ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

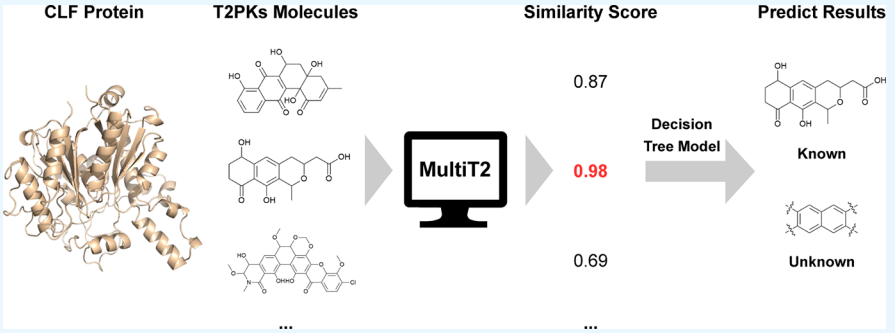
MultiT2: A Tool Connecting the Multimodal Data for Bacterial Aromatic Polyketide Natural Products

Liangjun Ge,[†] Qiandi Gao,[†] Jiayi He,[†] Xiaoyu Wang, Jiaquan Huang,* Heqian Zhang,* and Zhiwei Qin*

Cite This: *ACS Omega* 2025, 10, 5105−5110

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The integration of artificial intelligence (AI) into natural product science is an exciting and rapidly evolving area of research. By combining classical chemistry and biology with deep learning, these technologies have significantly improved research efficiency, particularly in overcoming laborious and time-consuming processes. Recently, there has been growing interest in leveraging multimodal algorithms to integrate biologically relevant yet mathematically disparate data sets in order to reorganize knowledge graphs. However, to the best of our knowledge, no studies have yet applied this approach specifically within the natural product field. This is largely because correlating multimodal natural product data is challenging due to their high degree of fragmentation. Here, we present MultiT2, an algorithm that connects these disparate data from bacterial aromatic polyketides, which form a medically important natural product family, as a showcase. Through a large-scale causal inference process, this approach aims to transcend mere prediction, unlocking new dimensions in the natural product discovery and research domains.

