UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA (CCET)

MARTIN ALBERTO SAVI GUALCO

ESTUDO COMPARATIVO ENTRE O FILTRO DE PARTÍCULAS E O FILTRO DE KALMAN ESTENDIDO PARA MONITORAMENTO ON-LINE DA SÍNTESE DE GALACTOOLIGOSSACARÍDEOS.

SÃO CARLOS -SP
2025

MARTIN ALBERTO SAVI GUALCO

ESTUDO COMPARATIVO ENTRE O FILTRO DE PARTÍCULAS E O FILTRO DE KALMAN ESTENDIDO PARA MONITORAMENTO ON-LINE DA SÍNTESE DE GALACTOOLIGOSSACARÍDEOS.

Trabalho de conclusão de curso apresentada ao Departamento de Engenharia Química da Universidade Federal de São Carlos, para obtenção do título de bacharel em Engenharia Química.

Orientador: Dr. Marcelo Perencin de Arruda Ribeiro

São Carlos-SP
2025

ERRATA

GUALCO, Martin. **Estudo comparativo entre o Filtro de Partículas e o Filtro de Kalman Estendido para monitoramento on-line da síntese de Galactooligossacarídeos.** 2025. nº de páginas. Natureza (Grau) - Departamento de Engenharia Química, Universidade Federal de São Carlos, São Carlos - SP, 2025.

Folha	Linha	Onde se lê	Leia-se
Indicar o n° da folha	Indicar o n° da linha	Indicar o erro	Indicar a correção

DEDICATÓRIA

Às mulheres da minha vida: Gabriela, Ana Paula e Tassia.

AGRADECIMENTO

Agradeço a todos aqueles que me ajudaram neste trajeto — meu orientador, os colegas do grupo de pesquisa e todos os professores com quem tive contato — pelo apoio, pelas lições e pelas experiências vividas. Agradeço também a doutoranda Nicole Maione pelo auxílio e fornecimento de dados experimentais, e aos órgãos de pesquisa que contribuiram para o financiamento do projeto: Coordenação de Aperfeiçoamento Pessoal de Nível Superior (CAPES) – Código de financiamento 001; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – bolsa #141150/2020-3 e bolsa PIBIC #145321/2022-3; Fundação de Apoio a Pesquisa de São Paulo (FAPESP) - projeto #2018/04933-5.

RESUMO

Os galactooligossacarídeos (GOS) são oligossacarídeos não digeríveis compostos por cadeias de 2 a 8 unidades de galactose e glicose. Sua obtenção tem início com a lactose, subproduto abundante na indústria de laticínios, por meio de uma reação enzimática denominada transgalactosilação da lactose pela β-galactosidase, gerando uma mistura complexa de di-, tri- e tetrassacarídeos. Dentre os benefícios dos GOS, destacam-se o estímulo à microbiota intestinal, o bloqueio da adesão de patógenos no trato digestivo e o auxílio na absorção de minerais. Além dos efeitos fisiológicos, os GOS são amplamente aplicados em formulações alimentícias graças à sua alta solubilidade em água, estabilidade térmica e resistência a variações de pH. Essas características conferem aos GOS elevado interesse por parte da indústria alimentícia e motivam o desenvolvimento de processos otimizados para sua produção. Tais processos, quando operados em escala, demandam um bom controle do reator, de modo a minimizar perdas deste produto por hidrólise e encontrar um ponto de parada satisfatório para processos em batelada. Surge, então a necessidade de métodos robustos e precisos de monitoramento, que sejam capazes de fornecer informação em tempo real sobre o processo. As técnicas diretas de medição on-line, como a aplicação de métodos espectrométricos, podem fornecer medidas em tempo real sobre algumas variáveis no processo. No entanto, estas técnicas possuem limitações tanto sobre as variáveis que conseguem medir, quanto sobre o ruído de suas medidas, de modo que o monitoramento realizado desta forma possa ser limitado para as finalidades de controle. Ao combinar um modelo fenomenológico do sistema, com as medições on-line do mesmo, surge a possibilidade de aproveitar mais informação do sistema, e portanto obter monitoramento e controle de melhor desempenho. Esta combinação é viabilizada através dos estimadores de estado, que, de modo geral, após realizarem uma estimativa a *priori* do estado com base no modelo, recebem uma medida do sistema e efetuam uma correção, fornecendo uma medida a *posteriori*. O presente estudo busca explorar as capacidades de dois tipos de estimadores de estado – Filtro de Partículas (PF, Particle Filter) e Filtro de Kalman Estendido (EKF, Extended Kalman Filter) – para realizar o monitoramento em tempo real de uma síntese de GOS, descrita matematicamente por um modelo cinético, e mensurada através de uma sonda UV-Vis. Apesar da maior complexidade e custo computacional do PF, os resultados mostraram que, partindo de condições iniciais bem definidas e com baixa incerteza, o desempenho de ambos estimadores foi similar. No entanto, ao utilizar uma condição inicial afastada da realidade e com maior incerteza, o PF apresentou um desempenho superior ao do EKF, convergindo de maneira mais rápida aos dados experimentais, e alcançando erros aproximadamente metade dos obtidos pelo EKF na maioria das espécies estimadas.

Palavras-chave: GOS. Filtro de Kalman Estendido. Filtro de Partículas. PLS. Softsensor.

RESUMO EM LÍNGUA ESTRANGEIRA

Galactooligosaccharides (GOS) are nondigestible oligosaccharides composed of chains of 2 to 8 units of galactose and glucose. Their production starts with lactose, an abundant by-product of the dairy industry, via an enzymatic reaction known as β-galactosidase-catalyzed lactose transgalactosylation, generating a complex mixture of di-, tri-, and tetrasaccharides. Key benefits of GOS include the stimulation of the intestinal microbiota, the blocking of pathogen adhesion in the digestive tract, and the enhancement of mineral absorption. In addition to their physiological effects, GOS are widely applied in food formulations due to their high water solubility, thermal stability, and resistance to pH variations. These characteristics make GOS highly attractive to the food industry and motivate the development of optimized processes for their production. Such processes, when operated at scale, require effective reactor control to minimize product losses from hydrolysis and determine a satisfactory stopping point for batch processes. This highlights the need for robust and precise monitoring methods capable of providing real-time process information. Direct online measurement techniques, such as spectrometric methods, can provide real-time measurements of some process variables. However, these techniques have limitations both in terms of the variables they can measure and the noise present in their measurements, which can restrict their usefulness for control purposes. By combining a fundamental model of the system with online measurements, it is possible to leverage more system information, thereby achieving improved monitoring and control performance. This combination is enabled through state estimators which, in general, first perform a prior estimate of the state based on the model and then receive a system measurement and apply a correction, providing a posterior estimate. The present study seeks to explore the capabilities of two types of state estimators - the Particle Filter (PF) and the Extended Kalman Filter (EKF) - for performing real-time monitoring of a GOS synthesis, described mathematically by a kinetic model and measured using a UV-Vis probe. Despite the greater complexity and computational cost of the PF, the results showed that, when starting from well-defined initial conditions with low uncertainty, the performance of both estimators was similar. However, when using an initial condition far from reality and with higher uncertainty, the PF showed superior performance compared to the EKF, converging more rapidly to the experimental data and achieving errors approximately half those obtained by the EKF for most of the estimated species.

Keywords: GOS. Extended Kalman Filter. Particle Filter. PLS. Softsensor.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Galactooligossacarídeos (GOS) são carboidratos prebióticos de interesse na área de alimentos e saúde, sendo descritos como oligossacarídeos não digeríveis constituídos por cadeias de galactose com uma glicose terminal. A forma mais comum de obter GOS se da através da conversão enzimática da lactose, que é um subproduto abundante na indústria laticínia (HERNÁNDEZ-HERNÁNDEZ et al., 2012; WHISNER et al., 2013). Do ponto de vista de benefícios à saúde, GOS são associados a efeitos como estímulo de bifidobactérias e lactobacilos, proteção contra infecções e aumento da absorção de cálcio, reforçando seu papel como ingredientes prebióticos, tanto para pessoas mais jovens quanto de idade (TORRES et al., 2010; DIAS et al., 2009; WHISNER et al., 2013; VAN LEUSEN et al., 2014). Além disso, alguns aspectos físico-químicos como estabilidade em meio ácido, boa solubilidade em água, doçura moderada e baixo valor calórico relativo tornam fazem dos GOS ingredientes atrativos para o uso em fórmulas infantis, e outros produtos alimentícios. Considerando a expansão do mercado global de alimentos pre e probióticos, surge a motivação para a busca por processos de síntese mais eficientes e economicamente competitivos (DIAS et al., 2009; TORRES et al., 2010; GOSLING et al., 2011; ALEXANDER et al., 2023; VAN LEUSEN et al., 2014).

Do ponto de vista de engenharia de processos, a reação de síntese de GOS é governada por uma rede de etapas simultâneas de transgalactosilação e hidrólise, fazendo com que em algum momento, a degradação do GOS seja dominante. Isso implica em um ponto de concentração máxima, para reações em batelada (GOSLING et al., 2011; SCHULTZ et al., 2021). Em escala industrial, isso faz com que o controle do reator permita identificar o equilíbrio na operação, evitando tanto a interrupção prematura quanto a extensão excessiva do tempo de reação, que leva à perda de produto por hidrólise (MAIONE, 2024). Para processos em batelada ou batelada alimentada, este equilíbrio se traduz na necessidade de monitorar em tempo real as principais espécies envolvidas, como a lactose, monossacarídeos e diferentes frações de GOS (DIAS et al., 2009; SCHULTZ et al., 2021). Alguns métodos de alta precisão, como a cromatografia líquida de alta performance (HPLC), são muito utilizados na caracterização das misturas de GOS e para a validação de modelos cinéticos, no entanto, demandam um alto custo e tempo para preparo de amostras,

portanto seu uso para fins de controle é inviável (HERNÁNDEZ-HERNÁNDEZ et al., 2012; MAIONE, 2024).

Devido às limitações dos métodos como HPLC, técnicas “on-line” baseadas em espectrometria, tais como UV-Vis, NIR e MIR, fornecem uma alternativa atraente para monitoramento de bioprocessos, por permitirem medições rápidas, não destrutivas e em tempo real, e são amplamente utilizadas para o monitoramento de GOS (DIAS et al., 2009; MISHRA et al., 2025; ALEXANDER et al., 2023). No entanto, os sinais espectrais geralmente apresentam alta colinearidade entre seus comprimentos de onda e estão sujeitos a ruídos significativos, isso gera uma grande necessidade de métodos robustos de calibração multivariada. Mesmo assim, o uso direto dessas variáveis para fins de controle pode ser insuficiente, dependendo do ruído e das variáveis que podem ser medidas (HERNÁNDEZ-HERNÁNDEZ et al., 2012; DIAS et al., 2009; ALEXANDER et al., 2023; MISHRA et al., 2025). Diversos estudos na área dos bioprocessos mostram que a combinação de modelos fenomenológicos com medições espectroscópicas, por meio da incorporação de estimadores de estado, pode mitigar essas limitações. A combinação de informação mecanística e correções com leituras empíricas foi demonstrada a fornecer estimativas com erros menores do que a utilização de apenas uma ou outra (KRÄMER; KING, 2019; LOPEZ et al., 2021; SCHIEMER et al., 2023; HERMANN; KREMLING, 2025; ALEXANDER et al., 2023; MISHRA et al., 2025).

Neste contexto, estimadores de estado como o Filtro de Kalman Estendido (EKF) e o Filtro de Partículas (PF) desempenham papel central, pois integram recursivamente predições de um modelo dinâmico não linear com medições ruidosas para fornecer estimativas a posteriori das variáveis de estado e de parâmetros (SIMUTIS et al., 2014; STELZER et al., 2017; ALEXANDER et al., 2023). O EKF, devido à seu funcionamento mais simples e suposições de distribuiçõs normais, fornece resultados de modo eficiente (com baixo custo computacional) porém sujeitos a erros de linearização em sistemas fortemente não lineares, enquanto o PF, ao representar explicitamente a distribuição a posteriori por um conjunto de partículas amostradas, pode alcançar maior acurácia em cenários não gaussianos e com grandes incertezas, a troco de um esforço computacional substancialmente maior (SIMUTIS et al., 2014; STELZER et al., 2017). Uma forma de avaliar a diferença entre estes estimadores para o monitoramento de GOS, é efetuando comparações

uniformes, através de simulações em conjuntos iguais de dados variando os parâmetros que são comuns a ambos os estimadores, de modo a observar o comportamento de cada um frente aos dados experimentais. Estas comparações podem fornecer uma base para determinar em quais casos o desempenho do PF justifica seu custo e complexidade adicionais, frente ao EKF. Portanto, o objetivo geral do trabalho é realizar comparações entre esses estimadores em suas capacidades de fornecer estimativas no monitoramento da síntese de GOS, tanto em condições iniciais bem definidas quanto em condições iniciais incertas. Em decorrência do trabalho, alguns objetivos intermediários foram alcançados, sendo eles o desenvolvimento de um sistema de simulações que permite simular o monitoramento com condições e métodos distintos de maneira consistente; e também a análise do modelo cinético e de calibração multivariada, pois os estimadores dependem fortemente deles.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. GALACTOOLIGOSSACARÍDEOS (GOS)

A formação de GOS pode ser dada tanto pelas rotas químicas quanto enzimáticas, uma das rotas químicas são as chamadas reações de reversão catálisadas por ácidos. A partir da lactose, esta reações geram misturas complexas de di- e trissacarídeos. As rotas químicas normalmente possuem múltiplas etapas, baixa seletividade e condições de alta temperatura e concentração de ácido, gerando também dificuldade na separação do GOS e das misturas ácidas. Estas dificuldades implicam em uma inviabilização econômica dessa rota produtiva (TORRES, 2010; ZENG, 2023; OSMAN, 2016). As rotas comerciais para a produção de GOS são as biocatalisadas, principalmente a transgalactosilação enzimática da lactose catalisada por β-galactosidases, em um mecanismo no qual se forma um complexo galactosil–enzima. A partir deste complexo, resíduo galactosil pode ser transferido tanto para outra molécula de açúcar presente no meio (como lactose ou um próprio GOS, formando novos oligossacarídeos) quanto para a água, o que acarreta em hidrólise de GOS. Isso implica em uma competição das reações entre vias de formação e degradação, e em processos em batelada leva a um perfil típico em que a concentração de GOS cresce até um máximo e, em seguida, decai (GOSLING et al., 2011; SCHULTZ et al., 2021). Em escala de processo, a forma mais comum de implementar essa rota é utilizando β-galactosidases de origem microbiana, empregando lactose como substrato em reatores operados em modo batelada, contínuo ou semi-contínuo. Uma matéria-prima comum para este processo é o permeado do soro do leite, devido a seu alto teor de lactose (DIAS et al., 2009; CHOCKCHAISAWASDEE et al., 2005; TORRES et al., 2010; RODRIGUEZ-COLINAS et al., 2016; MAIONE, 2024). A origem da enzima também possui um efeito no proceso, por exemplo, β-galactosidases de B. circulans, K. lactis e A. oryzae exibem perfis distintos de atividade e graus de polimerização diferentes (de di- a hexassacarídeos). Este fato acaba implicando em misturas de GOS com distribuições diferentes de di-, tri- e oligossacarídeos (GOSLING et al., 2010; YIN et al., 2017; TORRES et al., 2010). A caracterização fina dessas misturas, necessária para relacionar estrutura e quantificar as concentrações, recorre a técnicas de alta

resolução como HPLC e espectrometria de massa, capazes de separar e identificar oligosacarídeos com elevada precisão. Tais técnicas são associadas a um elevado custo, necessidade de preparação de amostras e por tanto não são capazes de fornecer respostas em tempo real (COULIER et al., 2009; HERNÁNDEZ-HERNÁNDEZ et al., 2012; DIAS et al., 2009).

2.2. MODELO CINÉTICO

Na literatura, existem diversos modelos cinéticos para a síntese de GOS (BOON et al.,1999; CHEN et al., 2003; MARTINS; LISBÔA, 2015; PALAI; BHATTACHARYA, 2013; RODRIGUEZ-FERNANDEZ et al., 2011, SCHULTZ et al., 2021). No entanto, a modelagem dessa cinética apresenta certas dificuldades, como a complexidade das reações simutâneas de hidrólise e transgalactosilação, o grande número potencial de espécies e parâmetros, fortes correlações paramétricas e, em muitos casos, omissão explícita da etapa de inibição ou mesmo violações de balanço de massa dos resíduos de sacarídeos (MARTINS; LISBÔA, 2015; SCHULTZ et al., 2021). O modelo de Boon et al., (1999) utiliza mecanismos padrão baseados em complexo enzima–galactosil com número moderado de parâmetros, descrevendo adequadamente hidrólise de lactose e formação de trissacarídeos e reduzindo o risco de sobreajuste, mas considera apenas GOS agregados com trissacarídeos, negligenciando tetrassacarídeos e galactobiose (BOON e tal., 1999; SCHULTZ et al., 2021). O trabalho de PALAI e BHATTACHARYA (2013), parte de uma forma simplificada de Michaelis–Menten em poucas etapas para relacionar conversão de lactose e rendimento global de GOS. Em complemento, MARTINS e LISBÔA (2015) propuseram um modelo em regime transitório com inibição reversível simultânea por glicose e galactose, formulado explicitamente para obedecer à conservação de massa para a soma da enzima livre e seus complexos ao longo do tempo. O modelo representa a fração de GOS apenas como soma de di- e trissacarídeos, sem discriminar espécies individuais como a galactobiose. Nesse contexto, o modelo de SCHULTZ et al. (2021) se diferencia por descrever explicitamente, para a β-galactosidase de *Kluyveromyces lactis* operando com lactose como substrato, e incluir a formação de tetrassacarídeos e GOS oriundos exclusivamente por resíduos

de galactose (em particular a galactobiose), incorporando um balanço de massa de sacarídeos residuais para a validação do perfil de galactobiose.

2.3. SOFTSENSORS E CALIBRAÇÃO MULTIVARIADA

Softsensors fornecem uma forma de extrair informação adicional de um sistema. Ao combinar informação de sensores físicos com um modelo, viabilizam estimativas em tempo real relevantes para otimização e controle de bioprocessos (HERMANN et al., 2025). Soft sensors podem ser baseados em dados, em comportamento mecanístico ou em ambos. Ao lidar com sinais de múltiplos sensores (como espectros), dependem de calibração multivariada, ou seja, de modelos que relacionam simultaneamente muitos preditores com uma ou mais variáveis de interesse usando a correlação entre eles. Normalmente, os dados de espectro possuem alta colinearidade, e alta dimensionalidade, acarretando em problemas de matrizes singulares ao utilizar métodos como mínimos quadrados comuns, portanto faz-se necessário a aplicação de métodos que possam reduzir a dimensionalidade, como PCR ou PLS (MISHRA et al., 2025). Normalmente, considera-se o soft sensor como todos os algoritmos associados para fornecer estimativas das variáveis do sistema em questão. No entanto, para fins de organização do trabalho, abordou-se o softsensor como sendo algo diferente dos estimadores de estado.

