**华东理工大学**

**毕业设计(开题报告)**

**课题** **基于LSTM生物过程监测**

**学院 生物工程学院**

**系别 生物工程系**

**专业 生物工程**

**年级 生工212**

**姓名 张福林**

**导师 李友元**

基于LSTM生物过程监测

# 选题背景、目的及意义

## 选题背景

生物发酵技术作为现代工业生物技术的核心组成部分，在制药、食品加工、生物能源、农业及环境保护等领域发挥着不可替代的作用。例如，抗生素、酶制剂、氨基酸、疫苗及生物燃料等高附加值产品的生产均依赖于高效、稳定的发酵过程。然而，生物发酵过程具有高度非线性、动态时变性和多变量耦合的特点，其反应机制涉及微生物代谢、环境参数波动、底物消耗与产物合成的复杂相互作用。这种复杂性导致发酵过程的精准监测与产物产量预测面临巨大挑战，直接影响工艺优化、质量控制及规模化生产的成本效益。

传统的发酵过程监测方法多依赖于离线实验室检测或基于经验模型的预测，存在实时性差、数据利用率低、难以捕捉动态变化规律等问题[1]。尽管统计模型（如PLS、PCA）和机器学习算法（如SVM、随机森林）在一定程度上提升了预测精度，但其对时序数据的长期依赖关系建模能力不足，难以适应发酵参数（如温度、pH、溶氧量、菌体浓度、底物浓度）在时间维度上的非线性演变特征。此外，发酵过程中传感器噪声、数据缺失及批次间差异进一步增加了模型构建的难度，导致现有方法在实际工业场景中的泛化性和鲁棒性有限。

近年来，深度学习技术的突破为时序数据建模提供了新的解决方案。长短期记忆网络（LongShort-TermMemory,LSTM）作为一种特殊的循环神经网络（RNN），凭借其独特的门控机制，能够有效捕捉时间序列中的长期依赖关系，并克服传统RNN的梯度消失问题[2]。将LSTM应用于生物发酵过程的动态监测与产量预测，有望通过实时整合多源传感器数据（如在线光谱、生物量传感器、代谢物浓度等），挖掘参数间的深层时序关联，从而建立高精度的产物合成动力学模型。例如，LSTM可学习发酵过程中菌体生长、底物消耗与产物积累之间的滞后效应，预测关键代谢拐点（如菌体生长停滞期、产物合成高峰期），为补料策略优化、异常状态预警提供决策依据。

本研究聚焦于LSTM在生物发酵过程监测与产量预测中的应用，旨在解决现有方法对时序动态建模不足的瓶颈问题。本次动态发酵过程是黑区霉发酵葡萄糖酸钠(sodiumgluconate)，葡萄糖酸钠的生产可以追溯到Hlasiwetz和Habermann在1870年发现了葡萄糖酸,一直到1952年,Blom等发明了添加NaOH、维持pH6.5以上的方法生产葡萄糖酸钠,使糖的转化率达95%以上,发酵时间缩短至20小时以内,这种工艺形成了现代工业深层发酵的基础[3]。葡萄糖酸钠是一种无腐蚀性、无挥发性、无毒、温和的有机酸钠盐。其分子式为C6H11NaO7,分子质量是218.14u,易溶于水,微溶于醇,同时不溶于乙醚及大多数有机溶剂。由于其无毒、热稳定性好、无潮解性等特点,在化工、食品、医药、轻工等行业具有广泛的应用[4]。

葡萄糖酸及其盐是相当重要的产品,被广泛应用于食品添加、制药及金属工业。目前工业上生产葡萄糖酸主要是通过黑曲霉(Aspergillusniger)发酵的方法,其过程是利用葡萄糖经葡萄糖氧化酶的脱氢作用形成葡萄酸-δ-内酯,再进一步水解为葡萄糖酸,同时发酵过程中补加氢氧化钠来维持发酵过程的pH,因而得到的产物为葡萄糖酸钠,若用碳酸钙来调节pH,则产物的形式为葡萄糖酸钙。[5]

