How Aligned Are Unimodal Language and Graph Encodings of Chemical Molecules?

Congfeng Cao Zhi Zhang Jelke Bloem Khalil Sima'an
Institute for Logic, Language and Computation
University of Amsterdam
{c.cao, z.zhang, j.bloem, k.simaan}@uva.nl

Abstract

Chemical molecules can be represented as graphs or as language descriptions. Training unimodal models on graphs results in different encodings than training them on language. Therefore, the existing literature force-aligns the unimodal models during training to use them in downstream applications such as drug discovery. But to what extent are *graph* and *language* unimodal model representations inherently aligned, i.e., aligned prior to any force-alignment training? Knowing this is useful for a more expedient and effective forced-alignment. For the first time, we explore methods to gauge the alignment of graph and language unimodal models. We find compelling differences between models and their ability to represent slight structural differences without force-alignment. We also present an unified unimodal alignment (**U2A**) benchmark for gauging the inherent alignment between graph and language encoders which we make available with this paper[1].

1 Introduction

In chemistry, molecules can be represented in two primary formats: a language description or a molecular graph (Zeng et al., 2022). Table 7 provides an example of the malic acid molecule in different formats (National Center for Biotechnology Information, 2025). The language description can be encoded using language models such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018, 2019). Molecular graphs can be encoded either explicitly or implicitly. Explicit encoding uses graph-based models (Hu* et al., 2020; Xia et al., 2023), such as Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) and Graph Isomorphism Networks (GINs) (Xu et al., 2019), which directly take the molecular graph as input. Alternatively, implicit encoding uses sequence models (Wang et al., 2019; Chithrananda et al., 2020), where the input sequence is derived from the graph, such as the Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988; Zeng et al., 2022). SMILES sequences are transformed from molecular graphs using depth-first tree traversal.

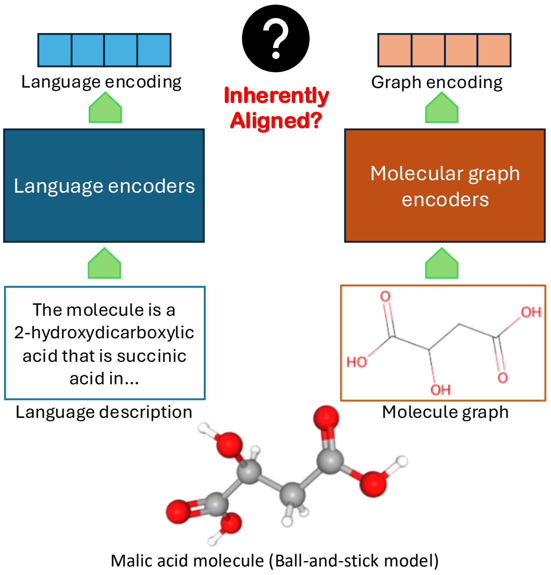


Figure 1: How Aligned Are Unimodal Language and Graph Encodings of Chemical Molecules?

The fact that different modalities, graph and language, demand different kinds of unimodal models and result in different encodings contrasts with the need for encodings of the same object to align together. Consequently, as shown in Figure 1, in this paper we ask whether unimodal graph and language encodings corresponding to the same chemical molecules are inherently aligned together, i.e., without specifically training them under an alignment objective.

Unlike unimodal language and vision encodings, gauging the alignment between graph and language unimodal encodings of molecules has not been studied before. Efforts have concentrated on

[1] GitHub link: U2A Benchmark Repository

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1084–1097
December 20-24, 2025 ©2025 Association for Computational Linguistics

force-aligning the unimodal encodings during training, to be used in downstream applications such as drug discovery (Seidl et al., 2023) and new material design (Jablonka et al., 2024). Force-aligning graph-language models aims to bring the two types of embeddings of the same object closer together, while pushing the embeddings of different objects further apart, thereby preserving the properties of both modalities. For example, MoleculeSTM integrates molecular structures and textual knowledge to enable zero-shot drug design based on text-based instructions (Liu et al., 2023).

In this paper, we gauge the inherent alignment between graph and language unimodal representations. We believe that encoders that are better aligned together (without the need for force-aligning them) could speed up forced-alignment training and reduce the amount of training data needed, as we show in our application experiment results. Creating paired graph and language (parallel) data for molecules is costly, requiring extensive experimental validation and expert knowledge (Mayr et al., 2018). For example, bioassays and bioactivities often come from wet-lab procedures with multiple chemical and biological steps.

In contrast with vision and language unimodal representations for which there are unified platforms (e.g., Hugging Face), there exist no unified platforms to gauge the alignment of graph and language representations. Therefore, we present a novel unified benchmark bringing together graph- and language-models used for encoding molecule representations from various platforms, ensuring their compatibility.

Gauging alignment between graph and language encodings of molecules provides a coarse-grained approach because the inputs to the unimodal models do not focus on the topology of the graphs. Here, we also take a fine-grained approach which focuses on the topology of the molecular representations. To do so, we leverage data concerning *Isomers* and *Tautomers*. Isomers share the same molecular formula, with identical atom counts for each element, but differ in structure. Tautomers, a specific type of isomers, result from interconversion, such as the relocation of a hydrogen atom within the molecule, creating distinct yet similar structures. Isomers and tautomers are derived from the original molecule graphs, consisting of language-graph pairs. We compare the structural alignment of molecular graphs using original molecules, isomers, and tautomers, and analyse their alignment scores.

Centered Kernel Alignment (CKA; Kornblith et al., 2019) is widely used for alignment evaluation, and in this work, we adopt Debiased CKA (Murphy et al., 2024), a more robust metric, with our experiments further validating its effectiveness.

