Unsupervised learning bioreactor regimes

Laborda, Víctor Puig I.; Puiman, Lars; Groves, Teddy; Haringa, Cees; Nielsen, Lars Keld

DOI
10.1016/j.compchemeng.2024.108891

Publication date
2025

Document Version
Final published version

Published in
Computers and Chemical Engineering

Computers and Chemical Engineering 194 (2025) 108891

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Unsupervised learning bioreactor regimes

Víctor Puig I Laborda [a,*], Lars Puiman [b], Teddy Groves [a], Cees Haringa [b], Lars Keld Nielsen [a,c,*]

[a] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
[b] Department of Biotechnology, Bioprocess Engineering group, Faculty of Applied Sciences, Delft University of Technology, van der Maasweg 9, 2629HZ, Delft, The Netherlands
[c] Australian Institute for Bioengineering and Nanotechnology (AIBN), The University of Queensland, St. Lucia, QLD 4067, Australia

ARTICLE INFO

Keywords:
Bioreactor Regimes
Compartmentalization
Unsupervised Machine Learning
Clustering Techniques
Computational Fluid Dynamics (CFD)
Mathematical Modeling

ABSTRACT

Efficient operation of bioreactors is crucial for the success of biomanufacturing processes. Traditional Computational Fluid Dynamics (CFD) simulations provide detailed insights but often involve lengthy computation times and complexity, hindering their practicality for real-time applications. This study introduces a novel multivariate unsupervised learning algorithm that clusters bioreactors into physically meaningful regions based on CFD-generated and real-world data. These clusters not only facilitate the determination of internal reactor regimes but also serve as a foundational step for developing compartment models. Our approach utilizes a custom k-means clustering algorithm, which ensures spatial continuity of clusters by incorporating geometric data, and optimizes the number of compartments to maximize physical significance and data retention. This optimization is guided by a Pareto front analysis, balancing the need for clear compartment definition with the preservation of maximum information from the dataset. The effectiveness and versatility of this methodology were verified through case studies involving a 202 m^3 Rushton impeller bioreactor (steady state simulation) and an 840 m^3 airlift reactor (dynamic simulation). In the airlift reactor, the clustering algorithm accounted for dynamic fluctuations by averaging the simulation results, providing a robust method for incorporating temporal variations into the compartment analysis. The findings highlight the advantages of 3-D compartmentalization in capturing the intricate dynamics of fluid motion and cellular activities, thereby advancing the design of bioreactors and scaling down experiments for more efficient industrial applications.

1. Introduction

The significance of industrial biotechnology in producing diverse compounds such as pharmaceuticals, enzymes, food products and commodity products has grown considerably in recent years. The performance of large-scale bioreactors is crucial to biotechnological processes, as the spatio-temporal variation in the conditions experienced by microorganisms directly affects process yield, productivity, and product quality. As a result, considerable effort has been devoted to understanding and predicting conditions within bioreactors, particularly focusing on the interaction between biological reactions and heterogeneous hydrodynamics arising from scale-related issues.

Computational Fluid Dynamics (CFD) simulations have been utilized to examine the inherent heterogeneous behavior of bioreactors under various conditions (Sharma et al., 2011). However, due to the complexity of the models and the high-density mesh necessary for high resolution results, computation times can be excessively long. Consequently, CFD is not an ideal solution when fast simulations are required for process control, optimization, or when simulations contain complex models, such as the combination of CFD with metabolic models.

Compartment Models (CMs) offer a suitable compromise between CFD simulations and the homogeneity assumption in reactors, reducing the computation time while sacrificing some accuracy in the results (Tajsoleiman et al., 2019). CMs represent non-ideal mixed systems as a network of well-mixed compartments connected by the transfer of momentum, heat and mass (Mann & Mavros, 1982). This allows for the use of continuity equations and transfer between compartments to simulate heterogeneous bioreactors, without the need for Navier-Stokes equations (Bezzo et al., 2004).

Creating a CM involves two important steps that affect the model's accuracy. First, compartments must be defined, with the size and position of each compartment chosen optimally. If a compartment is too small, then more compartments will be required, and the computation time will suffer. If a compartment is too big, or is incorrectly positioned,

* Corresponding authors.
E-mail addresses: viclab@dtu.dk (V.P.I. Laborda), lars.nielsen@uq.edu.au (L.K. Nielsen).

https://doi.org/10.1016/j.compchemeng.2024.108891
Received 19 July 2024; Received in revised form 7 October 2024; Accepted 9 October 2024
Available online 11 October 2024

Nomenclature

Variables

C_S	Substrate concentration $\left[\frac{kgS}{m^3}\right]$
$\mathscr{D}_S$	Diffusion coefficient of substrate in carrier $\left[\frac{m^2}{s}\right]$
ν_t	Turbulent viscosity $\left[\frac{m^2}{s}\right]$
Sc_t	Turbulence Schmidt number
S_S	Substrate uptake rate $\left[\frac{kgS}{m^3 s}\right]$
$q_{S,\max}$	Maximum substrate specific uptake rate $\left[\frac{kgS}{kgX \cdot s}\right]$
K_S	Half-saturation constant $\left[\frac{kgS}{m^3}\right]$
C_X	Biomass concentration $\left[\frac{kgX}{m^3}\right]$
P	Pressure [Pa]
F_S	Volumetric substrate feeding rate $\left[\frac{kgS}{m^3 s}\right]$
m	Total number of data points
K_H	Scaled kernel function
x	Evaluation point for density estimation
x_i	Individual data points from the dataset
$\overrightarrow{x_i}$	Individual data point, vector representation with the features and cell coordinates
$\overrightarrow{\mu_l}$	Centroid of cluster l with all features (vector)
$\overrightarrow{x_i}$	Individual data points in the clustered dataset with all the features (vector)
k	Total number of clusters
n	Total number of features
$\overrightarrow{w}$	Weight vector for the fetures used to cluster
$\overrightarrow{C_l}$	Datapoints assigned to cluster l
V	Matrix (example)
D	Distance matrix
D^{norm}	Normalized distance matrix
S_l	Average distance between all data points within a cluster l and its centroid
$d_{l,o}$	Inter-cluster distance between to distinct clusters l and o
$R_{l,o}$	Cluster similarity ratio between clusters l and o
DBI	Davies-Bouldin Index
SS	Average silhouette score
CV	Coefficient of variation
H	Average homogenenity score

Operators

$\oslash$	Element-wise division

Subscripts (Clustering)

f	Clustering features (exclude euclidean coordinates)
i	Datapoint
j	Datapoint different than i
t	Turbulence
e	Euclidean coordiantes
l	Cluster
o	Cluster, different than l

Functions

$J(C, \overrightarrow{\mu_l})$	Loss function evaluated for a set of centroids with different features
$\widehat{f_H}(x)$	Estimated density function evaluated at point x
$C(\overrightarrow{x_i})$	Clustering function per each datapoint
$d(p_1, p_2)$	Custom distance between points p_1 and p_2
$D(\overrightarrow{x_i})$	Distance from each data point to the nearest initialized cluster
$P(\overrightarrow{x_i})$	Probability of the centroid to be initialized at a certain datapoint
$C_{(l)}(\overrightarrow{x_i})$	Assignation of datapoint $\overrightarrow{x_i}$ to cluster l
$S(V)$	Normalization of matrix V
$\sigma(V)$	Variance of matrix V
$\overline{V}$	Mean of matrix V
$a(\overrightarrow{x_i})$	Average distance between point in cluster and all other points in the same cluster
$b(\overrightarrow{x_i})$	Minimum average distance betweena point and all points in any other cluster
$s(\overrightarrow{x_i})$	Silhouette function for single point

the assumption of within-compartment homogeneity will be invalid, leading to inaccurate results. Second, exchange flows between compartments must be identified; this step requires the calculation of momentum, mass, and energy exchange flows among compartments in the reactor.

CM models can be classified into three generations, corresponding to advances in the methods used to determine the compartments and flows and the compartmentalization objective variable. These are shown schematically in Fig. 1.

The first-generation models (Mann & Mavros, 1982), defined compartments (then also known as Network of Zones or NoZ) manually by looking at the hydrodynamic profile (Knysh & Mann, 1984), based solely on user expertise. The flows between compartments were calculated using global variables, which resulted in an underestimation of the complexity of turbulent flows and bioreactor gradients. Hydrodynamics was the primary variable used to define compartments. Enfors et al. (2001) used compartment models to study the physiological response of microorganisms in large-scale bioreactors, while Zahradník et al. (2001) employed NoZ analysis to investigate the mixing and mass transfer in three different industrial cases.

The second generation commenced with the work of Bezzo et al. (2004), who combined NoZ with CFD to compute flowrates between compartments (Bezzo et al., 2004; Rigopoulos & Jones, 2003). Bezzo & Macchietto (2004) developed a method for automatic compartment definition using aggregation techniques, albeit only for a structured mesh consisting of cells systematically arranged in a regular grid pattern. The primary variable for compartmentalization during this generation remained hydrodynamics-related, focusing primarily on the velocity field range. Wells & Ray (2005) automated the compartmentalization by assigning a tolerance range to variables in the entire volume; however, this resulted in non-spatially-continuous compartment definitions. Le Moullec et al. (2010) used CFD and CMs in combination to compute flowrates between compartments, albeit with a manual selection to determine NoZ.

The third generation began with the work of Delafosse et al. (2014) and has been further developed by Tajsoleiman et al. (2019) and Le Nepvou De Carfort et al. (2024). This generation is characterized by generalized automatic compartmentalization that applies to all types of meshes, albeit sometimes necessitating mesh transformation (Le Nepvou De Carfort et al., 2024; Tajsoleiman et al., 2019). This approach allowed the inclusion of variables beyond hydrodynamics for defining the compartments and the integration of CFD/CMs with dynamic models, population balances, stochastic particle tracking, and chemical and metabolic models (Delafosse et al., 2015; Haringa et al., 2022;

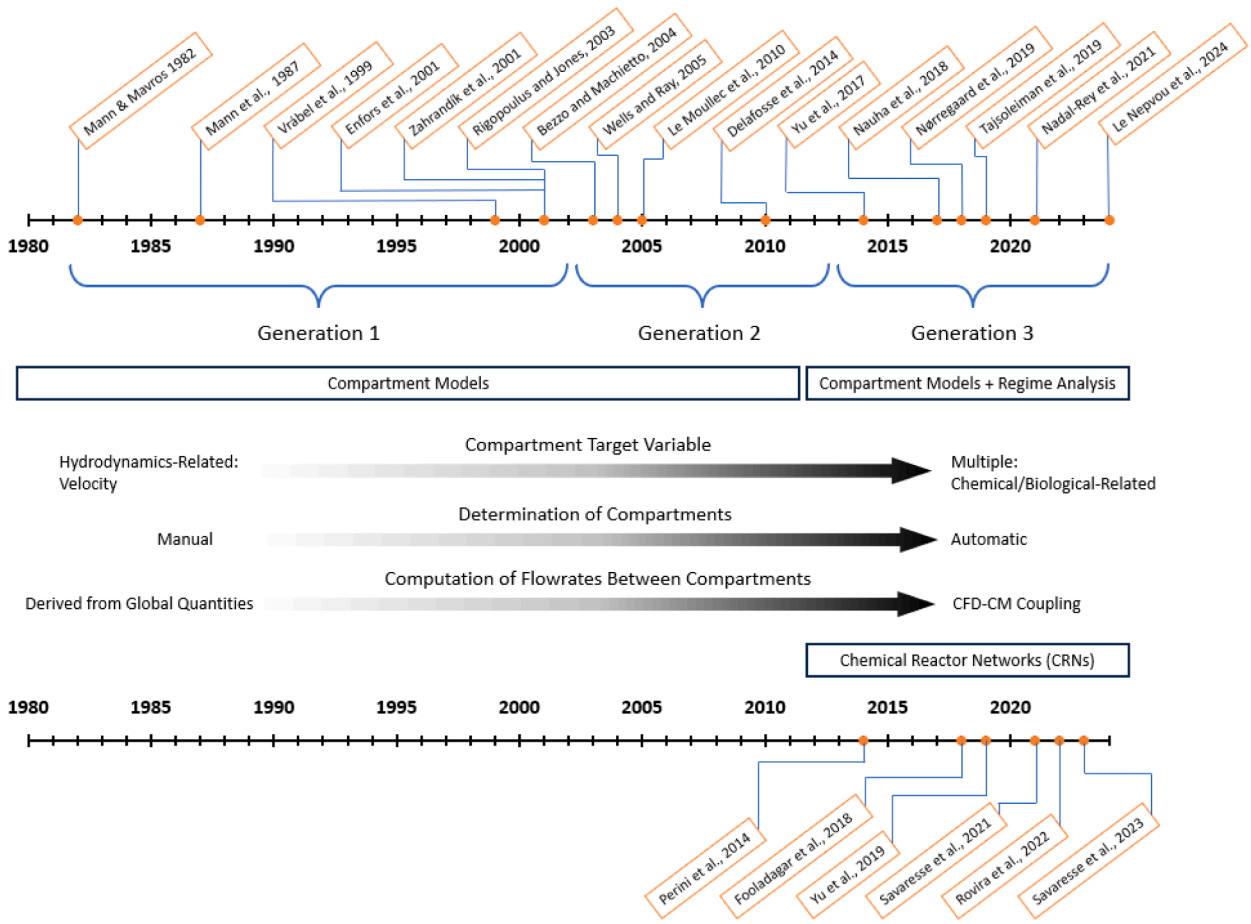

Fig. 1. Timeline illustrating the evolution of compartment models – regime analysis and chemical reactor networks based on the most cited literature, highlighting three design fields described above.

