Thin Bridges for Drug Text Alignment: Lightweight Contrastive Learning for Target Specific Drug Retrieval

Mallikarjuna Tupakula *
Rochester Institute of Technology
Rochester, NY 14623
mt3998@rit.edu
tmallikarjuna111@gmail.com

Abstract

Multimodal foundation models hold promise for drug discovery and biomedical applications, but most existing approaches rely on heavy pretraining or large scale multimodal corpora. We investigate whether *thin contrastive bridges*, lightweight projection heads over frozen unimodal encoders can align chemical and textual representations without training a full multimodal model. Using paired mechanisms from ChEMBL, we align *ECFP4 molecular fingerprints* with biomedical sentence embeddings through dual linear projections trained with a contrastive objective. To better handle drugs sharing the same therapeutic target, we incorporate hard negative weighting and a margin loss. Evaluation under scaffold based splits, which require generalization across disjoint chemical cores, demonstrates that our approach achieves non-trivial cross modal alignment and substantially improves within target discrimination compared to frozen baselines. These results suggest that thin bridges offer a compute efficient alternative to large scale multimodal pretraining, enabling scaffold aware drug text alignment and target specific retrieval in precision medicine.

1 Introduction

Multimodal foundation models have opened promising directions for biomedical research, particularly in drug discovery and precision medicine. These approaches seek to unify heterogeneous representations such as chemical structures, protein targets, and biomedical text under a shared embedding space. Recent advances in contrastive multimodal learning have demonstrated that cross-modal alignment can support drug target prediction, phenotype based screening, and chemical structure elucidation Xu et al. [2023], Wang et al. [2025], Rao et al. [2025], Rocabert-Oriols et al. [2025]. For instance, CLOOME showed that contrastive learning can unlock bioimaging databases for queries with chemical structures Sanchez-Fernandez et al. [2023], while MolCLR and related efforts emphasized molecular contrastive learning as an efficient path toward transferable molecular representations Wang et al. [2022], Pinheiro et al. [2022]. Beyond supervised pipelines, lightweight frameworks such as OneEncoder Faye et al. [2024] and cross-modal efficiency strategies Faye et al. [2024] suggest that thin projection heads over frozen encoders can serve as compute efficient bridges across modalities.

A growing line of work extends these retrieval based strategies toward drug design and discovery applications. Efforts such as multimodal protein ligand contrastive learning Wang et al. [2024],

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

retrieval augmented biomedical learning He et al. [2025], Gargari and Habibi [2025], and unified multimodal pipelines Luo et al. [2024], Dang et al. [2025] highlight the importance of efficiently linking chemical fingerprints with textual or biological context for downstream drug discovery tasks. Importantly, such methods emphasize that retrieval accuracy is not merely an evaluation metric but a foundation for enabling generation whether for novel molecule design or phenotype conditioned predictions Huang et al. [2024].

Parallel insights arise from **neuroscience** and **brain decoding**, where retrieval aligned models have been shown to enhance downstream generative reconstruction. For example, THINGS-data Hebart et al. [2023] provides a multimodal benchmark of fMRI, MEG, and behavioral similarity judgments, enabling representational alignment across modalities. Similarly, Défossez et al. [2023] demonstrated that contrastive alignment of MEG/EEG signals with pretrained speech embeddings supports zero-shot decoding of perceived speech segments with high accuracy. In vision, Benchetrit et al. [2023] and related work on fMRI-to-image decoding showed that improved retrieval alignment between brain signals and latent image embeddings leads to higher fidelity image reconstructions when paired with diffusion based generative models. These findings reinforce the idea that retrieval accuracy is tightly coupled with generative quality, motivating analogous strategies in biomedical domains.

Taken together, these advances suggest that compute efficient cross modal retrieval architectures can serve as scalable building blocks for more ambitious generative models. In this work, we investigate whether thin projection bridges lightweight contrastive heads over frozen unimodal encoders can align chemical fingerprints with biomedical mechanism text, while disambiguating drugs with shared targets through hard-negative sampling and margin based losses. By evaluating under scaffold splits on ChEMBLGaulton et al. [2012], we demonstrate non-trivial generalization and within target retrieval gains, supporting the broader view that such *thin bridges* can form the foundation for downstream generative pipelines in precision drug discovery.