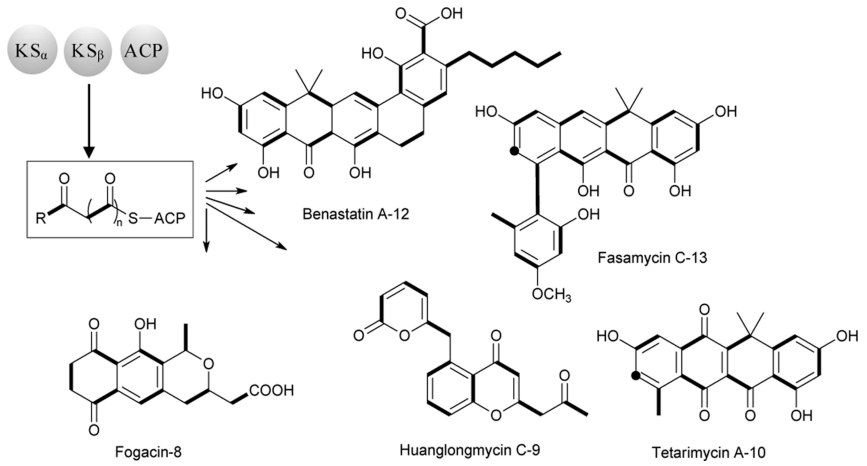
1. INTRODUCTION

Natural bacterial type II polyketide products (T2PKs), or aromatic polyketides, form an extensive family of biosynthetically related molecules that are characterized by significant structural diversity and often exhibit potent biological activities.[1] The carbon skeletons of T2PKs are universally derived from dissociated type II polyketide synthases (T2PKSs), and further structural diversity is introduced by tailoring enzymes following the initial polyketide biosynthesis process.[2] A T2PKS typically comprises a heterodimer of a ketosynthase (KS, or KS_α) and a chain length factor (CLF, or KS_β), an acyl carrier protein (ACP), and a malonyl-CoA transacylase (MAT). Together, the KS, CLF, and ACP, known as the minimal PKS, initiate polyketide biosynthesis by decarboxylating malonyl-ACP (Scheme 1).[3] A KS catalyzes the polyketide chain elongation process by performing Claisen condensation on acyl-thioester units such as malonyl-CoA, whereas the CLF regulates the chain length by controlling the C−C bond formation during each iterative cycle. These enzymes collectively determine whether to extend or terminate the chain after a specific number of building blocks are assembled. To date, approximately 163 types of T2PK families have been identified based on chain lengths of 8, 9, 10, 12, or 13 units.[4] This divergence may have been driven by environmental demands for greater polyketide diversity (Scheme 1).

Over the years, CLFs have been studied as pivotal markers that influence the chemical structures of T2PKs, either individually or in synergy with other enzymes in biosynthesis-related gene clusters (BGCs). For example, Hillenmeyer et al. noted a correlation between CLF (KS_β) phylogeny and T2PK chain length,[5] whereas Chen et al. leveraged a CLF as a biomarker to construct a coevolutionary statistical model, thus expanding the T2PK biosynthetic landscape.[6] Recently, Huang et al. introduced DeepT2, a deep learning-based protein language model (PLM) that enables the classification of CLFs from public databases and the prediction of structurally novel T2PKs.[4] However, these approaches rely exclusively on enzyme data derived from BGCs and are limited in their

Received: December 13, 2024
Revised: January 15, 2025
Accepted: January 23, 2025
Published: January 28, 2025

Scheme 1. Representative T2PK Families Based on the Chain Length (8, 9, 10, 12, and 13 Building Units)



