Powder Technology 413 (2023) 118042

Contents lists available at ScienceDirect

Powder Technology

journal homepage: www.journals.elsevier.com/powder-technology

# Neural network modeling of the dynamic inactivation of probiotics during single droplet drying for improved cell viability

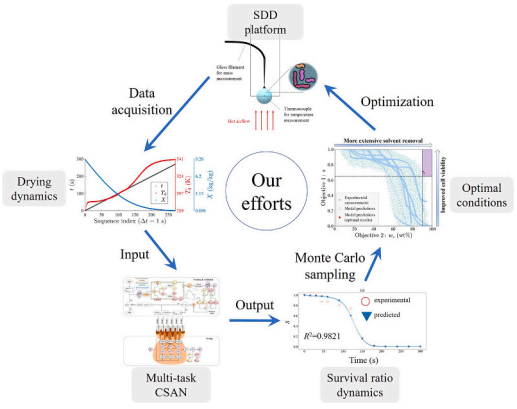
Hong Zhu, Dongbiao Jin, Nan Fu, Xiao Dong Chen, Jie Xiao *

*Suzhou Key Laboratory of Green Chemical Engineering, Particle Engineering Laboratory, School of Chemical and Environmental Engineering, College of Chemistry, Chemical Engineering and Materials Science, Soochow University, Suzhou 215123, China*

## HIGHLIGHTS

- A dynamic inactivation model for probiotics survival prediction during drying
- Effective learning of historical data by a convolutional self-attention neural network
- A double-branch architecture allows precise prediction of inactivation dynamics
- Optimized conditions identified for improved cell viability and sufficient solvent removal

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Keywords:*
Single droplet drying
Probiotic inactivation
Dynamic drying process
Neural network
Convolutional self-attention network (CSAN)

## ABSTRACT

Dairy products containing probiotics are often dried to improve shelf life and facilitate transportation. A reliable dynamic inactivation model has long been pursued to optimize the production by maximizing probiotics' survival during drying. How to take care of the dynamic drying process experienced by the cells for precise prediction of their survival remains a challenging task. In this work, a multi-task convolutional self-attention network (CSAN) has been developed for dynamic modeling of probiotics inactivation during single droplet drying (SDD). The convolution self-attention approach together with a unique double-branch architecture allows the neural network (NN) to learn effectively from historical data and predict inactivation dynamics throughout the whole drying process. In terms of prediction accuracy, our model ($R^2 > 0.96$) outperforms many other existing models ($R^2 < 0.6$ in most circumstances). By resorting to this model, two optimal SDD conditions have been identified with the resultant terminal solid contents higher than 90 wt% and cell survival ratios higher than 0.65.

* Corresponding author.
*E-mail address:* jie.xiao@suda.edu.cn (J. Xiao).

https://doi.org/10.1016/j.powtec.2022.118042
Received 14 July 2022; Received in revised form 30 September 2022; Accepted 20 October 2022
Available online 25 October 2022
0032-5910/© 2022 Elsevier B.V. All rights reserved.

## 1. Introduction

As functional ingredients, probiotics are commonly used in dairy products, including those in a powder form produced by spray drying. After being dried, they become easier for transportation and can have prolonged shelf life of up to decades. Such a practice can result in "anhydrobiosis" of probiotics. Upon rehydration, they will swell quickly and be reactivated [1]. While freeze drying is often applied to the dehydration of heat-sensitive constituents like probiotics for higher residual viability [2], hot air spray drying (hereafter referred to as spray drying) has drawn more attention than ever these years due to its cost efficiency [3]. Nevertheless, poor residual viability caused by spray drying remains a problem that limits its wide application [4]. It becomes crucial for the producer of these dairy products to seek a viable approach that can minimize damage to the cells from heat and dehydration and maximize probiotic viability at the dryer outlet.

There is an urgent need to develop a predictive model for the process of probiotic inactivation during spray drying. Tremendous efforts in multiphysics modeling of spray drying can be identified [5–8], which allowed us to track the dynamic drying process (hereafter referred to as drying history) of a large number of discrete droplets. Those models can predict particles' trajectory, size, moisture content, and temperature dynamics. Inactivation of probiotics in spray-dried droplets, however, was not considered in those models. Fu and Chen [9] summarized nine factors that can influence inactivation dynamics, including two hot air properties (i.e., temperature $T_a$ and humidity), five drying kinetics parameters (i.e., droplet temperature $T_d$, initial moisture content $X_0$, rates of droplet temperature rise and water removal, and exposure time $t$), and two carrier properties (composition and the cell location). As two main stresses, heat and dehydration inactivate the probiotics by different mechanisms: the former (reflected by $T_d$) denatures the gene that is crucial to cell reproduction [10], while the latter (reflected by $X$) can lead to increased osmotic pressure [11] and hence cell destabilization [12].

