Bridging chemical modalities by aligning embeddings

Adrian Mirza [1 2] **Erinc Merdivan** [3] **Sebastian Starke** [4] **Kevin Maik Jablonka** [1 5 2]

Abstract

Chemistry as a science is highly diverse in its ways of representing molecules, and many of these representations are highly abundant in the literature and, as such, underutilized. There is also a lack of frameworks that combine these different representations into a common one. Thus, we introduce the multimodal machine learning model `MoleculeBind`. It was trained with the goal of aligning five different modalities: SMILES, SELFIES, graphs, fingerprints, and 3D structures using contrastive learning. We investigate the retrieval metrics for the model and obtain high performance across all the different modalities. We also explore the potential of querying molecules with similar properties using the same approach. The retrieval of molecules with similar properties outperformed a random baseline significantly. We expect such a model to have a great impact on spectroscopy and improve the performance of existing generative methods.

1. Introduction

Chemical data is very diverse and highly complex. This is partly because different techniques are needed in different subfields to analyze specific phenomena. Even for representing one single small molecule, one can choose from plenty of options such as SELFIES (Krenn et al., 2020; 2022), SMILES (Weininger, 1988), IUPAC names, or InChI identifiers (Heller et al., 2015). Besides those line representations, molecules are often also reported as 3D structures or used via molecular fingerprints (e.g., Morgan fingerprint (Morgan, 1965)) or molecular graphs in computational modeling. Experimental scientists more commonly deal with molecules characterized via spectroscopic measurements such as nuclear magnetic resonance (NMR) and infrared (IR) spectroscopy. Thus, chemical data is highly fragmented across multiple modalities. This leads to massive underutilization of chemical data as, using current techniques, this multimodal data cannot be synergistically and jointly utilized. For instance, most machine learning models in the chemical sciences can only accept one input modality. In addition, most, if not all, research data management systems do not provide effective search functionalities across modalities, which would, however, lead to vast efficiency gains in chemical research. In this work, we build on contrastive embedding alignment techniques to bring chemical data into a shared data space. We show that this alignment allows us to perform effective cross-modal retrieval—even for pairs of modalities on which the model has not explicitly been trained.

Concretely, our main contributions are

- **MoleculeBind**: A model that aligns molecular graphs, 3D structures, molecular fingerprints, as well as SMILES and SELFIES, reusing existing encoder models.
- **Demonstration of emergent cross-modal retrieval in chemical data**: We demonstrate that our embeddings are performant even for recall across modality pairs for which the model has not been explicitly trained.

2. Related work

Binding embeddings through encoder alignment Outside the chemical domain, multimodal embedding alignment models have been introduced (Ma et al., 2022; Mu et al., 2022). Among these models, `ImageBind` (Girdhar et al., 2023) stands out for incorporating more than two modalities. The model was built from six existing encoders, and the embeddings were aligned using a symmetric InfoNCE loss (van den Oord et al., 2019) with respect to a central "binding" modality (e.g., images). `ImageBind` also introduced a term called *emergent alignment*, which represents

[1]Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldstr. 10, 07743 Jena, Germany [2]Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstraße 12-14, 07743, Jena, Germany [3]Helmholtz Munich, Munich, Germany [4]Computational Science group, Department of Information Services and Computing, Helmholtz-Zentrum Dresden-Rossendorf, Bautzner Landstraße 400, 01328 Dresden, Germany [5]Center for Energy and Environmental Chemistry Jena, Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany. Correspondence to: Adrian Mirza <andrian.mirza@uni-jena.de>, Kevin Maik Jablonka <mail@kjablonka.com>.

the alignment of pairs of modalities on which the model has not been explicitly trained. This represented an important departure from conventional alignment approaches, in which pairs for each relevant modality are required. In `ImageBind`, only data paired with the "binding" modality is required. The resulting embeddings have been used in various downstream applications, for example, interfaced with large language models (Han et al., 2023; Su et al., 2023).