2.3.1. PLS

De acordo com Wold et al. (2001), a regressão por PLS relaciona as matrizes X e Y por um modelo multivariado que modela a estrutura das mesmas. Essas variáveis são decompostas de acordo com as equações 13 e 14, onde T e U são os escores de X e Y, P é a matriz de loadings, C é a matriz de pesos de Y e E e F as matrizes de resíduos.

$$X = T\,P' + E \tag{1}$$

$$Y = U\,C' + F \tag{2}$$

Os escores de X são os preditores de Y, estes escores são estimados como combinações lineares das variáveis originais em X. A quantidade desses escores representa um hiperparâmetro ajustável, que controla a complexidade do modelo.

Este modelo fornece uma certa interpretabilidade ao analisar os escores T e U, e dos loadings/pesos P e C, pois permitem observar quais variáveis em X mais afetam a predição de Y. Uma das limitações do PLS é em problemas não-lineares: a relação entre os escores, preditores e variáveis preditas é essencialmente linear, logo a dependência entre X e Y é representada por correlações lineares. Existem outras formas do PLS que incluem transformações adicionai para captar comportamentos não-lineares (NELLES, 2001).

2.3.2. VALIDAÇÃO CRUZADA

A validação cruzada é uma metodologia robusta para validar modelos e realizar uma escolha de hiperparametros que tenha um bom equilíbrio de performance e complexidade, minimizando o sobreajuste (NELLES, 2001). Algumas formas comuns de validação cruzada são o K-Fold, Leave-One-Out (um caso do K-Fold em que K = N onde N é o número total de pontos experimentais), Bootstraping e critérios de informação. Segundo Lumbumba et al. (2024), a validação cruzada por Leave-One-Out possui menor viés porém maior variância que o K-fold, além de um custo computacional maior pela quantidade maior de divisões. Devido à simplicidade do modelo de PLS, a validação cruzada por K-fold com 4 folds foi escolhida.

O funcionamento desta metodologia de validação cruzada se baseia em dividir o conjunto de treinamento em três blocos (folds) de calibração e um de validação, então ajusta-se o modelo com os folds de calibração e realiza-se a predição em cima dos espectros correspondentes ao fold de validação, de modo a obter as concentrações preditas y_{pred}, que são comparadas com as concentrações desse mesmo fold de validação y_{ref}. Esse processo é repetido para todos os folds, e então calcula-se o RMSECV normalizado. Com essa métrica, é possível realizar um teste F (NELLES, 2001) para avaliar se a diminuição do RMSECV é estatísticamente significativa ao aumentar o número de regressores, escolhendo assim o modelo ótimo. Adotando um nível de significância de α = 0,10. As hipóteses nula H_0 e alternativa H_1 são dadas a seguir:

$$H_0: \theta_{adicional} = 0 \rightarrow \text{modelo simples é aceitável}$$
$$H_1: \theta_{adicional} \neq 0 \rightarrow \text{modelo simples não é aceitável}$$

Onde $\theta_{adicional}$ representa a adição de parâmetros. A hipótese nula é rejeitada de acordo com a inequação 3:

$$\frac{RMSECV^2_{simples}}{RMSECV^2_{complexo}} > F(\alpha,N,N) \quad (3)$$

Onde *N* são os graus de liberdade do numerador e denominador, igual ao número de pontos experimentais utilizados na multicalibração.

2.4. ESTIMADORES DE ESTADO

Estimadores de estado são algoritmos que combinam informações de um modelo dinâmico com medições on-line e obtem a distribuição a posteriori do vetor de estados de modo a fornecer, em tempo real, estimativas de variáveis internas não medidas diretamente, como concentrações ou taxas específicas (SIMUTIS et al., 2014; STELZER et al., 2017; DOUCET; JOHANSEN, 2008). O filtro de Kalman clássico resolve esse problema para sistemas lineares com ruído gaussiano, produzindo estimativas ótimas de estados medidos e não medidos (RAWLINGS; MAYNE; DIEHL, 2017). Para sistemas não lineares, foram propostas extensões como o filtro de Kalman estendido (EKF), que lineariza o modelo em torno da estimativa atual, e o filtro de Kalman não linear por pontos sigma (UKF ou SPKF), que propaga um conjunto de pontos sigma pela dinâmica para obter as estatísticas do estado (SIMUTIS et al., 2014; KRÄMER). Em problemas fortemente não lineares ou não gaussianos, filtros de partículas implementam métodos de Monte Carlo sequenciais que aproximam a distribuição a posteriori por um conjunto de amostras ponderadas (“partículas”), sem necessidade de linearização ou de hipóteses gaussianas (SIMUTIS et al., 2014; DOUCET; JOHANSEN, 2008). Alternativamente, a estimação por horizonte móvel (Moving Horizon Estimation, MHE) formula a estimação de estados como um problema de otimização restrita em uma janela finita de medições passadas, minimizando o desvio entre saídas medidas e previstas sujeito à dinâmica do modelo e a restrições físicas, sendo amplamente utilizada em conjunto com controle preditivo baseado em modelo (RAWLINGS; MAYNE; DIEHL, 2017).

Uma forma comum de utilização dos estimadores de estado em bioprocessos é acoplando softsensors que fornecem medições on-line a partir das varreduras de sondas espectrométricas, e junto a essas medições, utilizar um modelo dinâmico das reações no processo (SIMUTIS et al., 2014; HERMANN e KREMLING, 2025; SCHIEMER et al., 2023; ALEXANDER et al., 2023; KRAMER e KING, 2017; LOPEZ et al., 2021; STELZER et al., 2017). Nesse contexto, diversas abordagens híbridas têm sido propostas: Krämer e King (2017) combinaram espectroscopia NIR, regressão PLS e um SPKF em cultivos de S. cerevisiae, integrando espectros NIR e outras medições on-line no SPKF e mostrando que o método híbrido supera tanto o PLS isolado quanto o estimador puramente baseado no modelo na estimação de biomassa, glicose, etanol, amônio e fosfato; López et al. (2021), acoplaram espectroscopia ATR-MIR, modelos PLS e um EKF contínuo-discreto para monitorar fermentações lignocelulósicas de etanol, obtendo reduções significativas de erro e maior robustez em relação ao PLS e ao modelo cinético interno; Schiemer et al. (2023) fundiram um sensor quimiométrico baseado em UV/Vis-GPR com um modelo cinético via EKF para acompanhar uma reação de conjugação de anticorpos, com redução de até cerca de 23% no erro de predição frente ao modelo cinético sozinho; e Hermann e Kremling (2025) usaram previsões PLSR de variáveis on-line como medições em um UKF acoplado a um modelo coarse-grained, melhorando de forma marcante a estimação da concentração de L-fenilalanina em relação ao CGM isolado; Alexander et al. (2023) demonstraram que uma formulação de MHE baseada em parâmetros (P-MHE) reduz erro e tempo computacional em relação ao MHE tradicional, compete em erro com formulações baseadas em EKF e é adequada para monitoramento em tempo real graças à robustez, garantia de viabilidade e tempos de cálculo de poucos segundos por medição. Estudos comparativos em bioprocessos mostram ainda que filtros de partículas com número adequado de partículas (da ordem de dezenas a milhares, como 2500) podem alcançar erros menores que o EKF, embora com custo computacional substancialmente maior (STELZER et al., 2017; SIMUTIS et al., 2014).

2.4.1. FILTRO DE KALMAN ESTENDIDO (EKF)

O Filtro de Kalman Estendido (EKF, Extended Kalman Filter) é um estimador não linear recursivo, utilizado para inferir estados de um sistema dinâmico a partir de medições ruidosas, consiste principalmente em duas etapas (CAMPANI, 2018): etapa de predição no tempo e atualização com base em medidas do processo (Equações 1 e 2). Na primeira etapa, o estado anterior $x_{k-1|k-1} \sim N(\hat{x}_{k-1|k-1}, P_{k-1|k-1})$ é propagado um passo a frente no tempo, dando $x_{k|k-1} \sim N(\hat{x}_{k|k-1}, P_{k|k-1})$. A informação da medida atual $y_k = Cx_k + \upsilon_k$, contendo ruído $\upsilon \sim N(0, R)$, é então usada para melhorar a estimativa de $\hat{x}_{k|k-1}$ para $\hat{x}_{k|k} \sim N(\hat{x}_{k|k}, P_{k|k})$. P e R são as matrizes de covariância associadas aos estados e às medidas. Considerou-se que o ruído do processo $w_k = x_k - f(x_{k-1}, u_{k-1})$ apresenta igualmente distribuição normal com média zero e covariância Q_k. Também considerou-se a utilização da covariância dos parâmetros do modelo cinético Ω_k para estimar a covariância do estado. Mais informação sobre o EKF pode ser encontrada em Rawlings et al., (2020).

$$\text{predição pelo modelo} \begin{cases} \hat{x}_{k|k-1} = f\left(\hat{x}_{k-1|k-1}, u_{k-1}\right) \\ P_{k|k-1} = A_k P_{k-1|k-1} A_k^T + G_k Q G_k^T + B_k \Omega B_k^T \end{cases} \quad (4)$$

$$\text{atualização baseada na medida} \begin{cases} \hat{x}_{k|k} = \hat{x}_{k|k-1} + L_k(y_k - C\hat{x}_{k|k-1}) \\ P_{k|k1} = P_{k|k-1} A_k^T - L_k C P_{k|k-1} \end{cases} \quad (5)$$

$$\hat{x}_{0|0} = x_0,\ P_{0|0} = P_0 \quad (6)$$

$$L_k = P_{k|k-1} C^T \left(C P_{k|k+1} C^T + R\right)^{-1} \quad (7)$$

Na notação utilizada, $x \sim N(\hat{x}, P)$ significa que a variável x é considerada estocástica com distribuição normal, média $\hat{x}$ e matriz de covariância P. Definimos $\hat{x}_{k|k-1}$ como sendo a estimativa de estado no passo k dada a informação do processo antes de k, e $\hat{x}_{k|k}$ como sendo a estimativa em k dada a medida y_k.

2.4.2. FILTRO DE PARTÍCULAS (PF)

Um Filtro de Partículas (PF, Particle Filter) é um método de Monte Carlo sequencial (SMC) que aproxima em tempo real, a distribuição de estados de um modelo dinâmico a partir de um conjunto finito de amostras ponderadas (“partículas”), atualizadas à medida que chegam novas observações (JOHANSEN, 2009). O PF é especialmente útil para modelos não-lineares e não gaussianos. O PF pode ser implementado de várias maneiras, mas de uma forma geral é possível aplicá-lo em 3 etapas: Amostragem das partículas preditas pelo modelo, atualização dos pesos com base nas medidas e reamostragem sistemática (CHEN, 2003), demonstradas a seguir.

predição do modelo e amostragem $$\begin{cases} x_{k|k-1}^{(i)} = f\left(x_{k-1|k-1}^{(i)}, u_{k-1}, \theta_{k-1}, q_{k-1}\right) \\ q_k \sim N(0, Q), \theta_k \sim N(\hat{\theta}, \Omega) \end{cases} \quad (8)$$

onde o índice (i) se refere às particulas, e o conjunto $\left\{x_{k|k-1}^{(i)}, w_{k|k-1}^{(i)}\right\}$ aproxima a distribuição $p\,(\,x_k|y_{1:k-1})$. Após efetuar uma medida, y_k as partículas recebem um peso proporcional à verossimilhança da medida dado o estado:

Atualização de pesos baseada na medida $$\begin{cases} \widehat{w}_{k|k}^{(i)} = w_{k-1|k-1}^{(i)} l_k \\ w_{k|k}^{(i)} = \frac{\widehat{w}_{k|k}^{(i)}}{\Sigma \widehat{w}_{k|k}^{(i)}} \end{cases} \quad (9)$$

onde $l_k = p(y_k \mid x_{k|k-1}^{(i)})$ é a verossimilhança, calculada assumindo um ruído gaussiano nas medidas.

A etapa de reamostragem sistemática é uma forma de evitar a degenerescência dos pesos (CHEN et al., 2004). De forma simplificada, ela faz com que partículas com maior peso associado sejam mais frequentes na distribuição final das partículas, e que as de menor peso tendam a desaparecer. Isso é feito construindo posições igualmente espaçadas no intervalo [0,1) e, para cada posição, escolhendo o primeiro índice da função de distribuição cumulativa dos pesos que seja maior ou igual a ela.

3. MATERIAIS E MÉTODOS

3.1. SÍNTESE DE GOS

A síntese de GOS foi realizada por Maione (2024). Cada síntese foi feita em reator de volume útil de 25 mL encamisado por meio da hidrolização enzimática da lactose. A enzima utilizada foi a β-galactosidase de Aspergillus oryzae (Sigma-Aldrich). A temperatura de reação foi de 40 °C, estabilizada com o auxílio de banho termostático. O tampão utilizado para a estabilização do pH a 4,5 foi o citrato-fosfato 100 mM. Para este trabalho, os conjuntos de dados utilizados foram referentes a somente 6 bateladas diferentes (1, 2, 3, 4, 6 e 8), devido a problemas significativos nas varreduras e concentrações das demais bateladas. Suas especificações estão na Tabela 1:

Tabela 1 Condições iniciais de cada batelada

Batelada	Concentração inicial de lactose [M]	Concentração inicial de enzima [M]
1	0,691	$1{,}20 \times 10^{-7}$
2	0,500	$2{,}06 \times 10^{-8}$
3	0,500	$2{,}34 \times 10^{-8}$
4	0,324	$2{,}29 \times 10^{-8}$
5	0,103	$3{,}45 \times 10^{-8}$
6	0,691	$1{,}70 \times 10^{-7}$
7	0,500	$1{,}65 \times 10^{-8}$
8	0,500	$7{,}50 \times 10^{-9}$
9	0,103	$1{,}55 \times 10^{-8}$

Fonte: Adaptado de Maione 2024

3.2. OBTENÇÃO DE DADOS EXPERIMENTAIS

Os dados experimentais foram coletados por Maione (2024). O monitoramento espectroscópico do reator foi feito em tempo real, *in situ*, por meio de uma sonda UV-Vis acoplada a um espectrofotômetro de arranjo de diodos (Lambda 465, PerkinElmer). A sonda, com caminho óptico de 5 mm, foi imersa diretamente no meio reacional para aquisição contínua dos espectros; água destilada e tampão foram

utilizados como brancos. As varreduras UV-vis passaram por uma etapa de pré-tratamento. Utilizou-se o filtro de Savitzky-Golay (1ª derivada, polinômio de terceira ordem e comprimentos de onda entre 270-350 nm).

As análises off-line das concentrações foram realizadas por HPLC. O teor de lactose no permeado do soro foi determinado em coluna de troca iônica SUPELCOGEL Ca2+ (30 cm × 7,8 mm), mantida a 80 °C, com vazão de 0,5 mL/min e água Milli-Q como fase móvel. A quantificação empregou detector de índice de refração (RID) e curva de calibração externa obtida com padrões comerciais de lactose.

3.3. CALIBRAÇÃO DE MODELO CINÉTICO

O modelo cinético adotado neste trabalho foi proposto por Schultz et al. (2021). O modelo trata de um mecanismo fenomenológico para a síntese de GOS catalisada por β-galactosidase de *Kluyveromyces lactis*, com nove parâmetros ajustáveis e inativação enzimática de primeira ordem, mostrando bom ajuste e validação em altas concentrações de lactose. Como o modelo utiliza a β-galactosidase de K. Lactis, houve um reajuste para a reação utilizando a β-galactosidase de *Aspergillus oryzae*. A calibração do modelo para possibilitar o uso desta enzima foi realizada por Maione (2024). As seguintes variáveis de estado (x) são descritas pelo modelo: Concentração de lactose (Lac), glicose (Glu), galactose (Gal), galactobiose (Glb), GOS3 (Tri), GOS4 (Tet), galactotriose (Trig), galactotetraose (Tetg) e enzima total (ET). Tri e Tet correspondem a GOS com glicose terminal. As equações do modelo estão a seguir:

$$r(Lac) = E\left[\frac{k_h}{K_{MH}} Tri - Lac\left(\frac{k_{cat}}{K_M} + \frac{\gamma \times k_t}{K_{MT}}\right)\right] \quad (10)$$

$$r(Glu) = E\left(\frac{k_{cat} \times Lac}{K_M}\right) \quad (11)$$

$$r(Glb) = E\left[\gamma \times k_t \times \left(\frac{Gal}{K_{MGal}} - \frac{Glb}{K_{MT}}\right) + \frac{k_H}{K_{MT}} \times (Trig - Glb)\right] \quad (12)$$

$$r(Tri) = E\left[\gamma \times \frac{k_t}{K_{MT}} \times (Lac - Tri) + \frac{k_H}{K_{MH}} \times (Tet - Tri)\right] \quad (13)$$

$$r(Trig) = E\left[\gamma \times \frac{k_t}{K_{MT}} \times (Glb - Trig) + \frac{k_H}{K_{MH}} \times (Tetg - Trig)\right] \quad (14)$$

$$r(Tet) = E\left(\gamma \times \frac{k_t}{K_{MT}} \times Tri - \frac{k_H}{K_{MH}} \times Tet\right) \quad (15)$$

$$r(Tetg) = E\left(\gamma \times \frac{k_t}{K_{MT}} \times Trig - \frac{k_H}{K_{MH}} \times Tetg\right) \quad (16)$$

$$r(E_T) = -k_E \times E_T \quad (17)$$

$$\gamma = \frac{\frac{k_{cat}}{K_M} \times\ Lac + \frac{k_H}{K_{MH}} \times\ a}{k_{cat'} + k_T \times \left(\frac{Gal}{K_{MGal}} + \frac{b}{K_{MT}}\right)} \quad (18)$$

$$a = Glb + Tri + Trig + Tet + Tetg \quad (19)$$

$$b = Lac + Glb + Tri + Trig \quad (20)$$

$$E = \frac{E_T}{1 + \frac{Gal}{K_I} + \frac{Lac}{K_M} + \frac{a}{K_{MH}} + \gamma\left(1 + \frac{Gal}{K_{MGal}} + \frac{b}{K_{MT}}\right)} \quad (21)$$

Onde r(i) representa a velocidade de reação da espécie i ($mol.L^{-1}.min^{-1}$). Os parâmetros do modelo estão presentes na Tabela 2.

