PolyBind: Effectively Combining Datasets Indexed in Different Representations of Polymers

Sreekanth Kunchapu ✉
FSU Jena

Adrian Mirza
HIPOLE Jena, HZB

Kevin Maik Jablonka ✉
HIPOLE Jena, FSU Jena, CEEC Jena, JCSM Jena

✉ sreekanth.kunchapu@uni-jena.de and mail@kjablonka.com

Full affiliations are listed in the appendix.

Abstract

In polymer informatics, diverse datasets for the same material properties are available but often use different representations, posing challenges in meaningfully combining or utilizing them for machine learning (ML) models. This heterogeneity limits the predictive power of ML for material discovery. Here, we introduce **PolyBind**, a framework that leverages contrastive learning to align different polymer representations—including PSMILES, polymer names, and BigSMILES—within a shared latent space. PolyBind treats PSMILES as the anchor representation and maps polymer names and BigSMILES into the same embedding space, yielding a unified representation with richer chemical information than traditional fingerprint vectors. We demonstrate PolyBind's effectiveness on glass transition temperature prediction by successfully combining datasets with different polymer notations. Our framework offers a robust solution for integrating diverse polymer data sources.

1 Introduction

The rapid advancement of machine learning (ML) has enabled the efficient inference of material properties and accelerating the discovery of new polymers for diverse applications [1].

Polymers are materials with unique properties arising from their complex architecture of long chains with variable connectivity, repeating units, and end groups. Their behavior is best understood as statistical ensembles rather than individual molecules. There have been many attempts to digitally describe this complex nature. Historically, for instance in lab notebooks and handbooks such as the Polymer Handbook, IUPAC names have been utilized. However, for polymer informatics applications, alternative representations have emerged. Modern approaches include PSMILES, which represents the repeating structure of polymer macromolecules as a string, capturing their chemical connectivity and sequence [2]. BigSMILES extends traditional SMILES notation to represent polymer ensembles with variable chain lengths, branching patterns, and molecular weight distributions, specifically designed to describe complex polymer architectures [3, 4].

Currently, there is no tool that can convert between all these representations. This, effectively, leaves much data unused as polymer data is reported in different representations across the literature.

To address these challenges, we introduce PolyBind, a framework for generating unified embeddings from diverse polymer representations, such as PSMILES, BigSMILES, and polymer names (see Figure 1). We demonstrate the performance of this approach by applying PolyBind to combine datasets for glass transition temperature prediction across different modalities. The unified embedding space enables consistent representation and improved predictive performance compared to conventional Morgan fingerprints.

