Molecular Identification and Peak Assignment: Leveraging Multi-Level Multimodal Alignment on NMR

Anonymous Authors[1]

Abstract

Nuclear magnetic resonance (NMR) spectroscopy plays an essential role in deciphering molecular structure and dynamic behaviors. While AI-enhanced NMR prediction models hold promise, challenges still persist in tasks such as molecular retrieval, isomer recognition, and peak assignment. In response, this paper introduces a novel solution, Multi-Level Multimodal Alignment with Knowledge-Guided Instance-Wise Discrimination (K-M3AID), which establishes correspondences between two heterogeneous modalities: molecular graphs and NMR spectra. K-M3AID employs a dual-coordinated contrastive learning architecture with three key modules: a graph-level alignment module, a node-level alignment module, and a communication channel. Notably, K-M3AID introduces knowledge-guided instance-wise discrimination into contrastive learning within the node-level alignment module. In addition, K-M3AID demonstrates that skills acquired during node-level alignment have a positive impact on graph-level alignment, acknowledging meta-learning as an inherent property. Empirical validation underscores the effectiveness of K-M3AID in multiple zero-shot tasks.

1. Introduction

Nuclear magnetic resonance (NMR) spectroscopy has found broad applications in various scientific domains, such as chemistry, environmental science, food science, material science, and pharmaceuticals, by providing insights into molecular dynamics and structures (Gunther & Gunther, 1994; Claridge, 2016; Yu et al., 2021). The details of NMR spectra can be influenced by through-bond and through-space interactions, serving as ”fingerprints” to deduce atomic connectivity, relative stereochemistry, and conformations. The conventional approach for elucidating molecular structures and attributing peaks has long relied on manual determination by organic chemists (Guan et al., 2021). However, the interpretation of NMR spectra is not straightforward, particularly when dealing with isomers and complex molecules consisting of multiple stereogenic (chiral) centers (Wu et al., 2023; Chhetri et al., 2018). Even an expert chemist may encounter significant difficulties in accurately assigning isomeric compounds with extremely similar NMR spectra due to this complexity (Nicolaou & Snyder, 2005).

While recent AI-enhanced NMR spectrum prediction models show promise in generating spectra from candidate structures (Chen et al., 2020; Jonas et al., 2022; Kuhn, 2022), these models still face challenges in peak assignment due to their high error tolerance and a lack of precise point-to-point guidance. Since peak assignment is a determining step in isomer recognition, these models fall short in achieving accurate isomer recognition. Another contributing factor is the absence of quantitative ranking for candidate isomers in their implementation. In addition, the success of these models requires a good level of prior knowledge of molecular structures to construct candidates. However, real-world practice often demands spectral interpretation before detailed structural information is available. For instance, when identifying an unknown compound from a plant, there is limited or no knowledge of this compound. Thus, the interpretation of spectra should transition from spectral data to structural elucidation. Therefore, it is imperative to utilize advanced AI methodologies to simplify NMR spectral interpretation, particularly in tasks such as molecular retrieval, candidate ranking, and peak assignment (see Figure 1.a).

In the realm of data representations for NMR interpretation, two heterogeneous modalities come into play: NMR spectrum and molecular graph. A NMR spectrum is a sequence-based chemical modality that captures molecular structural and electronic details through an NMR spectrometer, translating such information into NMR peaks. A molecular

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. under review at ICML 2024 AI for Science workshop. Do not distribute.

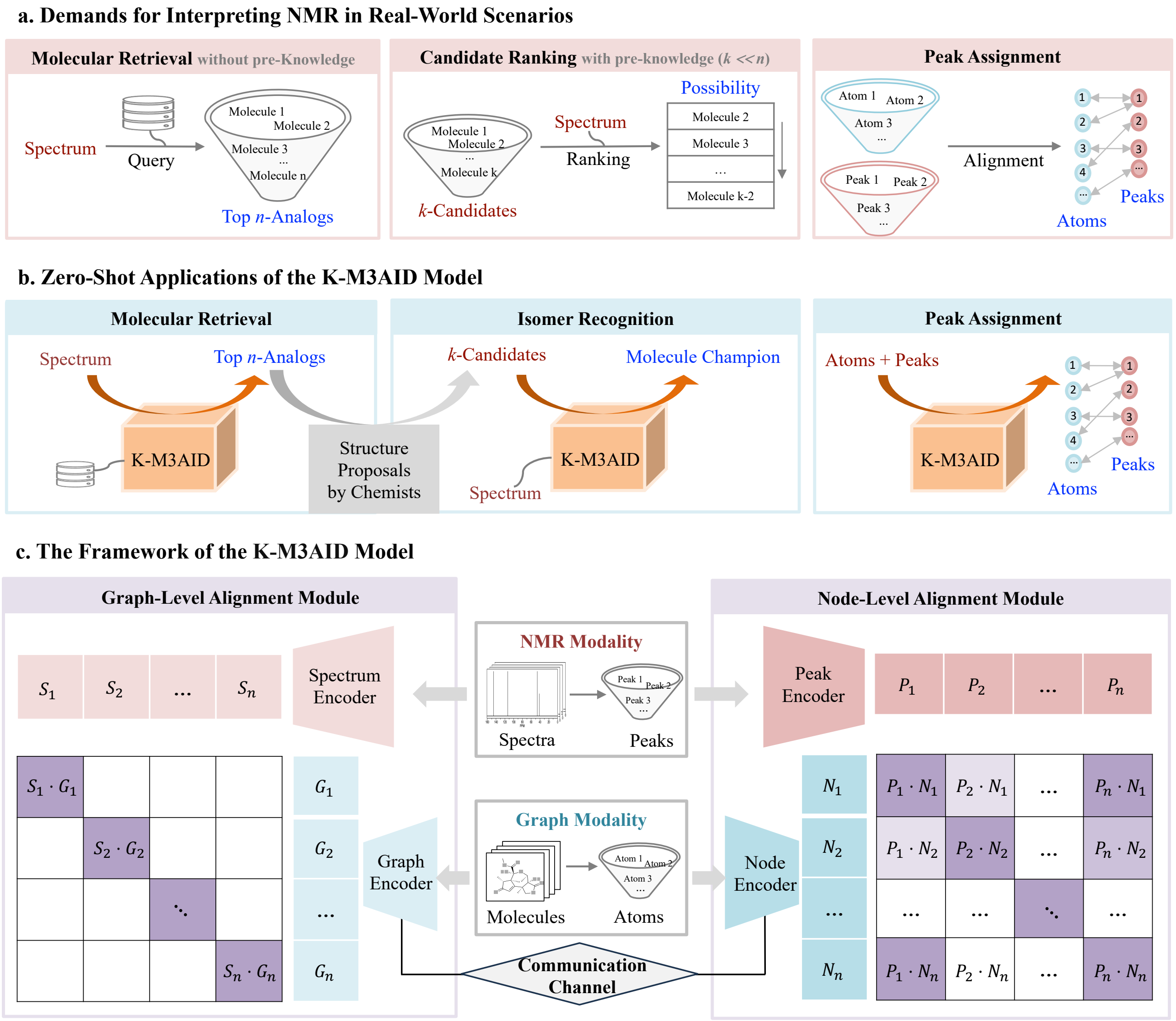

Figure 1: a. Demands for interpreting NMR spectra in real-world scenarios: molecular retrieval, candidate ranking, and peak assignment; b. Zero-shot applications of the K-M3AID model: molecular retrieval, isomer recognition, and peak assignment; c. The framework of K-M3AID model: the molecular alignment module is responsible for optimizing the the correspondence between modalities at the molecular level, the atomic alignment module focus on the fine-tuning of atomic positioning on the spectrum, and the communication channel dynamically adjusts the flow of gradients between node encoder and graph encoder during the training process. S for spectrum embedding, G for graph embedding, P for peak embedding and N for node embedding.

graph encapsulates molecular structural and electronic information through the arrangement of nodes and edges, along with their respective attributes. The analysis of molecular structure and peak assignment requires clear correspondence across these two heterogeneous modalities, which can be formulated as a multimodal alignment problem.

Molecules are distinguished by the distinctive configuration of atoms coupled with bonding patterns, giving rise to distinct spectra. As molecular diversity is extensive, it is impractical to include all molecules and their spectra in a training dataset. However, the corresponding atomic diversity is comparatively constrained. In the context of the multi-view nature of molecules, it is a sound approach to analyze molecular structures by interpreting spectra at the atomic level. Thus, this task can be formulated as a meta-learning problem, which is a branch of metacognition concerned with understanding one's own learning and learning processes.

