MOL-JEPA: A MULTIMODAL JOINT EMBEDDING PREDICTIVE ARCHITECTURE FOR MOLECULES

Florian Rottach[*1,2] **Sebastian Schieferdecker**[*2] **William Rudman**[3]
Randall Balestriero[4] **Carsten Eickhoff**[1]

[1]University of Tübingen [2]Boehringer Ingelheim
[3]The University of Texas at Austin [4]Brown University

florian.rottach@boehringer-ingelheim.com

[*]Equal contribution

ABSTRACT

Despite recent advances in molecular foundation models, several limitations remain, such as chemically invalid augmentations, modality collapse, and incomplete representation of biochemical environments. To address these challenges, we present **Mol-JEPA**, a scalable framework for learning molecular world models. Rather than relying on suboptimal molecular perturbations, our model uses modality masking to exploit information from molecular structures, cellular phenotypes, binding affinities, ADMET profiles, quantum chemistry simulations and other drug discovery data. Across various benchmarks, we show that the representations learned by Mol-JEPA deliver strong performance, demonstrating the value of incorporating biochemical context through latent space prediction.

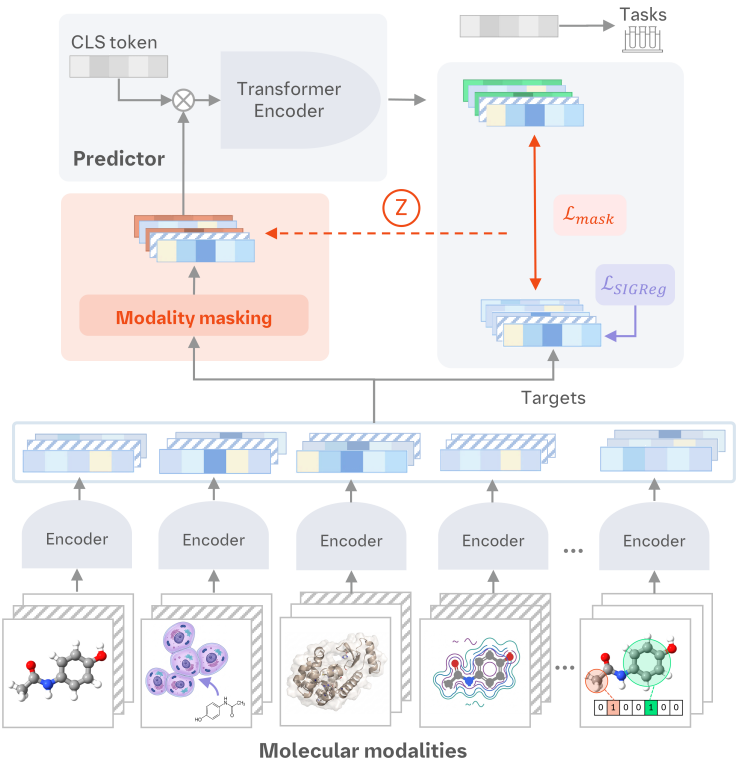


Figure 1: **Mol-JEPA:** Rather than representing molecules solely through their structure, our model learns from a spectrum of biochemical and physical modalities. During training, modalities are randomly masked and predicted from the remaining inputs using a Transformer-based predictor. To prevent representation collapse, we employ isotropic regularization through SIGReg.

1 INTRODUCTION

The development of novel drugs is a costly process characterized by high failure rates. Drug candidates may fail for numerous reasons, including weak target binding, poor absorption, rapid metabolic clearance, and adverse toxicity profiles Sun et al. (2022a). Over the past decades, cheminformatics research has led to the development of various molecular representations to address the challenging prediction problems in drug discovery. Examples range from basic physicochemical descriptors, such as atom and bond statistics to more advanced features including Extended Connectivity Fingerprints (ECFP) Rogers & Hahn (2010) and quantum-chemical descriptors Wang et al. (2021). In recent years, learned representations have emerged from various neural network variants, such as Graph Neural Networks Reiser et al. (2022), which operate directly on the molecular graph or Chemical Language Models Grisoni (2023), representing molecules as text. While these works are important milestones, several studies have found that pretrained representations often fail to consistently outperform traditional baselines or only show marginal performance gains Jiang et al. (2021); Sun et al. (2022b); Fischer et al. (2025); Praski et al. (2025); Guo (2026). On one hand, this can be attributed to the pretraining datasets, which either have poor quality or do not cover the relevant molecular space Rodrigues (2019); Chodera et al. (2026). On the other hand, models trained with self-supervised learning (SSL) methods use augmentation strategies that are unsuitable for molecular data. For example, many SSL approaches generate multiple views of a molecule through atom or bond masking and train the model to produce similar representations for these augmented views. This can be problematic because molecular property landscapes are often highly discontinuous, with small structural modifications leading to substantial changes in molecular properties, as demonstrated by property cliffs Stumpfe et al. (2019). This misalignment between augmentation bias and desired inductive invariance has been shown to harm SSL performance Van Assel et al. (2025) and offers an explanation why SSL approaches on molecular data have not been particularly successful.

In addition to unsuitable augmentation strategies, we argue that molecular structure alone is insufficient for SSL pretraining, because molecules are inherently *context-dependent*. Drug behavior emerges from complex interactions with biological systems, including protein networks, metabolic processes, cellular states, and physiological environments that are not encoded in molecular structure alone. Incorporating data that reflects molecular interactions enables models to navigate chemical space through a biological lens and may improve their ability to identify complex structure-property relationships, including property cliffs Dablander et al. (2023); Sanchez et al. (2026). In recent years, several attempts have been made to enrich molecular representations by incorporating additional information. Some of these models rely on multi-task pretraining Beaini et al. (2023), which however suffers from over-specialization due to discontinuities and errors in the data Sheridan et al. (2020). Joint embedding approaches have been shown to be more robust to variations in the input and target spaces Van Assel et al. (2025), making them particularly interesting for molecular data. Several works have adapted molecular joint embedding models using contrastive learning Liang et al. (2022); Masood et al. (2025); Xiong et al. (2026). While effective, these methods depend heavily on the selection of negative samples, which can substantially affect representation quality Jing et al. (2021). The recently presented Joint Embedding Predictive Architecture (JEPA) LeCun et al. (2022) has several advantages over previous methods, such as no need for negative sampling and robustness to noise, making it less prone to overfitting. However, JEPA still relies on meaningful data augmentations, which remain challenging to define for molecular structures.

To address these limitations, we present Mol-JEPA, a multimodal molecular world model based on the Joint Embedding Predictive Architecture. Rather than applying augmentations to the molecular structure, we mask entire modalities and predict the corresponding latent representations, enabling the model to learn predictive representations that capture molecular behavior across diverse biological and chemical contexts. Our considered modalities are derived from pretrained foundation models, quantum-chemical calculations and experimental datasets, spanning binding affinity, cellular effects and ADMET properties. Furthermore, by using the LeJEPA framework Balestriero & LeCun (2025) as anti-collapse mechanism, we ensure that each modality contributes during training. Across relevant downstream datasets, we show that our model learns effective molecular representations outperforming our baselines and reaching strong performance on public leaderboards. Overall, our results demonstrate the potential of latent-space molecular foundation models and establishes a scalable framework for integrating additional modalities and datasets. Our main contributions are as follows:

- We present **Mol-JEPA**, a scalable multimodal Joint Embedding Predictive Architecture that enables self-supervised learning through meaningful augmentations on molecular data in drug discovery.
- We introduce a novel multimodal dataset containing nearly 5 million molecules and spanning a diverse drug-like chemical space across a range of biological, chemical, and computational modalities.
- We demonstrate highly competitive performance on relevant benchmark datasets, statistically evaluate the quality of predictions, assess out-of-distribution performance and analyze the contributions of different modalities.

2 Related Work

Deep learning in chemistry In recent years, various deep learning methods have been applied in computational chemistry. Early works adapted popular neural network architectures to molecular data, yielding models such as ChemBERTa Chithrananda et al. (2020), Chemical VAE Gómez-Bombarelli et al. (2018) or Chemformer Irwin et al. (2022). Graph Neural Networks (GNNs) Kipf & Welling (2016) have become a prominent approach for molecular representation learning, as they integrate inductive priors that allow them to operate directly on molecular graphs. Others have represented molecules as text for training Chemical Language Models Grisoni (2023), most commonly using the SMILES notation Weininger (1988). While these approaches demonstrated the ability to learn representations directly from molecular data, they remain limited to string-based or graph-based inputs. To address this limitation, subsequent work incorporated geometric information such as interatomic distances in GEM Fang et al. (2022) and atomic coordinates in Uni-Mol Zhou et al. (2023), using a roto-translation equivariant transformer architecture. Following the trends in other domains, pretraining efforts have been scaled to increasingly large datasets, such as UMA Wood et al. (2026), an interatomic potential model trained on more than 500 million three-dimensional atomic structures.

Self-supervised molecular representation learning Popular self-supervised learning paradigms, such as masked prediction and contrastive learning have also found their way into the chemical sciences. For example, several methods have generated different views on a molecule by augmenting molecular graphs, which are jointly encoded or predicted Rong et al. (2020); Hafidi et al. (2020); Wang et al. (2022); Méndez-Lucio et al. (2024). We emphasize that such augmentations can introduce false positives; it cannot be guaranteed that two similar views *actually* behave similarly on a specific downstream tasks. Other works have adopted CLIP Radford et al. (2021) to fuse molecular graphs with other modalities, such as jointly encoding molecules with cellular images Sanchez-Fernandez et al. (2023); Masood et al. (2025) or images of molecules Harnik et al. (2025). These models overcome the augmentation bias of previous methods, but require negative samples, which can also introduce false negatives. Joint Embedding Predictive Architectures have been applied to augmented views of molecules Mizera et al. (2024); Piccoli et al. (2026); Orester (2025); Iyer & Sabar (2026), which avoid negative samples, but suffer from the same augmentation bias as the previous methods. Recent work introduced CheMeleon Green et al. (2026), a GNN pretrained to predict molecular descriptors. While this is a promising pretext task, their model only considers descriptors and does not readily scale to multimodal settings.

Multimodal architectures While several multimodal models integrate molecular graphs, SMILES, textual descriptions, and 3D structural information Mirza et al. (2024); Luo et al. (2023); Manolache et al. (2024), most focus solely on representations of the molecule itself, neglecting modalities that reflect its effects in biochemical environments. Furthermore, existing approaches typically fuse modalities using concatenation or attention-based aggregation, which can result in modality collapse, causing some modalities to be underutilized or ignored entirely. Joint embedding approaches offer an alternative that avoids these issues. For example, BioXMol aligns cellular, genetic, and molecular modalities within a shared embedding space encouraging the model to leverage information from all modalities during training Masood et al. (2025). While this work offers a promising direction, it relies on contrastive methods, which are known to suffer from modality gap and hubness effects, arising from its discriminative objective Liang et al. (2022). Additionally, such methods do not scale well to many modalities and struggle with sparse and only partially paired modalities.

Our architecture addresses all limitations of previous approaches. First, no chemically invalid molecule augmentations are introduced - instead we mask entire modalities. Second, we go beyond molecular structure by integrating knowledge about the biochemical environment. Third, all modalities are predicted, which avoids modality collapse and encourages the model to leverage all available information. Finally, the framework scales well to multiple modalities, is easy to optimize and does not require any heuristics, such as alternating gradients, stop-loss or student-teacher architectures, commonly found in other works. Overall, we find that the combination of multimodality and predictions in the latent space, both theoretically and empirically prove to be a promising direction towards molecular world models.

3 Method

3.1 Pretraining Data

We construct a pretraining dataset by merging and curating public datasets and augment them with additional modalities ranging from specialized model embeddings to quantum chemical calculations. Furthermore, we leverage experimental measurements to construct multidimensional molecular profiles that capture diverse aspects of a molecule's chemical and biological behavior. Table 1 provides a summary of the individual datasets, which result in a combined dataset size of 4.69 million unique compounds.

Table 1: Pretraining datasets with their final sample counts and average molecular weight (MW).