Nadal-Rey, McClure, Kavanagh, Cassells, et al., 2021; Nauha et al., 2018; Nørregaard et al., 2019; Pigou & Morchain, 2015). During this period, regime analysis emerged as a new approach derived from compartmentalization, focusing on identifying regimes based on various factors such as metabolic states, pH, and substrate concentration (Haringa et al., 2016; Nadal-Rey, McClure, Kavanagh, Cornelissen, et al., 2021; Nørregaard et al., 2019).

Despite these advancements, the use of manual compartment definitions persisted (Nauha et al., 2018; Nørregaard et al., 2019). While Tajsoleiman et al. (2019) and Delafosse et al. (2015) provided a framework for reactor compartmentalization using any design variable and automated computation of flow rates between compartments, their methods were confined to 2-D applications, assuming radial symmetry. Moreover, the required user expertise for determining the number of compartments based on design variables highlighted a gap in evaluating the optimal number of compartments. Recently, Le Nepvou De Carfort et al. (2024) proposed a compartmentalization method for 3-D models based on hydrodynamics, though it required mesh transformation to a structured format and user selection of the number of compartments.

In parallel developments within the field of combustion chemical engineering, new methodologies for dividing reaction domains into regimes based on unsupervised learning have emerged. Notably, the Chemical Reactors Network (CRN) introduced by Perini et al. (2014) employed clustering techniques in CFD simulations of fuel combustion to reduce computational demands. This approach was further advanced by Yu et al. (2019), who utilized clustering to define soot regimes in CFD results from combustion furnaces, and by Fooladgar & Duwig (2018) and Rovira et al. (2022), who aimed to reduce dataset dimensionality before clustering flame simulations into similar zones based on combustion characteristics. The problem of ensuring continuous spatial integrity of the clusters was addressed by Savarese et al. (2023), who combined clustering with graph analysis, although some manual curation remained necessary. Continuous spatial integrity of the clusters was less problematic for CRNs, as the clustering was applied to 2-D simulations characterized by distinct and discrete conditions. Additionally, the simulation space featured continuous geometries without abrupt or unusually shaped gradients, ensuring straightforward application of clustering techniques.

In this paper, we propose a new generalized regime analysis method based on unsupervised learning, specifically custom clustering algorithms. This approach merges knowledge from CRNs, modified to ensure spatial continuity in 3-D non-structured meshes without manual intervention, with regime analysis for bioreactors. This method enables clustering based on any target variable (or multiple ones) without causing an exponential increase in the number of compartments (Tajsoleiman et al., 2019). As this approach can cluster data using any target variable(s), its primary use case is like that of regime analysis in identifying zones with similar conditions, with the advantage of using the number of compartments/regimes as the sole design variable. These regimes can serve to characterize reactor performance, or as a basis to design scale-down experiments (Haringa et al., 2017). With the appropriate choice of clustering parameters, the approach can equally serve as an initial step for CMs, as this has in essence the same objective, dividing the domain into a limited number of spatial zones to capture certain key characteristics, in this case flow-related.

This work also proposes a general evaluation method for determining an optimal number of compartments or regimes that accurately describe the bioreactor's properties without the drawbacks of excessive

clustering. Our algorithm automatically identifies patterns within CFD datasets, eliminating the need for manual identification or tolerance-based algorithms.

2. Methods

This paper presents an application of unsupervised learning for discerning patterns in unlabeled datasets originating from CFD simulations or actual data cases. The procedure for identifying reactor compartments includes iterative steps that derive the optimal number of clusters to best represent the provided data. These steps can be automated through an optimum regime analysis where a Pareto front for the optimum number of compartments is obtained (Fig. 2).

Data can be read from a variety of standard formats, including comma-separated value (CSV) files. These files are loaded into dataframe structures using the programming language Python. A major goal of this research project was to ensure the algorithm's versatility across diverse datasets, regardless of the data-generating source, which could be, for example CFD software, soft-sensor, or static probes.

In this section, the main ideas behind the algorithm are described. Please refer to Section 3

Theory/CalculationTheory/Calculation, for a more complete mathematical formulation of all steps of the algorithm.

2.1. CFD Simulations, Data Generation

Although not the focus of this paper, a simple CFD simulation of a bioreactor was performed to generate an initial test case dataset for the clustering algorithm. The simulation models a representative geometry for a standard industrial stirred bioreactor (202 m^3) at the end of an industrial fed-batch fermentation. Fig. 3 represents a sketch of the bioreactor; the sizes of the different parameters are found in the geometry set-up table (Table 1).

Assuming radial symmetry, and pseudo-steady state (Multiple Frame

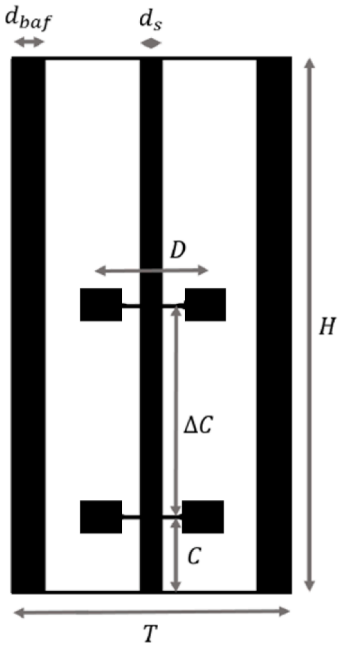

Fig. 3. Sketch of the Rushton-Rushton impeller bioreactor.

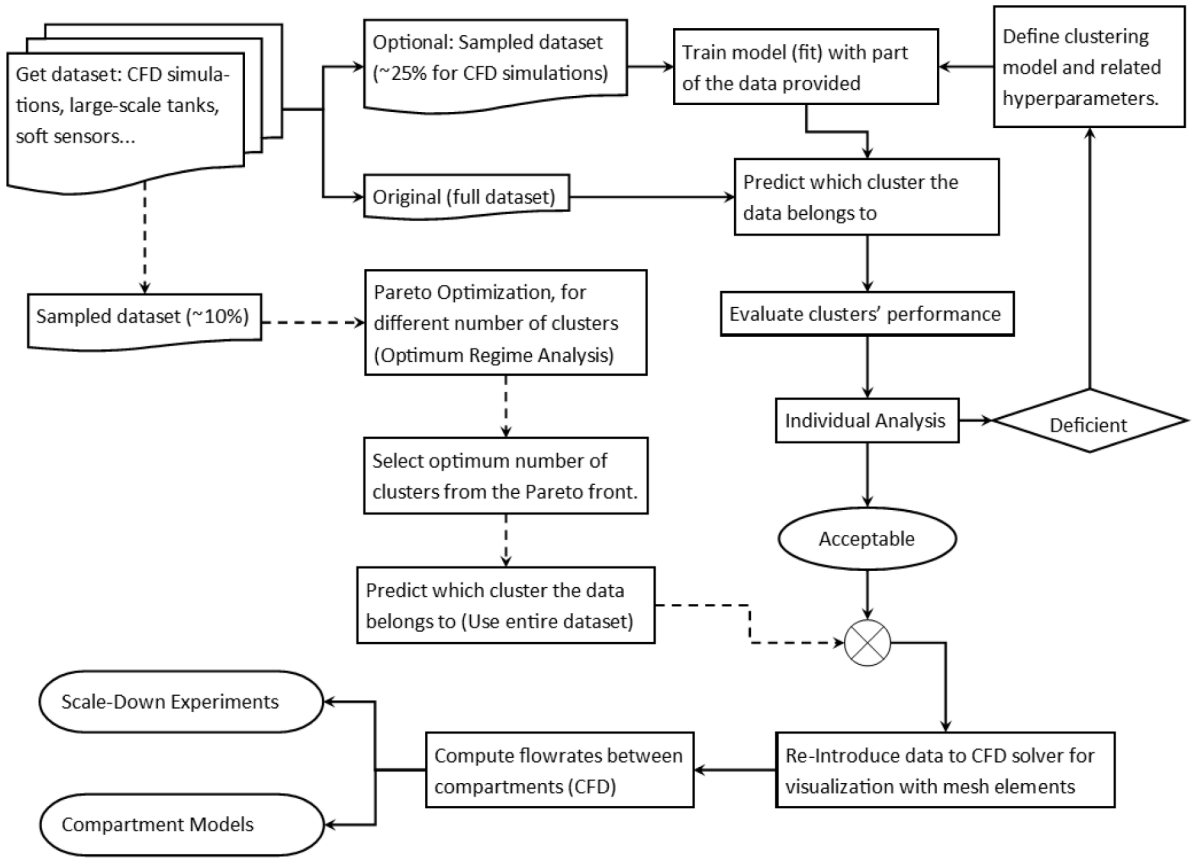

Fig. 2. Automatic compartmentalization process flowchart. Actions are defined in rectangles, input datasets in bottom-curved rectangles, outputs in sides-curved rectangles, merged actions in a circular cross and decisions in ovals/rhombus (acceptable/deficient). The solid lines represent an individual analysis path, whilst the dashed lines represent the optimum regime analysis path (Pareto Optimization for different number of clusters).

Table 1
Summary of geometric parameters for the Rushton-tank bioreactor.

Parameter	Value
Liquid height (H)	11.75 [m]
Tank diameter (T)	4.7 [m]
Lower impeller clearance (C)	$T/3$
Rushton impeller diameter (D)	$0.4{\cdot}T$
Distance between impellers (ΔC)	T
Baffle width (d_{baffle})	$0.1{\cdot}T$
Shaft diameter (d_S)	0.25 [m]

Reference) only 1/6 of the reactor (33 m^3) was simulated, including a unique baffle. The symmetry and pseudo-steady state assumptions mean that certain hydrodynamic phenomena such as macro-instabilities are not modelled even though they are relevant for accurate assessment of mixing in multi-impeller systems (Haringa et al., 2018). Still, these simplifications are acceptable for the purpose of demonstrating the regime analysis and testing the clustering algorithm.

A well-validated RANS approach was used for the single-phase CFD simulation. This combines k-ε for turbulence with MRF (Multiple-Reference Frame) for the impellers (Coroneo et al., 2011; Gunyol & Mudde, 2009). The substrate transfer is ruled by the general micro-balance for mass transport, as shown in Equation 1.

$$\frac{\partial C_s}{\partial t} + \nabla\cdot(uC_S) = -\nabla\cdot\left[-\left(\left(\mathscr{D} + \frac{\nu_t}{Sc_t}\right)\nabla C_S\right)\right] + S_S \tag{1}$$

The fluid flow and species transport were resolved using a coupled scheme for pressure-velocity coupling, alongside second-order upwind spatial discretization for all variables. A global timestep pseudo-time method was employed to ensure stability and enhance accuracy. While this approach yields superior results with higher accuracy and stability, it demands significantly higher computational resources.

The source term (S_S) was computed using simple Monod kinetics with the parameter values, maximum substrate uptake rate $\left(q_{S,\max}\right)$ and half-saturation constant (K_S), taken from Lin et al. (2001), to simulate *E. Coli* bio-kinetics. The feed used as substrate in the bioreactor was glucose. A summary of all the parameters used in the simulation can be found in Table 2.

$$S_S = q_{S,\max}\cdot\left(\frac{C_S}{C_S + K_S}\right)\cdot C_X \; + \; F_S \tag{2}$$

A summary of all the parameters used in the CFD simulation can be found in Table 2.