In light of these challenges and opportunities, we propose a novel framework, K-M3AID (Multi-Level Multimodal Alignment with Knowledge-Guided Instance-Wise Discrimination), aiming to achieve reliable analog retrieval, candidate ranking, and peak assignment in the interpretation of NMR spectra (see Figure 1.b). The overview of our K-M3AID framework features a dual-coordinated contrastive learning architecture, comprising three key components: a graph-level alignment module, a node-level alignment module, and a communication channel. The graph-level alignment module establishes correspondences between molecules and their individual ^{13}C NMR spectra. Given that each unique molecule produces a distinct spectral signature, this module employs a straightforward cross-entropy loss for effective contrastive learning. The node-level alignment module aligns each Carbon atom within the molecules with their signal peaks on the spectrum. Unlike the diverse and distinctive molecular spectral signatures, many atoms exhibit chemical symmetry and magnetic equivalence within the same molecule, corresponding to the same peaks. However, atoms with different local surroundings can still present significant similarity on the spectrum, introducing a heightened level of complexity. To address these complex scenarios, we introduce knowledge-guided instance-wise discrimination based on contrastive learning in the node-level alignment module (see Figure 2). The communication channel dynamically adjusts the flow of gradients between the node encoder and the graph encoder from two modules during the training process.

In summary, our contribution encompasses three significant aspects: ***Conceptually:*** We integrate cross-modal alignment at two architectural levels, namely graph and node levels, within the K-M3AID framework. This integration facilitates rapid adaptation, significantly boosting the efficiency of learning for zero-shot tasks. ***Methodologically:*** We introduce knowledge-guided instance-wise discrimination for cross-modal contrastive learning, leveraging continuous and domain-specific features with inherent natural ordering. To the best of our knowledge, this is the first demonstration of knowledge-guided instance-wise discrimination-based cross-modal contrastive learning, transforming discrete comparisons into a continuous paradigm. ***Empirically:*** We substantiate the effectiveness of K-M3AID through its successful application to various zero-shot tasks, including molecular retrieval, isomer recognition, and peak assignment.

2. Preliminaries

Multimodal Alignment: Multimodal alignment, as defined in the literature (Baltrusaitis et al., 2017), involves establishing relationships and correspondences among subcomponents of instances from two or more modalities. A typical example is identifying specific regions in an image that correspond to words or phrases in a given caption (Karpathy & Fei-Fei, 2015). This approach offers numerous benefits, including enhanced data interpretation, heightened accuracy and robustness, overcoming limitations of single-modal systems, and better addressing real-world complexity (Baltrusaitis et al., 2017), (Summaira et al., 2021), (Akkus et al., 2023). CLIP (Contrastive Language-Image Pretraining) (Radford et al., 2021; Li et al., 2021) is one of the most widely adopted frameworks for multimodal alignment. As highlighted in the introduction, molecular information originates from diverse sources such as molecule graphs and NMR spectroscopy. To leverage effective alignments of this multifaceted information across different modalities, we adopt the CLIP framework with graph neural networks (GNN) (Xu et al., 2018), (Wu et al., 2022) to encode molecular information and neural network encoders (Serra et al., 2018) to encapsulate NMR information.

Meta-Learning: Meta-learning is defined as the process of learning how to learn across tasks (Vilalta & Drissi, 2002). More specifically, it leverages skills previously acquired from related tasks to the current one (Lake et al., 2017). With more skills learned, acquiring new ones becomes easier, requiring fewer examples and less trial-and-error (Vanschoren, 2018; Finn et al., 2017). A meta-learner is trained on a diverse set of object recognition tasks. During this training, it learns common features, patterns, and strategies for recognizing objects. Once trained, when presented with a new, previously unseen object category, the meta-learner can rapidly adapt and achieve high recognition accuracy, leveraging the knowledge acquired from the diverse training tasks to perform in this novel recognition task (Finn et al., 2017). A profound understanding of atom properties allows us to extend our vision to previously unseen molecules, aligning with the principles of meta-learning in artificial intelligence.

Contrastive Learning: Contrastive learning focuses on discerning similarities and differences between items (Le-Khac et al., 2020b; Jaiswal et al., 2021; Liu et al., 2021). A fundamental aspect of this process involves instance-wise discrimination (Wu et al., 2018). Models incorporating instance-wise discrimination not only foster an understanding of the inherent data structure but also enhance generalization capabilities. This is attributed to the contrastive learning approach, which prioritizes distinguishing between instances rather than memorizing specific labeled examples. In chemistry, each molecule/atom is treated as a distinct instance, and the learning algorithm focuses on distinguishing each molecule/atom based on its context.

3. Our Method

In this section, we firstly introduce Knowledge-Guided Instance-Wise Discrimination. Then, we present the architecture of the K-M3AID framework, an end-to-end system designed for multi-level multimodal alignment, along with its loss function.

3.1. Knowledge-Guided Instance-Wise Discrimination Contrastive Learning

Knowledge Span, which we define as a continuous and domain-specific feature, exhibits natural ordering and is able to offer guidance for contrastive learning. As such, we introduce a novel approach into contrastive learning, termed Knowledge-Guided Instance-Wise Discrimination (see Figure 2). This approach expands the scope of contrastive learning from confined comparisons (predetermined negative and positive pairs) to unrestricted comparisons (no need for pre-determination). This extension removes the necessity of explicitly defining such pairs, thus mitigating the potential introduction of human bias.

Suppose $\mathcal{M}$ is the set of instances. $\mathcal{A} \subset \mathbb{R}^{d_1}$ is the set of tunable instances' embeddings in modality A, $\mathcal{B} \subset \mathbb{R}^{d_1}$ is the set of tunable instances' embeddings in modality B, and $\mathcal{K} \subset \mathbb{R}^{d_2}$ is the corresponding fixed knowledge span label that can guide the relative distance learning between components in $\mathcal{A}$ and $\mathcal{B}$. Thus, the size of $\mathcal{A}$, $\mathcal{B}$, $\mathcal{K}$ are $|\mathcal{M}|$, respectively.

Let $\mathcal{A}_i$ be the i^{th} instance embedding of $\mathcal{A}$, and $\mathcal{B}_j$ be the j^{th} instance embedding of $\mathcal{B}$. We define the distance function between $\mathcal{A}_i$ and $\mathcal{B}_j$ as $d_E(\mathcal{A}_i, \mathcal{B}_j) = \mathcal{A}_i \cdot \mathcal{B}_j \rightarrow \mathbb{R}^+$, and calibration function $d(\mathcal{K}_i, \mathcal{K}_j) \rightarrow \mathbb{R}^+$ with a monotonic property and constraint $\sum_{j=1}^{|\mathcal{M}|} d(\mathcal{K}_i, \mathcal{K}_j) = 1$, in which $\mathcal{K}_i$ and $\mathcal{K}_j$ serve as the designated Knowledge Span Label. We introduce the Knowledge Span Guided Loss (KSGL) as follows:

$$KSGL(i) = -\sum_{1\leq j\leq|\mathcal{M}|} d(\mathcal{K}_i, \mathcal{K}_j) \log \frac{e^{d_E(\mathcal{A}_i, \mathcal{B}_j)}}{\sum_{1\leq k\leq|\mathcal{M}|} e^{d_E(\mathcal{A}_i, \mathcal{B}_k)}} \tag{1}$$

$$= -\sum_{1\leq j\leq|\mathcal{M}|} d(\mathcal{K}_i, \mathcal{K}_j) \log(\text{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j))) \tag{2}$$

In particular, when it reaches ideal optimum, $d(\mathcal{K}_i, \mathcal{K}_j)$ and $d_E(\mathcal{A}_i, \mathcal{B}_j)$ reaches the following relation:

$$d(\mathcal{K}_i, \mathcal{K}_j) = \text{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j)) \tag{3}$$

For detail proof, please refer to Appendix A. As a result, the corresponding $CL_{instance}$ is expressed as following:

$$CL_{instance} = \frac{1}{|\mathcal{M}|} \sum_{1\leq i\leq|\mathcal{M}|} KSGL(i) \tag{4}$$

3.2. Architecture & Contrastive Learning Loss

The K-M3AID framework is a dual-CLIP architecture (see Figure 1), comprising three critical components: a graph-level alignment module, a node-level alignment module, and a communication channel. The graph-level alignment module adopts a gradient-asymmetric CLIP mechanism. While two unimodal encoders work in conjunction, only the from-scratch graph encoder (GIN, (Xu et al., 2018)) undergoes dynamic training throughout the process; the pre-trained spectrum encoder (Yang et al., 2021) remains fixed. Both encoders are complemented by dedicated projection layers, facilitating the mapping of embeddings into a joint space. The node-level alignment module adopts a gradient-symmetric CLIP mechanism. It is equipped with two from-scratch unimodal encoders, the node encoder and the peak encoder, as well as their dedicated projection layers. The graph encoder in the graph-level alignment module shares part of the weights with the node encoder in the node-level alignment module, serving as the communication channel.

The synergy between these two modules is pivotal, collectively contributing to the loss function, expressed as

$$L = CL_{graph} + CL_{node}, \tag{5}$$

where CL_{graph} represents the contrastive learning loss in the graph-level alignment module by Equation 7, and CL_{node} represents the contrastive learning loss in the node-level alignment module by Equation 11.