Different inactivation models have been developed, which can be generally classified into two categories (see Table S1.1 and Table S1.2 in the supplementary material). The first type is the reaction kinetics model. For over 100 years, the survival ratio of microbial cells has been found to decline exponentially with the exposure time according to a large number of experiments [13,14]. As a result, it is a common practice to model the process as a first-order reaction, and present cell inactivation in a lg$s$ vs. $t$ plot. The general form of the kinetics model is:

$$-\frac{\mathrm{d}s_t}{\mathrm{d}t} = k_{\mathrm{d},t}t \tag{1}$$

where $k_d$ denotes the inactivation rate constant and the subscript $t$ refers to a time instant $t$. Rahn [10] claimed that the first-order inactivation kinetics should result from the denaturation of the gene responsible for reproduction. In most cases, $k_{d,t}$ is correlated to the surroundings (like the ambient temperature) by the Arrhenius' equation, and is considered time-independent in the simplest model of this type [15] (Model 1 in Table S1.1). This model, however, was criticized to be an "exception rather than the rule" [16] owing to frequent deviations from theoretical log-linearity (i.e., the inactivation rate constant $k_{d,t}$ stays unchangeable) [17,18]. Such deviations are commonly shown in the form of a plateau near the start or the end of the lg$s$-$t$ curve, which is named the "shoulder" or "tailing" respectively (see Fig. 1). A few modifications have been made to correlate $k_{d,t}$ with drying kinetics (Models 2–9) [19–23], whereas the tailing is still hard to capture.

The second type is the probabilistic model. The death of viable cells is viewed as probabilistic rather than deterministic events in these models, e.g., the Weibull model (Model 10) [16], the logistic model (Model 11) [24], and the modified Gompertz model (Model 12) [25], where shoulder and tailing were well predicted.

The log scale and linear scale plots have their unique advantages in presenting a survival dynamics curve. As shown in Fig. 1, survival ratio values close to zero at the final stage of drying are differentiable on a log scale plot, which is not the case for the linear scale plot. Nevertheless, quite a few diagrams adopted linear scale plots [20,21], which can better capture cell inactivation in the initial stage of drying. Unfortunately, few models can fulfil the two tasks well simultaneously, namely predicting accurately on both the linear scale and the log scale.

Moreover, Marechal et al. [26] proved experimentally that the dynamic heating process could have prominent influence on bacterial survival. Hence, State$_t$ (described by $k_{d,t}$ or the survival rate $s_t$) should be a cumulative effect of the drying history. Specifically, the dynamic values of $x$ from time zero to $t$ should have influences on the current state to different degrees, i.e.,

$$\mathrm{State}_t = f(x_0, x_1, \cdots, x_t) \tag{2}$$

where $x_t$ can be $T_{d,t}$, $X_t$, and $pH_t$, etc.

The artificial neural network (ANN), as a machine learning technology, has demonstrated excellent performance in recent years in various research areas including chemical engineering [27]. With a large number of neurons, the ANN can approximate any arbitrarily complex function with any desired precision [28,29]. Additionally, it can extract information and learn knowledge from the given data automatically. The multilayer perceptron (MLP) [30] is the earliest and simplest ANN. MLP has been introduced into predictive microbiology since the last century for the prediction of microbial growth [31], where states at different time instants were merely regarded as independent samples. Few efforts can be identified to take into account the historical influence and the relations between different states. Over the past decades, more ANN architectures have been developed for time series

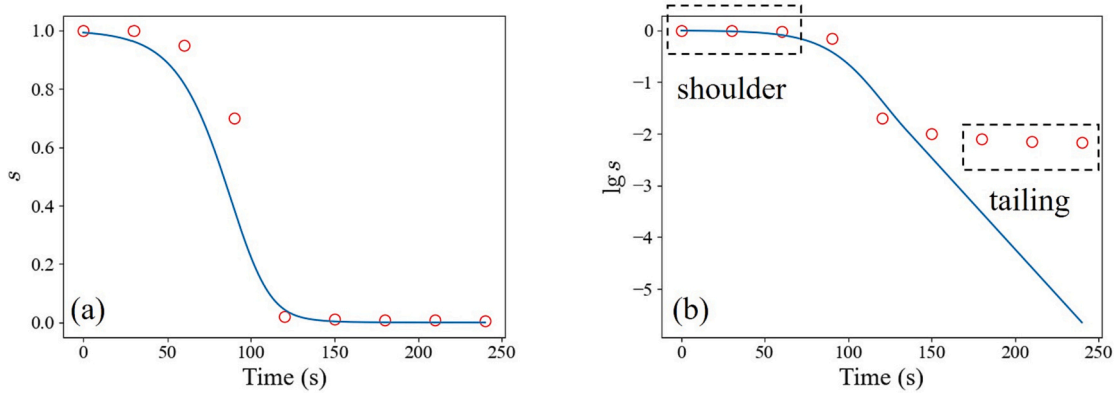



**Fig. 1.** Comparison between the diagram of the survival ratios versus time (a) on the linear scale and (b) the log scale. The lines in two subplots correspond to one data set.