Dataset	Domain	# Raw	# Processed	Avg. MW	Source
ChEMBL	Bioactivity	280,271	268,206	399.73	Gaulton et al. (2017)
TDC	ADMET	512,303	402,881	365.86	Huang et al. (2021)
PCBA-1328	Bioactivity	1,563,664	1,124,981	378.82	Beaini et al. (2023)
∇^2DFT	Quantum chemistry	1,936,931	1,270,806	307.46	Khrabrov et al. (2024)
Enamine REAL	Unlabeled	2,014,554	2,012,628	307.13	Grygorenko et al. (2020)

Public datasets First, we extract a subset of drug-like small molecules from ChEMBL Gaulton et al. (2017) and query the API to add bioactivity information for each molecule. For potency, we filter the activity types to fall into IC_{50}, K_i, K_d, EC_{50} or AC_{50} and use the pChEMBL value (-log10 M) for a consistent scale. For absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties, we standardize units and apply further processing steps, as detailed in Appendix A. We drop all activities with less than 30 labels, leaving us with 306 unique assays for around 30 thousand molecules. In addition, we integrate 21 ADME, 11 toxicity and 12 high-throughput screening endpoints from Therapeutic Data Commons (TDC) Huang et al. (2021). We aggregate all measurements per molecule and obtain 672 attributes, with on average 8 available labels per molecule. A full table of properties is provided in Appendix Table 5. We also use PCBA-1328 Beaini et al. (2023), a subset of PubChem Kim et al. (2016), containing 1,328 bioassays with "Active" or "Inactive" flags with at least 10 labels per assay. Further, we integrate quantum-chemical data from ∇^2DFT Khrabrov et al. (2024), containing 17 global density functional theory (DFT) properties, such as total energy, dipole moments and HOMO-LUMO gap, computed at the ωB97X-D/def2-SVP theory level. We extract the lowest energy conformer and assign it to each unique SMILES string. All DFT properties have been normalized to yield smooth label distributions, by applying scikit-learn's quantile transformer Kramer (2016). Finally, to increase the chemical space coverage relevant for drug discovery, we sample around 2 million unlabeled molecules from the Enamine REAL space Grygorenko et al. (2020). We compute international chemical identifiers (InChI), which are used to merge and de-duplicate the dataset, leaving us with a final size of 4.69 million compounds. All molecule SMILES are canonicalized and normalized to pH7 with MoKa (version 3.2.3), stereochemistry is validated and a low energy conformer is computed for each molecule using the ETKDGv3 algorithm Wang et al. (2020) followed by geometry optimization with the universal force field (UFF) Rappé et al. (1992).

Computed data Our model architecture allows us to distill information from other chemical foundation models, by predicting their learned representations. To this end, we extract final-layer embeddings from several pretrained molecular backbones. First, we compute embeddings for UMA, a large mixture-of-linear-experts graph network trained on DFT simulation data Wood et al. (2026). Due to computational constraints, we use the smallest model variant, *uma-s-1p2*. We further incorporate CLOOME Sanchez-Fernandez et al. (2023), which jointly encodes molecular structures and cell painting data, offering a deep phenotypic profile that explains the cellular bioactivity of a molecule. Similarly, we integrate BioXMol Masood et al. (2025), a multimodal model trained on cell painting and transcriptomics data using contrastive learning learning. As a general-purpose molecular language model, we additionally extract embeddings from ChemGPT (4.7M) Frey et al. (2023), which was pretrained on the PubChem10M dataset. Recent works emphasize that prioritizing binding affinity has the potential to lead to more successful drug candidates and better pharmacological understanding Murcko (2026). Therefore, we attempt to guide the model about adverse effects to improve downstream performance on ADMET properties, such as *in vivo* phenotypes. For this, we perform co-folding and affinity prediction using Boltz-2 Passaro et al. (2025) and concatenate the predictions with the ensemble embeddings from the final layer of the binding affinity module. Specifically, we predict against targets from the Bowes-44 panel Bowes et al. (2012), which are associated with adverse effects based on historical clinical failures. Due to substantial computational requirements, we restrict the predictions to a subset of 9 targets: 4 G-protein coupled receptors (serotonin 5-HT_{2A}, dopamine D_2, muscarinic M_2, α_1-adrenoceptor), 2 ion channels (hERG, NMDA GluN1), 1 kinase (Lck), 1 enzyme (COX-1) and 1 nuclear receptor (GR). Predictions were generated for more than 100,000 randomly selected molecules using eight RTX A6000 GPUs with cached multiple-sequence alignments.

In addition to the embeddings extracted from pretrained models, we computed quantum chemical properties using GFN2-xTB Bannwarth et al. (2019). We choose this semi-empirical method due to its balance of accuracy and performance, allowing us to compute descriptive vectors for a large number of samples. Lastly, we compute descriptors using the Molecular Operating Environment (MOE) Chemical Computing Group (2022) and Extended Connectivity Fingerprints Rogers & Hahn (2010), which are commonly used in computational chemistry. This allows us to integrate human expertise, built up over decades of drug discovery, integrating various relevant properties such as relevant substructures, charge distributions and 3D topology.

Table 2: Molecular modalities used in Mol-JEPA.

Modality	Description	Dim	Type	Encoder	# Samples	Source
UMA	Atomistic foundation model	128	Embedding	Atom	4,663,778	Wood et al. (2026)
ChemGPT	SMILES Transformer	2048	Embedding	Vector	4,688,836	Frey et al. (2023)
CLOOME	Cell painting model	512	Embedding	Vector	4,640,188	Sanchez-Fernandez et al. (2023)
BioXMol	Multi-assay phenotypic model	1024	Embedding	Vector	4,640,188	Masood et al. (2025)
Boltz-2	Boltz-2 embeddings	6912	Embedding	Vector	102,234	Passaro et al. (2025)
Boltz-2 (P)	Boltz-2 binding affinity	4	Embedding	Vector	102,234	Passaro et al. (2025)
GNN	Graph transformer model	512	Embedding	Graph	4,693,006	Shi et al. (2020)
MOE	Molecular descriptors	227	Descriptors	Vector	4,663,780	Chemical Computing Group
ECFP	Circular topology fingerprints	2048	Descriptors	Vector	4,663,780	Rogers & Hahn (2010)
xTB	GFN2-xTB simulations	7	Descriptors	Vector	3,435,704	Bannwarth et al. (2019)
∇^2DFT	ωB97X-D/def2-SVP DFT	17	Descriptors	Vector	1,270,722	Khrabrov et al. (2024)
ChEMBL	Bioactivity measurements	306	Experimental	Vector	268,188	Gaulton et al. (2017)
PCBA	Bioactivity measurements	1328	Experimental	Vector	1,124,829	Beaini et al. (2023)
TDC	ADMET measurements	672	Experimental	Vector	402,755	Huang et al. (2021)

3.2 Modality encoders

The combined dataset results in different *views* on a molecule, serving as distinct modalities in our model. Notably, we can treat multi-dimensional label vectors as modalities, as they describe how molecules behave in the chemical world. Combined with the embeddings and descriptors, we obtain 14 final modalities, which are summarized in Table 2. We employ three learnable encoders that project the modalities to the same dimensionality: An atom encoder, designed for atomic coordinates of the shape *(batch, #atoms, dim)*, a Graph encoder operating on molecular graphs with additional connectivity information and a Vector encoder, which transforms representations of shape *(batch, dim)*.

3.3 Model Architecture

Mol-JEPA is a multimodal joint embedding predictive architecture, which predicts masked modalities from available ones (see Figure 1). Given a masking ratio r, some of the encoded modalities are deactivated, while ensuring that always at least one modality is available. In the masked forward pass, we replace these embeddings by masking tokens and also replace all missing modalities through respective tokens. The prediction module is implemented as a transformer encoder and optimized to predict the masked embeddings. We use the *TransformerEncoderLayer* from PyTorch Paszke et al. (2019), which applies self-attention and a feedforward network. Further, we add a learnable CLS token as a global readout for downstream tasks and apply positional encodings to all predictor inputs, to track which embeddings need to be predicted. We experiment both with predicting the embeddings directly from the transformer (modality-based) and predicting them from the CLS token through another transformation (CLS-based). To avoid mode collapse, we apply Sketched Isotropic Gaussian Regularization (SIGReg) Balestriero & LeCun (2025). We consider multiple strategies for SIGReg, applying it either to the target embeddings, the predicted embeddings or the CLS token. Mol-JEPA is implemented using the stable-pretraining framework Balestriero et al. (2025) and hyperparameter-tuned with Optuna Akiba et al. (2019) over the search space specified in Appendix B, resulting in a final model with around 50 million parameters.

3.4 Pretraining Objectives

Assume a batch $\mathcal{B} = \{x_1, \ldots, x_B\}$ with a set of modalities $\mathcal{M} = \{1, \ldots, M\}$. For each sample i and modality $m \in \mathcal{M}$, an embedding $\mathbf{s}_i^m \in \mathbb{R}^d$ is obtained using encoder e^m. Further, let $\mathbf{a}_i = (a_i^1, \ldots, a_i^M)$, $a_i^m \in \{0, 1\}$ denote the modality-availability mask for each sample i, where $a_i^m = 1$ if modality m is present.

For each i, we sample a binary mask vector using masking ratio r, which determines the modalities to predict:

$$\mathbf{z}_i = (z_i^1, \ldots, z_i^M), \qquad z_i^m \sim \text{Bernoulli}(r\, a_i^m)$$

We ensure that at least one modality remains available after masking and denote the masked modalities m for sample i with $\hat{\mathbf{s}}_i^m$. A Transformer predictor is then trained to reconstruct the embeddings of the masked modalities from the available context. This reconstruction task encourages the model to combine information from different modalities

and reason about how samples are represented across alternative environments. The prediction error is implemented as the mean squared error between the masked modalities $\hat{\mathbf{s}}_i^m$ and the predicted modalities $\hat{\mathbf{s}}_{i,\text{pred}}^m$, across all samples in the batch of size B. For each modality $m \in \mathcal{M}$ we have:

$$\mathcal{L}_{\text{pred}}^m = \frac{1}{B} \sum_{i=1}^{B} \left\| \hat{\mathbf{s}}_i^m - \hat{\mathbf{s}}_{i,\text{pred}}^m \right\|_2^2 \tag{1}$$

To avoid the representations collapsing to a trivial subspace, Sketched Isotropic Gaussian Regularization (SIGReg) conditions the latent variables to approximate an isotropic distribution $\mathcal{N}(0, I_K)$. SIGReg computes this efficiently using randomized sketching. Given a batch of B embeddings $\{s_i^m\}_{i=1}^B$, the method projects them onto L random 1D directions $\{a_l\}_{l=1}^L$ drawn uniformly from the unit sphere $\mathbb{S}^{K-1}$. The Epps-Pulley statistic T is then applied to each 1D slice to measure the divergence from a standard 1D normal distribution. The total SIGReg loss is computed as the average of these univariate tests over all random projections:

$$\mathcal{L}_{\text{SIGReg}}^m = \frac{1}{L} \sum_{l=1}^{L} T\big(\{a_l^\top s_i^m\}_{i=1}^B\big) \tag{2}$$

For our experiments, we fix the number of random projections to 1024, the range of frequencies in Epps-Pulley to $T_{\max} = 3.0$, and the number of evaluated points to 17. The final Mol-JEPA objective combines the modality prediction and isotropic regularization losses across all modalities, balanced by λ:

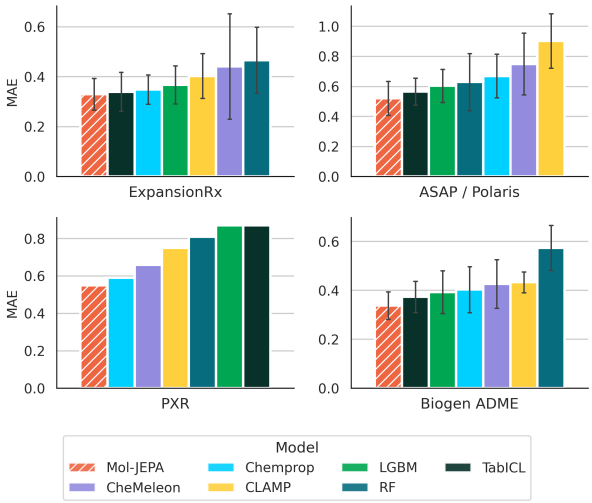
$$\mathcal{L}_{\text{JEPA}} = \sum_{m \in \mathcal{M}} \left(\mathcal{L}_{\text{pred}}^m + \lambda \, \mathcal{L}_{\text{SIGReg}}^m \right), \qquad \lambda > 0 \tag{3}$$

3.5 Benchmark Datasets

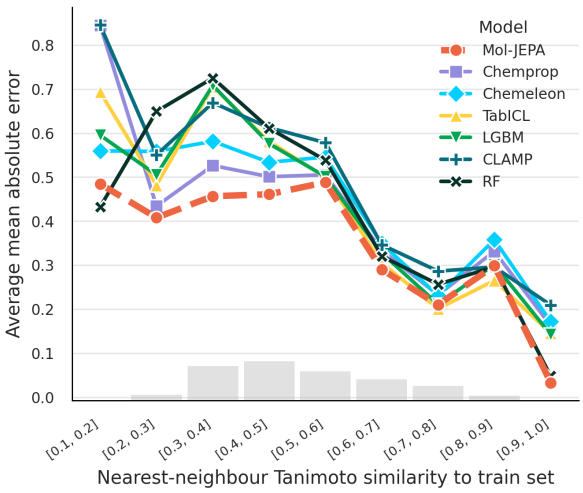
Traditional molecular benchmarks have been criticized for limited real-world relevance and data quality issues, including noisy labels, inconsistent chemical representations, and undefined stereochemistry Walters (2023); Liu et al. (2024). To ensure a realistic evaluation, we compare the model predictions across data from recent OpenADMET blind challenges along with a diverse collection of other high quality datasets. First, we use 9 endpoints from the OpenADMET ExpansionRx competition, which contain entirely novel compounds Castellanos & MacDermott-Opeskin. Similarly, we evaluate on 7 datasets from the ASAP-Polaris challenge MacDermott-Opeskin et al. (2026), including binding against two CoVID targets and multiple ADMET endpoints. Additionally, we evaluate on data from the OpenADMET Pregance-X Receptor (PXR) activity prediction blind challenge Fraser et al. (2026). Lastly, we include Biogen ADME Fang et al. (2023), which contains six high quality *in vitro* endpoints measured under the same experimental conditions. We construct 3 test-train splits using Taylor-Butina Clustering Butina (1999) on 1024 bit ECFP4 fingerprints, by grouping compounds based on a Tanimoto similarity threshold of 0.65 and assign entire clusters to either the training or test set. Furthermore, for the datasets where official competition splits are available, we report performance on these specific subsets. In Appendix A we report additional information about the benchmark datasets and splits.

