, R^{+,1}, \dots, R^{+,K}\}, \qquad R^{\mathrm{par},0} = R^0. \tag{19}$$

Each parent conformation is further perturbed to generate M corrupted conformations:

$$R^{-,k,m} = R^{\mathrm{par},k} + E^{-,k,m}, \qquad k = 0, \dots, K, \quad m = 1, \dots, M, \tag{20}$$

where $E^{-,k,m} \in \mathbb{R}^{N\times 3}$ denotes a structure-aware coordinate perturbation with a larger magnitude than the weak perturbation used to construct R^+. Atoms within the same ring share a ring-level Gaussian displacement and additionally receive atom-level Gaussian noise, while non-ring heavy atoms receive independent Gaussian displacements. Each hydrogen follows the displacement of its attached heavy atom. This preserves atom correspondence across views while introducing correlated geometric perturbations within ring structures.

We use $K = 2$ and $M = 1$, resulting in three parent conformations and one corrupted conformation for each parent.

A.2 ENCODERS, VIEW PROCESSING, AND READOUT

The topology branch embeds the categorical atom and bond features defined by OGB and processes them with a residual GINE encoder, while the geometry branch uses TorchMD-ET to produce invariant scalar features and equivariant vector features. The molecular graph is encoded once to obtain the topology representation $H^{(0)}$, while each geometric view R_q is independently encoded as $(S_q^{(0)}, V_q^{(0)})$. The resulting triplets $(H^{(0)}, S_q^{(0)}, V_q^{(0)})$ are processed independently by the shared RIME stack, with cross-view aggregation performed only after the final interaction block.

After the final RIME block, the view-specific topology representations are mean-aggregated over views for masked topology reconstruction, whereas geometric denoising operates directly on the scalar and vector atom representations of each individual view. This design allows topology reconstruction to integrate information across views while retaining view-specific representations for geometric supervision.

For molecular-level objectives, we construct invariant scalar and equivariant vector readouts for each geometric view. Let

$$a_{q,i} = \phi_{\mathrm{read}}\left([H_{q,i}^{(B)} \,\|\, S_{q,i}^{(B)} \,\|\, \|V_{q,i}^{(B)}\|_2]\right) \tag{21}$$

denote the invariant representation of atom i in view q, and let $\alpha_{q,i}$ denote its normalized attention weight within that view. The corresponding molecular representations are

$$s_q = \sum_{i=1}^{N} \alpha_{q,i} a_{q,i}, \tag{22}$$

$$v_q = \sum_{i=1}^{N} \alpha_{q,i} \left(V_{q,i}^{(B)} W_{\mathrm{read}} + \widetilde{r}_{q,i} c_{q,i}^{\top}\right), \tag{23}$$

where W_{read} is a bias-free channel projection, $c_{q,i}$ is predicted from $a_{q,i}$, and $\widetilde{r}_{q,i}$ denotes the centered position of atom i in view q. Since $\alpha_{q,i}$ and $c_{q,i}$ are derived from invariant features, s_q is invariant and v_q is equivariant. We further define the invariant conformer representation used by the multi-conformer objective as

$$z_q = s_q + \|v_q\|_2, \tag{24}$$

where the norm is taken over the spatial dimension independently for each feature channel.

A.3 Symmetry Verification

We verify that each RIME block preserves the transformation properties of the topology, scalar, and vector representations. Consider a rigid transformation of the coordinates,

$$r_i' = Qr_i + t, \qquad Q \in O(3). \tag{25}$$

By construction, $H^{(b)}$ and $S^{(b)}$ are invariant and $V^{(b)}$ is equivariant, i.e.,

$$H'^{(b)} = H^{(b)}, \qquad S'^{(b)} = S^{(b)}, \qquad V'^{(b)} = QV^{(b)}. \tag{26}$$