To the best of our knowledge, this is the first study to explore the inherent alignment between graph models and language models. Our key findings and contributions are as follows:

- Inherent alignment: Our study is the first to demonstrate that unimodal graph and language models exhibit inherent alignment. Moreover, we reveal that this alignment varies between different unimodal models.
- Structure-level alignment: We also show that these unimodal models align at both coarse-grained and fine-grained structural levels. MolCLR with GIN is more strongly structurally aligned with language models than with GCN.
- Unified benchmark: We contribute a benchmark for evaluating the inherent alignment between graph and language encoders.[2]

2 Related Work

Force-aligned Graph-Language Models Force-aligned graph-language models have gained attention due to their powerful multimodal capabilities. MoleculeSTM integrates molecular structures and textual knowledge, enabling drug design using text-based instructions and biological activity prediction (Liu et al., 2023). MolFM, a multimodal foundation model, is designed to facilitate joint representation learning from molecular structures, biomedical texts, and knowledge graphs (Luo et al., 2023). MoMu (Su et al., 2022) and CLAMP (Seidl et al., 2023) use contrastive learning to align molecular graphs and related textual data, or for activity prediction. All these models widely use transformer-based SMILES encoders, such as MegaMolBART (Irwin et al., 2022), and graph-based encoders, such as Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017), to encode molecules. Meanwhile, they also use transformer-based language encoders, such as SciBERT (Beltagy et al., 2019), to encode language descriptions. In this paper, we use MolFM as an upper-bound due to its outstanding performance on various downstream tasks.

[2] GitHub link: U2A Benchmark Repository

Molecular Models Molecular models have been widely explored in both graph and SMILES formats. Models such as MolBERT (Fabian et al., 2020) and ChemBERT (Zhang et al., 2022), inspired by the BERT architecture, are trained on SMILES representations to capture molecular features through self-supervised learning. Similarly, models like Chemformer (Irwin et al., 2022; Westerlund et al., 2024), MolBART (Irwin et al., 2022), and BARTSmiles (Chilingaryan et al., 2024) are based on the BART architecture, focusing on sequence-to-sequence learning, denoising, and molecular generation from SMILES data. MolFormer (Ross et al., 2022) is a transformer-based model trained on SMILES sequences of 1.1 billion unlabeled molecules from the PubChem and ZINC datasets. On the other hand, MolCLR (Wang et al., 2022) is a graph-based model that uses contrastive learning on molecular graphs and uses both Graph Convolutional Networks (GCN) and Graph Isomorphism Networks (GIN) for comparison. In this paper, we use MolFormer and MolCLR for the experiment, as they encode SMILES and molecular graphs respectively.

Language Models Language models encompass a broader range of architectures. ModernBERT represents the state-of-the-art in the BERT series, bringing modern optimizations to bidirectional encoder-only models and offering a significant Pareto improvement over older encoders (Warner et al., 2025). SciBERT, a BERT-based model, leverages unsupervised pretraining on a multi-domain scientific corpus to enhance scientific NLP tasks (Beltagy et al., 2019). SentenceBERT is a specialized BERT-based model that employs siamese and triplet network structures to generate semantically meaningful sentence embeddings (Reimers and Gurevych, 2019). The GPT series models, developed with an auto-regressive approach for text generation, have also been found effective for text embedding tasks in various studies (Radford et al., 2018, 2019). In this paper, we explore various BERT-series models, including SciBERT, SentenceBERT, and ModernBERT, as BERT-like models are widely used as language encoders in force-aligned models. Furthermore, we also explore GPT as an extension.

Alignment Metrics Several approaches have explored the alignment and similarity of different encoders. Fundamental similarity measures include cosine similarity and Canonical Correlation Analysis (CCA) (Hardoon et al., 2004) along with their extensions, such as Singular Vector Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017). Kornblith et al. (2019) proposed Centered Kernel Alignment (CKA) as a measure of representation similarity. CKA can be biased, particularly with limited data, and it may yield artificially high similarity scores even for random matrices due to its sensitivity to differing feature dimensions; *Debiased* CKA has been introduced to mitigate this issue (Murphy et al., 2024). We first validate the effectiveness of Debiased CKA compared to traditional CKA and then adopt it as the metric for subsequent experiments.

Molecular Graph Similarity The Jaccard index method measures graph similarity based on edge set overlap (Strehl et al., 2000). The Jaccard index method is also used to measure fingerprint overlap between molecules for assessing molecular function similarity (Chung et al., 2019). Graph edit distance counts the minimum operations (insertion, deletion, substitution) needed to transform one graph into another (Wilson and Hancock, 1997). Graph kernels compare substructures such as walks, paths, and subtrees (Gärtner et al., 2003). The spectral method uses eigenvalues of adjacency or Laplacian matrices to assess similarity (Wilson and Hancock, 1997; Wilson and Zhu, 2008). It is mathematically robust and invariant to node ordering. In this paper, we use both the spectral method and the Jaccard index to evaluate graph similarity and molecular functional similarity.

3 Data Collection, Processing and Analysis

3.1 Collection of Isomers and Tautomers

To explore structural-level inherent alignment between molecular graphs and language descriptions, we collect isomers and tautomers for each molecule and designate three of them as control groups. Our base dataset is ChEBI 20, an open-source dataset introduced in the Text2Mol paper (Edwards et al., 2021). This dataset contains $33,008$ molecular SMILES and language description pairs. Isomers can be found from molecular SMILES using the open-source toolkit RDKit (Landrum et al., 2025), while tautomers can be collected via the open API MAYGEN (Yirik et al., 2021). RDKit is a collection of cheminformatics and machine-learning software written in C++ and Python. MAYGEN is an open-source chemical structure generator that offers an API for tautomer generation. However, some molecules do not have isomers or tautomers, and the toolkit and API lack information for all $33,008$ molecules. After filtering, $3,515$ items remain. Since a single molecule can have multiple isomers and tautomers, we randomly select one isomer and one tautomer as candidate molecules.

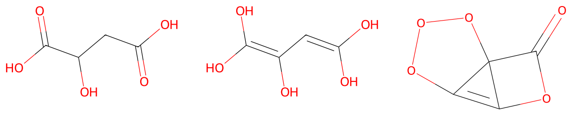
3.2 Extraction of Chemically Relevant Terms

Language descriptions often contain general terms that are not relevant to graph structures or molecular functions, such as phrases like *"The molecule is..."*. Removing these may yield better representations, though they may also encode important indirect associations. Therefore, in our experimental setup, we test both natural language descriptions and term-extracted descriptions, for which we apply an extraction process to identify and extract chemically relevant terms. Specifically, we first use ChatGPT-4o-mini to extract an initial list of terms related to graphs and chemistry from each language description. This list is then manually curated to ensure the relevance and accuracy of the extracted terms. Finally, we obtain the extracted chemical terms and pair them with the corresponding molecules. Section B provides details about the extraction prompt.

