



ELSEVIER

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng



Research
Artificial Intelligence—Article

基于大语言模型和量子计算开发沙门氏菌耐药性预测平台

游宇杰^a, 谭侃^a, 姜泽坤^{a,b}, 章乐^{a,b,*}

^a College of Computer Science, Sichuan University, Chengdu 610065, China

^b West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu 610065, China

ARTICLE INFO

Article history:

Received 18 October 2024

Revised 24 December 2024

Accepted 12 January 2025

Available online 28 January 2025

关键词

沙门氏菌耐药性预测

泛基因组学

大语言模型

量子计算

生物信息学

摘要

沙门氏菌是一种常见的食源性病原体,其抗微生物菌株的出现,对公共卫生安全构成了严重威胁。由于目前缺乏基于大语言模型的系统平台,难以实现沙门氏菌耐药性预测、数据呈现和数据共享。因此,为了克服以上问题,本研究首先提出了一种基于卡方检验和条件互信息最大化的两步特征选择过程,以在泛基因组学分析中找到关键的沙门氏菌耐药基因,并基于 Qwen2 大语言模型和低秩自适应开发了一种基于大语言模型的沙门氏菌抗生素耐药性预测模型 SARPLLM,实现准确的沙门氏菌耐药性预测。其次,本研究通过构建一个称为 QSMOTEN 的量子数据增强算法来优化 SMOTEN 算法的时间复杂度,以线性的计算复杂度计算样本间距离。再次,本研究建立了一个基于知识图的用户友好的沙门氏菌耐药性预测在线平台,该平台不仅便于用户进行在线耐药性预测,还能可视化呈现沙门氏菌数据集的泛基因组学分析结果。

© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. 引言

沙门氏菌是一种常见的食源性病原体,也是食源性疾病导致死亡的第三大原因[1]。尽管抗生素是治疗沙门氏菌引起的疾病的有效临床治疗方法,但它们的疗效受到基因突变和抗生素滥用的影响[2]。这些因素导致一些沙门氏菌菌株逐渐演变为耐药菌株,削弱了抗生素的治疗效果。因此,为了减少抗微生物沙门氏菌菌株对食品安全和公共卫生的影响,迫切需要为感染患者开发有针对性的抗菌治疗。

由于研究抗生素耐药性机制的过程既耗时又困难[3],因此通常用于检测细菌耐药性机制的细菌抗生素敏感性试

验效率十分低下。基于抗生素耐药性机制与细菌基因之间的密切关联,之前的研究[4–6]采用全基因组测序数据来预测沙门氏菌的耐药性。然而,由于全基因组测序数据会引入维度灾难的问题,导致目前基于机器学习和早期深度学习的预测模型[7–17]在训练过程中容易过拟合,使得预测结果难以令人满意。鉴于目前基于大语言模型的微调模型正成为强大的预测工具,可以在小样本量下实现高性能和鲁棒稳定性的预测。因此,我们提出了第一个科学问题:如何建立有效的特征选择过程,利用沙门氏菌全基因组测序数据挖掘关键的沙门氏菌耐药基因,并建立基于大语言模型的沙门氏菌耐药预测模型,以减轻小样本造成的维度灾难?

* Corresponding author.

E-mail address: zhangle06@scu.edu.cn (L. Zhang).

沙门氏菌耐药性预测也带来了另一个挑战：由于沙门氏菌全基因组测序数据中抗生素耐药性样本数量和敏感样本数量之间存在显著不平衡，抗生素耐药性预测模型的性能将显著降低。数据增强通常用于此类样本不平衡问题，因为它可以基于原始少数类样本生成伪样本以平衡样本大小。SMOTEN [18]是一种过采样数据增强算法，它寻找原始少数类样本的 k 个最近邻，然后在样本之间插值以生成新的样本。然而，SMOTEN及其衍生算法[18–20]具有较高的计算复杂性，以至于它们不适合高维全基因组测序数据。然而，量子计算的特性是其具有加速SMOTEN算法的潜力，因此我们提出了第二个科学问题：如何使用量子计算加速SMOTEN算法，以平衡沙门氏菌全基因组测序数据中的抗生素耐药性和敏感样本的数量？

尽管已经有研究基于ResFinder [21]和CARD [22]建立了在线抗生素耐药性基因分析网站，但这些网站并不能提供抗生素耐药性的预测功能。此外，现有的沙门氏菌耐药性预测模型[23–27]通常以源代码、工具包或实验流程图的形式呈现，这使得耐药性预测和分析变得困难。基于这些原因，我们提出了第三个科学问题：如何为沙门氏菌耐药性预测、数据展示和共享建立一个用户友好和方便的平台？

在本文中，我们提出了以下创新工作来解决以上科学问题：首先，我们提出了一个基于卡方检验和条件互信息最大化的两步特征选择过程，以研究泛基因组学分析中的关键沙门氏菌耐药基因，并开发了一种基于大语言模型沙门氏菌耐药性预测模型SARPLLM，以基于Qwen2 [28]大模型和低阶自适应算法(LoRA) [29]实现准确的耐药性预测；其次，我们通过构建量子数据增强算法QSMOTEN来优化从线性到对数水平的样本距离计算的时间复杂度；再次，我们基于知识图建立了一个用户友好的沙门氏菌耐药性预测在线平台[30–31]，此平台不仅便于用户进行在线耐药性预测，还可以可视化沙门氏菌数据集的泛基因组学分析结果。

为了评估我们的算法，我们在SARPLLM和其他抗生素耐药性预测模型之间进行了比较实验，结果表明SARPLLM不仅优于其他抗生素耐药预测模型，而且在沙门氏菌抗生素耐药性的预测方面也显示出高度的稳健性。然后，我们在虚拟量子机和物理量子机上模拟了所提出的QSMOTEN算法。模拟结果表明，QSMOTEN算法可以准确快速地计算沙门氏菌抗菌样本之间的距离，证明了量子计算方法加速SMOTEN算法的潜力。最后，我们构建了一个基于大语言模型沙门氏菌耐药性预测平台，为用户提供了4个关键模块，以方便他们研究沙门氏菌的耐药性。

2. 材料和方法

图1展示了本文研究方法的总流程。

2.1. 数据获取

沙门氏菌抗生素耐药性数据样本来自美国疾病控制与预防中心提供的国家抗生素耐药性监测系统数据库[32]。所有样本均为从美国患者身上分离出的鼠伤寒沙门氏菌。共有1167份沙门氏菌样本可获得抗生素耐药性和全基因组测序结果。根据抗生素耐药性结果，我们选择了所有1167个样本组成的耐药性矩阵，以指示样本是否具有耐药性。抗生素和抗生素耐药性结果的样本量见附录A中的表S1。附录A中的表S2提供了5种沙门氏菌抗生素和抗生素耐药性结果的样本量。

根据1167个沙门氏菌样本的样本号，我们从美国国家生物技术信息中心数据库中获得了沙门氏菌样本的全基因组测序数据[33]。这1167个沙门氏菌的基因注释数据是通过分级基因组组装过程[34]和美国国家生物技术信息中心数据库的原核生物基因组注释流程[35]从原始全基因组测序数据中生成的。数据采集信息详见附录A中的第S1节。

2.2. 泛基因组学分析和两步特征选择过程

图2(a)描述了泛基因组学分析的过程。首先，我们以沙门氏菌基因注释数据为输入，使用泛基因组学分析工具Roary [36]进行泛基因组学分析，以生成基因存在矩阵。然后，为了反映基因组之间的差异，我们从基因存在矩阵中删除了所有沙门氏菌基因组共享的核心基因，以获得附属基因存在矩阵。

其次，我们使用基因注释数据作为输入，通过基于快速傅里叶变换程序的多序列比对(MAFFT) [37]对沙门氏菌全基因组测序数据进行多序列比对，以获得核心基因比对数据。随后，我们使用单核苷酸多态性(SNP)位点[38]工具从核心基因比对数据中检测SNP，并最终输出核心SNP矩阵。