Tabela 2 Parâmetros obtidos para o modelo cinético utilizando o método de Monte-Carlo

Parâmetro	Unidade	Valor
$k_{cat} \times 10^{-7}$	(min^1)	1,76
$k_{cat'} \times 10^{-9}$	(min^1)	3,43
$K_{MH} \times 10^{5}$	(mol/L)	2,05
$K_{MT} \times 10^{6}$	(mol/L)	1,50
$k_H \times 10^{-4}$	(min^1)	4,41
$k_T \times 10^{-4}$	(min^1)	6,23
$K_I \times 10^{6}$	(mol/L)	6,12
$K_M \times 10^{1}$	(mol/L)	1,06
$K_{MGal} \times 10^{6}$	(mol/L)	1,30
$k_E \times 10^{6}$	(min^1)	7,04

Fonte: Adaptado de Maione 2024

3.4. MODELO DE CALIBRAÇÃO MULTIVARIADA

Entre os modelos de calibração, a Regressão por Mínimos Quadrados Parciais (PLS) é especialmente adequada para o tipo de dado em questão por lidar com muitos preditores colineares e ruidosos, mantendo um custo computacional moderado e boa interpretabilidade (WOLD et al., 2001). Em estudos de síntese de GOS, o uso de UV-Vis combinado a PLS já foi demonstrado com sucesso para quantificar lactose e GOS totais ao longo do processo (DIAS et al., 2009). Métodos mais complexos, como certas redes neurais, podem demandar maior ajuste de hiperparâmetros e conjuntos adicionais de validação, e mesmo um bom desempenho em teste não garante robustez preditiva em dados futuros (WESTAD; FLÅTEN, 2024). Para a calibração multivariada, foi utilizado um modelo de PLS para cada componente a ser inferido no sistema.

3.5. PARÂMETROS DO EKF

A matriz de covariância do erro das medidas R, foi obtida a partir do resultado do erro da validação cruzada. Para a matriz Q, os elementos de sua diagonal principal foram ajustados manualmente. Os testes utilizando a matriz Q em sua forma inteira (matriz simétrica) foram descartados com base em experiência prévia em um trabalho de iniciação científica, onde observou-se que seu efeito de utilizar uma covariância entre as Equações de estado foi pouco expressivo. A matriz Ω, de covariância paramétrica, foi obtida durante o ajuste por método baseado em Monte Carlo. Para simplificar as comparações, os efeitos das matrizes Q e Ω foram estudados de maneira isolada, portanto em nenhuma simulação houve utilização das duas simultaneamente, embora as equações permitam isso.

3.6. PARÂMETROS E CONFIGURAÇÕES DO PARTICLE FILTER

O Particle Filter utiliza, além dos mesmos parâmetros do EKF, um parâmetro adicional que é o número de partículas N, e também algumas configurações como a distribuição inicial de partículas e o modo de escolha da estimativa final. O número de partículas deve ser suficiente para obter uma boa representação da distribuição de estados e o aumento de partículas tende a fornecer melhores estimativas em troca

de um custo computacional maior. Para problemas similares, 2500 mostrou ser um número adequado (STELZER et al., 2017; SIMUTIS et al., 2014). Para evitar problemas oriundos de uma representaçao insuficiente das distribuições, um elevado número de partículas foi escolhido, com N = 20000, pois isso não prejudicou o tempo das simulações significativamente. A distribuição inicial de partículas foi escolhida como sendo normal, com média igual ao valor das concentrações iniciais e variância igual a 1% da concentração inicial para lactose, e 5% da concentração inicial para a enzima total (as demais variáveis são nulas no inicio). A outra configuração é a forma de fornecer a estimativa final, nesta implementação foi escolhida a média das partículas ponderadas pelos pesos.

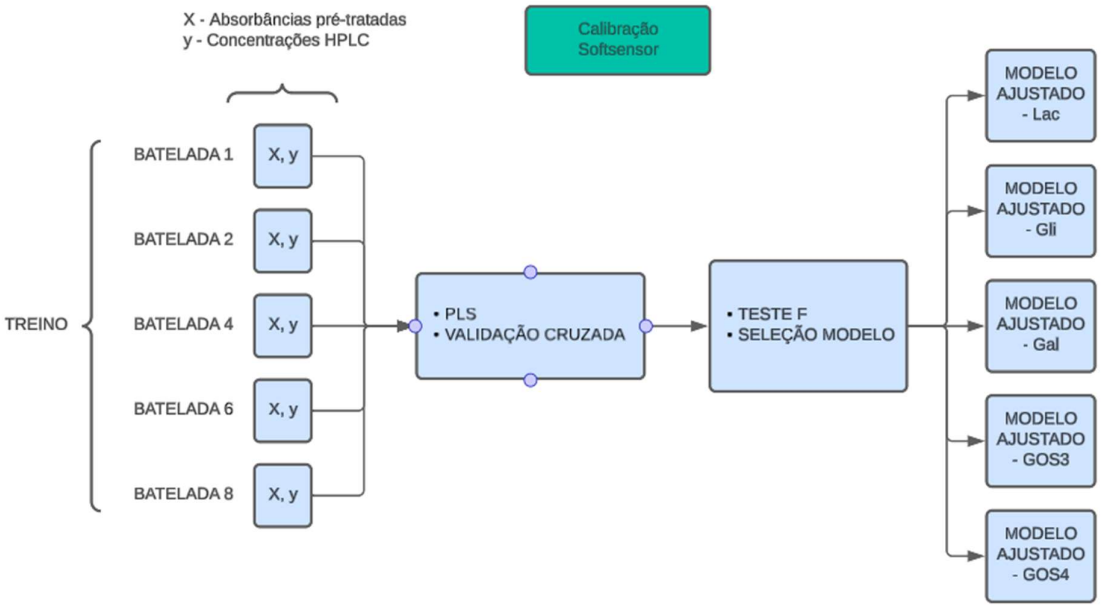
3.7. AVALIAÇÃO DE DESEMPENHO DOS ESTIMADORES

Para avaliar o desempenho dos estimadores, as configurações comuns a ambos foram alteradas da seguinte forma: utilização de R como diagonal e inteira; Ω como diagonal e inteira; Q como diagonal. Então realizou-se uma simulação por configuração utilizando cada uma das bateladas como conjuntos de teste, com ambos os estimadores. A métrica esolhida para mensurar desempenho foi o NRMSE, visto que fornece uma maneira normalizada e uniforme de comparar erros entre espécies. Como cada simulação estimou os valores de todas as espécies, comparações entre distintas simulações utilizaram a média do NRMSE entre as espécies. No caso das simulações testadas com bateladas presentes no conjunto de treino, realizou-se a média - daquele NRMSE médio entre as espécies – entre as simulações, como forma de agregação para representar o desempenho total do conjunto de treinamento. No total, com 2 opções para R; 1 para Q; 2 para Ω; 2 para o estimador (PF ou EKF) e 6 bateladas, foram realizadas 72 simulações (12 utilizando o conjunto de validação externa e 60 utilizando os demais). Adicionalmente, foram feitas duas simulações (uma para cada estimador, após encontrado suas respectivas configurações ótimas) considerando uma condição inicial quatro vezes maior na concentração e incerteza de enzima, como forma de simular um cenário onde o conhecimento do início da batelada é mais incerto e possui um maior desvio da realidade.

3.8. IMPLEMENTAÇÃO DOS MÉTODOS

A linguagem escolhida para efetuar as implementações foi python 3.12, devido à grande variedade de bibliotecas e funcionalidades oferecidas. A implementação do softsensor é dada em dois módulos: A calibração e a inferência. Na calibração, o algoritmo recebe todos os dados de treinamento iniciais, e por meio da validação cruzada e seleção de modelo, fornece os modelos ótimos de PLS para cada espécie, conforme o fluxograma da Figura 1.

Figura 1 - Fluxograma geral da calibração do softsensor.



Fonte: Elaboração própria

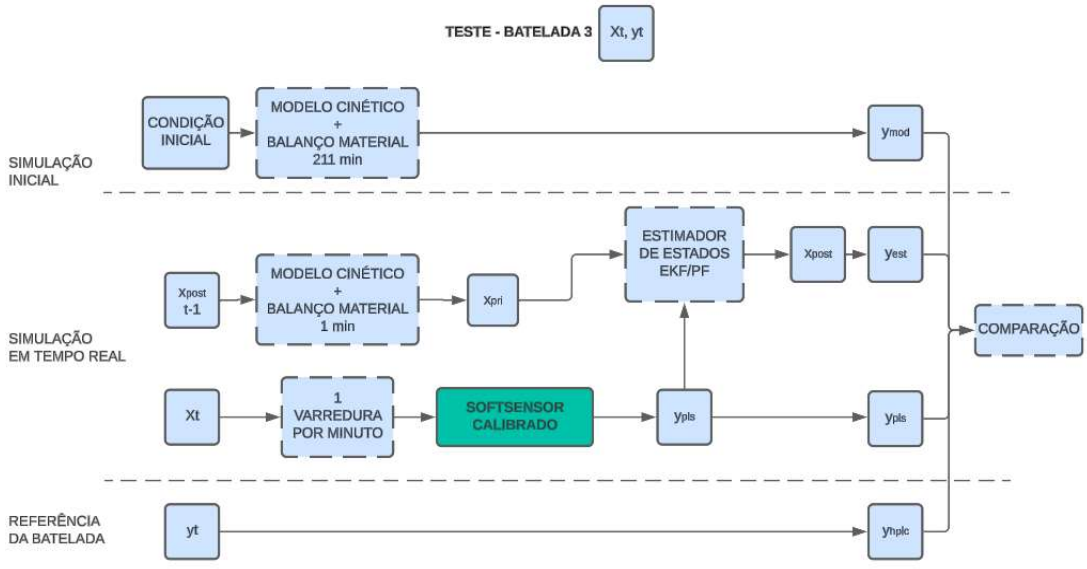
A Inferência do softsensor é realizada por uma função que recebe os dados do conjunto de teste, e aplica as transformações conforme os modelos de PLS ajustados para obter a inferência das concentrações.

Para os estimadores de estado, a implementação foi feita com classes independentes, o que possibilitou desacoplar a lógica de estimação de estados com a lógica do softsensor, bem como a utilização de diferentes estimadores no mesmo fluxo de simulação de processo. A implementação do EKF se deu de forma simples,

seguindo as respectivas equações (4, 5, 6, e 7), com uma etapa de previsão e uma de atualização. Por outro lado, o filtro de partículas demandou uma complexidade maior, e além de utilizar uma etapa adicional (reamostragem sistemática), sua implementação utilizou paralelização via GPU para funcionar em um tempo aceitável, visto que este estimador precisa resolver N sistemas de equações diferenciais não-lineares a cada intervalo de tempo.

A Figura 2 tem o fluxograma que representa o processo inteiro de simulação, inferência de concentrações por softsensor, estimativa de estados e comparação final para o conjunto de teste. A simulação inicial parte da condição inicial do conjunto de dados para obter as curvas do modelo cinético no tempo. A simulação em tempo real utiliza o modelo para obter as estimativas a priori x_{pri} (partindo das estimativas a posteriori anteriores $x_{post\,|\,t-1}$) e extrai dados do conjunto de varreduras pré-tratadas a cada minuto (com o intuito de simular uma sonda em funcionamento) para obter a inferência das concentrações y_{pls}, que são alimentadas ao estimador junto com x_{pri} a cada intervalo de tempo. A referência da batelada é apenas o conjunto de concentrações de HPLC que são usados para fins de comparação e avaliação de erros.

Figura 2 - Fluxograma da simulação, inferência do softsensor e estimativa de estados.



Fonte: Elaboração própria.

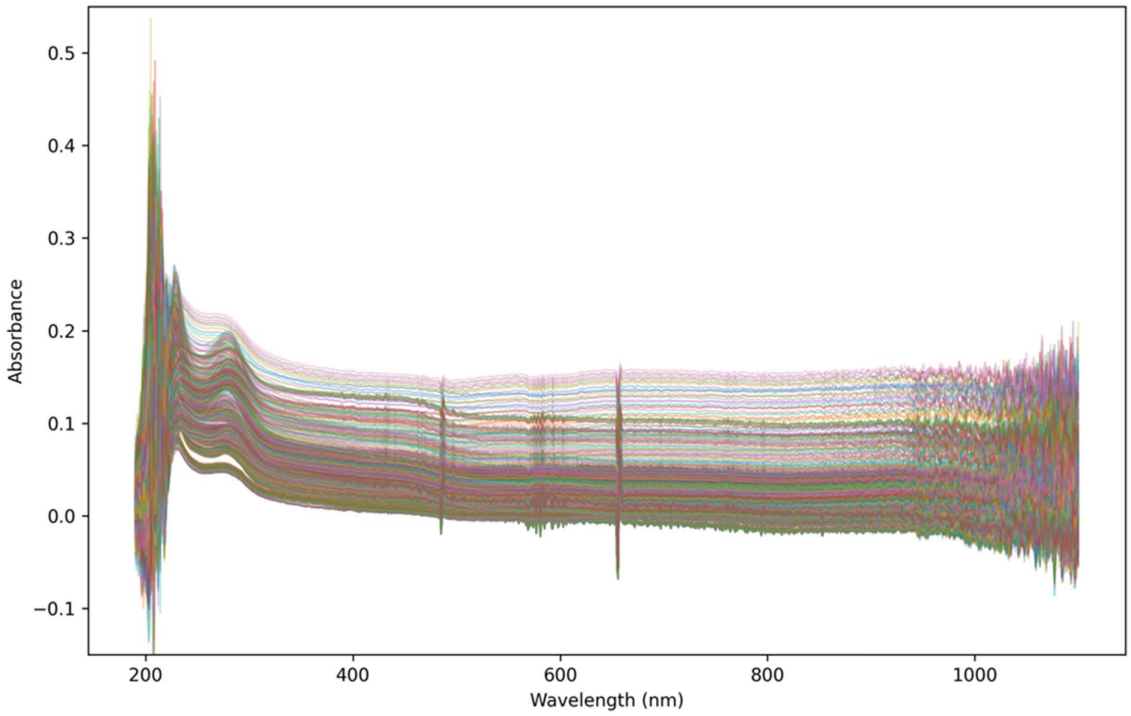
4. RESULTADOS

4.1. ANÁLISE DOS DADOS INICIAIS

4.1.1. VARREDURAS UV-VIS

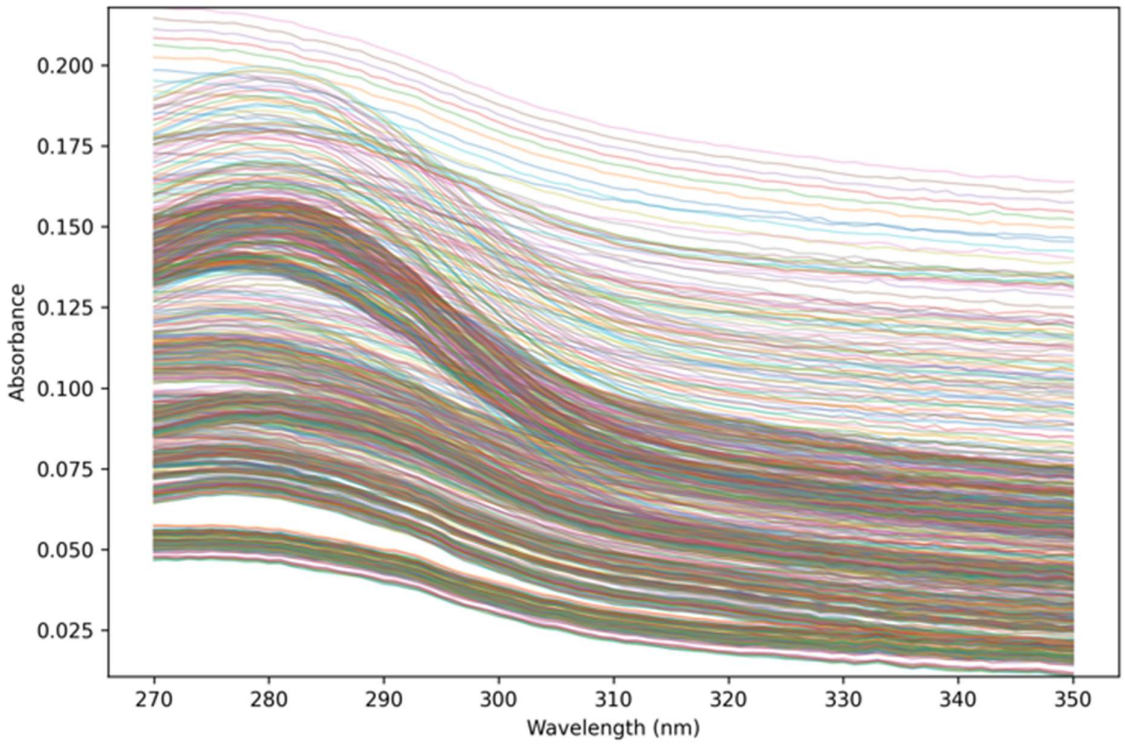
As varreduras brutas apresentaram uma região inicial muito ruidosa, e uma grande região com pouca variação para comprimentos de onda maiores que 350 nm. A Figura 3 mostra os espectros sobrepostos para todas as bateladas, os espectros individuais estão no Apêndice A. As Figuras 3 e 4 mostram os espectros brutos para todos os comprimentos de onda, e os selecionados, respectivamente

Figura 3 - Espectros brutos para todas as bateladas.



Fonte: Dados adaptados de Maione (2024).

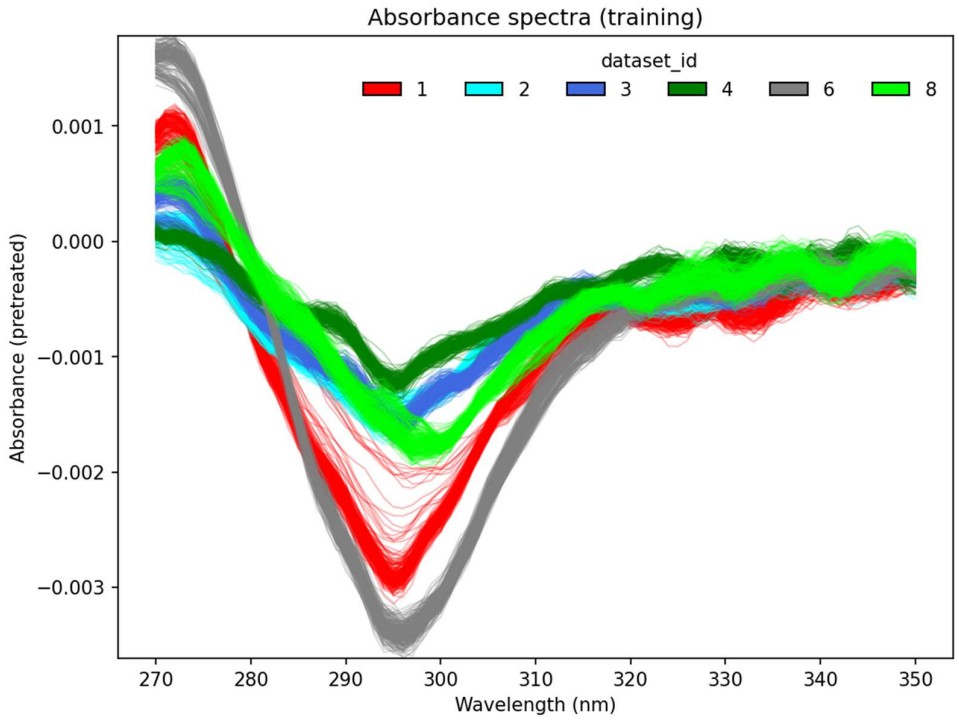
Figura 4 - Espectros brutos para todas as bateladas na região entre 270-350 nm.