Contributions. This paper makes the following contributions:

- We introduce **thin contrastive bridges** that align ECFP4 molecular fingerprints with biomedical mechanism text through lightweight dual projection heads, avoiding large-scale multimodal pretraining.
- We incorporate **hard-negative weighting and a margin based loss** to better disambiguate drugs that act on the same therapeutic target, addressing within target retrieval challenges.
- We evaluate on ChEMBL with a rigorous **scaffold-based split**, demonstrating non-trivial cross modal generalization and improved within target retrieval compared to frozen encoder baselines.
- We show that such thin bridges are **compute-efficient** (single GPU, short training time) and can serve as scalable foundations for downstream generative drug discovery pipelines.

2 Dataset

We constructed our dataset from **ChEMBL v28**, a curated database of bioactive molecules with drug like properties. To ensure clinical relevance, we extracted all *approved drugs* by filtering entries with `max_phase = 4`. From the `mechanism` table, we collected associations between drugs and their therapeutic targets, retaining only records with explicit molecular mechanisms.

To enrich these entries, we combined three information sources:

- **Molecule data**: Canonical SMILES strings retrieved from the ChEMBL molecule endpoint, corresponding to 2,970 unique molecules after deduplication.
- **Mechanism data**: Textual descriptions of drug mechanisms and annotated *action types* (e.g., inhibitor, agonist).
- **Target data**: Standardized ChEMBL target identifiers, mapped to preferred names and biological target types via the target endpoint.

To reduce redundancy, we removed duplicate drug target pairs and dropped incomplete rows lacking either SMILES or mechanism text. After filtering, the resulting dataset comprised **3,030 high quality drug–target pairs** with both chemical and textual representations. Each entry contains:

- Molecule ChEMBL ID
- Canonical SMILES
- Mechanism of action (free-text biomedical description)
- Target ChEMBL ID and target name
- Action type
- Maximum clinical phase

This dataset provides a paired multimodal resource aligning molecular fingerprints with biomedical text. The SMILES strings offer a cheminformatic view of drug structure, while the mechanism sentences encapsulate human curated biomedical knowledge, often specifying target level interactions. This dual representation makes the dataset well suited for **contrastive drug text alignment** tasks, where the objective is to learn lightweight bridges between unimodal embeddings.

By grounding the dataset in ChEMBL, a widely adopted benchmark in drug discovery, we ensure both interpretability and extensibility. The final corpus balances *domain richness* (multiple therapeutic areas, diverse targets) with *computational tractability* (single GPU scale), enabling efficient exploration of multimodal alignment methods for drug retrieval and, ultimately, generative drug design.

3 Methodology

3.1 Dataset Construction

We use drug target pairs from ChEMBL Gaulton et al. [2012], restricted to approved drugs (`max_phase = 4`). Each entry includes canonical SMILES, mechanism of action, target identifier, and action type. To reduce ambiguity in mechanism only descriptions and disambiguate drugs sharing targets, we constructed an enriched text field, `text_rich`, by concatenating mechanism, target name, action type, and drug preferred name (fetched from the ChEMBL molecule and target endpoints). Rows with missing SMILES or short descriptions were removed, yielding $\sim$3k high-quality pairs.

3.2 Molecular and Text Encoders

On the molecular side, we experimented with both ChemBERTa Chithrananda et al. [2020] embeddings and ECFP4 fingerprints (radius $= 2$, 2048 bits). On the text side, we used PubMedBERT Gu et al. [2021] and a similarity-tuned biomedical encoder (S-Biomed-RoBERTa-STSB) Deka et al. [2021]. All encoders remained frozen.

3.3 Thin Contrastive Bridge

We learn dual linear projection heads (one per modality) that map representations into a shared $d = 256$-dimensional space. Training minimizes a symmetric InfoNCE loss with temperature $T = 0.07$:

$$\mathcal{L} = \tfrac{1}{2}\Big(\mathrm{CE}\big(\tfrac{B_T B_M^\top}{T}, \mathrm{diag}\big) + \mathrm{CE}\big(\tfrac{B_M B_T^\top}{T}, \mathrm{diag}\big)\Big),$$

where $B_T = \mathrm{norm}(W_T Z_T)$ and $B_M = \mathrm{norm}(W_M X_M)$. We optimize with AdamW (lr $= 10^{-3}$, weight decay 10^{-4}) for 100 epochs with batch size 512. To address same-target confusion, we incorporate hard-negative weighting and a margin loss.

3.4 Evaluation Protocols

We report **Recall@1**, **Mean Reciprocal Rank (MRR)**, and **Grouped Recall@1** (within-target discrimination, excluding groups with fewer than 3 compounds). We also evaluate under a rigorous **scaffold split**, partitioning compounds by Bemis-Murcko scaffold into disjoint train/test sets to measure generalization across unseen chemical cores. Bootstrap resampling provides 95% confidence intervals.