3.3 Molecular Similarity Analysis

Theoretically, tautomerization conversion commonly results from the relocation of compounds within the molecule. Therefore, tautomers have a more similar structure to the original molecular graph, whereas isomers can be completely different. A concrete example can be seen in the molecular formula $C_4H_6O_5$, illustrated in Figure 4. To verify whether our collected dataset aligns with this theoretical expectation, we assessed the similarity between the original molecules and their isomers, as well as their tautomers. We evaluate two kinds of similarities: graph similarity and functional similarity.

Graph similarity We use spectral distance (Wilson and Zhu, 2008), which ranges from zero to positive infinity, as a metric for evaluating graph similarity, as mentioned in Section 2. In our dataset, the mean spectral distance between the original molecular graphs and their isomers is 2.69, while the mean spectral distance to their tautomers is 2.03. These values indicate that tautomer graphs are more similar to the original molecular graphs than isomer graphs.



(a) Original (b) Tautomer (c) Isomer

Figure 2: Original $C_4H_6O_5$ and its tautomer and isomer. The original molecular is more similar to the tautomer compared with the isomer. The structure shows nodes (carbon implied at vertices) connected by bonds.

Functional similarity To assess molecular functional similarity, we use the Jaccard index to measure fingerprint overlap (ranging from zero to one) (Chung et al., 2019). Molecular fingerprints are molecular descriptors that describe a molecule's features or functional groups as binary digits. The functional similarity between original molecules and their isomers is 0.029, whereas that between original molecules and their tautomers is 0.4178. These values, consistent with graph similarity results, further indicate that tautomers are functionally more similar to the original molecules.

After data processing, we obtain a dataset in different molecular graphs (original molecules, isomers and tautomers) and different corpora (language description and extraction of chemical terms), resulting in six control groups (see in section C). An example of the dataset is shown in Appendix D, and the structural visualizations of isomers and tautomers are in Figures 2b and 2c. More examples are provided in Appendix E.

4 Experimental Setting

4.1 Questions and hypotheses

Is There Inherent Alignment Between Unimodal Graph and Language Encodings? We hypothesize that there exists a degree of inherent alignment between molecule graph and language encodings. This may vary across different graph and language encoders and it is influenced by the representation capacity and pretraining of the unimodal encoders. Moreover, inherent alignment is also affected by the architecture, as prior research (Csiszárik et al., 2021) suggests that models with similar architectures tend to exhibit higher alignment.

To What Extent Is Structure-Level Inherent Alignment Present? We further hypothesize that the inherent alignment between graph and language encodings is sensitive to structural variations due to the topological encoding capability of the graph encoder.

4.2 Setup

In Section 5.1, we use MolFM and its randomly initialized counterpart as the upper- and lower- bound, respectively, because MolFM achieves state-of-the-art performance with force-aligned training and randomly initialized encoders lack training and modality-specific properties. Our experiment for the two models builds three control groups respectively as reference bounds (see Appendix G for details). We gauge the alignment of MolFM between the same language descriptions and different molecular structures (original molecules, tautomers, and isomers). As described in Section 3.3 and Section 4.1, tautomers are more similar to original molecules than isomers. The alignment score between the language description and original molecules, tautomers, and isomers should theoretically decrease progressively (three-descending pattern). We use this pattern as a reference for gauging structural inherent alignment.

In Section 5.2, using the defined bounds, we gauge inherent alignment between unimodal graph and language within six different language encoders (BERT, SciBERT, SentenceBERT, ModernBERT, and GPT) and two molecular graph encoders (MolCLR and MolFormer) (Appendix H).

In Section 5.3, we gauge different inherent alignment across different structures, considering structural changes from the original molecules to isomers and tautomers. Theoretically, sharing the same language encoding, inherent alignment should decrease as structures change from original molecules to isomers and tautomers. We evaluated whether unimodal graph and language models can detect major structure changes from the original molecular graph to isomers with decreasing Debiased CKA scores and whether the models' inherent alignments follow the three-descending pattern as well as force-aligned models.

In Section 5.4, we present our application experiments. To investigate whether better-aligned encodings can accelerate forced-alignment training, we train a two-layer multilayer perceptron (MLP) to align graph and language encodings of the same dimensionality using two different training regimes (1,000 and 10,000 steps), and measure alignment improvement using Debiased CKA scores. To evaluate whether higher inherent alignment translates into better retrieval performance, we train another two-layer MLP to perform cross-modality retrieval between the graph and language encodings. In both experiments, we randomly sample 200 unimodal graph–language encoding pairs, using 100 for training and 100 for testing. To ensure robustness, we conduct each experiment over 10 runs and report the mean Top-10 accuracy for the second task.

4.3 Encoders

We selected multiple language and molecular encoders for our experiments based on their performance and their use in force-aligned models.

Language Encoders As mentioned in Section 2, BERT-like models (Devlin et al., 2019; Reimers and Gurevych, 2019; Beltagy et al., 2019) are widely used in molecular graph-language models as language encoders for encoding language descriptions. Therefore, we select BERT, the domain-specific SciBERT, and Sentence-BERT as candidate language models. In addition, we explore the recent ModernBERT (Warner et al., 2025), which brings modern model optimizations to encoder-only models. GPT-series models (Radford et al., 2018, 2019) are autoregressive and are widely used for generative tasks. These models can also be used for language embedding (Jiang et al., 2024). We employ the [CLS] token and mean pooling as language encodings for BERT-like models, whereas for GPT-like models, we use mean pooling as language encodings.

Molecular Graph Encoders As mentioned in Section 2, there are two main categories of molecular models. Adapted language models trained on SMILES sequences, such as MolFormer (Ross et al., 2022), implicitly encode molecular graphs, while GNN-based encoders, such as Graph Convolutional Networks (GCN) and Graph Isomorphism Networks (GIN), explicitly encode molecular graphs, as seen in MolCLR (Wang et al., 2022). Therefore, we employ both MolFormer and MolCLR as molecular encoders to explore both implicit and explicit methods.