如图2(b)所示，为了解决维度灾难导致的抗生素耐药性预测不准确的问题，我们提出了一种基于卡方检验[39]和条件互信息最大化算法[27]的两步特征选择过程，以快速筛选出与抗生素耐药性高度相关的基因。首先，基于对5种沙门氏菌抗生素[即阿莫西林-克拉维酸钾(AUG)、头孢曲松(AXO)、氯霉素(CHL)、氨苄西林(AMP)和头孢西丁(FOX)]的耐药性，我们进行了卡方检验[式(1)]，从辅助基因和核心SNP中筛选 p 值小于0.05的沙门氏菌耐药基因。然后，我们采用条件互信息最大化算

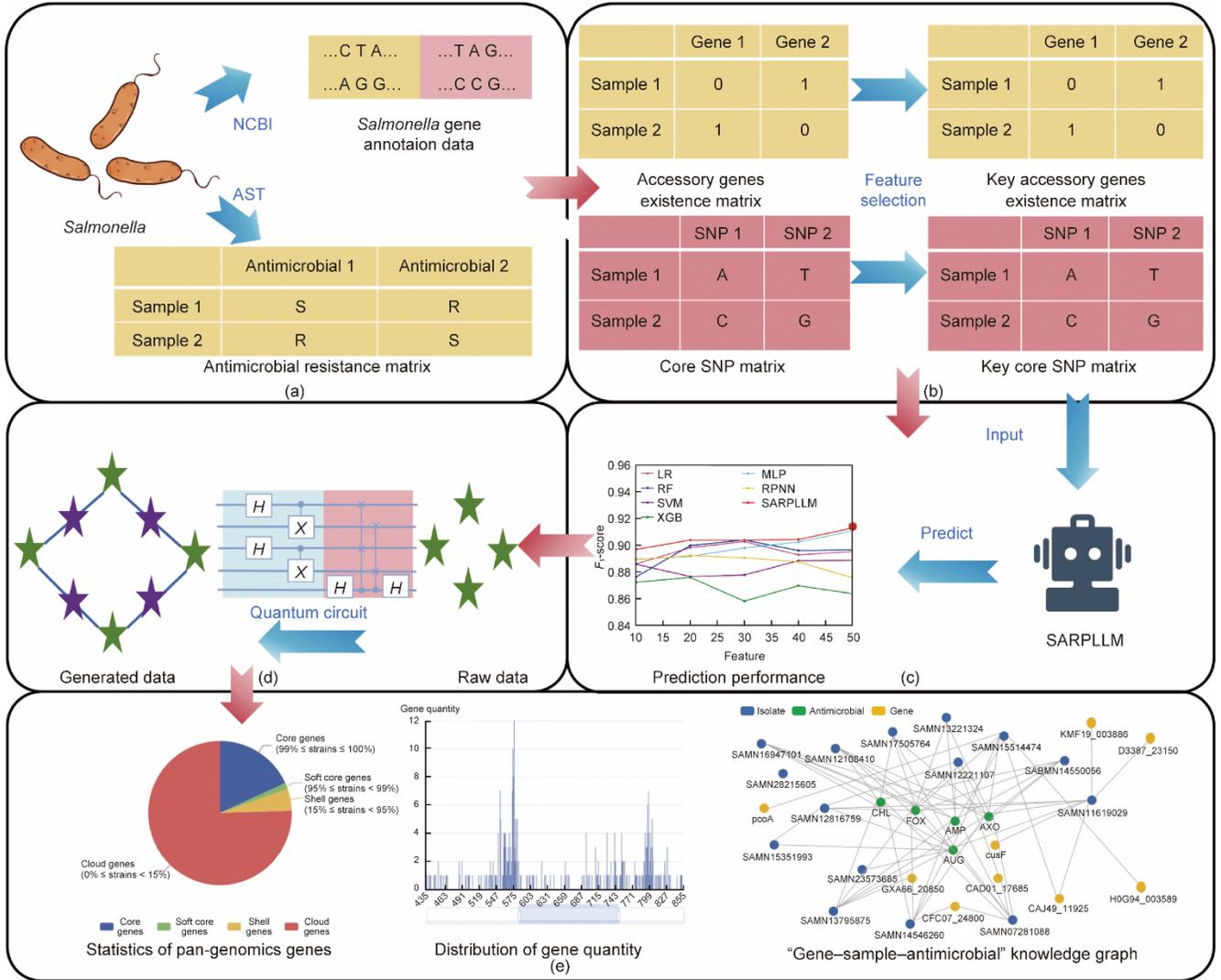


图1. 本文研究方法的总流程。(a) 数据获取；(b) 泛基因组学分析和两步特征选择过程；(c) SARPLLM模型；(d) QSMOTEN算法；(e) 沙门氏菌耐药性预测平台。A: 腺嘌呤；T: 胸腺嘧啶；G: 鸟嘌呤；C: 胞嘧啶；S: 敏感；R: 抗性；NCBI: 美国国家生物技术信息中心；SNP: 单核苷酸多态性；LR: 逻辑回归；RF: 随机森林；SVM: 支持向量机；XGB: 基于梯度提升决策树的集成学习算法；RPNN: 耐药预测神经网络模型；H: 量子Hadamard门；X: 量子Pauli-X门；AUG: 阿莫西林-克拉维酸；AXO: 头孢曲松；CHL: 氯霉素；AMP: 氨苄西林；FOX: 头孢西丁。

法[式(2)]来评估沙门氏菌耐药关键基因对每种抗生素的相对重要性。基于相对重要性，我们进一步筛选了与5种抗生素高度相关的沙门氏菌耐药基因。附录A中的第S2节详细介绍了条件信息最大化算法。

$$\chi^2(x,y) = \frac{N(AD-BC)^2}{(A+C)(A+B)(B+D)(C+D)} \quad (1)$$

$$I(x;y|z) = \text{IE}(x,z) + \text{IE}(y,z) - \text{IE}(x,y,z) - \text{IE}(z) \quad (2)$$

式中， x 和 y 分别表示来自沙门氏菌样品的基因和沙门氏菌耐药性的标签； N 表示沙门氏菌样本总数； A 表示带 x 基因的沙门氏菌抗生素耐药性的样本数量； B 表示带 x 基因的沙门氏菌抗菌敏感性样本数量； C 代表无 x 基因的沙门氏菌抗菌耐药性样本数量； D 表示无 x 基因的沙门氏菌抗微生物敏感性样本数量。 IE 表示信息熵[40]， I 表示在 z

基因条件下具有 x 基因和沙门氏菌抗生素耐药性标签 y 的沙门氏菌样本的条件互信息。

2.3. SARPLLM模型

在此，我们提出了一种SARPLLM模型，该模型使用沙门氏菌附属基因特征和从两步特征选择过程中获得的SNP特征作为输入，并学习潜在的抗生素耐药性关系以预测沙门氏菌样本的抗生素耐药性。如图3所示，SARPLLM由三个步骤组成：数据整合和提示工程、建模和微调，以及SARPLLM预测。