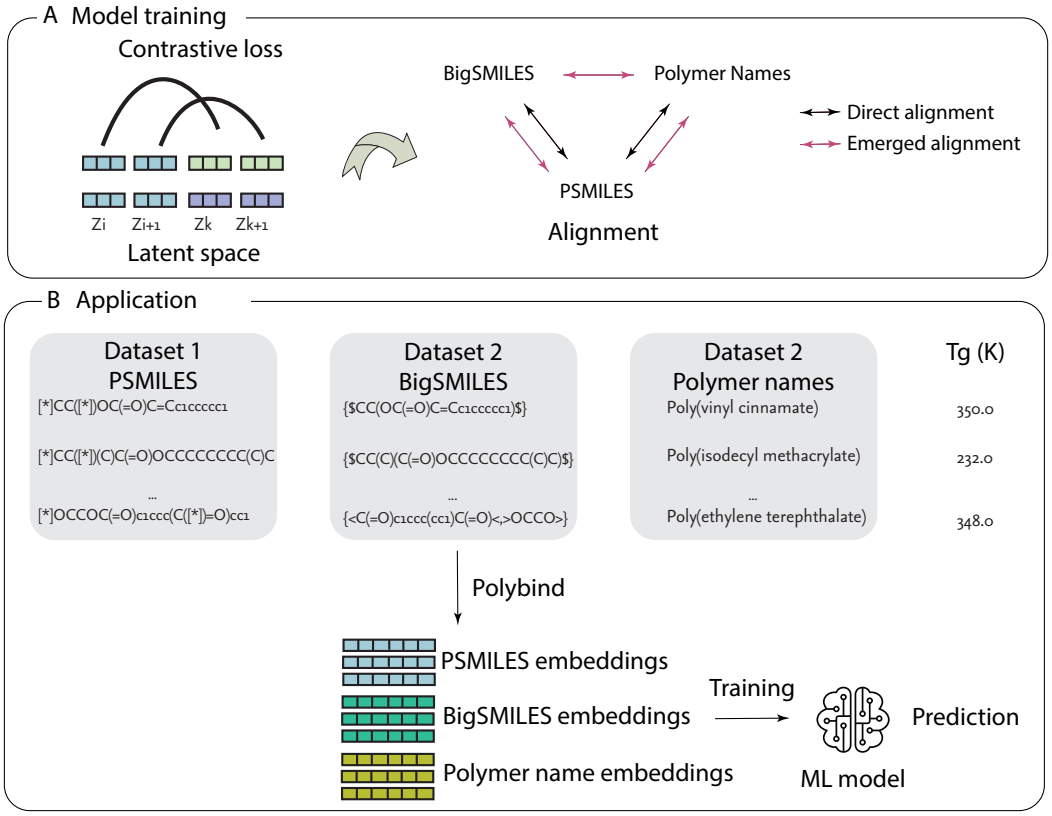

Figure 1: **Overview of PolyBind:** The framework integrates polymer informatics through a two-step process. (1) Model training employs self-supervised contrastive learning to align PSMILES, BigSMILES, and polymer name representations in a shared latent space, using PSMILES as the anchor. (2) Application combines datasets (PSMILES, BigSMILES, and polymer names) into a unified embedding space via PolyBind, enabling ML model training and prediction of property (T_g), with corresponding T_g values.

Concretely, our contributions are:

1. The first unified self-supervised contrastive learning approach that uses PSMILES as the central modality and aligns polymer names and BigSMILES within the same latent space, enabling integration of previously incompatible polymer datasets.
2. Dense embeddings that capture richer, more predictive information than traditional Morgan fingerprint vectors for downstream polymer property prediction tasks.

2 Related Work

2.1 Polymer representations and identifiers

The challenge of representing polymers computationally spans multiple categories of approaches, each with distinct strengths and limitations. Unlike small molecules that can be uniquely described

by single structures, polymers exist as ensembles of similar molecules with varying chain lengths, architectures, and compositions, necessitating specialized representation strategies.

Text-based identifiers form the foundation of polymer nomenclature and data exchange. IUPAC names provide systematic chemical identification but become unwieldy for complex architectures. Modern string-based identifiers include PSMILES [2], which extends SMILES notation to represent polymer repeat units with terminal asterisks, and BigSMILES [4], which uses specialized syntax to describe polymer ensembles with stochastic objects, like random sequences in copolymers or probabilistic branching in dendrimers.

Recent developments have introduced PSELFIES [5], extending SELFIES notation [6, 7] to polymers, and specialized representations like Group SELFIES [8] for systematic polymer description.

Graph-based representations have emerged as powerful approaches for capturing polymer connectivity and chemical relationships. Aldeghi and Coley [9] developed molecular ensemble representations specifically tailored for polymers. This foundation has been extended through multiple polymer-specific graph approaches: PolymerGNN [10] addresses multitask property learning using graph representations of monomer compositions; polyGNN [11] introduces invariant transformations for repeat unit graphs, ensuring that the graph neural network processes the molecular structure of polymer repeat units in a way that remains consistent under symmetries like rotations or node permutations and recent work has developed polymer-unit graphs [12] that provide coarse-grained representations. Self-supervised approaches for graph neural networks have also shown promise for polymer property prediction [13].

Fingerprint-based representations convert chemical structures into fixed-length numerical vectors. Morgan fingerprints [14] remain widely used but often produce sparse, insufficiently informative representations for polymers. Polymer-specific fingerprint approaches [15, 16] have demonstrated improved performance over traditional molecular fingerprints.

Currently, no unified framework exists for converting between these diverse representation modalities. This fragmentation effectively leaves much polymer data unused, as datasets employing different identifiers (PSMILES, BigSMILES, and polymer names) cannot be meaningfully combined without loss of essential information.

2.2 Self-supervised learning for aligning representations

Self-supervised learning (SSL) has demonstrated remarkable success in aligning diverse data modalities within unified embedding spaces. ImageBind [17] creates a single embedding space that binds images, text, audio, depth, thermal, and inertial measurement unit data through contrastive learning. This approach enables zero-shot cross-modal retrieval and emergent alignment between modalities that were never directly paired during training. Recent work has extended multimodal alignment to chemical domains: MoleculeBind [18] demonstrates contrastive learning across five molecular modalities (SMILES, SELFIES, graphs, fingerprints, and 3D structures), showing that unified embeddings can enable cross-modal retrieval and property-based molecular search.

2.3 Self-Supervised learning in polymer informatics

polyBERT [2] pioneered the application of transformer architectures to polymers by training on PSMILES strings using masked language modeling, where the model learns to predict missing tokens in polymer sequences.

Self-supervised graph neural networks [13] have demonstrated that SSL pre-training on unlabeled polymer graphs can reduce prediction errors by 28% in data-scarce scenarios compared to supervised models trained solely on limited labeled data. Uni-Poly [19] showed the potential of multimodal learning by combining structural and textual modalities for polymers, though it does not address the fundamental challenge of unifying different identifier systems.

The key limitation across existing polymer SSL approaches is their focus on single representation modalities. While these methods have shown promise for property prediction tasks, none address the critical bottleneck of integrating datasets that employ different polymer identifiers.

3 Methods

3.1 Data acquisition

Data were collected from multiple sources, as summarized in Table 1. The complete dataset comprises 4.91M data points, consisting of 970k homopolymers and 3.94M copolymers. Among the sources listed in the table, only two originally provided both polymer names and PSMILES representations. The remaining sources contained PSMILES along with corresponding monomer structures (either single monomers or two monomers for copolymers).