## 选题目的

核心目标：本次研究分析了发酵法生产葡萄糖酸钠过程中的各参数的变化规律,构建了一种基于LSTM的智能化发酵过程监测与产量预测框架，通过深度融合时序数据特征与发酵过程机理，突破传统方法在动态建模及实时决策中的局限性。通过在线监测和离线分析检测,得出各参数的变化规律，为工厂进一步优化发酵工艺、缩短发酵周期提供原始的理论依据。

具体目标包括：

1. **动态时序建模与预测精度提升**  
   设计多变量LSTM网络架构，整合黑曲霉发酵过程中多维传感器数据（如温度、pH、溶氧量、生物量、底物/产物浓度等），捕捉参数间的时序关联性与非线性耦合特征，解决传统模型对发酵动态演变建模不足的问题，显著提升产物产量预测的精度与鲁棒性。
2. **实时监测与工艺优化闭环集成**  
   构建“预测-反馈-优化”闭环系统，将LSTM的实时预测结果与发酵过程控制策略（如动态补料、pH调节）相结合，探索基于预测模型的在线优化算法，缩短工艺调整滞后时间，提升产物合成效率与资源利用率。

文献综述**：**

一直以来生物产业的发酵预测就是一大关注重点。生物工程尤其是发酵过程是技术密集型的产业,涉及到微生物学，生物化学,生化工程,自动控制技术和计算机技术等。当前，生物发酵工业的自动检测技术主要局限于易测参数（如温度、pH值等），依赖成熟仪器仪表实现实时监测，而其他关键参数仍依赖离线检测或缺乏经济高效的方法。针对这一瓶颈，软测量技术通过构建可测变量与被估变量间的数学模型（软测量器/软仪表），利用计算机软件模型实时估计难以直接测量的重要工艺参数。其核心逻辑为：以可测变量为输入，通过数据驱动模型输出被估变量（如产物浓度、菌体活性等），从而为过程优化控制提供实时决策依据。

传统的软测量方法基于计算流体力学(CFD)，CFD能够对发酵罐内的流体流动、传热、传质、混合等过程进行高分辨率仿真，提供温度场、浓度场、流速场等物理量的空间分布信息。这种物理场数据可显著提升软测量模型对复杂反应器内局部参数（如溶氧梯度、底物分布）的估计精度。

Takamatsu等用非结构模型表征黑曲霉合成葡萄糖酸钠的过程。而结构模型则在更广泛的发酵工艺上有更精确的应用。Liu等首次用非结构模型构建了黑曲霉产葡萄糖酸钠的模型,模型是基于logistic方程的基础上构建的。此外张小昊等人也使用logistic方程进行了模拟预测。

此外，孙宗海和孙优贤2004[6]的研究聚焦于支持向量机（SVM）及其在微生物发酵领域的应用。支持向量机作为一种基于统计学习理论的模式识别方法，在处理小样本、非线性及高维模式识别问题中表现出显著优势。也是作为生物发酵领域的软测量模型之一

赵学庆等人2000[7]研究了神经网络在生物发酵过程状态预测中的应用。由于生物发酵过程的复杂性和不确定性，传统建模与预测方法往往难以奏效。为此，研究者提出了一种基于数据驱动的滚动学习—预测算法，并将其应用于工业生产规模的青霉素流加发酵过程，成功实现了对产量和糖耗的多步超前预测。实验结果表明，该神经网络模型能够有效弥补传统方法的不足，取得了令人满意的预测效果。然而，该方法的性能高度依赖于数据的质量和数量，并且在更复杂的发酵过程中可能面临适用性挑战。

## 发酵数据预测参数

发酵数据作为经典的时序数据，有着多种的预测手段。目前,对于发酵法生产葡萄糖酸钠的研究主要集中在菌种选育、培养基及培养条件优化等方面。而生产局限主要在产量和周期上,这使得发酵培养优化的试验有一定的盲目性。对发酵过程各个参数进行检测主要是在线实时检测,研究不同发酵阶段的各参数变化特点与产量及周期的关系,这对发酵优化的研究具有重要意义。

