Supplementary Materials

Integrating sequence and chemical insights: a co-modeling AI prediction framework for peptides

Zihan Liu[1,#], Meiru Yan[2,3,#], Zhihui Zhu[2,3,#], Yongfu Guo[4], Mouzheng Xu[4], Jiaqi Wang[2,3,*]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

[2]Wisdom Lake Academy of Pharmacy, Xi’an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China.

[3]Jiangsu Province Higher Education Key Laboratory of Cell Therapy Nanoformulation, Xi’an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China.

[4]XJTLU High Performance Computing Platform, Management Information Technology and System Office, Xi’an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China.

[#]Authors contributed equally.

[*]**Correspondence to:** Dr. Jiaqi Wang, Wisdom Lake Academy of Pharmacy, Xi’an Jiaotong-Liverpool University, No 111, Renai Road, Suzhou 215123, Jiangsu, China. E-mail: Jiaqi.Wang02@xjtlu.edu.cn

Supplementary Methods

Contrastive learning-based co-modelling framework

We conducted empirical evaluations of non-co-modeling and co-modeling frameworks on downstream datasets for both regression and classification tasks. Our findings substantiate the superior performance of co-modeling approaches, with particular performance enhancement on the contrastive learning-based co-modelling framework. Building on these results, we provide a detailed elaboration of our contrastive learning-based co-modeling framework as depicted in **Figure 1** in the main text.

Normally, chemical structural information depicted by molecular graphs offers a higher level of detail compared to amino acid sequences. However, GNNs exhibit weaker long-range dependency with molecular graphs compared to attention mechanisms over message passing. Our objective is to enrich the sequence encoder with molecular information via contrastive learning, and to leverage attention mechanisms for capturing long-range relationships among amino acids, thereby achieving more accurate predictions of peptide properties and functions.

Let the primary structure of a peptide x be denoted by its sequence x_{seq} and its molecular graph x_{graph}, with the corresponding label being y. For a given peptide input (x_{seq}, x_{graph}), sequence-based encoder module $f_e(x_{seq})$ and graph-based encoder $g_e(x_{graph})$ produce the representations h_{seq} and h_{graph}, respectively. The downstream predictors, $f_p(h_{seq})$ and $g_p(h_{graph})$, which are based on MLP, take these sequences and graph representations as inputs, respectively. To optimize the predictive performance, we minimize the prediction loss associated with the downstream tasks as follows:

$$\mathrm{L}_{prd} = \tilde{a}_{(x,y)\sim D^{tr}} \left[\mathrm{L}_{\mathrm{sup}} \left(f_p \left(f_e \left(x_{seq} \right) \right), y \right) + \mathrm{L}_{sup} \left(g_p \left(g_e \left(x_{graph} \right) \right), y \right) \right] \quad \text{(S1)}$$

where D^{tr} represents the training dataset, and L_{sup} denotes the supervised loss associated with downstream tasks. To enhance the interdependence of the outputs from sequence and graph encoders, we employ an unsupervised loss based on contrastive learning, specifically utilizing the InfoNCE loss as described in reference[1]. For the i-th peptide chain, its contrastive loss is presented as follows:

$$\mathrm{L}_{con} = \tilde{\mathrm{a}}_{x\sim D^{tr}}\left[-\log\frac{\exp\left(h_{seq}^{T}\cdot h_{graph}/\tau\right)}{\sum_{j=0}^{K}\exp\left(h^{T}\cdot h^{-}/\tau\right)}\right] \qquad \text{(S2)}$$

where h (*i.e.*, outputs from the encoders) denotes either h_{seq} or h_{graph} of the peptide sample x, K denotes the number of negative samples, h^{-} is the encoder output of another peptide. The pair (h, h^{-}) forms a negative sample pair. The parameter τ is the temperature hyperparameter used in the contrastive learning framework. The theoretical justification for the enhancement of the discriminative capability within the co-modeling framework, based on the contrastive loss as presented in **Equation (S2)** is demonstrated as below.