3.6 Baselines

We compare the performance of our model with several established state-of-the-art baselines. As classical approaches, we fit a Random Forest and a Light Gradient Boosting Machine using scikit-learn Pedregosa et al. (2011) and LightGBM Ke et al. (2017) respectively. We experiment with ECFP4 Rogers & Hahn (2010), AlvaDesc descriptors Mauri (2020) and 3D pharmacophore fingerprints as features and select the best representation for the evaluation. For in-context predictions, we use the same feature sets along with different variants of TabICLv2 Qu et al. (2025), a tabular foundation model that doesn't require any significant hyperparameter tuning. Further, we evaluate three deep learning models: the descriptor foundation model CheMeleon Green et al. (2026), the GNN architecture Chemprop Heid et al. (2023) and the contrastive language-assay trained molecular foundation model CLAMP Seidl et al. (2023). All models are finetuned and hyperparameter tuned using a 5-fold cross-validation on the train split, across the search spaces defined in B. For CLAMP, we use the learned representations directly as features and evaluate both linear and non-linear predictors.



(a) **Mol-JEPA demonstrates strong performance across tasks.** The bars represent the aggregated mean absolute error of models within each benchmark family along with the standard deviation across the individual datasets.



(b) **Mol-JEPA generalizes better.** The lines show the average MAE for molecules across all datasets, binned by Tanimoto similarity to the train dataset. Mol-JEPA shows the best out-of-distribution performance.

Figure 2: Mol-JEPA evaluation results.

3.7 Mol-JEPA downstream prediction

There exist multiple strategies for using Mol-JEPA embeddings for downstream tasks. First, we use the CLS token to extract a global summary of all modalities. We feed this representation of size 512 into a linear and non-linear probe and experiment with TabICLv2 as a predictor. We use default parameters and do not perform any hyperparameter optimization. Secondly, we use the modality-specific predicted embeddings in an aggregative probe. We choose a 2 layer transformer encoder that receives tokens of size 512 for each modality and applies attention to combine the information. The outputs embeddings are mean pooled and fed into a simple linear output layer. Lastly, we perform LoRA fine-tuning Hu et al. (2022) with rank 16 and target layers of the transformer module and output projections of the Mol-JEPA architecture. Our experiments, model checkpoints, and Hugging Face integration are publicly available at github.com/Boehringer-Ingelheim/mol-jepa.

4 Results

Mol-JEPA outperforms the baselines on multiple datasets. As summarized by Figure 2a, Mol-JEPA demonstrates strong performance across all benchmark families. Specifically, our model demonstrates a clear advantage on benchmarks with smaller datasets, such as ASAP/Polaris and Biogen ADME. This finding is especially important in drug discovery, where obtaining large, high-quality labeled datasets is often prohibitively expensive. In addition, we evaluate out-of-distribution performance in Figure 2b and find that Mol-JEPA consistently outperforms the baselines as the distance from the training data increases. This advantage likely arises from pretraining on a large and diverse set of drug-like compounds, allowing Mol-JEPA to capture broad chemical space relevant to drug discovery. A detailed comparison for mean absolute errors is provided in Table 3 and for additional metrics in Appendix D.

Statistical significance. To assess whether performance differences are statistically significant, we compare the sample-wise absolute errors of all methods using paired Wilcoxon signed-rank tests Wilcoxon (1945). We perform pairwise comparisons between Mol-JEPA and all baseline models and compute mean MAE differences and FDR-corrected p-values. We count the statistically significant pairwise victories across datasets and summarize the win rates in Figure 3a, along with the win rates based on MAE and R^2 results. We find that Mol-JEPA has the highest win rates when using classical metrics and is close to the highest win rate when using wilcoxon rank tests. Specifically, it achieves superior performance in 38% of the comparisons, compared to 47% for TabICL with AlvaDesc descriptors. Further analysis indicates that TabICLv2 derives much of its advantage from the larger ExpansionRx datasets, while Mol-JEPA tends to achieve more significant wins on the smaller ASAP and Biogen benchmarks. These results suggest that Mol-JEPA may offer particular benefits in low-data settings.

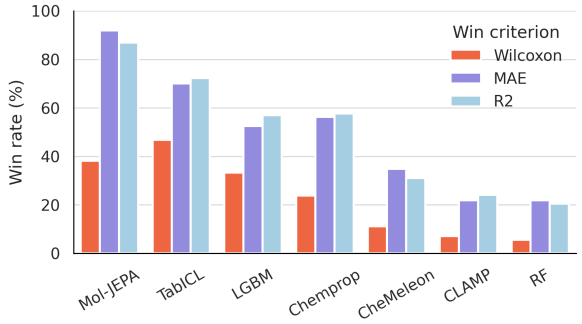
Table 3: Mean absolute error (MAE) comparison across benchmark datasets for three cluster splits. Mol-JEPA (Best) denotes the best performance achieved across all probe and embedding variants, while Mol-JEPA (Transformer) uses a Transformer head with modality-specific embeddings.

Dataset	**Mol-JEPA** Best	**Mol-JEPA** Transformer	**CLAMP** Nonlinear	**CheMeleon** Finetuned	**Chemprop** Finetuned	**TabICLv2** AlvaDesc	**RF** ECFP4	**LGBM** AlvaDesc
ExpansionRx								
Caco-2 Permeability	**0.34** ± .02	0.35 ± .02	0.40 ± .05	0.37 ± .05	0.39 ± .04	0.42 ± .02	0.43 ± .02	0.42 ± .02
Caco-2 Efflux	0.25 ± .02	0.26 ± .01	0.31 ± .03	0.27 ± .05	**0.24** ± .03	**0.24** ± .02	0.28 ± .04	**0.24** ± .02
LogD	**0.33** ± .02	0.46 ± .04	0.50 ± .05	0.35 ± .02	0.74 ± .06	0.36 ± .02	0.73 ± .07	0.43 ± .04
KSOL	**0.45** ± .03	0.49 ± .04	0.53 ± .04	**0.45** ± .06	0.60 ± .05	0.46 ± .06	0.62 ± .04	0.49 ± .03
HLM CLint	**0.39** ± .02	0.43 ± .03	0.50 ± .02	0.40 ± .06	0.46 ± .01	**0.39** ± .05	0.47 ± .02	**0.39** ± .03
MLM CLint	0.40 ± .05	0.38 ± .07	**0.36** ± .04	0.39 ± .06	0.42 ± .04	0.41 ± .09	0.44 ± .08	0.38 ± .07
MBPB	0.30 ± .02	0.30 ± .03	0.31 ± .01	0.29 ± .06	0.45 ± .10	**0.26** ± .03	0.43 ± .01	0.34 ± .04
MGMB	0.30 ± .05	**0.28** ± .06	0.35 ± .08	0.31 ± .13	0.41 ± .13	0.30 ± .09	0.39 ± .12	0.33 ± .10
MPPB	**0.26** ± .04	0.28 ± .05	0.31 ± .03	0.31 ± .02	0.41 ± .01	**0.26** ± .05	0.41 ± .06	0.29 ± .05
ASAP								
MERS-CoV-2 Potency	0.59 ± .10	0.70 ± .08	0.96 ± .04	0.86 ± .09	**0.52** ± .08	0.54 ± .09	0.58 ± .04	0.59 ± .13
SARS-CoV-2 Potency	**0.62** ± .06	**0.62** ± .06	1.02 ± .12	0.71 ± .13	0.73 ± .18	0.64 ± .17	0.77 ± .25	0.68 ± .16
LogD	0.68 ± .09	0.72 ± .11	1.09 ± .14	0.84 ± .19	**0.64** ± .14	0.89 ± .07	0.99 ± .05	0.74 ± .14
KSOL	**0.42** ± .19	0.53 ± .08	0.91 ± .19	0.60 ± .02	0.67 ± .19	0.65 ± .12	0.58 ± .06	0.70 ± .11
HLM	0.39 ± .11	0.39 ± .11	0.58 ± .03	0.50 ± .24	**0.36** ± .09	0.42 ± .15	0.43 ± .07	0.43 ± .10
MLM	0.53 ± .01	0.53 ± .01	1.02 ± .37	0.67 ± .13	0.64 ± .08	0.59 ± .15	**0.50** ± .06	0.51 ± .09
MDR1 Efflux	**0.42** ± .13	0.44 ± .14	0.74 ± .14	0.51 ± .30	0.45 ± .22	0.48 ± .26	0.56 ± .36	0.58 ± .41
PXR								
PXR Activity	**0.55** ± .04	**0.55** ± .06	0.75 ± .06	0.59 ± .13	0.79 ± .11	0.87 ± .20	0.81 ± .11	0.87 ± .17
Biogen ADME								
Solubility	**0.30** ± .02	**0.30** ± .02	0.38 ± .02	0.35 ± .03	0.42 ± .04	0.33 ± .01	0.43 ± .01	0.33 ± .01
HLM CLint	**0.33** ± .02	0.34 ± .02	0.45 ± .06	0.35 ± .01	0.51 ± .09	0.37 ± .03	0.55 ± .08	0.36 ± .02
RLM CLint	**0.37** ± .04	0.38 ± .04	0.46 ± .04	0.39 ± .01	0.57 ± .07	0.39 ± .03	0.56 ± .04	0.39 ± .04
HPPB	**0.33** ± .06	0.34 ± .07	0.45 ± .05	0.48 ± .08	0.56 ± .07	0.35 ± .04	0.67 ± .11	0.40 ± .07
RPPB	**0.43** ± .07	0.45 ± .08	0.48 ± .18	0.55 ± .13	0.57 ± .18	0.49 ± .12	0.68 ± .15	0.56 ± .12
MDR1 Efflux	**0.27** ± .03	0.29 ± .01	0.38 ± .05	0.30 ± .06	0.39 ± .02	0.31 ± .04	0.55 ± .08	0.32 ± .04
Average MAE	**0.402**	0.427	0.576	0.471	0.519	0.453	0.559	0.468

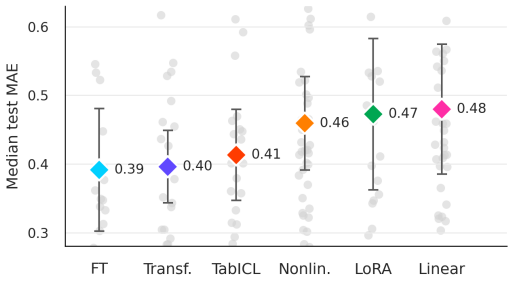
Downstream probe evaluation. Figure 3b summarizes the performance of Mol-JEPA embeddings using different downstream strategies. While full fine-tuning has the best median error, we find that a lightweight multimodal Transformer and TabICLv2 on CLS tokens are the strongest downstream models. More light-weight models, such as a linear or non-linear probes generally lead to worse performance, which suggests that the representations are too complex for these models. Finally, LoRA-finetuning of the Mol-JEPA transformer module under-performs most other methods. This is presumably due to hyperparameter sensitivity or overfitting.