Let i denote the i^{th} instance, and j denote the j^{th} instance. Then x_i denotes the raw input in modality A for the i^{th} instance and y_j denotes the raw input in modality B for the j^{th} instance. Suppose $f_x(\cdot)$ represent the encoding function for modality A, and $f_y(\cdot)$ denote the encoding function for modality B. In graph-level alignment module, these two encoding functions, should map x_i and y_j to a proximate location in the joint embedding (inter-modality) if $i = j$.

$$CL_{graph}(i) = -\log \frac{e^{\delta(x_i, y_i)}}{\sum_{1\leq j\leq N} e^{\delta(x_i, y_j)}} \tag{6}$$

$$= -\log(\text{softmax}(\delta(x_i, y_i))) \tag{7}$$

Where $\delta(x_i, y_j) = \left(f_x(x_i)^T \cdot f_y(y_j)\right)$, N is the total number of instances from the current batch.

Thus, the total CL_{graph} is expressed as following:

$$CL_{graph} = \frac{1}{N} \sum_{1\leq i\leq N} CL_{graph}(i) \tag{8}$$

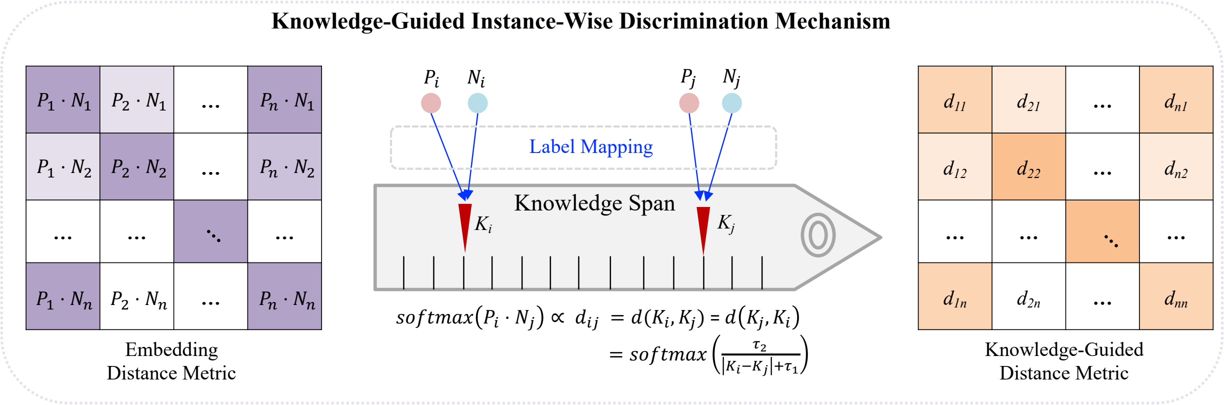
Knowledge-Guided Instance-Wise Discrimination Mechanism
$P_1 \cdot N_1$ $P_2 \cdot N_1$ … $P_n \cdot N_1$
$P_1 \cdot N_2$ $P_2 \cdot N_2$ … $P_n \cdot N_2$
… … ⋱ …
$P_1 \cdot N_n$ $P_2 \cdot N_n$ … $P_n \cdot N_n$
Embedding Distance Metric
P_i N_i P_j N_j
Label Mapping
Knowledge Span
K_i K_j
$softmax(P_i \cdot N_j) \propto d_{ij} = d(K_i, K_j) = d(K_j, K_i) = softmax\left(\frac{\tau_2}{|K_i - K_j| + \tau_1}\right)$
d_{11} d_{21} … d_{n1}
d_{12} d_{22} … d_{n2}
… … ⋱ …
d_{1n} d_{2n} … d_{nn}
Knowledge-Guided Distance Metric

Figure 2: Knowledge-Guided Instance-Wise Discrimination Mechanism. K_i and K_j represent the corresponding knowledge span labels for i^{th} and j^{th} items.

This design for the loss aims to match the same instance cross different modalities.

4. Experiments

To thoroughly evaluate the performance of K-M3AID, we compare it with other baselines across various zero-shot downstream tasks, including molecular retrieval, isomer recognition, and peak assignment. Please refer to the detailed settings of pre-training and downstream tasks in Appendix B.

4.1. Chosen Knowledge Span-ppm

^{13}C NMR uncovers molecular structures by providing the chemical environments of carbon atoms and their magnetic responses to external fields, quantifying these features in parts per million (ppm) relative to a reference compound like tetramethylsilane (TMS), simplifying comparisons across experiments. Thus, continuous peak positions, measured in ppm, can serve as a robust knowledge span to facilitate instance-wise discrimination for this contrastive learning task.

For the node-level alignment module, $\mathcal{A}$ is the set of node embeddings for Carbon atoms in the molecular graph modality, and $\mathcal{B}$ is the set of peak embeddings for respective Carbon atoms in the NMR modality. $\mathcal{K}$ is the set of ppm values for each corresponding Carbon atom in $\mathcal{A}$ and $\mathcal{B}$. Suppose ppm_i is the ppm for the i^{th} Carbon Atom, and ppm_j is the corresponding ppm for the j^{th} peak. $d(\cdot, \cdot)$ is then defined as follows:

$$d(\mathcal{K}_i, \mathcal{K}_j) = d(ppm_i, ppm_j) \tag{9}$$

$$= softmax(\frac{\tau_2}{|ppm_i - ppm_j| + \tau_1}) \tag{10}$$

where τ_1 and τ_2 are temperature hyper-parameter. For further discussion of selection about τ_1 and τ_2, please refer to Appendix C.2. Then, the final form of contrastive loss for node-level alignment according to Equation 2 and Equation 4 is as following:

$$CL_{node} = -\frac{1}{|\mathcal{M}|}\sum_{j=1}^{|\mathcal{M}|} d(\mathcal{K}_i, \mathcal{K}_j) \cdot log\frac{e^{d_E(\mathcal{A}_i, \mathcal{B}_j)}}{\sum_{k=1}^{|\mathcal{M}|} e^{d_E(\mathcal{A}_i, \mathcal{B}_k)}} \tag{11}$$

Here, i and j are indices of atoms. $\mathcal{A}_i \in \mathcal{A}$ represents the embedding of $i - th$ atom in modality A while $\mathcal{B}_i \in \mathcal{B}$ represents the embedding of $i - th$ atom in modality B, $\mathcal{K}_j \in \mathcal{K}$ represents peaks, and th is the abbreviation for the threshold.

4.2. Baselines

No Communication: In contrast to the communicative mechanism of K-M3AID, one of the baselines is established without the utilization of a communication channel (denoted as No Comm.).

Strong-Pair-based Instance-Wise Discrimination: We explore an alternative baseline where the knowledge-guided instance-wise discrimination in the node-level alignment module is replaced with strong-pair-based instance-wise discrimination (denoted as SP-ID). SP-ID enforces a precise match in node-level (atom-peak) alignment, ensuring that only correct pairs established during the training process are considered. The mathematical definition of a strong pair is as follows:

$$\text{Strong Pair: } |ppm_i - ppm_j| = 0, \tag{12}$$

where i, j represent the indices of nodes (atoms).

Weak-Pair-based Instance-Wise Discrimination: We replace the knowledge-guided instance-wise discrimination with weak-pair-based instance-wise discrimination (denoted as WP-ID) in the node-level alignment module. WP-ID broadens the matching criteria of SP-ID, allowing for multiple matches within a specified threshold set for the distance of their corresponding parts per million (ppm, referenced in Section 4.1). The mathematical definition of a weak pair is as follows:

$$\textit{Weak Pair: } |ppm_i - ppm_j| \leq th, \tag{13}$$

where i, j represent the indices of nodes (atoms).

4.3. Results

4.3.1. Validation Performance

The K-M3AID model showcases an impressive validation accuracy of 95.5% in aligning molecules with spectra within the graph-level alignment module. In direct comparison, K-M3AID outperforms alternative models such as SP-ID, WP-ID, and the model without a communication mechanism (No Comm.), demonstrating superior performance with a margin ranging from approximately 1% to 6% in the graph-level alignment module (refer to Table 1). Notably, SP-ID significantly outperforms WP-ID, and as the matching criteria threshold widens, the performance of the latter deteriorates.

In the context of peak-atom alignment within the node-level alignment module, K-M3AID and the model without a communication mechanism exhibit comparable accuracies (refer to Table 1). However, K-M3AID showcases slightly better stability across 5-fold cross-validation. Moreover, K-M3AID demonstrates superiority in peak-atom alignment when compared to SP-ID and WP-ID. This superiority may arise from the inherent limitations of both strong and weak pair definitions, which fail to precisely calibrate the diverse relationships among the elements. This finding is further supported by the significant decreases in the accuracy of peak-atom alignment as the threshold of weak pair increases.