To generate polymer names for the entries lacking them, we first translated the SMILES representations of monomers to IUPAC names using the inhouse model smiles-to-iupac-translator. For homopolymers, we then wrapped these names in the poly(monomer) format, while for copolymers, we applied the standard IUPAC polymer nomenclature convention poly(monomer-co-monomer) [20] using custom Python scripts. Subsequently, we converted all PSMILES representations to BigSMILES format using the corresponding Python package [4].

This process resulted in three distinct modalities for polymer representation. Following the ImageBind framework [17], we treat PSMILES as the central modality and align both polymer names and BigSMILES representations to this central reference point.

Table 1: **Data Sources.** This table summarizes the sources of the dataset, listing the number of data points available from each source along with the presence of polymer names, chemical representations (PSMILES), and corresponding references.

Source	Number of Points	Polymer Name	PSMILES	Reference
PolyUniverse	3,799,195	✗	✓	[21]
PI1M	944,346	✗	✓	[22]
SMiPoly	151,920	✗	✓	[23]
PolyMetrix	7,367	✗	✓	[15]
Catauggie	6,273	✓	✗	[24]
PN2S	4,885	✓	✓	[25]

3.2 Models

3.2.1 Joint model architecture.

Our framework enables the integration of established models from existing literature. For PSMILES string encoding, we utilize the MolFormer model [26], which implements a transformer-based architecture [27] enhanced with rotary positional encoding [28]. For the encoding of polymer names, we opt for RoBERTa-base from FacebookAI [29], a language model pretrained on a large corpus of English text that is well-suited for processing natural language of polymer names. To address the dimensional mismatch between different encoder outputs, we incorporate a linear projection layer for each modality encoder, following the approach demonstrated in [17].

3.2.2 Training

PolyBind framework's primary objective is to achieve semantic alignment across multiple polymer representation modalities. Our training procedure employs the InfoNCE contrastive loss function [30], which is formulated as follows:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \tag{1}$$

- $\mathbf{z}_i, \mathbf{z}_j$: Embeddings of a positive pair (e.g., the same polymer represented as PSMILES and polymer name).
- $\mathrm{sim}(\cdot, \cdot)$: Cosine similarity, which quantifies how close the embeddings are in the shared latent space.

- τ: Temperature parameter that controls the concentration of the distribution—lower values make the model focus more on the closest (most similar) pairs.
- $2N$: The total number of representations (for a batch of N samples, each positive pair gives two views/modalities).
- $\mathbb{1}_{[k \neq i]}$: Indicator function that is 1 if $k \neq i$ and 0 otherwise; ensures that only representations other than i are considered as negatives.

3.2.3 Hyperparameters

We use a learning rate of 10^{-4}. For contrastive learning and set the temperature parameter (τ) to 0.07, following the approach established by [31]. This temperature parameter regulates the penalty strength applied to negative sample pairs (i.e., mismatched PSMILES and polymer name embeddings) [32]. Training convergence is determined using an early-stopping criterion [33] with a patience threshold of 10 epochs.

4 Results and Discussion

PolyBind uses contrastive learning to create a unified embedding space where different representations of the same polymer are positioned close together. The framework employs InfoNCE loss (see Equation (1))[30] with PSMILES as the anchor modality: for each PSMILES string, the model learns to pull the embeddings of its corresponding BigSMILES and polymer name closer together in the embedding space while pushing away the embeddings of different polymers. For example, the PSMILES "[*]CC[*]", BigSMILES "{CC}", and name "polyethylene" form positive pairs whose embeddings should be similar, while the embeddings of any representations of "polystyrene" serve as negative examples that should be distant.

4.1 Unified embedding space quality

Figure 2 shows a UMAP visualization of PolyBind embeddings colored by polymer chemical families. Chemically similar polymer classes form distinct, well-separated clusters. For instance, polyesters, polyamides, and polyimides each occupy distinct regions, while aromatic polymers (polyphenylenes) separate clearly from aliphatic chains (polyolefins).The tight intra-family clustering demonstrates that PolyBind preserves the chemical knowledge from the pretrained SMILES models, capturing shared structural motifs and chemical properties.

4.2 Cross-modal alignment and retrieval performance

Figure 3 shows the results of retrieval experiments. The plot shows top-k candidates on the x-axis and the fraction of correct matches (recall) on the y-axis for different PSMILES as queries and retrieval responses.

For PSMILES-to-BigSMILES alignment, PolyBind achieves 92% recall at Top@1, reaching 100% at Top@20. PSMILES-to-polymer name retrieval, a more complex task due to the shift from structured notation to natural language, yields 71% recall at Top@1 and 98% at Top@20. For the emergent alignment BigSMILES-to-PSMILES, the recall is close to that of PSMILES-to-BigSMILES because there is only a slight difference between the notations of PSMILES and BigSMILES. The lower performance for BigSMILES-to-polymer name may be attributed to MolFormer's model [26] treatment of tokens like {, }, <, and > as unknowns during tokenization, disrupting alignment with RoBERTa's natural language embeddings.