Since orthogonal transformations preserve Euclidean norms,

$$\|QV_{i,:,c}^{(b)}\|_2 = \|V_{i,:,c}^{(b)}\|_2, \tag{27}$$

the invariant geometric representation $G^{(b)}$ is unchanged. Consequently, the reciprocal predictions, innovations, bottleneck messages, and gates used in the topology and scalar updates are all invariant, implying

$$H'^{(b+1)} = H^{(b+1)}, \qquad S'^{(b+1)} = S^{(b+1)}. \tag{28}$$

It remains to verify the vector update in Equation 13. Centering removes the translation component, so

$$\widetilde{r}_i' = Q\widetilde{r}_i. \tag{29}$$

Moreover, $W_V^{(b)}$ acts only on feature channels and therefore commutes with the spatial transformation:

$$(QV_i^{(b)})W_V^{(b)} = Q\big(V_i^{(b)}W_V^{(b)}\big). \tag{30}$$

Because the coefficients $a_{V,i}^{(b)}$ and $a_{R,i}^{(b)}$ are computed from invariant features, they are themselves invariant. Each term in the vector update therefore transforms equivariantly, yielding

$$V_i'^{(b+1)} = QV_i^{(b+1)}. \tag{31}$$

Thus, each RIME block preserves invariant topology and scalar representations together with an $O(3)$-equivariant vector representation, while remaining invariant to global translations.

A.4 Pretraining Objectives

Geometric denoising and coordinate recovery. For each atom i in a corrupted conformation, let $\epsilon_i \in \mathbb{R}^3$ denote the coordinate perturbation relative to its parent conformation. We normalize the perturbation using running element-wise statistics:

$$\epsilon_{i,\text{norm}} = \frac{\epsilon_i - \mu_\epsilon}{\sigma_\epsilon}, \tag{32}$$

where $\mu_\epsilon, \sigma_\epsilon \in \mathbb{R}^3$ denote the running mean and standard deviation, respectively. The denoising objective is

$$\mathcal{L}_{\text{den}} = \frac{1}{3|\mathcal{A}^-|} \sum_{i \in \mathcal{A}^-} \|\widehat{\epsilon}_{i,\text{norm}} - \epsilon_{i,\text{norm}}\|_2^2, \tag{33}$$

where $\mathcal{A}^-$ indexes atom instances across all corrupted conformations.

After reversing the normalization, the predicted perturbation $\widehat{\epsilon}_i$ is used to recover the corresponding parent coordinates:

$$\widehat{R}_i^{\text{par}} = R_i^- - \widehat{\epsilon}_i. \tag{34}$$

The coordinate-recovery objective is then

$$\mathcal{L}_{\text{rec}} = \frac{1}{3|\mathcal{A}^-|} \sum_{i \in \mathcal{A}^-} \left\| \widehat{R}_i^{\text{par}} - R_i^{\text{par}} \right\|_2^2. \tag{35}$$

Masked topology reconstruction. We independently mask atom and bond inputs with probability 0.15 and replace the masked entries with learned mask tokens. After RIME interaction and cross-view aggregation, masked atom types are predicted from the corresponding node representations, while masked bond types are predicted from the representations of their incident atoms. The topology objective is

$$\mathcal{L}_{\text{top}} = \mathcal{L}_{\text{atom}} + \mathcal{L}_{\text{bond}}, \tag{36}$$

where $\mathcal{L}_{\text{atom}}$ and $\mathcal{L}_{\text{bond}}$ are cross-entropy losses evaluated only over masked atoms and bonds, respectively.

Reciprocal prediction. The innovation objective $\mathcal{L}_{\text{inn}}$, defined in Equation 16, is applied at every RIME block to all molecular views, including corrupted conformations. Stop-gradient is applied only to the prediction targets. The reciprocal predictors therefore receive gradients from $\mathcal{L}_{\text{inn}}$ and, through the exchanged innovations, from the downstream pretraining objectives.