Fonte: Dados adaptados de Maione (2024).

Observou-se também que os espectros brutos possuem alguns problemas como diferentes linhas de base e curvas ruidosas, o que dificulta a identificação de picos para o ajuste. Ambos problemas são minimizados pela ação da derivada e interpolação na etapa de pré-tratamento. Para as varreduras pré-tratadas observou-se uma grande diferença entre a maioria dos espectros de cada batelada, tanto em tamanho de picos/vales quanto na posição dos mesmos, indicando uma região ampla de ajuste que inclui comportamentos não-lineares. Também observou-se significativa sobreposição de espectros para comprimentos de onda superiores a 320 nm, como revelado na Figura 5. Uma exceção foram as bateladas 2 e 3, para estas, percebeu-se que há uma grande semelhança entre as duas. Uma possível explicação para esse fato é que ambas possuem condições iniciais muito similares.

Figura 5 - Varreduras pré-tratadas coloridas por batelada.

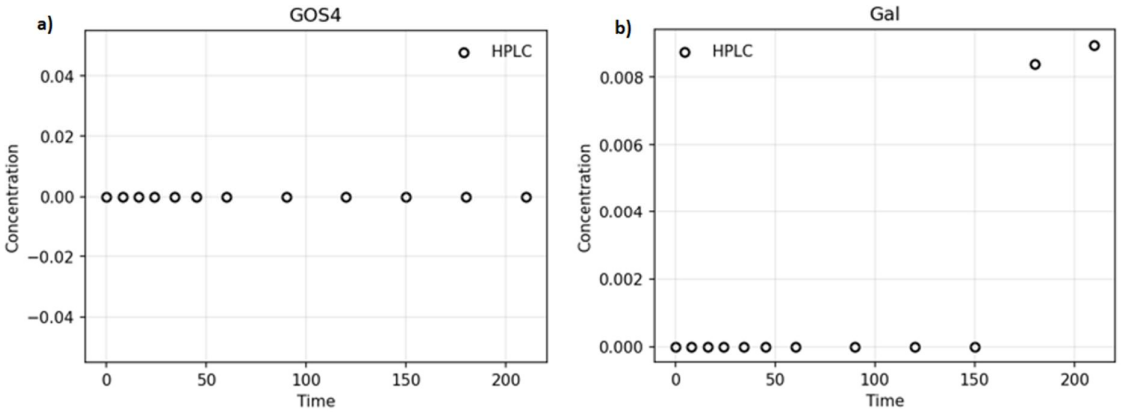


Fonte: Dados adaptados de Maione, 2024

4.1.2. CONCENTRAÇÕES HPLC

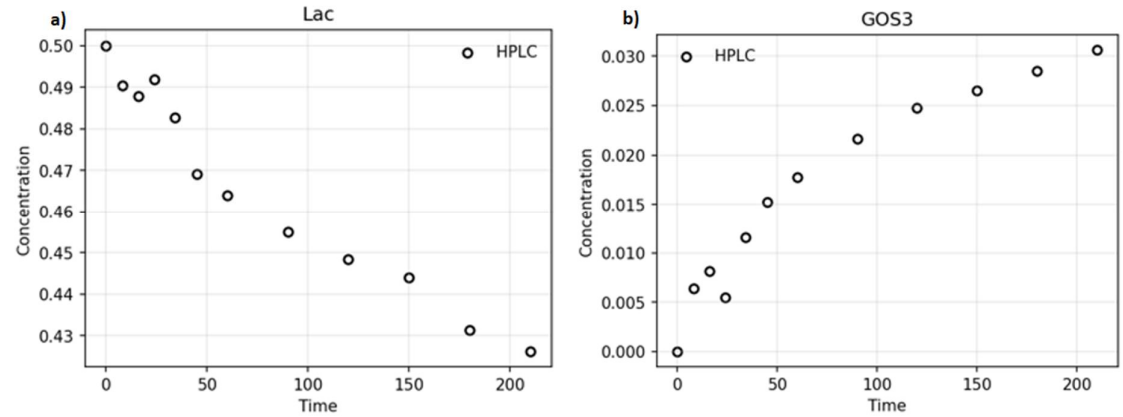
As concentrações obtidas através do HPLC, de modo geral, apresentarm perfis esperados com os da reação em questão. No entanto, em algumas bateladas observaram-se dois problemas: espécies com concentrações abaixo do limite de detecção do HPLC, o que resultou em zeros indevidos; e problemas de quantificação das áreas do cromatograma, ambos evidenciados na batelada 8, conforme mostrado na Figura 6 e Figura 7.

Figura 6 - Espécies abaixo do limite de detecção do HPLC - batelada 8.



Fonte: Dados adaptados de Maione (2024).

Figura 7 - Erros de quantificação de áreas do cromatograma no quarto ponto - batelada 8.

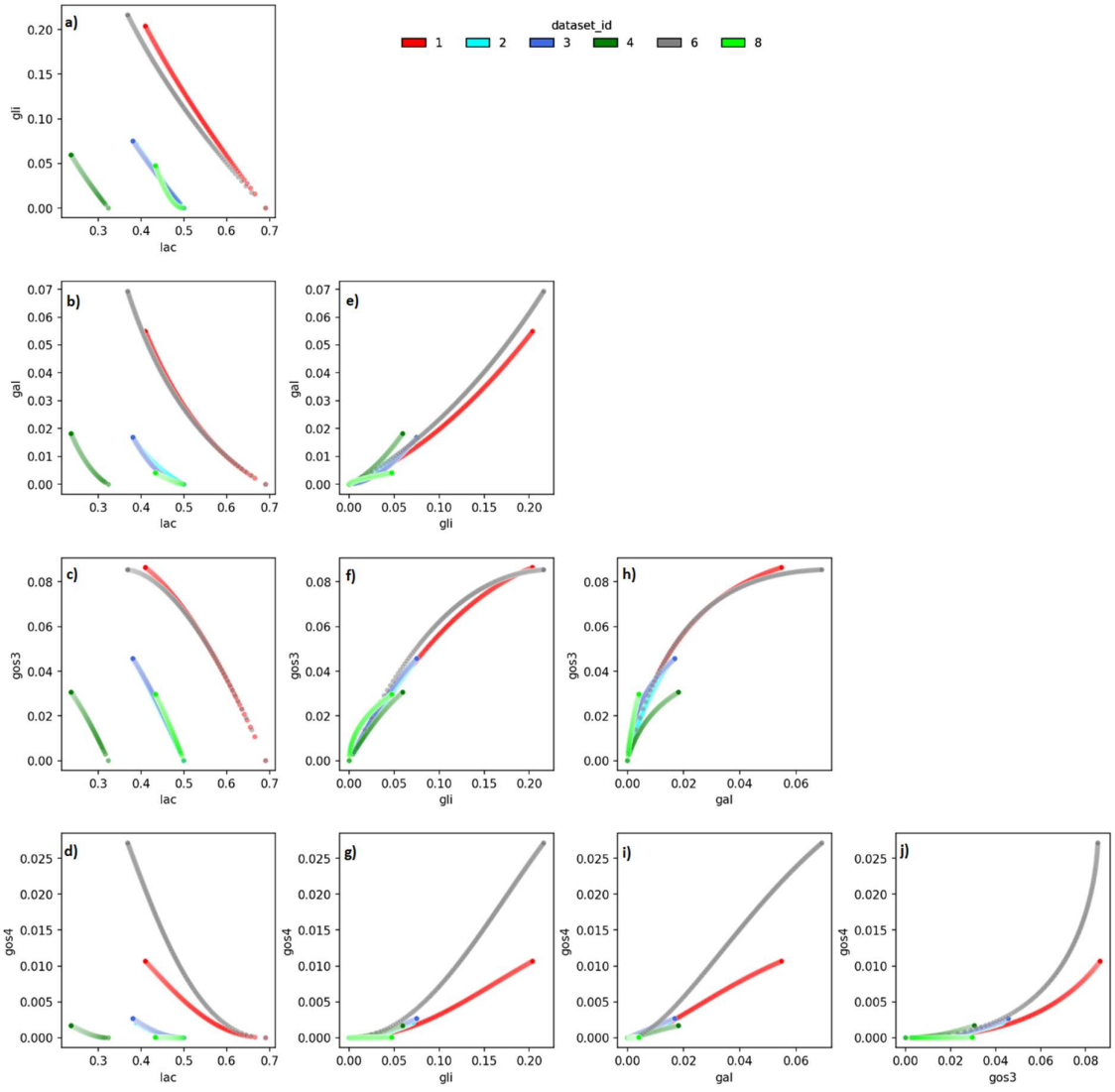


Fonte: Dados adaptados de Maione (2024).

Uma análise realizada com base nos dados de HPLC é a de correlação entre concentrações de espécies, presente na Figura 8. Esta análise permitiu observar como as concentrações entre algumas espécies foram descorrelacionadas ao considerar todo o conjunto de dados de treinamento, bem como mostrar a semelhança entre as bateladas 2 e 3 em termos de perfis de espécies. A Figura 9 mostrou como a lactose foi totalmente descorrelacionada das demais espécies ao considerar as bateladas utilizadas no ajuste, bem como a forte correlação negativa ao considerar uma única batelada. Esta análise também mostra a importância de utilizar diversos conjuntos de dados oriundos de experimentos com condições iniciais

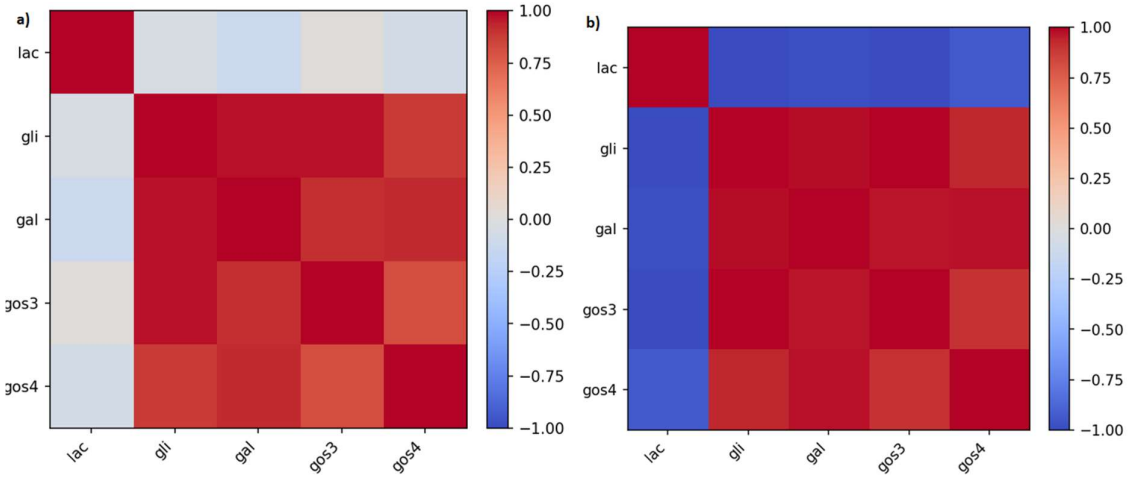
variadas. Tomando como exemplo a primeira coluna de gráficos da Figura 8, observou-se que em cada batelada a lactose possui forte correlação negativa com as demais espécies, o que é esperado pois ela é a única espécie consumida na reação. Entretanto ao observar todo o conjunto, a adição de bateladas em diferentes condições iniciais fez com que variações na lactose não impliquem estritamente em aumento ou diminuição de outras espécies. Para o ajuste do PLS, uma forte correlação poderia implicar em um modelo em que uma variável determina o comportamento e as demais se tornam redundantes.

Figura 8 - Análise de correlação entre espécies, eixos em mol/L da respectiva espécie, colorido por batelada.



Fonte: Elaboração própria.

Figura 9 - Matriz de correlação entre concentração de espécies do conjunto de treinamento (esquerda) e de validação externa (direita).

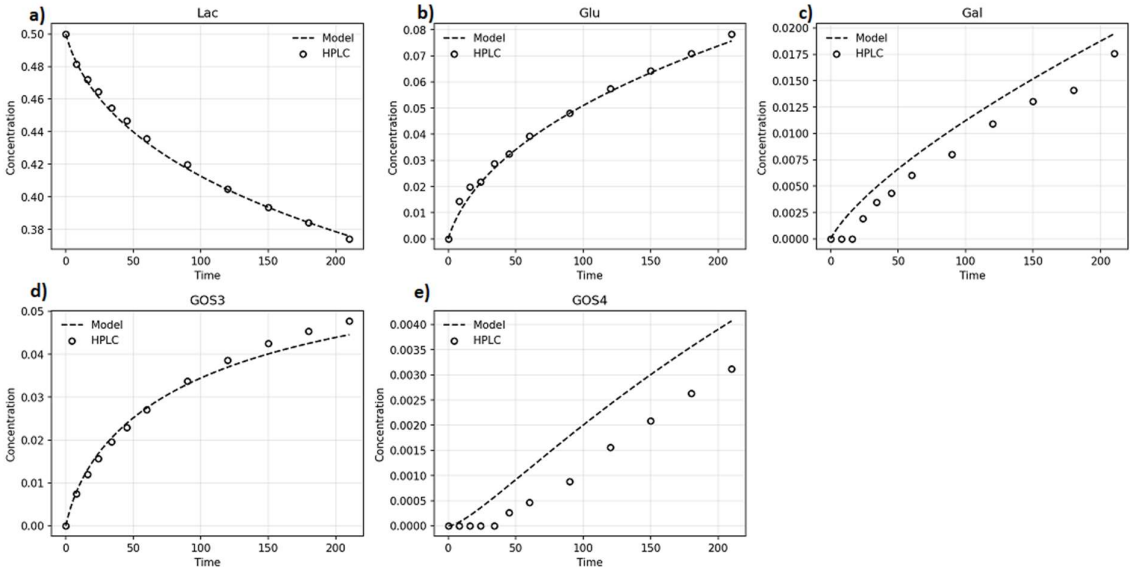


Fonte: Elaboração própria.

4.2. MODELO CINÉTICO

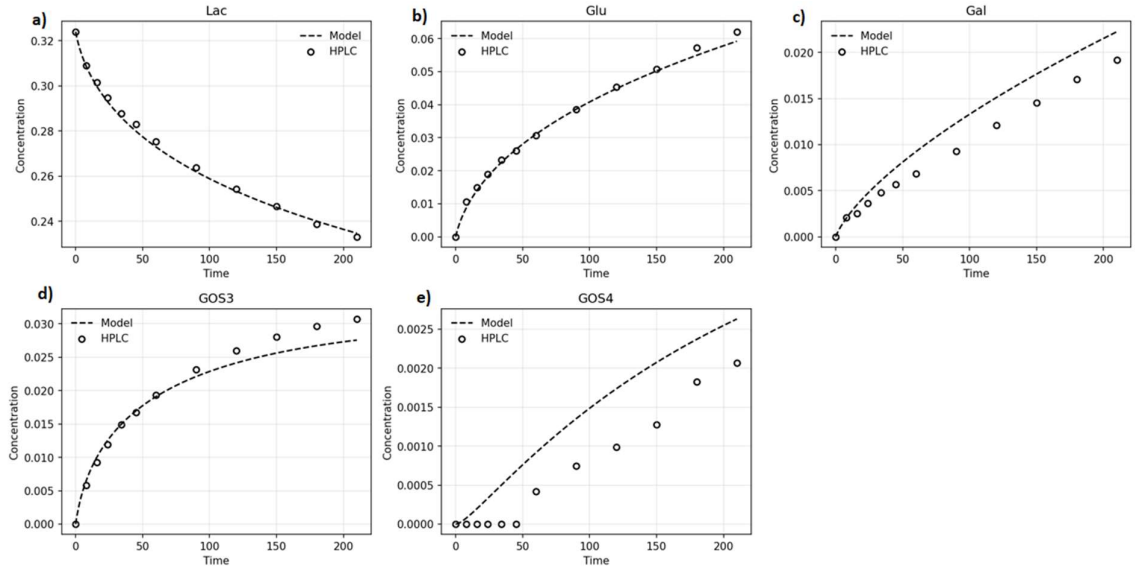
O modelo cinético apresentou uma boa resposta para a maioria das espécies frente aos dados experimentais de acordo com as simulações. Notou-se ocorrência de desvio sistemático para o GOS4 em todas as bateladas, e também para a galactose nas bateladas 3 e 4, embora em uma intensidade menor. Isso pode ser explicado pela dificuldade do ajuste dessas variáveis, visto que sua presença no meio reacional é significativamente menor que as outras, e frequentemente passam por problemas de detecção. A Figura 11 também mostra um pequeno desvio do modelo nos pontos finais para o GOS3.

Figura 10 - Concentrações no tempo para batelada 3. Modelo cinético vs HPLC.



Fonte: Elaboração Própria

Figura 11 - Concentrações no tempo para batelada 4. Modelo cinético vs HPLC.



Fonte: Elaboração Própria

O valor do NRMSE presente na Tabela 3, ilustra o desempenho do modelo para todas as espécies em cada batelada, e na média de todas as bateladas. Observou-se que de modo geral, a lactose, glicose e GOS3 obtiveram respostas muito boas, enquanto o GOS4 obteve um erro significativo em todas as bateladas, e a galactose

teve um erro médio elevado devido à batelada 8, onde a maioria dos pontos sofreu por problemas de detecção conforme ilustrado na Figura 6 a) e b).

Tabela 3 - NRMSE da resposta do modelo comparada com dados de HPLC para todas as espécies, em todas as bateladas.