生化反应的关键参数可分为物理参数、化学参数和生物参数三类，其测量技术成熟度差异显著：

物理参数包含指标：发酵罐温度、压力、体积、空气流量、冷却水流量及进出口温度、搅拌马达转速与电流、泡沫高度等。 已有成熟的仪器仪表支持全自动实时监测，可根据工艺需求灵活选配。

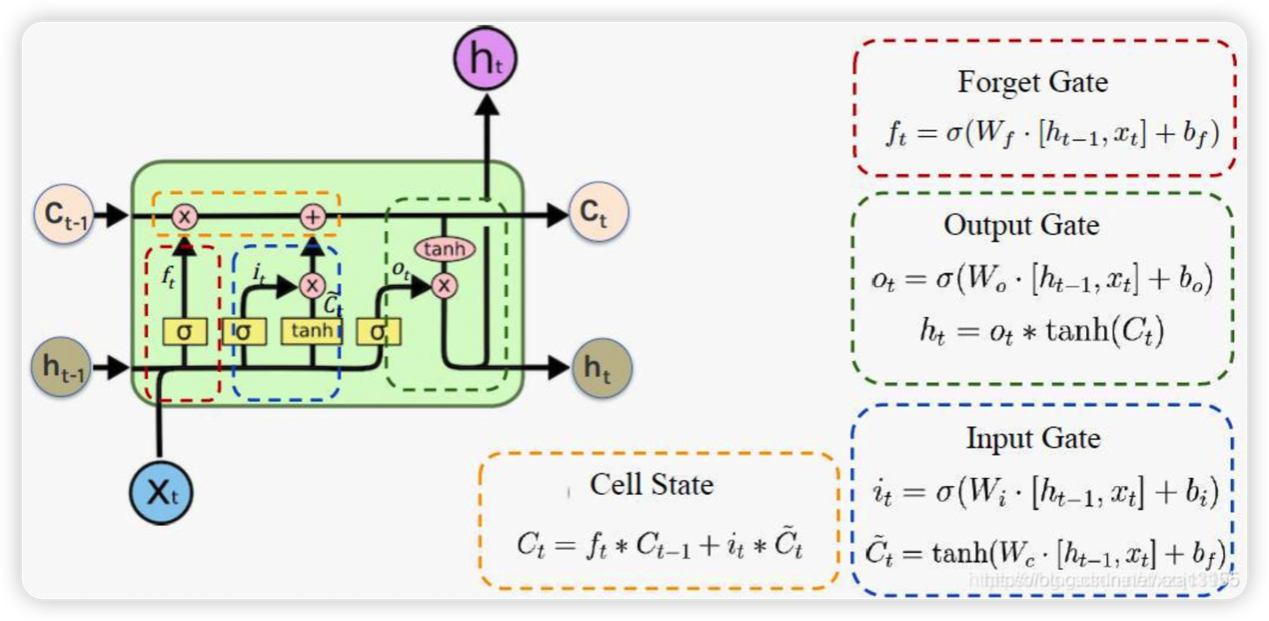
化学参数包含指标：pH值、溶解氧（DO）浓度、氧化还原电势（ORP）。 测量现状：通过专用电极（如pH电极、DO电极、ORP电极）可直接实现在线实时测量，技术成熟且广泛应用。

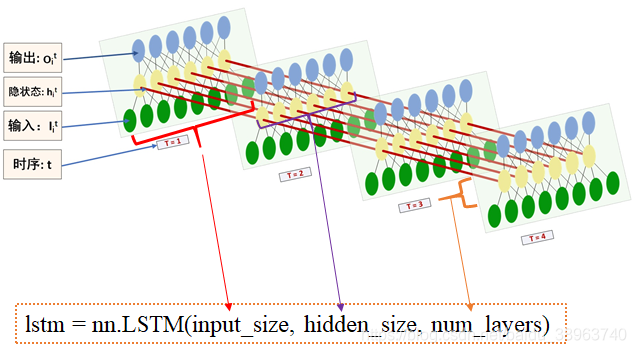
生物参数包含指标：生物质浓度、代谢物浓度（如产物、副产物）、底物浓度、比生长速率（μ）、底物消耗速率（qs）、产物形成速率（qp）等。 测量现状： 缺乏实时仪表：目前尚无经济高效的实时检测设备，主要依赖离线实验室分析（如高效液相色谱、生物传感器）或软测量技术（基于模型或数据驱动的间接估计）。物理与化学参数的实时监测技术已较为完善，而生物参数的实时精准测量仍是工业发酵过程的瓶颈，需通过软测量、新型传感器开发或多模态数据融合技术突破这一限制，以实现工艺优化与智能控制。