2.2. Bi-Variate KDE Analysis

For the exploratory analysis, multivariate Kernel Density Estimation (KDE) of the response variable was performed,

$$\widehat{f_H}(x) = \frac{1}{m}\sum_{i=1}^{m} K_H(x - x_i) \tag{3}$$

Table 2
Parameters used in the CFD simulation for the Rushton-tank bioreactor.

Parameter	Value
Maximum Substrate Uptake Rate $\left(q_{S,\max}\right)$	$1.5\left[\frac{kg_S}{kg_X\cdot s}\right]$
Half-saturation constant (K_S)	$0.02\left[\frac{kg}{m^3}\right]$
Biomass concentration (C_X)	$82\left[\frac{kg}{m^3}\right]$
Volumetric feeding rate (F_S)	$0.0025\left[\frac{kg}{m^3\cdot s}\right]$
Stirring speed	150 [rpm]
Number of mesh elements	$2{\cdot}10^5$

Scott's method was used to determine the smoothing passed to the Gaussian KDE (Scott, 2015) with the multiplicative scale factor (K_H) set to 1.

2.3. Custom k-Means Algorithm

A modified k-means clustering algorithm was used for the compartmentalization steps. K-means is a popular unsupervised machine learning algorithm that partitions a dataset into k distinct clusters based on similarity measures. It seeks to minimize the within-cluster similarity distance, also known as inertia, thereby ensuring that points within each cluster are as similar as possible, while simultaneously maximizing the separation between distinct clusters (Lloyd, 1982). The customization involves the method used to determine similarity: our algorithm uses a new distance metric specifically designed for clustering continuous compartments in the bioreactor. Our metric accounts for the similarities across multiple features.

The loss function (inertia), which is minimized during model training, is modified as follows:

$$J(C, \vec{\mu_l}) = \sum_{l=1}^{k}\sum_{C(\vec{x_i})=l} d(x_i, \vec{\mu_l}) \tag{4}$$

In this project we used a composite distance metric $d(\vec{x_i}, \vec{\mu_l})$ in multidimensional space (Equation 5), between the clusters centroids $(\vec{\mu_l})$ and the data points (x_i). This metric combines a spatial Euclidean distance for the 3-D mesh elements with a Manhattan distance for various weighted features. This dual approach balances geometric proximity and feature similarity, preventing bias towards outliers, which is a common problem in CFD simulations of bioreactors. For example, it is common to observe a very high substrate uptake rate around the feed point. This method offers a robust solution for clustering where both spatial $(\vec{p_{i,e}})$ and feature-based $\left(\vec{p_{i,f}}\right)$ relationships are key.

$$d(\vec{p_1}, \vec{p_2}) = \left|\left|\vec{p_{1,e}} - \vec{p_{2,e}}\right|\right| + \sum_{f=1}^{n_f} \vec{w}\cdot\left|\vec{p_{1,f}} - \vec{p_{2,f}}\right| \tag{5}$$

Categorical variables can be included in the custom distance metric along with continuous variables by employing one-hot encoding. This technique converts each category value into a binary vector. Each category is represented by its own unique binary column, where the presence of a category is marked by '1′ and the absence by '0′, allowing both categorical and continuous data to be integrated seamlessly for analysis. Clustering can be conducted with multiple objective features, which can be weighted $(\vec{w})$ to construct a comprehensive model including multiple features: for example, both glucose and oxygen gradients can be included.

The clustering process employed k-means++ for initializing the centroids' positions, a method particularly effective due to its strategy of placing initial centroids to maximize the diversity of starting points, thereby significantly improving the likelihood of achieving a global optimum compared to random initialization (Arthur & Vassilvitskii, 2007). We set the number of initializations to 10 for each run (n_{init}). This approach yielded consistent results in approximately 90% of cases (lowest inertia), effectively avoiding local minima.

The full description of the custom K-Means algorithm developed using NumPy with vectorization can be found in Section 3.1.

2.4. Performance Evaluation

In evaluating clustering effectiveness and determining the optimal number of clusters, we applied the scoring metric at different levels: individual data points, compartments (clusters) and whole simulations. This approach allows for a comprehensive assessment, aiding fine-tuning of the cluster hyperparameters to best represent the target

variable's continuous body.

An optimal regime analysis was conducted to identify the ideal number of clusters representing the geometry and target variables, also called features. This involved testing multiple compartment numbers and utilizing whole simulation scores to establish a Pareto front or identify elbow points where significant score changes occurred. A clear elbow point indicates a threshold beyond which adding additional clusters yields diminishing returns in model improvement. The Pareto front was used to determine a compromise solution for inversely proportional metrics.

2.4.1. Inertia

Already described above in the discussion of Equation 4, inertia is the sum of the distances from each data point to the centroid of its assigned cluster. Inertia is a whole-simulation metric which can be used to compare simulations with different numbers of compartments.

2.4.2. Davies-Bouldin Index

The Davies-Bouldin Index (DBI) evaluates the overall quality of clustering across the entire simulation by calculating the ratio of within-cluster scatter to between-cluster separation, as described by (Davies & Bouldin, 1979). Lower DBI values indicate superior clustering quality, reflecting a lower average of the maximum similarity measures across all clusters. The description of the DBI used in the algorithm can be found in Section 3.2.1.

2.4.3. Silhouette Score.

The Silhouette score is a measure of how similar an object is to its own cluster compared with similar it is to the other clusters based on the distance metric used to determine the similarity. It ranges from -1 to 1, with a high value indicating that the object is well matched to its own cluster and distinct from neighboring clusters. Thus, values close to 1 imply ideal clustering (Rousseeuw, 1987).

The metric is data point specific, however the average values of the metric can give general scores for each specific cluster and for the whole clustering process.

Description of the Silhouette score as computed in our algorithm can be found in Section 3.2.2.

2.4.4. Compartment Homogeneity.

This custom metric assesses compartment homogeneity independently, focusing on internal similarity rather than overall clustering performance. It evaluates similarity within a compartment but not clustering performance. Since it is specific to each cluster, cluster homogeneity can be applied to the entire simulation through a weighted average, with weights based on data point distribution. This metric considers only the clustering features, excluding the spatial coordinates.

Description of compartment homogeneity as computed in our algorithm can be found in Section 3.2.3.

2.5. Case Studies

Two distinct studies were conducted to verify the performance of our algorithm, with modifications tailored to the specific decision variables set prior to execution. These decision or objective variables are the ones used to cluster the dataset. Given that our datasets originated from physics simulations (CFD), they encompass multiple variables from the 3-D continuous body as results. For our analysis, biology-related variables were chosen because they encapsulate the integrated effects of multiphysics phenomena (fluid flow, turbulence, transport of species and bio-kinetics), providing a holistic view compared to variables like velocity profiles which only reflect fluid dynamics and turbulence. Additionally, the chosen objective variable was rendered dimensionless, and logarithmic transformations were applied to enhance its correlation with the distance to the feeding point, a gradient-related variable, as detailed in Section 4.1.1 (Exploratory Analysis) and Section 4.2.1 (Exploratory Analysis). The specific choices made are outlined in Table 3.

The data from the different CFD results were extracted using in-house user defined functions. The datasets were subsequently analyzed in python scripts.

3. Theory/Calculation

Some sections of the materials and methods can be further developed, based on the mathematical description of some variables.

3.1. Custom k-Means Computation

Custom k-means algorithm. Each step in the process represents a function in the code, vectorized using NumPy.

The data given to the algorithm is defined as a set of m datapoints with n features per each one.

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} = \begin{bmatrix} \overrightarrow{x_1} \\ \vdots \\ \overrightarrow{x_m} \end{bmatrix} \in \mathbb{R}^{n \times m} \tag{6}$$

3.1.1. k-Means Algorithm

The custom k-means algorithm has been coded in python, using NumPy broadcasting for vectorization (Harris et al., 2020). This factor is especially important for computation efficiency and allows to scale entire vectors before advancing to the next step.

The custom k-means performs as follows:

1. Initialize centroids: Use k-means++ as initialization method (Arthur & Vassilvitskii, 2007), it gives better results than the random initialization method, reaching faster convergence. It consists of four steps:
   a. Select first centroid ($\overrightarrow{\mu_1} \in \mathbb{R}^n$) at random from the data points:

$$\overrightarrow{\mu_1} = \overrightarrow{x_i}, \quad \text{With probability} = \frac{1}{m}, \quad i \in \{1, 2, \ldots, m\} \tag{7}$$

   b. For each remaining centroid, $\overrightarrow{\mu_l}$ ($l = 2, 3, \ldots, k$), calculate the distance from each data point $\overrightarrow{x_i}$ to the nearest centroid that has already been initialized. Use a custom distance metric $d(\overrightarrow{x_i}, \overrightarrow{\mu_l})$, also referred as d.

$$D(\overrightarrow{x_i}) = \min\{d(\overrightarrow{x_i}, \overrightarrow{\mu_l}) : l \in \{1, \ldots, k-1\}\} \tag{8}$$

   c. Select the next centroid $\overrightarrow{\mu_l}$, from the dataset with probability proportional to the distance calculated in the previous step. Data points which are farther from initialized centroids will have a higher probability of being chosen as next centroids:

Table 3
Clustering example from the CFD simulations, using different decision variables available in the algorithm.

Reactor Set-Up	Number of Clusters	Objective Variable	Optimum Regime Analysis Performed
Rushton Stirred Tank	3-100	$\log_{10}\left(\frac{q_S}{q_{S,max}}\right)$	Yes
	3	$\log_{10}\left(\frac{q_S}{q_{S,max}}\right)$, position respect to impeller	No
Bubble Column	3-100	$\log_{10}\left(\frac{q_S}{q_{S,max}}\right)$	Yes

$$P(\overrightarrow{x_i}) = \overrightarrow{x_i} = \frac{D(\overrightarrow{x_i})}{\sum_{i'=1}^{m} D(\overrightarrow{x_{i'}})},\ i \in \{1, 2, \ldots, m\} \tag{9}$$

d. Repeat steps 2-3 until all k centroids have been initialized.

2. Assign datapoints to clusters based on similarity using the custom distance metric.

$$C_{(l)}(\overrightarrow{x_i}) = \text{argmin}_{(l)} d(\overrightarrow{x_i}, \overrightarrow{\mu_l}) =$$

$$= l : \{\overrightarrow{x_i} : d(\overrightarrow{x_i}, \overrightarrow{\mu_l}) \le d(\overrightarrow{x_i}, \overrightarrow{\mu_o}) \forall o, l \in \{1, 2, \ldots, k\}\}, \forall i \in \{1, 2, \ldots, m\} \tag{10}$$

3. Update clusters' centroids by computing the mean of all data points assigned to each cluster, use all features measured per observation.

$$\overrightarrow{\mu_l} = \frac{1}{\left|\overrightarrow{C_l}\right|} \sum_{\overrightarrow{x_i} \in \overrightarrow{C_l}} \overrightarrow{x_i},\ \forall l \in \{1, 2, \ldots, k\},\ i \in \{1, 2, \ldots, m\} \tag{11}$$

4. Repeat steps 2-3 till convergence of the loss function (inertia):

$$J\left(\overrightarrow{C}, \overrightarrow{\mu_l}\right) = \sum_{l=1}^{k} \sum_{\overrightarrow{C(x_i)} = l} d(x_i), \overrightarrow{\mu_l}$$

Once the model is trained, so the cluster centroids are well defined we can perform a prediction for any other data point in the set. This step is only useful when we have not used all the data points to define the clusters (centroids), so we can predict the ones not used so far. For the prediction we just need to repeat step 2 to assign the closest cluster to the data points.

In case of too many data points to train the model, there is the option to just select a fraction of the dataset as a training set and then predict the rest of it.

The k-means steps explained above are repeated n_{init} times (10) to assure we do not get into a local optimum for the loss function (inertia), around 90% of the times, global optimum was reached.