2.3.1. 数据整合和提示工程

提示(prompt)是大语言模型的输入文本。角色扮演是一种流行的提示工程技术，即在提示中包含了对大语言模

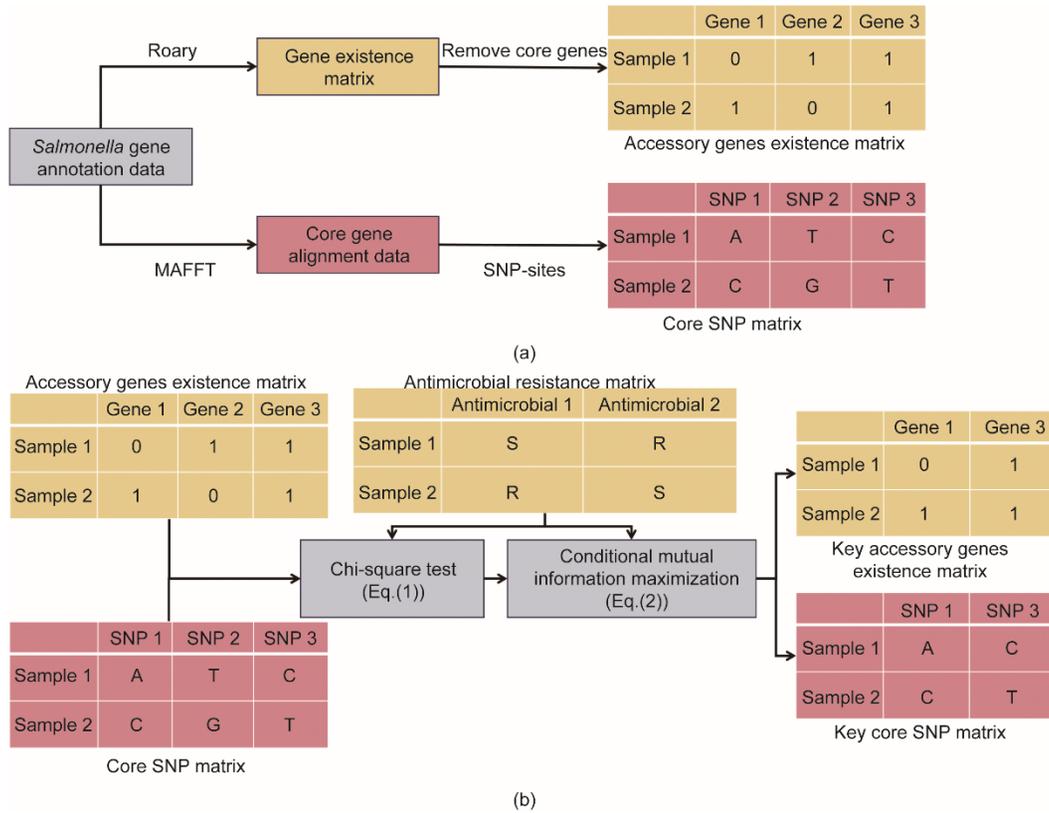


图2. 泛基因组学分析和两步特征选择的过程。(a) 泛基因组学分析的工作流程；(b) 两步特征选择的工作流程。MAFFT: 基于快速傅里叶变换的多序列比对。

型在完成时应扮演角色的描述。在本任务中，我们提供以下提示，使 SARPLLM 理解输入数据的含义和预期输出：

“Prompt”: “You are an expert in *Salmonella* antimicrobial-resistance prediction, and you will receive gene feature sequences. Please output the prediction results.”

由于大语言模型性能对自然语言输入的精确细节非常敏感，并且由于核心 SNP 特征和附属基因存在特征的模式不同，因此我们试图将这两种特征的模式转换为自然语言描述来实现数据整合。在这里，我们将核心 SNP 特征的元素值转换为 “*” “A” “G” “C” 和 “T”，从而表示沙门氏菌样本的核心 SNP 位点分别为 “缺失值” “腺嘌呤” “鸟嘌呤” “胞嘧啶” 或 “胸腺嘧啶”。我们还将辅助基因存在特征中的元素值设置为 “1” 或 “0”，以表示基因是否出现在沙门氏菌样本的基因组中。

由于先前的研究[41–42]指出，大语言模型的预测性能更多地依赖于正确的值，而不是特征名称，因此我们首先列出了需要预测的抗生素名称，然后以公式 (2) 评估得到的相对重要性顺序列出沙门氏菌耐药性特征的值。在此，我们使用空格符号来间隔不同的耐药特征。如 “Input”: “AMP: A C * 1 ...”。通过这种方式，我们将数据集转换为 SARPLLM 可以识别的句子。

2.3.2. 建模和微调

SARPLLM 是基于 Qwen2 大语言模型[28]和 LoRA [29] 建模的。更具体地说，SARPLLM 采用名为 Qwen2 [28] 的预训练大语言模型作为其基本分类器。然后，我们在沙门氏菌耐药性数据集上通过 LoRA 方法训练 SARPLLM，以专门化其预测沙门氏菌耐药性的功能，并提高其预测性能。其中，LoRA 是一种参数高效微调方法，将权重矩阵更新限制在低秩[29]。SARPLLM 优越的预测性能在于它能够利用预训练 Qwen2 中编码的广泛知识，因此仅需要少量的沙门氏菌抗生素耐药标记数据。

由于语言模型在不改变架构或损失函数的情况下对非语言任务进行微调和处理[41]，因此我们使用默认的交叉熵损失[32]来微调 SARPLLM。对于用于微调的每个训练样本，我们定义训练模板如下：

“Prompt”: “You are an expert in *Salmonella* antimicrobial-resistance prediction, and you will receive gene feature sequences. Please output the prediction results.”

“Input”: “AMP: A C * 1 ... 0”, “Output”: “1”.

2.3.3. SARPLLM 预测

在微调 SARPLLM 之后，我们从 SARPLLM 解析每个

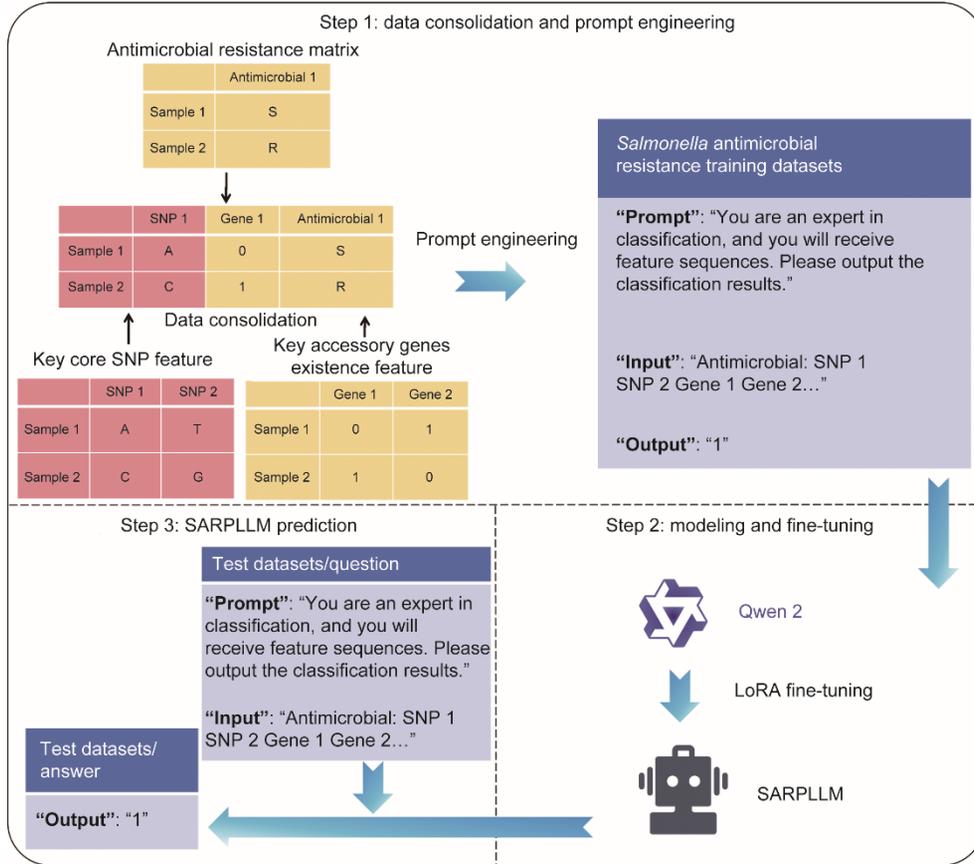


图3. SARPLLM工作流程包括三个步骤：数据整合和提示工程、建模和微调，以及SARPLLM预测。

输入样本的预测输出结果。由于大语言模型的预测性能更多地依赖于正确的值，而不是特征名称[41–42]，因此SARPLLM只需要输出“1”或“0”来指示样本是否具有耐药性或敏感性，而不需要输出“resistance”或“sensitivity”的文本字符串。例如，如果SARPLLM输出“1”，那么该样本的最终预测结果解析为具有抗生素耐药性。

2.4. QSMOTEN算法

在本节中，我们基于SMOTEN算法构建了优化算法QSMOTEN，并提供了一种线路简化和线路映射方法来实现QSMOTEN算法，从而为缓解抗生素耐药性预测中耐药和敏感沙门氏菌样本数量的显著差异提供了解决方案。

SMOTEN算法[18]是针对无序数据的SMOTE算法的改进。为了优化大规模沙门氏菌样本的SMOTEN算法的时间复杂度，本研究提出了基于量子计算的QSMOTEN算法，该算法使用相似性作为 k 近邻计算的度量，将特征的名称和值编码为量子态，并使用SWAP测试[33]线路计算样本之间的距离。QSMOTEN算法及其相关量子线路分别由算法1和图4描述。

算法1 QSMOTEN算法

Input: Original dataset $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, x_i is the sample feature, y_i is the classification label, N represents the total sample size, and i is the i th sample. m is the multiple of new sample, and $|\phi\rangle_i$ is the quantum state of the minority class sample characteristic x_i .