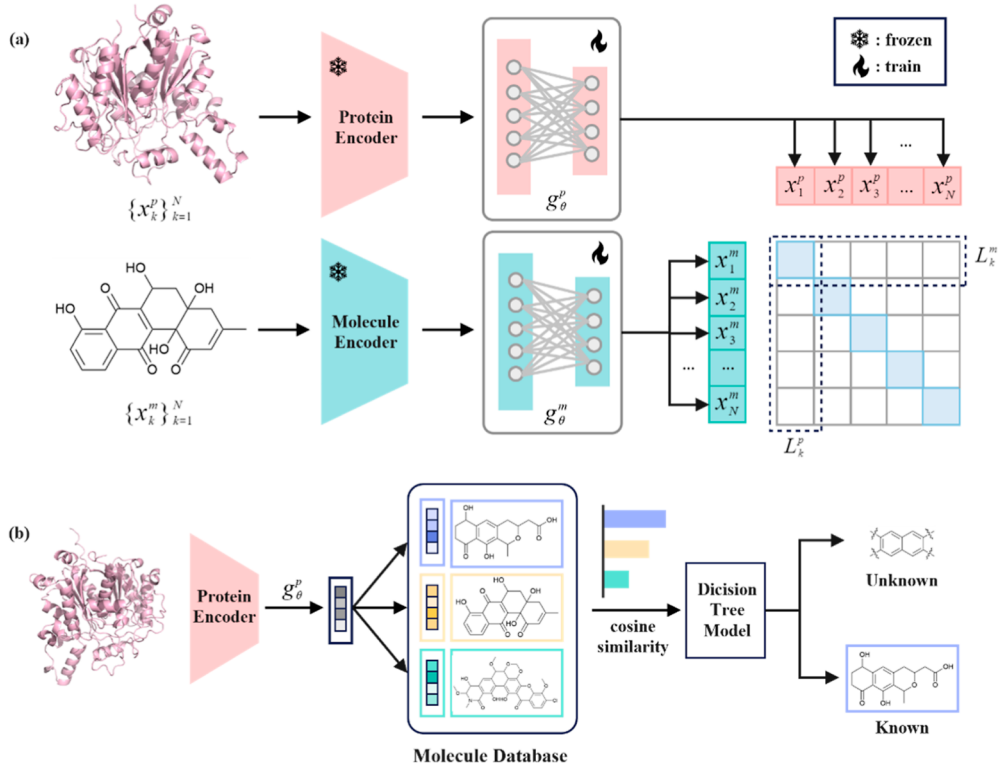


Figure 1. Overall workflow of MultiT2 showing a two-phase training. (a) During the training phase, two encoder layers are frozen, while the nonlinear protein layer (g_θ^p) and the nonlinear molecule layer (g_θ^m) are trained. The model learns feature representations by predicting the correct protein-molecule pairs (x_k^p, x_k^m) within each batch. (b) In the prediction phase, the protein encoder and trained nonlinear protein layer (g_θ^p) are used to encode query protein sequences. These encoded sequences are then matched with candidate molecules from the molecular database for similarity. A decision tree is employed to predict whether the corresponding T2PK is known.

ability to incorporate multimodal data sources, such as genomic, proteomic, metabolomic, spectroscopic, or (bio)-chemical data. Importantly, each data type offers unique insights into the same biochemical entities from different perspectives, with the potential to enrich and inform one another.

The integration of diverse natural product data types presents a major challenge, as these data sets are often scattered and fragmented, despite offering complementary insights. The existing deep learning architectures struggle to fully leverage these multimodal and unbalanced data, limiting their potential for discovering new natural products. In this study, we address this issue by employing multimodal algorithms to unify these data sets. To the best of our knowledge, this represents the first application of such a strategy in natural product research.

2. METHODS

During our search for suitable candidate algorithms, CLIP captured our attention. The CLIP model introduces a novel paradigm that effectively integrates multimodal data—specifically images and text—through contrastive learning, and eventually yielding high-precision zero-shot predictions.[7] This innovation has sparked a revolution in multimodal artificial

https://doi.org/10.1021/acsomega.4c11266
ACS Omega 2025, 10, 5105−5110

intelligence. Inspired by CLIP, our study focuses on the significant role of CLFs in influencing the structures of T2PKs. We propose the MultiT2 model, which employs a protein language model called ESM2[8] and a chemical language model called MoLFormer[9] as encoders to integrate two distinct data types: protein sequences and the SMILES representations of T2PK structures (Figure 1 and Table 1). To optimize the

Table 1. Pseudocode Showing the Overall Algorithms for MultiT2

```
Algorithm: MultiT2 pseudocode
Input: D_mol = {m1, m2, ..., mn};
       D_prot = {p1, p2, ..., pn};
       D_test = {t1, t2, ..., tm}
Output: y_pred = {y1, y2, ..., ym}
Procedure:
1. Embedding Generation
   M ← MolFormer(D_mol)
   P ← ESM2(D_prot)

2. Coembedding Model Training
   For epoch ← 1 to max_epochs:
      L_con ← ContrastiveLoss(M, P)
      L_cls ← ClassificationLoss(M, P)
      update_model_weights()
      save_best_checkpoint()

3. Decision Tree Training
   S ← Coembedding(M, P)
   pos_samples ← diagonal(S)
   neg_samples ← mean(S - diagonal(S))
   X ← concatenate(pos_samples, neg_samples)
   y ← concatenate(ones(N), zeros(N))
   tree.fit(X, y)

4. Batch Prediction
   For each batch b in D_test:
      P_test ← ESM2(b)
      S_test ← Coembedding(M, P_test)
      max_sim ← max(S_test, axis=1)
      y_pred ← tree.predict(max_sim)
Return y_pred
```

embeddings of these two data types within a high-dimensional space, we apply a contrastive loss, ensuring that the embedding of a given CLF closely resembles the corresponding SMILES embedding. Ultimately, we assess the similarity between the CLF embeddings output by the model and the SMILES embeddings to determine whether a compound is biosynthesized by the associated protein. This methodology facilitates the prediction of known natural product structures and aids in the discovery of novel type II polyketides.

2.1. Data Sets Preparation. In our previous work, we utilized 163 CLF sequences to develop the DeepT2 model.[4] In this study, we identified the natural product structures corresponding to these 163 CLFs from the MIBiG and PubChem databases, as well as the relevant literature. Utilizing the Rdkit toolkit, we converted these structures into standard SMILES representations, ultimately creating a comprehensive data set comprising 146 pairs of CLF sequences and T2PK SMILES data (Table S1). This data set was then partitioned into training and test sets at an 8:2 ratio.