3.1.2. Custom Distance Normalization

As we are working on a "vectorized manner" the distances represent matrices of distances between all points and all centroids for all features. These are normalized by the mean and variance before being added together to have the same effect on the clustering algorithm.

$$S(V) = (V - \overline{V}) \oslash \sigma(V) \quad \rightarrow \quad D_b^{\text{norm}} = S(D_b) \quad \forall b \in \{e, f\}, \quad \text{with } D_b^{\text{norm}}, D_b \in R^{m \times k \times n} \tag{12}$$

The normalized distances are finally summed up to compute the final distance between two data points, or in this case the data points and the cluster centroids.

$$d(\overrightarrow{x_i}, \overrightarrow{\mu_l}) = d_p^{\text{norm}}\left(\overrightarrow{\mu_{l,e}}, \overrightarrow{x_{i,e}}\right) + \overrightarrow{w} \cdot d_f^{\text{norm}}\left(\overrightarrow{x_{i,f}}, \overrightarrow{\mu_{l,f}}\right) \tag{13}$$

3.2. Clustering Performance Evaluation

Once the clustering has been predicted, there is a need to evaluate how good the clustering has been and evaluate what is the optimum number of clusters to describe the continuous body for the target variable.

For the evaluation step, multiple custom k-means can be performed with a different number of clusters at each step. Per each prediction, four scores/techniques are used to evaluate how good is the clustering and how representative are they from the continuous body.

3.2.1. Davies-Bouldin Index Computation (DBI)

1. Compute intra-cluster distances: Average distances between all data points ($\overrightarrow{x_i}$) within a cluster l and its centroid $\overrightarrow{\mu_l}$.

$$S_l = \frac{1}{|C_l|} \sum_{\overrightarrow{x_i} \in C_l} d(\overrightarrow{x_i}, \overrightarrow{\mu_l}) \tag{14}$$

2. Compute inter-cluster distances: Distances between centroids of two distinct clusters l and o, given by:

$$d_{l,o} = d(\overrightarrow{\mu_l}, \overrightarrow{\mu_o}) \tag{15}$$

3. Compute cluster similarity: A ratio that compares the sum of intra-cluster distances of two clusters (l and o), to their inter-cluster distance ($d_{l,o}$):

$$R_{l,o} = \frac{S_l + S_o}{d_{l,o}} \tag{16}$$

4. Compute DBI:

$$\text{DBI} = \frac{1}{k} \sum_{l,o=1}^{k} \max(R_{l,o}) \quad \forall l \neq o \tag{17}$$

3.2.2. Silhouette Score

1. Compute the average distance between a point in a cluster and all other points within the same cluster. This measures the cohesion within a cluster.

$$a(\overrightarrow{x_i}) = \frac{1}{|C_l| - 1} \sum_{\overrightarrow{x_i}, \overrightarrow{x_j} \in C_l} d(\overrightarrow{x_i}, \overrightarrow{x_j}) \quad \forall i \neq j \tag{18}$$

2. Compute the minimum average distance between a point and all points in any other cluster, of which the original point is not a member. This quantifies the separation of the point from its nearest neighboring cluster.

$$b(\overrightarrow{x_i}) = \min\left[\frac{1}{|C_l|} \sum_{\overrightarrow{x_j} \in C_l,\ \overrightarrow{x_i} \notin C_l} d(\overrightarrow{x_i}, \overrightarrow{x_j})\right] \tag{19}$$

3. Compute the silhouette values for each data point:

$$s(\overrightarrow{x_i}) = \frac{b(\overrightarrow{x_i}) - a(\overrightarrow{x_i})}{\max(a(\overrightarrow{x_i}), b(\overrightarrow{x_i}))} \tag{20}$$

4. Compute the average silhouette score for all the data points:

$$\text{SS} = \frac{1}{m} \sum_{i=1}^{m} s(\overrightarrow{x_i}) \tag{21}$$

3.2.3. Homogeneity Score

1. Compute the coefficient of variation for each cluster:

$$\mathrm{CV}_l = \frac{\sigma(C_l)}{\overline{C_l}} \tag{22}$$

2. Compute a weighted average based on how many points has each cluster for the CV in all clusters. This one is the general (average) homogeneity score. Subtract it from 1 so it is a score between 0 and 1, being 1 the perfect homogeneity (no standard deviation):

$$H = 1 - \mathrm{CV}_{\mathrm{weighted}} = 1 - \frac{1}{m}\sum_{l=1}^{k}\frac{|C_l|}{m}\mathrm{CV}_l \tag{23}$$

3. Compute the homogeneity score for each individual cluster. Which is the CV relative to the H:

$$h_l = \frac{\mathrm{CV}_l}{H} \tag{24}$$

3.3. Dataset Sampling and Computation Scalability

The custom k-means algorithm and associated clustering metrics such as silhouette scores demand substantial computational resources, often exceeding the capabilities of standard laptops, particularly when processing entire datasets from high-resolution CFD simulations. The computational complexity of this algorithm typically scales linearly with the number of features, clusters, data points and initializations, represented as $O(m{\cdot}n{\cdot}k{\cdot}n_{in})$. However, under certain conditions where exhaustive pairwise distances are computed or in specific algorithm implementations, this complexity can escalate to $O(m^2)$ (Pakhira, 2014). Effective management of computational resources is crucial, especially in the context of optimum regime analysis within the Pareto front, where performance metrics are computed for each "experiment".

Despite the algorithm's efficient vectorization, RAM limitations remain a challenge. Serialization techniques like looping in silhouette score computations help manage memory by avoiding large pairwise distance matrices, though this slightly reduces processing speed. The greatest computational demands occur when managing many clusters or conducting extensive optimum regime analyses, involving multiple simulations to assess different clustering scenarios and metrics.

For performance evaluation of the clustering, using about 20% of the dataset is practical, especially with a high number of clusters. Using 5-10% of the dataset for CFD simulation models keeps the distribution of the variables almost un-changed, as verified by Kernel Density Estimation (KDEs). For training with very high cluster counts (over 300), using up to 25% of the data is advisable. Prediction phases, which are less resource-intensive, can efficiently utilize the entire dataset.

Ensuring that the KDE of the sampled dataset matches that of the full dataset is crucial, as it confirms that the essential characteristics of the data are preserved. This method guarantees that even if specific points are not directly sampled, the surrounding volumes still accurately represent the reactor dynamics. This balance between computational efficiency and analytical accuracy is essential for the effective application of k-means clustering in large-scale bioreactor simulations. Importantly, sampling the datasets does not alter the distribution of variables, given the extensive size of datasets derived from CFD simulations.

For this study, we utilized several cores of a high-performance computing (HPC) system to conduct optimum regime analysis. The clustering range was serialized, but within each clustering operation, parallelization was employed using 16 cores—the maximum number supported by our algorithm. The total RAM consumption reached approximately 24 GB, primarily due to the computation of pairwise distance matrices required for score calculation. The comprehensive optimum regime analysis, spanning 2 to 300 clusters, was completed in about one day. In cases where the dataset was clustered with a fixed number of clusters, the processing time varied significantly: clustering with fewer clusters (e.g., 4) was nearly instantaneous, whereas clustering with 300 clusters required up to 30 minutes, depending on the fraction of the original dataset used.

4. Results & Discussion

We evaluated our algorithm's effectiveness using two case studies: One from a 202 m^3 bioreactor and another one from an external loop gaslift reactor (840 m^3). For the first case study we compared two different clustering approaches based on different objective variables. In the presented case studies, the focus lies on the application of the approach towards identification of coherent regimes of similar reaction conditions in the bioreactor.

4.1. Case Study: 202 m^3 Rushton-Impeller Bioreactor

4.1.1. Exploratory Analysis

The exploratory analysis started with the creation of a correlation matrix that covered essential metrics for reactor operation, such as response variables (like concentrations), positional variables, and flow field characteristics (Fig. 4). This correlation matrix enabled the identification of variables with significant correlations with the response or "target" variables.

For this study, cell performance was used as the response variable. It was normalized from 0 to 1 against the maximum uptake rate, allowing for a comprehensive analysis of the cell efficiency $\left(q_S/q_{S,max}\right)$.

The correlation matrix presented in Fig. 4 indicates pronounced correlations between positional features, notably between the axial positional variable $\left(y_{cyl}\right)$ and various response variables. This strong association underlines the presence of an axial gradient, a distinctive attribute of Rushton-impeller reactors. Such reactors are known for their powerful radial mixing efficiency. The results of the analysis extend to the categorical variables that define the spatial relation of exam measurement point to the impellers, thereby delineating the distinct mixing zones within the bioreactor: above the top impeller, between impellers, and below the bottom impeller.

The observed correlations intensified upon the inclusion of a variable representing the Euclidean distance from each point in the bioreactor to the feeding point. We attribute this enhancement in correlation strength to this variable's encapsulation of information from other positional variables, specifically the cylindrical coordinates r and θ. However, it is predominantly influenced by y_{cyl}, reflecting the reactor's elongated cylindrical shape. Upon comparison with alternative distance metrics such as Manhattan or Haversine, which is tailored to account for radial shapes, the Euclidean distance shows better correlation strength with the response variable.

Equally surprising is the minimal correlation observed between the fluid flow variables and the response variables. This highlights the key importance of the species transport and microbial reaction kinetic models in the simulations.

Having chosen the strongest correlated variable we visualized the gradients in a bivariate KDE plot (Fig. 5), which gave us a "fingerprint" of the substrate gradient in the bioreactor.

The KDE plot does not show a perfect correlated trend but normal-shaped distributions, oval and circular, along the axis. These performance clusters appear due to the complicated fluid dynamics inside the stirred tank reactor, and the associated flow circulation patterns (Haringa et al., 2016).

As these clusters indicate spatially contiguous regions of similar cell response $\left(\frac{q_S}{q_{S,\max}}\right)$, Fig. 5 gives a clear view of how the reactor can be compartmentalized using clustering techniques (unsupervised machine learning), while ensuring spatial connectivity.

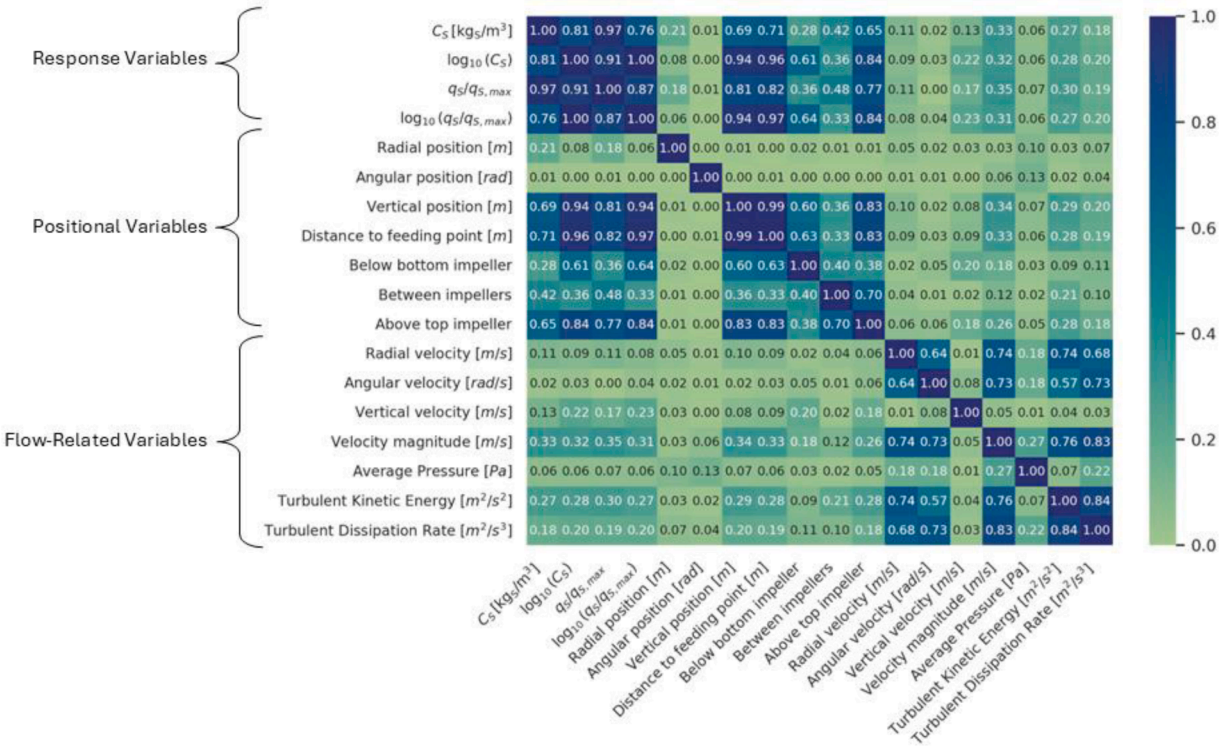

Fig. 4. Correlation matrix delineating the primary features extracted from CFD simulations. The features are categorized into three primary groups based on their intrinsic correlations. The values of the matrix are Pearson's correlation coefficients.