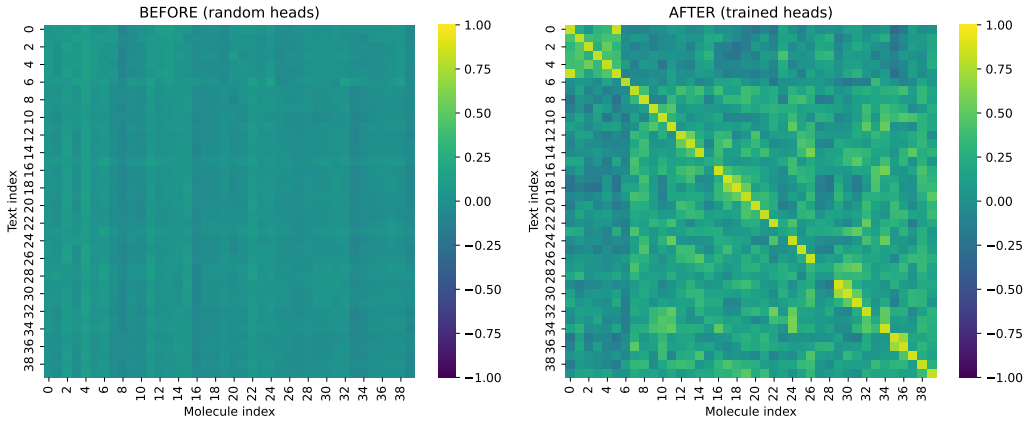

Figure 1: ECFP4 bridge with enriched text (`text_rich`). **Left:** before training. **Right:** after training, showing clear diagonal alignment.

4 Results

4.1 Baseline with Frozen Encoders

Direct alignment using ChemBERTa for SMILES and PubMedBERT for biomedical text proved ineffective. Retrieval accuracy was near random (`Recall@1 = 0.000-0.001, MRR = 0.003`), with similarity tuned S-Biomed-RoBERTa offering no meaningful improvement. This confirms that pretrained unimodal encoders alone cannot provide cross-modal alignment.

4.2 Contrastive Bridge Improves Alignment

Introducing a lightweight dual projection bridge significantly improved retrieval. On the ChEMBL drug mechanism dataset, the bridge achieved **Recall@1 = 0.188** and **MRR = 0.338**. Cosine similarity matrices exhibited a clear diagonal structure after training, indicating successful one-to-one molecule text alignment (Figure 6).

4.3 ECFP4 Bridge with Enriched Text

Combining ECFP4 molecular fingerprints with enriched `text_rich` descriptions substantially boosted alignment. The model achieved **Recall@1 = 0.762** and **MRR = 0.863**, with strong diagonal structure emerging in cosine similarity matrices (Figure 1). Including drug names in `text_rich` further improved disambiguation.

4.4 Generalization under Scaffold Split

When evaluated under scaffold splits, performance decreased but remained meaningful: **Recall@1 = 0.150**, **MRR = 0.228**, and **Grouped Recall@1 = 0.317**. This more than tripled grouped Recall@1 compared to frozen baselines, demonstrating that the bridge generalizes beyond memorized scaffolds (Figure 2).

4.5 Top-k Retrieval and Ablations

Cumulative Match (Recall@k) curves show steady improvements with k: global Recall@10 ≈ 0.39, while within-target Recall@10 > 0.80, indicating that the correct molecule is typically found in a small candidate set. Ablations confirmed robustness across temperature T and margin m, with best grouped Recall@1 at *WithDrug*, T=0.05, m=0.15 (Figure 3, Figure 4).

4.6 Summary of Results

Table 1 summarizes performance across settings.

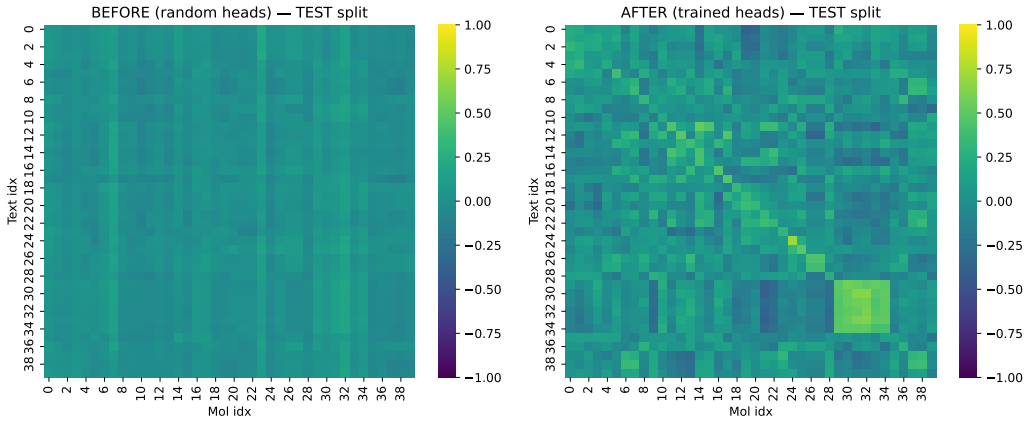

Figure 2: Cosine similarity matrices on the scaffold split test set (first $K=40$ pairs). **Left:** random heads. **Right:** trained bridge with strong diagonal alignment.