2.2. Model Architecture. The MultiT2 model comprises two distinct pretrained transformer encoders, which are complemented by nonlinear layers. It accepts CLF protein sequences and compound SMILES sequences as inputs. The protein sequences are transformed into protein embeddings via ESM2, whereas the SMILES representations are converted into molecular embeddings via MoLFormer. Each embedding is projected into a shared embedding space through separate nonlinear layers (g_θ^p, g_θ^m). To enhance the performance of the model, we calculate the contrastive loss between the outputs of these two nonlinear layers. This process aims to maximize the similarity between positive sample pairs (protein sequences and their corresponding compound molecules) and minimize the similarity between negative sample pairs (protein sequences and unrelated compound molecules). The ultimate goal is to ensure that the CLF protein sequences and their corresponding compound SMILES embeddings are highly similar (Figure 1a).

2.3. Training Strategy. Our training methodology alternates between classification and contrastive learning phases. In the classification phase, we optimize the model using cross-entropy loss. In the contrastive phase, we compute the similarity metric for each protein-molecule pair using cosine similarity as the measure between their embeddings (hereafter referred to as "similarity").[10] Next, we employ a batch sampling strategy that is analogous to that used in CLIP. For a given batch of paired data $\{(x_k^p, x_k^m)\}_{k=1}^N$, which consists of N entries, we combine the protein data $\{x_k^p\}_{k=1}^N$ and molecular data $\{x_k^m\}_{k=1}^N$ to create N^2 pairs of protein-molecule data (x_i^p, x_j^m) $(i,j \in \{1, ..., N\})$. When a pair is defined with $i = j$, it constitutes a positive sample pair; conversely, when it is defined with $i \neq j$, it represents a negative sample pair. Ultimately, we utilize the contrastive loss function $\mathcal{L}_{con}$ to maximize the similarity $s(x_i^p, x_i^m)$ between the positive sample pairs while minimizing the similarity $s(x_i^p, x_j^m)$ between the negative sample pairs, as defined from previous work but modified as showing below:[11]

$$\mathcal{L}_k^p = -\frac{1}{N}\log\frac{\exp(s(x_k^p, x_k^m))}{\sum_{i=1}^N \exp(s(x_k^p, x_i^m))} \quad (1)$$

$$\mathcal{L}_k^m = -\frac{1}{N}\log\frac{\exp(s(x_k^p, x_k^m))}{\sum_{i=1}^N \exp(s(x_i^p, x_k^m))} \quad (2)$$

$$\mathcal{L}_{con} = \frac{1}{2}\sum_{k=1}^N (\mathcal{L}_k^p + \mathcal{L}_k^m) \quad (3)$$

where x_k^p represents the k-th input protein data point; x_k^m represents the k-th input molecular data point; $s(x_i^p, x_j^m)$ represents the similarity between the protein data x_i^p and the molecular data x_j^m; $\mathcal{L}_k^p$ represents the partial loss calculated based on the protein data x_k^p; $\mathcal{L}_k^m$ represents the partial loss calculated based on the molecular data x_k^m; and $\mathcal{L}_{con}$ represents the final contrastive loss function, which is obtained by averaging the losses of all sample pairs.

Utilizing the aforementioned method, we trained the model by freezing the weights of the two pretrained encoders. The parameters of the two nonlinear layers were optimized via the Adam optimizer. The batch size was set to 32, and the initial learning rate was configured to 1×10^{-4}. Additionally, we implemented a cosine annealing strategy to dynamically adjust the learning rate throughout the training process.