Multimodal models in chemistry Multimodal models for chemistry have been proposed in the literature. For example, Han et al. (2024) introduced a framework for polymer property prediction (graph and PSMILES); Kaufman et al. (2024) built a bimodal encoder-decoder model based on aligned text and 3D representations; Ock et al. (2024) introduced a bimodal model for catalysis, where the use of graph and language models into a single framework aided data efficiency for energy predictions tasks. Seidl et al. (2023) introduced `CLAMP`, a bimodal model based on text descriptions of assays and molecules (i.e., string notations, graphs, or fingerprints).

Cross-modal retrieval Cross-modal retrieval refers to retrieving semantically related data stored in one modality (e.g., images) by querying with another modality (e.g., text). Wang et al. (2016) provide an overview of cross-modal retrieval. Explicitly aligning chemical modalities was tried by Sanchez-Fernandez et al. (2023). Their `CLOOME` retrieval model includes microscopy images and text modalities. Liu et al. (2023) use cross-modal retrieval for text descriptions and small molecules for drug discovery. Xiao et al. (2024) introduced `MolBind` to align four chemistry-related modalities with the primary focus on proteins.

3. MoleculeBind Model

Architecure We built `MoleculeBind` using `PyTorch` and `PyTorchLightning` based on the architecture proposed by Girdhar et al. (2023). In contrast to their work, we utilized a simple InfoNCE loss without symmetrization for computational efficiency:

$$L_{R,M} = -\log \frac{\exp\left(\mathbf{q}_i^{\mathbf{T}}\mathbf{k}_i/\tau\right)}{\exp\left(\mathbf{q}_i^{\mathbf{T}}\mathbf{k}_i/\tau\right) + \sum_{i\neq j}\exp\left(\mathbf{q}_i^{\mathbf{T}}\mathbf{k}_j/\tau\right)}, \tag{1}$$

where *R* is the central representation (modality) to which other modalities are aligned, and *M* is any other modality $M \neq R$.

We align five modalities: SELFIES (central modality), SMILES, 3D structures, molecular graphs, and Morgan fingerprints enhanced by 93 cheminformatics descriptors (e.g., QED, LogP) computed using `RdKit`. The model architecture is shown in Figure 1, highlighting the direct alignments (i.e., trained on) and a part of the emergent alignments (i.e., not trained on).

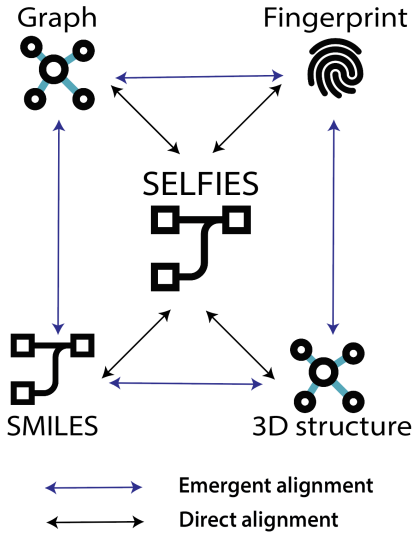

Figure 1. MoleculeBind architecture. Purple arrows indicate emergent alignment, and black arrows indicate direct alignment (i.e., pairs of modalities on which the model is explicitly trained). For illustration purposes, not all emergent links between modalities are shown. The hidden state sizes for the models are as follows: SELFIES, SELFIES, fingerprint - 768, graphs - 256, 3D structures - 128. All the embeddings are projected into a 1024-dimensional hidden state.

Encoders The model is built using five encoders available in the literature. For encoding the central modality SELFIES, we use `SELFormer` (Yüksel et al., 2023). As the structure encoder, we use `Dimenet` (Gasteiger et al., 2020). We use ChemBERTa from (Chithrananda et al., 2020) to encode SMILES strings, and `MolCLR` (Wang et al., 2022) for graph encoding. To encode molecular fingerprints, we trained a variational autoencoder (VAE) (Higgins et al., 2017; Kingma & Welling, 2022). Besides the fingerprint-VAE, all the model configurations are based on the original implementation. All encoders except the SMILES and SELFIES transformers were trained from scratch.

