华大BGI 华大生命科学研究院 BGI·Research

量子大模型

李岳仪

2024.5.30

量子大模型优劣

华大生命科学研究院
BGI·Research

1.量子随机存取存储器
实现难

2.大模型参数非常多，量子线路
训练过程对算法要求很高

3.量子态的不可克隆原理导致数
据管理比较麻烦

1. 高维表示能力

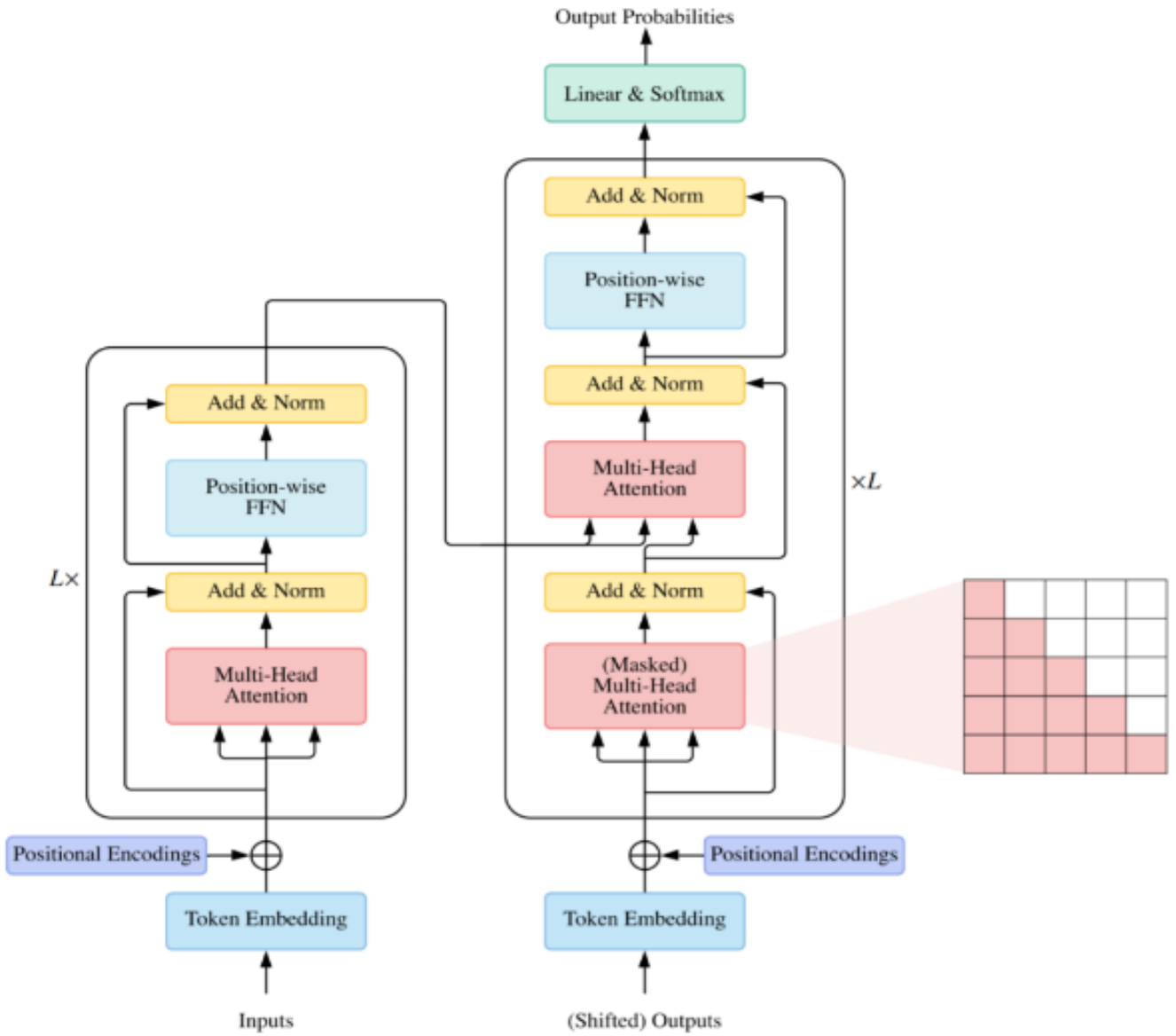
2.量子叠加和纠缠

3.降低网络复杂度

Transformer

华大生命科学研究院
BGI·Research

Transformer结构图：



Transformer重点结构

华大生命科学研究院
BGI·Research

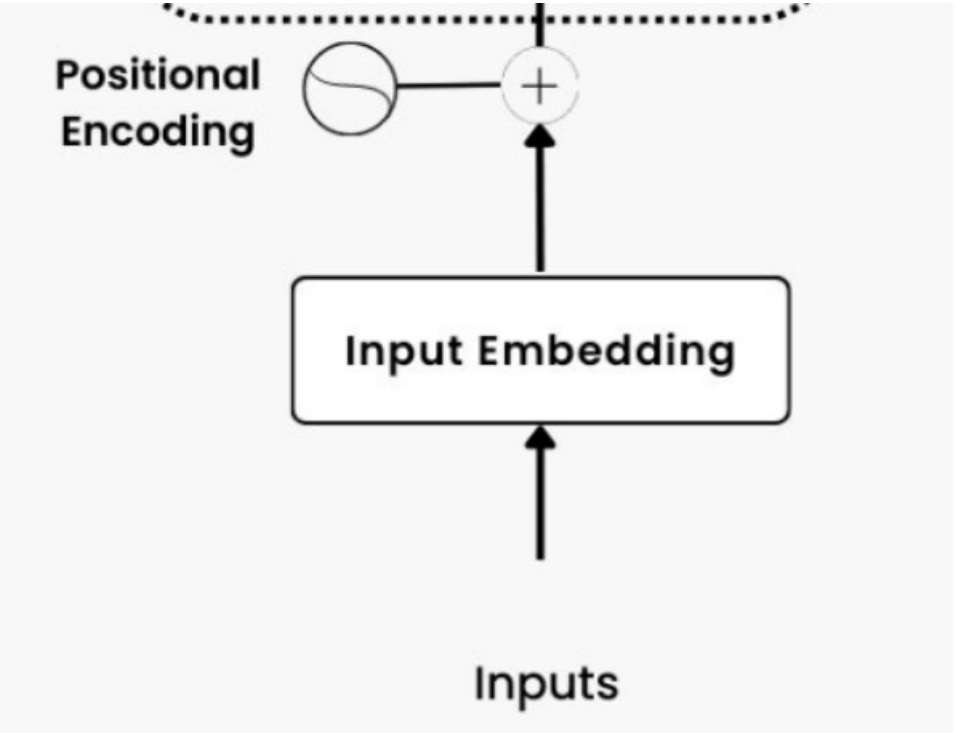