4.3 Correlation between cosine similarity and recall

To estimate how well calibrated the cosine similarities are in PolyBind's embedding space, we analyzed their relationship with Top@1 recall across cross-modal polymer pairs, as shown in Figure 4. The analysis uses 30 quantile-based bins, with perfect calibration (dashed diagonal) as reference.

For PSMILES-to-BigSMILES retrieval, the pipeline calibration curve consistently lies above the perfect calibration diagonal, indicating systematic overestimation of recall performance.

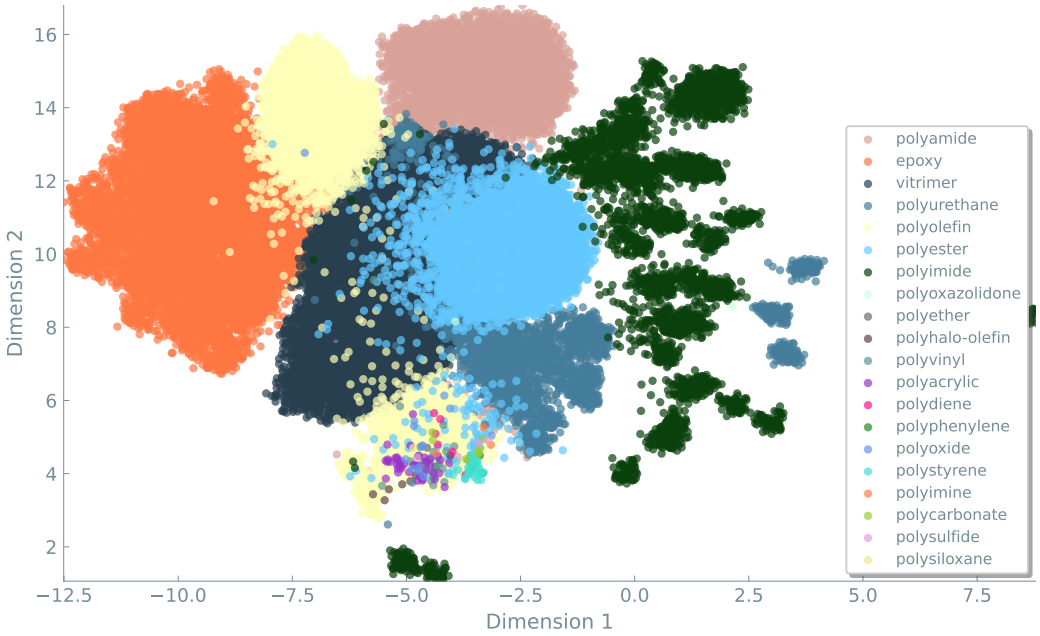

Figure 2: **UMAP projection of PolyBind embeddings along the first two dimensions.** Points correspond to individual polymers, colored by their chemical family (e.g., polyamide, epoxy, vitrimers, polyurethane, polyester, polyimide, polyolefin, polyacrylic, polystyrene, polycarbonate, polysiloxane). The separation across diverse classes highlights that the learned representation encodes meaningful chemical distinctions across polymer space.

For PSMILES-to-polymer name retrieval, the pipeline calibration curve lies below the perfect calibration diagonal in the 0.6–0.8 similarity range, indicating underestimation where actual recall exceeds cosine similarity predictions. Above 0.8, the calibration approaches the diagonal.

4.4 Integrating datasets using aligned identifier representations

To demonstrate PolyBind's utility in integrating datasets, we applied it to glass transition temperature (T_g) prediction. The datasets contain the polymers described using different notations: Poly(isodecyl methacrylate) appears as "[*]CC([*])(C)C(=O)OCCCCCCCCC(C)C" (PSMILES), "$CC(C)(C(=O)OCCCCCCCCC(C)C)$" (BigSMILES) and "Poly(isodecyl methacrylate)". Traditional approaches would treat these as separate entries or require manual mapping.

After alignment, we generated unified embeddings for polymers across PSMILES, BigSMILES, and polymer names. For regression, we used Gradient Boosting Regression (GBR) [34] with default hyperparameters. Polymers appearing in multiple representations were included multiple times in training to reinforce semantic relationships.

We compared PolyBind against baseline methods using Extended Connectivity Fingerprints (ECFP) [35] and PolyBERT [2] on the same T_g property dataset from PolyMetriX [15]. PolyBERT is a DeBERTa-based encoder-only Transformer trained on 100 million hypothetical polymer SMILES strings using masked language modeling, generating 600-dimensional dense fingerprint vectors from PSMILES strings [2]. All models were evaluated on fixed test set using Mean Absolute Error (MAE).

As shown in Table 2, PolyBind achieves a mean MAE of 45.85 ± 0.014 K, compared to the baseline ECFP's 54.56 ± 0.172 K and PolyBERT's 48.507 ± 0.007 K, representing 16.0% and 5.5% error reductions respectively. To isolate the effect of representational diversity from dataset size, we conducted controlled experiments (see Table 3).