Modality scaling in Mol-JEPA As shown in Figure 4a, training time increases approximately linearly with the number of modalities, although it also depends on the specific modalities used and their input dimensionalities. For instance, when pretraining on a dataset of 100k molecules for 300 epochs on a single GPU, using two modalities (CLOOME and Graph) requires roughly 9 hours, four modalities (adding UMA and MOE) require about 13 hours, and eight modalities (further adding ECFP, Boltz, BioXMol, and ChemGPT) increase the training time to approximately 27 hours. In all experiments, we observe a better performance when increasing the number of modalities, as annotated in Figure 4a and Appendix Figure 15. These results highlight the effectiveness of Mol-JEPA as a multimodal learning framework, showing that the integration of additional modalities consistently enhances downstream task performance.

Leave-one-modality-out analysis To better understand the contribution of individual modalities, we perform a leave-one-modality-out analysis. We observe only minor performance differences when removing a single modality at inference time, suggesting that the learned representations contain a certain degree of redundancy. This behavior is likely encouraged by the reconstruction objective, which promotes the encoding of overlapping information across modalities. To better quantify the contribution of individual modalities, we perform a leave-one-modality-out analysis during training on a 100k subset of the pretraining data. In each experiment, one modality is completely removed, allowing us to measure the impact of its full absence on representation learning. Figure 4 summarizes the resulting downstream errors, where larger error increases indicate a greater contribution of the removed modality to downstream performance. We find that the graph modality contributes the most to downstream performance. One possible expla-

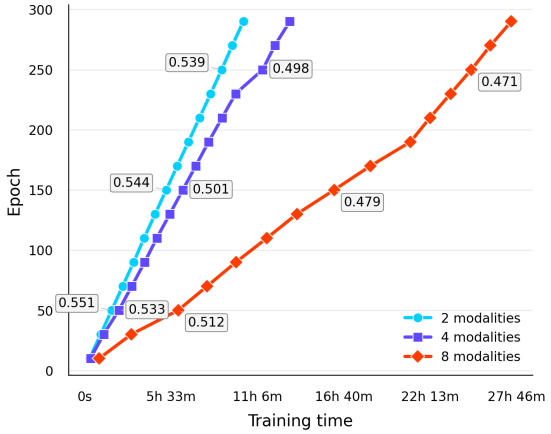


(a) **Comparison win rates**. The bars show how often a model wins across all pair-wise comparisons. While Mol-JEPA is best using MAE and R^2, it wins slightly less often when using Wilcoxon test for evaluation.

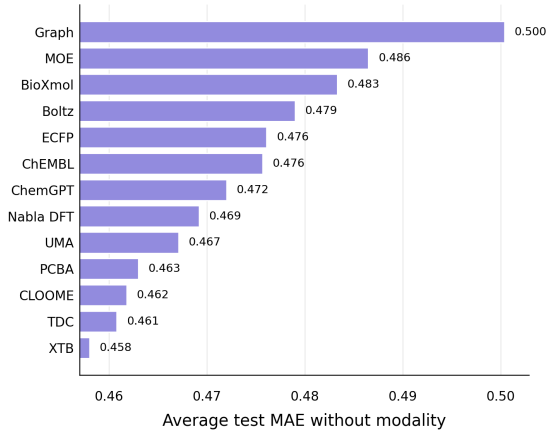


(b) **Downstream probe performance.** Besides full fine-tuning (FT), the Transformer probe on multimodal embeddings and TabICLv2 on CLS tokens are the best downstream predictors.

Figure 3: Mol-JEPA training results.



(a) **Performance increases with more modalities.** The lines represent the training time and epochs on a 100k subset dataset, annotated with the average downstream MAE.



(b) **Leave-one-modality-out analysis.** We analyze the effect of removing modalities during training. The reported scores are the average downstream probe MAEs.

Figure 4: Modality analysis

nation is that it is the only modality with a trainable GNN backbone, whereas the remaining modalities rely on frozen representations. We also find that the molecular fingerprint modalities MOE and ECFP4 rank among the most important modalities, indicating that their expert-designed features contain valuable information for downstream tasks. Lastly, we find that modalities capturing chemical knowledge through binding affinities and cellular interaction profiles also contribute to the learned representations, highlighting the value of incorporating biological context alongside molecular structure. We discuss these results and further experiments in Appendix F.

Multimodal training dynamics. Our proposed architecture enables stable multimodal training, and we observe a consistent relationship between the prediction loss, the SIGReg loss, and downstream performance, as illustrated in Figure 6. However, as shown in Figure 5, the training dynamics differ across modalities. While the effective rank increases consistently for all modality-specific embeddings, the prediction loss and isotropic regularization exhibit more heterogeneous behavior, suggesting that modalities contribute differently to the learning process. For instance, the experimental ChEMBL vectors first experience a rise in the prediction error and then improve steadily throughout training. In contrast, UMA embeddings show a decreasing prediction loss but oscillating regularization loss. We also find that sparse modalities are generally more difficult to optimize as their loss signal is less dominant compared to modalities that are always available. Lastly, we find that the inclusion of specific modalities, such as CLOOME can decrease the downstream performance. The experienced complexity of multimodal pretraining is in line with recent

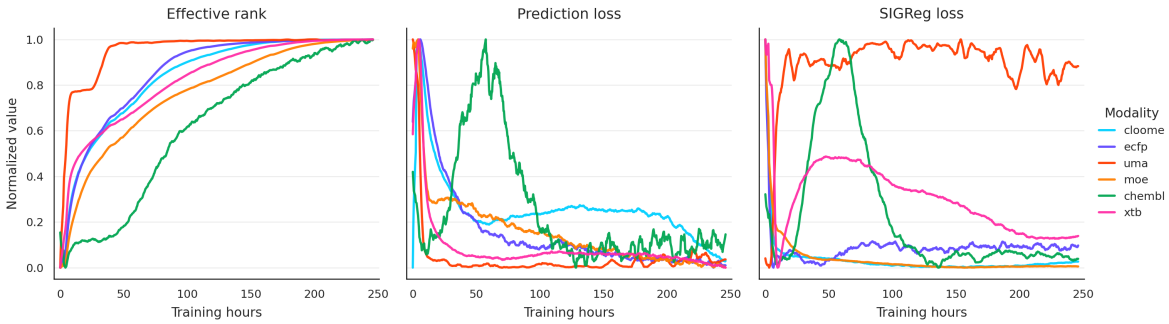


Figure 5: **Multimodal training dynamics.** Effective rank, prediction loss and SIGReg loss for different modalities during training. While effective rank is continuously increasing, the losses exhibit instabilities for some modalities.

works Kamai et al. (2026), showing that different modalities might require different approaches and not all might be suitable for joint embedding predictive architectures.

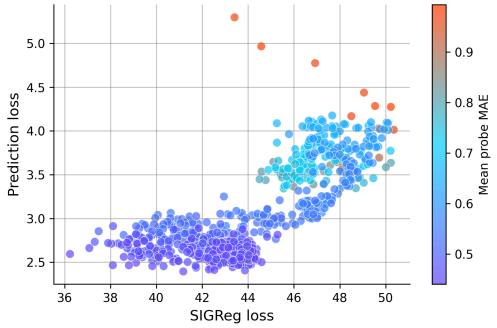


Figure 6: Losses and downstream error.

5 Conclusion

In this work, we presented **Mol-JEPA**, a multimodal foundation model that enables stable learning from many molecular modalities by predicting their latent representations. Using various high quality benchmarks, we show that Mol-JEPA achieves strong performance across different downstream tasks. Furthermore, we statistically compare the predictions and find that our model is particularly better on smaller datasets and in out-of-distribution settings. Additionally, we analyze the impact of individual modalities on training dynamics and downstream performance, suggesting that the selection of the most effective modality set remains an important direction for future research. Given the limited data availability in drug discovery, we believe this work represents an important step toward molecular world models that combine diverse sources of information to reason about physical dynamics and biochemical interactions. More broadly, Mol-JEPA demonstrates a general approach to multimodal representation learning that could be applied across scientific domains, including materials science, biology, and other data-scarce fields.

6 Limitations

There are several limitations and opportunities for future research. First, the current reconstruction objective does not account for the fact that some modalities may be inherently unable to recover information contained in others. Incorporating modality-specific dependencies and advanced masking strategies into the training objective may therefore be beneficial. Second, as with most machine learning models, out-of-distribution generalization remains a key challenge. Future work should investigate the robustness of Mol-JEPA when applied to molecules, tasks, and data distributions that differ substantially from those encountered during pretraining. Finally, our understanding of multimodal learning dynamics during training and inference remains limited. In particular, it is still unclear how individual modalities contribute to downstream performance and how factors such as data sparsity, modality dimensionality,

information content, and redundancy affect representation learning. Addressing these questions could lead to more effective modality selection and improved multimodal training strategies. Lastly, due to computational constraints, most ablation studies were conducted on a considerably smaller dataset than the full pretraining data, potentially impacting some conclusions.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Randall Balestriero, Hugues Van Assel, Sami BuGhanem, and Lucas Maes. stable-pretraining-v1: Foundation model research made simple. *arXiv preprint arXiv:2511.19484*, 2025.

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, 2019. doi: 10.1021/acs.jctc.8b01176.

Dominique Beaini, Shenyang Huang, Joao Alex Cunha, Zhiyi Li, Gabriela Moisescu-Pareja, Oleksandr Dymov, Samuel Maddrell-Mander, Callum McLean, Frederik Wenkel, Luis Müller, et al. Towards foundational models for molecular learning on large-scale multi-task datasets. *arXiv preprint arXiv:2310.04292*, 2023.

Christopher M Bishop, Markus Svensén, and Christopher KI Williams. Gtm: The generative topographic mapping. *Neural computation*, 10(1):215–234, 1998.

Joanne Bowes, Andrew J. Brown, Jacques Hamon, Wolfgang Jarolimek, Arun Sridhar, Gareth Waldron, and Steven Whitebread. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nature Reviews Drug Discovery*, 11(12):909–922, 2012. ISSN 1474-1776. doi: 10.1038/nrd3845.

Darko Butina. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39 (4):747–750, 1999.

Maria Castellanos and Hugo MacDermott-Opeskin. Lessons learned from the openadmet-expansionrx blind challenge: Can we trust zero-shot admet predictions? lessons and reflections from the openadmet-expansionrx blind challenge—evaluating whether zero-shot admet predictions can be trusted in real drug discovery settings.

Chemical Computing Group. Molecular Operating Environment (MOE), 2022. URL `https://www.chemcomp.com`.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

John D Chodera, W Patrick Walters, Sriram Kosuri, and James S Fraser. Blind challenges let us see the path forward for predictive models. *Journal of Chemical Information and Modeling*, 66(4):1947–1949, 2026.

Markus Dablander, Thierry Hanser, Renaud Lambiotte, and Garrett M Morris. Exploring qsar models for activity-cliff prediction. *Journal of Cheminformatics*, 15(1):47, 2023.

Cheng Fang, Ye Wang, Richard Grater, Sudarshan Kapadnis, Cheryl Black, Patrick Trapa, and Simone Sciabola. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling*, 63(11):3263–3274, 2023.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Yaëlle Fischer, Thibaud Southiratn, Dhoha Triki, and Ruel Cedeno. Deep learning vs classical methods in potency and adme prediction: Insights from a computational blind challenge. *Journal of Chemical Information and Modeling*, 65(24):13115–13131, 2025.

James S Fraser, Steven Edgar, L Naomi Handly, Sriram Kosuri, John D Chodera, Mark Murcko, and W Patrick Walters. Mapping the avoid-ome: a systematic open-science approach to predictive admet. *Nature Communications*, 17(1): 4644, 2026.

Nathan C Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gomez-Bombarelli, Connor W Coley, and Vijay Gadepally. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11):1297–1305, 2023.

Anna Gaulton, Anne Hersey, Michał Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

William Green, Jackson Burns, Akshat Shirish Zalte, Charlles Abreu, Jochen Sieg, Christian Feldmann, and Miriam Mathea. Deep learning foundation models from classical molecular descriptors. 2026.

Francesca Grisoni. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology*, 79:102527, 2023.

Oleksandr O. Grygorenko, Dmytro S. Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E. Gubina, and Yurii S. Moroz. Generating multibillion chemical space of readily accessible screening compounds. *iScience*, 23(11): 101681, 2020. doi: 10.1016/j.isci.2020.101681.