2.4. Workflow for MultiT2. The prediction script in the GitHub repository https://github.com/Qinlab502/MultiT2 demonstrates how to utilize the model for prediction tasks. The workflow is as follows:

- Input: The model accepts a FASTA file containing CLF sequences as input.

https://doi.org/10.1021/acsomega.4c11266
ACS Omega 2025, 10, 5105−5110

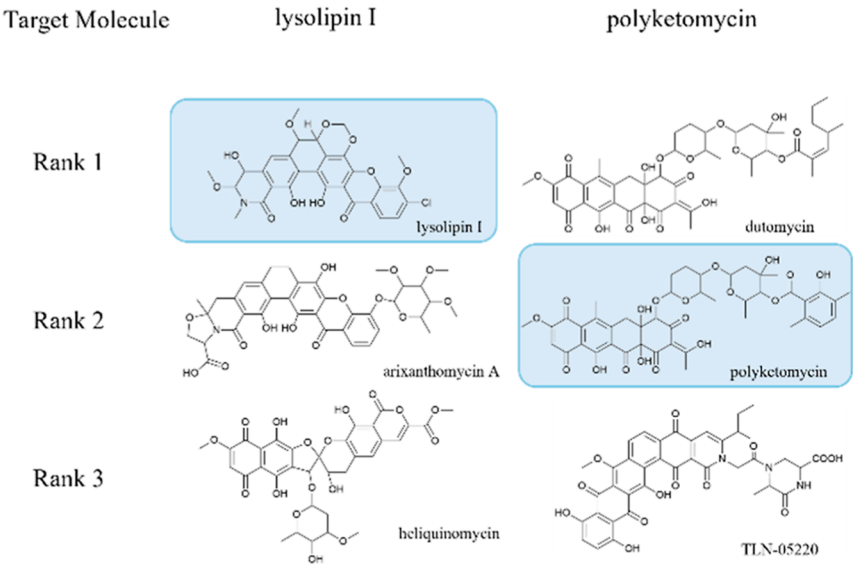


Figure 2. Top three scoring predictions by MultiT2 for the target molecules corresponding to two different CLF sequences.

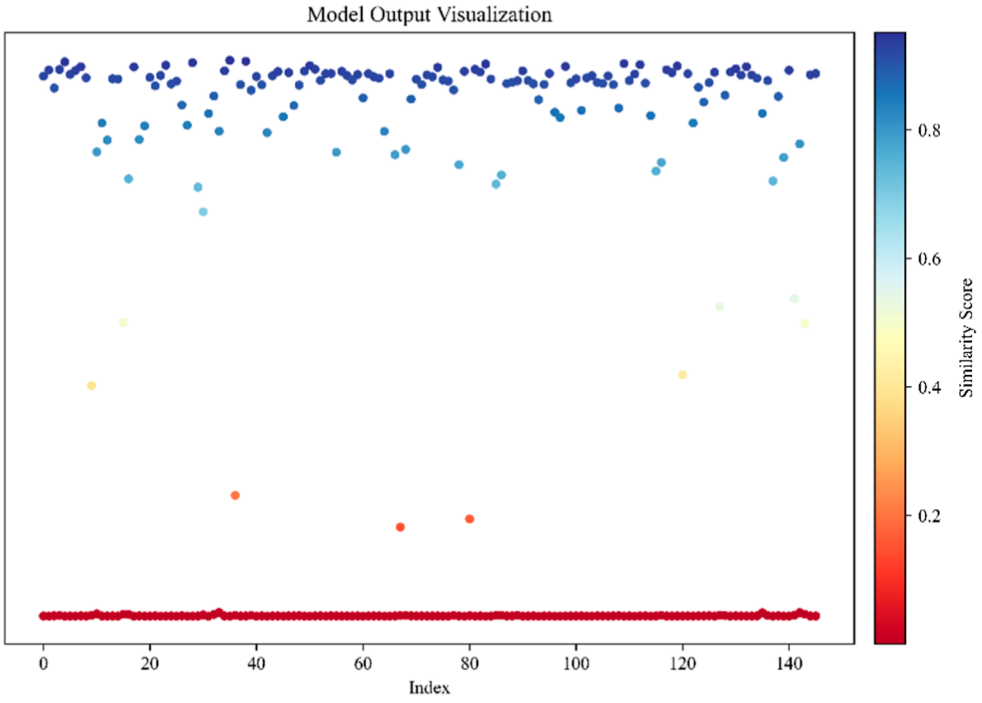


Figure 3. Distribution of similarity scores between sample pairs output by MultiT2.