物理与化学参数的实时监测技术已较为完善，而生物参数的实时精准测量仍是工业发酵过程的瓶颈，需通过软测量、新型传感器开发或多模态数据融合技术突破这一限制，以实现工艺优化与智能控制。

## LSTM

长短时记忆(Long-Shokl-TermMemory,LSTM)网络是RNN的一种,能够解决一般RNN存在的长期依赖问题,并且可以解决长序列训练过程中出现的梯度爆炸和梯度消失的问题,因而LSTM适用于处理、预测时间序列中时间间隔较长的问题。LSTM的结构如图所示。一个经典的LSTM网络由单元或者记忆块组成,并且LSTM单元一般会有两个输出:单元状态C和隐藏状态h,将其传递到下一个LSTM单元,因此LSTM有三个输入:t-1时刻的单元状态Ct-1、隐藏状态h，以及当前时刻的单元状态C。





LSTM的计算更新过程可以表示为:

it=σ(Wxixt+Whiht-1+bi)(3)

ft=σ(Wxfxt+Whfht-1+bf)(4)

C~t=tanh(Wxcxt+Whcht-1+bc)(5)

Ct=ft⊙Ct-1+it⊙C~t(6)

ot=σ(Wxoxt+Whoht-1+bo)(7)

ht=ot⊙tanh(Ct)(8)

其中it、ft、ot分别为输入门、遗忘门、输出门,W为权重矩阵,b为偏置矩阵,C槇t为用于更新的备选内容,σ为sigmoid激活函数,tanh为激活函数,⊙表示逐元素相乘。若此时的LSTM单元为网络中的最后一个单元,那么,网络最终的输出为:

y=softmax(W⊙ht+b)(9)

其中,softmax为激活函数,W为权重矩阵,b为偏置矩阵。

## 对时序数据的处理方法

时间序列（TimeSeries）[8]可以定义为“一组按顺序排列的实数值”，它是按照时间顺序取得的一系列观测值。这里的“时间”具有广义坐标轴的含义，既可以指按时间的先后顺序排列的数据，也可以指按空间的前后顺序排列的随机数据。从经济到工程技术，从天文到地理和气象，几乎在各个领域都会遇到时间序列。而发酵过程中产生的各种数据，就是在一个发酵周期里的时间序列。

时间序列的特点在于其数据点之间存在时间或空间上的依赖关系，这种依赖关系使得时间序列分析成为研究数据变化规律、预测未来趋势的重要工具。

在工业发酵过程中获得的数据往往存在多种质量问题，如噪声数据、不完整数据和不一致数据，这些问题可能导致数据分析结果不准确，进而影响决策、建模、优化和控制的效果。噪声数据指数据中存在的错误或异常值，偏离了期望范围；不完整数据指某些感兴趣的属性缺失值；不一致数据则指数据内涵出现矛盾或不一致的情况。因此，数据预处理技术成为了不可或缺的环节，尤其是在处理包含噪声、不完整甚至是不一致数据时，必须通过数据预处理来提高数据质量，从而提升数据挖掘所获模式和知识的准确性与可靠性。数据预处理的主要任务包括数据清洗、数据集成、数据转换和数据约简[8]，其中数据清洗作为核心环节，通过去噪、填补缺失值、纠正错误等操作，有效提高数据质量，为后续的数据分析、挖掘和决策提供可靠的基础。

常见的数据预处理方法，如数据对齐，填补缺失值等

前向填充：使用上一个数据的值来填补缺失值

移动平均法：使用缺失值前一些数据的平均值填充

线性插值法：两点之间的距离来估计缺失值

样条插值法：使用多项式拟合曲线

最大值填充：使用当前序列的最大值填充数据

在发酵过程中，由于每批次发酵数据的时间长度和采样频率可能不同，数据对齐成为一项关键任务，以确保不同批次的数据能够进行有效比较和分析。常见的数据对齐方法包括以下几种：