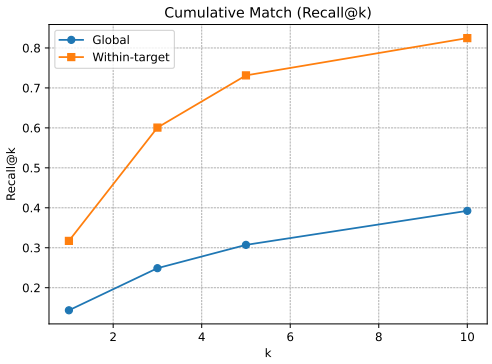

Figure 3: Cumulative Match (Recall@k) on the scaffold split test set. Global retrieval (blue) improves steadily with k, while within-target retrieval (orange) climbs steeply, showing most correct matches appear in small sets.

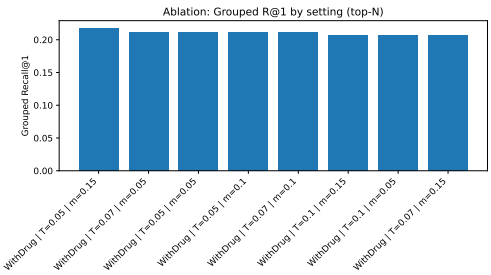

Figure 4: Ablation on grouped Recall@1 across temperature T, margin m, and drug name inclusion in `text_rich`. Best: *WithDrug*, $T=0.05$, $m=0.15$.

Table 1: Retrieval performance on ChEMBL drug–text alignment.

Setting	**Recall@1**	**MRR**	**Grouped R@1**
Frozen encoders (ChemBERTa + PubMedBERT)	0.001	0.003	–
Frozen encoders (ChemBERTa + S-Biomed)	0.000	0.003	–
Contrastive bridge (ChemBERTa + PubMedBERT)	0.188	0.338	0.098
ECFP4 + `text_rich` bridge	**0.762**	**0.863**	–
Scaffold split (ECFP4 + `text_rich`)	0.150	0.228	0.317

4.7 Future Directions

Our results suggest that cross modal alignment can accelerate drug discovery. By enabling efficient retrieval of molecule mechanism pairs, the framework supports rapid in silico screening and provides interpretable links between chemical structure and biological function. Moreover, scaffold split generalization indicates potential for discovering compounds with novel scaffolds. Coupling this approach with generative molecular models could further enable *de novo* design of candidate drugs.

References

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*, 2023.

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

Tien Dang, Viet Thanh Duy Nguyen, Minh Tuan Le, and Truong-Son Hy. Multimodal contrastive representation learning in augmented biomedical knowledge graphs. *arXiv preprint arXiv:2501.01644*, 2025.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10): 1097–1107, 2023.

Pritam Deka, Anna Jurek-Loughrey, and Deepak. Unsupervised keyword combination query generation from online health related content for evidence-based fact checking. In *The 23rd International Conference on Information Integration and Web Intelligence*, pages 267–277, 2021.

Bilal Faye, Hanane Azzag, and Mustapha Lebbah. Oneencoder: A lightweight framework for progressive alignment of modalities. *arXiv preprint arXiv:2409.11059*, 2024.

Omid Kohandel Gargari and Gholamreza Habibi. Enhancing medical ai with retrieval-augmented generation: A mini narrative review. *Digital health*, 11:20552076251337177, 2025.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Jiawei He, Boya Zhang, Hossein Rouhizadeh, Yingjian Chen, Rui Yang, Jin Lu, Xudong Chen, Nan Liu, Irene Li, and Douglas Teodoro. Retrieval-augmented generation in biomedicine: A survey of technologies, datasets, and clinical applications. *arXiv preprint arXiv:2505.01146*, 2025.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Lei Huang, Zheng Yuan, Huihui Yan, Rong Sheng, Linjing Liu, Fuzhou Wang, Weidun Xie, Nanjun Chen, Fei Huang, Songfang Huang, et al. A unified conditional diffusion framework for dual protein targets-based bioactive molecule generation. *IEEE Transactions on Artificial Intelligence*, 5(9):4595–4606, 2024.