Multi-conformer regularization. We apply multi-conformer regularization only to the $C = K + 1$ parent conformations. Given their invariant representations $\{z_k\}_{k=1}^{C}$, we define the shared and conformation-specific components as

$$z_{\text{shared}} = \frac{1}{C}\sum_{k=1}^{C} z_k, \qquad z_{k,\text{private}} = z_k - z_{\text{shared}}. \tag{37}$$

The shared component is aligned with the stop-gradient topology representation z_{top}, while the private components are regularized to preserve conformation-specific variation and reduce their directional overlap with the shared component:

$$\mathcal{L}_{\text{shared}} = \text{SmoothL1}\left(z_{\text{shared}}, \text{sg}(z_{\text{top}})\right), \tag{38}$$

$$\mathcal{L}_{\text{margin}} = \max\left(0,\, m - \sqrt{\frac{1}{Cd}\sum_{k=1}^{C}\|z_{k,\text{private}}\|_2^2}\right), \tag{39}$$

$$\mathcal{L}_{\text{orth}} = \frac{1}{C}\sum_{k=1}^{C}\cos^2(z_{k,\text{private}}, z_{\text{shared}}), \tag{40}$$

$$\mathcal{L}_{\text{conf}} = \mathcal{L}_{\text{shared}} + \mathcal{L}_{\text{margin}} + 0.1\mathcal{L}_{\text{orth}}. \tag{41}$$

We set $m = 0.05$. Corrupted conformations are excluded from this regularization but remain involved in the shared RIME stack and all other applicable pretraining objectives.

A.5 Pretraining Setup

We pretrain MolRIME on PCQM4Mv2, which provides approximately 3.4 million molecules with DFT-optimized equilibrium structures. We use RDKit to extract each molecule's 2D topology and directly use its provided 3D structure as the reference conformation R^0. We randomly select 100,000 molecules for validation and use the remaining molecules for pretraining. Molecular views are constructed following Appendix A.1 and generated online rather than stored as a fixed set of augmented conformations.

The reported MolRIME model uses a hidden dimension of 256, 64 radial basis functions, eight TorchMD-ET layers, five GINE layers, and three RIME blocks with a bottleneck dimension of 16. The topology and interaction dropout rates are 0.10 and 0.05, respectively. Unless otherwise specified, the objective weights in Equation 17 are

$$\lambda_{\text{den}} = 1.0, \qquad \lambda_{\text{rec}} = \lambda_{\text{top}} = \lambda_{\text{inn}} = \lambda_{\text{conf}} = 0.1. \tag{42}$$

We optimize MolRIME using AdamW with a batch size of 64, a learning rate of 5×10^{-4}, weight decay of 10^{-5}, and gradient-norm clipping at 5.0. Pretraining runs for 15 epochs without a learning-rate scheduler. After each epoch, we evaluate the total pretraining objective on the validation split and retain the checkpoint with the lowest validation loss for downstream initialization.

B Downstream Experimental Details

B.1 Data Splits and Evaluation Protocols

MD22. For each MD22 system, we randomly split trajectory frames into training, validation, and test sets with an 80%/10%/10% ratio and fine-tune a separate model. The molecular topology is constructed once per system and shared across all trajectory frames. The model predicts molecular energy, with atomic forces obtained as its negative gradient with respect to the input coordinates. Fine-tuning uses Smooth L1 losses for energy and forces with weights of 0.2 and 0.8, respectively, without auxiliary pretraining objectives. We select the checkpoint with the lowest validation force MAE and report test force MAE averaged over atomic Cartesian components in $\mathrm{kcal\,mol^{-1}\,\AA^{-1}}$. Fine-tuning hyperparameters are provided in Table 5.

Table 5: Hyperparameters for fine-tuning MolRIME on MD22.