1.时间插值对齐

通过插值方法（如线性插值、样条插值等）将不同批次的数据统一到相同的时间点上。例如，将较短批次的数据通过插值扩展到与最长批次相同的时间范围，或者将所有批次的数据插值到固定的时间间隔上。

3.动态时间规整（DTW）

动态时间规整是一种非线性对齐方法，适用于时间序列长度不同且变化速率不一致的情况。它通过计算数据点之间的最小距离路径，将不同批次的数据对齐到最优的时间轴上。

5.归一化时间对齐

将每批次的时间轴归一化到相同的范围（如0到1之间），然后将数据映射到归一化时间轴上。这种方法适用于时间长度差异较大的批次数据。

# 研究内容和研究方法

## 研究内容

发酵过程是生物工程、食品工业和制药等领域的关键环节[2]，其数据通常以时间序列的形式记录。

本次研究主要以黑曲霉发酵葡萄糖酸钠为研究对象，通过在线监测和离线分析技术统计分析不同时刻参数的变化规律，以此为特征构建LSTM神经网络，单步预测某发酵批次下一时刻的残糖以及葡萄糖酸钠的含量，为后续的动态补料和发酵检测提供理论指导。本次发酵的工艺流程如下: 茄子瓶种子 →15 L 种子罐 →35 °C pH5.2发酵至耗氧速率( OUR) 涨至最高点→转接 13% ( 占 50 L 罐量)到50L发酵罐→38 °C pH 5.2 发酵( 用 NaOH 与发酵产生的葡萄糖酸中和生成葡萄糖酸钠) → 至溶氧 DO 回升至 90% 以上发酵结束。种子罐 与发 酵 罐 实 施 在 线 监 测 尾 气 数 据,残 糖、酶活、酸钠及菌体浓度等离线数据每 4 h 取样检 测。

然而，真实的发酵时序数据往往存在数据缺失、取样时间长度不一致等问题，直接使用这些数据进行建模会严重影响模型的精度。其次，LSTM在处理时间序列数据时，虽然可以捕捉长期依赖关系，自动学习时间步之间的模式，但它的性能确实会受到输入特征的影响。如果现有特征不足以描述数据中的复杂关系，模型可能表现不佳。例如，在发酵过程中，简单的温度、pH值可能无法完全反映微生物代谢状态，可能需要结合其他特征如变化率、滑动窗口统计量等。