Yizhen Luo, Xing Yi Liu, Kai Yang, Kui Huang, Massimo Hong, Jiahuan Zhang, Yushuai Wu, and Zaiqing Nie. Toward unified ai drug discovery with multimodal knowledge. *Health Data Science*, 4:0113, 2024.

Gabriel A Pinheiro, Juarez LF Da Silva, and Marcos G Quiles. Smiclr: contrastive learning on multiple molecular representations for semisupervised and unsupervised representation learning. *Journal of Chemical Information and Modeling*, 62(17):3948–3960, 2022.

Jiahua Rao, Hanjing Lin, Leyu Chen, Jiancong Xie, Shuangjia Zheng, and Yuedong Yang. Multimodal contrastive learning with negative sampling calibration for phenotypic drug discovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30752–30762, 2025.

Pau Rocabert-Oriols, Núria López, and Javier Heras-Domingo. Multi-modal contrastive learning for chemical structure elucidation with vibraclip. 2025.

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.

Yifei Wang, Yunrui Li, Lin Liu, Pengyu Hong, and Hao Xu. Advancing drug discovery with enhanced chemical understanding via asymmetric contrastive multimodal learning. *Journal of chemical information and modeling*, 2025.

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.

Zhen Wang, Zhanfeng Wang, Maohua Yang, Long Pang, Fangyuan Nie, Siyuan Liu, Zhifeng Gao, Guojiang Zhao, Xiaohong Ji, Dandan Huang, et al. Enhancing challenging target screening via multimodal protein-ligand contrastive learning. *bioRxiv*, pages 2024–08, 2024.

Hao Xu, Yifei Wang, Yunrui Li, and Pengyu Hong. Asymmetric contrastive multimodal learning for advancing chemical understanding. *arXiv preprint arXiv:2311.06456*, 2023.

A Appendix

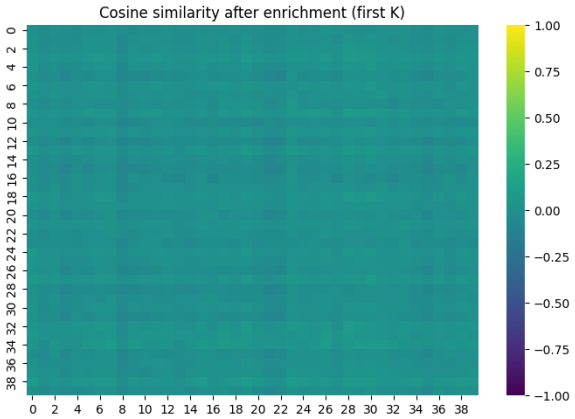

Figure 5: Cosine similarity heatmap between the first $K = 40$ drug–text pairs after enrichment. The uniform pattern indicates that frozen encoders alone do not yield meaningful alignment.

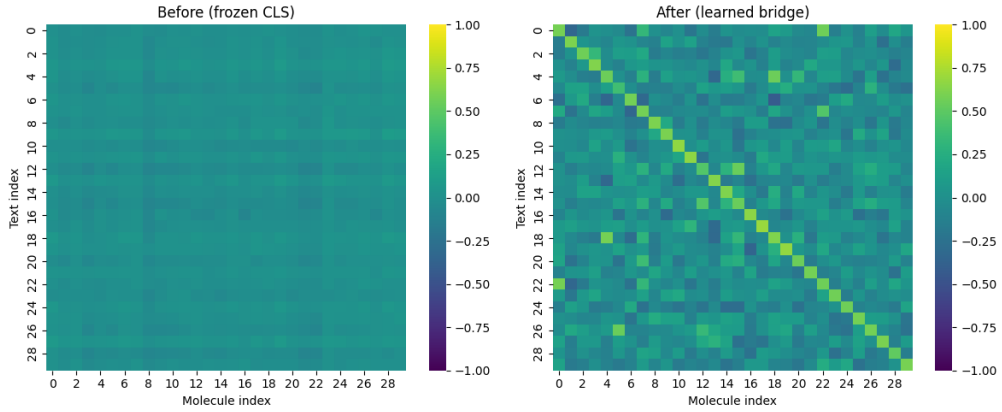

Figure 6: Cosine similarity comparison between the first $K = 30$ drug–text pairs before and after training. (**Left**) Frozen CLS embeddings show uniform similarity with no meaningful alignment. (**Right**) After training the learned bridge, a strong diagonal emerges, indicating accurate molecule–text alignment.