prediction, e.g., time-delay neural network (TDNN) [32], which is nearly equivalent to one-dimensional convolutional neural network (1-D CNN) and the simple recurrent neural network (S-RNN) [33]. As a variant of S-RNN, the long short-term memory network (LSTM) [34] has been widely applied to modeling chemical engineering processes. Kim et al. [35] employed LSTM to locate the leakage source of a chemical plant. Fang et al. [36] used the same method to predict the moisture content of biomass during drying.

Transformer [37] has shown excellent performance and overshadowed all the other modules in computer vision and natural language processing since its invention in 2017. Compared with LSTM, Transformer with self-attention is much more powerful and efficient in learning arbitrarily long-term dependencies and making predictions at different time instants in parallel. Li et al. [38] improved Transformer by introducing convolution and developed the so-called convolutional self-attention network (CSAN). This new network can focus more on the short-term dependencies (through convolution) while learning the long-term ones (through self-attention) simultaneously. Very few ANN-based models can be identified for dynamic modeling of probiotics inactivation during spray drying. CSAN is a promising network for modeling this dynamic process.

Spray drying is often regarded as a black box process since it is difficult to measure online droplets' dynamic information from inlet to outlet through experiments. Even if $T_d$ and $X$ can be estimated by a reliable computational fluid dynamics (CFD) model, it is impossible to count the viable cells in real time. Moreover, the drying media conditions vary along the droplets' flying trajectories, and the drying time of droplets cannot be uniformly and precisely controlled [39,40]. In this work, to obtain dynamic experimental data, we chose to start with the single droplet drying (SDD) experiment [41]. It is a well-developed technique and allows us to measure $T_d$, $X$, and $s$ throughout the drying process of a single droplet suspended on a glass filament. For the SDD setup, drying conditions can be well controlled, including drying air temperature, air velocity, air humidity, chamber pressure, and the total drying time.

In this work, a CSAN model will be developed for dynamic modeling of probiotic inactivation during the whole process of SDD. The influence of drying history on probiotics survival will be fully taken into account. Moreover, our model will be compared with the existing ones to demonstrate its performance. Finally, such a model will be applied to the identification of optimal drying conditions that can offer enhanced cell viability with sufficient solvent removal.

## 2. Raw data acquisition and preprocessing

### 2.1. Data acquisition

Single droplet drying (SDD) experiments were carried out on the Dong Concept SDD platform [41] to obtain the dynamic data. As shown in Fig. 2, a droplet is suspended by a glass filament with hot air flowing upwards. The following data were obtained. Droplet temperature $T_d$ was monitored in situ by a thermocouple. Moisture content $X$ was calculated based on the droplet mass change measured through the displacement of the filament. Since counting viable cells is difficult and laborious, the viable cell count $N$ is acquired by counting the colony forming unit (CFU) on the plate at a certain time interval. The survival ratio $s$ was then calculated by dividing $N$ by the initial cell count $N_0$. Therefore, the dynamic measurement data of $T_d$ and $X$ are dense and available every second during the drying process (see Fig. 3(a1)), but the data of $s$ are sparse and only available at certain time instants (see Fig. 3(b1)). For more experimental details, please refer to Fu et al. [19].

The *Lactococcus lactis* ssp. *cremoris* (LLC) and *Lactobacillus rhamnosus* GG (LGG) were selected as experimental subjects with reconstituted skim milk as the carrier. The data set consists of data from nine groups of samples with four dynamic quantities, namely time ($t$), the droplet temperature ($T_d$), the average moisture content on the dry basis ($X$) and the survival ratio ($s$) during the process of SDD lasting for 120 to 400 s. Among them, G1-G6 are retrieved from Fu et al. [19], and G7-G9 from Mao et al. [42]. Experimental conditions are listed in Table 1. The air pressure was kept at 101325 Pa.