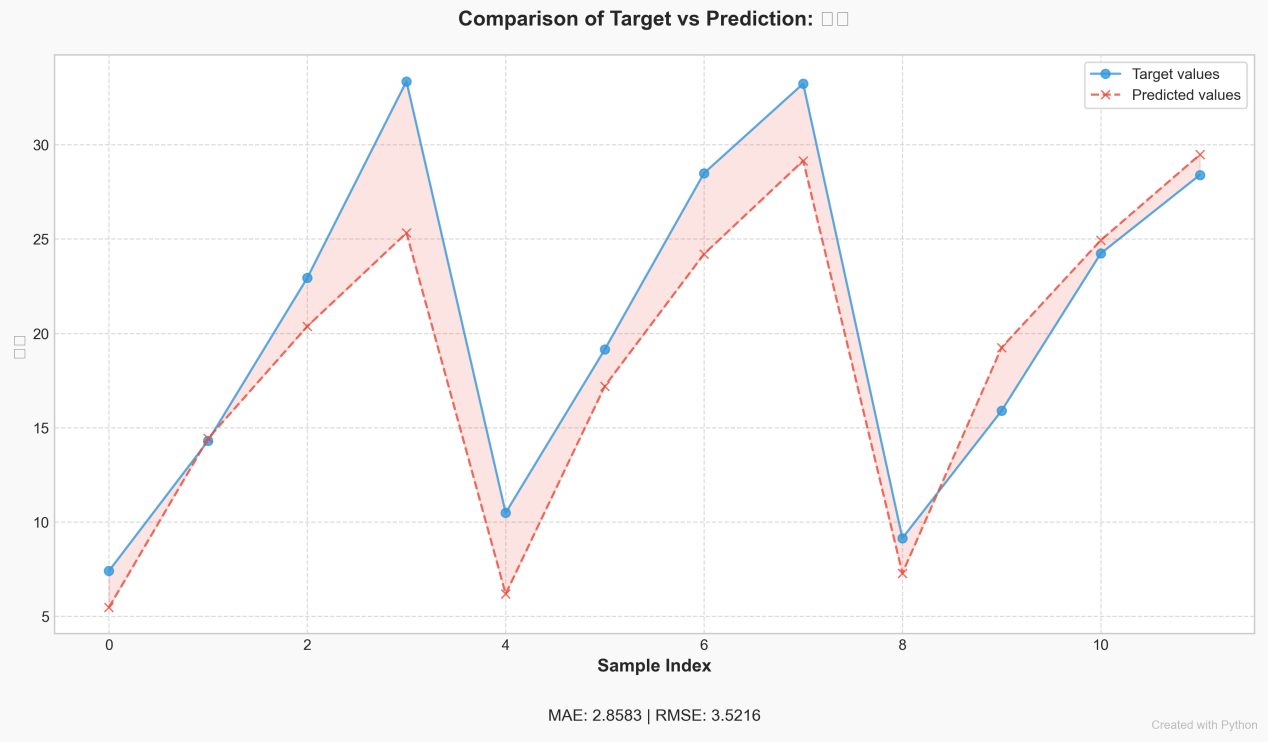
在线监测数据包括物理、化学参数，但离线检测的生物参数如生物质浓度等可能难以实时测量，可能需要通过软测量或特征工程来间接表示这些变量。例如，通过计算溶氧的变化率来反映菌体活性，或者结合多个传感器的数据生成新的综合指标。据此可以考虑特征工程的具体方法。例如，时间滞后特征可以捕捉历史信息的影响，滑动窗口统计量（均值、方差）可以平滑噪声并突出趋势，交互特征如温度与搅拌速度的乘积可能揭示协同效应。此外，基于领域知识的特征，引入领域先验知识（如黑曲霉生长阶段划分）构建特征（如代谢阶段标签、补料事件标记），可提升模型可解释性和泛化性。可以有效结合机理模型，提升模型的可解释性。