4.1.2. Optimum Regime Analysis

Clustering techniques offer a unique benefit in their ability to self-evaluate performance through several metrics. These metrics can be inversely related, particularly as cluster numbers increase. Higher cluster counts usually capture more information while producing more regime volumes in the reactor (diffuse boundaries between compartments).

In this complex scenario, no single metric could be optimized without affecting another. This is depicted in Fig. 6, which illustrates the correlation of these scores across a range of 2 to 150 clusters. Broadly, the metrics diverged into two trends:

Inertia and homogeneity score improve with more clusters, capturing more information from the continuous body and creating more uniform compartments. On the other hand, silhouette score and DBI (Davies Bouldin Index) deteriorate as the number of clusters increases, leading to a potential overlap and less separation, diminishing cluster (compartment) clarity.

The trade-off between these metrics produced a Pareto front (Fig. 6), allowing for the identification of an optimal balance between information retention and clear differentiation of the compartments. This strategy facilitates efficient multivariate optimization in clustering analysis. For the continuous scores, which is only inertia in this case, a sweet spot is observed by the presence of a knee (inflection point) in the analysis, beyond which adding more clusters did not significantly improve the performance.

Fig. 6 illustrates a Pareto front for optimal compartment/regime number selection in a bioreactor, balancing design efficiency and accuracy. This methodology works with homogeneity and DBI as well as with inertia and silhouette score: the user can choose which variables to analyze this way. The analysis extends to a 3-D Pareto front, although as seen in the score's correlation (Fig. 6), two scores should be sufficient.

The Pareto front method demonstrates remarkable adaptability across a broad spectrum of experimental conditions, offering a diverse array of optimal solutions for cluster analysis. At the higher end of the spectrum, it favors well-defined clusters that are ideal for scale-down experiments where clarity and manageability are paramount. These clusters not only offer enhanced physical meaningfulness, as evidenced by higher silhouette scores, but also ensure greater compartment clarity. Conversely, the lower end of the Pareto front presents solutions with a larger number of compartments, resulting in finer granularity and more diffuse clusters. While this retains more data from the original dataset, it does so at the expense of some physical meaningfulness, indicated by lower silhouette scores.

Although alternative multi-objective optimization techniques such as genetic algorithms or simulated annealing might effectively navigate complex landscapes to avoid local minima, they introduce significant complexities and greater computational demands. Moreover, these techniques lack the Pareto front's ability to provide a clear, visual representation of trade-offs, from which users can select the most suitable option based on their specific requirements. Opting for a high number of clusters might be impractical for experimental scale-downs, but such configurations are crucial for developing compartment models that effectively capture the heterogeneity of the reactor. These models significantly reduce computational costs compared to traditional computational fluid dynamics (CFD) simulations.

Ultimately, the Pareto front method is employed to optimize the number of clusters' selection, utilizing a dataset enriched with physics simulations (CFD) and reactor geometry data. This approach not only ensures that the clusters retain a high degree of physical relevance but also maximizes the informational content of the original dataset. By leveraging the Pareto front, we identify clusters that exhibit the most substantial physical meaning, evidenced by high silhouette scores and low inertia. Clusters with lower silhouette scores, though less distinct and possibly less physically meaningful, are crucial for maintaining the integrity of the dataset's information. Rather than amalgamating these points into more distinct clusters, defining them separately allows for a more comprehensive representation of the original data's complexity. This method balances the clarity and physical significance of each cluster against the overarching goal of data preservation.

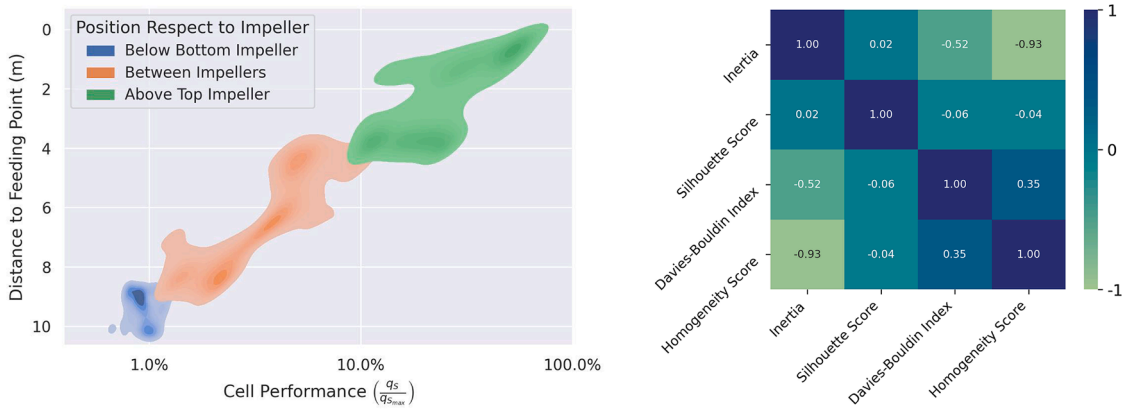

Fig. 5. (A). Kernel Density Estimate (KDE) plot to visualize the distribution of observations between the two most correlated variables in the bioreactor. This correlation is explained due to the bad axial mixing resulting from the Rushton turbines. The plot includes a classification of each mesh element position with respect to the impellers (legend). The KDE plots are weighted using the volumes corresponding to each element in the CFD simulation. The scale for the cell performance is logarithmic. (B) Correlation matrix showing the relationships between scores. The scores came from clustering the bioreactor using a range of 2 to 150 clusters and computing each individual score. Positive values indicate performance improvements with more clusters, while negative coefficients highlight score trade-offs.

4.1.3. *Regimes Definition for a Rushton Stirred Tank*

Optimal regime analysis identified five clusters as an effective representation of a Rushton Impeller bioreactor (Pareto front), tailored to its geometry, fluid dynamics and objective variables, so five regimes are identified based on the cell performance inside the reactor.

In Fig. 7, the clusters are mapped onto the CFD simulation, highlighting their spatial distribution. As observed, the clustering algorithm effectively segregates the bioreactor into distinct zones, each characterized by the cell performance variable $\left(\frac{q_S}{q_{S,\ max}}\right)$. These regimes are spatially continuous and exhibit well-defined shapes and boundaries, whilst making physical meaning as the objective variable derived from physics simulations are used to cluster the reactor. A notable observation from the 3-D regime analysis is the inadequacy of a simple 2-D radial symmetry approach. The influence of the baffle on compartment formation is particularly evident in Fig. 7 (left), where a rear view of the compartments demonstrates their

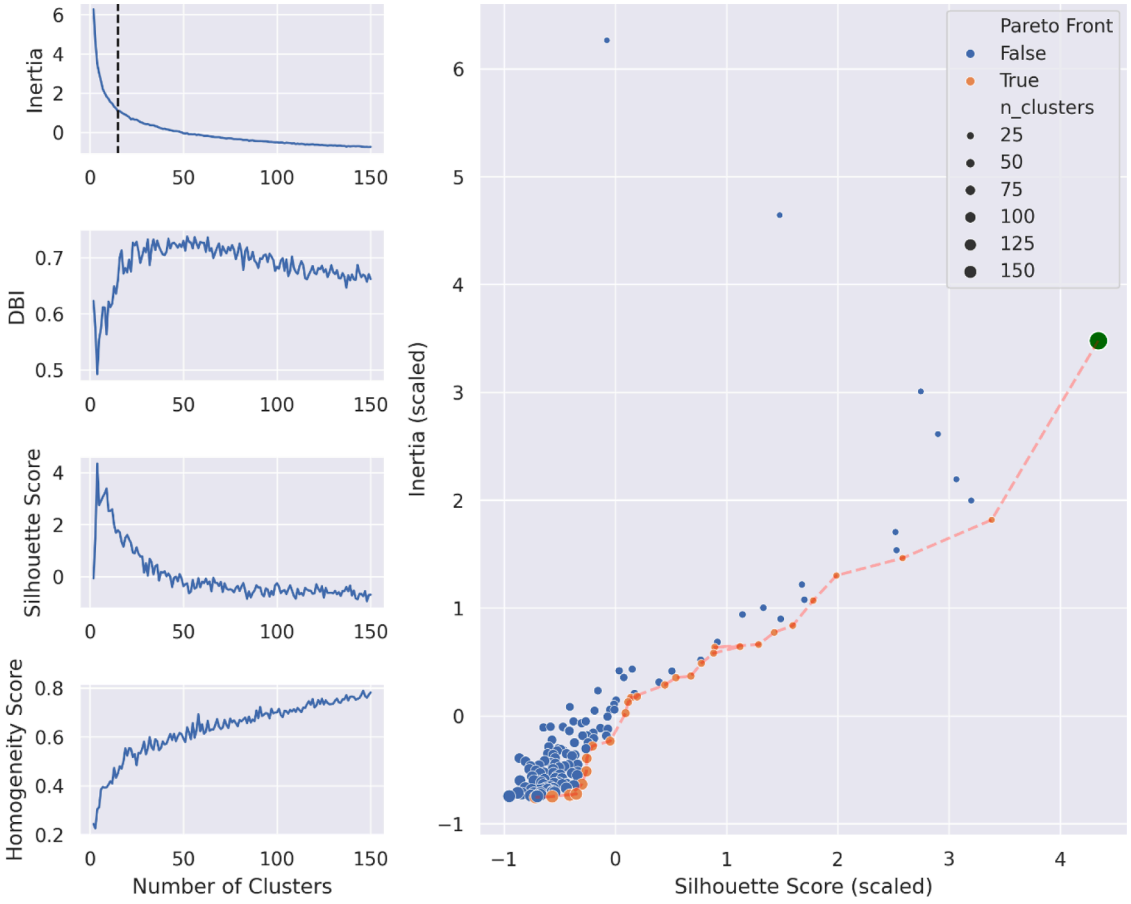

Fig. 6. Performance metrics for the Rushton bioreactor compartmentalization over 150 clusters (compartments). The left plots describe each individual score per each clustering with a different number of clusters. The inflection point in the inertia plot is indicated by the vertical dotted line. The right plot identifies the Pareto front, which is highlighted in orange, and it is defined by the inertia and silhouette score. These two metrics are inversely proportional regarding their performance.

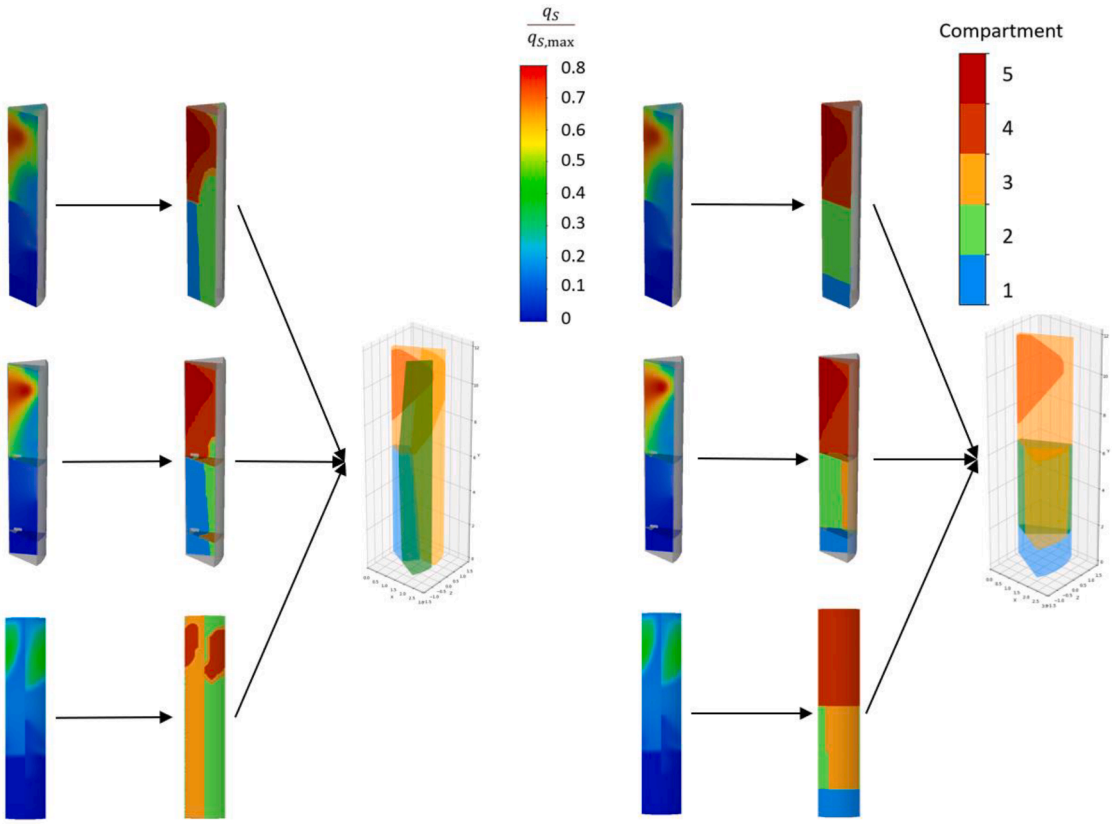

Fig. 7. Compartmentalization results for the Rushton tank using cell performance as a single clustering objective (left) or using it combined with the position of each data point with respect to the impeller, which is a categorical variable. The continuous solution is clustered into compartments which are continuous as seen in the transparent plot of the reactor (1/6 symmetry).

spatial border over the vertical plane (baffle).