	1	2	3	4	6	8	Média
Lac	0.047	0.038	0.015	0.019	0.030	0.050	0.033
Glu	0.044	0.051	0.026	0.021	0.041	0.120	0.051
Gal	0.054	0.064	0.114	0.116	0.039	0.345	0.122
GOS3	0.051	0.043	0.034	0.052	0.034	0.078	0.049
GOS4	0.322	0.336	0.215	0.253	0.158	N/A	0.257

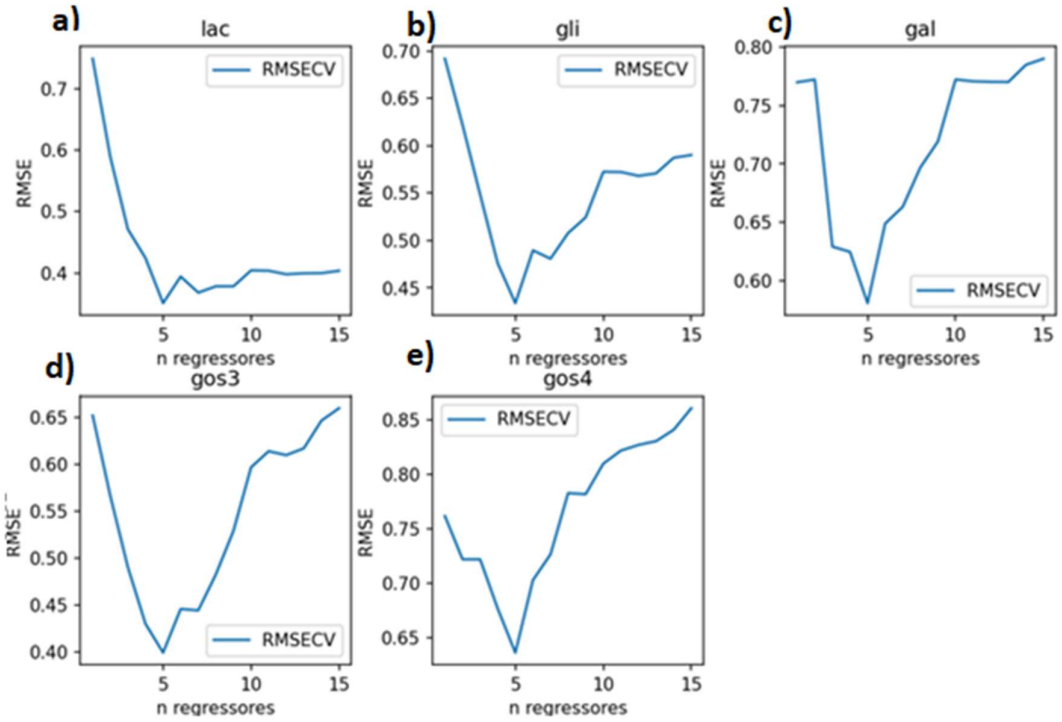
Fonte: Elaboração própria.

4.3. SOFTSENSOR

4.3.1. SOFTSENSOR – CALIBRAÇÃO

A calibração do softsensor pode ser avaliada em termos dos resultados da validação cruzada. Os resultados obtidos mostraram que foi alcançado um balanço entre complexidade dos modelos de PLS e capacidade de inferência. A Figura 12 mostra um claro ponto de aumento de erro de validação a partir de 5 regressores nos modelos de cada espécie.

Figura 12 - Gráficos de RMSE normalizado vs número de regressores para a validação cruzada.



Fonte: Elaboração própria.

A realização do teste F selecionou 5 regressores em todos os modelos, mostrando que o ganho em performance de 4 para 5 regressores é estatisticamente significativo para o nivel de significância escolhido. Este resultado implica na redução de dimensionalidade do problema: em vez de utilizar 81 comprimentos de onda diretamente, os modelos de PLS utilizaram 5 regressores que condesam a informação dos espectros. Este ponto reforça uma das vantagens do PLS que é a sua utilização em problemas que lidam com matrizes de alta colinearidade. Embora modelos ótimos de PLS tenham sido selecionados, os erros de validação obtidos ainda foram relativamente altos, como mostrados pela Tabela 4.

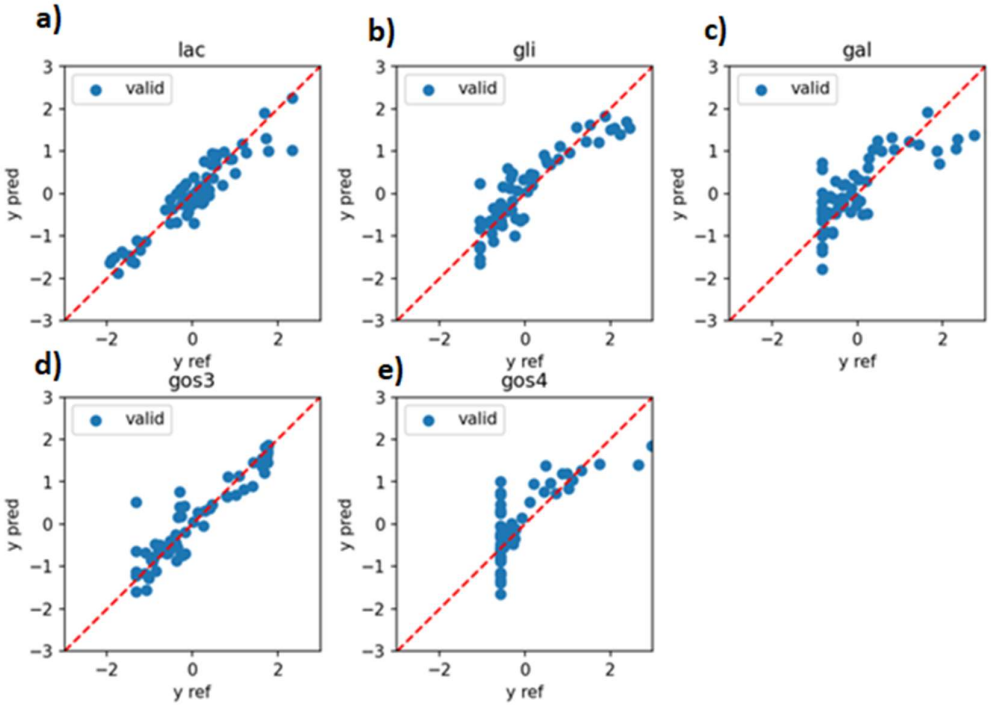
Tabela 4 - NRMSE obtido comparando concentrações preditas com as de referência na validação cruzada, para o melhor modelo de cada espécie.

	Lac	Gli	Gal	GOS3	GOS4
RMSECV	0.351	0.434	0.581	0.400	0.636

Fonte: Elaboração Própria.

Esse fato pode ser explicado pela abrangência dos conjuntos de treinamento, que possuem espectros muito variados, e regiões não lineares que dificultam o ajuste do PLS, que é essencialmente um modelo linear. As espécies GOS4 e galactose apresentaram erros especialmente altos, devido à presença de zeros indevidos em algumas bateladas. Estes erros ficam evidentes na Figura 13, onde é possível ver fileiras verticais para o menor valor de y_{ref}. Um possível ponto de melhoria para futuros trabalhos seria realizar a remoção destes zeros antes da realização do ajuste.

Figura 13 - Gráficos de concentrações normalizadas previstas para cada fold da validação cruzada (y_{pred}) vs reservadas para validação (y_{ref}) com linha de bissetriz para comparação.

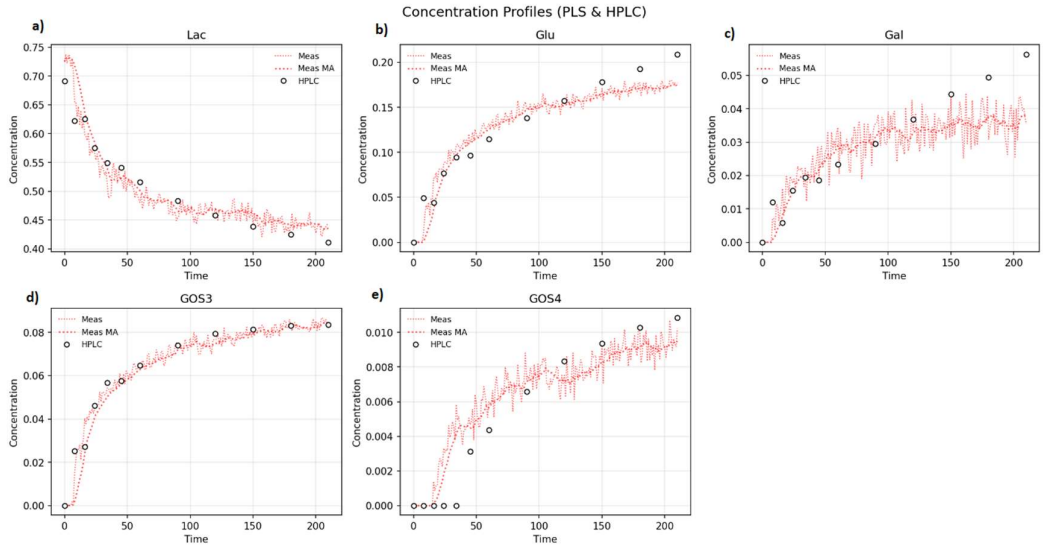


Fonte: Elaboração própria.

4.3.2. SOFTSENSOR – INFERÊNCIA

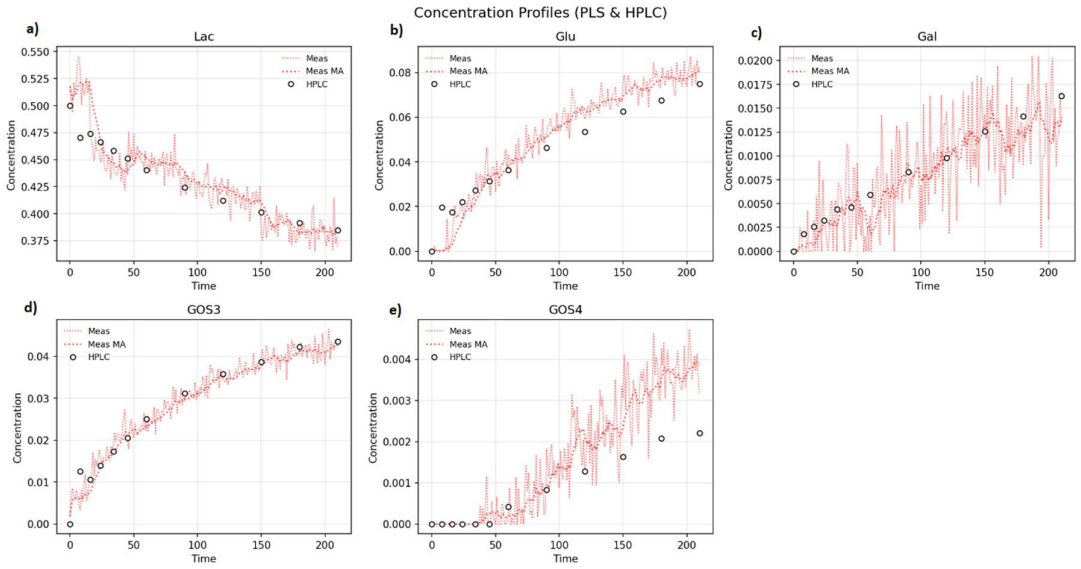
A inferência predita pelo PLS para as bateladas presentes no conjunto de treinamento forneceu resultados mistos, conforme mostram as Figuras 14, 15, 16, 17 e 18. A batelada utilizada como validação externa está presente na Figura 19. Para as bateladas 1 e 2, o resultado foi bom, com uma inferência que se aproximou dos pontos experimentais.

Figura 14 - Gráficos de concentração no tempo para batelada 1 - PLS vs HPLC.



Fonte: Elaboração própria.

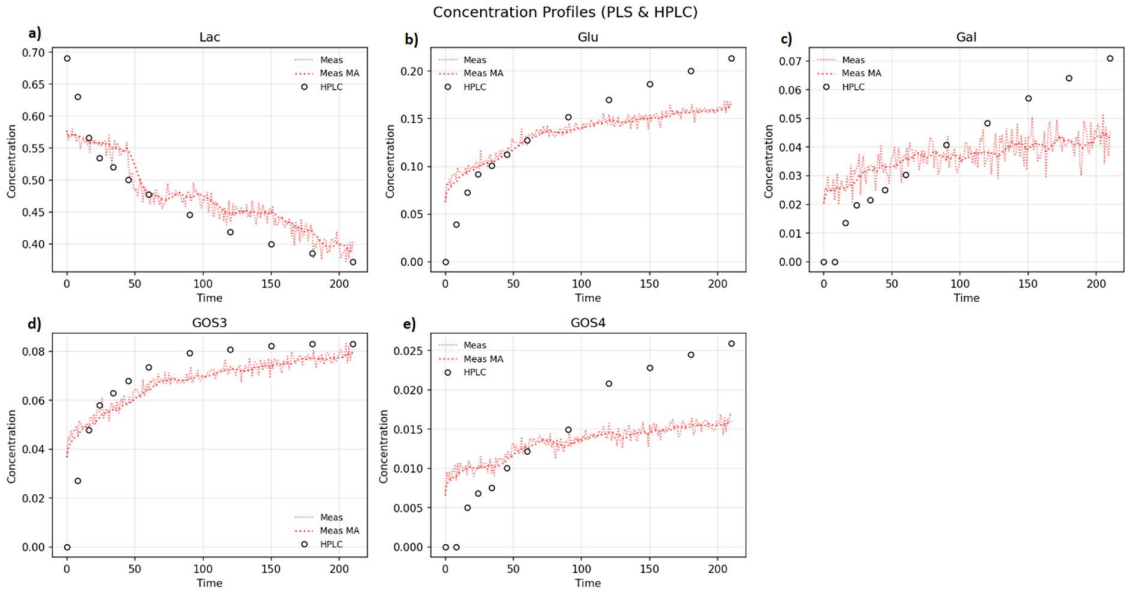
Figura 15 - Gráficos de concentração no tempo para batelada 2 - PLS vs HPLC.



Fonte: Elaboração própria.

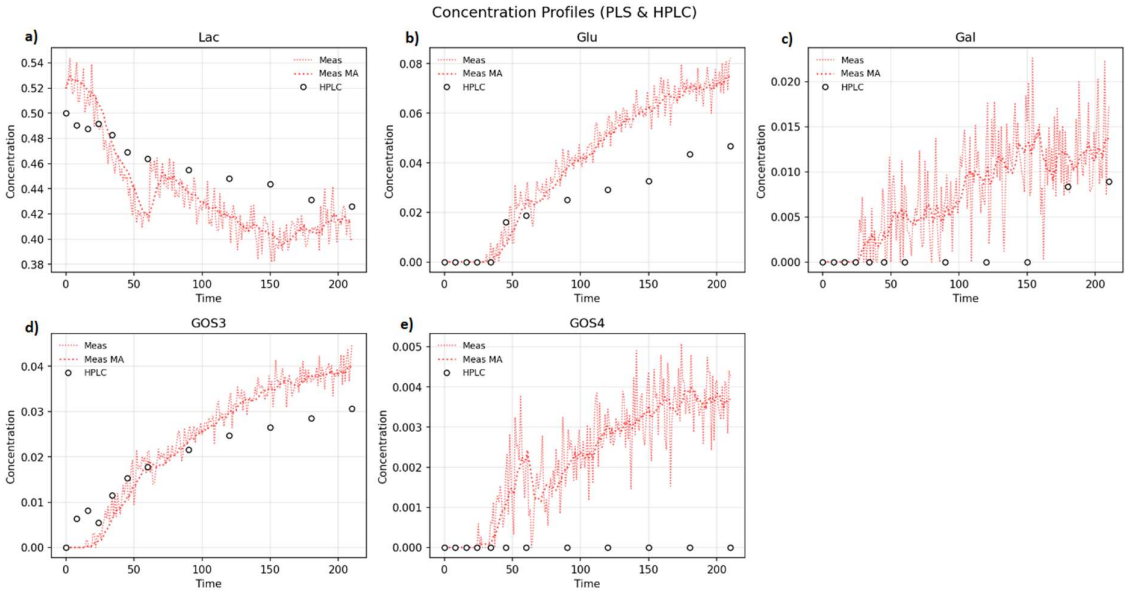
Para as bateladas 6 e 8, foi possível observar as tendências de queda da lactose e de aumento nos produtos, no entanto com um desvio significativo dos dados experimentais.

Figura 16 - Gráficos de concentração no tempo para batelada 6 - PLS vs HPLC.



Fonte: Elaboração própria.

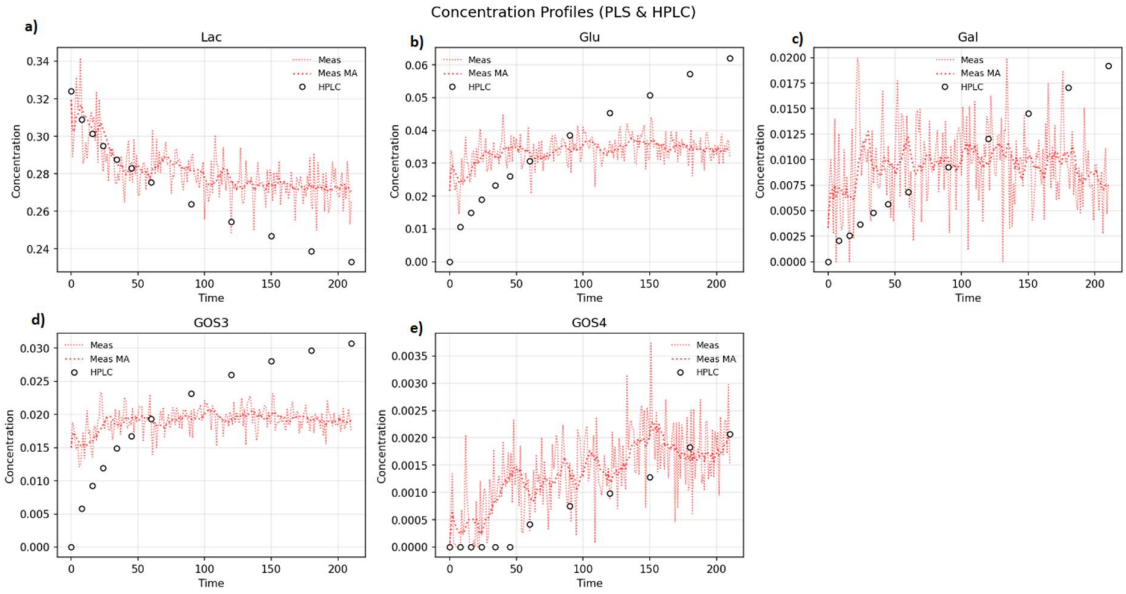
Figura 17 - Gráficos de concentração no tempo para batelada 8 - PLS vs HPLC.



Fonte: Elaboração própria.

Já para a batelada 4, o PLS conseguiu captar parte da tendência para a lactose e GOS4, falhando para as demais.

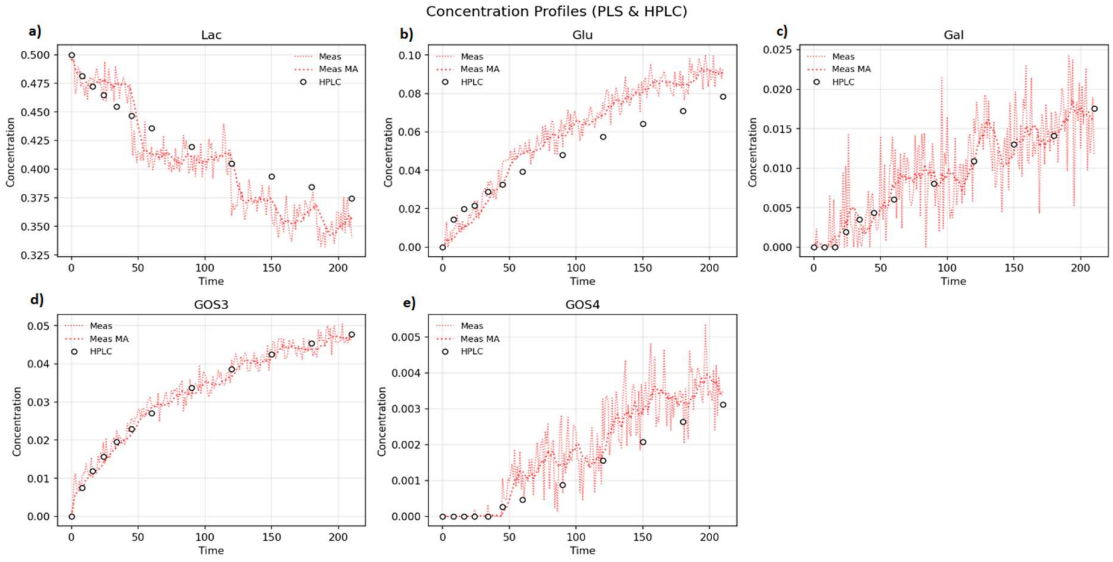
Figura 18 - Gráficos de concentração no tempo para batelada 4 - PLS vs HPLC.