Output: Generated dataset G

1: Use the SWAP-test circuit to obtain the similarity between quantum states $|\phi\rangle_i$ and $|\phi\rangle_j$, and then find the top k nearest neighbor of each sample (x_i, y_i) . K is the set of nearest neighbors, and j is the j th sample.

$$K = \{(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2}), \dots, (x_{i_k}, y_{i_k})\}$$

2: Randomly select m nearest neighbors in set K to get a set M :

$$M = \{(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2}), \dots, (x_{i_m}, y_{i_m})\}$$

3: Interpolate between each sample (x_j, y_j) in (x_i, y_i) and M to obtain a new class sample y_j . α is a coefficient:

$$x_{\text{new}} = \alpha x_i + (1 - \alpha) \times x_j, \alpha \sim \text{Bernoulli}(0.5)$$

4: Add all generated samples to T and obtain generated dataset G

$$G = \{(x_1, y_1), (x_2, y_2), \dots, (x_{n \times m}, y_i)\}$$

2.4.1. 量子态编码

对于样本集 $S = \{\phi_i | i = 0, 1, \dots, (N-1)\}$ ，每个样本由 F 个特征组成，每个特征最多可以有 T 个取值 $\phi_i =$

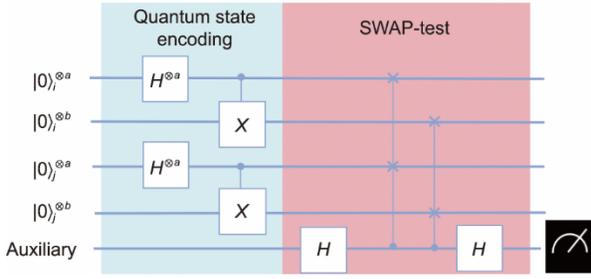


图4. QSMOTEN算法的量子线路由两部分组成：量子态编码和SWAP测试。 i : 第 i 个样本; j : 第 j 个样本; a : 量子位的数量; b : 量子位的数量。

$\{x_{ij} | j=0, 1, \dots, F-1; x_{ij}=0, 1, \dots, T-1; F>1, T>1\}$ 。这里， N 表示总样本量； ϕ_i 是样本集 S 中的第 i 个元素； x_{ij} 是 ϕ_i 中的第 j 个样本特征。QSMOTEN通过公式(3)将特征编码为以下量子态：

$$|\phi\rangle_i = \frac{1}{2^{a-1}} \sum_{v=0}^{2^a-1} |v\rangle |x_{iv}\rangle \quad (3)$$

式中， $a = \lceil \log_2 F \rceil$ ， $b = \lceil \log_2 T \rceil$ ，并且 $|x_{iv}\rangle \in \{|t\rangle | t=0, 1, 2, \dots, (b-1)\}$ 。 a 是量子位的数量， b 也是量子位的数量， v 表示第 v 个特征的位置；当 $v \geq F$ 时， $|x_{iv}\rangle = |0\rangle^{\otimes b}$ 。 x_{iv} 是无序的样本特征。 $|\phi\rangle_i$ 是样本 ϕ_i 的量子态。

在这里，我们展示了量子态的制备过程：

(1) 我们将量子态初始化为 $|0\rangle^{\otimes(a+b)}$ 。

(2) 我们将Hadamard门(H)应用于 a 位量子比特处[公式(4)]，以表示特征名称的编码。例如， $|0110\rangle$ 表示第六个(二进制下用0110表示)特征。

$$H^{\otimes a} |0\rangle^{\otimes(a+b)} = \frac{1}{2^{a-1}} \sum_{v=0}^{2^a-1} |v\rangle |0\rangle^{\otimes b} \quad (4)$$

(3) 我们应用多重控制NOT门(MCX)将 a 位量子比特设置为控制位。然后，我们将Pauli- X 门应用于 b 位量子比特以反转其值。量子态的结果可以用公式(5)表示。例如，当 a 位量子比特表示 $|v\rangle$ 时， b 位量子比特被反转为 $|x_{iv}\rangle$ 。

$$|\phi\rangle_i = \text{MCX} \left(\frac{1}{2^{a-1}} \sum_{v=0}^{2^a-1} |v\rangle |0\rangle^{\otimes b} \right) = \frac{1}{2^{a-1}} \sum_{v=0}^{2^a-1} |v\rangle |x_{iv}\rangle \quad (5)$$

2.4.2. 沙门氏菌样本之间的相似性

QSMOTEN算法通过计算样本之间的相似度来找到前 k 个最相似的邻居样本。相似性定义如下：

$$\text{Sim}(\phi_i, \phi_j) = \sum_{k=0}^{F-1} (x_{ik} \odot x_{jk}) \quad (6)$$

式中， x_{iv} 为无序样本特征； \odot 表示等价操作； ϕ_j 为样本集 S 中的第 j 个元素。

如图4所示，我们采用SWAP测试线路，通过以下三个步骤计算相似度[公式(6)]：

步骤1: 该线路以两个量子态 $|\phi\rangle_i$ 和 $|\phi\rangle_j$ 使用相同数量的量子比特作为输入，同时将一个辅助量子比特初始化为 $|0\rangle$ 。

步骤2: 该线路将Hadamard门(H)应用于辅助量子比特，并在辅助量子比特为 $|1\rangle$ 时，应用Fredkin门(CSWAP)以交换 $|\phi\rangle_i$ 和 $|\phi\rangle_j$ 。

步骤3: 该线路再次对辅助量子比特施加Hadamard门(H)，然后对辅助量子比特进行测量。

如附录A中的第S3节所示，测量结果的概率与相似性之间的关系可以用公式(7)来描述：

$$\text{Sim}(\phi_i, \phi_j) = F - 2^a \left(1 - \sqrt{1 - 2P(|1\rangle)} \right) \quad (7)$$

这表明两个样本之间的相似性随着测量概率 $P(|1\rangle)$ 的降低而增加。因此，两个样本之间的相似性可以通过测量概率 $P(|1\rangle)$ 来确定。这里， P 表示测量得到某种状态的概率。

2.4.3. QSMOTEN算法的时间复杂度分析

假设总共有 N 个样本，每个样本由 F 个序列特征组成，并且每个序列特征可以有 T 个可选值。SMOTEN算法需要计算每对样本之间的距离共 $N^2/2$ 次，每次计算都需要进行比较 F 次。因此，距离计算的时间复杂度为 $O(N^2F)$ 。

QSMOTEN算法首先将每个样本编码为量子长度为 $\lceil \log_2 F \rceil + \lceil \log_2 T \rceil$ 的量子态，因此所需时间复杂度为 $O(N^2(\log F + \log T))$ 。然后，SMOTEN算法使用带有 $(\lceil \log_2 F \rceil + \lceil \log_2 T \rceil)$ 个Fredkin门的SWAP测试线路来计算样本之间的距离。这个过程的时间复杂度为 $O(N^2(\log F + \log T))$ 。因此，QSMOTEN算法距离计算的总时间复杂度为 $O(N^2(\log F + \log T))$ 。

基因样本通常具有 $T=4$ 种特征，其中，每个特征的值是“A”“T”“G”或“C”。所以， $\log T$ 是一个具有较小值的常数，通常会被忽略。因此，与SMOTEN算法相比，QSMOTEN算法可以将时间复杂度由 $O(N^2F)$ 降低到 $O(N^2(\log F + \log T))$ 。