By integrating multiple variables, including a categorical one representing each data point's position relative to the impeller and the cell performance, the clustering algorithm effectively compartmentalized the bioreactor based on more complex situations. As shown in Fig. 7 (right), this approach ensures the cell performance-based zones do not intersect with axial impeller planes, demonstrating the algorithm's ability to recognize multidimensional spatial regions while balancing between multiple objective variables. In this case the a 'constraint' is applied in the compartmentalization (no clusters between impeller planes). The physical sense of the clusters is now defined by both the objective variables, cell performance and position respect to the impeller, as seen in the new clusters generated. Using multiple variables as objective for clustering could describe the bioreactor in a more complete way and could lead to easier validation.

Although we do not perform an extensive sensitivity analysis in this study, altering certain operational parameters in the CFD simulations leads to slight variations in the resulting clusters. For a specific bioreactor geometry and turbine design, the fluid flow and gradient profiles remain quite similar regardless of the operational conditions. However, changing these conditions affects the optimal number of clusters needed to accurately define the bioreactor, necessitating a new optimum regime analysis. Consequently, the shapes of the clusters may also change. Interestingly, when we increase the stirring speed by 100% and perform clustering with five clusters, the results are quite similar to the original case. This suggests that, as the tank becomes more homogeneous due to increased stirring, a smaller number of clusters might better describe the reactor, potentially leading to clusters with different shapes

Validating the compartments generated by any algorithm in a real-world bioreactor setting presents challenges due to the current difficulty of obtaining detailed spatial measurements in large-scale reactors. Typically, measurements are limited to integrated parameters like power input or mixing time, which are useful for validating overall CFD performance but lack the spatial resolution needed for direct compartmentalization validation. Our primary objective in this paper is to introduce and validate a clustering algorithm that effectively partitions bioreactors into physically meaningful compartments based on available data. The algorithm is designed to utilize real spatial 3-D data when available, making it ready for future applications with such datasets. While the practical applicability is currently limited due to the scarcity of real-world spatial data, our platform is developed to be applicable to both CFD and real data, ensuring it is prepared for future experiments. In the meantime, clustering models rely on CFD simulations that have been validated against experimental data for these global parameters (Bach et al., 2017; Brannock et al., 2010; Puiman et al., 2022). When a CFD model accurately predicts these integrated variables, it increases confidence in its ability to represent the reactor's internal fluid dynamics. Consequently, the compartments derived from such validated CFD data are considered realistic representations of spatial heterogeneities; essentially, the clusters will reflect reality as accurately as the quality of the input data permits. As explained in the Methods Section 2, the CFD simulation used here is designed to showcase the application of the clustering algorithm rather than to provide a precise description of a bioreactor. Simplifying assumptions, such as pseudo-steady state and radial symmetry, are employed, which can overlook phenomena like macro-instabilities.

Future work could involve employing advanced measurement techniques or soft sensors to obtain spatially resolved data, providing direct validation of the compartmentalization (Jiang et al., 2020) and enabling real-time, comprehensive biological measurements across the entire reactor space (e.g., spatial-based proteomics). Although a major advantage of using CFD-generated synthetic data is the high spatial and temporal resolution it offers, experimental data are often much sparser due to constraints inherent in experimental work, particularly at large scales. Nonetheless, the algorithm should still be capable of capturing the regimes accurately, provided that the data are representative of the overall bioreactor state. To ensure the

representativeness of the experimental low-resolution data, multiple datasets may need to be collected, and inferential statistical methods applied to assess and enhance data reliability.

4.1.4. Metabolic States Inside the Rushton Stirred Tank

This methodology can also be extended to cluster metabolic states as categorical objective variables if a suitable kinetic or metabolic model is integrated. However, clustering based on metabolic states within the reactor is complex, primarily because obtaining comprehensive metabolic data across the entire reactor space is currently challenging due to experimental limitations. To address this challenge, one approach is to use computational models that couple fluid dynamics with metabolic processes. For example, CFD-derived data can be integrated with metabolic flux models, facilitated by hydrodynamic compartment models, to estimate the intracellular flux distributions using Genome-Scale Metabolic Models (GEMs) and Flux Balance Analysis (FBA) (Promma et al., 2024). This approach allows the algorithm to map out volumes related to different metabolic states based on predicted intracellular clusters. A significant challenge in this integration is the high dimensionality of metabolic data, which involves thousands of reactions and entails high computational demands. To manage this complexity, dimensionality reduction techniques like Principal Component Analysis (PCA) can be applied, as metabolic fluxes are often highly correlated.

Alternatively, hybrid modeling techniques that combine first-principles models with data-driven approaches, such as neural networks, can be employed to model the complexity of biological systems (Pinto et al., 2022; Shah et al., 2022). This method would lead to a more straightforward clustering process due to the reduced number of features. However, it could lack the deep description of the metabolic networks provided by GEMs.

Assuming there is a direct link between the conditions outside a cell and its internal metabolic state, the algorithm can map out areas related to key metabolic processes. In our study, we identified regions corresponding to two key processes: overflow metabolism, marked in maroon, and starvation, marked in blue. We achieved this by using cell performance, a continuous variable, as the main clustering criterion. Both states were the best-classified compartments based on their silhouette scores (Section A.1). These findings align with those of (Nadal-Rey et al., 2023), reaffirming the high physical relevance of clusters with high silhouette scores.

4.2. Case Study: 840 m^3 Airlift Loop with Down Comer (Bubble Column)

The second case study represents an airlift loop with a down-comer and a total volume of 840 m^3 for syngas-to-ethanol fermentation, as described by Puiman et al. (2022).

4.2.1. Exploratory Analysis

In this study, like the Rushton case, cell performance emerged as the variable most correlated with positional factors, though less strongly than before. Thus, it was again chosen as the objective variable. Here, cell performance is defined as the ratio of the observed to the maximum apparent carbon monoxide uptake rate $\left(\frac{q_{CO}}{q_{CO,\max(app)}}\right)$, reflecting the fact that kinetic inhibition prevents the cells from the theoretical maximum $\left(q_{CO,max}\right)$.

Notably, Kernel Density Estimation (KDE) analysis shows a clear departure in gradient shapes from those observed in the Rushton tank, indicating distinct kinetic and performance characteristics under different tank configurations (Fig. 8).

The comparative analysis of the Rushton tank and the bubble column case studies reveals distinct differences in flow dynamics and their correlation with cell performance. The Rushton tank exhibits a low radial gradient but a high axial gradient, in contrast with the bubble column, which demonstrates the opposite pattern. This variation can be attributed to the bubble column's lower average liquid phase kinetic energy and the absence of forced radial flow.

Furthermore, the correlation between the distance to the feeding point and cell performance is weaker in the bubble column compared with the Rushton tank. This weaker correlation is due to two main factors: a modest pressure gradient and variations in substrate composition, as detailed by Puiman et al. (2022). Additionally, since the substrate is a dissolved gas rather than a liquid solute, the maximum concentration in the liquid phase is capped by (local) solubility, which varies slightly through the domain due to the pressure gradient. This means substrate is introduced in the liquid phase throughout the domain via gas-liquid mass transfer, rather than a concentrated feed at a single point. Consequently, the correlation between local cell performance and 'feed point' (sparger) distance is much weaker.

Kernel Density Estimation (KDE) analysis of the bubble column reveals that gradient variability is influenced by turbulence levels and re-circulation patterns (Fig. 8). High radial gradients, challenging for the clustering algorithm, correlate with re-circulation patterns and low turbulence (zones 5 and 9). In contrast, areas with smaller re-circulation patterns and higher turbulence (regions 2, 4 and 7) display more consistent radial gradients, depicted as two-dimensional Gaussian distributions in the KDE plot. Special regions like 1 and 8 represent the headspace and liquid-gas interface.

4.2.2. Regime Definition for a Bubble Column

Optimum regime analysis for the column identified 9 compartments as the optimum number for the "cell performance" variable. The result is however plotted with 10 regimes to account for the headspace also considered in the CFD simulations (Fig. 9).

The algorithm successfully partitioned the column into the optimum nine distinct zones, showcasing a pattern different from the more orderly zones observed in the Rushton tank. This difference is attributed to the complex re-circulation patterns typical in gas-mixed reactors. Notably, Compartment 1 precisely captured the primary uplift gas flow from the inlet and Compartment 3 comprises the whole down-comer, ensuring seamless spatial integration among the zones.

Prior to this clustering, the headspace compartment (Compartment 10) was specified to prevent any disturbance of the liquid phase patterns. Compartment 3 consolidates regions 2 and 3 (as depicted in Fig. 8), highlighting the interaction between these areas. Similarly, Compartment 9 focuses on region 4, the down-comer re-circulation inlet (also detailed in Fig. 8).

It is worth noting that the compartments created account for dynamics, as they are based on the average values from dynamic simulations spanning 100 seconds. This approach effectively integrates dynamic data into the clustering process. One potential strategy for incorporating dynamic data is to average time-series data from these simulations, as has been done so far. Alternatively, capturing snapshots at different stages of the simulation could provide insights into how compartments evolve over time in response to changes in fermentation variables and fluid flow within the reactor (Nadal-Rey, McClure, Kavanagh, Cassells, et al., 2021). As these operational conditions vary, the compartment configurations are also likely to change, reflecting the dynamic nature of the process. This dual approach of averaging data for stability and analyzing snapshots for temporal changes offers a comprehensive method for understanding and adapting to the dynamics within bioreactors.

The analysis of the airlift loop with a down-comer provides a blueprint for optimizing syngas-to-ethanol fermentation processes. By dividing the system into distinct compartments, we not only gain insight in the interplay between cell performance and fluid dynamics but also identify opportunities for efficiency improvements and give a proper start point for a scale-down replica of the industrial-scale case. This division facilitates targeted interventions, allowing for the precise control of conditions across different reactor zones. These findings underscore the potential of compartmentalization to enhance bioreactor designs, making them more effective for specific industrial applications and scale-down processes.

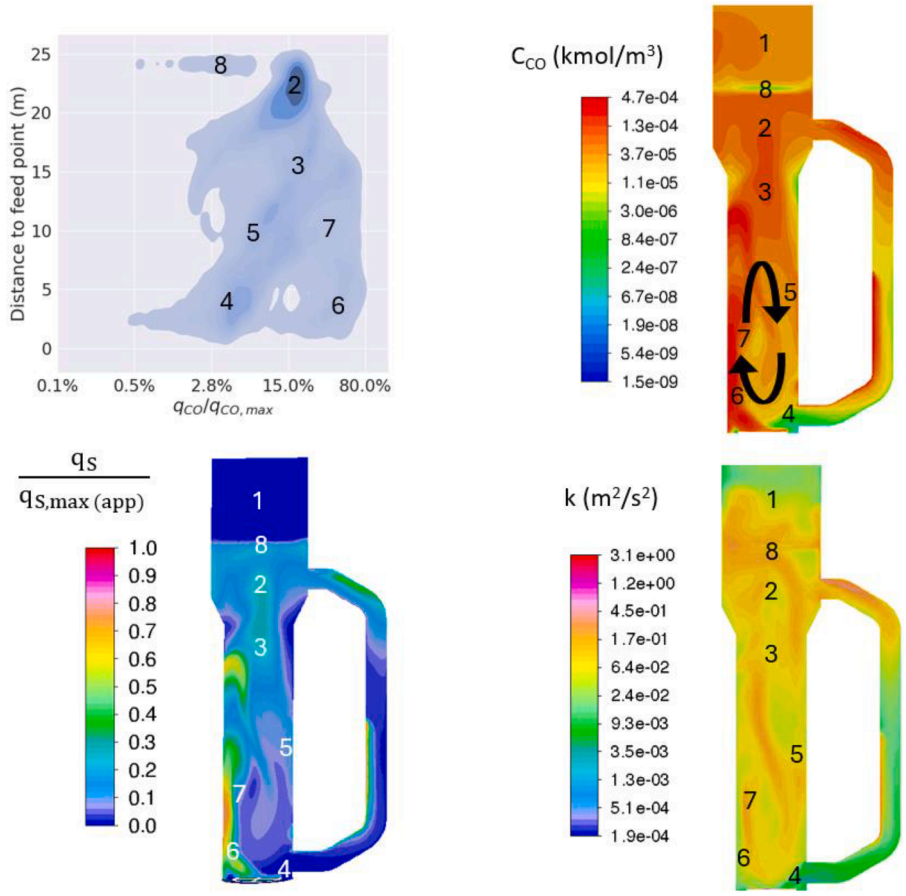

Fig. 8. Gradient exploratory analysis of the bubble column. KDE plot of cell performance, correlated against the distance to the feed point, without the headspace mesh elements (top-left). Contour plot of the CO concentration profile (top-right). Cell performance contour (bottom-left). Turbulent kinetic energy profile (bottom-right).