3.1. Data

The data that has been used in this work comes from multiple sources.

3D Structures For the training of the structure encoder, we sampled 100k molecules from the QMugs dataset (Isert et al., 2022). The target property of choice was the HOMO-LUMO gap computed with DFT single-point calculations on structures optimized with the semi-empirical GFN2-xTB method (Grimme et al., 2017; Bannwarth et al., 2019; 2021).

Graph For the graph model, we used 100k molecules that were presented as an example by Wang et al. (2022) for training `MolCLR`. The full model has been trained on ≈10M molecules.

SELFIES and SMILES 200k SELFIES and SMILES strings were sampled from the CHEMBL (v30) database (Gaulton et al., 2017), 100k from the QMugs dataset (Isert et al., 2022) and 100k from the ChemNLP IUPAC-SMILES dataset (Jablonka, 2024). The total number of unique molecules ended up at ca. 393k.

MoleculeBind dataset Starting from SELFIES or SMILES, the dataset for `MoleculeBind` is prepared as follows. We compute graphs based on SMILES at runtime. For structures, we load 3D geometries from SDF files (one per SMILES or SELFIES) as provided by Isert et al. (2022). We precompute the Morgan fingerprints and descriptors using the `RdKit` package (Landrum et al., 2013) and concatenate these to form a fingerprint of length 2141 (2048 bits, 93 descriptors).

4. Results

Retrieval metrics The *top-1* and *top-5* retrieval recall have been assessed on a complete validation set (i.e., all modality pairs are complete). These metrics are shown in Table 1. Since the loss we used is not symmetric, different results were obtained for the same pair of modalities but with swapped query modalities. For example, when building the vector database from SELFIES embeddings and querying it with graph embeddings, the performance is lower than vice versa.

We observe emergent alignment between the pairs of modalities that the model was never trained on. For example, a retrieval rate of more than 80% is obtained for the pairs 3D structure - graph, graph - fingerprint, and fingerprint-3D structures. This highlights the efficiency of this architecture and how we can leverage data for which plenty of pairs are available to perform effective retrieval across modality pairs for which only little data is available.

Property space for queries Following Frey et al. (2023), we showcase a few examples of how `MoleculeBind` can assist in retrieving molecules with similar properties (see Figure 2). The average distance between the properties of the molecule used for queries and the queried molecules as a distance in the 2D property space is calculated using

$$d_{\text{prop}} = \sqrt{\frac{(P_{1@20} - P_{1\text{query}})^2}{\mu_{P_1}} + \frac{(P_{2@20} - P_{2\text{query}})^2}{\mu_{P_2}}},$$

where P_1 and P_2 are property 1 and property 2, respectively. μ_{P1} and μ_{P2} are the means across the population for the properties. $P_{1@20}$ and $P_{2@20}$ are the means of respective properties P_1 and P_2 for the *top-20* retrieved molecules using a query molecule. On average, queries are within 0.30 distance (normalized) in the 2D property space of QED and LogP compared to 0.53 in a random baseline. A Kolmogorov-Smirnov test shows that the difference between the two distributions is significant ($p - value = 0$).

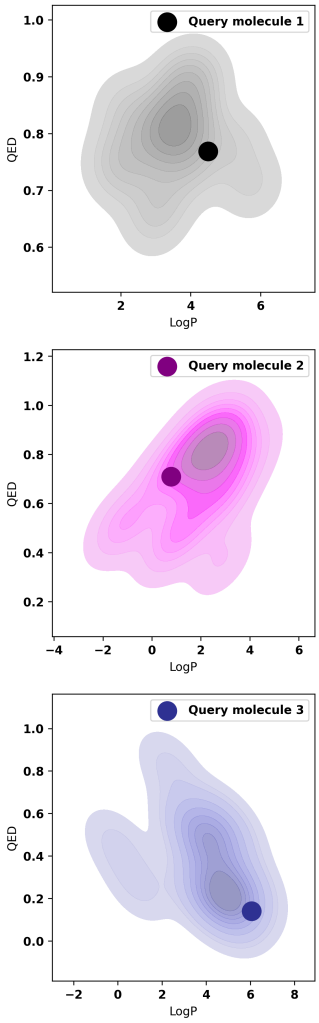

Figure 2. Property queries examples. The *top-20* retrieved molecules are displayed as a kernel density estimation (KDE). The figure shows the retrieval between graph and SELFIES embeddings, for which the dataset sizes can be found in Table 1.