2.5. 沙门氏菌耐药性预测平台

基于泛基因组学分析结果和SARPLLM模型，我们建立了一个基于知识图谱[29]的沙门氏菌耐药性预测在线平台。沙门氏菌耐药性预测在线平台采用Django[43]作为后端服务架构，以实现对用户访问的监测和响应。前端使用

Echart进行知识图可视化。沙门氏菌耐药性预测在线平台可以根据用户上传的沙门氏菌基因特征文件对多个抗生素耐药性进行在线预测。此外，我们的平台不仅显示沙门氏菌数据集的泛基因组学分析结果，还为用户提供了一种下载原始和分析数据的便捷方式。该平台有4个主要模块。

(1) **抗生素耐药性预测模块：**该模块为用户提供了一个上传沙门氏菌基因特征文件的界面；并且，该模块还提供 SARPLLM 模型进行沙门氏菌耐药性在线预测和结果可视化的服务。

(2) **泛基因组学分析结果模块：**该模块将沙门氏菌抗生素耐药性数据集的泛基因组学分析统计结果可视化。

(3) **基因样本抗菌知识图模块：**该模块构建了一个有向图，描述基因、样本和抗生素之间的关系，并可视化它们之间的关系。

(4) **数据下载模块：**该模块提供原始数据和泛基因组学分析数据的下载功能。

3. 结果

3.1. 抗生素耐药性预测模型的实验结果

为了回答我们提出的第一个科学问题，我们展示了泛基因组学分析过程、两步特征选择过程和抗生素耐药性预测模型比较实验的结果。首先，通过执行第2.2节中描述的泛基因组学分析过程，我们在附录A的第S4.1节中展示获取的附属基因存在矩阵。其次，我们在附录A的第S4.2节中展示获取的核心SNP矩阵。然后，通过两步特征选择，我们筛选了分别对应于每种抗生素（即AUG、AXO、CHL、AMP和FOX）的沙门氏菌耐药基因的强相关特征。这些相关特征在附录A的第S4.3节和第S4.4节中呈现。通过执行两步特征选择算法，我们选择前5个重要基因特征在附录A的表S3中展示。再次，我们使用SARPLLM、逻辑回归（LR）、随机森林（RF）[44]、极限梯度增强（XGBoost）[45]、支持向量机（SVM）[46]、多层感知器（MLP）和耐药性预测神经网络（RPNN）模型（详见附录A中的第S5节）对5种抗生素（AUG、AXO、CHL、AMP和FOX）进行抗生素耐药性预测。LR、RF、XGBoost、SVM和RPNN的配置参数见附录A中的第S6.1节和第S6.2节。

SARPLLM采用批处理学习，其中，批处理大小为4，且训练周期为4。训练使用AdamW优化器[47]，学习率为 10^{-5} 。学习调度器类型设置为多项式。SARPLLM模型的更多架构参数在附录A中的第S6.3节呈现。考虑到沙门氏菌数据集中耐药和敏感抗菌标签的不平衡，以及大语言模

型在预测任务中常用的综合评价指标，我们使用 F_1 评分来评估模型的预测性能。

对于每种抗生素，我们分别将前 $N=10, 20, 30, 40, 50$ 个与沙门氏菌耐药性相关的特征作为输入。然后，我们进行4次重复测试的三重交叉验证，并统计确定了预测指标的平均值。图5显示了AUG、AXO、CHL、AMP和FOX的预测结果。

图5显示，在只有10个耐药特征时，所有模型的预测性能都大于85%。此外，随着耐药特征的数量从10个逐渐增加到50个，所有预测模型的评估指标都会出现波动，然后趋于稳定，这表明我们提出的两步特征选择方法可以准确地筛选和保留与抗生素耐药性相关的特征，从而减小输入所需的耐药特征的数量，同时保持所有模型的高性能预测能力。

为了进一步分析SARPLLM与其他模型之间的性能差异，我们针对5种抗生素，对具有50个输入耐药特征的特征的SARPLLM与其他模型之间进行了 T 检验[48–50]。预测结果的统计信息记录在附录A中的第S7节。检验结果见表1和附录A中的表S4。表1显示，在大多数情况下，SARPLLM的 F_1 评分明显优于LR、RF、XGB、SVM、MLP和PRNN模型（ $p < 0.05$ ），表明我们提出的SARPLLM模型在多个抗生素数据集下，其抗生素耐药性预测方面具有显著的性能优势，因此具有较好的泛化能力。此外，由于我们的输入数据中存在一定量的缺失值，但SARPLLM仍然具有最佳的预测性能，这意味着SARPLLM能够应对具有缺失数据的情况，具有出色的鲁棒性。

3.2. QSMOTEN的实验结果

为了回答我们提出的第二个科学问题，我们分别使用虚拟量子机和物理量子机展示了实验结果。

3.2.1. 在虚拟机上的实验结果

我们假设有4个样本 ϕ_i ($i \in [0-3]$)。每个样本有4个特征，每个特征的值为“A”“T”“G”或“C”。本实验以 $\phi_0 = \text{“ATCG”}$ 、 $\phi_1 = \text{“ATGC”}$ 、 $\phi_2 = \text{“AAGT”}$ 和 $\phi_3 = \text{“TATA”}$ 为例，验证QSMOTEN算法的有效性。

基于上述假设，我们提供了一个量子线路（图6）来计算 ϕ_0 和 ϕ_1 之间的相似性。附录A中的第S8节提供了计算剩余样本对（ $\phi_0\phi_2$ 、 $\phi_0\phi_3$ 、 $\phi_1\phi_2$ 、 $\phi_1\phi_3$ 和 $\phi_2\phi_3$ ）之间对比相似性的量子线路。对于每个样本对，我们为每个实验设置了 10^4 次测量。Qiskit [51]使用“aer_simulator”模拟器在理想条件下模拟了实验。我们采用公式（7）计算相似性，模拟结果记录在表2中。