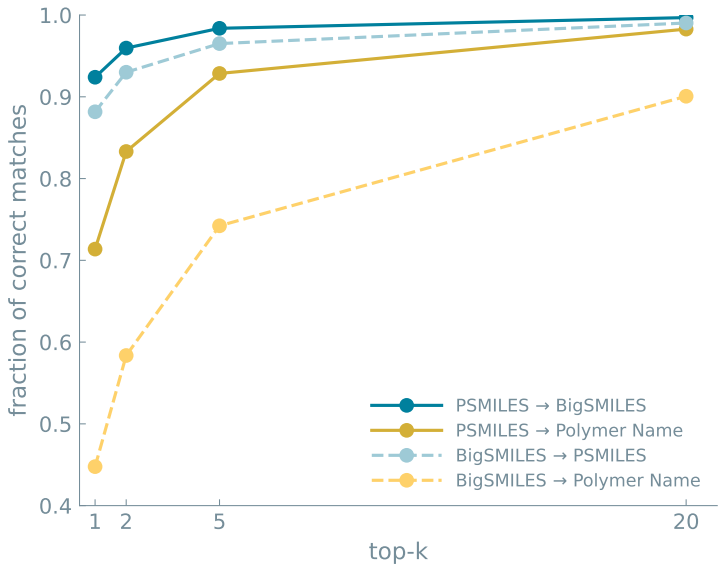

Figure 3: **Cross-modal retrieval performance of PolyBind framework :** Direct alignment pairs (PSMILES-to-BigSMILES and PSMILES-to-polymer name) were explicitly trained during contrastive learning, while emergent alignment pairs (BigSMILES-to-PSMILES and BigSMILES-to-polymer name) arise naturally from the shared embedding space without direct training.

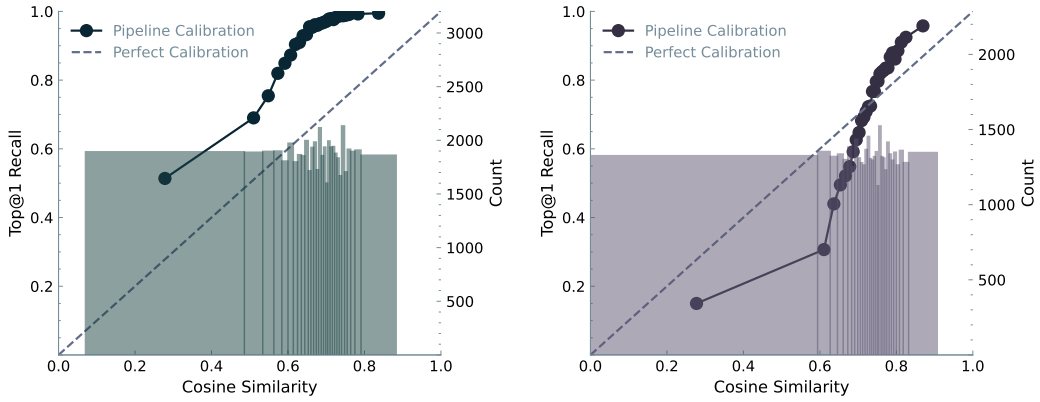

Figure 4: **Calibration analysis of cosine similarity as a predictor of retrieval performance. Left: PSMILES-to-BigSMILES** retrieval shows systematic overestimation with pipeline calibration curve above the diagonal. **Right: PSMILES-to-polymer name** retrieval demonstrates underestimation in mid-range similarities (0.6–0.8) with better calibration at higher similarities.

5 Conclusions

The representational diversity in polymer science has long hindered effective data integration and machine learning applications. PolyBind demonstrates that self-supervised contrastive learning can successfully bridge this gap by creating unified embeddings that transcend notational boundaries. This approach fundamentally shifts the paradigm from representation-specific modeling to universal polymer understanding.

The framework establishes a new standard for handling heterogeneous polymer datasets in computational materials science. By treating different notations as complementary views of the same underlying chemical reality, PolyBind enables researchers to leverage the full spectrum of available polymer data regardless of its original format. This capability transforms previously fragmented datasets into coherent, integrated resources for machine learning.

Table 2: **Comparison of T_g prediction performance between baseline ECFP, PolyBERT and PolyBind.** All the models are evaluated on the fixed test set to ensure fair comparison. Standard deviations are from four runs with different model initialization seeds. PolyBind integrates diverse representations, yielding a larger training set and lower MAE.

Model	Mean MAE (K)	Train Size	Test Size
Baseline ECFP	54.56 ± 0.172	6252	1115
PolyBERT	48.50 ± 0.007	6252	1115
PolyBind	$\mathbf{45.85 \pm 0.014}$	12858	1115

The success of PolyBind points toward broader implications for materials informatics, where similar representational challenges exist across different material classes.