4.3.2. Performance on Zero-Shot Molecular Retrieval

We conduct a systematic evaluation of the effectiveness of our K-M3AID model, comparing it with baseline models in the zero-shot molecular retrieval task across datasets of varying magnitudes. Detailed results are presented in Table 2. The K-M3AID model consistently attains an impressive top-1 accuracy of approximately 95.8% in molecular retrieval when the molecular reference library comprises 100 entries. This performance surpasses that of alternative mechanisms such as SP-ID (95.3%), WP-ID (92.9%), and No Comm. (94.8%). As the molecular reference library expands to 1000 entries, the K-M3AID model exhibits notable superiority, achieving accuracy levels of 80.4%, 1.8%, 8.7%, and 2.8% higher than SP-ID, WP-ID, and No Comm. mechanisms, respectively. The advantage of K-M3AID becomes even more pronounced when the library size reaches 10,000 entries. In this scenario, K-M3AID yields 46.3% at top-1 accuracy, showcasing advancements of 10.5%, 13.6%, and 6.2% over SP-ID, WP-ID, and No Comm. mechanisms, respectively. Even with larger molecular reference libraries, such as 100,000 and 1,000,000 entries, K-M3AID consistently outshines SP-ID, WP-ID, and No Comm. mechanisms. These compelling results distinguish the K-M3AID model as an exceptional choice in scenarios demanding robust performance in molecular retrieval tasks.

4.3.3. Performance on Zero-Shot Isomer Recognition

K-M3AID stands out prominently when compared to SP-ID, WP-ID, and no communication approaches in the task of zero-shot isomer recognition, achieving an exceptional 100% accuracy across given groups of isomers (refer to Table 3). These empirical observations underscore the advantages of K-M3AID in the context of isomer recognition. The superiority of K-M3AID over the no communication baseline demonstrates the positive impact of node-level alignment on graph-level alignment, emphasizing the potency of meta-learning.

4.3.4. Performance on Zero-Shot Peak Assignment

The K-M3AID model demonstrates a validation accuracy surpassing 90% for peak assignment (peak-atom alignment) within the node-level alignment module after 200 epochs (see Figure 3.A). Notably, the model achieves a 100% accuracy rate in 74.1% of molecules containing fewer than 10 carbon atoms (see Figure 3.B). For molecules with carbon atom counts ranging from 10 to 20, the model attains 100% accuracy in 37.2% of cases (see Figure 3.C). Furthermore, it achieves an accuracy exceeding 80% in more than 50% of cases pertaining to molecules containing more than 20 carbon atoms (see Figure 3.D). Additionally, to further illustrate the power of K-M3AID on peak assignment, we present two complex natural product molecules featuring multiple rings (4 and 4, respectively) and stereogenic (chiral) centers (6 and 8, respectively) in Figure F.1.

K-M3AID demonstrates superior performance in peak assignment compared to SP-ID and WP-ID. Our case studies reveal that the limitations of SP-ID and WP-ID become particularly pronounced in two scenarios: 1) When local contexts of specific atoms exhibit a high degree of similarity. 2) When certain atoms display symmetric mapping within the same molecule.

In the former scenario, exemplified by molecular A in Fig-

Table 1: Batch-wise validation accuracy (%) of K-M3AID and baselines with $epochs = 200$. For WP-ID, the threshold is configured at 1, 5, and 10 ppm.

Alignment	SP-ID	WP-ID(*th*=1)	WP-ID(*th*=5)	WP-ID(*th*=10)	No Comm.	K-M3AID
Graph-Level	93.5±0.6	91.3±0.8	90.3±0.6	88.4±1.4	94.6±0.4	**95.5±0.4**
Node-Level	89.3±0.4	83.7±0.6	79.8±0.5	66.1±2.5	90.4±0.2	**90.3±0.1**

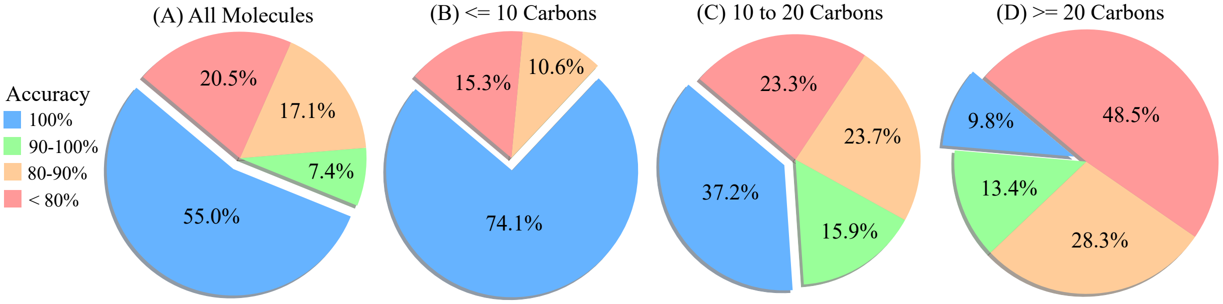

Figure 3: The statistics of zero-shot peak assignment.

Table 2: Zero-shot molecular retrieval at top 1 accuracy across datasets of varying sizes. (For more statistics, such as top 5, top 10 and top 25, please refer to Appendix Table D.1. For similarity comparison between the molecules and the Top 1 neighbor by different retrieval methods, please consult Appendix Table D.2.)

Method	10^2	10^3	10^4	10^5	10^6
K-M3AID	95.8±1.0	80.4±3.9	46.3±1.2	18.0±0.8	5.8±1.7
SP-ID	95.3±0.8	78.6±2.7	35.8±3.8	12.9±1.6	3.4±0.9
WP-ID (*th* = 1)	92.9±0.6	71.7±1.0	32.7±1.3	10.7±0.5	3.6±0.7
No Comm.	94.8±1.2	77.6±1.4	40.1±1.2	14.4±0.9	4.1±1.1

Table 3: Zero-shot isomer recognition accuracy (%) of K-M3AID and baselines.

Formula	#Isomers	SP-ID	WP-ID (th=1)	No Comm.	K-M3AID
C_4H_6O	15	86.7	86.7	86.7	**100.0**
C_9H_9N	15	86.7	80.0	**100.0**	**100.0**
$C_7H_{11}NO_3$	14	78.6	85.7	85.7	**100.0**
$C_6H_{13}NO$	23	91.3	91.3	**100.0**	**100.0**
$C_8H_7NO_4$	13	92.3	84.6	92.3	**100.0**
$C_{15}H_{24}O$	16	93.8	93.8	**100.0**	**100.0**
$C_{11}H_{14}$	10	90.0	80.0	70.0	**100.0**
$C_7H_{15}NO$	14	85.7	85.7	**100.0**	**100.0**
$C_{10}H_{16}O_2$	26	92.3	84.6	**100.0**	**100.0**
$C_8H_{15}N$	11	81.8	90.9	**100.0**	**100.0**

ure 4, atom 0 and atom 4 are secondary carbons (attaching to 2 carbons and 2 hydrogens), nearly symmetric on the same 5-member ring, corresponding to the peak position measured in ppm of 27.0 and 29.8, respectively (for the definition of ppm, please refer to Section 4.1). The similar local content of these two atoms fools SP-ID and WP-ID. Meanwhile, atom 1 and atom 3 are tertiary carbons (attaching to 3 carbons and 1 hydrogen), nearly symmetric on the same 5-member ring, corresponding to the peak position measured in ppm of 54.5 and 44.1, respectively. Only WP-ID fails to distinguish and align them.

In the latter scenario, exemplified by molecular B in Figure 4, there exist instances one-to-one and one-to-many for atomic-level alignment within the molecular configuration. Both SP-ID and WP-ID methods misalign certain atoms with other atoms with small ppm differences (less than 3 ppm in this case), rather than aligning them with themselves or their symmetric counterparts. In contrast, the K-M3AID approach excels in both scenarios by discerning each one of the atoms, which is attributed to the full utilization of ppm difference distance learning. (For additional cases, please refer to Appendix Figure F.2)

5. Related Work

Multimodal Instance-Wise Discrimination: As mentioned in the preliminaries, instance discrimination (Le-Khac et al., 2020a; Zolfaghari et al., 2021; Morgado et al., 2021; Liu et al., 2023), an important part of contrastive learning, distinguishes individual instances without explicit class labels. Transitioning into multimodal contrastive learning, it can be categorized into two general approaches: strong-pair-based (van den Oord et al., 2019; Jaiswal et al., 2021; Liu et al., 2023) and weak-pair-based (Salakhutdinov & Hinton, 2007; Frosst et al., 2019; Liang et al., 2021) instance-wise discrimination. The strong-pair-based approach, such as the Noise Contrastive Estimation (NCE) method, enforces a precise one-to-one correspondence for real samples with artificially generated noise samples. An example of a positive pair can be a noise-added picture of a zebra with the text description of a zebra. Instead of one-to-one correspondences, the weak-pair-based ap-