Fonte: Elaboração própria.

Para o conjunto de validação externa (batelada 3), os resultados foram próximos dos pontos experimentais, mesmo o PLS não tendo sido ajustado com informações deste conjunto. Este ponto revela a capacidade preditiva do modelo quando a região de ajuste possui conjuntos suficientemente parecidos.

Figura 19 - Gráficos de concentração no tempo para batelada 3 (validação externa) - PLS vs HPLC.



Fonte: Elaboração própria.

Independente da batelada, o PLS produziu resultados com grande ruído, mostrado pela variação dos picos e vales em torno de sua média móvel. A utilização da média móvel para realizar comparações forneceu resultados mais estáveis no cálculo das métricas. A Tabela 5 mostra as métricas de NRMSE entre a inferência do PLS e HPLC, nela o erro do GOS4 para a batelada 8 foi removido devido aos zeros indevidos nos dados de HPLC, vale notar que nesta batelada o erro da galactose permaneceu alto devido ao mesmo problema.

Tabela 5 - NRMSE entre inferência do PLS e dados de HPLC para cada espécie em todas as bateladas.

	Conjuntos de treino					Validação Externa
	1	2	4	6	8	3
Lac	0.118	0.186	0.208	0.145	0.371	0.118
Glu	0.094	0.121	0.245	0.153	0.351	0.127
Gal	0.142	0.101	0.289	0.227	0.629	0.060
GOS3	0.089	0.054	0.265	0.164	0.200	0.025
GOS4	0.144	0.327	0.286	0.240	N/A	0.135

Fonte: Elaboração própria.

Uma observação importante ao comparar a Tabela 5 com a Tabela 3 é que o erro do modelo cinético tende a ser significativamente menor que o erro do PLS em quase todos os casos. Vale ressaltar que o modelo cinético é de natureza fenomenológica, e foi ajustado especificamente para esses conjuntos de dados (incluindo o conjunto que está sendo usado como validação externa no PLS) utilizando relações não-lineares de maior complexidade.

4.4. ESTIMADORES DE ESTADO

Os parâmetros obtidos para os estimadores são as matrizes de covariância R, Q e Ω. Seus valores podem ser encontrados nas Tabelas 6, 7 e 8 respectivamente. Para a matriz Q apenas os elementos da diagonal principal foram utilizados.

Tabela 6 - Elementos da matriz R.

1.41E-03	-8.10E-04	-3.13E-04	-3.11E-04	-1.09E-04
-8.10E-04	6.98E-04	2.68E-04	2.21E-04	9.18E-05
-3.13E-04	2.68E-04	1.27E-04	7.28E-05	3.72E-05
-3.11E-04	2.21E-04	7.28E-05	1.15E-04	2.40E-05
-1.09E-04	9.18E-05	3.72E-05	2.40E-05	1.69E-05

Fonte: Elaboração própria.

Tabela 7 - Elementos da diagonal principal da matriz Q.

1.00E-10	1.00E-16	1.00E-08	1.00E-08	1.00E-08	1.00E-16	1.00E-16	1.00E-16	1.00E-22	1.00E-08

Fonte: Elaboração própria.

Tabela 8 - Elementos da Matriz Ω

1.68E+13	3.95E+15	-5.34E+00	-6.84E-02	-1.66E+07	1.47E+11	-4.41E+00	-4.25E+03	-1.38E-01	0.00E+00
3.95E+15	1.99E+19	-1.50E+03	-7.45E+02	-1.36E+12	3.16E+14	-2.73E+03	-3.43E+07	-1.68E+01	0.00E+00
-5.34E+00	-1.50E+03	2.93E-12	1.39E-13	9.30E-05	-3.04E-02	5.50E-13	6.41E-08	9.57E-14	0.00E+00
-6.84E-02	-7.45E+02	1.39E-13	2.81E-13	-1.52E-04	1.74E-02	2.34E-13	4.75E-09	1.37E-16	0.00E+00
-1.66E+07	-1.36E+12	9.30E-05	-1.52E-04	7.78E+05	2.73E+07	-2.54E-04	2.40E+00	3.59E-05	0.00E+00
1.47E+11	3.16E+14	-3.04E-02	1.74E-02	2.73E+07	2.29E+10	-1.96E-02	1.45E+03	7.27E-03	0.00E+00
-4.41E+00	-2.73E+03	5.50E-13	2.34E-13	-2.54E-04	-1.96E-02	7.14E-12	8.07E-08	7.61E-14	0.00E+00
-4.25E+03	-3.43E+07	6.41E-08	4.75E-09	2.40E+00	1.45E+03	8.07E-08	7.65E-03	7.97E-09	0.00E+00
-1.38E-01	-1.68E+01	9.57E-14	1.37E-16	3.59E-05	7.27E-03	7.61E-14	7.97E-09	1.15E-13	0.00E+00
0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.56E-10

Fonte: Elaboração própria.

O último parâmetro relacionado a matriz de covariância é a constante de ativação enzimática, k_e, que foi ajustada de forma independente aos demais. Por isso não há correlação entre o ajuste de k_e com os demais parâmetros. Uma análise complementar para a matriz Ω, é sobre a correlação dos parâmetros do modelo cinético, presente na Figura 20. Esta figura revelou que existem elementos correlacionados fora da diagonal principal, o que indicou que a utilização desta matriz na sua forma inteira pode implicar em um melhor aproveitamento da informação disponível.

Figura 20 - Módulo da correlação dos parâmetros do modelo cinético, colorida do vermelho ao verde de acordo com a escala.

1.00	0.22	0.76	0.03	0.00	0.24	0.40	0.01	0.10	0.00
0.22	1.00	0.20	0.32	0.35	0.47	0.23	0.09	0.01	0.00
0.14	0.20	1.00	0.15	0.06	0.12	0.12	0.43	0.16	0.00
0.03	0.32	0.15	1.00	0.33	0.22	0.17	0.10	0.00	0.00
0.00	0.35	0.06	0.33	1.00	0.20	0.11	0.03	0.12	0.00
0.24	0.47	0.12	0.22	0.20	1.00	0.05	0.11	0.14	0.00
0.40	0.23	0.12	0.17	0.11	0.05	1.00	0.35	0.08	0.00
0.01	0.09	0.43	0.10	0.03	0.11	0.35	1.00	0.27	0.00
0.10	0.01	0.16	0.00	0.12	0.14	0.08	0.27	1.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Fonte: Elaboração própria.

A variação de cada parâmetro no estimador resultou em 12 configurações possíveis para o estimador, e com um conjunto fixo de treinamento e 6 possíveis bateladas para realização do teste (5 presentes no conjunto de treino e uma reservada para validação externa), o número de simulações realizadas foi de 72, 12 para as que foram testadas no conjunto de validação externa e 60 para as demais. A performance de cada configuração de estimador foi avaliada pela média do NRMSE entre cada espécie e cada batelada nos conjuntos de simulações. Gerando as tabelas 9 e 10.

Tabela 9 - Resultados agregados para simulações sobre o conjunto de treinamento. Da esquerda pra direita, identificador da configuração do estimador; NRMSE médio para Estimador, PLS e Modelo Cinético; forma de utilização da matriz R, Q e Ω.

Config	NRMSE_est	NRMSE_pls	NRMSE_mod	Estimador	Matriz R	Matriz Q	Matriz Ω
0	0.107	0.215	0.100	PF	Diag.	Diag.	
6	0.107	0.215	0.100	PF	Int.	Diag.	
1	0.111	0.215	0.100	EKF	Diag.	Diag.	
7	0.119	0.215	0.100	EKF	Int.	Diag.	
3	0.139	0.215	0.100	EKF	Diag.		Diag.
10	0.140	0.215	0.100	PF	Int.		Int.
5	0.142	0.215	0.100	EKF	Diag.		Int.
9	0.146	0.215	0.100	EKF	Int.		Diag.
11	0.146	0.215	0.100	EKF	Int.		Int.
4	0.149	0.215	0.100	PF	Diag.		Int.
8	0.157	0.215	0.100	PF	Int.		Diag.
2	0.160	0.215	0.100	PF	Diag.		Diag.

Fonte: Elaboração própria

Tabela 10 - Resultados agregados para simulações sobre o conjunto de validação externa

Config	NRMSE_est	NRMSE_pls	NRMSE_mod	Estimador	Matriz R	Matriz Q	Matriz Ω
1	0.101	0.093	0.081	EKF	Diag.	Diag.	
6	0.102	0.093	0.081	PF	Int.	Diag.	
0	0.105	0.093	0.081	PF	Diag.	Diag.	
7	0.113	0.093	0.081	EKF	Int.	Diag.	
5	0.116	0.093	0.081	EKF	Diag.		Int.
3	0.119	0.093	0.081	EKF	Diag.		Diag.
4	0.134	0.093	0.081	PF	Diag.		Int.
2	0.147	0.093	0.081	PF	Diag.		Diag.
10	0.157	0.093	0.081	PF	Int.		Int.
8	0.170	0.093	0.081	PF	Int.		Diag.
11	0.186	0.093	0.081	EKF	Int.		Int.
9	0.191	0.093	0.081	EKF	Int.		Diag.

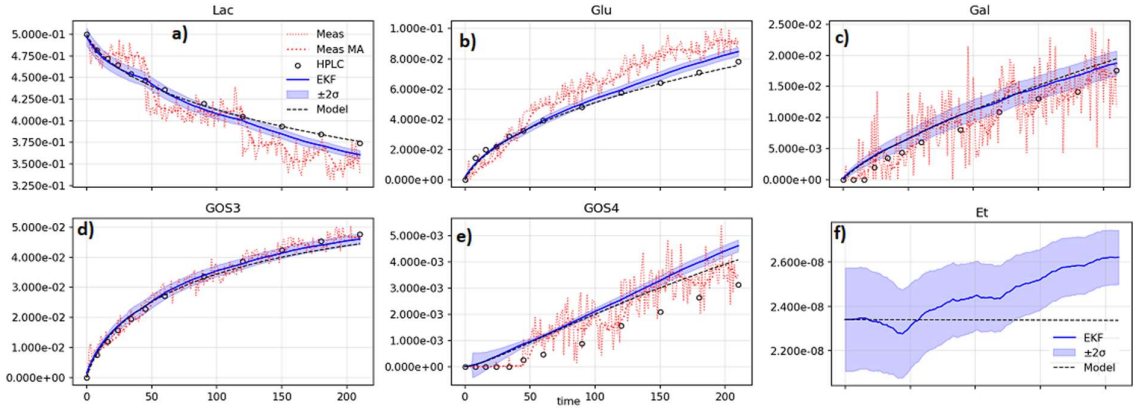
Fonte: Elaboração própria

As tabelas 9 e 10 mostraram um resultado interessante, para as quatro melhores configurações de estimador, a diferença em performance não foi tão significativa. O principal efeito notado é na diferença de utilizar a matriz Q versus Ω, os resultados parecem indicar que a utilização de Q fornece maior precisão, tanto no conjunto de treinamento quanto no de validação externa.

4.4.1. EKF

Abordando o EKF, escolheu-se a configuração 1 para avaliar a simulação sobre o conjunto de validação externa, mostrada nas figuras 21 e 22. As métricas das simulações estão na Tabela 11.

Figura 21 - Gráficos para a simulação da batelada 3 utilizando a configuração 1. Concentrações medidas por PLS; média móvel de 10 pontos do PLS; HPLC; EKF; bandas de incerteza e modelo cinético.



Fonte: Elaboração própria.

Tabela 11 - NRMSE do EKF, PLS e Modelo Cinético para todas as espécies e suas médias. Simulação da batelada 3 utilizando configuração 1.

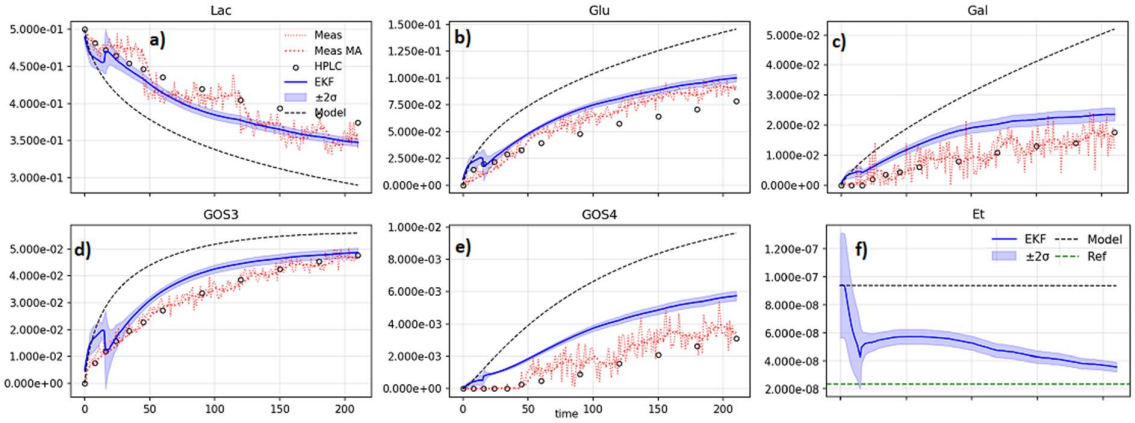
	NRMSE_est	NRMSE_pls	NRMSE_mod
Lac	0.053	0.118	0.015
Glu	0.044	0.127	0.026
Gal	0.106	0.060	0.114
GOS3	0.025	0.025	0.034
GOS4	0.275	0.135	0.215
Média	0.101	0.093	0.081

Fonte: Elaboração Prórpria

A análise da Figura 21 e Tabela 11 permitem observar que o estimador inicialmente tende ao comportamento do modelo e gradualmente vai sendo corrigido pelas leituras do softsensor, e para as espécies de maior concentração, a estimativa de estados fornece um resultado mais preciso do que a utilização puramente do softsensor. Por incorporar informações do modelo, o EKF também fornece um resultado com menor variância, assemelhando-se mais do comportamento real do processo. Outra vantagem do EKF (e dos estimadores de estado em geral) é revelada na Figura 21: a estimativa da atividade da enzima em tempo real. Essa é uma informação que o softsensor não tem capacidade de inferir, e o modelo cinético é capaz de estimar, mas somente para uma dada condição inicial. A estimativa em tempo real dessa variável permite uma análise mais profunda a respeito do processo, neste caso, observou-se que ela aumentou comparado com a sua condição inicial. Uma possível explicação para esse fato é que as medidas do softsensor indicam que a concentração de lactose é menor que o previsto pelo modelo, e o oposto ocorre para as concentrações de produto, indicando que a reação ocorreu a uma taxa um pouco mais elevada do que a prevista pelo modelo inicialmente, o que corresponderia a uma estimativa maior da enzima.

Uma análise adicional oriunda desta capacidade de estimar a enzima total, é a de verificar a capacidade de estimador de corrigir a atividade enzimática quando um valor incorreto é fornecido na condição inicial. A Figura 22 mostra a resposta da simulação abordada para uma concentração inicial de enzima total, e sua incerteza, ambas com quatro vezes o valor original.

Figura 22 - Gráficos para a simulação da batelada 3 utilizando a configuração 1 (EKF). Concentração inicial da enzima e sua incerteza aumentadas para quatro vezes seu valor original (linha verde no último gráfico).



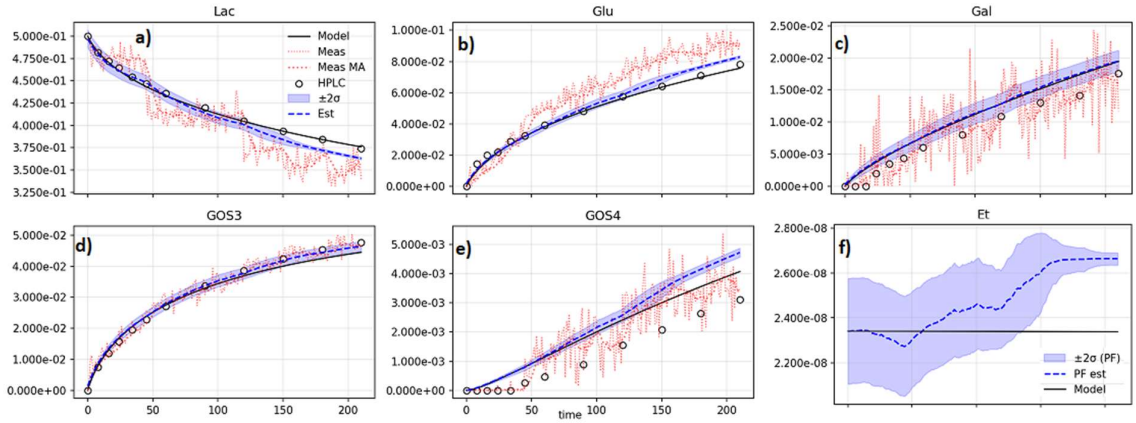
Fonte: Elaboração própria.

A Figura 22 mostra uma forte tendência em seguir o modelo nos pontos iniciais, seguida de uma rápida correção em direção aos pontos experimentais proporcionada pelo efeito das medições. Na estimativa da enzima essa correção manifestou-se como uma forte queda nos tempos iniciais seguida de um leve aumento e então uma tendência gradual ao valor original.

4.4.2. PARTICLE FILTER

No Filtro de Partículas, a configuração escolhida para avaliar a simulação sobre o conjunto de validação externa foi a 0, visto que forneceu a melhor performance nas simulações de treinamento. Os perfis gerados pelas simulações estão nas figuras 23, suas métricas nas tabelas 12 e 13.

Figura 23 - Gráficos para a simulação da batelada 3 utilizando a configuração 0. Concentrações medidas por PLS; média móvel de 10 pontos do PLS; HPLC; PF; bandas de incerteza e modelo cinético.



Fonte: Elaboração própria.

Tabela 12 - NRMSE do PF, PLS e Modelo Cinético para todas as espécies e suas médias. Simulação da batelada 3 utilizando configuração 0.