4.4 Benchmark Configuration and Release

We present an unified unimodal alignment (**U2A**) benchmark, which supports three encoder types: graph, language, and force-aligned graph-language models. The U2A benchmark includes an evaluation toolkit for alignment scores, such as Centered

Kernel Alignment (CKA) (Kornblith et al., 2019), Debiased CKA (Murphy et al., 2024), and other metrics. Our U2A benchmark reduces the effort required to find and configure encoders and evaluation tools, as various graph and language models originate from different platforms and applications. The U2A benchmark unifies these models, ensuring compatibility and facilitating further research. Appendix F provides runtime details.

5 Results and Analysis

In our experiments, we first establish the upper and lower bounds. Then, we explore the inherent alignment between molecular graphs and language encoders, followed by an assessment of structural-level alignment.

5.1 Upper-, Lower-Bound and Comparison

As set up in Section Section 4.2, we use MolFM to build the upper-bound and its randomly initialized counterpart to build the lower-bound.

CKA v.s. Debiased CKA As shown in Table 1, the randomly initialized encoder shows a low CKA score, which should be zero, indicating a bias in the general CKA compared to Debiased CKA (Murphy et al., 2024). To further assess these two metrics, we compare them using ten independent random samples, each selecting $2,000$ out of $3,515$ molecular graph and language pairs. The results are presented in Appendix K. The results show that beyond reducing bias, Debiased CKA is more stable than general CKA, with lower standard deviation. Therefore, we employ **Debiased CKA** as the metric for our subsequent experiments.

Alignment Comparison As shown in Table 1, all Debiased CKA scores from MolFM outperform those from the randomly initialized encoder, demonstrating that force-alignment training effectively aligns the two encoders. The Debiased CKA score between the original molecule and language description from MolFM is nearly 0.5, while that from randomly initialized encoders is close to zero.

Tautomers v.s. Isomers In a comparison between the original molecule, tautomers, and isomers, the scores exhibit three-descending pattern: 0.497, 0.431, and 0.021 (Table 1). This result aligns with our theoretical expectation, demonstrating that the force-aligned graph-language model can effectively detect structural differences.

Model	Method	Language Type	Molecule Type		
			Ori.	Iso.	Tau.
MolFM	CKA	Language des.	0.499	0.022	0.434
	De. CKA	Language des.	0.497	0.021	0.431
	CKA	Extraction	0.508	0.023	0.443
	De. CKA	Extraction	0.505	0.021	0.440
Random Initial	CKA	Language des.	0.068	0.068	0.068
	De. CKA	Language des.	0.000	0.000	0.000
	CKA	Extraction	0.068	0.068	0.069
	De. CKA	Extraction	0.000	0.000	0.001

Table 1: Performance comparison of upper- and lower-bound with CKA and Debiased CKA (De. CKA). Ori., Iso. and Tau. represent original molecule, isomers and tautomers respectively.

Language Description v.s. Chemical Extraction The results also show a slight improvement in alignment between molecular graphs and extracted chemical descriptions compared to that between molecular graphs and language descriptions, indicating that MolFM is more sensitive to chemical information.

5.2 Is There Inherent Alignment Between Unimodal Graph and Language Encodings?

5.2.1 MolCLR

Inherent Alignment between MolCLR and Language Encoders MolCLR employs two types of graph encoders: Graph Convolutional Networks (GCN) (Kipf and Welling, 2017) and Graph Isomorphism Networks (GIN) (Xu et al., 2019). Therefore, we explore the inherent alignment between these two encoders and language encoders. As presented in Table 2, all Debiased CKA scores between graph embeddings and language embeddings fall within the upper- and lower- bound. For example, Debiased CKA scores are 0.246 (MolCLR (GCN) and SciBERT mean pooling), 0.256 (MolCLR (GCN) and SentenceBERT), 0.206 (MolCLR (GIN) and SciBERT mean pooling), and 0.222 (MolCLR (GIN) and SentenceBERT). These results demonstrate the inherent alignment that already exists in MolCLR and language encoders.

Different Degrees of Alignment Between MolCLR and Language Encoders The results reveal that different unimodal models exhibit varying degrees of alignment. For example, the Debiased CKA score of embeddings from GPT mean pooling and MolCLR (GCN) is 0.235 whereas the Debiased CKA score of embeddings from SciBERT mean pooling and MolCLR (GCN) is 0.246.

Model	Embedding	MolCLR		MolFormer
		GCN	GIN	
SciBERT	CLS	0.215	0.184	0.267
	Mean	0.246	0.206	**0.387**
	Extraction CLS	0.199	0.171	0.299
	Extraction Mean	0.228	0.193	0.361
SentenceBERT	Mean Embedding	**0.256**	**0.222**	0.356
	Extraction Mean	0.202	0.181	0.313
BERT	CLS	0.156	0.132	0.209
	Mean	0.109	0.093	0.165
	Extraction CLS	0.101	0.091	0.157
	Extraction Mean	0.094	0.083	0.146
ModernBERT	CLS	0.141	0.105	0.143
	Mean	0.152	0.114	0.157
	Extraction CLS	0.117	0.083	0.131
	Extraction Mean	0.071	0.053	0.096
GPT	Mean	0.235	0.198	0.298
	Extraction Mean	0.254	**0.222**	0.344

Table 2: Debiased CKA score comparison between different language models with different embedding types and MolCLR (GCN and GIN), as well as between that and MolFormer. The double and single underlines mark the highest and second-highest CKA scores. CLS and Mean represent the language embedding extracted from [CLS] embedding and mean pooling over all output embedding, respectively. with and without Extraction means the language model takes full language description and the extracted chemical terms as input.

ModernBERT, the state-of-the-art model, does not demonstrate an advantage in inherent alignment while SentenceBERT achieves the highest Debiased CKA scores. In MolCLR, the results also show that GCN get higher Debiased CKA scores across all language models and embedding methods. For example, Debiased CKA score of embeddings from ModernBERT mean pooling and MolCLR (GCN) is 0.152 while Debiased CKA score of embeddings from ModernBERT mean pooling and MolCLR (GIN) is 0.114.

Language Description v.s. Chemical Extraction Unlike the force-aligned graph-language model, the extraction of chemical information (Extraction condition) does not provide advantages across BERT-series models. The embeddings of extracted information show a decline across all BERT-series language encoders, such as the score between SciBERT mean pooling and MolCLR (GCN) decrease from 0.246 to 0.228.