- Embedding Calculation: MultiT2 computes the embeddings of both the input CLF sequences and a set of 146 known T2PKs. Subsequently, it calculates the similarity between the embeddings of the CLF sequences and the known T2PKs.
- Decision Tree Classification: For each CLF sequence, the T2PK with the highest similarity is identified and used as input to a decision tree model. This model determines whether the CLF sequence belongs to the known in-domain samples.
- Label Assignment: If the sequence is classified as in-domain, it is assigned a label of "1", and the SMILES of the most similar T2PK from the 146 known compounds is provided as output. While if the sequence is classified as out-of-domain, it is assigned a label of "0", indicating that the sequence may be associated with the biosynthesis of a potentially novel T2PK scaffold.
- Output: The prediction results, including the assigned labels and the corresponding SMILES (if in-domain), are written to a text file.

3. RESULTS AND DISCUSSION

3.1. Evaluation of MultiT2. We utilized MultiT2 to calculate the embeddings of the CLF protein sequences and the SMILES representations of known compound molecules, projecting the embeddings in a shared embedding space (Figures S1 and S2). We then computed the similarity scores between the protein embeddings and all the molecular embeddings in batches, sorting the molecules according to

https://doi.org/10.1021/acsomega.4c11266
ACS Omega 2025, 10, 5105−5110

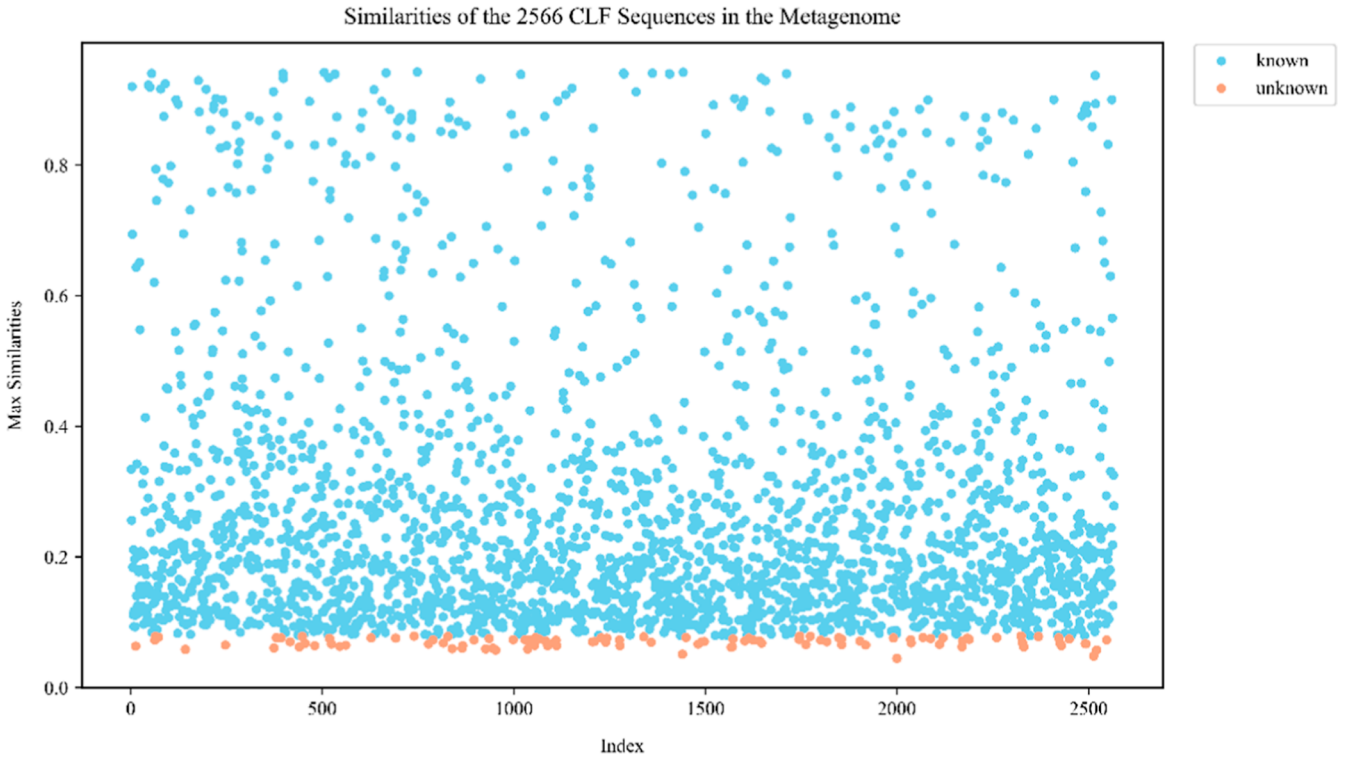


Figure 4. Prediction of corresponding T2PKs from CLFs in metagenomes based on similarity score distribution. Blue dots represent CLF sequences for known compounds, and orange dots indicate those for novel compounds.

these scores (Figure 1b). The top-*k* accuracy rates were subsequently calculated, with top-1, top-3, and top-5 accuracy rates on the test set being 86.67%, 100%, and 100%, respectively. Importantly, in some instances, the highest-ranked molecule may not correspond to the actual target product of the protein; instead, it may share certain structural similarities with other molecules. However, as the number of molecules considered in the ranking process increases, the true target product is consistently found among the top-ranked candidates.

3.2. Query Retrieval Task Involving Known T2PKs. To further evaluate the generalization ability of the developed model, we tested it using the CLFs that were not part of the training data set but their corresponding chemical structures are already known. For example, the first CLF sequence is involved in the biosynthesis of the target molecule lysolipin, and MultiT2 ranked the target molecule first among the top three compounds. Similarly, for the second CLF sequence involved in synthesizing polyketomycin, the target molecule ranked second. Interestingly, in the latter case, the highest-ranked molecule, dutomycin, shares a similar carbon skeleton with polyketomycin but differs in its sugar moiety (Figure 2). This suggests that the model can predict the