表2显示，QSMOTEN计算的相似度接近样本对的实

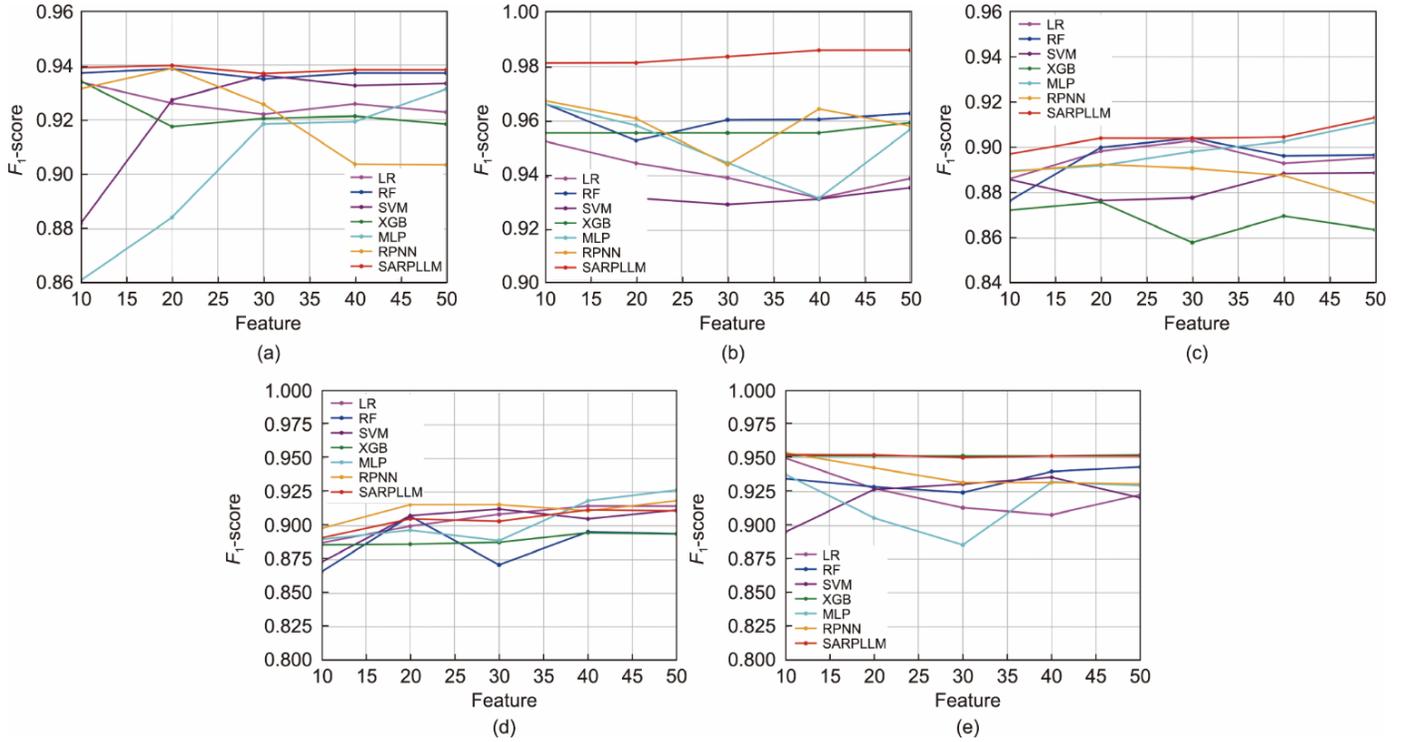


图5. 在5种抗生素下, 沙门氏菌耐药性预测的 F_1 评分预测结果。(a) AUG的预测结果; (b) AXO的预测结果; (c) CHL的预测结果; (d) AMP的预测结果; (e) FOX的预测结果。

表1 AUG、AXO和CHL抗生素在具有50个特征的预测模型上的 T 检验结果

Model	T -test results for AUG			T -test results for AXO			T -test results for CHL		
	Mean	Variance	p value	Mean	Variance	p value	Mean	Variance	p value
LR	0.923	0.002	4.0×10^{-5}	0.939	0.023	1.8×10^{-2}	0.895	0.003	8.3×10^{-4}
RF	0.937	0.002	4.0×10^{-1}	0.962	0.013	4.5×10^{-2}	0.897	0.012	7.8×10^{-2}
SVM	0.933	0.002	8.1×10^{-3}	0.936	0.009	1.3×10^{-3}	0.889	0.016	5.3×10^{-2}
MLP	0.931	0.009	2.2×10^{-1}	0.957	0.011	1.1×10^{-2}	0.911	0.008	6.7×10^{-1}
XGB	0.918	0.008	1.2×10^{-2}	0.959	0.004	1.2×10^{-3}	0.864	0.010	1.2×10^{-3}
RPNN	0.903	0.021	4.5×10^{-2}	0.958	0.008	5.5×10^{-3}	0.876	0.012	4.2×10^{-3}
SARPLLM	0.938	0.002	—	0.985	0	—	0.913	0.004	—

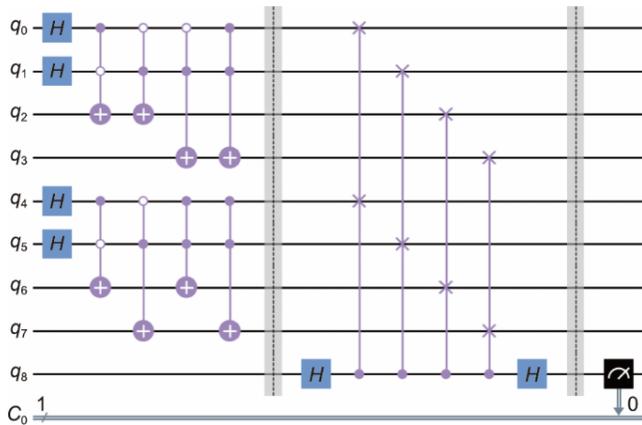


图6. 在虚拟机上执行QSMOTEN的量子线路, 以计算 ϕ_0 ="ATCG"和 ϕ_1 ="ATGC"的相似性。量子位 q_0 - q_3 编码样本 ϕ_0 的特征, 量子位 q_4 - q_7 编码样本 ϕ_1 的特征。量子位 q_8 是一个辅助量子位。 C_0 是一个用于存储测量结果的经典位寄存器。

表2 QSMOTEN计算的样本对 $\phi_0\phi_1$ 、 $\phi_0\phi_2$ 、 $\phi_0\phi_3$ 、 $\phi_1\phi_2$ 、 $\phi_1\phi_3$ 和 $\phi_2\phi_3$ 之间的相似性

Sample pairs	$P(1)$	$\text{Sim}(x_i, x_j)$	Actual similarity
(ϕ_0, ϕ_1)	0.3769	1.982	2
(ϕ_0, ϕ_2)	0.4706	1.075	1
(ϕ_0, ϕ_3)	0.4997	0.233	0
(ϕ_1, ϕ_2)	0.3740	2.025	2
(ϕ_1, ϕ_3)	0.4981	0.187	0
(ϕ_2, ϕ_3)	0.4652	1.088	1

际相似度, 这表明我们提出的QSMOTEN算法可以正确计算样本之间的相似度, 为量子计算机进行数据增强提供了一种有效的方法。

3.2.2. 在量子物理机上的实验结果

由于量子计算机可用的硬件支持有限，目前物理量子机的保真度较低。因此，我们采用 ϕ_0 ="A"、 ϕ_1 ="T" 和 ϕ_2 ="A" 作为示例，以验证 QSMOTEN 算法的有效性。这三个样本以以下形式编码： $|\phi\rangle_0 = |\phi\rangle_2 = |0\rangle$ 和 $|\phi\rangle_1 = |1\rangle$ 。基于上述假设，我们开发了图 7 (a) 和 (b) 所示的量子线路，分别计算 ϕ_0 和 ϕ_1 之间的相似度以及 ϕ_0 和 ϕ_2 之间的相似性。

对于每个样本对，我们设置了 10^4 次测量次数。实验在“骁鸿”量子计算机上进行测量。“骁鸿”量子计算机的量子处理器参数详见附录 A 中的第 S9 节。图 7 (c) 和 (d) 分别显示了样本 ϕ_0 ="A" 和 ϕ_1 ="T" 相似性的测量结果，以及样本 ϕ_0 ="A" 和 ϕ_2 ="A" 的测量结果。图 7 (c) 显示了量子线路[图 7 (a)]的测量结果分别为 $P(|0\rangle) = 0.4808$ 和 $P(|1\rangle) = 0.5192$ 。将测量结果代入方程式 (7)，计算得到 ϕ_0 ="A" 和 ϕ_1 ="T" 的相似度为 0，这与实际相似度 0 相同。图 7 (d) 显示了量子线路[图 7 (b)]的测量结果分别为 $P(|0\rangle) = 0.8018$ 和 $P(|1\rangle) = 0.1982$ 。通过将测量结果代入方程式 (7)，发现样本 ϕ_0 ="A" 和 ϕ_2 ="A" 之间的相似度为 0.7769，接近实际相似度 1。因此，“骁鸿”量子计算机上的实验结果表明，QSMOTEN 算法可以正确计算样本对之

间的相似性，证明了量子计算方法在量子物理机器上加速 SMOTEN 算法的潜力。

3.3. 沙门氏菌耐药性预测在线平台

为了回答第三个科学问题，我们建立了一个基于知识图谱的沙门氏菌耐药性预测在线平台。该平台提供 4 种在线服务：耐药性预测模块、泛基因组学分析结果模块、基因-样本-抗生素知识图谱模块，以及数据下载模块。图 8 (a) 展示了抗生素耐药性预测模块，该模块有两个功能：一个是选择文件以指定格式上传沙门氏菌基因特征文件；另一个是预测并可视化抗生素耐药性预测的结果。图 8 (b) 展示了泛基因组学分析结果模块，该模块展示了沙门氏菌泛基因组中基因的数量和分布。用户可以将鼠标移动到饼图中的某一基因上，以获取该基因的详细信息。此外，用户可以将鼠标移动到直方图上，以确定基因组中存在多少泛基因组中的基因。图 8 (c) 展示了基因-样本-抗生素知识图谱模块。图中的蓝点代表菌株样本实体，绿点代表抗生素实体，黄点代表基因实体，线条代表实体之间的关系，线条上的文本标签描述了关系的类型。用户可以将鼠标移动到实体或关系上，以突出显示该实体或相邻关系。此外，当鼠标选择时，实体的属性也会显示出来。图 8 (d) 显示了数据下载模块，用户可以通过点击“下载”按钮获得相应的数据。