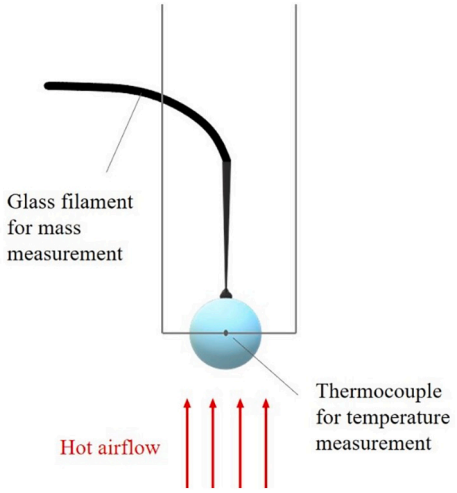



**Fig. 2.** A schematic diagram of the drying chamber of a SDD rig.

### 2.2. Data preprocessing

In this part, a notation with a tilde symbol "~" indicates the preprocessed quantity.

#### 2.2.1. Data set split

The data are split into two groups respectively for training and testing. With different environmental temperatures and solid contents, G3 does not show the tailing but G5 does. These two representative groups are included in the testing data. Also, to inspect the generalization performance of our model over different probiotic strains, G9 (carried out for LGG) is selected as the testing data, with the remaining six to form the training-validation data, i.e.,

$$\begin{aligned} \mathscr{D}_{\text{test}} &= \{\text{G3},\text{G5},\text{G9}\},\\ \mathscr{D}_{\text{train+vali}} &= \{\text{G1},\text{G2},\text{G4},\text{G6},\text{G7},\text{G8}\} \end{aligned} \tag{3}$$

#### 2.2.2. Standardization of inputs

Among four types of dynamic quantities, the first three, i.e., $t$, $T_d$, $X$, form the input data; while the remainder, i.e., the survival ratio $s$, forms the ground truth (see discussions in Section 2.2.3). It is believed that our ANN can calculate the gradients implicitly with given $T_d$ history, $X$ history, and sampling time instants if the rates are of any significance.

To facilitate convergence of simulation, z-score standardization is performed on $t$, $T_d$, $X$ respectively, giving dimensionless $\widetilde{t},\widetilde{T}_d,\widetilde{X}$,

$$\widetilde{x} = \frac{x-\mu_{\text{train+vali}}}{\sigma_{\text{train+vali}}}, \quad x: t \text{ or } T_d \text{ or } X \tag{4}$$

where the mean ($\mu_{\text{train+vali}}$) and standard deviation ($\sigma_{\text{train+vali}}$) of a certain variable $x$ are calculated with all $x$ data in $\mathscr{D}_{\text{train+vali}}$.

#### 2.2.3. Creation of labels

In our study, the "ground truth" denotes the experimental survival ratio. The interpolation over the ground truth will be discussed soon, after which the "label" will denotes both the experimental and the interpolated survival ratios. The labels will be used for the NN training, validation and test.

As to the ground truth $s^{\text{grd}}$, some abnormal results exist, in which LLC seems to "revive" and $s^{\text{grd}}$ is greater than one at some time instants. It may be attributed to either cell breakup or measurement error.