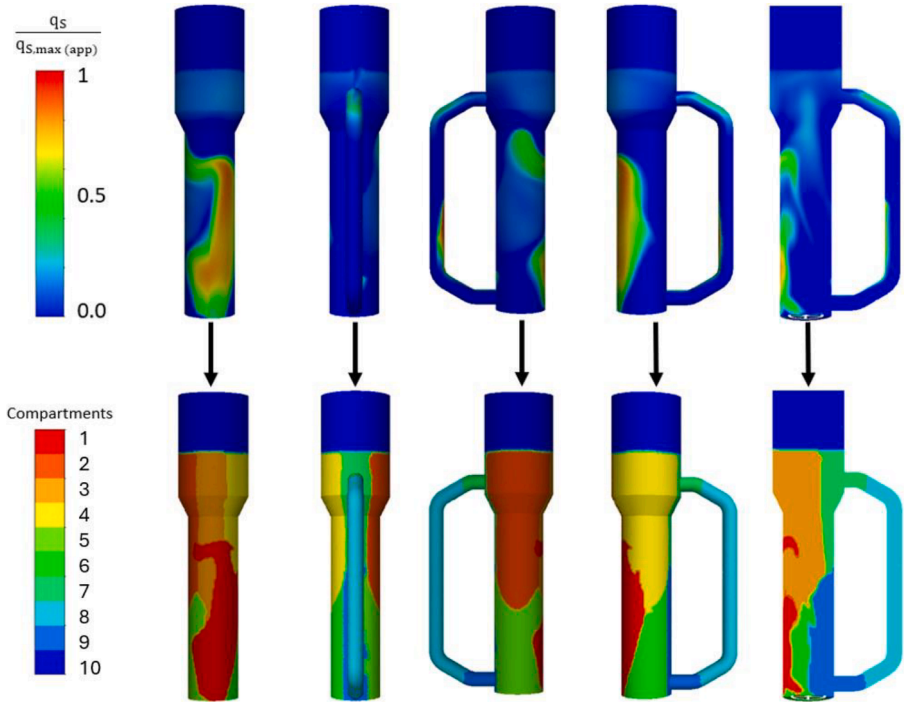

Fig. 9. Clustering of the bubble column based on the average of the cell performance over 203 seconds of transient simulation time. The compartments are ordered from high to low cell performance, being compartment 1 the one with the highest and 10 the lowest value (headspace).

5. Conclusions

We developed a novel multivariate algorithm for segmenting bioreactors, or similar 3-D finite element volume geometries, into spatially coherent compartments/regimes. Our method offers a precise way to identify the ideal number of compartments/regimes for specific conditions, based on amount of information captured and cluster clarity. Thus, the number of compartments/regimes can stop being a design variable based on user expertise. This capability enhances spatial analysis and lays a robust groundwork for subsequent scale down experiments, focusing on balancing the granularity of regimes against the comprehensiveness of the data captured by the clusters. In this work we focused on the use of algorithm to define clusters of coherent conditions, akin to regime analysis. These new clusters/compartments can be used alongside using any desired objective variable, including hydrodynamics-related variables, to further perform a classical compartmentalization to replicate the CFD simulations (Haringa et al., 2018).

The versatility and effectiveness of the algorithm were demonstrated through its application to two case studies: A 202 m^3 Ruston impeller bioreactor and an 840 m^3 airlift reactor (Puiman et al., 2022). In the Rushton tank, it identified five distinct compartments, illustrating the significance of 3-D compartmentalization due to the influence of the baffle and the impact of the axial gradient on cell performance and metabolism. In contrast, for the airlift reactor, the algorithm delineated nine compartments, uncovering the intricate dynamics of radial gradients and re-circulation patterns characteristic of gas-mixed systems (Tabib et al., 2008). The selection of the optimal number of compartments and their configuration was determined through rigorous analysis of inertia and silhouette score, which together formed a Pareto front. These compartments can be the basis for further down-scale experiments or stochastic parcel tracking, to gain a fine-grained understanding of the cell environment during fermentation.

In future work, the algorithm could be enhanced to better leverage real-world data, potentially improving its predictive accuracy and practical utility. For instance, integrating soft sensors or other data sources could refine their applicability to real-world scenarios. The data could also create hybrid-CFD models which could lead to a 3-D validated CFD in-situ as it would use real data to define some of the simulation parameters, as it has started to be done in homogeneous lab-scale reactor simulations (Bangi et al., 2022).

Additionally, incorporating a complex metabolic model within the CFD simulations could deepen our understanding of the interplay between reactor conditions and cellular metabolism, which is critical for optimizing bioprocess outcomes. Further exploration could also focus on the dynamic aspect of the bioreactor checking how the compartments change over a fed-batch run.

Bioreactors represent a fusion of biological complexity and fluid dynamics, resulting in an intricate environment that poses challenges for replication across different scales and detailed modeling (i.e., for use in digital twins for control purposes). This algorithm constitutes a novel tool which may help to facilitate the identification of regimes across multiple scales by integrating both biological and physical insights.

Author contribution

For transparency, we require corresponding authors to provide co-author contributions to the manuscript using the relevant CRediT roles. The CRediT taxonomy includes 14 different roles describing each contributor's specific contribution to the scholarly output. The roles are: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Roles/Writing - original draft; and Writing - review & editing. Note that not all roles may apply to every manuscript, and authors may have contributed through multiple roles. More details and an example.

Term	Definition
Conceptualization	Ideas; formulation or evolution of overarching research goals and aims
Methodology	Development or design of methodology; creation of models
Software	Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components
Validation	Verification, whether as a part of the activity or separate, of the overall replication/ reproducibility of results/ experiments and other research outputs
Formal analysis	Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data
Investigation	Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection
Resources	Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools
Data Curation	Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse
Writing - Original Draft	Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation)
Writing - Review & Editing	Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre-or post-publication stages
Visualization	Preparation, creation and/or presentation of the published work, specifically visualization/ data presentation
Supervision	Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team
Project Administration	Management and coordination responsibility for the research activity planning and execution
Funding acquisition	Acquisition of the financial support for the project leading to this publication

CRediT authorship contribution statement

Víctor Puig I Laborda: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Lars Puiman:** Writing – review & editing, Validation, Resources. **Teddy Groves:** Writing – review & editing, Conceptualization. **Cees Haringa:** Writing – review & editing, Supervision. **Lars Keld Nielsen:** Writing – review & editing, Resources, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Víctor Puig I Laborda reports financial support was provided by Technical University of Denmark. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements / Funding

This work was supported by the Novo Nordisk Foundation (NNF20CC0035580 and NNF14OC0009473) within the framework of the Fermentation-based Biomanufacturing Initiative (FBM), grant number: NNF17SA0031362.

Lars Puiman contributed as part of the MicroSynC research program (project number P16–10/5) and is (partly) financed by the Netherlands Organization for Scientific Research (NWO).

Appendices

A.1. Individual Silhouette Score

As the Silhouette score can be computed individually for each data point, we can plot them directly to the continuous body to see which datapoints are the best classified (Fig. Appendix 1).

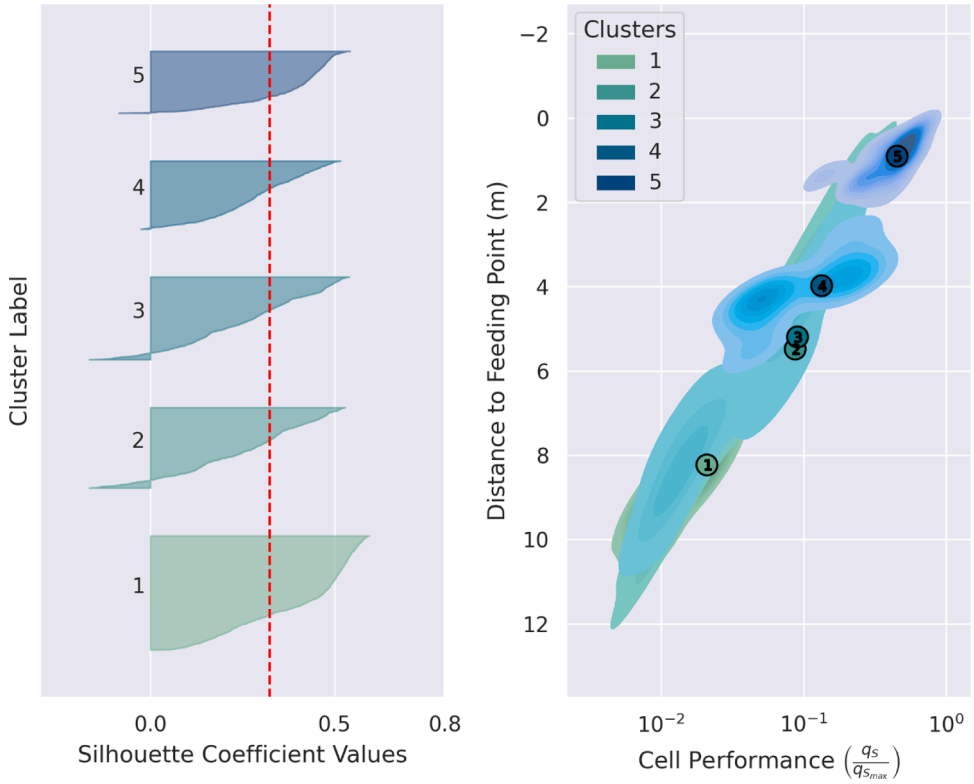

Fig. Appendix 1. Silhouette profile of each data point belonging to each cluster and spatial distribution of the clusters in a KDE plot. Distinct centroids of clusters are marked by encircled numbers. The average silhouette score for the whole analysis is set as the red dashed line.

As seen in the Figure, clusters 1 and 5 are the best classified with most of the datapoints being over the average silhouette score, and the centroids being very separated from the other ones. This can be seen as the spatial distribution of the silhouette score in the 3-D CFD simulation body (Fig. Appendix 2).

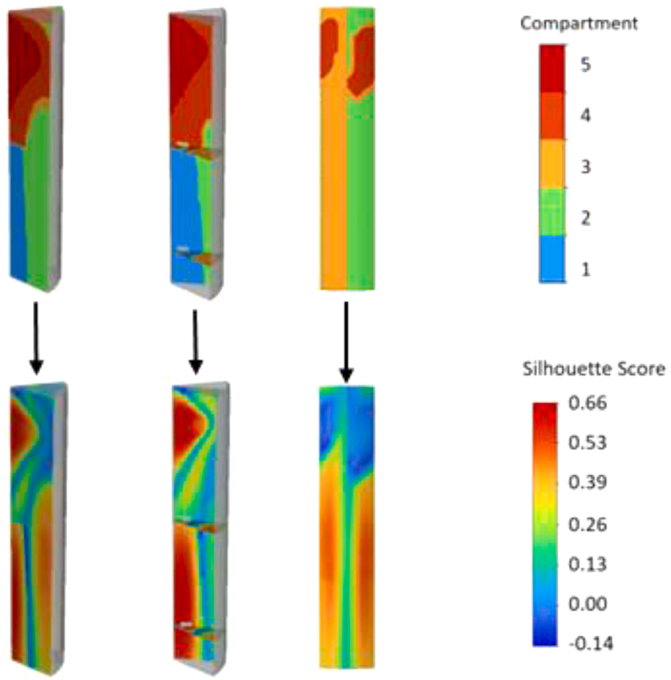

Fig. Appendix 2. Silhouette scores for each individual data point represented in each mesh element in the 3-D geometry body.