6 Acknowledgments

This work was supported by the Carl Zeiss Foundation.

Parts of A.M.'s work were supported by the Helmholtz Association within the framework of the Helmholtz Foundation Model Initiative (project SOL-AI).

K.M.J. is part of the NFDI consortium FAIRmat funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project 460197019.

References

[1] Yuankai Zhao et al. "A review on the application of molecular descriptors and machine learning in polymer design". In: *Polymer Chemistry* 14.29 (2023), pp. 3325–3346.

[2] Christopher Kuenneth and Rampi Ramprasad. "polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics". In: *Nature communications* 14.1 (2023), p. 4099.

[3] Sunho Choi et al. "Automated BigSMILES conversion workflow and dataset for homopolymeric macromolecules". In: *Scientific data* 11.1 (2024), p. 371.

[4] Tzyy-Shyang Lin et al. "BigSMILES: a structurally-based line notation for describing macromolecules". In: *ACS central science* 5.9 (2019), pp. 1523–1531.

[5] Anagha Savit et al. "polyBART: A Chemical Linguist for Polymer Property Prediction and Generative Design". In: *arXiv preprint arXiv:2506.04233* (2025).

[6] Mario Krenn et al. "SELFIES and the future of molecular string representations". In: *Patterns* 3.10 (Oct. 2022), p. 100588. ISSN: 2666-3899. DOI: 10.1016/j.patter.2022.100588. URL: http://dx.doi.org/10.1016/j.patter.2022.100588.

[7] Mario Krenn et al. "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation". In: *Machine Learning: Science and Technology* 1.4 (2020), p. 045024.

[8] Austin H Cheng et al. "Group SELFIES: a robust fragment-based molecular string representation". In: *Digital Discovery* 2.3 (2023), pp. 748–758.

[9] Matteo Aldeghi and Connor W Coley. "A graph representation of molecular ensembles for polymer property prediction". In: *Chemical Science* 13.35 (2022), pp. 10486–10498.

[10] Owen Queen et al. "Polymer graph neural networks for multitask property learning". In: *npj Computational Materials* 9.1 (2023), p. 90.

[11] Rishi Gurnani et al. "Polymer informatics at scale with multitask graph neural networks". In: *Chemistry of Materials* 35.4 (2023), pp. 1560–1567.

[12] Xinyue Zhang et al. "Polymer-unit graph: advancing interpretability in graph neural network machine learning for organic polymer semiconductor materials". In: *Journal of Chemical Theory and Computation* 20.7 (2024), pp. 2908–2920.

[13] Qinghe Gao et al. "Self-supervised graph neural networks for polymer property prediction". In: *Molecular Systems Design & Engineering* 9.11 (2024), pp. 1130–1143.

[14] H. L. Morgan. "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service." In: *Journal of Chemical Documentation* 5.2 (May 1965), pp. 107–113. ISSN: 1541-5732. DOI: 10.1021/c160017a018. URL: http://dx.doi.org/10.1021/c160017a018.

[15] S Kunchapu and KM Jablonka. "PolyMetriX: An Ecosystem for Digital Polymer Chemistry". In: *ChemRxiv* (2025). This content is a preprint and has not been peer-reviewed. DOI: 10.26434/chemrxiv-2025-s2f2r.

[16] Chiho Kim et al. "Polymer genome: a data-powered polymer informatics platform for property predictions". In: *The Journal of Physical Chemistry C* 122.31 (2018), pp. 17575–17585.

[17] Rohit Girdhar et al. "Imagebind: One embedding space to bind them all". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 15180–15190.

[18] Adrian Mirza et al. "Bridging chemical modalities by aligning embeddings". In: *AI for Accelerated Materials Design - Vienna 2024*. 2024. URL: https://openreview.net/forum?id=KGhnnDR8uv.

[19] Qi Huang et al. "Unified multimodal multidomain polymer representation for property prediction". In: *npj Computational Materials* 11.1 (2025), pp. 1–11.

[20] Erick Isidro. "International Union of Pure and Applied Chemistry Polymer Division Subcommittee on Polymer Terminology A Brief Guide to Polymer Nomenclature". In: *IUPAC Polymer Division, Subcommittee on Polymer Terminology* (2012).

[21] Tianle Yue, Jianxin He, and Ying Li. "Polyuniverse: generation of a large-scale polymer library using rule-based polymerization reactions for polymer informatics". In: *Digital Discovery* 3.12 (2024), pp. 2465–2478.

[22] Ruimin Ma and Tengfei Luo. "PI1M: a benchmark database for polymer informatics". In: *Journal of Chemical Information and Modeling* 60.10 (2020), pp. 4684–4690.

[23] Mitsuru Ohno et al. "SMiPoly: generation of a synthesizable polymer virtual library using rule-based polymerization reactions". In: *Journal of Chemical Information and Modeling* 63.17 (2023), pp. 5539–5548.

[24] Ivan Pavlovich Malashin et al. "Estimation and prediction of the polymers' physical characteristics using the machine learning models". In: *Polymers* 16.1 (2023), p. 115.