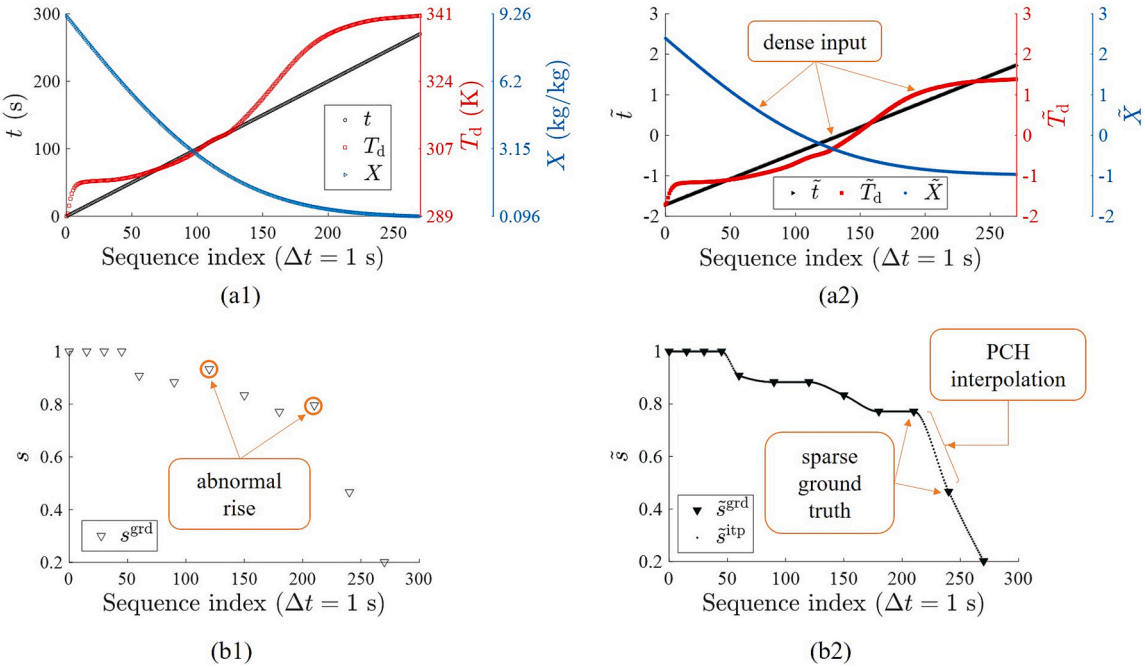



**Fig. 3.** The first group (G1) of experimental data before and after preprocessing: (a1) original input; (a2) preprocessed input; (b1) original ground truth; (b2) preprocessed labels.

**Table 1**
Experimental conditions of nine groups of samples.

| Index | Strain | Hot air temperature (°C) | Hot air velocity (m/s) | Initial solid content (wt%) | Initial droplet volume (μL) | Duration (s) | Absolute humidity (kg/kg) | Sampling number of survival ratio |
|---|---|---|---|---|---|---|---|---|
| G1 | LLC | 70 | 0.75 | 10 | 2 | 270 | 0.0001 | 12 |
| G2 | LLC | 90 | 0.75 | 10 | 2 | 270 | 0.0001 | 12 |
| G3 | LLC | 110 | 0.75 | 10 | 2 | 120 | 0.0001 | 7 |
| G4 | LLC | 70 | 0.75 | 20 | 2 | 300 | 0.0001 | 13 |
| G5 | LLC | 90 | 0.75 | 20 | 2 | 300 | 0.0001 | 13 |
| G6 | LLC | 110 | 0.75 | 20 | 2 | 120 | 0.0001 | 7 |
| G7 | LLC | 70 | 0.45 | 20 | 1 | 400 | 0.0005 | 10 |
| G8 | LLC | 70 | 0.45 | 20 | 2 | 400 | 0.0005 | 10 |
| G9 | LGG | 70 | 1.00 | 20 | 2 | 360 | 0.0005 | 11 |

Abnormal fluctuations will confuse our ANN. Theoretically, $s$ is by definition to be non-negative, and it is reasonable to assume that probiotics do not propagate, let alone revive after inactivation. Accordingly, $s^{\mathrm{grd}}$ is first clamped to ensure $\widetilde{s}^{\mathrm{grd}} \geq 10^{-7}$. Note that a sufficiently small number $10^{-7}$ is used instead of zero because we want to know how inactivation goes on in a lg$s$-$t$ graph, and lg$s$ requires $s$ to be positive. Another constraint is that $\widetilde{s}^{\mathrm{grd}}$ should be monotonically decreasing. If $\widetilde{s}^{\mathrm{grd}}_{t[i]}$ is used to denote the preprocessed ground truth at the $i$-th sampling point, it follows that $\widetilde{s}^{\mathrm{grd}}_{t[0]} = 1$, and for each $\widetilde{s}^{\mathrm{grd}}_{t[i]}$ $(i \geq 1)$,

$$\widetilde{s}^{\mathrm{grd}}_{t[i]} = \begin{cases} \widetilde{s}^{\mathrm{grd}}_{t[i-1]}, & \text{if } s^{\mathrm{grd}}_{t[i]} > \widetilde{s}^{\mathrm{grd}}_{t[i-1]}; \\ max\left\{s^{\mathrm{grd}}_{t[i]}, 10^{-7}\right\}, & \text{else.} \end{cases} \tag{5}$$

According to our description in Section 2.1, the input (every second) is much denser than the ground truth (every >15 s). With hundreds of thousands of neurons, our ANN is inclined to overfit the data with only 95 (the sum of sampling numbers in Table 1) samples, which requires the data set to be augmented somehow. As described above, $s$ has been assumed to be monotonically decreasing with time. It inspired us to continualize $s^{\mathrm{grd}}$ by a monotonic interpolation. The Piecewise Cubic Hermite (PCH) interpolation can achieve monotonicity and smoothness simultaneously. So in this study, we chose to fill the preprocessed ground truth at each second by PCH interpolation. Then 95 original labels (ground truth) $s^{\mathrm{grd}}$ and 2454 interpolated labels $\widetilde{s}^{\mathrm{itp}}$ together formed the label set. G1 is used as an example to show our preprocessed data in Fig. 3 (a2 and b2). For convenience, the tilde symbol "~" of $\widetilde{s}$ will be omitted henceforth.