Jinjiang Guo. Do larger models really win in drug discovery? a benchmark assessment of model scaling in ai-driven molecular property and activity prediction. *arXiv preprint arXiv:2604.26498*, 2026.

Hakim Hafidi, Mounir Ghogho, Philippe Ciblat, and Ananthram Swami. Graphcl: Contrastive self-supervised learning of graph representations. *arXiv preprint arXiv:2007.08025*, 2020.

Yonatan Harnik, Hadas Shalit Peleg, Amit H Bermano, and Anat Milo. Data efficient molecular image representation learning using foundation models. *Chemical Science*, 16(24):10833–10841, 2025.

Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J McGill. Chemprop: a machine learning package for chemical property prediction. *Journal of chemical information and modeling*, 64(1):9–17, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

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

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.

Karthik Iyer and Nasser Sabar. M-jepa: Predictive self-supervised learning for molecular graphs with scaffold-shift evaluation on tox21. *Journal of Chemical Information and Modeling*, 2026.

Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of cheminformatics*, 13(1):12, 2021.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Ilay Kamai, Hugues Van Assel, Aviv Regev, Hagai B Perets, and Randall Balestriero. When to align, when to predict: A phase diagram for multimodal learning. *arXiv preprint arXiv:2606.11190*, 2026.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Kuzma Khrabrov, Anton Ber, Artem Tsypin, Konstantin Ushenin, Egor Rumiantsev, Alexander Telepov, Dmitry Protasov, Ilya Shenbin, Anton Alekseev, Mikhail Shirokikh, et al. Nabla dft: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials. *Advances in Neural Information Processing Systems*, 37:36869–36889, 2024.

Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1): D1202–D1213, 2016.

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

Oliver Kramer. Scikit-learn. In *Machine learning for evolution strategies*, pp. 45–53. Springer, 2016.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1): 1–62, 2022.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Yunchao Liu, Ha Dong, Xin Wang, Rocco Moretti, Yu Wang, Zhaoqian Su, Jiawei Gu, Bobby Bodenheimer, Charles David Weaver, Jens Meiler, and Tyler Derr. WelQrate: Defining the gold standard in small molecule drug discovery benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks*, 2024. URL https://arxiv.org/abs/2411.09820.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*, 2023.

Hugo MacDermott-Opeskin, Jenke Scheen, Cas Wognum, Joshua T Horton, Devany West, Alexander Matthew Payne, Maria A Castellanos, Sean Colby, Edward Griffen, David Cousins, et al. A computational community blind challenge on pan-coronavirus drug discovery data. *Journal of Chemical Information and Modeling*, 66(6):3129–3149, 2026.

Andrei Manolache, Dragos Tantaru, and Mathias Niepert. Molmix: a simple yet effective baseline for multimodal molecular representation learning. *arXiv preprint arXiv:2410.07981*, 2024.

Muhammad Arslan Masood, Markus Heinonen, and Samuel Kaski. Multi-modal representation learning for molecules. In *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*, 2025.

Andrea Mauri. alvadesc: A tool to calculate and analyze molecular descriptors and fingerprints. In *Ecotoxicological QSARs*, pp. 801–820. Springer, 2020.

Oscar Méndez-Lucio, Christos A Nicolaou, and Berton Earnshaw. Mole: a foundation model for molecular graphs using disentangled attention. *Nature communications*, 15(1):9431, 2024.

Adrian Mirza, Sebastian Starke, Erinc Merdivan, and Kevin Maik Jablonka. Bridging chemical modalities by aligning embeddings. In *AI for Accelerated Materials Design-Vienna 2024*, 2024.

Mikolaj Mizera, Arkadii Lin, Eugene Babin, Yury Kashkur, Tatiana Sitnik, Ien An Chan, Arsen Yedige, Maksim Vendin, Shamkhal Baybekov, and Vladimir Aladinskiy. Graph transformer foundation model for modeling admet properties. 2024.

Mark A Murcko. The affinity advantage. *Journal of Medicinal Chemistry*, 69(3):1963–1969, 2026.

Mark Orester. Counterfactual planning in latent chemical space. 2025.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Francesco Piccoli, Gabriel Vogel, and Jana M Weber. Joint embedding predictive architecture for self-supervised pretraining on polymer molecular graphs. *Digital Discovery*, 2026.

Mateusz Praski, Jakub Adamczyk, and Wojciech Czech. Benchmarking pretrained molecular embedding models for molecular representation learning. *arXiv preprint arXiv:2508.06199*, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Anthony K Rappé, Carla J Casewit, Kent S Colwell, William A Goddard III, and W Mason Skiff. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society*, 114(25):10024–10035, 1992.

Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, et al. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1):93, 2022.

Tiago Rodrigues. The good, the bad, and the ugly in chemical and biological data for machine learning. *Drug Discovery Today: Technologies*, 32:3–8, 2019.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

Maxime Sanchez, Nicolas Bourriez, Ihab Bendidi, Ethan Cohen, Ivan Svatko, Elaine Del Nery, Hamza Tajmouati, Guillaume Bollot, Laurence Calzone, and Auguste Genovesio. Large scale compound selection guided by cell painting reveals activity cliffs and functional relationships. *Communications Biology*, 2026.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, July 2023.

Robert P Sheridan, Prabha Karnachi, Matthew Tudor, Yuting Xu, Andy Liaw, Falgun Shah, Alan C Cheng, Elizabeth Joshi, Meir Glick, and Juan Alvarez. Experimental error, kurtosis, activity cliffs, and methodology: What limits the predictivity of quantitative structure–activity relationship models? *Journal of chemical information and modeling*, 60(4):1969–1982, 2020.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*, 2020.

Dagmar Stumpfe, Huabin Hu, and Jurgen Bajorath. Evolving concept of activity cliffs. *ACS omega*, 4(11):14360–14368, 2019.

Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062, 2022a.

Ruoxi Sun, Hanjun Dai, and Adams Wei Yu. Does gnn pretraining help molecular representation? *Advances in Neural Information Processing Systems*, 35:12096–12109, 2022b.

Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. *arXiv preprint arXiv:2505.12477*, 2025.

W. Patrick Walters. We need better benchmarks for machine learning in drug discovery. Practical Cheminformatics (blog), August 2023. URL `http://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html`.

Liangliang Wang, Junjie Ding, Li Pan, Dongsheng Cao, Hui Jiang, and Xiaoqin Ding. Quantum chemical descriptors in quantitative structure–activity relationship models and their applications. *Chemometrics and Intelligent Laboratory Systems*, 217:104384, 2021.

Shuzhe Wang, Jagna Witek, Gregory A Landrum, and Sereina Riniker. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *Journal of chemical information and modeling*, 60(4):2044–2058, 2020.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

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.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Brandon Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John Kitchin, Daniel Levine, et al. Uma: A family of universal models for atoms. *Advances in Neural Information Processing Systems*, 38:129391–129427, 2026.

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 pre-training for molecular representation learning. *Nature Communications*, 2026.

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.

APPENDICES

A Dataset details **17**

B Implementation details **18**

C Baseline models **20**

D Results and Model evaluation **20**

E Pretraining data scaling **23**

F Ablation studies **23**

G Representational similarity **24**

H In-context learning comparison **25**

A DATASET DETAILS

ChEMBL processing We construct a feature matrix by pivoting ChEMBL bioassay activity data, representing each molecule as a vector of assay measurements. To ensure sufficient coverage of individual assay endpoints while maintaining a diverse set of biological readouts, we analyze the trade-off between assay coverage and data sparsity. As shown in Table 4, increasing the minimum number of required measurements per assay reduces the number of retained endpoints. Based on this analysis, we select a threshold of 30 measurements per assay, resulting in a feature vector with 306 dimensions. The retained assays encompass a broad range of ADMET and physicochemical properties, including microsomal stability, solubility, hERG channel inhibition, receptor binding activities, and other pharmacologically relevant endpoints across multiple species.

Table 4: Assay threshold filtering.

Threshold	Remaining assays
1	37,178
10	904
20	349
50	103
100	34
500	17
1000	24

Therapeutics Data Commons processing Table 5 provides an overview of all processed datasets. Analogous to the ChEMBL data, measurements are aggregated at the molecule level, resulting in a feature vector of all available experimental observations. Since several datasets contain multiple endpoints, the resulting feature vectors have 672 dimensions. The number of available measurements varies substantially across datasets, ranging from as few as 50 observations for certain toxicity endpoints to more than 300,000 measurements in high-throughput screening (HTS) datasets.

Table 5: TDC dataset summary

Dataset	**Description**	**Category**
Caco-2	Cell permeability measure	ADME
PAMPA	Parallel artificial membrane permeability	ADME
HIA	Human intestinal absorption classification	ADME
P-gp	P-glycoprotein inhibition activity	ADME
Bioavailability	Rate and extent of drug absorption	ADME
Lipophilicity	Lipid solubility measure	ADME
Solubility (AqSolDB)	Aqueous solubility in water	ADME
Hydration Free Energy	Solvation energy in water	ADME
BBB Penetration	Blood-Brain Barrier passage	ADME
PPBR	Plasma protein binding percentage	ADME
VDss	Volume of distribution at steady state	ADME
CYP Inhibition	Cytochrome P450 inhibition (1A2, 2C9, 2C19, 2D6, 3A4)	ADME
CYP Substrates	Cytochrome P450 substrate activity (2C9, 2D6, 3A4)	ADME
Half-Life	Elimination half-life in body	ADME
Hepatocyte Clearance	Hepatic metabolic clearance rate	ADME
LD50	Median lethal dose (acute toxicity)	Tox
hERG Cardiotoxicity	Potassium channel blockade & cardiac risk	Tox
Ames Test	Bacterial mutagenicity assay	Tox
DILI	Drug-induced liver injury risk	Tox
Skin Reaction	Dermal allergic response	Tox
Carcinogenicity	Cancer-promoting potential	Tox
Tox21 / ToxCast	High-throughput nuclear & pathway toxicity	Tox
ClinTox	Clinical trial failures due to toxicity	Tox
SARS-CoV-2	In-vitro viral inhibition & 3CLPro protease screen	HTS
HIV	HIV replication inhibition	HTS
Orexin-1 Receptor	Target binding & modulation	HTS
M1 Muscarinic	Receptor agonist & antagonist activity	HTS
Ion Channels	Potassium (Kir2.1, KCNQ2) and Calcium (Cav3) channels	HTS
Choline Transporter	Neurotransporter binding	HTS
STK33 & TDP1	Kinase and phosphodiesterase enzymatic screens	HTS

Benchmark data processing All benchmark datasets are downloaded and used in the form provided by the challenge organizers, as the majority of the required preprocessing has already been performed. For ExpansionRx, it was

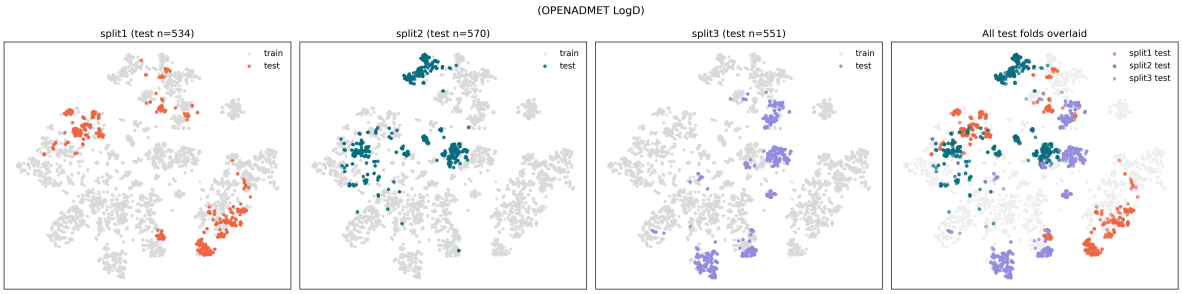


Figure 7: Cluster splits t-SNE visualization for ExpansionRx LogD dataset using ECFP4 fingerprints.

Table 6: Train/test sample counts for the three cluster splits of each benchmark task.