5. Conclusions

`MoleculeBind` is the first model to align five different chemical modalities. We obtained high performance on retrieval recall metrics on the validation set. Weobserved an emergent alignment of the embeddings between pairs

Table 1. Recall rates for different molecular representations and their combinations (≈ 10k molecules). The second modality in the pair represents the query, while the embeddings of the first are used to create the vector database. Text colors correspond to either direct or emergent alignment, as shown in Figure 1.

Modality Pair	Recall@1	Recall@5	Random Recall @5
SELFIES - SMILES	0.941	0.997	0.000
SELFIES - Graph	0.807	0.956	0.001
SELFIES - Fingerprint	0.819	0.964	0.000
SELFIES - 3D structure	0.947	0.998	0.000
SMILES - SELFIES	0.947	0.998	0.000
SMILES - Graph	0.900	0.983	0.000
SMILES - Fingerprint	0.872	0.976	0.000
SMILES - 3D structure	0.943	0.997	0.000
Graph - SELFIES	0.846	0.977	0.000
Graph - SMILES	0.926	0.992	0.000
Graph - Fingerprint	0.845	0.969	0.000
Graph - 3D structure	0.885	0.987	0.000
Fingerprint - SELFIES	0.828	0.969	0.000
Fingerprint - SMILES	0.882	0.986	0.000
Fingerprint - Graph	0.817	0.960	0.000
Fingerprint - 3D structure	0.862	0.981	0.000
3D structure - SELFIES	0.955	0.999	0.000
3D structure - SMILES	0.944	0.998	0.000
3D structure - Graph	0.842	0.967	0.001
3D structure - Fingerprint	0.866	0.978	0.000

of modalities the model was never trained on. This has implications for how we leverage data. For example, in our study, we worked with a limited amount of structural data compared to the other modalities. However, the retrieval metrics for this modality were similar.

Moreover, we showcased how the embeddings `MoleculeBind` generated can be used to query molecules with similar properties and that we observed a significant performance increase from a random baseline.

This illustrates that an architecture such as `MoleculeBind` can be effective in bridging the multitude and diversity of embeddings we find in chemical data. Doing so will allow us to leverage the data collectively and, ultimately, might also enable researchers to uncover unknown links that have been recorded in different modalities.

6. Future work

Since it is likely that the performance of our approach can be improved by scaling models and datasets, we plan to evaluate the scaling laws for the model beyond the 393k dataset. Also, we aim to augment `MoleculeBind` with new modalities (e.g., nuclear magnetic resonance (NMR), infrared (IR), and Ultraviolet-visible (UV-VIS) spectroscopy). Importantly, our approach can be reused on modalities for other classes of compounds besides small molecules (e.g., in materials science or biochemistry). Having a general representation of molecules would be useful for the many subfields and applications that chemistry has found its way into.

7. Acknowledgment and Disclosure of Funding

The research of A.M. and K.M.J. was supported by the Carl-Zeiss Foundation as well as the SOL-AI project of the Helmholtz-Foundation Model Initiative.

This work was supported by the Helmholtz Association's Initiative and Networking Fund on the HAICORE@FZJ partition.

References

Bannwarth, C., Ehlert, S., and Grimme, S. 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.

Bannwarth, C., Caldeweyher, E., Ehlert, S., Hansen, A., Pracht, P., Seibert, J., Spicher, S., and Grimme, S. Extended tight-binding quantum chemistry methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 11(2):e1493, 2021.

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Frey, N. C., Soklaski, R., Axelrod, S., Samsi, S., Gomez-Bombarelli, R., Coley, C. W., and Gadepally, V. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11):1297–1305, 2023.