## 3. CSAN model development

### 3.1. Capturing historical information

The self-attention mechanism is depicted graphically in Fig. 4. As discussed in Introduction, we want to know the relation between two states at two arbitrary time points, e.g., State$_i$ and State$_j$. This can be achieved by the matrices $Q$ and $K$. For the ease of understanding, they can be simply viewed as a "query" base and a "key" base respectively. Let $l$ be the duration of drying. For each $i$ from 0 to $l$, State$_i$ suggests a query $\mathbf{q}_i$ to each State$_j$ $(0 \leq j \leq l)$ on how it is influenced by State$_j$, and State$_j$ replies with a key $\mathbf{k}_j$. Mathematically, our ANN just needs to calculate the inner product $\mathbf{q}_i \cdot \mathbf{k}_j$. In this way, the correlation between two arbitrary states can be acquired. The model can make multi-step time series prediction in parallel by matrix multiplication (abbreviated as "matmul" in Fig. 4) $QK^T$. At the same time, it can avoid the problem of vanishing gradient due to long-term dependencies. The side effect,

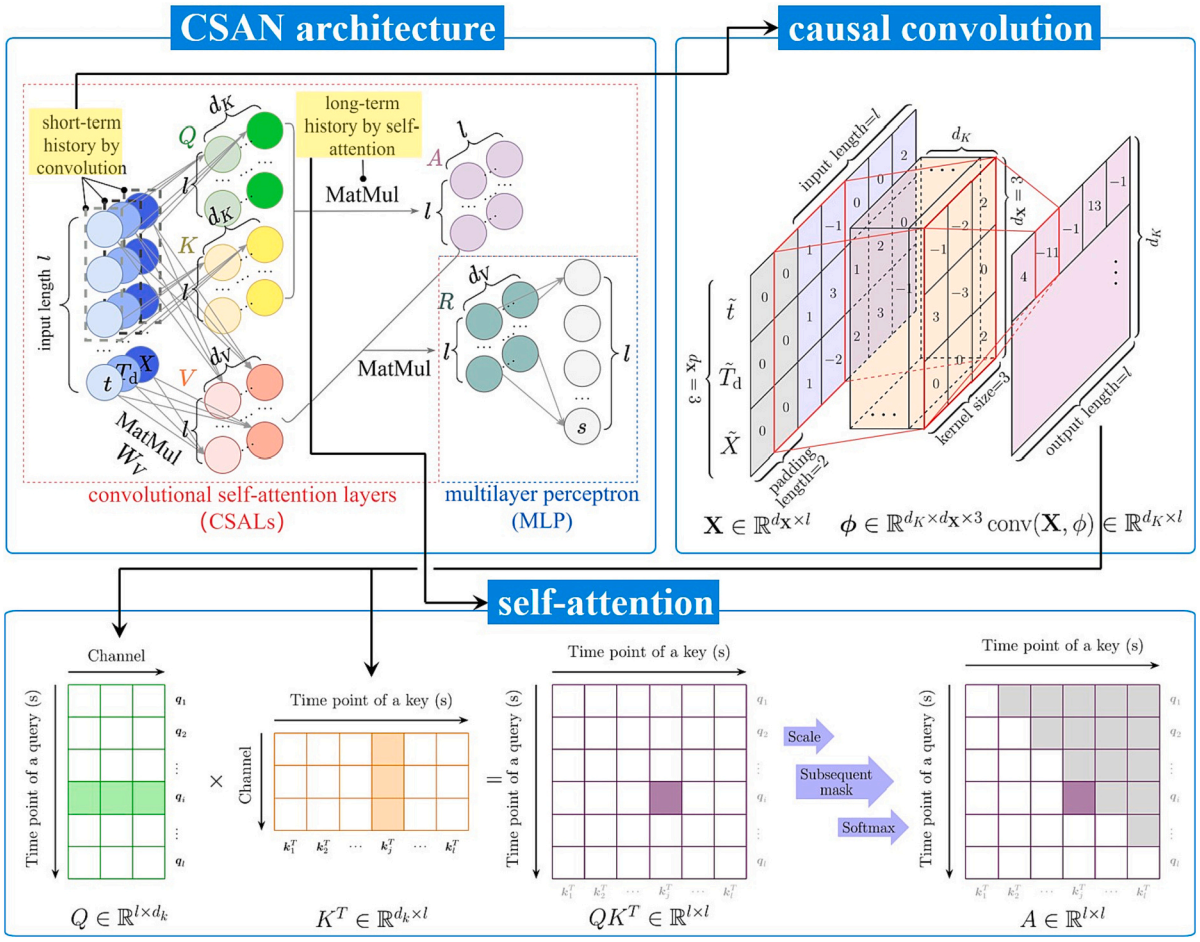



**Fig. 4.** The working principle of CSAN.

however, is that the influence of the future will also be considered. To block the influence of the future, a subsequent mask is added, or rather, off-diagonal elements of the upper right of the scaled $QK^T$ are set to zeros, resulting finally in a lower triangular attention matrix $A$.