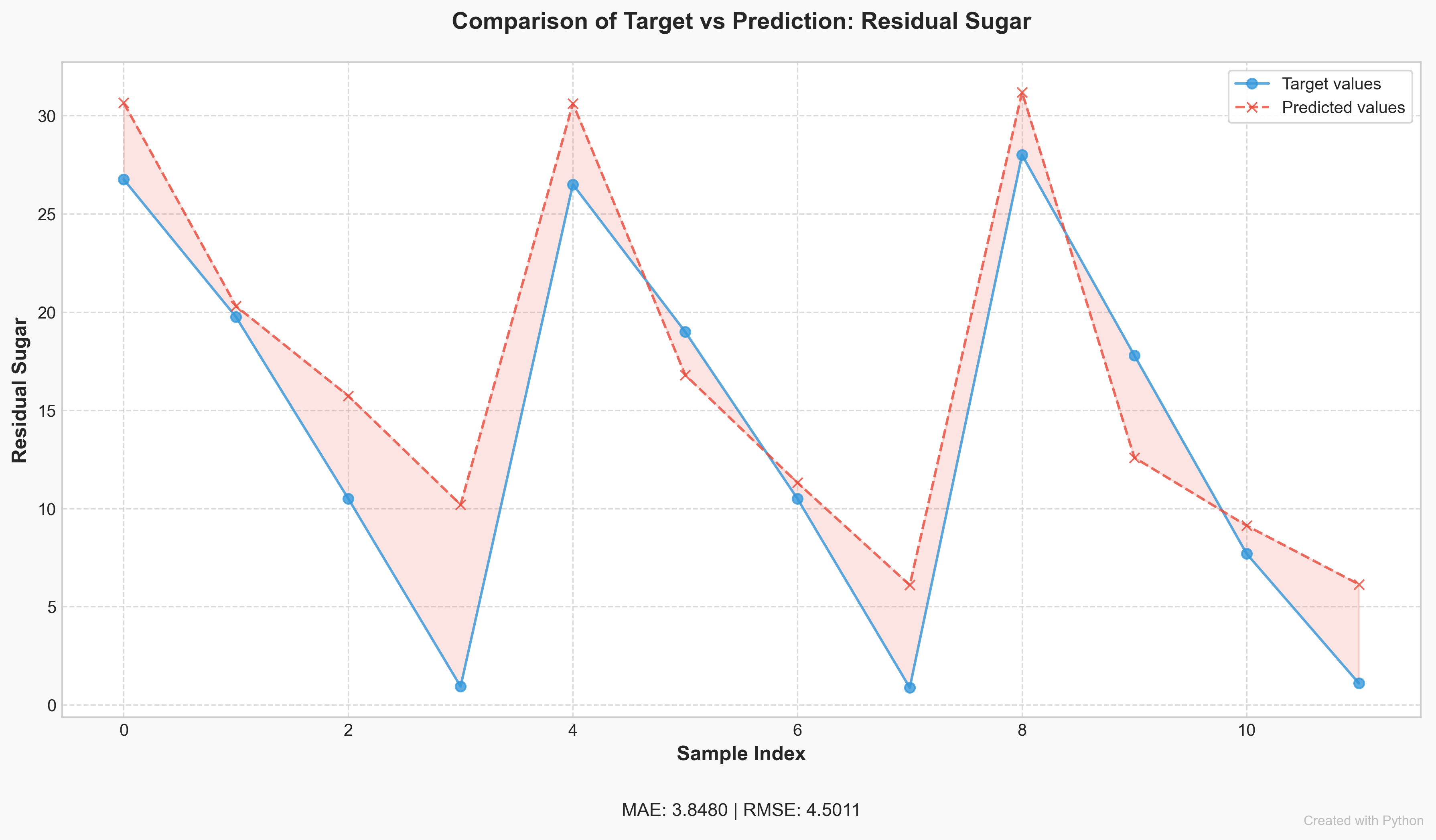
## 研究方法

此次研究依据实验室获取的四十余批黑曲霉发酵葡萄糖酸钠的真实发酵样本，数据的预处理主要是通过对进行缺失值的填补，时序特征的扩展等。对于网络模型架构主要通过超参数的调整来提高模型预估的准确率，减少模型损失。

1. LSTM的自动特征提取能力在小样本数据下受限，需人工特征补充先验知识，进行时序特征的拓展，如：添加滞后特征（Lag Features）：提取历史时刻的参数值（如前1h的溶氧均值），反映参数惯性。变化率与导数：计算参数的瞬时变化率（如溶氧下降速率），捕捉代谢拐点。滑动窗口统计量：窗口均值、方差、极值等，平滑噪声并突出趋势（例如：发酵期溶氧的滑动方差反映混合均匀性）。
2. 模型构建：利用torch.nn模块下的长短期长短时记忆(Long-Short-TermMemory,LSTM)网络来构建模型。将发酵数据以8:2切分训练集与测试集。使用丰富后的特征作为参数作为输入特征，以下一时刻的残糖和酸钠作为标签，代入模型进行训练。
3. 模型优化：通过对模型超参数参数（学习率，记忆细胞个数，隐藏层数，Dropout层等）进行调整，从而达到最佳的预测效果。
4. 在LSTM后端添加时间注意力层（Temporal Attention），依据黑曲霉在不同的发酵时间，划分为不同生长时期，自动加权不同时间步的特征重要性，例如强化对数生长期的特征权重。