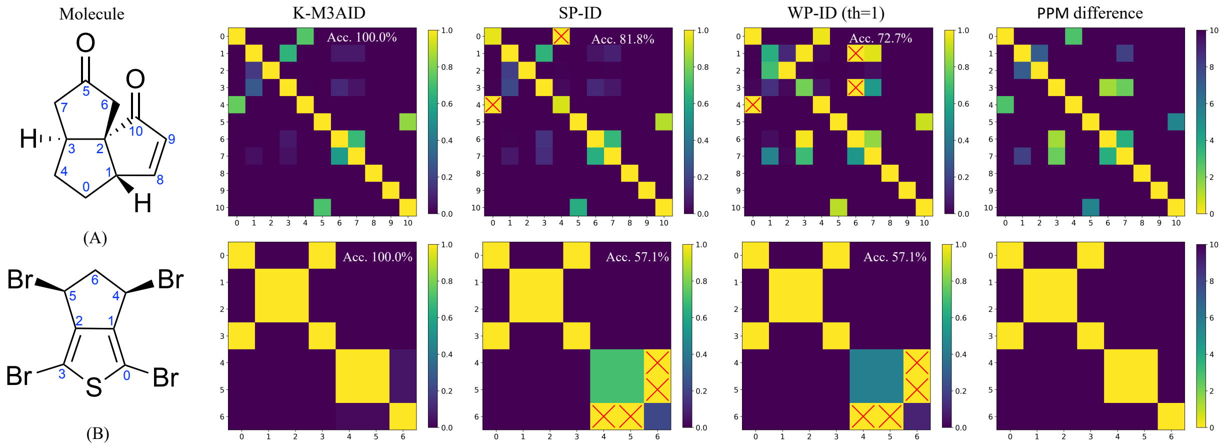

Figure 4: Case study of peak assignment. Yellow cells in PPM difference represent the ground truth alignment, and red cross represents the wrong alignment. For the definition of ppm, please refer to 4.1. For additional cases, please refer to Appendix Figure F.2

proach relaxes the positive pairs to broader semantic correspondences. An example of a positive pair can be a picture of a zebra with the text description of a horse but not with the text description of a tiger.

Multimodal Meta-Alignment: Within the realm of multimodal alignment, multimodal meta-alignment is a novel method for aligning representation spaces using paired cross-modal data with different similarity levels while ensuring quick generalization to new tasks across different modalities (Liang et al., 2021). This approach can be observed at different levels, including the intermediate and fundamental (irreducible) element level. Examples of this method at the intermediate level can be found in research on Cross-Modal Generalization (Chen et al., 2017; Li et al., 2020; Liang et al., 2021; Zhang et al., 2021) and Livestreaming Product Recognition (Yang et al., 2023). While these studies showcase how multimodal meta-alignment operates at a broad objective level, the application of multimodal meta-alignment at the most fundamental element level remains underexplored in current research.

6. Conclusion and Future Work

In this paper, we introduced the K-M3AID (Knowledge-Guided Multi-Level Multimodal Alignment with Instance-Wise Discrimination) framework, incorporating both graph-level and node-level alignment. Its effectiveness was demonstrated through multiple zero-shot tasks, including molecular retrieval, isomer recognition, and peak assignment. The significance of knowledge-guided instance-wise discrimination is underscored through various metrics and case studies. Moreover, the findings from molecular retrieval and isomer recognition highlight the favorable influence of node-level alignment on graph-level alignment. This emphasizes the successful integration of meta-learning within our hierarchical alignment framework. While our framework achieves an atomic-level alignment overall accuracy of 100% for 55% of cases, it drops significantly to 9.8% when handling molecules with more than 20 carbon atoms. Currently, our graph encoder operates on 2D molecular graphs with basic node and edge features. This implementation potentially constrains its ability to generate precise node embeddings for distinguishing atoms in highly complex scenarios. Future developments could benefit from incorporating a 3D-based graph, holding substantial potential to enhance performance in such complex situations.

Accessibility

The code and dataset will be made available upon the date of publication.

References

Akkus, C., Chu, L., Djakovic, V., Jauch-Walser, S., Koch, P., Loss, G., Marquardt, C., Moldovan, M., Sauter, N., Schneider, M., Schulte, R., Urbanczyk, K., Goschenhofer, J., Heumann, C., Hvingelby, R., Schalk, D., and Aßenmacher, M. Multimodal deep learning, 2023.

Baltrusaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy, 2017.

Chen, D., Wang, Z., Guo, D., Orekhov, V., and Qu, X. Review and prospect: deep learning in nuclear magnetic resonance spectroscopy. *Chemistry–A European Journal*, 26(46): 10391–10401, 2020.

Chen, L., Srivastava, S., Duan, Z., and Xu, C. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 349–357, 2017.

Chhetri, B. K., Lavoie, S., Sweeney-Jones, A. M., and Kubanek, J. Recent trends in the structural revision of natural products. *Natural product reports*, 35(6):514–531, 2018.

Claridge, T. D. *High-resolution NMR techniques in organic chemistry*, volume 27. Elsevier, 2016.

Dice, L. R. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Frosst, N., Papernot, N., and Hinton, G. Analyzing and improving representations with the soft nearest neighbor loss, 2019.

Guan, Y., Sowndarya, S. S., Gallegos, L. C., John, P. C. S., and Paton, R. S. Real-time prediction of 1 h and 13 c chemical shifts with dft accuracy using a 3d graph neural network. *Chemical Science*, 12(36):12012–12026, 2021.

Gunther, H. and Gunther, H. *NMR spectroscopy: basic principles, concepts, and applications in chemistry*. John Wiley & Sons Chichester, UK, 1994.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. A survey on contrastive self-supervised learning. *Technologies*, 9(1), 2021. ISSN 2227-7080. doi: 10.3390/technologies9010002. URL https://www.mdpi.com/2227-7080/9/1/2.

Jonas, E., Kuhn, S., and Schlörer, N. Prediction of chemical shift in nmr: A review. *Magnetic Resonance in Chemistry*, 60(11): 1021–1031, 2022.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380, 2023.

Kuhn, S. Applications of machine learning and artificial intelligence in nmr, 2022.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Le-Khac, P. H., Healy, G., and Smeaton, A. F. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020a. doi: 10.1109/ACCESS.2020.3031549.

Le-Khac, P. H., Healy, G., and Smeaton, A. F. Contrastive representation learning: A framework and review. *Ieee Access*, 8: 193907–193934, 2020b.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

Liang, P. P., Wu, P., Ziyin, L., Morency, L.-P., and Salakhutdinov, R. Cross-modal generalization: Learning in low resource modalities via meta-alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475247. URL https://doi.org/10.1145/3474085.3475247.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1): 857–876, 2021.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35 (1):857–876, 2023. doi: 10.1109/TKDE.2021.3090866.

Morgado, P., Vasconcelos, N., and Misra, I. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12475–12486, 2021.

Nicolaou, K. and Snyder, S. A. Chasing molecules that were never there: misassigned natural products and the role of chemical synthesis in modern structure elucidation. *Angewandte Chemie International Edition*, 44(7):1012–1044, 2005.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Russel, J. A. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

Salakhutdinov, R. and Hinton, G. Learning a nonlinear embedding by preserving class neighbourhood structure. In Meila, M. and Shen, X. (eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pp. 412–419, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR. URL https://proceedings.mlr.press/v2/salakhutdinov07a.html.

Salton, G. and McGill, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

Serra, J., Pascual, S., and Karatzoglou, A. Towards a universal neural network encoder for time series. In *CCIA*, pp. 120–129, 2018.

Sokal, R. R. and Sneath, P. H. A. *Principles of Numerical Taxonomy*. W. H. Freeman and Company, 1963.

Steinbeck, C., Krause, S., and Kuhn, S. Nmrshiftdb constructing a free chemical information system with open-source components. *Journal of chemical information and computer sciences*, 43(6):1733–1739, 2003.

Summaira, J., Li, X., Shoib, A. M., Li, S., and Abdul, J. Recent advances and trends in multimodal deep learning: A review, 2021.

Tanimoto, T. T. An elementary mathematical theory of classification and prediction. *Internal Report 17, IBM*, 1957.

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

Vanschoren, J. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

Vilalta, R. and Drissi, Y. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18:77–95, 2002.

Wu, A., Ye, Q., Zhuang, X., Chen, Q., Zhang, J., Wu, J., and Xu, X. Elucidating structures of complex organic compounds using a machine learning model based on the 13c nmr chemical shifts. *Precision Chemistry*, 1(1):57–68, 2023.

Wu, S., Sun, F., Zhang, W., Xie, X., and Cui, B. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5):1–37, 2022.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Yang, W., Chen, Y., Li, Y., Cheng, Y., Liu, X., Chen, Q., and Li, H. Cross-view semantic alignment for livestreaming product recognition, 2023.

Yang, Z., Song, J., Yang, M., Yao, L., Zhang, J., Shi, H., Ji, X., Deng, Y., and Wang, X. Cross-modal retrieval between 13c nmr spectra and structures for compound identification using deep contrastive learning. *Analytical Chemistry*, 93(50): 16947–16955, 2021.

Yu, H.-Y., Myoung, S., and Ahn, S. Recent applications of benchtop nuclear magnetic resonance spectroscopy. *Magnetochemistry*, 7(9):121, 2021.

Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., and Yang, Y. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021.