Then, canonical self-attention generates $Q$ and $K$ by affine transformation

$$Q = W^Q\mathbf{X} + \mathbf{b}^Q, \quad K = W^K\mathbf{X} + \mathbf{b}^K \tag{6}$$

where $\mathbf{X}$ denotes the input matrix, $W^Q$ and $W^K$ the weight matrices, $\mathbf{b}^Q$ and $\mathbf{b}^K$ the bias vectors. Li et al. [38] introduced causal convolution to the self-attention module to enhance the significance of the local context. As shown in Fig. 4, there are $d_{\mathbf{X}}$ (i.e., 3 in this case corresponding to the three input variables) input channels and $d_K$ output channels, and the shape of the kernel is $d_K \times d_{\mathbf{X}} \times 3$, where the number 3 is called the "kernel size". For each output channel c ($1 \le c \le d_K$), the corresponding convolutional kernel $\phi_{c,\star,\star}$ scans from the start to the end over the input vector $\mathbf{X}$ to generate the corresponding output $Q_{c,\star}$ or $K_{c,\star}$, where the subscript $\star$ denotes all the elements of one dimension. To obtain the $Q$ and $K$ elements mathematically, for each c and each time instant $i$, carry out Hadamard product (aka point-wise product) of the kernel $\phi_{c,\star,\star}$ and the input submatrix $\mathbf{X}_{\star,i-1:i+1}$ containing the $i-1$-th to $i+1$-th elements of $\mathbf{X}$ and then sum up, i.e.,

$$\begin{aligned} Q_{c,i} &= \mathrm{conv}\left(\mathbf{X}, \phi^Q_{c,\star,\star}\right)_i + \mathbf{b}^Q_c = \mathbf{b}^Q_c + \sum \mathbf{X}_{\star,i-1:i+1} \odot \phi^Q_{c,\star,\star} \\ K_{c,i} &= \mathrm{conv}\left(\mathbf{X}, \phi^K_{c,\star,\star}\right)_i + \mathbf{b}^K_c = \mathbf{b}^K_c + \sum \mathbf{X}_{\star,i-1:i+1} \odot \phi^K_{c,\star,\star} \end{aligned} \tag{7}$$

where $\odot$ denotes the Hadamard product. Because such an operation with a kernel of size 3 will reduce the sequence length from $l$ to $l-2$, the input is extended in advance at the start with zero paddings of length two to maintain the sequence length, which is called "causal convolution" [43].

Generally, compared with self-attention attending to two arbitrary positions, causal convolution can pay more attention to the locality, namely the previous two states $\mathrm{State}_{i-1}$ and $\mathrm{State}_{i-2}$ in our study. Therefore, the CSAN architecture is able to capture both long-term and short-term historical information.

### 3.2. Multi-task prediction on the linear scale and log scale

From this part, a superscript "^" will be used to denote a predicted value. As mentioned earlier, our model needs to predict well on both linear and log scales, which is a difficult task for a single CSAN. Multi-task learning with multi-label data is needed. A possible scheme is to train two scale-specific CSANs individually, one for the linear scale and the other for the log scale, and finally integrate them. Notwithstanding, now that the only difference between the two CSANs lies in the scale, there is a reasonable prospect that representations learned by the bottom layer, namely the convolutional self-attention layers (CSALs, see Fig. 4) of the two CSANs should be similar. Hence, a better scheme is to design a double-branch CSAN with shared CSALs near the input, but individual MLPs near the output (see Fig. 5), to produce linear and log output separately. In this way, our CSANs is expected to learn more general representation when computing time and costs are saved. In the double-branch model, each branch will pay close attention to the local features near the start or the end of the inactivation process. Integrating them will give us reliable prediction of the dynamics.

