**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。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（Long Short-Term Memory, 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。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。通过构建融合多变量发酵参数的LSTM预测框架，探索其对不同发酵阶段（延迟期、对数生长期、稳定期）产物合成规律的建模能力，并评估其在跨批次、跨菌种场景下的泛化性能。

## 选题目的

核心目标：本研究旨在构建一种基于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、跨场景泛化及实时决策中的局限性。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. **跨批次与跨工艺泛化能力验证**  
   开发基于迁移学习或元学习的LSTM优化策略，针对不同发酵批次（如初始条件差异、菌种代谢波动）或工艺类型（如分批发酵、补料分批发酵），建立自适应参数更新机制，验证模型在跨场景下的泛化能力，降低对历史数据完备性的依赖。
3. **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 ( C F D )，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。

此外，孙宗海和孙优贤 2004[1]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[2]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。

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。

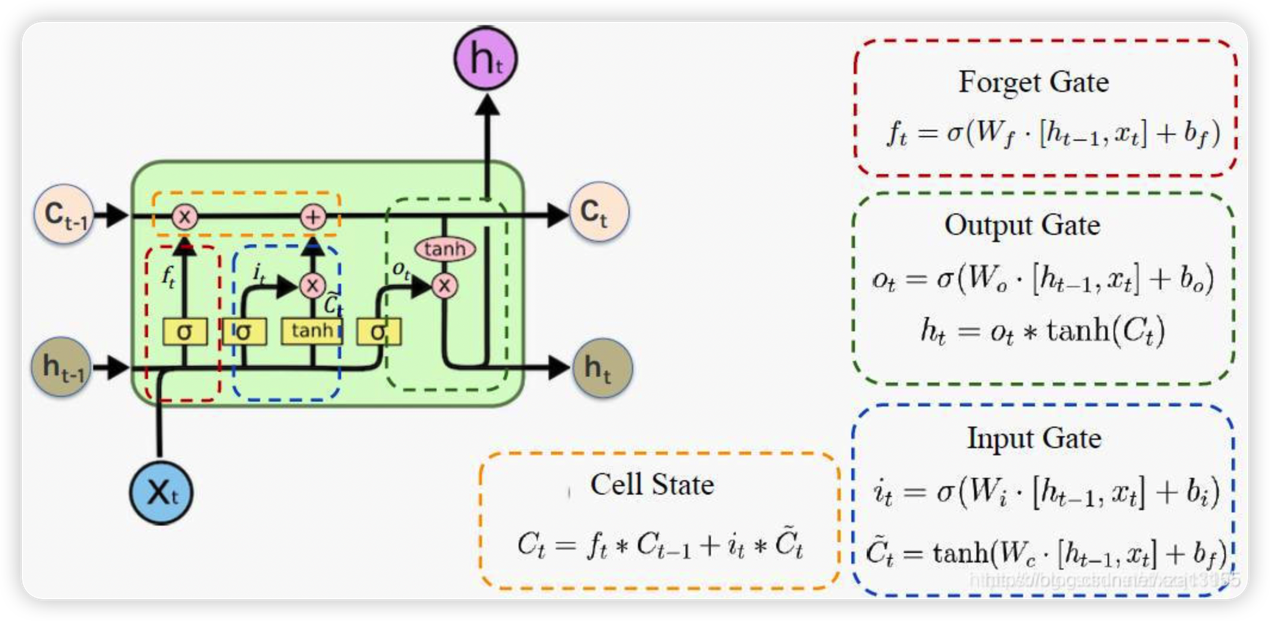
## Pytorch框架

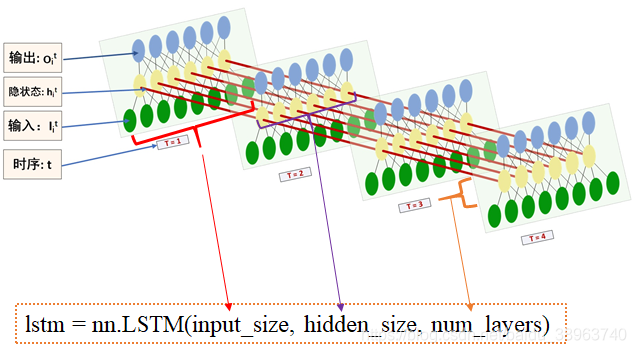
Torch 是一个开源的科学计算框架，最初由 Facebook 的人工智能研究团队（FAIR） 及其他团队共同开发，专注于深度学习和计算机视觉。其设计和实现基于 Lua 语言，最早版本可追溯到 2002 年，但直到 2010 年 才在深度学习社区中广泛流行。Torch 的主要优势在于其高效性，它利用高效的 C 和 Fortran 库（如 BLAS 和 ATLAS）进行数值计算，同时提供了灵活的工具，便于构建和实现机器学习模型，并且代码完全开放，便于全球研究者和开发者协作。然而，尽管 Torch 具有这些优势，但由于其基于 Lua 语言开发，普及度相对较低，尤其是在 Python 已成为机器学习领域主流语言的背景下，这大大限制了 Torch 的使用范围。

为了解决这一问题，PyTorch 应运而生。PyTorch 是基于 Torch 设计的 Python 版本，由 Soumith Chintala 等人于 2016 年 创建。其设计宗旨是继承 Torch 的灵活性，同时结合 Python 的普及性、易用性及其强大的生态系统。自推出以来，PyTorch 迅速成为深度学习和工业应用中最受欢迎的框架之一。PyTorch 的核心特点包括动态计算图（Dynamic Computational Graph），与 TensorFlow 1.x 的静态计算图不同，PyTorch 使用动态计算图，计算图在执行时动态生成，极大提高了灵活性，特别适合调试、修改模型结构及研究新算法。此外，PyTorch 深度集成 Python，其核心使用 Python 编写，并与 Python 的数据处理库（如 NumPy、SciPy、Pandas 等）深度集成，对 Python 开发者非常友好。PyTorch 还提供了强大的自动求导功能（Autograd），可以自动计算梯度，极大简化了反向传播算法的实现。

## LSTM

长短时记忆( Long-Shokl-Term Memory,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( Wxi xt + Whi ht-1 + bi ) ( 3)

ft = σ( Wxf xt + Whf ht-1 + bf ) ( 4)

C ~ t = tanh ( Wxc xt + Whc ht-1 + bc ) ( 5)

Ct = ft ⊙ Ct-1 + it ⊙ C ~ t ( 6)

ot = σ( Wxo xt + Who ht-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（Time Series）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[3]，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，其数据通常以时间序列的形式记录。然而，真实发酵数据往往存在缺失值、时间长度不一致等问题，直接使用这些数据进行建模会影响模型精度。本研究旨在通过对30余批真实发酵时序数据进行预处理（如缺失值填补、数据对齐等），构建并优化LSTM（长短期记忆）神经网络模型，以提高发酵过程预测的准确性和可靠性。

## 研究方法

主要分为数据预处理、模型构建、模型优化比较三个部分。

1. 数据预处理：使用线性差值法对缺失值进行填充，之后使用DTW动态时间规划进行时间步的统一。
2. 模型构建：利用torch.nn模块下的长短期长短时记忆( Long-Shokl-Term Memory,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。使用物理化学等当前参数作为输入特征，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：通过对模型参数（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。

# 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] 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.

[2] 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.

[3] 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].