[25] Chieh Lin et al. "Essential step toward mining big polymer data: polyname2structure, mapping polymer names to structures". In: *ACS Applied Polymer Materials* 2.8 (2020), pp. 3107–3113.

[26] Jerret Ross et al. "Large-scale chemical language representations capture molecular structure and properties". In: *Nature Machine Intelligence* 4.12 (2022), pp. 1256–1264.

[27] Vaswani Ashish. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017), p. I.

[28] Jianlin Su et al. "Roformer: Enhanced transformer with rotary position embedding". In: *Neurocomputing* 568 (2024), p. 127063.

[29] Yinhan Liu et al. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[31] Zhirong Wu et al. "Unsupervised feature learning via non-parametric instance discrimination". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3733–3742.

[32] Feng Wang and Huaping Liu. "Understanding the behaviour of contrastive loss". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2495–2504.

[33] Lutz Prechelt. "Early stopping-but when?" In: *Neural Networks: Tricks of the trade*. Springer, 2002, pp. 55–69.

[34] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[35] David Rogers and Mathew Hahn. "Extended-connectivity fingerprints". In: *Journal of chemical information and modeling* 50.5 (2010), pp. 742–754.

Appendix

A Full affiliations

FSU Jena Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstrasse 10, 07743 Jena, Germany

- Sreekanth Kunchapu
- Kevin Maik Jablonka

HIPOLE Jena Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstrasse 12-14, 07743 Jena, Germany

- Adrian Mirza
- Kevin Maik Jablonka

HZB Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Hahn-Meitner-Platz 1, 14109

- Adrian Mirza

CEEC Jena Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany

- Kevin Maik Jablonka

JCSM Jena Jena Center for Soft Matter (JCSM), Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany

- Kevin Maik Jablonka

A.1 Data efficiency gains for retrieval

Figure A.1 shows how data efficiency gains in cross-modal retrieval change across different training dataset sizes, while keeping the validation dataset (242k samples) and test dataset (56k samples) fixed.

The results reveal distinct learning patterns between the two cross-modal tasks. PSMILES-to-BigSMILES retrieval demonstrates exceptional data efficiency, achieving high recall@1 performance (0.91) even with minimal training data (2.42k samples). Performance remains consistently stable across all training sizes, showing only marginal improvements as data increases (0.92). This stability arises because PSMILES and BigSMILES are based on SMILES notation and share the same SMILES encoder.

In contrast, PSMILES-to-polymer name retrieval exhibits a fundamentally different learning curve. Performance starts near zero with small datasets and demonstrates continuous improvement as training data increases, reaching 0.71 recall@1 at maximum data size.

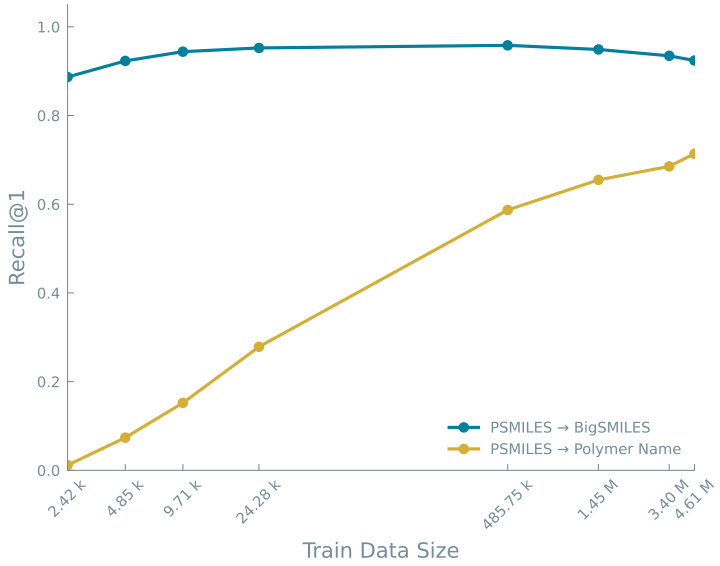

Figure A.1: Data efficiency analysis for cross-modal retrieval tasks. PSMILES-to-BigSMILES retrieval (blue) achieves high performance with minimal training data due to structural similarity between notations. PSMILES-to-polymer name retrieval (gold) requires substantially more training data to bridge the semantic gap between chemical structures and natural language descriptions.

A.2 Success and failure analysis of cross-modal Retrieval by polymer class

Figure A.2 shows cross-modal retrieval performance across polymer classes, revealing patterns that correlate with training data abundance. Training data is dominated by epoxy (1M samples), vitrimer (987k), polyester (518k), and polyamide (424k) classes. The test set maintains similar distributions with vitrimer (10,008), epoxy (9,895), polyester (5,798), polyimide (5,425), and polyamide (4,825) as the most represented classes.

For PSMILES-to-Polymer Name retrieval, epoxy and vitrimer polymers achieve the highest successful retrieval counts, directly reflecting their training abundance and test set representation.