In MolCLR, both MolCLR(GIN) and MolCLR(GCN) combined with SentenceBERT and SciBERT achieve the highest and second-highest Debiased CKA scores. This result aligns with our hypothesis that SentenceBERT and SciBERT have advantages in representation expression that achieve better inherent alignment with MolCLR(GIN) and MolCLR(GCN). The results also show that GCN consistently obtains higher Debiased CKA scores across all language models and embedding methods, which is a new phenomenon that has not been reported before.

5.2.2 MolFormer

Inherent Alignment between MolFormer and Language Encoders MolFormer, which uses SMILES as input, is adapted from masked language models. As seen in Table 2, the results also fall within the upper- and lower- bound, revealing inherent alignment between MolFormer and various language models, such as 0.267 (MolFormer and SciBERT) and 0.356 (MolFormer and SentenceBERT)

Different Degrees of Alignment Between MolFormer and Language Encoders The highest Debiased CKA score, 0.387, is observed between MolFormer and SciBERT with mean pooling, followed by SentenceBERT with mean pooling at 0.356. The results indicate that different unimodal models exhibit varying degrees of alignment. SciBERT with MolFormer achieves the highest score, while ModernBERT achieves the lowest scores across all embedding types.

Language Description v.s. Chemical Extraction Apart from SciBERT and GPT, the extraction of chemical information leads to a decline in the Debiased CKA score, such as a decrease from 0.356 to 0.313 with SentenceBERT.

The results are consistent with our hypothesis. On the graph side, MolFormer is pretrained on 1.1 billion SMILES sequences using a Transformer-based architecture and has the largest parameter size (46.8M). Therefore, MolFormer exhibits the strongest inherent alignment across language encoders. In contrast, MolCLR is pretrained on 10 million unique molecules, with a GCN variant having 3.98M parameters and a GIN variant with 9.18M parameters. MolCLR (GCN) and MolCLR (GIN) generally demonstrate similar inherent alignment. On the language side, BERT-based encoders with similar parameter sizes tend to exhibit comparable representational capacity. SentenceBERT (Reimers and Gurevych, 2019), which is trained to enhance sentence representations, may retain advantages in representing chemical language. SciBERT (Beltagy et al., 2019), pretrained on a

		GIN			GCN		
Model	**Emb.**	**Ori.**	**Iso.**	**Tau.**	**Ori.**	**Iso.**	**Tau.**
SciBERT	CLS	0.184	0.058	0.120	0.215	0.035	0.222
	Mean	0.206	0.084	0.148	0.246	0.054	0.261
	Extraction CLS	0.171	0.043	0.111	0.199	0.026	0.199
	Extraction Mean	0.193	0.075	0.141	0.228	0.047	0.237
SentenceBERT	Mean	0.222	0.043	0.126	0.256	0.026	0.250
	Extraction Mean	0.181	0.044	0.113	0.202	0.027	0.197
BERT	CLS	0.132	0.048	0.091	0.156	0.031	0.161
	Mean	0.093	0.058	0.076	0.109	0.041	0.128
	Extraction CLS	0.091	0.028	0.070	0.101	0.018	0.098
	Extraction Mean	0.083	0.049	0.067	0.094	0.034	0.104
ModernBERT	CLS	0.105	0.060	0.070	0.141	0.037	0.141
	Mean	0.114	0.077	0.082	0.152	0.046	0.158
	Extraction CLS	0.083	0.062	0.069	0.117	0.039	0.125
	Extraction Mean	0.053	0.046	0.047	0.071	0.031	0.076
GPT	Mean	0.198	0.058	0.115	0.235	0.040	0.253
	Extraction Mean	0.222	0.062	0.131	0.254	0.043	0.266

Table 3: Structural levels inherent alignment between different language models and the MolCLR with GIN or GCN. Ori., Iso. and Tau. represent original molecule, isomers and tautomers respectively.

large-scale scientific corpus within the biomedical domain, may also have an advantage in representing chemical language, leading to better inherent alignment with graph encoders.

5.3 To What Extent Is the Structure-Level Inherent Alignment Present?

5.3.1 MolCLR

Structure-Level Inherent Alignment Between MolCLR and Language Encoders The results from Table 3 show that MolCLR with GIN follows the same three-descending pattern as the force-aligned model MolFM across all graph and language encoders, as the Debiased CKA score follows the three-descending pattern for original molecules, tautomers, and isomers. For example, the Debiased CKA scores between MolCLR (GIN) and mean pooling of SciBERT are 0.206, 0.148, and 0.084, respectively.

However, the results also show varying performance under the same settings with MolCLR (GCN). For example, the scores between MolCLR (GCN) and mean pooling of SciBERT are 0.246, 0.261, and 0.054 for the original molecular graph, tautomers, and isomers, respectively. This pattern diverges from that of the force-aligned model. The results indicate that MolCLR (GCN) is not sensitive to slight structural differences between the original molecule and its tautomers, but can identify major structural differences between the original molecule and its isomers.

Language Description v.s. Chemical Extraction In Table 3, for extraction of chemical terms, the Debiased CKA scores between MolCLR (GIN) and mean pooling of SciBERT are 0.193, 0.141, and 0.075 for the original molecular graph, tautomers, and isomers, respectively, while the scores between MolCLR (GCN) and mean pooling of SciBERT are 0.228, 0.237, and 0.047, respectively. The extraction of chemical information with MolCLR follows the same pattern: MolCLR (GIN) can identify structural changes across original molecules, tautomers, and isomers, whereas MolCLR (GCN) can only detect major structural changes.

5.3.2 MolFormer

Structure-Level Inherent Alignment Between MolFormer and Language Encoders As seen in Table 4, the results show that the Debiased CKA scores between MolFormer and language models, including ModernBERT and GPT, follow the same pattern as the force-aligned model MolFM. With GPT, the Debiased CKA scores are 0.298, 0.276, and 0.107. This pattern aligns with the force-aligned model when using chemical extraction. However, MolFormer with other language models does not follow the same pattern, indicating that while they can identify major structural changes between the original molecule and its isomers, they cannot detect slight structural changes between the original molecule and its tautomers.

Language Description v.s. Chemical Extraction As seen in Table 4, the extraction of chemical information with MolFormer does not show advantages in detecting structural changes. For example, the Debiased CKA scores between MolFormer and mean pooling of SciBERT are 0.361, 0.377, and 0.131 for the original molecular graph, tautomers, and isomers, respectively, which does not exhibit the same structural pattern with force-aligned model when use chemical extraction. The Debiased CKA scores between MolFormer and GPT follow the same pattern as the force-aligned model MolFM, whereas that scores between MolFormer and other language encoders do not.