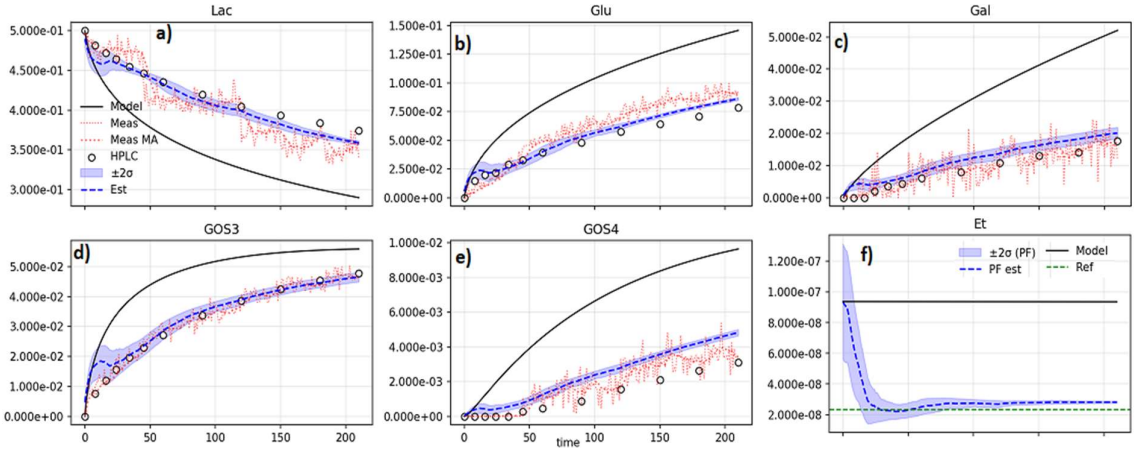
	NRMSE_est	NRMSE_pls	NRMSE_mod
Lac	0.049	0.118	0.015
Glu	0.038	0.127	0.026
Gal	0.122	0.060	0.114
GOS3	0.023	0.025	0.034
GOS4	0.295	0.135	0.215
Média	0.105	0.093	0.081

Fonte: Elaboração própria.

Observando a Figura 23 e Tabela 12 é possível notar um comportamento muito similar ao do EKF, com o estimador seguindo o modelo no inicio e sendo gradualmente guiado pelas medidas. Também observou-se que as bandas de incerteza no Particle Filter são mais estreitas, indicando uma grande aglomeração das partículas em torno da média. A estimativa da enzima também seguiu uma tendência muito similar à do EKF, com a diferença que no Particle Filter ela convergiu para um valor constante, indicado pelo encolhimento das bandas.

Realizando o mesmo teste da atribuição de valor incorreto na enzima, o Particle Filter revelou uma grande diferença em termos de capacidade de estimar a atividade enzimática, como evidenciado na Figura 24.

Figura 24 - Gráficos para a simulação da batelada 3 utilizando a configuração 0 (PF). Concentração inicial da enzima e sua incerteza aumentadas para quatro vezes seu valor original (linha verde no último gráfico).



Fonte: Elaboração própria.

A Figura 24 mostra uma correção muito rápida na atividade enzimática, e uma convergência muito próxima com o valor original. Para as demais espécies, essa correção se traduziu em uma estimativa significativamente mais próxima dos pontos experimentais do que na simulação equivalente utilizando o EKF, quantitativamente demonstrado na Tabela 13.

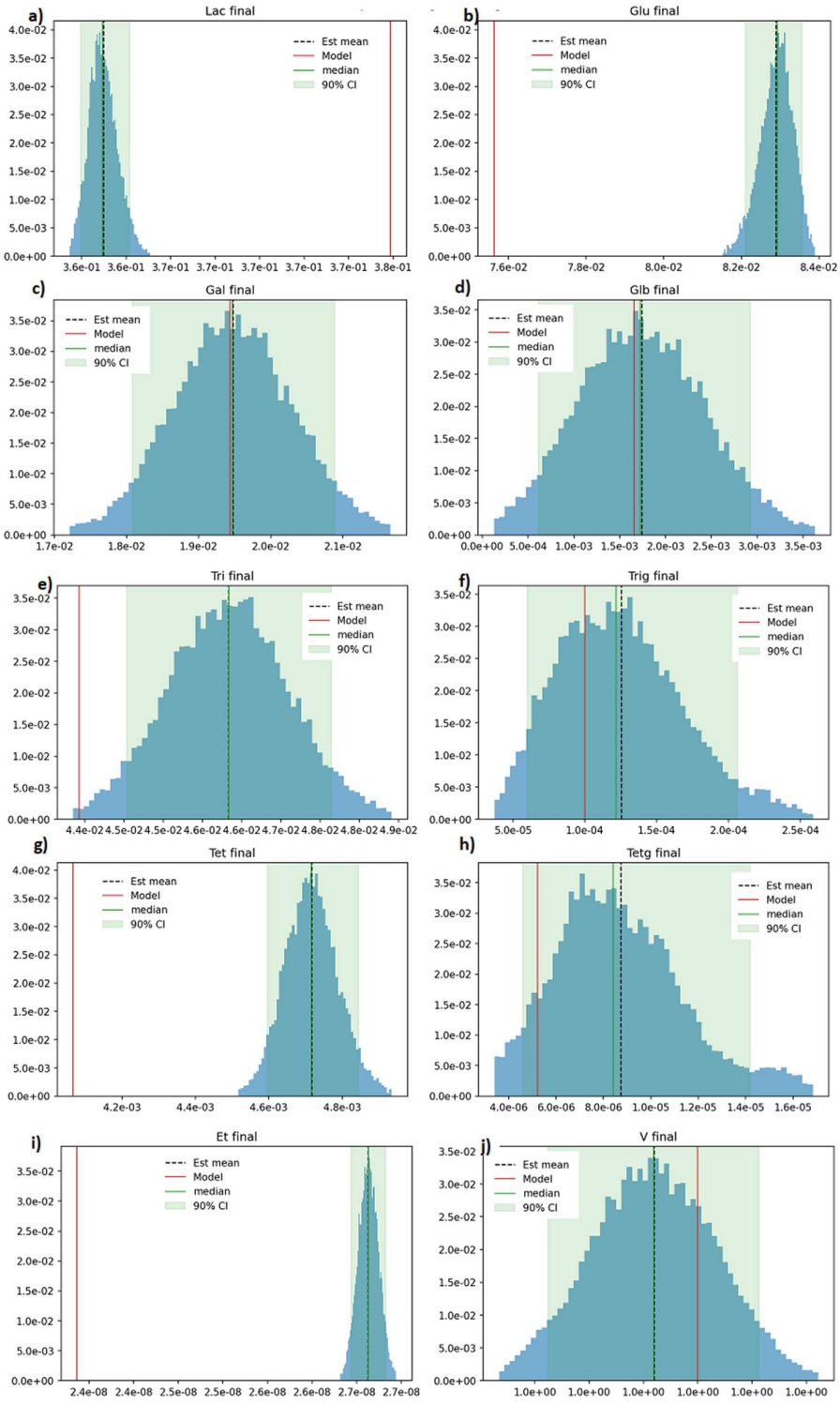
Tabela 13 - NRMSE do EKF, PF e erro relativo na condição de alto erro e incerteza de enzima total.

	NRMSE_EKF	NRMSE_PF	$\frac{\text{NRMSE\_PF}}{\text{NRMSE\_EKF}}$
Lac	0.166	0.086	51.5%
Glu	0.210	0.067	32.0%
Gal	0.386	0.168	43.4%
GOS3	0.110	0.077	69.7%
GOS4	0.613	0.326	53.2%
Média	0.297	0.145	48.7%

Fonte: Elaboração própria.

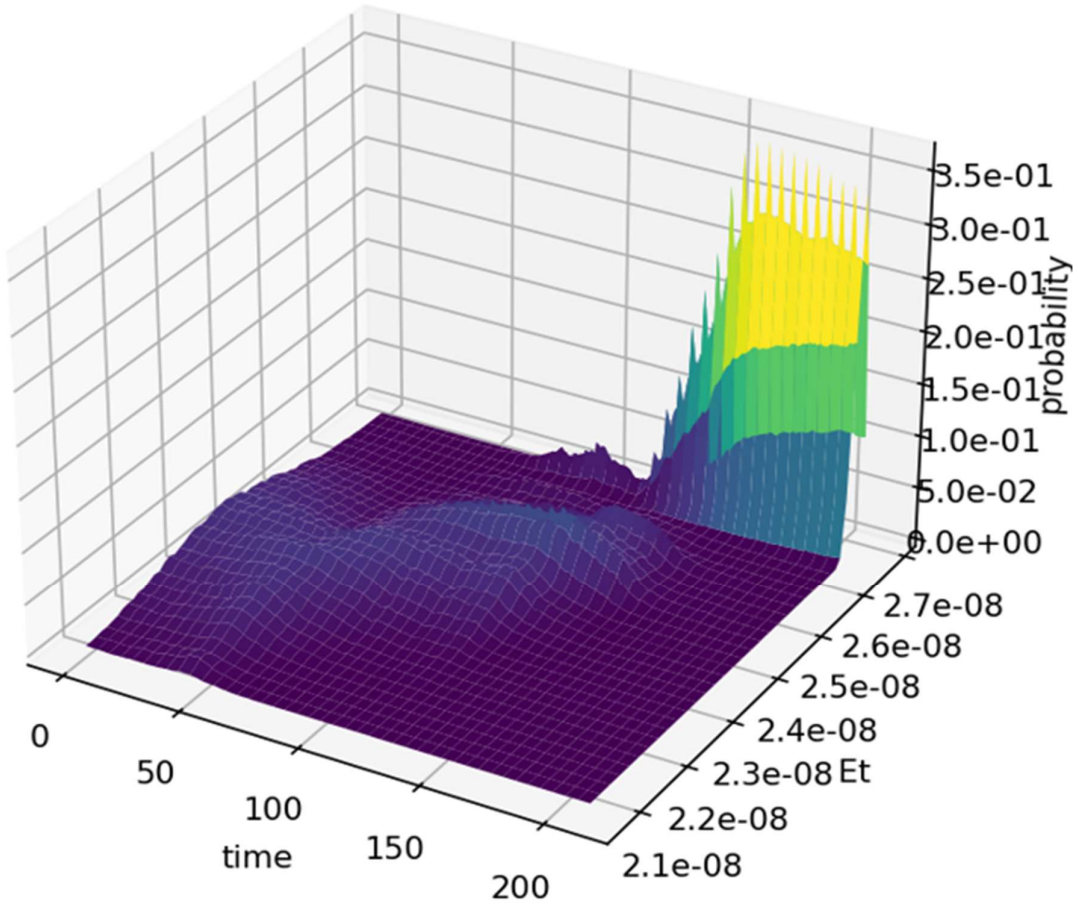
O Particle Filter também pode ser analisado não só em termos de precisão da estiamtiva final, mas também no comportamento de suas partículas. A Figura 25 mostra a distribuição final de partículas para a simulação escolhida (nas condições originais), e com ela é possivel observar que a todas variáveis de estado tiveram distribuições de partículas que se aproximam de uma gaussiana, com média e mediana muito próximas. Adicionalmente, é possivel observar que a estimativa do modelo se distancia de todas as partículas para a lactose, glicose tetrassacarídeos e enzima no final, o que é corroborado pelos gráficos da Figura 23 onde se observa esse distanciamento combinado com o estreitamento das bandas. Apesar da distribuição final ficar próxima de uma gaussiana para as variáveis, a evolução das partículas para a enzima total no tempo (Figura 26) mostra que existem momentos onde a distribuição é multimodal. Esta figura também condiz com o estreitamento das bandas de incerteza da enzima da Figura 23, visto que ocorre uma grande aglomeração de partículas no mesmo ponto ao final. A distribuição de partículas é um dos principais aspectos do PF, pois permite representar distribuições multimodais e capturar mais hipóteses sobre o estado, fornecendo mais flexibilidade na estimativa. Já no EKF, devido à suposição de distribuições gaussianas, no momento em que o filtro converge para uma média, as linearizações e correções são baseadas nessa média.

Figura 25 - Histogramas das partículas para a estimativa final de todas as variáveis de estado. Simulação da batelada 3 utilizando configuração 0.



Fonte: Elaboração própria.

Figura 26 - Evolução das partículas de enzima total com o tempo. Simulação da batelada 3 utilizando configuração 0.



Fonte: Elaboração própria.

5. DISCUSSÃO DOS RESULTADOS

Foi possível perceber que para os estimadores, a variação no desempenho para as melhores configurações não foi tão expressiva, no entanto ao considerar todas as configurações, o máximo aumento no erro foi de 49% e 89% para as simulações no conjunto de treinamento e de validação externa, respectivamente. Isso mostra a necessidade de uma boa escolha dos parâmetros, tanto para o EKF quanto para o PF. Quanto a utilização das matrizes Q e Ω, o fato dos resultados apresentarem melhor desempenho para Q indica que este parâmetro teve um ajuste satisfatório para o conjunto de dados em questão, no entanto o erro adicional no caso de utilizar Ω não foi tão expressivo, indicando que o uso deste parâmetro pode ser adequado quando um ajuste não é viavel, como por exemplo no caso de utilizar um conjunto diferente de dados. Uma outra observação importante é a semelhança e desempenho entre o PF e o EKF nas simulações, o que pode ser explicado pelas estimativas posteriores possuindo distribuição primordialmente gaussiana. O único caso onde se observou grande diferença entre o PF e o EKF foi no teste da atribuição incorreta da enzima, onde o PF demonstrou um poder de correção expressivamente maior do que o EKF. Nesta condição de maior incerteza da concentração inicial da enzima, o PF rapidamente convergiu para o valor verdadeiro, fornecendo estimativas acuradas das demais espécies. Este resultado está de acordo com o observado na literatura e pode ser explicado pelo fato do EKF depender em linearizações locais que podem ser incorretas em condições iniciais diferntes das verdadeiras (LIU; NIRANJAN, 2012; WHITELEY, 2012; CHEN et al., 2005).

6. CONCLUSÕES/CONSIDERAÇÕES FINAIS

De modo geral, ambos estimadores demonstraram bom desempenho frente aos dados experimentais, e para uma aplicação real, a complexidade e custo adicional associados ao PF poderiam ser justificados pela sua maior capacidade de lidar com condições iniciais de maior incerteza. O trabalho possui diversas linhas adicionais que podem ser exploradas. Na parte de calibração multivariada, uma seleção mais refinada dos dados iniciais poderia ser aplicada para remover os dados que possuem problemas de detecção causados pelo HPLC. Ainda nesta parte, outros modelos empíricos poderiam ser testados no lugar do PLS, como redes neurais, que permitiriam captar regiões não-lineares nos dados de treinamento e permitir uma inferência mais precisa. Quanto aos estimadores, mais testes variando suas configurações seriam possíveis, incluindo a utilização das matrizes de covariância Q e Ω simultâneamente, permitindo assim o uso da incerteza dos parâmetros do modelo cinético combinado com um parâmetro ajustado para obter melhores estimativas a priori.

REFERÊNCIAS

ALEXANDER, Ronald et al. State and covariance estimation of a semi-batch reactor for bioprocess applications. **Computers & Chemical Engineering**, v. 172, p. 108180, 2023.

AZCARATE-PERIL, M. A. et al. Impact of short-chain galactooligosaccharides on the gut microbiome of lactose-intolerant individuals. Proceedings of the National Academy of Sciences, [S. l.], v. 114, n. 3, p. E367-E375, 2017.

BOON, M. A.; JANSSEN, A. E. M.; VAN DER PADT, A. Modelling and parameter estimation of the enzymatic synthesis of oligosaccharides by β-galactosidase from Bacillus circulans. **Biotechnology and bioengineering**, v. 64, n. 5, p. 558-567, 1999.

CHEN, C. W.; OU-YANG, C. C.; YEH, C. W. Synthesis of galactooligosaccharides and transgalactosylation modeling in reverse micelles. **Enzyme and Microbial Technology**, v. 33, n. 4, p. 497-507, 2003.

CHEN, T.; MORRIS, J.; MARTIN, E. Particle filters for state and parameter estimation in batch processes. Journal of Process Control, [S.l.], v. 15, n. 6, p. 665-673, 2005.

CHEN, Wen-shiang et al. Bayesian estimation via sequential Monte Carlo sampling: unconstrained nonlinear dynamic systems. **Industrial & engineering chemistry research**, v. 43, n. 14, p. 4012-4025, 2004.

CHEN, Z. **Bayesian filtering: From Kalman filters to particle filters, and beyond**. *Statistics*, [S.l.], v. 182, n. 1, p. 1-69, 2003.

CHOCKCHAISAWASDEE, S. et al. Synthesis of galacto-oligosaccharide from lactose using β-galactosidase from Kluyveromyces lactis: studies on batch and continuous UF membrane-fitted bioreactors. **Biotechnology and Bioengineering**, [S. l.], v. 89, n. 4, p. 434-443, 2005.

COULIER, L. et al. In-depth characterization of prebiotic galacto-oligosaccharides by a combination of analytical techniques. **Journal of Agricultural and Food Chemistry**, [S. l.], v. 57, n. 18, p. 8488-8495, 2009.

DIAS, L. G.; VELOSO, A. C.; CORREIA, D. M.; ROCHA, O.; TORRES, D.; ROCHA, I.; PERES, A. M. UV spectrophotometry method for the monitoring of galacto-oligosaccharides production. ***Food Chemistry***, v. 113, n. 1, p. 246–252, 2009.

FRENZEL, Monika et al. Comparison of the galacto-oligosaccharide forming activity of different β-galactosidases. **LWT-Food Science and Technology**, v. 60, n. 2, p. 1068-1071, 2015.

GOSLING, A. et al. Effect of the substrate concentration and water activity on the yield and rate of the transfer reaction of β-galactosidase from Bacillus circulans. **Journal of Agricultural and Food Chemistry**, [S. l.], v. 59, n. 7, p. 3366-3372, 2011.

GOSLING, A. et al. Recent advances refining galactooligosaccharide production from lactose. **Food Chemistry**, [S. l.], v. 121, n. 2, p. 307-318, 2010.

HERMANN, L.; KREMLING, A. A hybrid soft sensor approach combining partial least-squares regression and an unscented Kalman filter for state estimation in bioprocesses. **Bioengineering**, Basel, v. 12, n. 6, p. 654, jun. 2025.

HERNÁNDEZ-HERNÁNDEZ, O. et al. Hydrophilic interaction liquid chromatography coupled to mass spectrometry for the characterization of prebiotic galactooligosaccharides. **Journal of Chromatography A**, [S. l.], v. 1220, p. 57-67, 2012.

JOHANSEN, A. **A tutorial on particle filtering and smoothing: Fifteen years later**. 2009.

KRÄMER, D.; KING, R. A hybrid approach for bioprocess state estimation using NIR spectroscopy and a sigma-point Kalman filter. Journal of Process Control, [S.l.], v. 82, p. 91-104, out. 2019.