PSMILES-to-BigSMILES retrieval shows substantially different patterns. Success rates remain high across most polymer classes with dramatically reduced failure counts. This improvement stems from the structural similarity between PSMILES and BigSMILES representations, both using standardized chemical syntax for connectivity encoding.

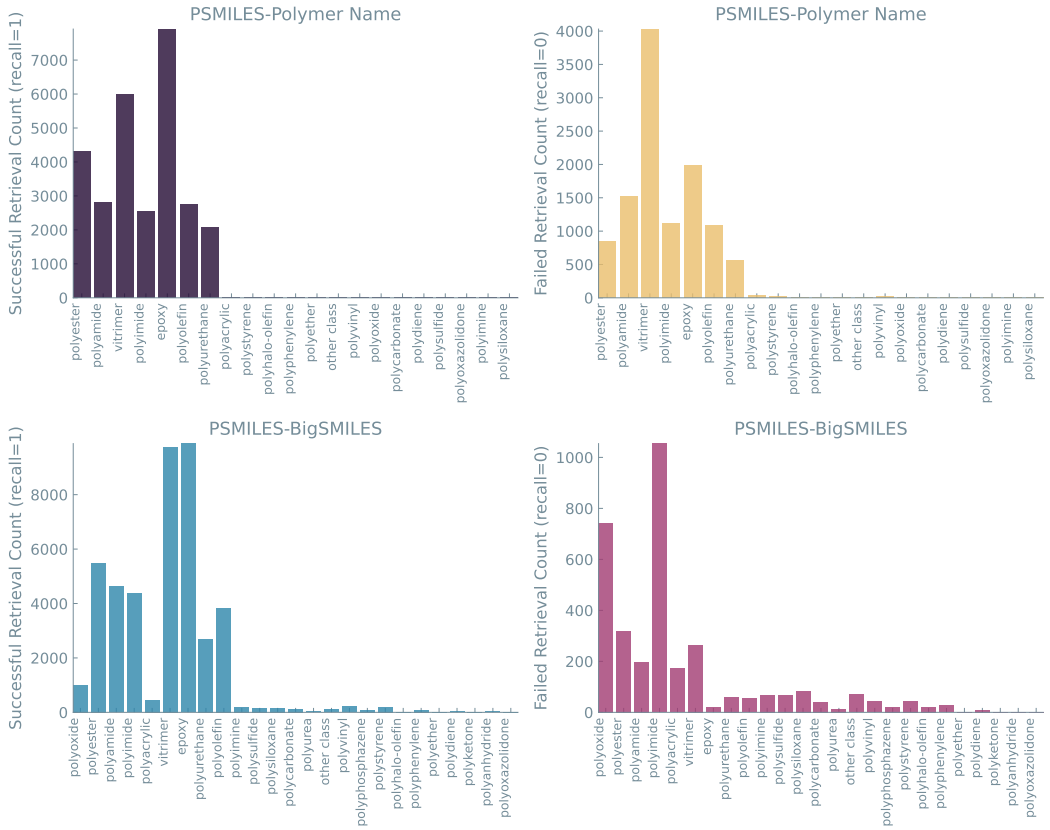

Figure A.2: Cross-modal retrieval performance by polymer class. PSMILES-to-Polymer Name retrieval (top) shows high success for training-abundant classes but notable failure rates due to nomenclature variability. PSMILES-to-BigSMILES retrieval (bottom) demonstrates consistently high success across polymer families due to structural representation similarity.

A.3 Controlled Experiments: dataset size vs. representational diversity

To isolate the effects of dataset size and representational diversity, we conducted controlled experiments using the fixed test set.

ECFP-Doubled doubles the baseline T_g dataset to 12,504 samples by sampling with replacement from the original 6252 samples, using ECFP fingerprint representations. This tests whether increased dataset size improves performance with traditional Morgan fingerprints.

PolyBERT-Doubled similarly doubles the T_g dataset to 12504 samples using PolyBERT fingerprint representations, evaluating the impact of larger dataset size on transformer-based polymer representations.

PolyBind-Mix combines embeddings from multiple polymer representations (PSMILES, BigSMILES, and polymer names), maintaining the original train dataset size of 6252 samples by randomly selecting points across modalities. This assess whether representational diversity alone enhances performance.

Results show that PolyBind-Mix, with 6,252 samples, reduces MAE by 17.8% compared to ECFP-Doubled, while doubling the ECFP dataset size yields no improvement over the baseline. This indicates performance gains arise from improved representations rather than increased dataset size.

Table 3: Comparison of dataset size and representational diversity effects on a fixed test set of 1,115 samples. Mean MAE (K) and standard deviation are reported over four runs with different seeds for the model.

Model	**Mean MAE (K)**	**Train Size**	**Test Size**
ECFP-Doubled	55.353 ± 0.065	12,504	1,115
PolyBERT-Doubled	47.024 ± 0.018	12,504	1,115
PolyBind-Mix	$\mathbf{45.099 \pm 0.009}$	6,252	1,115