The results are consistent with our hypothesis. Structural-level alignment is influenced by the topological encoding capability of the encoders. MolFormer encodes graphs as SMILES, which loses the topological information of the graph and thus limits its ability to capture topology. Therefore, although MolFormer exhibits the strongest inherent alignment, it can only detect significant structural changes. GCN is weaker than GIN in distinguish-

Model	Embedding	Ori.	Iso.	Tau.
SciBERT	CLS	0.267	0.108	0.278
	Mean	0.387	0.149	0.394
	Extraction CLS	0.299	0.086	0.315
	Extraction Mean	0.361	0.131	0.377
BERT	CLS	0.209	0.085	0.215
	Mean	0.165	0.099	0.177
	Extraction CLS	0.157	0.054	0.166
	Extraction Mean	0.146	0.081	0.157
SentenceBERT	Mean	0.356	0.103	0.365
	Extraction Mean	0.313	0.085	0.321
ModernBERT	CLS	0.143	0.093	0.137
	Mean	0.157	0.119	0.155
	Extraction CLS	0.131	0.098	0.138
	Extraction Mean	0.096	0.090	0.110
GPT	Mean	0.298	0.107	0.276
	Extraction Mean	0.344	0.112	0.328

Table 4: Structural levels inherent alignment between MolFormer and different language models.

	Before	1,000 steps		10,000 steps	
Model	Training	After	Δ ↑	After	Δ ↑
SentenceBERT+MolCLR	0.263	0.292	0.029	0.347	0.084
SciBERT+MolFormer	0.484	0.542	0.058	0.670	0.186

Table 5: Debiased CKA before training and after 1,000/10,000 steps. Δ ↑ indicates the improvement over the initial value.

ing graph structures. Theoretically, GCN is non-injective due to mean aggregation, whereas GIN achieves injective mappings through sum aggregation, matching the expressiveness of the Weisfeiler-Lehman test (Xu et al., 2019). Empirical studies (Hu et al., 2020; Xiao et al., 2024) also demonstrate GIN's superior structural performance. Consequently, MolCLR (GCN) and MolCLR (GIN) achieve similar levels of inherent alignment with language encoders, while MolCLR (GIN) is expected to achieve better structural alignment.

5.4 Application Experiments

Better Inherently Aligned Encodings Accelerate Forced-Alignment Training As shown in Table 5, for both training regimes, encoding pairs from SciBERT+MolFormer, initially exhibiting higher Debiased CKA, achieve greater improvements after training compared to SentenceBERT+MolCLR (GCN). These results indicate that, given the same number of training steps, encodings with better inherent alignment can accelerate forced-alignment training.

Model	Debiased CKA	Top-10 Accuracy
Random Initialization	0.000	6.4%
SentenceBERT + MolCLR (GCN)	0.245	24.0%
SciBERT + MolFormer	0.388	36.8%
MolFM	0.576	59.5%

Table 6: Higher inherent alignment (as measured by Debiased CKA) corresponds to better retrieval performance.

Better Inherently Aligned Encodings Benefit Cross-Modal Retrieval As shown in Table 6, encoding pairs with higher inherent alignment (SciBERT + MolFormer, Debiased CKA of 0.388) achieve better retrieval performance (36.8%) than those with lower inherent alignment (SentenceBERT + MolCLR, Debiased CKA of 0.245, accuracy 24.0%). Encoding pairs from MolFM, a forced-alignment model, serve as a baseline and outperform both in terms of Debiased CKA and retrieval accuracy. In contrast, encoding pairs from randomly initialized models achieve only 6.4% mean Top-10 accuracy. These results further demonstrate a strong positive correlation between inherent alignment and retrieval performance.

6 Conclusion

In this paper, we propose, for the first time, that inherent alignment exists in unimodal graph and language encoders and further verify the existence of structural alignment. Additionally, we introduce a novel U2A benchmark to facilitate further research.

Force-aligned graph-language models, such as MoleculeSTM, MoMu, and MolFM, widely use SciBERT as the language encoder and the Graph Isomorphism Network (GIN) as the molecular encoder. Although these models select encoders based on downstream tasks rather than inherent alignment, their choices align with our findings. Better-aligned encodings accelerate forced-alignment training, therefore our benchmark and method can be used to decide on encoder pairs for graph-language modelling when more powerful unimodal encoders are released in the future.

We believe this research opens new directions for the alignment of molecular graph-language models, and our findings should also apply to the graph-LM interface in other domains.

7 Limitations

As both graph and language are encoded implicitly, there is no transparent way to determine what is aligned and what is not. Specifically, in structural-level alignment, while the CKA score allows us to identify structural changes, the encoding process remains unknown.

Despite these promising results, our study focuses on the inherent alignment stage. Therefore, more fine-grained alignment tasks involving unimodal graph and language encoders, such as molecule retrieval and molecule captioning, remain unexplored and will be the focus of our future work.

Moreover, the metrics for graph similarity and molecular function similarity measurement are limited, particularly since molecular function similarity based on the Jaccard similarity primarily considers common functional groups while overlooking other functional elements.

Additionally, our study focuses on small molecules due to data limitations. In our dataset, the mean length of molecule SMILES strings is 36, with the maximum length being 119. This constraint may affect the generalizability of our findings to larger or more complex molecules.

8 Ethics Statement

We do not foresee any particular ethical concerns with our study, which explores existing models and is unlikely to lead to unforeseen uses of those models. The base dataset, ChEBI 20, is publicly available and was introduced in the Text2Mol paper (Edwards et al., 2021). RDKit, which is available under the BSD license (Landrum et al., 2025), and MAYGEN (Yirik et al., 2021) are also publicly available. Additionally, the encoders used in our study are open-sourced from various platforms.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Gayane Chilingaryan, Hovhannes Tamoyan, Ani Tevosyan, Nelly Babayan, Karen Hambardzumyan, Zaven Navoyan, Armen Aghajanyan, Hrant Khachatrian, and Lusine Khondkaryan. 2024. Bartsmiles: Generative masked language models for molecular representations. *Journal of Chemical Information and Modeling*, 64(15):5832–5843.