Dataset	Task	Total	Split 1		Split 2		Split 3	
			Train	Test	Train	Test	Train	Test
ASAP ADMET	HLM Clearance	166	148	18	153	13	153	13
	Kinetic Solubility	214	183	31	186	28	185	29
	LogD	203	171	32	163	40	169	34
	MDR1 Efflux	242	210	32	205	37	213	29
	MLM Clearance	189	150	39	157	32	145	44
ASAP Potency	MERS M^{pro}	421	383	38	366	55	366	55
	SARS M^{pro}	356	327	29	307	49	335	21
Biogen ADME	HLM Clearance	3087	2983	104	3012	75	3012	75
	MDR1 Efflux	2642	2565	77	2567	75	2566	76
	Human PPB	194	169	25	174	20	173	21
	Rat PPB	168	148	20	148	20	148	20
	RLM Clearance	3054	2947	107	2930	124	2950	104
	Solubility	2173	2075	98	2060	113	2048	125
OpenADMET	Caco-2 Efflux Ratio	2161	1822	339	1901	260	1953	208
	Caco-2 $P_{\text{app},A\to B}$	2156	1918	238	1942	214	1954	202
	HLM Clearance	3595	3152	443	3186	409	3169	426
	Kinetic Solubility	5128	4520	608	4632	496	4608	520
	LogD	5039	4505	534	4469	570	4488	551
	Mouse Brain Unbound Fraction	973	853	120	860	113	797	176
	Mouse Gut Homogenate Binding	221	206	15	199	22	177	44
	MLM Clearance	4375	3758	617	3847	528	3720	655
	Mouse Plasma Unbound Fraction	1292	1159	133	1145	147	1146	146
	PXR (pEC_{50})	4288	4043	245	4055	233	4052	236

additionally necessary to convert some endpoints to a logarithmic scale by applying $\log_{10}(x + 0.001)$, consistent with the preprocessing used in the competition, where an epsilon of 0.001 was added prior to the transformation.

Dataset splitting In Table 6 we provide the number of samples for each benchmark dataset along with the dataset sizes for each split. We move specific clusters to the test data and keep the remaining ones in the training data. In Figure 7, we visualize the three splits on the OpenADMET LogD dataset. Although Butina clustering at a Tanimoto similarity threshold of 0.65 reduces structural overlap between training and test sets, it does not strictly enforce a maximum cross-split similarity of 0.65 and some molecule pairs assigned to different clusters may still exceed this threshold. To assess the impact of such cases, we additionally applied a strict similarity-based filtering procedure to remove all train-test pairs above the threshold and found that this had no substantial effect on the experimental results.

B Implementation Details

Hyperparameter tuning Table 7 summarizes the hyperparameter search space used for tuning Mol-JEPA. Hyperparameter optimization was performed on a subset of 100,000 molecules by evaluating 300 randomly sampled configurations in parallel and selecting the model with the best online probe performance. The most important architectural

Component	Hyperparameter	Search Space	Best Value
Optimizer	Learning rate	{0.01, 0.005, 0.001, 0.0005}	0.0005
Optimizer	Weight decay	{0.0, 0.0001, 0.001, 0.01}	0.01
Mol-JEPA Transformer	Hidden dimension	{32, 128, 512}	512
Mol-JEPA Transformer	Masking ratio	{0.1, 0.3, 0.6, 0.9}	0.3
Mol-JEPA Transformer	λ	{0.1, 0.3, 0.5, 0.7}	0.7
Mol-JEPA Transformer	Dropout	{0.1, 0.4, 0.7}	0.1
Graph Encoder	Number of layers	{1, 2, 3}	3
Graph Encoder	Pooling	{`mean`, `max`}	`max`
Graph Encoder	Attention heads	{2, 4, 8}	8
Embedding Encoder	Number of layers	{1, 2, 3}	3
Atom Encoder	Number of layers	{1, 2, 3}	1
Atom Encoder	Attention heads	{2, 4, 8}	4
Data Loader	Batch size	{32, 64, 128}	128

Table 7: Hyperparameter search space used for model tuning.

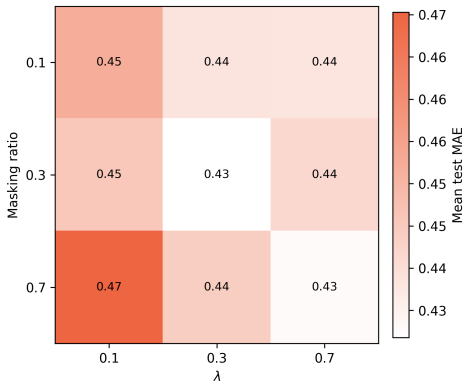


Figure 8: Loss parameters performance.

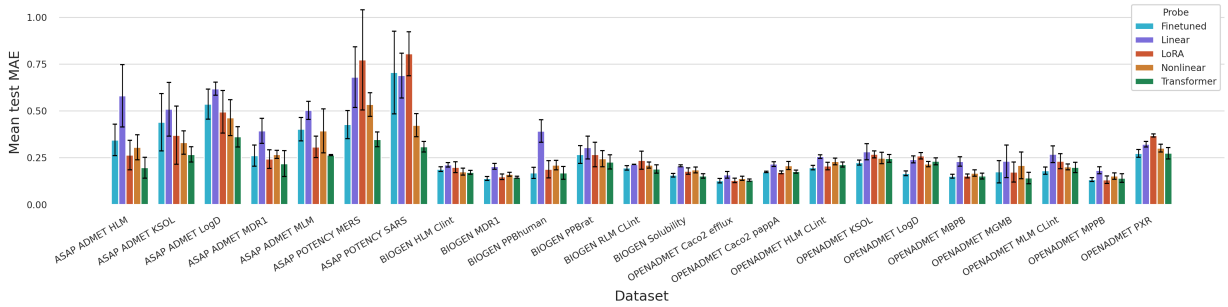


Figure 9: Mean absolute error of different probe variants across datasets.

hyperparameters include the dimensions of the modality-specific encoders, the structure of the Mol-JEPA aggregation module, and standard neural network training parameters. In addition, we explored different masking ratios and values of the loss-balancing parameter λ. As shown in Figure 8, performance deteriorates for small values of λ, whereas no consistent relationship between masking ratio and downstream performance could be observed.

Probing model details We evaluate several strategies for leveraging Mol-JEPA representations in downstream prediction tasks. The most straightforward approach is to use the CLS token as a global representation of the molecule. This 512-dimensional embedding is provided as input to either linear or nonlinear neural network predictors. Additionally, we employ TabICLv2 as a more expressive downstream model, using $n_{\text{estimators}} = 8$. A second strategy utilizes the modality-specific embeddings predicted by Mol-JEPA. Beyond using the final-layer embeddings, we investigate whether intermediate representations contain complementary information by extracting features prior to the final readout layer. We further evaluate a concatenated representation that combines intermediate and final-layer embeddings. As a third approach, we fine-tune the Mol-JEPA transformer module using low-rank adaptation (LoRA). For all fine-tuning experiments, we set the LoRA rank to 16, train for 60 epochs, and use a learning rate of 5×10^{-5}. Figure 9 compares the performance of the different feature extraction strategies, while Figure 10 reports the predictive performance of the downstream models. Overall, the best results are obtained either by combining the CLS token representation with TabICLv2 or by using multimodal embeddings with the transformer-based downstream model.

Computational requirements Mol-JEPA is trained on 12 RTX-6000ADA GPUs with parallelized data loading using 8 CPUs and 128GB of memory. The model is optimized with single precision (float32) and we use 2 gradient accumulation steps along with an aggregated batch size of 1024. The training converges after 3 days at around 150 epochs. After this point, we observe overfitting and degradation of online probe performance.

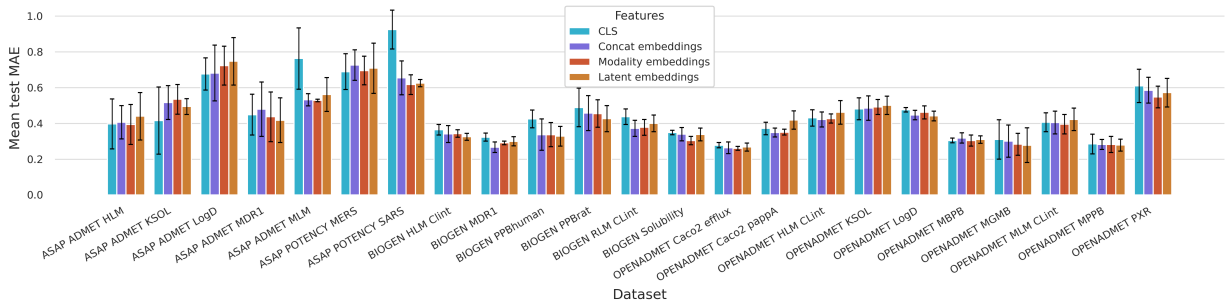


Figure 10: Mean absolute error of different feature variants across datasets.

C Baseline models

Table 8 summarizes the hyperparameter search spaces considered for tuning the baseline models. All feature-based models were trained using ECFP4 fingerprints, AlvaDesc descriptors, and 3D pharmacophore fingerprints as input representations. To reduce the risk of overfitting, the number of AlvaDesc descriptors (nearly 6000) was limited to at most one quarter of the dataset size. Feature selection was performed using Scikit-learn's `SelectKBest` method with both mutual information and F-test scores, after which the selected descriptor sets were combined. For each dataset split, hyperparameter optimization was conducted using five-fold cross-validation on the training data, and the best-performing model was subsequently evaluated on the corresponding test split.

Table 8: Hyperparameter search space used for baseline model tuning.

Model	**Hyperparameter**	**Search Space**
Random Forest	Number of estimators	{100, 200, 300, 400, 500}
Random Forest	Criterion	{`squared_error`, `absolute_error`, `friedman_mse`, `poisson`}
Random Forest	Maximum depth	{5, 7}
LightGBM	Number of estimators	{100, 200, 300, 400, 500}
LightGBM	Learning rate	{0.01, 0.05, 0.1, 0.2}
LightGBM	Maximum depth	{5, 7}
XGBoost	Number of estimators	{100, 200, 300, 400, 500}
XGBoost	Learning rate	{0.01, 0.05, 0.1, 0.2}
XGBoost	Maximum depth	{5, 7}
Chemprop	Depth	{2, 3, 4}
Chemprop	Message hidden dimension	{300, 600, 1200}
Chemprop	Dropout	$\mathcal{U}(0, 0.5)$
Chemprop	Aggregation	{`sum`, `mean`, `norm`}
Chemprop	FFN hidden dimension	{300, 600, 1200}
Chemprop	Number of FFN layers	{1, 2, 3}
Chemeleon	FFN hidden dimension	{300, 600, 1200}
Chemeleon	Number of FFN layers	{1, 2, 3}
TabICLv2	Number of estimators	{8, 16, 32}
TabICLv2	Normalization method	{`none`, `power`, `quantile`, `quantile_rtdl`, `robust`}
CLAMP	Probe model	{`linear`, `nonlinear`}

D Results and Model Evaluation

Additional results Table 9 reports the benchmark results using R^2 as evaluation metric. Consistent with the previous findings, Mol-JEPA achieves the strongest overall performance on average across datasets. Figures 12a and 12b high-

Table 9: R^2 comparison across benchmark datasets. Negative R^2 values are clipped to zero.