Parameter	Value or description
Batch size	32; 16 for buckyball catcher; 8 for double-walled nanotube
Optimizer	AdamW
Learning rate (encoder / head)	10^{-3} / 10^{-3}
Learning rate for buckyball catcher and double-walled nanotube	5×10^{-3} / 5×10^{-3}
Weight decay	10^{-2}
Learning-rate schedule	Linear warm-up, then ReduceLROnPlateau
Warm-up duration	1,000 optimizer steps
Plateau reduction factor	0.8
Plateau patience	20 epochs; 30 for buckyball catcher and double-walled nanotube
Minimum learning rate	10^{-7}
Maximum number of epochs	600; 900 for AT-AT-CG-CG, buckyball catcher, and double-walled nanotube
Supervised loss	Smooth L1 for energy and forces
Force / energy loss weights	0.8 / 0.2
Downstream auxiliary losses	Not used

QM9. QM9 contains approximately 134K small organic molecules annotated with quantum-chemical properties computed using density functional theory. We randomly assign 110,000 molecules to training and 10,000 to validation, with the remainder used for testing, and use the same partition across all 12 targets. A separate model is fine-tuned for each target using Smooth L1 loss without auxiliary pretraining objectives. For energy-related targets, the prediction pipeline incorporates the corresponding atomic reference values. We select checkpoints by validation MAE and report test MAE in the target-specific units given in Table 2. Fine-tuning hyperparameters are provided in Table 6.

Table 6: Hyperparameters for fine-tuning MolRIME on QM9.

Parameter	Value or description
Batch size	64
Optimizer	AdamW
Learning rate (encoder / head)	5×10^{-4} / 5×10^{-4}
Weight decay	10^{-2}
Learning-rate schedule	Linear warm-up, then cosine decay
Warm-up duration	1% of scheduled optimizer steps
Minimum learning-rate ratio	0.01
Supervised loss	Smooth L1
Downstream auxiliary losses	Not used

Table 7: Hyperparameter search space for MoleculeNet fine-tuning. Not all combinations were evaluated.

Hyperparameter	Values considered
Head learning rate	$\{2 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$
Batch size	$\{32, 64, 128, 256\}$
Head dropout rate	$\{0, 0.2, 0.4\}$
Trainable encoder layers	0-8
Encoder learning-rate scale	$\{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$
Weight decay	$\{10^{-5}, 10^{-2}\}$

MoleculeNet. We evaluate BBBP, Tox21, MUV, BACE, ToxCast, SIDER, ClinTox, and HIV using scaffold-based training, validation, and test splits reproduced with the publicly released M2UMol splitting code. The same split for each dataset is used across all training seeds. Fine-tuning uses binary cross-entropy over observed labels only, with missing labels excluded from the loss. We select checkpoints by validation ROC-AUC and report test ROC-AUC as the mean and standard deviation over three seeds. For multi-task datasets, ROC-AUC is computed for each task containing both positive and negative test examples and then averaged across evaluable tasks. The fine-tuning hyperparameter search space is provided in Table 7.

B.2 Frozen-Encoder Transfer under Limited Labels

We provide additional details of the Direct Exchange baseline and the complete results for the frozen-encoder experiment in Section 4.5. Direct Exchange is pretrained separately from MolRIME while retaining the same modality encoders, low-rank message projections, gated update paths, pretraining protocols, and optimization settings. The only difference is the signal communicated between modalities. In each interaction block, Direct Exchange uses the invariant geometry representation $G^{(b)}$ as the source of the 3D-to-2D message instead of the corresponding prediction residual $G^{(b)} - \hat{G}^{(b)}$, and uses the topology representation $H^{(b)}$ as the source of the 2D-to-3D message instead of $H^{(b)} - \hat{H}^{(b)}$. Thus, Direct Exchange communicates the full source-modality representations, whereas MolRIME communicates cross-modal prediction residuals. All downstream training hyperparameters, including the learning rate and batch size, follow the original QM9 evaluation setup, except that the maximum number of training epochs is set to 200.

For downstream evaluation, the pretrained encoders are kept frozen and only the prediction heads are optimized. The validation and test sets are unchanged across label fractions, and the two models use identical labeled subsets for each seed. Table 8 reports the absolute test MAEs and standard deviations over three seeds corresponding to the relative reductions reported in Table 4.