structure of compounds to some extent. Since MultiT2 considers only the CLF as its input, more accurate predictions would require the model to consider the contributions of other enzymes in a gene cluster, such as KS, ACP, and cyclase/aromatase, which shape the final product structure. These findings support the hypothesis that CLFs interact with other key functional enzymes during T2PK biosynthesis, further highlighting the significant role of CLF sequences in predicting the chemical structures of T2PKs.

3.3. Prediction Task of Novel T2PKs Scaffolds. In natural product discovery cases, identifying compounds with novel scaffolds has always been one of the major challenges. In this work, we established MultiT2 to demonstrate the ability to mine novel T2PK scaffolds. Notably, in a query retrieval task involving known compounds, MultiT2 can output the molecule that most closely matches the known molecular structure library (positive sample) based on the input CLF sequence features, providing the corresponding best match similarity score. In contrast, for molecules that clearly do not match (negative samples), the model typically assigns lower similarity scores. The definition of a molecule with novel scaffold can be understood as follows: if no molecule in the known structure library closely matches the features of the input CLF sequence, the sequence may be involved in the biosynthesis of a new molecule with a completely distinct scaffold. In this work, we trained a decision tree model using similarity scores from positive samples and the average similarity scores from negative samples (Figure 3). When novel scaffolds were detected, MultiT2 calculated the similarity scores between the embedding of the input CLF sequence and all known molecules in the library and used the highest score as inputs for the decision tree model, which then determined whether the CLF sequence corresponds to a novel T2PK scaffold (Figure 1b). By using this method, we tested whether the CLF sequences derived from the metagenome, which have not been chemically characterized, could correspond to novel T2PK products. This led to the identification of 107 CLF sequences that may warrant further attention for the task discussed above, and these sequences will be the subject of future wet lab experiments in our laboratory (Figure 4 and Table S2).

https://doi.org/10.1021/acsomega.4c11266
ACS Omega 2025, 10, 5105−5110

4. CONCLUSION

While this work focused only on multimodal natural product data, including data on polyketide synthases and their associated chemical structures, building a more advanced model that can leverage the diverse data types of T2PKs is a significant next step and currently underway in our laboratory. Nevertheless, as a proof of concept, we have demonstrated the feasibility of integrating diverse data types within a biology-computing-chemistry cycle. We therefore anticipate that this work will support drug development processes based on natural products and facilitate the discovery of new chemical entities.