The effectiveness of contrastive loss in co-modeling framework

Theorem 1: Given the sequence x_{seq} and the chemical structure x_{graph} as two distinct depictions of a peptide primary structure, by employing InfoNCE-based contrastive loss, the fusion of learned representations h_{seq} and h_{graph} derived from x_{seq} and x_{graph}, can effectively enhance the discriminative capacity of the model.

Proof. Assume D_c is a distribution of peptides with label c on a given downstream task from which sequence data x_{seq} and graph data x_{graph} with label c can be drawn as:

$$\begin{aligned} x_{seq} &\sim D_c\left(X \mid \tau = seq\right) \\ x_{graph} &\sim D_c\left(X \mid \tau = graph\right) \end{aligned} \qquad \text{(S3)}$$

x_{seq} and x_{graph} are two distinct observation forms of a specific peptide $X \sim D_c$. The condition variable τ serves as a determining factor to classify whether the data is observed as sequence or molecular graph. Within our method, x_{seq} and x_{graph} above are considered as a positive pair (x, x^{+}), and (h, h^{+}) are considered as the corresponding positive pair of learned representations from encoders. We illustrate the aforementioned definition within a classification downstream task. This approach can be generalized to regression tasks by conceptualizing each regression value as a distinct soft label in a binary classification context.

In the subsequent proof, we aim to construct a function that quantifies the discriminative capacity of the model, and we demonstrate such a function can be upper bounded by the contrastive loss. By minimizing this function, the model will be enabled to effectively distinguish representations with different labels in downstream tasks.

Therefore, we reformulate and simplify the contrastive loss $\mathcal{L}_{con}$ in **Equation (S2)** as following:

$$
\begin{aligned}
\mathcal{L}_{con} &= \mathbb{E}_{\substack{c,\{c_i^-\}_{i=1}^{k} \\ \sim C}} \mathbb{E}_{\substack{(x,x^+)\sim D_c \\ x_i^- \sim D_{c_i^-}}} \left[-log\left(\frac{e^{\left(h^T h^+\right)}}{e^{\left(h^T h^+\right)} + \sum_{i=1}^{k} e^{\left(h^T h_i^-\right)}} \right) \right] \\
&= \mathbb{E}_{\substack{c,\{c_i^-\}_{i=1}^{k} \\ \sim C}} \left[\mathbb{E}_{\substack{(x,x^+)\sim D_c \\ x_i^- \sim D_{c_i^-}}} \left[-log\left(\frac{1}{1 + \sum_{i=1}^{k} e^{\left(h^T \left(h_i^- - h^+\right)\right)}} \right) \right] \right] \\
&= \mathbb{E}_{\substack{c,\{c_i^-\}_{i=1}^{k} \\ \sim C}} \left[\mathbb{E}_{\substack{(x,x^+)\sim D_c \\ x_i^- \sim D_{c_i^-}}} \left[l\left(h^T\left(h^+ - h_i^-\right)\right) \right] \right]
\end{aligned}
\tag{S4}
$$

$c,\{c_i^-\}_{i=1}^{k} \sim C$ is the sampling process from a label distribution C. We take $k + 1$ labels from this process, and we aim to separate label c from other labels in $\{c_i^-\}_{i=1}^{k}$ to demonstrate discriminative ability of our method. $(x, x^+) \sim D_c$ denotes two observation forms (sequence or graph) of the same peptide sampled from D_c; $x_i^- \sim D_{c_i^-}$ denotes x_i^- is a negative sample of (x, x^+) from another peptide with downstream label c_i^-.

l is a convex function with respect to x_i, which can be expressed as:

$$
l\left(\{x_i\}_{i=1}^{n}\right) = -log\left(\frac{1}{1 + \sum_{i=1}^{n} e^{-x_i}} \right)
\tag{S5}
$$