B.3 Cross-Modal Predictability and Innovation Dynamics

We further analyze how cross-modal prediction and innovation evolve across the three RIME interaction blocks. Using approximately 2,000 held-out pretraining molecules, we run the pretrained MolRIME encoder without parameter updates and examine the representations and reciprocal predictions associated with each interaction block. Specifically, we extract the topology representation H, the invariant geometry representation G, and their corresponding cross-modal predictions $\widehat{H}$ and $\widehat{G}$ at each block, following the definitions in Section 3.3. Following Equation 6, the corresponding directional innovations are

$$\Delta^{(b)}_{3D\to 2D} = G^{(b)} - \widehat{G}^{(b)}, \qquad \Delta^{(b)}_{2D\to 3D} = H^{(b)} - \widehat{H}^{(b)}. \tag{43}$$

Since the predictions are recomputed from the evolving representations at every interaction block, these innovations are dynamically re-estimated rather than propagated unchanged through the interaction stack.

Cross-modal predictability. Figure 2(a) measures how well each modality representation can be predicted from the other modality. For molecule m with N_m atoms, we compute the cosine similarity between each atom-level representation and its cross-modal prediction, and then average

Table 8: Complete results for frozen-encoder transfer on QM9 with limited labels. Values are test MAE (mean ± sample standard deviation over three seeds); lower is better. MolRIME and Direct Exchange use the same labeled subset for each seed and identically configured prediction heads. Only the prediction heads are trained; the pretrained encoders remain frozen.

Property	Method	1% labels	5% labels	10% labels
R^2	Direct Exchange	37.264 ± 1.409	19.250 ± 0.141	14.906 ± 0.069
	MolRIME	34.267 ± 0.490	17.872 ± 0.365	13.647 ± 0.177
ZPVE	Direct Exchange	0.06520 ± 0.00165	0.03603 ± 0.00025	0.02773 ± 0.00012
	MolRIME	0.06313 ± 0.00103	0.03440 ± 0.00040	0.02603 ± 0.00025
HOMO	Direct Exchange	0.1907 ± 0.0021	0.1397 ± 0.0012	0.1150 ± 0.0010
	MolRIME	0.1873 ± 0.0031	0.1393 ± 0.0006	0.1153 ± 0.0006
α	Direct Exchange	1.288 ± 0.028	0.719 ± 0.003	0.559 ± 0.003
	MolRIME	1.286 ± 0.007	0.696 ± 0.006	0.532 ± 0.007
μ	Direct Exchange	0.5773 ± 0.0117	0.3550 ± 0.0050	0.2870 ± 0.0017
	MolRIME	0.5627 ± 0.0181	0.3303 ± 0.0076	0.2643 ± 0.0015
G	Direct Exchange	0.3603 ± 0.0115	0.1820 ± 0.0044	0.1257 ± 0.0023
	MolRIME	0.3247 ± 0.0095	0.1660 ± 0.0026	0.1177 ± 0.0012