Transformer数据输入：

1.向量的位置编码

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right)$$

2.语句的词嵌入编码（用预训练的词嵌入模型输出）



Transformer 量子实现

Transformer数据输入： 量子振幅编码、量子线路参数编码、块编码

1.量子计算里对向量的位置编码处理：旋转门+量子纠缠

2.语句的词嵌入编码处理：块编码，态准备编码

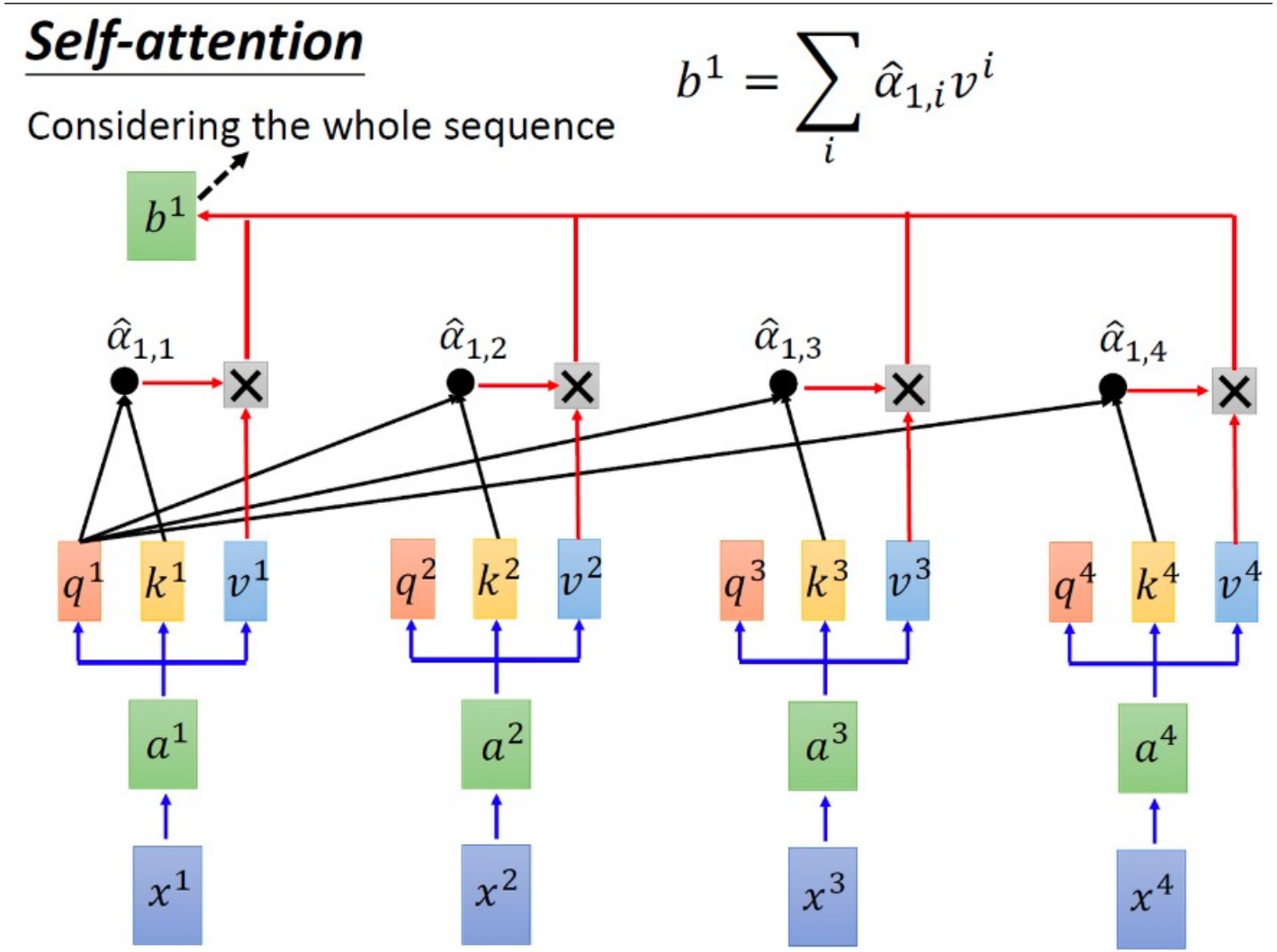
$$|\psi\rangle = U_{\text{enc}}(\mathbf{y})H^{\otimes n}|0^n\rangle$$