Zolfaghari, M., Zhu, Y., Gehler, P., and Brox, T. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1450–1459, 2021.

Appendix

A. Revisiting Knowledge Span Guided Loss

Theorem 1 (Knowledge Span Guided Loss)**.** *Suppose $\mathcal{M}$ is the set of instances. $\mathcal{A} \subset \mathbb{R}^{d_1}$ is the set of tunable instances' embeddings in modality A, $\mathcal{B} \subset \mathbb{R}^{d_1}$ is the set of tuable instances' embeddings in modality B, and $\mathcal{K} \subset \mathbb{R}^{d_2}$ is the corresponding fixed knowledge span label that can guide the relative distance learning between components in $\mathcal{A}$ and $\mathcal{B}$. Thus, the size of $\mathcal{A}$, $\mathcal{B}$, $\mathcal{K}$ are $|\mathcal{M}|$, respectively.*

Let $\mathcal{A}_i$ be the i^{th} instance embedding of $\mathcal{A}$, and $\mathcal{B}_j$ be the j^{th} instance embedding of $\mathcal{B}$. We define the distance function between $\mathcal{A}_i$ and $\mathcal{B}_j$ as $d_E(\mathcal{A}_i, \mathcal{B}_j) = \mathcal{A}_i \cdot \mathcal{B}_j \to \mathbb{R}^+$, and calibration function $d(\mathcal{K}_i, \mathcal{K}_j) \to \mathbb{R}^+$ with a monotonic property and constraint $\sum_{j=1}^{|\mathcal{M}|} d(\mathcal{K}_i, \mathcal{K}_j) = 1$, in which $\mathcal{K}_i$ and $\mathcal{K}_j$ serve as the designated Knowledge Span Label. We introduce the Knowledge Span Guided Loss (KSGL) as follows:

$$KSGL(i) = -\sum_{1 \le j \le |\mathcal{M}|} d(\mathcal{K}_i, \mathcal{K}_j) \log \frac{e^{d_E(\mathcal{A}_i, \mathcal{B}_j)}}{\sum\limits_{1 \le k \le |\mathcal{M}|} e^{d_E(\mathcal{A}_i, \mathcal{B}_k)}} \tag{A.1}$$

$$= -\sum_{1 \le j \le |\mathcal{M}|} d(\mathcal{K}_i, \mathcal{K}_j) \log(\textit{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j))) \tag{A.2}$$

Proof. In order to optimize the loss $KSGL(i)$, we need to set the following partial derivative to be 0 for each $d_E(\mathcal{A}_i, \mathcal{B}_j)$ with $1 \le j \le |\mathcal{M}|$. Here are the detail process:

$$\frac{\partial KSGL(i)}{\partial d_E(\mathcal{A}_i, \mathcal{B}_j)} = \frac{\partial}{\partial d_E(\mathcal{A}_i, \mathcal{B}_j)} \underbrace{\left(-d(\mathcal{K}_i, \mathcal{K}_j) \log \frac{e^{d_E(\mathcal{A}_i, \mathcal{B}_j)}}{e^{d_E(\mathcal{A}_i, \mathcal{B}_j)} + \sum_{k \ne j} e^{d_E(\mathcal{A}_i, \mathcal{B}_k)}} \right)}_{\text{When the numerator includes } e^{d_E(\mathcal{A}_i, \mathcal{B}_j)}}$$

$$+ \frac{\partial}{\partial d_E(\mathcal{A}_i, \mathcal{B}_j)} \underbrace{\left(\sum_{k \ne j} -d(\mathcal{K}_i, \mathcal{K}_k) \log \frac{e^{d_E(\mathcal{A}_i, \mathcal{B}_k)}}{e^{d_E(\mathcal{A}_i, \mathcal{B}_j)} + \sum_{k \ne j} e^{d_E(\mathcal{A}_i, \mathcal{B}_k)}} \right)}_{\text{When the numerator does not include } e^{d_E(\mathcal{A}_i, \mathcal{B}_j)}}$$

$$= -(d(\mathcal{K}_i, \mathcal{K}_j) - d(\mathcal{K}_i, \mathcal{K}_j) \cdot \text{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j))$$
$$- \sum_{k \ne j} d(\mathcal{K}_i, \mathcal{K}_k) \cdot \text{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j))$$
$$= -\left(d(\mathcal{K}_i, \mathcal{K}_j) - (d(\mathcal{K}_i, \mathcal{K}_j) + \sum_{k \ne j} d(\mathcal{K}_i, \mathcal{K}_k)) \cdot \text{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j)) \right)$$

Since $\sum_{l=1}^{|\mathcal{M}|} d(\mathcal{K}_i, \mathcal{K}_l)$ = 1, we can further simplify it as

$$\frac{\partial KSGL(i)}{\partial d_E(\mathcal{A}_i, \mathcal{B}_j)} = -(d(\mathcal{K}_i, \mathcal{K}_j) - \text{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j))$$

In order to optimize, we need to set the respective partial derivative to be 0:

$$\frac{\partial KSGL(i)}{\partial d_E(\mathcal{A}_i, \mathcal{B}_j)} = -(d(\mathcal{K}_i, \mathcal{K}_j) - \text{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j)) = 0$$

In addition, the corresponding second partial derivative denoted as $\frac{\partial KSGL(i)}{\partial d_E^2(\mathcal{A}_i, \mathcal{B}_j)}$ manifests as follows:

$$\frac{\partial KSGL(i)}{\partial d_E^2(\mathcal{A}_i, \mathcal{B}_j)} = \text{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j))(1 - \text{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j)))$$

As softmax$(d_E(\mathcal{A}_i, \mathcal{B}_j))$ takes values within the open interval (0,1), it follows that $\frac{\partial KSGL(i)}{\partial d_E^2(\mathcal{A}_i,\mathcal{B}_j)}$ is always positive. Consequently, the pinnacle of optimization emerges as a global minimum.
Furthermore, when it comes to optimum:

$$\begin{aligned} d(\mathcal{K}_i, \mathcal{K}_j) &= \text{softmax}(d_E(\mathcal{A}_i, \mathcal{B}_j)) \\ d_E(\mathcal{A}_i, \mathcal{B}_j) &= \log(d(\mathcal{K}_i, \mathcal{K}_j)) + \log\left(\sum_{1\le l\le|\mathcal{M}|} e^{d_E(\mathcal{A}_i,\mathcal{B}_l)}\right) \end{aligned}$$

It is easy to show that when it reaches optimum, $d_E(A_i, B_j)$ is consistent with Knowledge Span Guidance $d(\mathcal{K}_i, \mathcal{K}_j)$. Without loss of generosity, suppose $d(\mathcal{K}_i, \mathcal{K}_j) > d(\mathcal{K}_i, \mathcal{K}_{j'})$:

$$\begin{aligned} d_E(\mathcal{A}_i, \mathcal{B}_j) - d_E(\mathcal{A}_i, \mathcal{B}_{j'}) &= \log(d(\mathcal{K}_i, \mathcal{K}_j)) + \log\left(\sum_{1\le l\le|\mathcal{M}|} e^{d_E(\mathcal{A}_i,\mathcal{B}_l)}\right) \\ &\quad - \left(\log(d(\mathcal{K}_i, \mathcal{K}_{j'})) + \log\left(\sum_{1\le l\le|\mathcal{M}|} e^{d_E(\mathcal{A}_i,\mathcal{B}_l)}\right)\right) \\ &= \log(d(\mathcal{K}_i, \mathcal{K}_j)) - \log(d(\mathcal{K}_i, \mathcal{K}_{j'})) \\ &= \log\left(\frac{d(\mathcal{K}_i, \mathcal{K}_j)}{d(\mathcal{K}_i, \mathcal{K}_{j'})}\right) > 0 \end{aligned}$$

□

B. Experimental Setting

B.1. Pre-training

Dataset: We use ^{13}C NMR spectra of about 20,000 molecules sourced from nmrshiftdb2 (Steinbeck et al., 2003)), a public access database that contains NMR spectra of organic molecules. In the collected dataset, molecule are aligned with their respective ^{13}C NMR spectra, and atomic alignments with peaks are also included. Notably, the dataset contains 12,771 molecules with fewer than 10 carbon atoms, 7,043 molecules featuring carbon atom counts ranging from 10 to 20, and 1,138 molecules incorporating more than 20 carbon atoms. The quality of the dataset was further validated by experienced organic chemists. We randomly sample 80% of the molecules for training and the rest for evaluation.

Training: We concurrently leverage both graph- and node-level alignment tasks in the pre-training of K-M3AID. Graph-level alignment focuses on aligning molecules with their spectra, accompanied by the utilization of cross-entropy loss for contrastive purposes. On the other hand, node-level alignment entails aligning atoms with their corresponding peaks, implemented through knowledge-guided instance-wise discrimination to achieve contrastive loss. A diverse set of molecular features is employed for training, including atomic number (node feature), chiral tags (node feature), hybridization (node feature), bond types (edge feature), and bond direction (edge feature). The spectral features are derived from peak intensity, peak position (chemical shift measured in ppm), and peak type.