Dataset	Mol-JEPA Best	Mol-JEPA Embeddings	CLAMP Nonlinear	CheMeleon Finetuned	Chemprop GNN	TabICLv2 AlvaDesc	RF ECFP4	LGBM AlvaDesc
ExpansionRx								
Caco-2 Permeability	**0.43** ± .06	0.37 ± .11	0.24 ± .18	0.29 ± .15	0.22 ± .20	0.14 ± .15	0.13 ± .07	0.14 ± .12
Caco-2 Efflux	0.30 ± .10	0.22 ± .16	0.06 ± .24	0.16 ± .12	0.23 ± .11	**0.38** ± .09	0.08 ± .03	0.37 ± .10
LogD	**0.86** ± .01	0.74 ± .07	0.69 ± .09	0.85 ± .01	0.74 ± .03	0.83 ± .06	0.41 ± .09	0.79 ± .05
KSOL	0.51 ± .06	0.38 ± .11	0.41 ± .07	**0.53** ± .08	0.19 ± .26	0.51 ± .06	0.27 ± .05	0.47 ± .04
HLM CLint	0.29 ± .06	0.18 ± .10	0.00 ± .00	**0.31** ± .16	0.00 ± .00	0.29 ± .07	0.07 ± .09	0.30 ± .02
MLM CLint	0.31 ± .22	0.31 ± .25	**0.42** ± .18	0.35 ± .24	0.40 ± .31	0.26 ± .20	0.21 ± .32	0.37 ± .27
MBPB	0.65 ± .08	0.63 ± .14	0.64 ± .03	0.66 ± .11	0.60 ± .08	**0.72** ± .10	0.30 ± .14	0.56 ± .15
MGMB	0.42 ± .28	0.38 ± .22	0.30 ± .27	**0.44** ± .29	0.30 ± .27	0.35 ± .48	0.11 ± .22	0.35 ± .40
MPPB	**0.62** ± .05	0.54 ± .09	0.46 ± .05	0.48 ± .15	0.41 ± .18	**0.62** ± .04	0.15 ± .24	0.51 ± .11
ASAP								
MERS-CoV-2 Potency	0.07 ± .12	**0.19** ± .24	0.00 ± .00	0.00 ± .00	0.00 ± .00	0.14 ± .12	0.03 ± .14	0.14 ± .03
SARS-CoV-2 Potency	0.19 ± .26	0.12 ± .23	0.00 ± .00	0.00 ± .00	0.00 ± .00	**0.20** ± .35	0.00 ± .00	0.13 ± .38
LogD	0.57 ± .08	0.53 ± .14	0.00 ± .00	0.31 ± .23	0.36 ± .23	**0.64** ± .14	0.13 ± .17	0.52 ± .13
KSOL	**0.16** ± .35	**0.16** ± .52	0.00 ± .00	0.00 ± .00	0.00 ± .00	0.00 ± .00	0.00 ± .00	0.00 ± .00
HLM	**0.18** ± .38	**0.18** ± .38	0.00 ± .00	0.11 ± .35	0.00 ± .00	0.11 ± .35	0.08 ± .17	0.09 ± .20
MLM	0.00 ± .00	0.00 ± .00	0.00 ± .00	0.00 ± .00	0.00 ± .00	0.00 ± .00	**0.10** ± .13	0.00 ± .00
MDR1 Efflux	**0.11** ± .09	0.09 ± .25	0.00 ± .00	0.00 ± .00	0.00 ± .00	0.10 ± .11	0.00 ± .00	0.00 ± .00
PXR								
PXR Activity	**0.66** ± .08	0.63 ± .12	0.41 ± .13	0.57 ± .23	0.52 ± .17	0.20 ± .43	0.31 ± .26	0.22 ± .39
Biogen ADME								
Solubility	**0.50** ± .06	0.44 ± .07	0.33 ± .05	0.33 ± .09	0.27 ± .15	0.44 ± .01	0.18 ± .08	0.41 ± .06
HLM CLint	**0.54** ± .09	0.50 ± .14	0.33 ± .06	0.51 ± .13	0.44 ± .09	**0.54** ± .06	0.12 ± .12	**0.54** ± .11
RLM CLint	0.58 ± .09	0.57 ± .06	0.42 ± .00	0.57 ± .07	0.49 ± .07	0.58 ± .06	0.15 ± .11	0.57 ± .08
HPPB	**0.73** ± .07	0.68 ± .07	0.51 ± .12	0.51 ± .08	0.47 ± .08	0.68 ± .09	0.12 ± .17	0.57 ± .14
RPPB	0.43 ± .27	0.41 ± .20	**0.47** ± .21	0.27 ± .24	0.20 ± .23	0.28 ± .40	0.00 ± .00	0.14 ± .42
MDR1 Efflux	**0.60** ± .07	0.56 ± .10	0.43 ± .07	0.55 ± .11	0.53 ± .15	0.53 ± .21	0.12 ± .12	0.51 ± .24
Average R^2	**0.42**	0.38	0.27	0.34	0.28	0.37	0.13	0.33

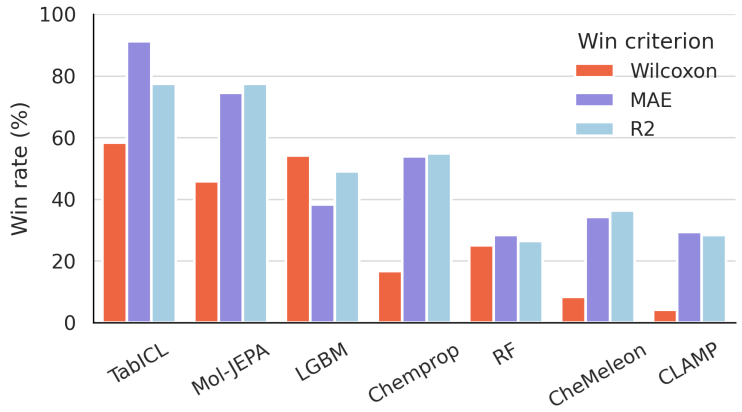
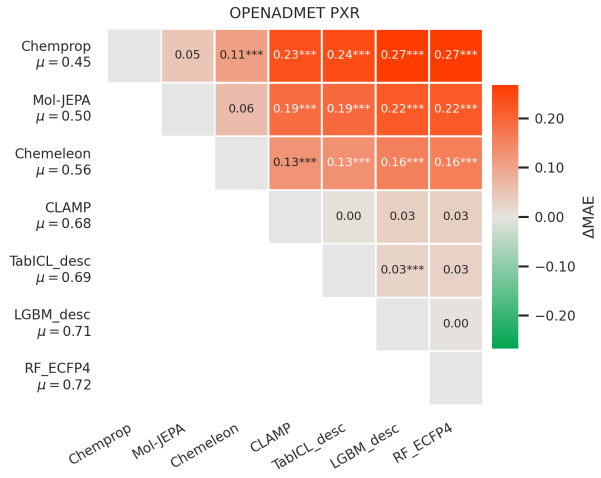


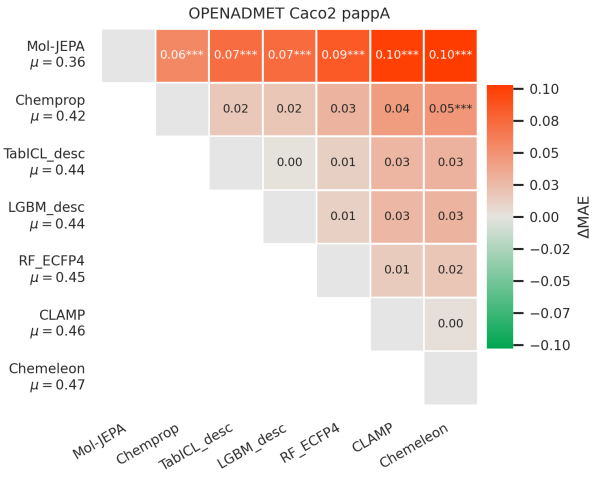
Figure 11: **Win rates on public competition splits**

light two representative examples on the cluster splits, demonstrating Mol-JEPA's ability to significantly outperform the evaluated baseline methods in predictive accuracy.

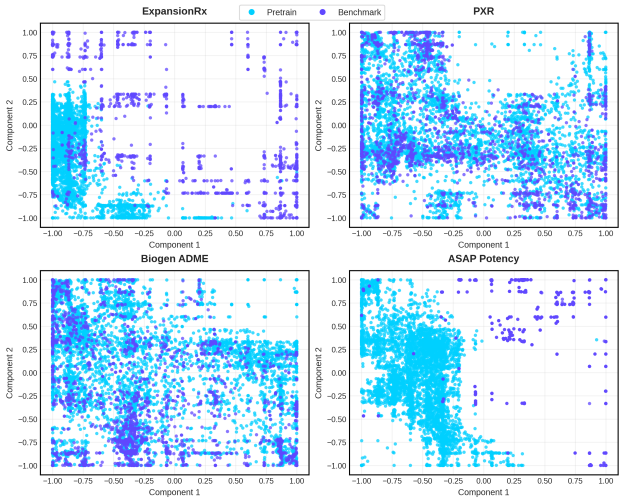
We further evaluate model win rates on the public splits of ExpansionRX, ASAP, and PXR. These splits are generated using a temporal strategy, which, according to the competition organizers, better represents realistic drug discovery settings by preserving the chronological order of data generation. As shown in Figure 11, TabICLv2 using AlvaDesc descriptors achieves higher Wilcoxon and MAE win rates than Mol-JEPA. However, Mol-JEPA delivers the strongest performance in terms of R^2. While these results indicate that there is still room for improvement through the incorporation of larger pretraining datasets and additional modalities, Mol-JEPA already demonstrates highly competitive performance. In particular, it consistently surpasses other foundation model approaches, which achieve substantially lower scores on both public and cluster-based splits.



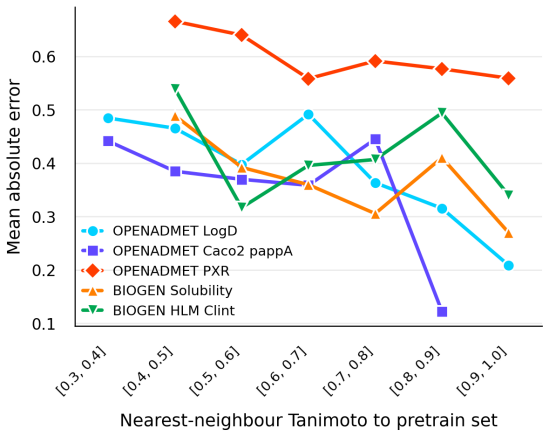
(a) Pairwise statistical comparison using Wilcoxon test.



(b) Pairwise statistical comparison using Wilcoxon test.



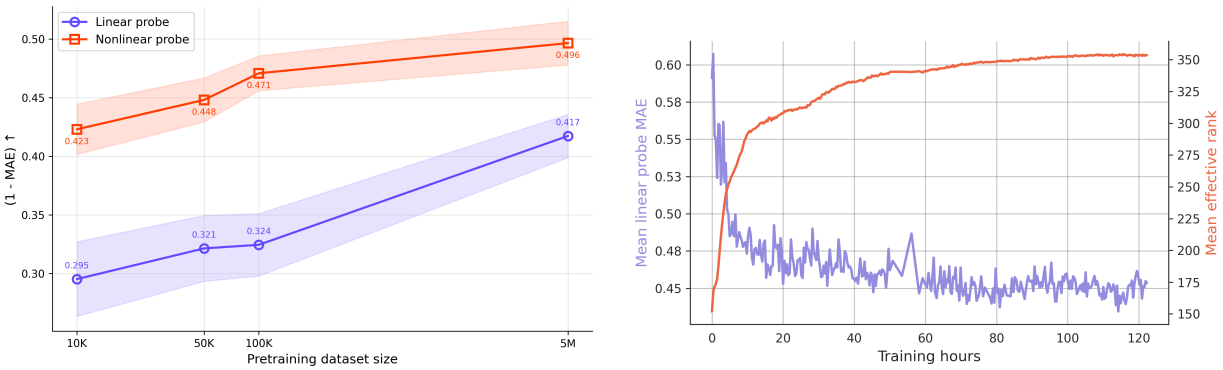
(a) **Generative Topographical Mappings** for different benchmark families and pretraining data.



(b) **Pretraining data coverage determines downstream performance**.

Figure 13: Pretraining data analysis

Molecular space coverage To understand the role of the pretraining data for the downstream tasks, we analyze the molecular similarity using generative topographic mappings (GTM) Bishop et al. (1998). We randomly sample 5,000 molecules from the pretraining corpus and 2,000 molecules from the benchmark datasets, convert them into ECFP4 fingerprints, and project them into a shared GTM space. The resulting visualization, shown in Figure 13a, reveals that while some benchmark datasets are well represented within the pretraining distribution, others occupy more distant regions of chemical space. Notably, Mol-JEPA tends to achieve stronger performance on benchmark families with more pretraining data coverage, suggesting that representation of relevant chemical space during pretraining contributes to downstream predictive accuracy. To further investigate this relationship, we selected five datasets and computed, for each molecule, its distance to the full pretraining corpus using ECFP4 fingerprints. Figure 13b illustrates the relationship between prediction error and binned molecular similarity. The results show a clear trend of increasing Mol-JEPA prediction error with increasing distance from the pretraining data, indicating that molecules located further from the pretraining distribution are more challenging to predict accurately. These findings suggest that expanding the size and diversity of the pretraining corpus could further improve model performance by increasing coverage of relevant chemical space.



(a) **Effect of pretraining dataset size on performance.** (b) **Effective rank and downstream performance.**

Figure 14: Training configurations

E PRETRAINING DATA SCALING

We analyze the effect of pretraining dataset size on downstream performance. As shown in Figure 14a, performance steadily improves with increasing pretraining data for both linear and nonlinear probes across all downstream datasets (reflected by the confidence bands). To better highlight this trend, performance is reported as 1-MAE, where higher values indicate better performance. The observed scaling behavior suggests that Mol-JEPA benefits from additional pretraining data and further increases in dataset size are expected to yield additional gains in downstream performance.

F ABLATION STUDIES

We perform several experiments to assess the sensitivity of different parameter and architectural choices, which are summarized in Table 10. For computational reasons, the experiments are conducted using a subset of 100,000 randomly sampled data points.