Transformer重点结构

Self-attention：

$$\text{Attention}(Q,K,V)=\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,...,\text{head}_\text{h})W^O$$
$$\text{where head}_\text{i}=\text{Attention}(QW_i^Q,KW_i^K,VW_i^V)$$



Transformer 量子实现

Self-attention：

1.修改注意力机制里面的注意力的计算方法

$$\alpha_{s,j} := e^{-(\langle Z_q \rangle_s - \langle Z_k \rangle_j)^2} \tag{1}$$

$$\langle Q_i | K_j \rangle := \bigoplus_h (Q_{i,h} \wedge K_{j,h}) \tag{2}$$

2.修改Q，K，V三个矩阵，用量子电路生成

$$|Q_i\rangle = U_q |W_i\rangle \tag{1}$$

$$|K_j\rangle = U_k |W_j\rangle \tag{2}$$

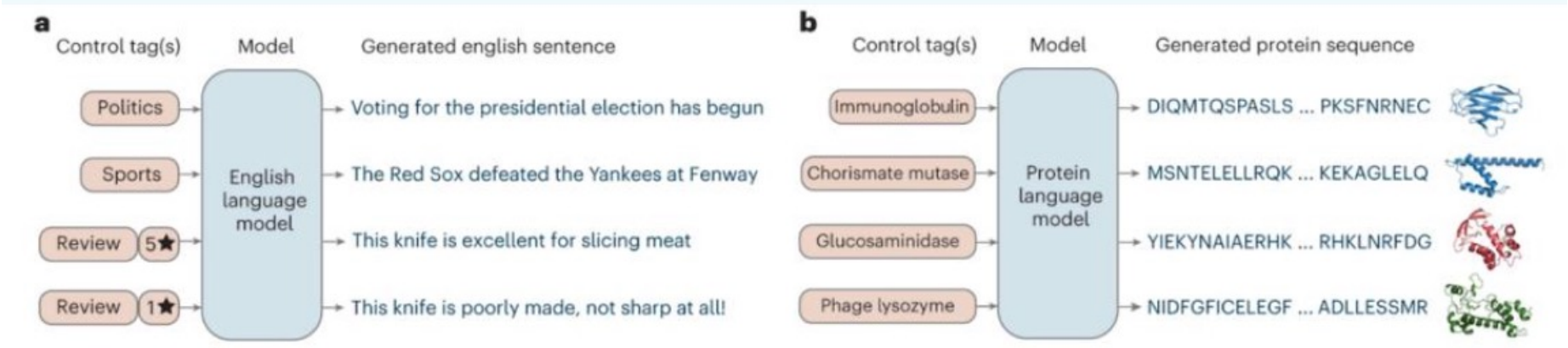
$$|V_j\rangle = U_v |W_j\rangle \tag{3}$$

大模型在蛋白质组学应用

蛋白质模型条件化：

序列条件化：2023年的 ProtGen模型

该模型可以生成特定蛋白质结构域的变体



结构条件化：Facebook提出的ESM-IF模型

大模型在蛋白质组学应用

难点 :可读性和理解性差
功能单元识别难
长度变化复杂

部分难点经典解决方案可参考：《Generative Pretrained Autoregressive Transformer Graph Neural Network applied to the Analysis and Discovery of Novel Proteins》

创新点：结合了图神经网络来解决蛋白质识别问题

量子大模型在蛋白质组学优势

可读性和理解性差：量子计算的并行处理能力，开发更高效的蛋白质序列解析算法。

功能单元识别难：量子优化算法在大规模蛋白质序列数据中识别功能单元，优化结构预测和功能识别（或者考虑量子图神经网络也行）

长度变化复杂：映射至高维希尔伯特空间中表示并且能操作大量状态，适应蛋白质序列长度的高度可变性（开发合适的量子卷积网络也行）

捕捉复杂关系：利用量子计算机模拟蛋白质序列中的量子态，捕捉氨基酸间的复杂相互作用和关系。

华大生命科学研究院
BGI·Research

THANKS

OMICS FOR ALL

基 因 科 技 造 福 人 类