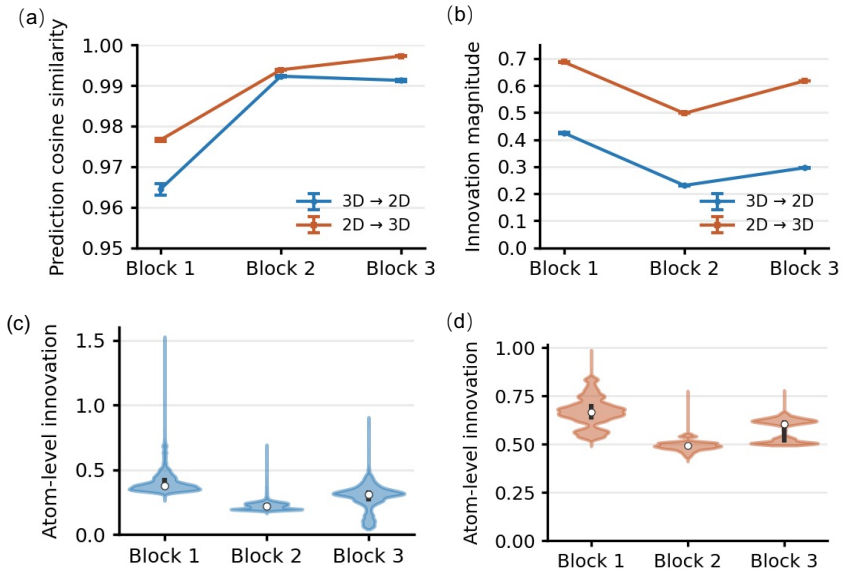

Figure 2: **Cross-modal predictability and innovation dynamics in pretrained MolRIME.** (a) Atom-wise cosine similarity between each representation and its cross-modal prediction, averaged within each molecule. (b) Molecule-level normalized innovation magnitude. (c-d) Atom-level normalized innovation distributions associated with the two directional predictions. Error bars in (a-b) denote 95% molecule-level bootstrap confidence intervals; white dots and dark bars in (c-d) denote medians and interquartile ranges, respectively.

over atoms:

$$C_{H,m}^{(b)} = \frac{1}{N_m} \sum_{i=1}^{N_m} \cos\left(H_{m,i}^{(b)}, \widehat{H}_{m,i}^{(b)}\right), \tag{44}$$

$$C_{G,m}^{(b)} = \frac{1}{N_m} \sum_{i=1}^{N_m} \cos\left(G_{m,i}^{(b)}, \widehat{G}_{m,i}^{(b)}\right). \tag{45}$$

We report the mean across held-out molecules, with 95% confidence intervals obtained from 5,000 molecule-level bootstrap resamples. The prediction similarity increases substantially from Block 1 to Block 2 and remains high in Block 3, indicating increased cross-modal predictability after the initial interaction stage.

Molecule-level innovation magnitude. While cosine similarity measures directional alignment between a representation and its cross-modal prediction, it does not capture the relative magnitude of their difference. We therefore quantify the molecule-level prediction residual in Figure 2(b). For each molecule, we compute

$$R_{H,m}^{(b)} = \frac{\left\| H_m^{(b)} - \widehat{H}_m^{(b)} \right\|_F}{\left\| H_m^{(b)} \right\|_F + \epsilon}, \tag{46}$$

$$R_{G,m}^{(b)} = \frac{\left\| G_m^{(b)} - \widehat{G}_m^{(b)} \right\|_F}{\left\| G_m^{(b)} \right\|_F + \epsilon}, \tag{47}$$

where $\epsilon = 10^{-8}$ is used for numerical stability. Each molecule contributes one value per block for each representation, and we report the mean across held-out molecules with 95% confidence intervals obtained from 5,000 molecule-level bootstrap resamples. The normalized residual magnitude decreases from Block 1 to Block 2 and partially increases in Block 3. This pattern is compatible with the high cosine similarity in Figure 2(a), since cosine similarity captures directional alignment, whereas the normalized residual additionally reflects differences in magnitude.

Atom-level innovation distributions. Figures 2(c-d) further characterize the prediction residuals at the atom level. For atom i, we compute the normalized residual magnitude

$$r_{H,i}^{(b)} = \frac{\left\| H_i^{(b)} - \widehat{H}_i^{(b)} \right\|_2}{\left\| H_i^{(b)} \right\|_2 + \epsilon}, \tag{48}$$

$$r_{G,i}^{(b)} = \frac{\left\| G_i^{(b)} - \widehat{G}_i^{(b)} \right\|_2}{\left\| G_i^{(b)} \right\|_2 + \epsilon}. \tag{49}$$

The violin plots pool atoms from all held-out molecules, with white dots indicating the median and dark vertical bars indicating the interquartile range. The distributions contract from Block 1 to Block 2 and broaden or reorganize in Block 3, indicating that the block-wise changes in prediction residuals are not explained by a uniform rescaling across atoms.

Overall, Figure 2 shows that cross-modal predictability increases early in the interaction stack, while the prediction residuals remain non-zero and continue to change across blocks. This behavior is consistent with RIME's block-wise re-estimation of cross-modal predictions and innovations.