Gasteiger, J., Groß, J., and Günnemann, S. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Grimme, S., Bannwarth, C., and Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (z= 1–86). *Journal of chemical theory and computation*, 13(5):1989–2009, 2017.

Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., Lu, X., Ren, S., Wen, Y., Chen, X., Yue, X., Li, H., and Qiao, Y. Imagebind-llm: Multi-modality instruction tuning, 2023.

Han, S., Kang, Y., Park, H., Yi, J., Park, G., and Kim, J. Multimodal transformer for property prediction in polymers. *ACS Applied Materials & Interfaces*, 16(13):16853–16860, 2024. doi: 10.1021/acsami.4c01207. PMID: 38501934.

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., and Tchekhovskoi, D. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7:1–34, 2015.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

Isert, C., Atz, K., Jiménez-Luna, J., and Schneider, G. Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(1):273, 2022.

Jablonka, K. chemnlp_iupac_smiles (revision 3753ca6), 2024. URL https://huggingface.co/datasets/kjappelbaum/chemnlp_iupac_smiles.

Kaufman, B., Williams, E. C., Underkoffler, C., Pederson, R., Mardirossian, N., Watson, I., and Parkhill, J. Coati: Multimodal contrastive pretraining for representing and traversing chemical space. *Journal of Chemical Information and Modeling*, 64(4):1145–1157, 2024. doi: 10.1021/acs.jcim.3c01753. PMID: 38316665.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2022.

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

Krenn, M., Ai, Q., Barthel, S., Carson, N., Frei, A., Frey, N. C., Friederich, P., Gaudin, T., Gayle, A. A., Jablonka, K. M., et al. Selfies and the future of molecular string representations. *Patterns*, 3(10), 2022.

Landrum, G. et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281, 2013.

Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., and Anandkumar, A. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., and Ji, R. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 638–647, 2022.

Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, May 1965. ISSN 1541-5732. doi: 10.1021/c160017a018. URL http://dx.doi.org/10.1021/c160017a018.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pp. 529–544. Springer, 2022.

Ock, J., Magar, R., Antony, A., and Farimani, A. B. Multimodal language and graph learning of adsorption configuration in catalysis. *arXiv preprint arXiv:2401.07408*, 2024.

Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S., and Klambauer, G. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1), November 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-42328-w. URL http://dx.doi.org/10.1038/s41467-023-42328-w.

Seidl, P., Vall, A., Hochreiter, S., and Klambauer, G. Enhancing activity prediction models in drug discovery with the ability to understand human language, 2023.

Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., and Cai, D. Pandagpt: One model to instruction-follow them all, 2023.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding, 2019.

Wang, K., Yin, Q., Wang, W., Wu, S., and Wang, L. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

Wang, Y., Wang, J., Cao, Z., and Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Weininger, D. 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.

Xiao, T., Cui, C., Zhu, H., and Honavar, V. G. Molbind: Multimodal alignment of language, molecules, and proteins. *arXiv preprint arXiv:2403.08167*, 2024.

Yüksel, A., Ulusoy, E., Ünlü, A., and Doğan, T. Selformer: molecular representation learning via selfies language models. *Machine Learning: Science and Technology*, 4 (2):025035, 2023.

Appendix

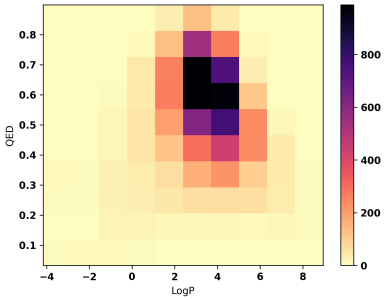

Figure 3. Property distribution for the validation set of 9937 molecules shown in Table 1.

Retrieval methods

Before computing the retrieval metrics on the entire dataset, we compute embeddings for each batch and then aggregate them into one tensor per modality. The indexing for pairwise retrieval calculations is acquired from the stored binarized dataset. `chromaDB` is used as the vector database of choice due to ease of use and sufficient performance.