According to Jensen’s Inequality, we have:

$$
\begin{aligned}
\mathcal{L}_{con} &= \mathbb{E}_{\substack{c,\{c_i^-\}_{i=1}^{k} \\ \sim C}} \left[\mathbb{E}_{\substack{(x,x^+)\sim D_c \\ x_i^- \sim D_{c_i^-}}} \left[l\left(h^T\left(h^+ - h_i^-\right)\right) \right] \right] \\
&\geq \mathbb{E}_{\substack{c,\{c_i^-\}_{i=1}^{k} \\ \sim C, x\sim D_c}} \left[l\left(\mathbb{E}_{\substack{x^+\sim D_c \\ x_i^- \sim D_{c_i^-}}} \left[h^T\left(h^+ - h_i^-\right) \right] \right) \right] \\
&= \mathbb{E}_{\substack{c^+,c^- \\ \sim C}} \mathbb{E}_{x\sim D_c} \left[l\left(h^T\left(\mu_{c^+} - \mu_{c^-}\right)\right) \right] \\
&= \lambda\, \mathbb{E}_{c,x} \left[l\left(\left\{g(h)_c - g(h)_{c'}\right\}_{c\neq c'}\right) \right]
\end{aligned}
\tag{S6}
$$

Here, $\lambda = \text{ð}\left(c' \neq c \mid c\right)$, $g\left(h\right)_c = h^T \mu_c = \left[W^{\mu} h\right]_c$, $W^{\mu} \in \text{ò}^{C \times d}$ whose c^{th} row is the mean of embedding representation with label *c* (*i.e.*, $\left[W^{\mu}\right]_c = \mu_c = \text{ã}_{x \sim D_c}\left[h\right]$). In terms of neural networks, we adopt a fully connected layer with no bias as an example downstream task predictor with the input of representation $h \in \text{ò}^{d}$. We approximate *W* as W^{μ} as *W* is optimized during training ideally. By taking the inner product between *h* (assume having label *c*) and each row of W^{μ}, the output vector has a larger value in c^{th} row, so it acts as a classifier *W*.

Thus, by minimizing contrastive loss L_{con}, the classifier W^{μ} can better discriminate *x* with label *c* from the other label c'.

Follow-up theoretical support for contrastive loss from a mutual information perspective

Theorem 2: Minimizing contrastive learning loss L_{con} can improve the information correlation between sequence representation and graph representation $I(h;\, h^{+})$.

Proof. As per **Theorem 1**, the model can distinguish representations with different downstream task labels by optimizing contrastive loss. It is reasonable to posit that $e^{\left(h,h'\right)}$ is proportional to $\frac{\text{ð}\left(h' \mid h\right)}{\text{ð}\left(h'\right)}$, where $\text{ð}\left(h' \mid h\right)$ denotes the probability that yield representation $h.\, h'$ shares the same label as *h*. The denominator $\text{ð}\left(h'\right)$ ensures the permutation invariant under h' and *h*, thereby we can reformulate the contrastive learning loss as:

$$
\begin{aligned}
\mathrm{L}_{contrastive} &= \underset{\substack{c,\left\{c_i^-\right\}_{i=1}^{k} \\ \sim C}}{\text{ã}}\ \underset{\substack{\left(x,x^+\right)\sim D_c \\ x_i^- \sim D_{c_i^-}}}{\text{ã}} \left[-log\left(\frac{e^{\left(h^T h^+\right)}}{e^{\left(h^T h^+\right)} + \sum_{i=1}^{k} e^{\left(h^T h_i^-\right)}}\right)\right] \\
&= \text{ã}\left[-log\left(\frac{\frac{\text{ð}\left(h^+ \mid h\right)}{\text{ð}\left(h^+\right)}}{\frac{\text{ð}\left(h^+ \mid h\right)}{\text{ð}\left(h^+\right)} + \sum_{i=1}^{k} \frac{\text{ð}\left(h^- \mid h\right)}{\text{ð}\left(h^-\right)}}\right)\right] \\
&\approx \text{ã}\left[log\left(1 + \frac{\text{ð}\left(h^+\right)}{\text{ð}\left(h^+ \mid h\right)} k\, \text{ã}\left[\frac{\text{ð}\left(h_i^- \mid h\right)}{\text{ð}\left(h_i^-\right)}\right]\right)\right] \\
&\overset{1}{=} \text{ã}\left[log\left(1 + \frac{\text{ð}\left(h^+\right)}{\text{ð}\left(h^+ \mid h\right)} k\right)\right] \\
&\geq \text{ã}\left[log\left(\frac{\text{ð}\left(h^+\right)}{\text{ð}\left(h^+ \mid h\right)} k\right)\right] \\
&= -I\left(h^+; h\right) + log\left(K\right)
\end{aligned}
\qquad \text{(S7)}
$$