B.2. Zero-Shot Molecular Retrieval

Dataset: We randomly collected 1 million molecules from PubChem (Kim et al., 2023) to form a molecular reference library. Subsequently, we carefully selected 1000 spectra, ensuring that they had not appeared in the training dataset, from an external dataset to serve as query spectra. Following this, the corresponding molecules associated with these 1000 spectra were added into the existing reference library.

Evaluation: We perform molecular retrieval using each of the selected spectra to determine if the correct corresponding molecular entity can be retrieved from the reference library. The model's performance is assessed at top-1 accuracy, as well as at accuracy of top 5, top 10, and top 25.

B.3. Zero-Shot Isomer Recognition

Dataset: We categorize isomers from the validation dataset to guarantee their absence from the training dataset. To assess effectiveness, we perform isomer recognition on each isomer group containing at least 10 molecules. Within the same group, isomers may be structural or spatial isomers of each other. Structural isomers refer to molecules with the same molecular formula but different structural arrangements of atoms, resulting in distinct chemical structures. On the other hand, spatial isomers, also known as stereoisomers, have the same molecular formula and arrangement of atoms but differ in the spatial orientation of their atoms in three-dimensional space, leading to different stereoisomeric forms. (For an elucidation of an isomer group and in-depth insights into isomers with NMR, please refer to the details provided in Appendix E.)

Evaluation: We conduct isomer recognition for each isomer group, aiming to assess the correct alignment of each spectrum with its respective molecule within each isomer group.

B.4. Zero-Shot Peak Assignment

Dataset: The dataset utilized for evaluating the overall performance of K-M3AID on zero-shot peak assignment is the validation dataset from the pre-training phase. In order to highlight the capabilities of K-M3AID in zero-shot peak assignment, the case studies include complex natural products featuring multiple fused rings, stereogenic (chiral) centers, and symmetric structures.

Evaluation: We conduct peak assignment within each molecule, aiming to assess the accurate alignment of each atom with its corresponding peak on the spectrum. It's important to note that this alignment process is confined to each individual molecule and not across different molecules.

C. Further ablation study about parameter choices

C.1. Ablation study about the choice of GIN structure and projection.

We choose GIN(Xu et al., 2018) as our graph encoder. By Table C.1, "GIN Depth" signifies the number of layers in the GIN, "GIN Embedding Dim" denotes the dimensionality of the embeddings generated by the GIN model, and "Projection Dim" indicates the resulting dimensionality after transforming the GIN-produced embeddings. In particular, the best performance is observed when the GIN model has 5 layers, GIN Embedding Dim is 128, and projection Dim is 512.

Table C.1: GIN structure and projection ablation study

GIN Depth	GIN Embedding Dim	Projection Dim	Validation accuracy (%)
3	128	128	86.6
3	256	128	86.8
3	512	128	86.3
5	128	128	89.4
5	256	128	89.6
5	512	128	89.3
3	128	256	86.6
3	256	256	86.8
3	512	256	86.3
5	128	256	89.4
5	256	256	89.6
5	512	256	89.3
3	128	512	86.6
3	256	512	86.5
3	512	512	86.2
5	32	512	84.0
5	64	512	87.5
5	128	512	**90.0**
5	256	512	89.4
5	512	512	88.9

C.2. Ablation Study on the Choice of τ_1 and τ_2

We also conducted a further ablation study exploring different combinations of τ_1 and τ_2 as shown in Table C.2. For this analysis, we fixed the GIN depth at 5, set the GIN embedding dimensionality to 128, and maintained a projection dimension of 512. We observe that the best performance is achieved when $\tau_1 = 10^{-5}$ and $\tau_2 = 10^1$.

Table C.2: Ablation study about tau_1 and tau_2. We have 5 layers and 128 dimension as the final representation.

τ_1	τ_2	**Molecular Alignment Accuracy (%)**	**Atom Alignment Accuracy (%)**
10^{-1}	10^1	94.9	89.6
10^{-1}	10^2	95.2	89.8
10^{-1}	10^3	95.6	89.6
10^{-1}	10^4	95.1	88.9
10^{-1}	10^5	95.0	89.3
10^{-2}	10^1	95.5	89.8
10^{-2}	10^2	94.8	89.8
10^{-2}	10^3	95.4	88.8
10^{-2}	10^4	94.8	87.2
10^{-2}	10^5	95.1	89.4
10^{-3}	10^1	95.0	89.2
10^{-3}	10^2	95.1	89.1
10^{-3}	10^3	95.2	89.0
10^{-3}	10^4	95.3	89.7
10^{-3}	10^5	95.0	89.4
10^{-4}	10^1	95.0	89.8
10^{-4}	10^2	95.1	89.7
10^{-4}	10^3	95.0	89.8
10^{-4}	10^4	95.3	89.5
10^{-4}	10^5	95.1	88.4
10^{-5}	10^1	**95.4**	**90.0**
10^{-5}	10^2	95.0	89.5
10^{-5}	10^3	95.8	89.6
10^{-5}	10^4	95.2	89.7
10^{-5}	10^5	95.0	89.7

D. Additional Results on Molecular Retrieval

Table D.1: Zero-shot molecular retrieval top 5, 10, 25 accuracy (%) with K-M3AID and baselines

Method	Accuracy	10^2	10^3	10^4	10^5	10^6
Knowledge Guide	Top 1(%)	95.8±1.0	80.4±3.9	46.3±1.2	18.0±0.8	5.8±1.7
	Top 5(%)	99.8±0.2	96.8±0.5	77.8±1.1	41.6±1.6	16.6±2.3
	Top 10(%)	100.0±0.0	98.8±0.3	87.7±1.0	53.9±1.8	25.2±3.2
	Top 25(%)	100.0±0.0	99.6±0.2	94.8±0.6	71.6±0.6	37.8±4.1
SP-ID	Top 1(%)	95.3±0.8	78.6±2.7	35.8±3.8	12.9±1.6	3.4±0.9
	Top 5(%)	95.4±0.1	77.3±0.7	44.7±2.3	16.2±2.4	4.4±1.5
	Top 10(%)	100.0±0.0	97.3±0.7	77.5±2.3	40.2±2.4	12.3±1.5
	Top 25(%)	100.0±0.0	99.1±0.2	85.9±1.0	53.1±3.0	18.5±1.8
WP-ID(th=1)	Top 1(%)	92.9±0.6	71.7±1.0	32.7±1.3	10.7±0.5	3.6±0.7
	Top 5(%)	99.6±0.1	93.8±0.8	63.9±1.5	29.3±1.5	10.2±1.2
	Top 10(%)	99.9±0.0	97.1±0.4	76.8±0.7	39.3±0.9	15.7±1.5
	Top 25(%)	100.0±0.0	99.1±0.2	88.2±0.6	55.7±1.1	26.5±2.0
No communication	Top 1(%)	94.8±1.2	77.6±1.4	40.1±1.2	14.4±0.9	4.1±1.1
	Top 5(%)	99.8±0.1	96.2±0.5	73.6±2.2	35.8±1.3	11.4±1.0
	Top 10(%)	99.9±0.1	98.6±0.3	84.1±1.4	47.3±1.8	17.3±1.8
	Top 25(%)	100.0±0.0	99.7±0.2	92.9±0.9	65.1±2.2	27.9±2.5

Table D.2: Comparing sampled molecules to their Top 1 neighbors using K-M3AID, SP-ID, WP-ID (th = 1), and K-M3AID without communication across datasets of varying sizes and employing different similarity metrics (%) including Cosine (Salton & McGill, 1986)), Dice (Dice, 1945)), Russel (Russel, 1980), Sokal (Sokal & Sneath, 1963) and Tanimoto (Tanimoto, 1957).