Experimental modalities. Overall, we find that the performance differences introduced by the experimental modalities are moderate. This is likely because the modality data contains many missing values, imbalanced distributions, and potentially noisy measurements, limiting the amount of useful information that can be extracted. To better understand how this information should be incorporated, we conduct two additional ablation studies. First, we investigate whether modality information should be used as labels in a semi-supervised setting or incorporated through latent prediction within the JEPA framework. As summarized in Table 10, the semi-supervised alternative consistently underperforms JEPA across both linear and non-linear probing settings, increasing the MAE from 0.436 to 0.467 and from 0.393 to 0.417, respectively. These results suggest that JEPA more effectively exploits the available modality information, supporting the hypothesis that learning predictive latent representations is a more robust strategy for integrating noisy and incomplete experimental data than directly predicting modality labels. Second, we assess the impact of removing the experimental modalities entirely. In this setting, downstream performance consistently deteriorates, demonstrating that the modalities still provide valuable information that improves the quality and transferability of the learned molecular representations (MAE 0.436 vs. 0.449 and 0.393 vs. 0.407).

Regularization There are several ways to apply the SIGReg loss within the JEPA framework. First, isotropy can be enforced on the CLS embeddings. Second, the loss can be applied directly to the predicted embeddings. Third, it can be applied to the target embeddings. We observe training collapse when jointly optimizing the prediction objective and the SIGReg regularization on the predicted embeddings. In contrast, applying the regularization to the target embeddings yields stable training and the best downstream performance. These results suggest that regularizing the target representations encourages a well-structured latent space while allowing the predictor to focus on the alignment objective.

Loss projection Similar to contrastive learning approaches such as SimCLR, we investigate whether applying the training objective in a dedicated projection space is beneficial. As shown in Table 10, removing the projection head leads to improved downstream performance for both the linear probe (0.468 vs. 0.436) and the non-linear probe (0.488

Table 10: Model architecture and training ablations on a subset of 100k. The reported numbers are the best downstream MAEs after training up to 1000 epochs, averaged across all benchmark datasets.

Group	Variant	Linear probes mean	Nonlinear probes mean
Modalities			
	All modalities	**0.436**	**0.393**
	Semi-Supervised	0.467	0.417
	No experimental	0.449	0.407
Regularization			
	SIGReg on targets	**0.436**	**0.393**
	SIGReg on CLS	0.445	0.406
Loss projection			
	W/o projection	**0.436**	**0.393**
	W/ projection	0.468	0.488
Embedding dimension			
	256	0.483	0.465
	512	0.436	**0.393**
	1024	**0.435**	0.397

vs. 0.393). These results suggest that, unlike in contrastive learning, introducing a separate projection space is not advantageous in our JEPA-based setting. A possible explanation is that the JEPA objective already promotes informative latent representations, making an additional projection head unnecessary and potentially causing the backbone representations to lose information relevant for downstream tasks.

Embedding dimension. We further investigate the impact of the embedding dimension on downstream performance. As shown in Table 10, increasing the embedding dimension from 256 to 512 substantially improves performance, reducing the MAE for both the linear probe (0.483 to 0.436) and the non-linear probe (0.465 to 0.393). Further increasing the dimension to 1024 yields nearly identical results (0.435 and 0.397, respectively), indicating diminishing returns from additional model capacity. Overall, these findings suggest that larger embedding dimensions are beneficial up to a certain point, after which performance largely saturates. Consequently, an embedding dimension of 512 appears to provide a favorable trade-off between representation capacity and downstream performance.

Multimodal benefits To quantify the benefit of incorporating additional experimental modalities, we perform a controlled comparison using the full pretraining dataset. Specifically, we train (1) a model using all available modalities and (2) a model using only molecular graphs and ECFP4 fingerprints, which can be computed directly from molecular structure. We evaluate both models on the downstream benchmark datasets using linear and non-linear probe models based on the CLS representation. As shown in Figure 15, we find that incorporating the additional modalities reduces the MAE by 14% for the linear probe and 13% for the non-linear probe. These results demonstrate that modalities provide valuable complementary information that significantly improves the quality of the learned molecular representations.

G Representational similarity

Representation similarity analysis. Figure 16 shows the pairwise Centered Kernel Alignment (CKA) similarities between modality-specific embeddings. Clear clusters emerge among modalities that capture related information. Most notably, the quantum chemistry modalities, including Nabla, MOE, and xTB, exhibit the highest mutual similarities, reflecting their shared focus on physicochemical and electronic properties. The cellular profiling modalities CLOOME and BioXMol form a distinct cluster, which is expected given their overlap in underlying datasets and biological measurements. Interestingly, BioXMol also shows similarity with Boltz, suggesting that cellular phenotypic profiles may implicitly encode information related to protein binding interactions. In contrast, ChEMBL and TDC show comparatively low similarity to most other modalities, indicating that they contribute complementary information not captured by the remaining data sources. Overall, these results suggest that Mol-JEPA organizes modality-

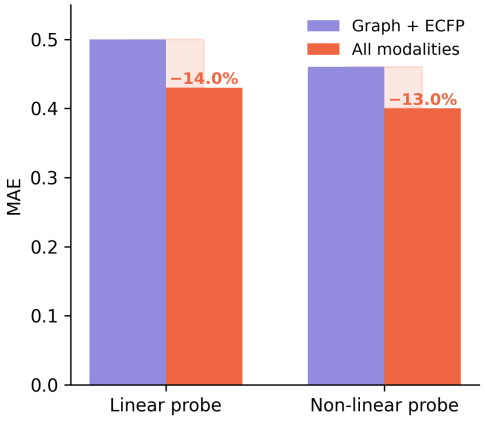


Figure 15: **Multimodal performance gains**. We re-train the model on the full dataset using only Graph and ECFP4 modalities. We find that the downstream performance is significantly better when using all modalities.

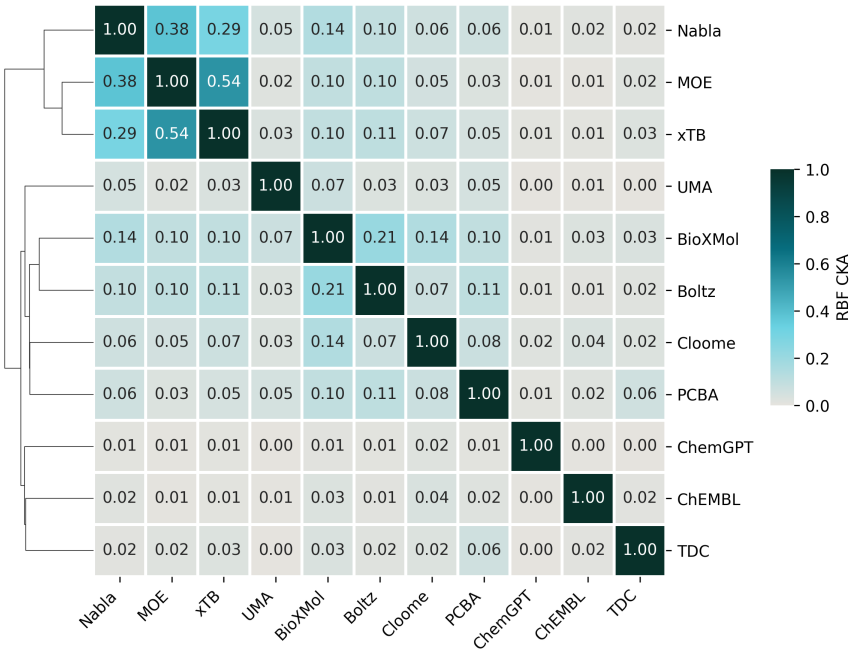


Figure 16: Centered Kernel Alignment

specific representations according to their underlying biological and physicochemical relationships while preserving information unique to individual modalities.

H IN-CONTEXT LEARNING COMPARISON

We assess the effectiveness of Mol-JEPA representations for in-context learning by comparing them with widely used molecular representations. As illustrated in Figures 17-19, Mol-JEPA CLS-token embeddings achieve superior performance across most tasks for different TabICLv2 hyperparameters. This finding indicates that the multimodal architecture effectively combines complementary information from different molecular modalities, resulting in richer representations that are particularly beneficial for in-context learning.

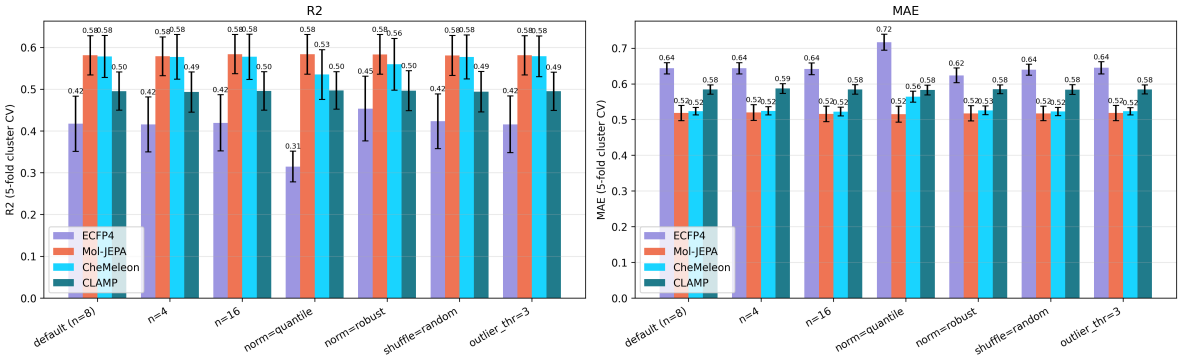


Figure 17: OpenADMET PXR TabICLv2 comparison.

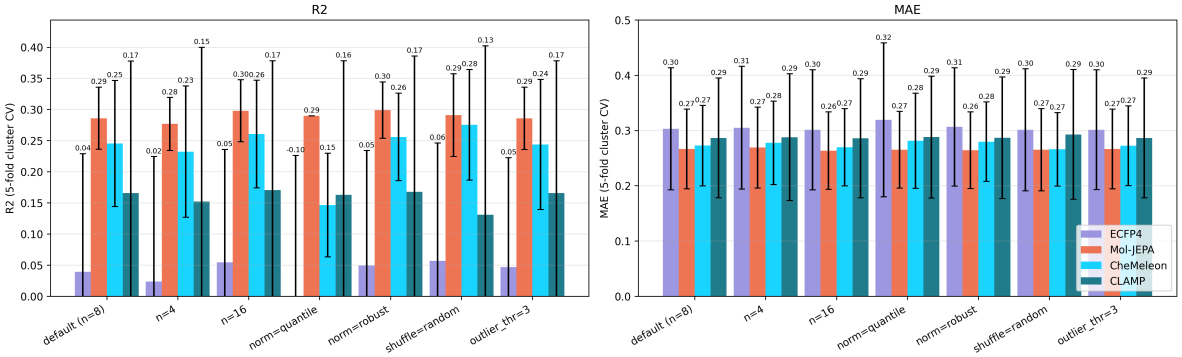


Figure 18: OpenADMET Caco2 Efflux TabICLv2 comparison.

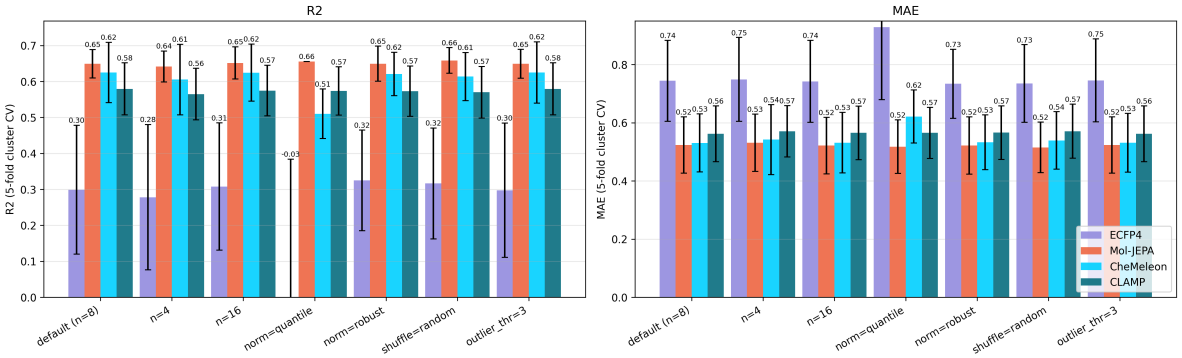


Figure 19: OpenADMET LogD TabICLv2 comparison.