based on the fact that h_i^- is independently sampled, *i.e.*, $\text{ə}\left(h_i^- \mid h\right) = \text{ə}\left(h_i^-\right)$.

Considering both performance and efficiency, the contrastive loss is incorporated into the model's supervised training as a regularization component. The final loss is given by:

$$\mathcal{L}_{train} = \mathcal{L}_{prd} + \lambda \mathcal{L}_{con} \qquad \text{(S8)}$$

where λ is a hyperparameter that weighs the contrastive loss component. It ensures that the model's training process is not solely dominated by the supervised loss, but also benefits from the regularizing effect of the contrastive loss, which encourages the model to learn more discriminative representations.

The pseudocode outlining the training and testing procedures of the aforementioned co-modeling framework is presented in **Figure 2** in the main text. By employing contrastive learning, we integrate the chemical information extracted by the GNN into the long-range dependencies within the amino acid sequences obtained through the attention mechanism. Upon the model is properly trained, it activates the sequence encoder and its predictor for prediction. It is important to note that, the co-modelling framework based on contrastive learning involves the joint training of two end-to-end models, however, only certain modules need to be activated during the prediction phase. In contrast, co-modelling frameworks based on other fusion methods must activate all modules. Therefore, our approach is superior in terms of time efficiency and computational cost among various implementations.

Implementations of other representation fusion methods

In addition to contrastive learning, the implementation principles of fusion module also include techniques such as WS, Concat, CA, and CBP. The fusion module taks the representations h_{seq} and h_{graph} as input, which are derived from the sequence information and chemical structure of a peptide, respectively. Let h_{co-rep} denote the output representation from fusion module, the implementation details of the baseline fusion methods are as follows:

(1) WS simply merges the input representations, denoted as:

$$h_{co-rep} = \delta h_{seq} + \left(1-\delta\right) h_{graph} \qquad \text{(S9)}$$

Where δ is a hyperparameter, set as 0.5 in our implementation.

(2) Concat directly aligns two representations, denoted as:

$$h_{co-rep} = \left[h_{seq}, h_{graph} \right] \qquad \text{(S10)}$$

(3) CA is based on multi-head attention mechanism used for aligning two representations. The mathematical expression can be denoted as:

$$h_{co-rep} = soft\,max\left(\frac{h_{graph} h_{graph}^{T}}{\sqrt{d_{h_{graph}}}} \right) h_{seq} \qquad \text{(S11)}$$

where *h_{graph}* is considered as the key and query, *h_{seq}* is considered as the value, and $d_{h_{graph}}$ is the dimension of h_{graph}. The number of heads is set as 8 in our implementation.

(4) CBP combines features from different sources or modalities, aiming to capture rich interactions. The formulation of CBP is denoted as:

$$h_{co-rep} = \mathrm{F}^{-1}\left(\mathrm{F}\left(h_{seq} \right) \odot \mathrm{F}\left(h_{graph} \right) \right) \qquad \text{(S12)}$$

Where $\mathrm{F}(\cdot)$ represents the Fourier Transform operator, $\mathrm{F}^{-1}(\cdot)$ represents the inverse Fourier Transform operator, and $\odot$ represents element-wise multiplication.

Reference

1. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020; pp. 9729-9738.