padded 与 mask out 处理后，在测试集上预测结果





原始时序信号可能包含噪声（如传感器抖动）和冗余信息，通过特征工程可提取高信息量特征（如滑动窗口统计量、频域特征），增强模型对关键动态的捕捉能力。

引入领域先验知识（如黑曲霉生长阶段划分）构建特征（如代谢阶段标签、补料事件标记），可提升模型可解释性和泛化性。

时序特征扩展：

滞后特征（Lag Features）：提取历史时刻的参数值（如前1h的溶氧均值），反映参数惯性。

变化率与导数：计算参数的瞬时变化率（如溶氧下降速率，菌生长速率，比生长速率，糖消耗速率，产物变化速率。）。

滑动窗口统计量：窗口均值、方差、极值等，平滑噪声并突出趋势（例如：发酵期溶氧的滑动方差反映混合均匀性）。

交互特征构建：

参数间耦合关系：例如，溶氧（DO）与搅拌转速的比值反映氧传递效率；温度与pH的乘积表征代谢环境稳定性。

代谢平衡指标：基于质量守恒定律构建虚拟特征，如底物消耗与产物生成的摩尔比

阶段化特征标记：！！！！

根据黑曲霉生长周期划分阶段（适应期、对数期、稳定期），为各阶段赋予独热编码（One-hot Encoding），辅助LSTM识别阶段特异性模式。

标记关键事件（如补料操作、pH调节），作为二进制特征输入模型。（尤其指发酵末尾溶氧升高的阶段）

非线性变换：对参数进行对数化、平方根变换，缓解数据偏态分布对模型的影响。

LSTM自动特征学习的局限性

LSTM的自动特征提取能力在小样本数据下受限，需人工特征补充先验知识。

若输入特征间存在强共线性（如菌体干重与湿重），可能导致模型权重不稳定。

引入产物抑制模型，作为特征。

预处理阶段：优先提取领域关键特征（如溶氧下降速率、残糖消耗量），确保基础信息完整性。

模型迭代阶段：通过特征重要性分析（如SHAP值、注意力权重，PCA）筛选高贡献特征，逐步剔除冗余特征。

结合注意力机制：

在LSTM后端添加时间注意力层（Temporal Attention），自动加权不同时间步的特征重要性，例如强化对数生长期的特征权重。

混合建模：

将机理模型输出（如基于Monod方程的比生长速率估算值,产物抑制等）作为特征输入LSTM，实现“白盒+黑盒”协同建模，平衡数据驱动与机理约束。

# 时间安排

（1）2024.10-2024.11查阅文献，学习pytorch框架deep-learning基础原理

（2）2024.11-2024.12查阅相关资料，学习deep-learning相关算法和原理

（3）2024.12-2025.4进行数据预处理，构建LSTM网络流程

（4）2025.4-2025.6进行改进，优化模型参数，整理数据，开始论文写作。

# 参考文献

[1] 冯文红, 周生民, 赵伟. 发酵法生产葡萄糖酸钠过程中参数检测[J]. 生物加工过程, 2014, 12(4): 20-23.

[2] FOULADGAR N, FRÄMLING K. A novel LSTM for multivariate time series with massive missingness[J/OL]. Sensors, 2020, 20(10): 2832. DOI:10.3390/s20102832.

[3] 张静. 黑曲霉发酵生产葡萄糖酸钠的研究[D/OL]. 江南大学, 2009[2025-03-10]. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD2009&filename=2009013570.nh.

[4] 张同建, 赵祥颖, 张家祥, 等. 黑曲霉发酵生产葡萄糖酸钠残糖成分解析[J]. 中国酿造, 2016, 35(4): 88-91.

[5] 郭茸恺, 王泽建, 庄英萍, 等. 葡萄糖酸钠发酵中底物-产物作用对氧传递的影响及发酵过程优化[J/OL]. 安徽农业科学, 2012, 40(22): 11151-11153, 11198. DOI:10.13989/j.cnki.0517-6611.2012.22.022.

[6] 孙宗海. 支持向量机及其在控制中的应用研究[D/OL]. 浙江大学, 2004[2025-03-03]. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CDFD&dbname=CDFD9908&filename=2004042562.nh.

[7] 赵学庆,袁景淇,周又玲,贺松. 生物发酵过程神经网络状态预报器的验证[J]. 无锡轻工大学学报, 2000(6): 542-546.

[8] 秦天杰. 时序数据处理及其应用系统的开发[D/OL]. 北京化工大学, 2006[2025-03-03]. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD0506&filename=2006125266.nh.