Methods	Similarity Metric	10^2	10^3	10^4	10^5	10^6
K-M3AID	Cosine	96.1±0.9	81.9±3.5	50.0±1.2	23.9±0.7	13.5±1.5
	Dice	96.1±0.9	81.9±3.5	50.0±1.2	23.7±0.6	13.1±1.5
	Russel	95.8±0.1	80.4±3.9	46.4±1.1	18.1±0.8	5.9±1.7
	Sokal	95.9±0.9	80.8±3.8	47.2±1.2	19.5±0.7	7.7±1.7
	Tanimoto	96.0±0.9	81.2±3.7	48.2±1.2	21.0±0.7	9.6±1.6
SP-ID	Cosine	95.4±0.8	78.6±2.4	42.5±3.6	20.9±1.5	12.2±0.7
	Dice	95.5±0.8	78.6±2.5	42.4±3.6	20.3±1.5	11.8±0.8
	Russel	95.1±0.8	76.9±2.7	38.3±3.8	14.4±1.6	4.3±0.9
	Sokal	95.2±0.8	77.4±2.6	39.4±3.8	15.9±1.6	6.3±0.8
	Tanimoto	95.3±0.8	77.8±2.6	40.4±3.7	17.4±1.6	8.2±0.8
WP-ID(th=1)	Cosine	93.4±0.6	73.9±0.1	37.5±1.2	17.2±0.4	11.4±0.5
	Dice	93.3±0.6	73.7±0.9	37.3±1.2	16.9±0.4	11.1±0.5
	Russel	92.9±0.6	71.7±1.0	32.7±1.3	10.8±0.5	3.7±0.6
	Sokal	93.0±0.6	72.3±3.4	33.9±1.2	12.4±0.5	5.6±0.6
	Tanimoto	93.1±0.6	72.8±1.0	35.1±1.2	14.0±0.5	7.5±0.6
No communication	Cosine	95.2±1.1	79.2±1.3	44.2±1.2	20.6±0.7	12.2±0.9
	Dice	95.1±1.1	79.1±1.3	44.0±1.2	20.4±0.7	11.8±0.9
	Russel	94.8±1.2	77.6±1.4	40.2±1.2	14.5±0.9	4.2±1.1
	Sokal	94.9±1.1	78.0±1.4	41.2±1.2	16.0±0.8	6.2±1.1
	Tanimoto	95.0±1.1	78.4±1.3	42.2±1.2	17.6±0.8	8.2±1.0

E. Additional Discussion about Isomers

E.1. Isomer Category

Isomers typically fall into two main categories: constitutional (structural) isomers, which share the same chemical formula but display distinct atom connectivity, and stereoisomers (spatial isomers), which share the same topology graph but diverge in their three-dimensional arrangement (see Figure E.1). Constitutional isomers are NMR-variant, meaning that different isomers produce distinct NMR spectrum. In the sub-categories of stereoisomers, enantiomers are NMR-invariant, but diastereomers and cis-trans isomers are NMR-variant.

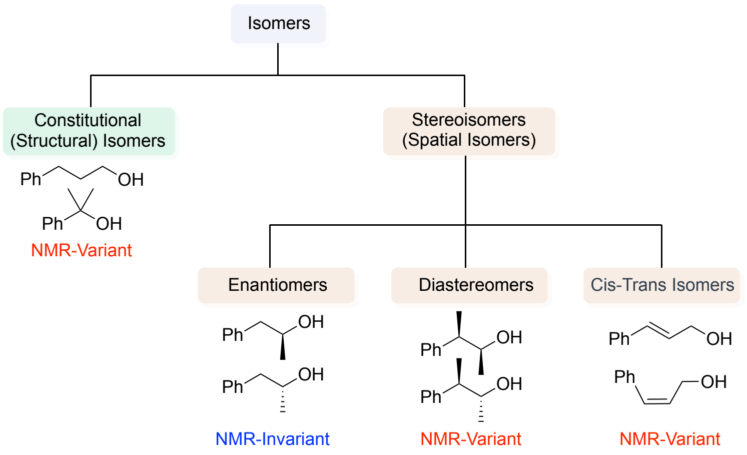

Figure E.1: NMR Variability in Isomers

E.2. Isomers Group for $C_7H_{11}NO_3$

Here is an example for isomer groups. In this isomer group of $C_7H_{11}NO_3$, they all share the same chemical formula in Figure E.2. The first 10 are constitutional (structural) isomers of each other (cycled green), the last 4 are two pairs of diastereomers (cycled brown). Each of these isomers corresponds to a distinct NMR spectrum.

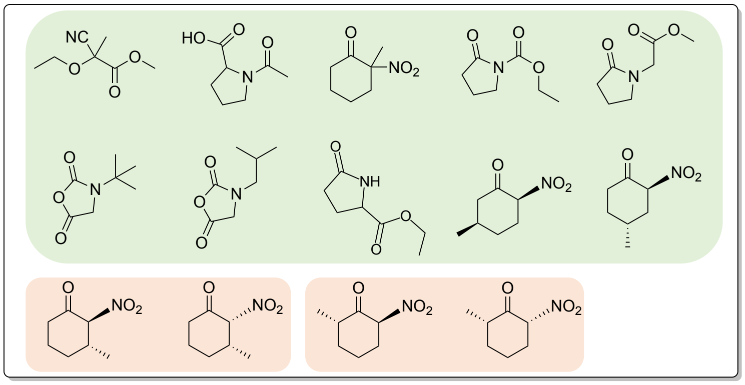

Figure E.2: Isomer demo for $C_7H_{11}NO_3$

F. Additional Results on Peak Assignment

In complex natural product molecules, it is a common situation that the local contents of some atoms within the same molecule exhibit a high degree of similarity. It gives rise to challenges for the atomic alignment, as some atoms correspond to ppm values in close proximity. However, our K-M3AID model is capable of recognizing each of the atoms with effective learnt embeddings and deciphering the correspondences among the atoms and the peaks at zero-shot. Two complex natural product molecules with multiple rings (4 and 4, respectively) and multiple chiral centers (6 and 8, respectively) are taken to showcase the effectiveness of atomic alignment (see Appendix Figure F.1).

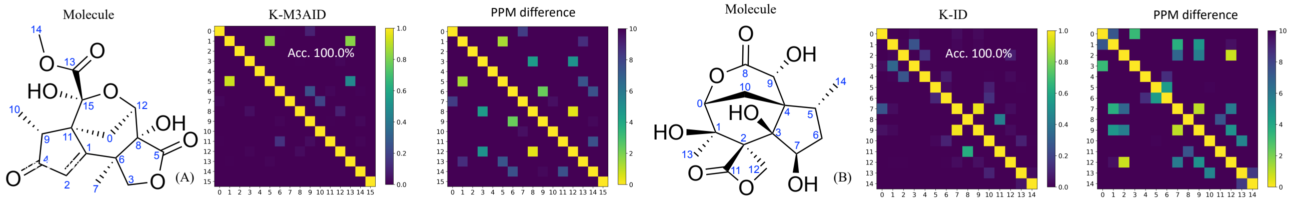

Figure F.1: Examples of Zero-shot Atomic Alignment for Complex Natural Products. Yellow cells in the PPM difference represent the ground truth alignment.

In molecular A in Figure F.2, atom 13 and atom 14 are tertiary carbons (attaching to 3 carbons and 1 hydrogen) and on the same 5-member ring, corresponding to the ppm of 34.3 and 35.6, respectively. The similar local content of these two atoms fools SP-ID and WP-ID. In addition, WP-ID fails with more atomic alignments. The molecular B is chemical symmetric regarding atom 0. Thus, atom 1 and atom 3 correspond to the same peak on the spectra. The ppm of atom 1 and atom 3 is 114.2, the ppm of atom 2 and atom 4 is 110.0. While there is 4.2 difference, SP-ID and WP-ID fails to pick up right alignment for atom 1 and atom 3. In contrast, K-ID succeed to align the atoms with peaks in both molecules.

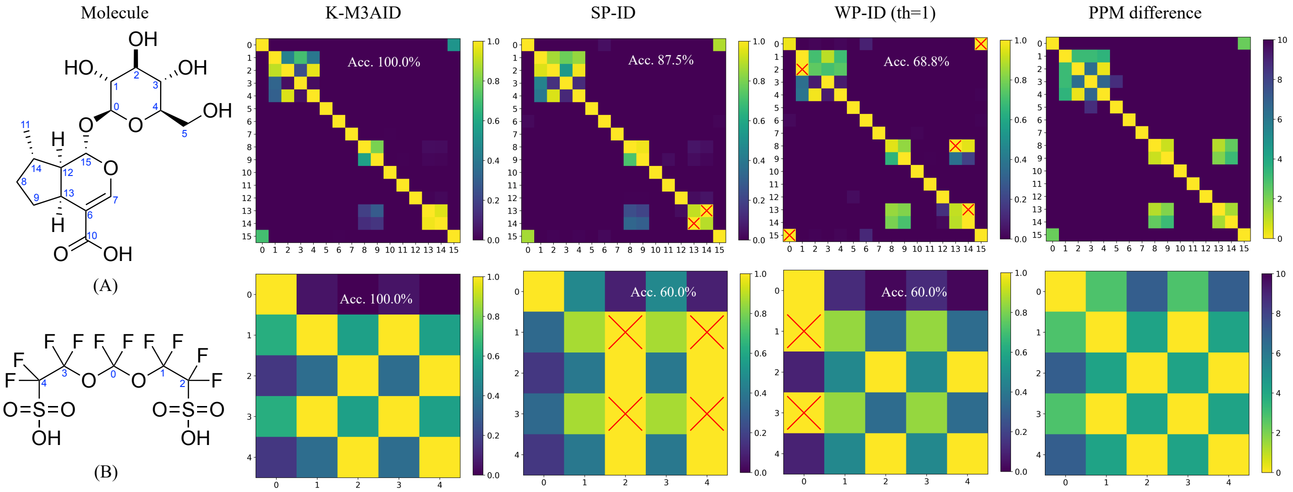

Figure F.2: Extra case studies of IE-Meta-MMA.Yellow cells in the PPM differerence represent the ground truth alignment, and red cross represents the wrong alignment.