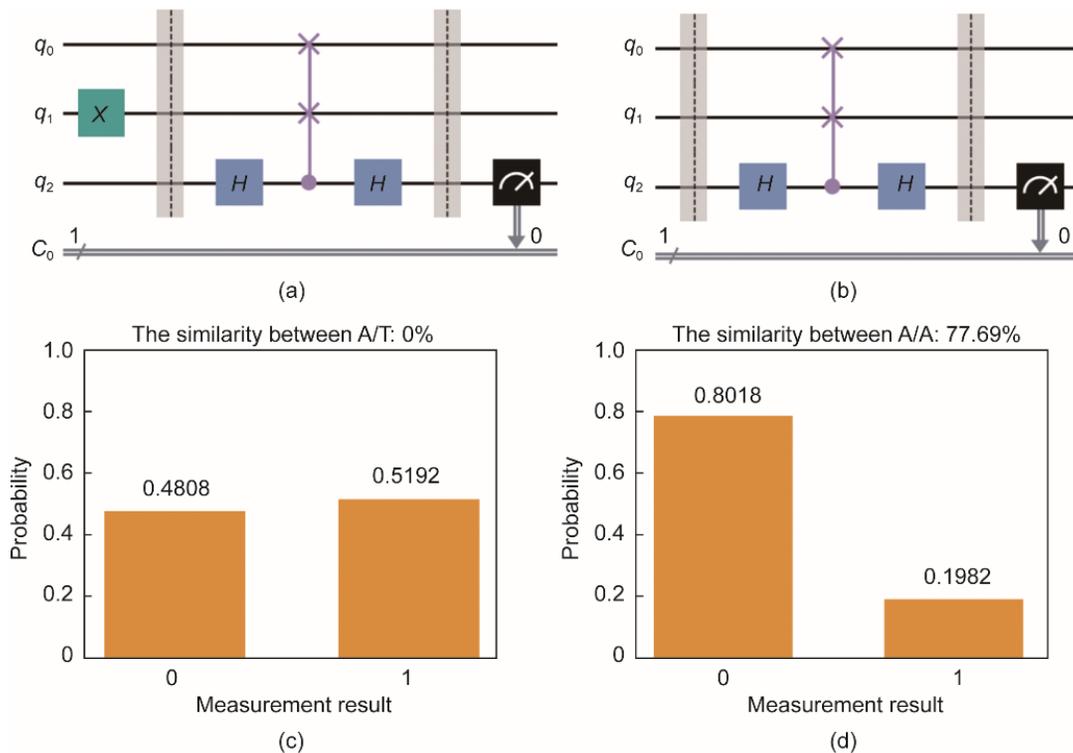


图 7. QSMOTEN 算法的量子线路用于计算 ϕ_0 ="A" 和 ϕ_1 ="T" 之间的相似性 (a)、 ϕ_0 ="A" 和 ϕ_2 ="A" 之间的相似性 (b)；(c) ϕ_0 ="A" 和 ϕ_1 ="T" 之间的相似性的测量结果；(d) ϕ_0 ="A" 和 ϕ_2 ="A" 之间的相似性的测量结果。

理大规模高维全基因组序列数据中，存在的计算效率低下的问题，本研究提出了基于 SMOTEN 算法和 SWAP 测试量子线路的 QSMOTEN。QSMOTEN 算法的时间复杂度分析表明，该算法将计算样本之间距离的关键步骤的时间复杂度从线性水平 $O(N^2F)$ 降低到对数水平 $O(N^2(\log F + \log T))$ 。此外，在虚拟机（表2）和物理机[图7（c）和（d）]上的模拟实验表明，QSMOTEN 算法既可以准确计算样本之间的距离，又可以通过量子计算加速 SMOTEN 算法。

为了解决目前缺乏专门的沙门氏菌耐药性预测在线平台的问题，我们在研究中通过整合网络技术和知识图谱技术构建了一个大语言模型平台。该平台由4个模块组成：抗生素耐药性预测模块、泛基因组学分析结果模块、基因-样本-抗生素知识图谱模块以及数据下载模块。该平台不仅为用户提供方便的沙门氏菌耐药性预测服务，还将泛基因组学分析结果可视化。此外，该平台采用知识图谱技术以高可扩展性存储沙门氏菌基因组数据和抗生素耐药性数据。

尽管我们在研究中基于大语言模型和量子计算开发了沙门氏菌抗生素耐药性的预测平台（代码在附录A中的第S10节），但这项工作有两个缺点。首先，沙门氏菌耐药性预测涉及复杂的生物学和遗传学知识，目前的大语言模型很难完全理解和准确地表示这些知识，这降低了预测的准确性。此外，大语言模型的性能在很大程度上取决于训练数据集的质量和数量。沙门氏菌耐药性预测领域可用的高质量 and 多样化的数据集不足，限制了大语言模型的预测性能。其次，量子计算技术仍处于早期发展阶段，量子计算受到量子硬件处理器能力的限制。大多数量子计算算法只能进行数学分析或在量子计算机的高性能模拟器上运行。因此，在这些算法能够在物理机器上验证并应用于工程之前，还有很长的路要走。

总之，我们未来的研究将侧重于整合多源数据和领域知识，以提高基于大语言模型的沙门氏菌抗生素耐药性预测平台的准确性。我们还将尝试开发更稳定可靠的量子硬件，以增加量子计算机在沙门氏菌耐药性数据增强中的应用。

CRediT authorship contribution statement

Yujie You: Writing-review & editing, Writing-original draft, Validation, Methodology. **Kan Tan:** Writing-original draft, Visualization, Investigation. **Zekun Jiang:** Supervision, Resources, Investigation. **Le Zhang:**

Writing-review & editing, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

致谢

本研究得到了国家科技重大专项(2021YFF1201200)、国家自然科学基金(62372316)和四川省科技计划重点项目(2024YFHZ0091)的支持。我们感谢 QuantumCTek 提供的“骁鸿”量子计算机。

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.eng.2025.01.013>.

References

- [1] Ferrari RG, Rosario DKA, Cunha-Neto A, Mano SB, Figueiredo EES, Conte CA, et al. Worldwide epidemiology of *Salmonella* serovars in animal-based foods: a meta-analysis. *Appl Environ Microbiol* 2019;85(14):e00591–19.
- [2] Qin XJ, Yang MZ, Cai H, Liu YT, Gorris L, Aslam MZ, et al. Antibiotic resistance of *Salmonella typhimurium* monophasic variant 1, 4, 5, 12: i- in China: a systematic review and meta-analysis. *Antibiotics* 2022;11(4):532.
- [3] Anahtar MN, Yang JH, Kanjilal S. Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *J Clin Microbiol* 2021;59(7):e01260–20.
- [4] Botelho J, Schulenburg H. The role of integrative and conjugative elements in antibiotic resistance evolution. *Trends Microbiol* 2021;29(1):8–18.
- [5] Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. *Mol Syst Biol* 2013;9:640.
- [6] Su M, Satola SW, Read TD. Genome-based prediction of bacterial antibiotic resistance. *J Clin Microbiol* 2019;57(3):e01405–18.
- [7] Wang CC, Hung YT, Chou CY, Hsuan SL, Chen ZW, Chang PY, et al. Using random forest to predict antimicrobial minimum inhibitory concentrations of nontyphoidal *Salmonella* in Taiwan. *Vet Res* 2023;54(1):11.
- [8] Ren Y, Chakraborty T, Doijad S, Falgenhauer L, Falgenhauer J, Goesmann A, et al. Deep transfer learning enables robust prediction of antimicrobial resistance for novel antibiotics. *Antibiotics* 2022;11(11):1611.
- [9] Gao J, Lao QH, Liu P, Yi HH, Kang QB, Jiang ZK, et al. Anatomically guided cross domain repair and screening for ultrasound fetal biometry. *IEEE J Biomed Health Inform* 2023;27(10):4914–25.
- [10] Lai X, Zhou J, Wessely A, Heppt M, Maier A, Berking C, et al. A disease network based deep learning approach for characterizing melanoma. *Int J Cancer* 2022;150(6):1029–44.
- [11] Song H, Chen L, Cui Y, Li Q, Wang Q, Fan J, et al. Denoising of MR and CT images using cascaded multi-supervision convolutional neural networks with progressive training. *Neurocomputing* 2022;469:354–65.
- [12] Zhang Q, Zhang H, Zhou K, Zhang L. Developing a physiological signal-based, mean threshold and decision-level fusion algorithm (PMD) for emotion