LIU, X.; NIRANJAN, M. State and parameter estimation of the heat shock response system using Kalman and particle filters. **Bioinformatics**, Oxford, v. 28, n. 11, p. 1501-1507, jun. 2012.

LOPEZ, P. et al. Transforming data to information: a parallel hybrid model for real-time state estimation in lignocellulosic ethanol fermentation. **Biotechnology and Bioengineering**, Hoboken, v. 118, n. 2, p. 579-591, fev. 2021.

LUMUMBA, V. W.; KIPROTICH, D.; MPAINE, M. L.; MAKENA, N. G.; KAVITA, M. D. Comparative analysis of cross-validation techniques: LOOCV, K-folds cross-validation, and repeated K-folds cross-validation in machine learning models. **American Journal of Theoretical and Applied Statistics**, 1 jun. 2024.

MAIONE, Nicole Ribeiro. **MONITORAMENTO DA SÍNTESE ENZIMÁTICA DE GALACTOOLIGOSSACARÍDEOS POR UV-VIS**. 2024. 68 f. Tese (Doutorado em Engenharia Química) - Departamento de Engenharia Química, Universidade Federal de São Carlos, São Carlos, 2024.

MARTINS, A. R.; LISBÔA, C. R. Mathematical model for the conversion of lactose and synthesis of galacto-oligosacharides (GOS) with simultaneous reversible inhibition by

glucose and galactose. **International Journal of Engineering Research and Technology**, Gujarat, v. 4, n. 4, p. 299-305, 2015.

MEEUSEN, E. et al. Gram-scale chemical synthesis of galactosyllactoses and their impact on infant gut microbiota in vitro. **Organic & Biomolecular Chemistry**, [S. l.], v. 22, n. 10, p. 2091-2097, 2024.

NELLES, Oliver. **Nonlinear system identification**. 2 ed. Berlin: Springer, 2001. 1233 p.

OSMAN, A. Synthesis of prebiotic galacto-oligosaccharides: science and technology. In: SHAHIDI, F. (ed.). **Probiotics, prebiotics, and synbiotics**: bioactive foods in health promotion. [S. l.]: Springer, 2016. p. 135-154.

PALAI, T.; BHATTACHARYA, P. K. Kinetics of lactose conversion to galacto-oligosaccharides by β-galactosidase immobilized on PVDF membrane. **Journal of Bioscience and Bioengineering**, v. 115, n. 6, p. 668-673, 2013.

RAWLINGS, James Blake et al. **Model predictive control: theory, computation, and design**. Madison, WI: Nob Hill Publishing, 2020.

RODRIGUEZ-COLINAS, B. et al. Continuous packed bed reactor with immobilized β-galactosidase for production of galactooligosaccharides (GOS). **Catalysts**, [S. l.], v. 6, n. 12, p. 189, 2016.

RODRIGUEZ-FERNANDEZ, M. et al. Detailed kinetic model describing new oligosaccharides synthesis using different B-galactosidases. **Journal of Biotechnology**, v. 153, n. 3-4, p. 116-124, 2011.

SCHIEMER, R. et al. An adaptive soft-sensor for advanced real-time monitoring of an antibody-drug conjugation reaction. **Biotechnology and Bioengineering**, Hoboken, v. 120, n. 7, p. 1914-1928, jul. 2023.

SCHULTZ, Guilhermina et al. Kinetic modeling of the enzymatic synthesis of galacto-oligosaccharides: describing galactobiose formation. **Food and Bioproducts Processing**, v. 127, p. 1-13, 2021.

SIERRA, C. et al. Prebiotic effect during the first year of life in healthy infants fed formula containing GOS as the only prebiotic: a multicentre, randomised, double-blind and placebo-controlled trial. **European Journal of Nutrition**, [S. l.], v. 54, n. 1, p. 89-99, 2015.

SIMUTIS, R. et al. State estimation of a biotechnological process using extended Kalman filter and Particle filter. In: **VETERINARY AND AGRICULTURAL ENGINEERING**, 2014, [S.l.]. p. 920-924.

STELZER, Ines V.; KAGER, Julian; HERWIG, Christoph. Comparison of particle filter and extended kalman filter algorithms for monitoring of bioprocesses. In: **Computer Aided Chemical Engineering**. Elsevier, 2017. p. 1483-1488.

TORRES, D. P. et al. Galacto-oligosaccharides: production, properties, applications, and significance as prebiotics. **Comprehensive Reviews in Food Science and Food Safety**, [S. l.], v. 9, n. 4, p. 308-320, 2010.

VAN LEUSEN, E. et al. Industrial applications of galactooligosaccharides. In: VAN LEUSEN, E. et al. **Food oligosaccharides**: Production, analysis and bioactivity. [S. l.]: Wiley, 2014. Cap. 20, p. 470-491.

VULEVIC, J. et al. Influence of galacto-oligosaccharide mixture (B-GOS) on gut microbiota, immune parameters and metabonomics in elderly persons. **British Journal of Nutrition**, [S. l.], v. 114, n. 4, p. 586-595, 2015.

WHISNER, C. M. et al. Galacto-oligosaccharides increase calcium absorption and gut bifidobacteria in young girls: a double-blind cross-over trial. **British Journal of Nutrition**, [S. l.], v. 110, n. 7, p. 1292-1303, 2013.

WHITELEY, N. Sequential Monte Carlo samplers: error bounds and insensitivity to initial conditions. **Stochastic Analysis and Applications**, [S.l.], v. 30, n. 5, p. 774-798, 2012.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. ***Chemometrics and Intelligent Laboratory Systems***, v. 58, n. 2, p. 109–130, 2001.

ZENG, M. et al. Novel galacto-oligosaccharides from lactose: Chemical synthesis, structural characterization, and in vitro assessment of prebiotic activity. **ACS Sustainable Chemistry & Engineering**, [S. l.], v. 11, n. 38, p. 14031-14045, 2023.

ZHAO, J. et al. Structure, enzymatic production, biological activities, and food applications of galacto-oligosaccharides: A review. **The Journal of Nutrition**, [S. l.], 2025.

APÊNDICE A – DADOS ADICIONAIS

Figura 27 - Espectros brutos para batelada 1.

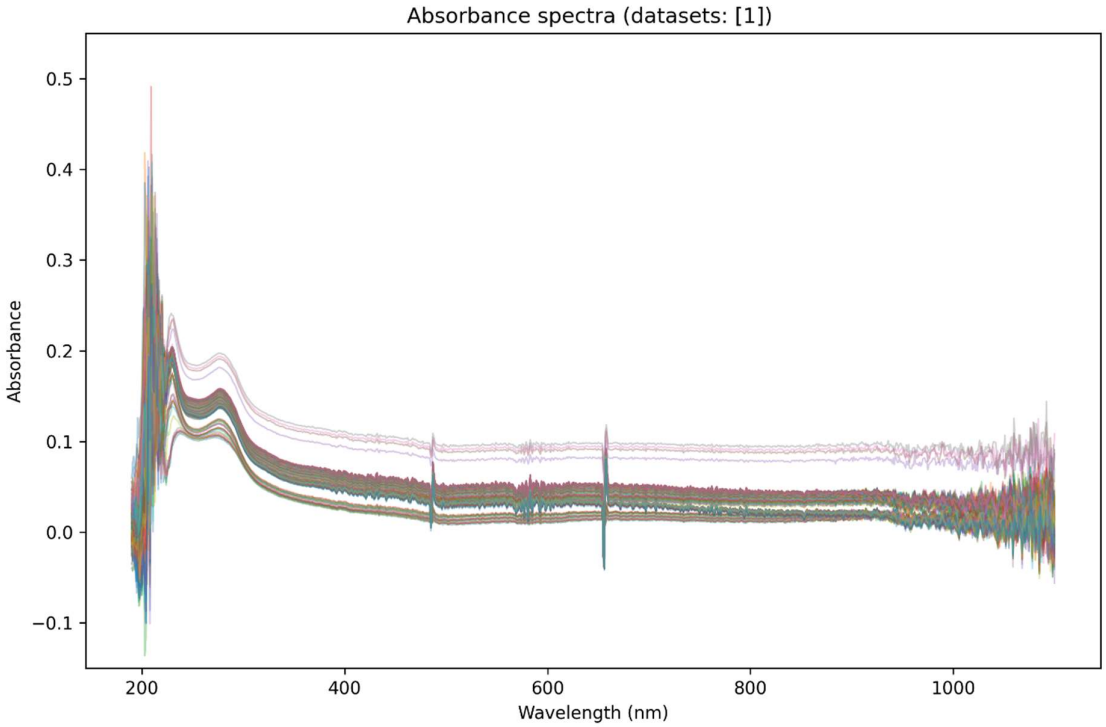


Figura 28 - Espectros brutos para batelada 2.

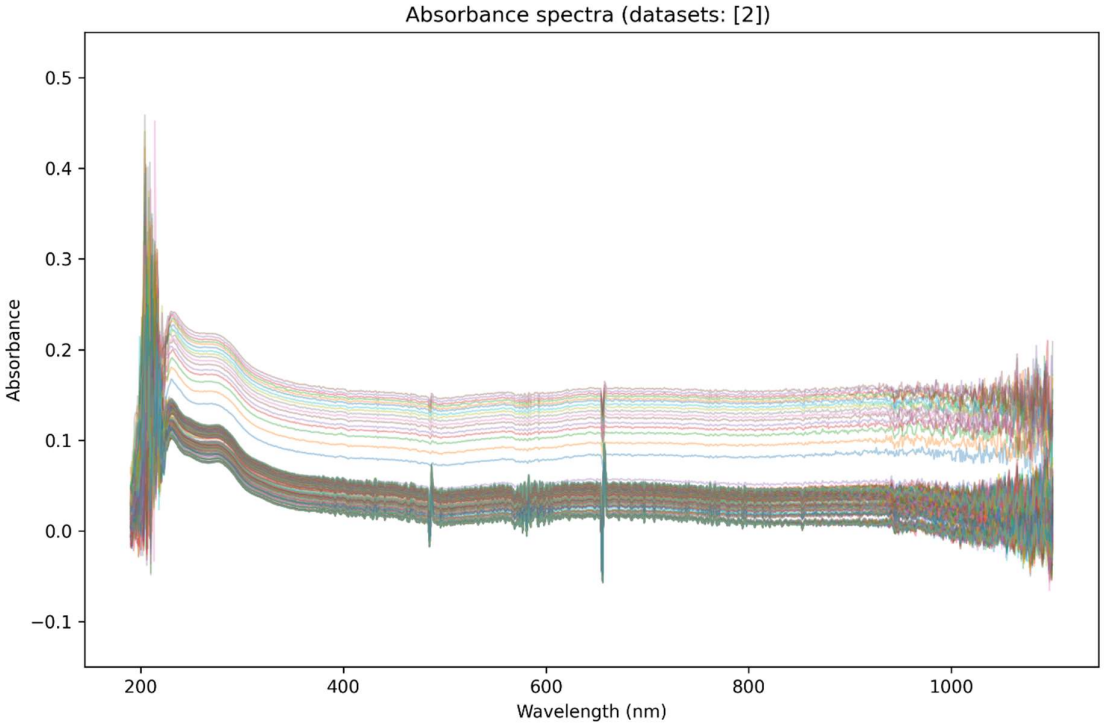


Figura 29 - Espectros brutos para batelada 3.

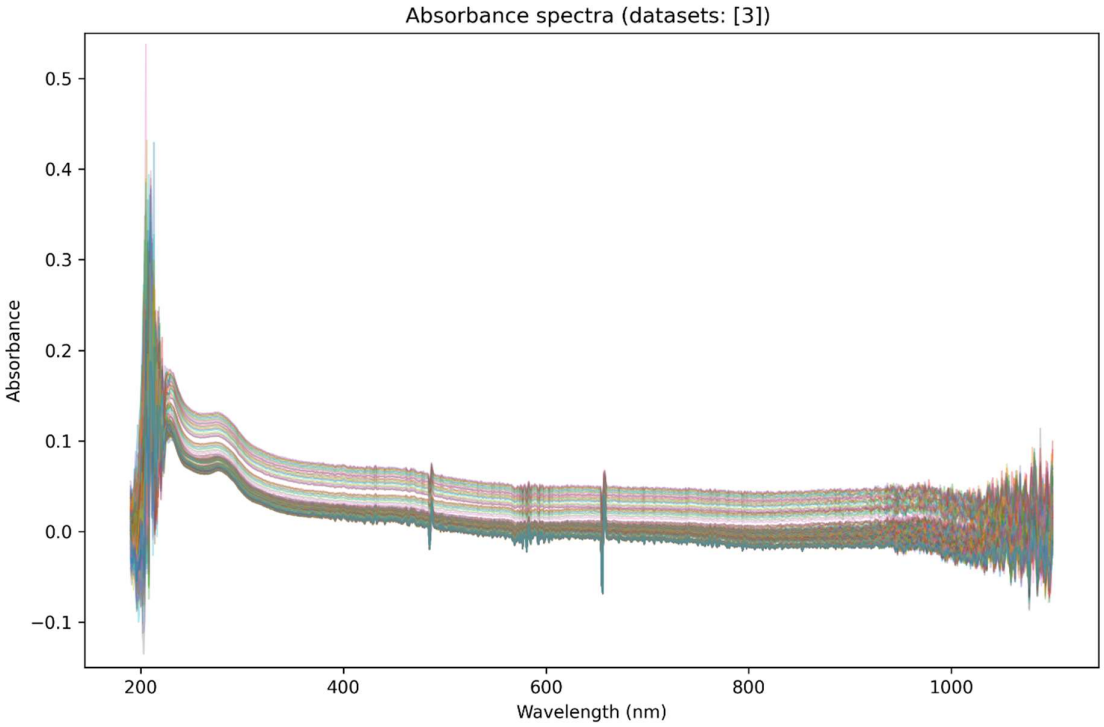


Figura 30 - Espectros brutos para batelada 4.

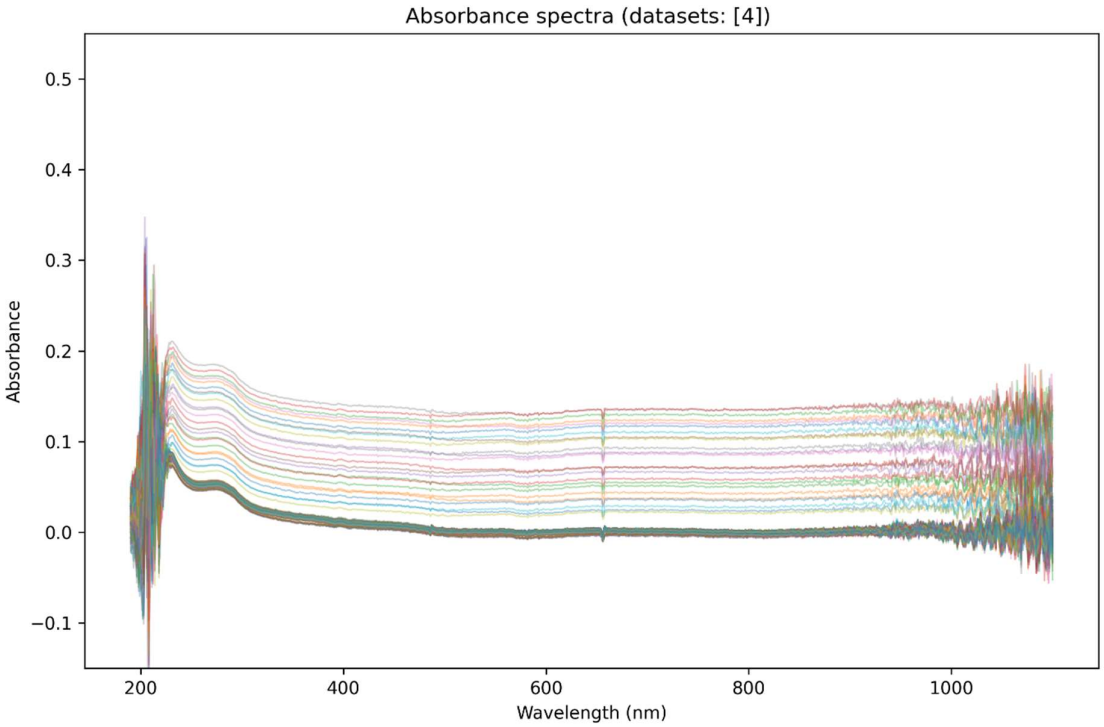


Figura 31 - Espectros brutos para batelada 6.

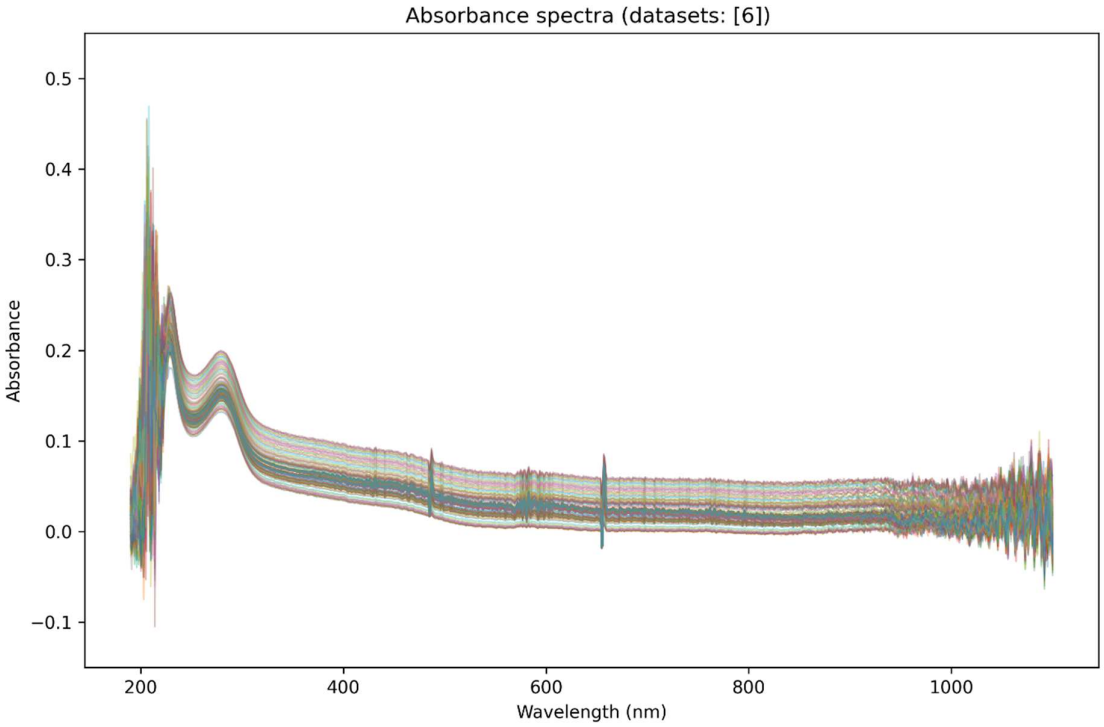


Figura 32 - Espectros brutos para batelada 8.