As seen in Fig. Appendix 2, provides a 3-D view of the silhouette scores, highlighting lower silhouette scores at cluster borders, suggesting possible classification overlaps. Clusters 1 and 5 exhibit high accuracy, likely representing overflow and starvation states.

Data availability

Data will be made available on request.

References

Arthur, D., Vassilvitskii, S., 2007. K-means++ the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035.

Bach, C., Yang, J., Larsson, H., Stocks, S.M., Gernaey, K.V, Albaek, M.O., Krühne, U., 2017. Evaluation of Mixing and Mass Transfer in a Stirred Pilot Scale Bioreactor Utilizing CFD. Chemical Engineering Science 171, 19–26. https://doi.org/10.1016/j.ces.2017.05.001.

Bangi, M. S. F., Kao, K., & Kwon, J. S.-I. (2022). *Physics-informed neural networks for hybrid modeling of lab-scale batch fermentation for -carotene production using Saccharomyces cerevisiae.*

Bezzo, F., Macchietto, S., 2004. A General Methodology for Hybrid Multizonal/CFD Models: Part II. Automatic Zoning Computers and Chemical Engineering 28 (4), 513–525. https://doi.org/10.1016/j.compchemeng.2003.08.010.

Bezzo, F., Macchietto, S., Pantelides, C.C., 2004. A General Methodology for Hybrid Multizonal/CFD Models: Part I. Theoretical Framework. Computers and Chemical Engineering 28 (4), 501–511. https://doi.org/10.1016/j.compchemeng.2003.08.004.

Brannock, M., Wang, Y., Leslie, G., 2010. Mixing characterisation of full-scale membrane bioreactors: CFD modelling with experimental validation. Water Research 44 (10), 3181–3191.

Coroneo, M., Montante, G., Paglianti, A., Magelli, F., 2011. CFD prediction of fluid flow and mixing in stirred tanks: Numerical issues about the RANS simulations. Computers & Chemical Engineering 35 (10), 1959–1968.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 2, 224–227.

Delafosse, A., Calvo, S., Collignon, M.-L., Delvigne, F., Crine, M., Toye, D., 2015. Euler–Lagrange approach to model heterogeneities in stirred tank bioreactors–Comparison to experimental flow characterization and particle tracking. Chemical Engineering Science 134, 457–466.

Delafosse, A., Collignon, M.-L., Calvo, S., Delvigne, F., Crine, M., Thonart, P., Toye, D., 2014. CFD-based Compartment Model for Description of Mixing in Bioreactors. Chemical Engineering Science 106, 76–85. https://doi.org/10.1016/j.ces.2013.11.033.

Enfors, S.-O., Jahic, M., Rozkov, A., Xu, B., Hecker, M., Jürgen, B., Krüger, E., Schweder, T., Hamer, G., O'Beirne, D., Noisommit-Rizzi, N., Reuss, M., Boone, L., Hewitt, C., McFarlane, C., Nienow, A., Kovacs, T., Trägårdh, C., Fuchs, L., Manelius, Å., 2001. Physiological Responses to Mixing in Large Scale Bioreactors. Journal of Biotechnology 85 (2), 175–185. https://doi.org/10.1016/S0168-1656(00)00365-5.

Fooladgar, E., Duwig, C., 2018. A new post-processing technique for analyzing high-dimensional combustion data. Combustion and Flame 191, 226–238. https://doi.org/10.1016/j.combustflame.2018.01.014.

Gunyol, O., Mudde, R.F., 2009. Computational study of hydrodynamics of a standard stirred tank reactor and a large-scale multi-impeller fermenter. International Journal for Multiscale Computational Engineering 7 (6).

Haringa, C., Deshmukh, A.T., Mudde, R.F., Noorman, H.J., 2017. Euler-Lagrange analysis towards representative down-scaling of a 22 m3 aerobic S. cerevisiae fermentation. Chemical Engineering Science 170, 653–669. https://doi.org/10.1016/j.ces.2017.01.014.

Haringa, C., Tang, W., Deshmukh, A.T., Xia, J., Reuss, M., Heijnen, J.J., Mudde, R.F., Noorman, H.J., 2016. Euler-Lagrange Computational Fluid Dynamics for (Bio) Reactor Scale down: An Analysis of Organism Lifelines. Engineering in Life Sciences 16 (7), 652–663. https://doi.org/10.1002/elsc.201600061.

Haringa, C., Tang, W., Noorman, H.J., 2022. Stochastic parcel tracking in an Euler–Lagrange compartment model for fast simulation of fermentation processes. Biotechnology and Bioengineering 119 (7), 1849–1860.

Haringa, C., Vandewijer, R., Mudde, R.F., 2018. Inter-compartment interaction in multi-impeller mixing. Part II. Experiments, sliding mesh and large Eddy simulations. Chemical Engineering Research and Design 136, 886–899.

Jiang, Y., Yin, S., Dong, J., Kaynak, O., 2020. A review on soft sensors for monitoring, control, and optimization of industrial processes. IEEE Sensors Journal 21 (11), 12868–12881.

Knysh, P., Mann, R., 1984. Utility of Networks of Interconnected Backmixed Zones to Represent Mixing in a Closed Stirred Vessel. *FLUIDMIXINGII SYMP., (BRADFORD,U.K.: APR. 3-5, 1984), RUGBY, U.K., INST. CHEM. ENGRS., 1984, P.127-145. (ICHEME SYMP., 89) (ISBN 0-85295-171X)*.

Le Moullec, Y., Gentric, C., Potier, O., Leclerc, J.P, 2010. Comparison of Systemic, Compartmental and CFD Modelling Approaches: Application to the Simulation of a Biological Reactor of Wastewater Treatment. Chemical Engineering Science 65 (1), 343–350. https://doi.org/10.1016/j.ces.2009.06.035.

Le Nepvou De Carfort, J., Pinto, T., Krühne, U, 2024. An Automatic Method for Generation of CFD-Based 3D Compartment Models: Towards Real-Time Mixing Simulations. Bioengineering (2), 11. https://doi.org/10.3390/bioengineering11020169.

Lin, H.Y., Mathiszik, B., Xu, B., Enfors, S.-O., Neubauer, P., 2001. Determination of the maximum specific uptake capacities for glucose and oxygen in glucose-limited fed-batch cultivations of Escherichia coli. Biotechnology and Bioengineering 73 (5), 347–357.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Transactions on Information Theory 28 (2), 129–137.

Mann, R., & Mavros, P. (1982). *Analysis of Unsteady Tracer Dispersion and Mixing in a Stirred Vessel Using Interconnected Networks of Ideal Flow Zones.*

Nadal-Rey, G., Kavanagh, J.M., Cassells, B., Cornelissen, S., Fletcher, D.F., Gernaey, K.V, McClure, D.D., 2023. Modelling of industrial-scale bioreactors using the particle lifeline approach. Biochemical Engineering Journal, 108989.

Nadal-Rey, G., McClure, D.D., Kavanagh, J.M., Cassells, B., Cornelissen, S., Fletcher, D.F., Gernaey, K.V., 2021. Development of Dynamic Compartment Models for Industrial Aerobic Fed-Batch Fermentation Processes. Chemical Engineering Journal 420. https://doi.org/10.1016/j.cej.2021.130402.

Nauha, E.K., Kálal, Z., Ali, J.M., Alopaeus, V., 2018. Compartmental Modeling of Large Stirred Tank Bioreactors with High Gas Volume Fractions. Chemical Engineering Journal 334, 2319–2334. https://doi.org/10.1016/j.cej.2017.11.182.

Nørregaard, A., Bach, C., Krühne, U., Borgbjerg, U., Gernaey, K.V., 2019. Hypothesis-Driven Compartment Model for Stirred Bioreactors Utilizing Computational Fluid Dynamics and Multiple pH Sensors. Chemical Engineering Journal 356, 161–169. https://doi.org/10.1016/j.cej.2018.08.191.

Pakhira, M.K., 2014. A linear time-complexity k-means algorithm using cluster shifting. In: 2014International Conference on Computational Intelligence and Communication Networks, pp. 1047–1051.

Perini, F., Krishnasamy, A., Ra, Y., Reitz, R.D., 2014. Computationally efficient simulation of multicomponent fuel combustion using a sparse analytical jacobian chemistry solver and high-dimensional clustering. Journal of Engineering for Gas Turbines and Power 136 (9). https://doi.org/10.1115/1.4027280.

Pigou, M., Morchain, J., 2015. Investigating the interactions between physical and biological heterogeneities in bioreactors using compartment, population balance and metabolic models. Chemical Engineering Science 126, 267–282.

Pinto, J., Mestre, M., Ramos, J., Costa, R.S., Striedner, G., Oliveira, R., 2022. A general deep hybrid model for bioreactor systems: Combining first principles with deep neural networks. Computers & Chemical Engineering 165, 107952.

Promma, I., Aucoin, M.G., Abukhdeir, N.M., Budman, H., 2024. A coupled metabolic flux/compartmental hydrodynamic model for large-scale aerated bioreactors. Computers & Chemical Engineering 189, 108806.

Puiman, L., Abrahamson, B., van der Lans, R.G.J.M., Haringa, C., Noorman, H.J., Picioreanu, C., 2022. Alleviating mass transfer limitations in industrial external-loop syngas-to-ethanol fermentation. Chemical Engineering Science 259, 117770.

Rigopoulos, S., Jones, A., 2003. A Hybrid CFD-reaction Engineering Framework for Multiphase Reactor Modelling: Basic Concept and Application to Bubble Column Reactors. Chemical Engineering Science 58 (14), 3077–3089. https://doi.org/10.1016/S0009-2509(03)00179-9.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65.

Rovira, M., Engvall, K., Duwig, C., 2022. Identifying key features in reactive flows: A tutorial on combining dimensionality reduction, unsupervised clustering, and feature correlation. Chemical Engineering Journal 438. https://doi.org/10.1016/j.cej.2022.135250.

Savarese, M., Cuoci, A., De Paepe, W., Parente, A., 2023. Machine learning clustering algorithms for the automatic generation of chemical reactor networks from CFD simulations. Fuel 343. https://doi.org/10.1016/j.fuel.2023.127945.

Scott, D.W., 2015. Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons.

Shah, P., Sheriff, M.Z., Bangi, M.S.F., Kravaris, C., Kwon, J.S.-I., Botre, C., Hirota, J., 2022. Deep neural network-based hybrid modeling and experimental validation for an industry-scale fermentation process: Identification of time-varying dependencies among parameters. Chemical Engineering Journal 441, 135643.

Sharma, C., Malhotra, D., Rathore, A.S., 2011. Review of Computational Fluid Dynamics Applications in Biotechnology Processes. Biotechnology Progress 27 (6), 1497–1510. https://doi.org/10.1002/btpr.689.

Tajsoleiman, T., Spann, R., Bach, C., Gernaey, K.V, Huusom, J.K., Krühne, U., 2019. A CFD Based Automatic Method for Compartment Model Development. Computers and Chemical Engineering 123, 236–245. https://doi.org/10.1016/j.compchemeng.2018.12.015.

Wells, G.J., Ray, W.H., 2005. Methodology for Modeling Detailed Imperfect Mixing Effects in Complex Reactors. AIChE Journal 51 (5), 1508–1520. https://doi.org/10.1002/aic.10407.

Yu, W., Zhao, F., Yang, W., Xu, H., 2019. Integrated analysis of CFD simulation data with K-means clustering algorithm for soot formation under varied combustion conditions. Applied Thermal Engineering 153, 299–305. https://doi.org/10.1016/j.applthermaleng.2019.03.011.

Zahradník, J., Mann, R., Fialová, M., Vlaev, D., Vlaev, S.D., Lossev, V., Seichter, P., 2001. A Networks-of-Zones Analysis of Mixing and Mass Transfer in Three Industrial Bioreactors. Chemical Engineering Science 56 (2), 485–492. https://doi.org/10.1016/S0009-2509(00)00252-9.

Further reading

Nadal-Rey, G., McClure, D.D., Kavanagh, J.M., Cornelissen, S., Fletcher, D.F., Gernaey, K.V., 2021. Understanding Gradients in Industrial Bioreactors. Biotechnology Advances 46. https://doi.org/10.1016/j.biotechadv.2020.107660.