- recognition. *Tsinghua Sci Technol* 2023;28(4):673–85.
- [13] Zhang L, Song W, Zhu T, Liu Y, Chen W, Cao Y, et al. ConvNeXt-MHC: improving MHC-peptide affinity prediction by structure-derived degenerate coding and the ConvNeXt model. *Brief Bioinform* 2024;25(3):bbae133.
- [14] Jiang Z, Cheng D, Qin Z, Gao J, Lao Q, Li K, et al. TV-SAM: increasing zero-shot segmentation performance on multimodal medical images using GPT-4 generated descriptive prompts without human annotation. *Big Data Min Anal* 2024;7(4):1199–211.
- [15] Gao J, Lao Q, Kang Q, Liu P, Du C, Li K, et al. Boosting your context by dual similarity checkup for in-context learning medical image segmentation. *IEEE Trans Med Imaging* 2025;44(1):310–9.
- [16] You Y, Zhou F, Yue Y. The classical iterative HHL-based hemodynamic simulation quantum linear equation algorithm for abdominal aortic aneurysm. *Eur Phys J Spec Top*. In press.
- [17] Xiao M, Wei R, Yu J, Gao C, Yang F, Zhang L, et al. CpG island definition and methylation mapping of the T2T-YAO genome. *Genom Proteom Bioinform* 2024;22(2):zqae009.
- [18] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;16:321–57.
- [19] He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*; 2008 Jun 1–8; Hong Kong, China; 2008. p. 1322–8.
- [20] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: *Proceedings of the International Conference on Intelligent Computing*; 2005 Aug 23–26; Hefei, China. Berlin: Springer Nature; 2005. p. 878–87.
- [21] Zankari E, Hasman H, Cosentino S, Vestergaard A, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012;67:2640–4.
- [22] McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 2013;57:3348–57.
- [23] Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L, et al. Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan genome data. *PLoS Comput Biol* 2018;14(12):e1006258.
- [24] Ha SM, Lin EY, Klausner JD, Adamson PC. Machine learning to predict ceftriaxone resistance using single nucleotide polymorphisms within a global database of *Neisseria gonorrhoeae* genomes. *Microbiol Spectr* 2023; 11(6): e0170323.
- [25] Yang Y, Walker TM, Kouchaki S, Wang C, Peto TEA, Crook DW, et al. An end-to-end heterogeneous graph attention network for *Mycobacterium tuberculosis* drug-resistance prediction. *Brief Bioinform* 2021;22(6):bbab29.
- [26] Jiang Z, Lu Y, Liu Z, Wu W, Xu X, Dinnyés A, et al. Drug resistance prediction and resistance genes identification in *Mycobacterium tuberculosis* based on a hierarchical attentive neural network utilizing genome-wide variants. *Brief Bioinform* 2022;23(3):bbac041.
- [27] Shi JH, Yan Y, Links MG. Antimicrobial resistance genetic factor identification from whole-genome sequence data using deep feature selection. *BMC Bioinformatics* 2019;20(Suppl 15):535.
- [28] Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen technical report. 2023. arXiv:2309.16609.
- [29] Ma F, Xiao M, Zhu L, Jiang W, Jiang J, Zhang PF, et al. An integrated platform for Brucella with knowledge graph technology: from genomic analysis to epidemiological projection. *Front Genet* 2022;13:981633.
- [30] Zhang L, Dai Z, Yu J, Xiao M. CpG-island-based annotation and analysis of human housekeeping genes. *Brief Bioinform* 2021;22(1):515–25.
- [31] Zhang L, Zhang L, Guo Y, Xiao M, Feng L, Yang C, et al. MCDB: a comprehensive curated mitotic catastrophe database for retrieval, protein sequence alignment, and target prediction. *Acta Pharm Sin B* 2021;11(10):3092–104.
- [32] Kline DM, Berardi VL. Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Comput Appl* 2005;14(4):310–8.
- [33] Barenco A, Berthiaume A, Deutsch D, Ekert AK, Jozsa R, Macchiavello C, et al. Stabilization of quantum computations by symmetrization. *SIAM J Comput* 1997;26:1541–57.
- [34] Chin CS, Alexander D, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013;10(6):563–9.
- [35] Tatusova T. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* 2016;44(14):6614–24.
- [36] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31(22):3691–3.
- [37] Katoh K, Misawa K, Ki K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;30(14):3059–66.
- [38] Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2016;2(4):e000056.
- [39] Dong X, Sun F, Han X, Hou R. Study of positive and negative association rules based on multi-confidence and chi-squared test. In: Li X, Zaïane OR, Li Z, editors. *Advanced data mining and applications*. Berlin: Springer Nature; 2006. p. 100–9.
- [40] Liang J, Shi Z, Li D, Wierman MJ. Information entropy, rough entropy and knowledge granulation in incomplete information systems. *Int J Gen System* 2006;35:641–54.
- [41] Dinh T, Zeng Y, Zhang R, Lin Z, Gira M, Rajput S, et al. LIFT: language-interfaced fine-tuning for non-language machine learning tasks. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*; 2022 Nov 28–Dec 9; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2022. p. 11763–84.
- [42] Hegselmann S, Buendia A, Lang H, Agrawal M, Jiang X, Sontag D, et al. TabLLM: few-shot classification of tabular data with large language models. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*; 2023 Apr 25–27; Valencia, Spain. PMLR. p. 5549–58.
- [43] Putnam J. Python Web development with Django. *Comput Rev* 2010;51(6):330.
- [44] Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
- [45] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016 Aug 13–17; FranciscoSan, CA, USA. New York City: Association for Computing Machinery (ACM); 2016. p. 785–94.
- [46] Cortes C, Vapnik VN. Support-vector networks. *Mach Learn* 1995;20:273–97.
- [47] Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *Proceedings of the International Conference on Learning Representations*; 2019 May 6–9; New Orleans, LA, USA. Wadern: dblep; 2019.
- [48] Xia Y, Yang C, Hu N, Yang Z, He X, Li T, et al. Exploring the key genes and signaling transduction pathways related to the survival time of glioblastoma multiforme patients by a novel survival analysis model. *BMC Genomics* 2017; 18(Suppl 1):950.
- [49] Zhang L, Liu G, Kong M, Li T, Wu D, Zhou X, et al. Revealing dynamic regulations and the related key proteins of myeloma-initiating cells by integrating experimental data into a systems biological model. *Bioinformatics* 2021;37(11):1554–61.
- [50] You Y, Lai X, Pan Y, Zheng H, Vera J, Liu S, et al. Artificial intelligence in cancer target identification and drug discovery. *Signal Transduct Target Ther* 2022;7(1):156.
- [51] Aleksandrowicz G, Alexander T, Barkoutsos P, Bello L, Ben-Haim Y, Bucher D, et al. Qiskit: an open-source framework for quantum computing [Internet]. Genève: Zenodo; 2019 Jan 23 [cited 2024 Jan 22]. Available from: <https://zenodo.org/records/2562111>.
- [52] Zha J, Su J, Li T, Cao C, Ma Y, Wei H, et al. Encoding molecular docking for quantum computers. *J Chem Theory Comput* 2023;19(24):9018–24.
- [53] Shu G, Shan Z, Xu J, Zhao J, Wang S. A general quantum algorithm for numerical integration. *Sci Rep* 2024;14:10432.
- [54] Liu F, Bian K, Meng F, Zhang W, Dahlsten O. Information compression via hidden subgroup quantum autoencoders. *npj Quantum Inf* 2024;10:74.