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

Neo Christopher Chung, BłaŻej Miasojedow, Michał Startek, and Anna Gambin. 2019. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC bioinformatics*, 20(Suppl 15):644.

Adrián Csiszárik, Péter Kőrösi-Szabó, Ákos K. Matszangosz, Gergely Papp, and Dániel Varga. 2021. Similarity and matching of neural network representations. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*.

Thomas Gärtner, Peter Flach, and Stefan Wrobel. 2003. On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 129–143. Springer.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: datasets for machine learning on graphs. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Weihua Hu*, Bowen Liu*, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2020. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*.

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

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, Peter Gedeck, Gareth Jones, NadineSchneider, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, tadhurst cdd, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Vincent F. Scalfani, Rachel Walker, Kazuya Ujihara, Daniel Probst, Juuso Lehtivarjo, Hussein Faara, guillaume godin, Axel Pahl, and Jeremy Monat. 2025. rdkit/rdkit: 2024_09_5 (q3 2024) release.

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

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

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. 2018. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical science*, 9(24):5441–5451.

Alex Graeme Murphy, Joel Zylberberg, and Alona Fyshe. 2024. Correcting biased centered kernel alignment measures in biological and artificial neural networks. In *ICLR 2024 Workshop on Representational Alignment*.

National Center for Biotechnology Information. 2025. Pubchem compound summary for cid 525, malic acid. Accessed: Feb. 10, 2025.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. 2023. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *International Conference on Machine Learning*, pages 30458–30490. PMLR.

Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. 2000. Impact of similarity measures on web-page clustering. In *Workshop on artificial intelligence for web search (AAAI 2000)*, volume 58, page 64.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *Preprint*, arXiv:2209.05481.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436.

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

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

David Weininger. 1988. SMILES, a chemical language and information system. *Journal of chemical information and computer sciences*, 28(1):31–36.

Annie M Westerlund, Siva Manohar Koki, Supriya Kancharla, Alessandro Tibo, Lakshidaa Saigiridharan, Mikhail Kabeshov, Rocío Mercado, and Samuel Genheden. 2024. Do Chemformers dream of organic matter? Evaluating a transformer model for multistep retrosynthesis. *Journal of Chemical Information and Modeling*, 64(8):3021–3033.

Richard C Wilson and Edwin R Hancock. 1997. Structural matching by discrete relaxation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):634–648.

Richard C Wilson and Ping Zhu. 2008. A study of graph spectra for comparing graphs and trees. *Pattern Recognition*, 41(9):2833–2841.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. 2023. Mole-BERT: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*.

Jianping Xiao, Li Yang, and Shuqun Wang. 2024. Graph isomorphism network for materials property prediction along with explainability analysis. *Computational Materials Science*, 233:112619.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks? In *International Conference on Learning Representations*.

Mehmet Aziz Yirik, Maria Sorokina, and Christoph Steinbeck. 2021. MAYGEN: an open-source chemical structure generator for constitutional isomers based on the orderly generation principle. *Journal of Cheminformatics*, 13:1–14.

Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. 2022. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016.

Xiao-Chen Zhang, Cheng-Kun Wu, Jia-Cai Yi, Xiang-Xiang Zeng, Can-Qun Yang, Ai-Ping Lu, Ting-Jun Hou, and Dong-Sheng Cao. 2022. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning BERT enhanced by SMILES enumeration. *Research*, 2022:0004.

A Example: Malic Acid Molecule

This is an example of a molecule, malic acid, shown in different formats, including a graph, a language description, SMILES, and a chemical formula (Table 7).

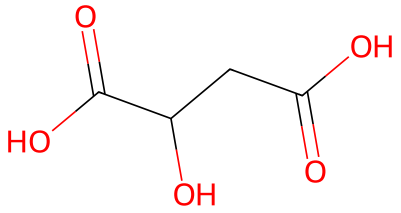
Name: Malic Acid
Molecular Formula: $C_4H_6O_5$
Molecule Graph: The structure shows nodes (carbon implied at vertices) connected by single and double bonds, with two carboxyl (-COOH) groups and a hydroxyl (-OH) group labeled in red. 
Language Description: The molecule is a 2-hydroxydicarboxylic acid that is succinic acid in which one of the hydrogens is replaced by a hydroxyl group. It has a role as a food acidity regulator and a fundamental metabolite. It is a 2-hydroxydicarboxylic acid and a C4-dicarboxylic acid. It derives from succinic acid and is a conjugate acid of malate(2-) and malate.
SMILES: C(C(C(=O)O)O)C(=O)O

Table 7: An example of a malic acid molecule with different modalities.

B Prompt

Figure 3 is the prompt for extracting chemistry-relevant terms.

C Control Groups

After extracting chemical terms and collecting isomers and tautomers, we can pair our data into six groups with different combinations (see Table 8). We use these six groups to gauge inherent alignment under different controlled conditions.

(Instrution) Process the following molecular description text to extract chemistry-related keywords. Please do this carefully, ensuring only chemistry-related words are retained while keeping as many chemistry-relevant words as possible. Extract the chemical terms one by one from the entire sentence.

(Example) Here is an example: original text: The molecule is a 2-hydroxydicarboxylic acid that is succinic acid in which one of the hydrogens attached to a carbon is replaced by a hydroxy group. It has a role as a food acidity regulator and a fundamental metabolite. It is a 2-hydroxydicarboxylic acid and a C4-dicarboxylic acid. It derives from a succinic acid. It is a conjugate acid of a malate(2-) and a malate.

(Answer of example) Extracted text: 2-hydroxydicarboxylic acid, succinic acid, hydroxy group, a food acidity regulator and a fundamental metabolite, a 2-hydroxydicarboxylic acid and a C4-dicarboxylic acid, conjugate acid, malate(2-), malate.

(prompt final answer) Please extract terms from this text and return the results only: { **Language Description**}

Figure 3: Example prompt for extracting the chemical terms

Corpus	Molecule
Language description	original molecules
Language description	tautomers
Language description	isomers
Chemical extraction	original molecules
Chemical extraction	tautomers
Chemical extraction	isomers

Table 8: Control groups used in the experiments.