■ ASSOCIATED CONTENT

Data Availability Statement

The source code of MultiT2 is available at the GitHub repository https://github.com/Qinlab502/MultiT2. This repository also includes detailed documentation on the installation process, input files and usage examples.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c11266.

Supplementary Figures S1 and S2, t-SNE visualization of the latent space representations for CLF sequences and their corresponding compound SMILES before and after MultiT2 training (PDF)

Supplementary Tabel S1, general information on known T2PKs SMILESs and corresponding KS_{β} sequences (XLSX)

Supplementary Tabel S2, general information on metagenome T2PKs prediction results (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Zhiwei Qin − *Center for Biological Science and Technology, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong 519087, China;* orcid.org/0000-0002-1444-3684; Email: z.qin@bnu.edu.cn

Heqian Zhang − *Center for Biological Science and Technology, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong 519087, China;* Email: zhangheqian@bnu.edu.cn

Jiaquan Huang − *Center for Biological Science and Technology, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong 519087, China;* Email: jiaquan_terry@bnu.edu.cn

Authors

Liangjun Ge − *Center for Biological Science and Technology, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong 519087, China*

Qiandi Gao − *Center for Biological Science and Technology, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong 519087, China*

Jiayi He − *Center for Biological Science and Technology, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong 519087, China*

Xiaoyu Wang − *Center for Biological Science and Technology, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong 519087, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c11266

Author Contributions

[†]L.G., Q.G., and J.H. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (32170079 to Z.Q., 32200035 to H.Z., and 32400235 to J.H.), the Natural Science Foundation of Guangdong (2021A1515012026 and 2024A1515012593 to Z.Q., and 2023A1515110175 to J.H.), Guangdong Talent Scheme (2021QN020100 to Z.Q.). The authors would like to thank the Interdisciplinary Intelligence Super Computer Center, Beijing Normal University, for High Performance Computing for access to computational resources.

■ REFERENCES

(1) Staunton, J.; Weissman, K. J. Polyketide biosynthesis: a millennium review. *Nat. Prod. Rep.* **2001**, *18*, 380−416.

(2) Hertweck, C. The biosynthetic logic of polyketide diversity. *Angew. Chem., Int. Ed.* **2009**, *48*, 4688−4716.

(3) Bräuer, A.; Zhou, Q.; Grammbitter, G. L.; Schmalhofer, M.; Rühl, M.; Kaila, V. R.; Bode, H. B.; Groll, M. Structural snapshots of the minimal PKS system responsible for octaketide biosynthesis. *Nat. Chem.* **2020**, *12*, 755−763.

(4) Huang, J.; Gao, Q.; Tang, Y.; Wu, Y.; Zhang, H.; Qin, Z. A deep learning model for type II polyketide natural product prediction without sequence alignment. *Digital Discovery* **2023**, *2*, 1484−1493.

(5) Hillenmeyer, M. E.; Vandova, G. A.; Berlew, E. E.; Charkoudian, L. K. Evolution of chemical diversity by coordinated gene swaps in type II polyketide gene clusters. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 13952−13957.

(6) Chen, S.; Zhang, C.; Zhang, L. Investigation of the molecular landscape of bacterial aromatic polyketides by global analysis of type II polyketide synthases. *Angew. Chem., Int. Ed.* **2022**, *61*, No. e202202286.

(7) Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of Machine Learning Research*, 2021; Vol. *139*.

(8) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.

(9) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.* **2022**, *4*, 1256−1264.

(10) Singh, R.; Sledzieski, S.; Bryson, B.; Cowen, L.; Berger, B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120*, No. e2220778120.

(11) Gao, B.; Qiang, B.; Tan, H.; Jia, Y.; Ren, M.; Lu, M.; Liu, J.; Ma, W.-Y.; Lan, Y. Drugclip: Contrasive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 2024; Vol. *36*.