D Example: Data Details

The example of the data obtained after processing is shown in Table 9. It presents the language description, chemical extraction, tautomers, and isomers, respectively.

Category	Details
Language description	The molecule is a 2-hydroxydicarboxylic acid that is succinic acid in which one of the hydrogens attached to a carbon is replaced by a hydroxyl group. It has a role as a food acidity regulator and a fundamental metabolite. It is a 2-hydroxydicarboxylic acid and a C4-dicarboxylic acid. It derives from a succinic acid. It is a conjugate acid of a malate(2-) and a malate.
Extraction	2-hydroxydicarboxylic acid, succinic acid, hydrogen, carbon, hydroxyl group, food acidity regulator, fundamental metabolite, 2-hydroxydicarboxylic acid, C4-dicarboxylic acid, conjugate acid, malate(2-), malate.
Original molecule	`C(C(C(=O)O)O)C(=O)O`

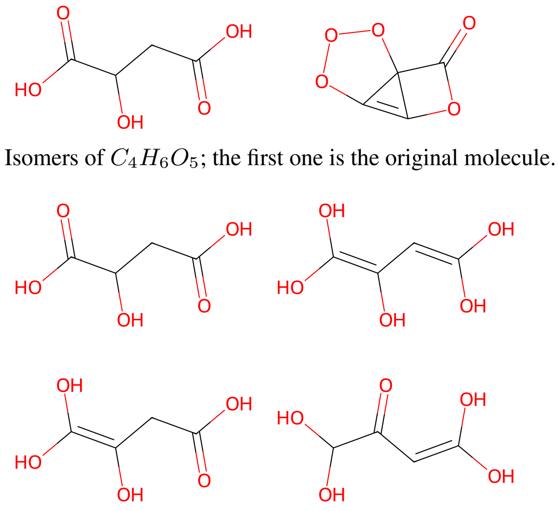
Tautomers	`OC(O)=CC(O)=C(O)O`
	`O=C(O)CC(O)=C(O)O`
	`O=C(C=C(O)O)C(O)O`

Isomers	`O=C1OC2=C3OOOC132`

Table 9: Example of data details.

E Structural Visualization of $C_4H_6O_5$ and Its Isomers and Tautomers

We show the isomers and tautomers in Figure 4. Tautomers exhibit a structure similar to the original molecule, whereas isomers do not.



(a) Isomers of $C_4H_6O_5$; the first one is the original molecule.

(b) Tautomers of $C_4H_6O_5$; the first one is the original molecule.

Figure 4: Structure visualization of $C_4H_6O_5$ and its isomers and tautomers.

F Benchmark and Experimental Environment

Our benchmark and experimental environment is as follows:

- Python 3.9
- GPUs: Nvidia A100
- Operating system: CentOS 7
- Platform: Snellius

G Upper- and Lower- Bound Setup

The experimental setup is shown in Table 10. It presents the upper-bound model along with its counterpart and six control groups.

H Inherent Alignment Setup

The setup can be seen in Table 11. It includes two graph models (MolCLR and MoLFormer), five language models (BERT, SciBERT, SentenceBERT, ModernBERT, and GPT), and two control groups (language description with original molecules and chemical extraction with original molecules).

Experiment	Model	Control Group
Upper-bound	MolFM	Language description with original molecules Language description with tautomers Language description with isomers Chemical extraction with original molecules Chemical extraction with tautomers Chemical extraction with isomers
Lower-bound	Random Initial	Language description with original molecules Language description with tautomers Language description with isomers Chemical extraction with original molecules Chemical extraction with tautomers Chemical extraction with isomers

Table 10: Experimental setup of upper- and lower-bound with MolFM, random initial encoder, and control groups.

Graph Encoder	Language Encoder	Control Group
MolCLR MolFormer	BERT SciBERT SentenceBERT ModernBERT GPT	Language description with original molecules Chemical extraction with original molecules

Table 11: Experimental setup of inherent alignment with molecule graph encoders, language encoders, and control groups.

I Structural Alignment Setup

The setup can be seen in Table 12. It includes two graph models (MolCLR and MoLFormer), five language models (BERT, SciBERT, SentenceBERT, ModernBERT, and GPT), and six control groups (Language description with original molecules , Language description with tautomers , Language description with isomers , Chemical extraction with original molecules , Chemical extraction with tautomers , Chemical extraction with isomers).

Graph Encoder	Language Encoder	Control Group
MolCLR MolFormer	BERT SciBERT SentenceBERT ModernBERT GPT	Language description with original molecules Language description with tautomers Language description with isomers Chemical extraction with original molecules Chemical extraction with tautomers Chemical extraction with isomers

Table 12: Experimental setup of structural alignment with molecule graph encoders, language encoders, and control groups.

J Encoder Pairings and Configurations

We present the different embedding sizes of various encoders in Table 13.

Molecular Model	NLP Model	Embedding
MolFormer (768)	ModernBERT (768)	CLS token Mean pooling
MolFormer (768)	BERT (768)	CLS token Mean pooling
MolFormer (768)	SciBERT (768)	CLS token Mean pooling
MolFormer (768)	Sentence-BERT (768)	Mean pooling
MolFormer (768)	GPT-2 (768)	Mean pooling Last word token Summarized prompt
MolCLR (512)	ModernBERT (768)	CLS token Mean pooling
MolCLR (512)	BERT (768)	CLS token Mean pooling
MolCLR (512)	SciBERT (768)	CLS token Mean pooling
MolCLR (512)	Sentence-BERT (512)	Mean pooling
MolCLR (512)	GPT-2 (768)	Mean pooling Last word token Summarized prompt

Table 13: Encoder Pairings and Configurations. The number in brackets represents the output size from different encoders.

K Comparison of CKA and Debiased CKA

Table 14 presents the mean and standard deviation of CKA and Debiased CKA scores, calculated from ten independent random samples.

		Original molecule	Isomers	Tautomers
MolFM	CKA	0.499 ± 0.007	0.024 ± 0.001	0.436 ± 0.006
	Debiased CKA	0.496 ± 0.006	0.021 ± 0.001	0.430 ± 0.006
Random Initial	CKA	0.113 ± 0.001	0.113 ± 0.001	0.113 ± 0.001
	Debiased CKA	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.000

Table 14: Comparison of CKA and Debiased CKA. This value represents the mean ± standard deviation.