For simplicity here, the two branches are combined linearly to pro-

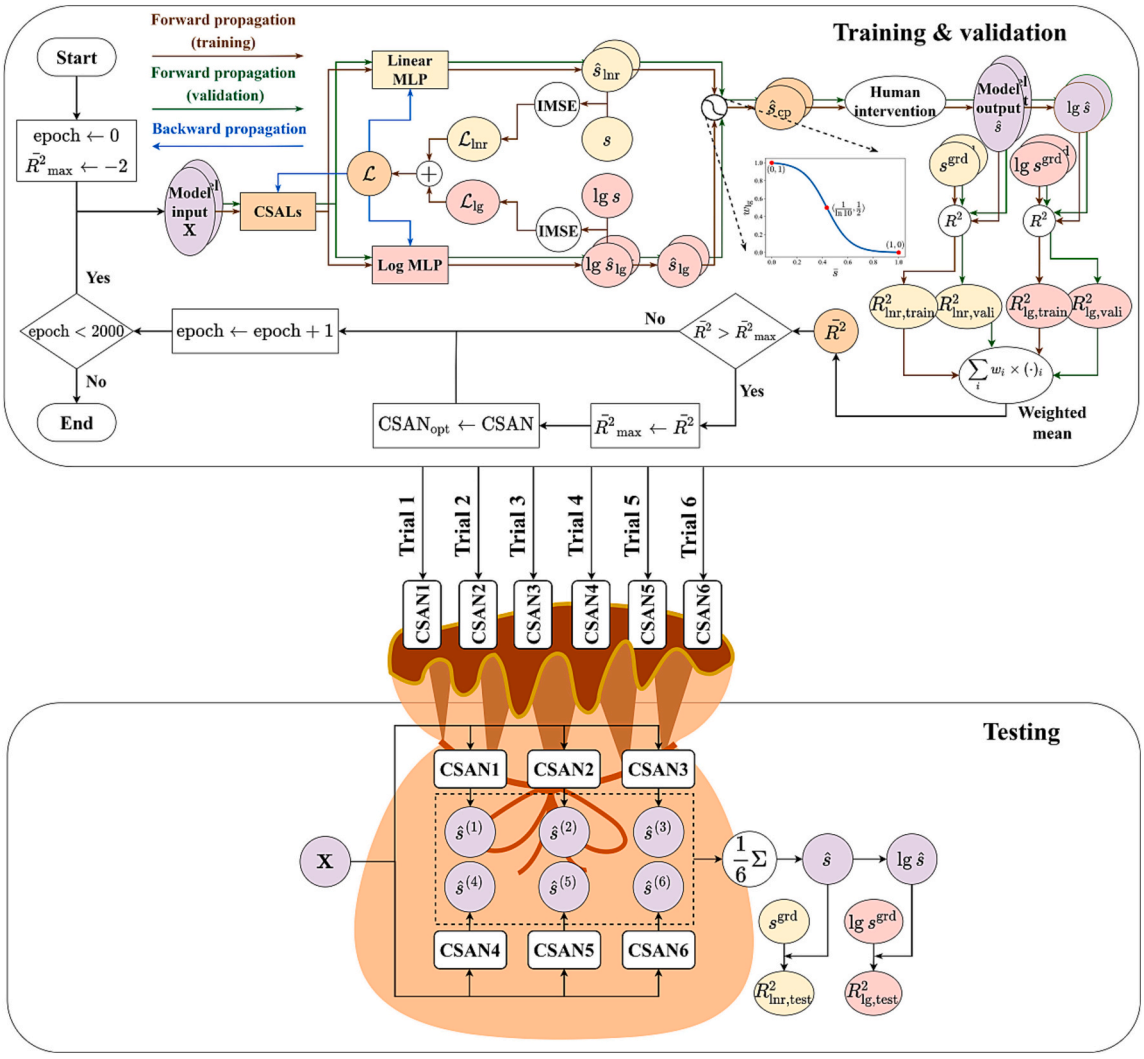



**Fig. 5.** The working flow of our double-branch CSAN from training and validation to testing.

duce the coupled output $\hat{s}_{\text{cp}}$. To measure comprehensively whether the output is closer to zero or one, the arithmetical mean $\bar{s}$ of the linear output $\hat{s}_{\text{lnr}}$ and the log output on the linear scale $\hat{s}_{\text{lg}}$ is used in Eq. (8).

$$\begin{aligned} \hat{s}_{\text{cp}} &= \left[1 - w_{\text{lg}}(\bar{s})\right] \times \hat{s}_{\text{lnr}} + w_{\text{lg}}(\bar{s}) \times \hat{s}_{\text{lg}}, \\ \bar{s} &= \frac{\hat{s}_{\text{lnr}} + \hat{s}_{\text{lg}}}{2} \end{aligned} \tag{8}$$

How to design the weight function of the log branch $w_{\text{lg}}$ depends on our confidence in each branch. Through training, it is believed that the linear branch is more reliable near the start of the process whereas the log branch is more reliable near the end, with $1/\ln 10$ as the cut-off point where they are equally reliable (see the explanation in the Section S2 in the supplementary material). Therefore, $w_{\text{lg}}$ should satisfy three conditions:

i) $w_{\text{lg}}$ monotonically decreases with the increase of the arithmetical mean $\bar{s}$;
ii) $w_{\text{lg}}$ ranges from zero to one;
iii) $1 - w_{\text{lg}}(1/\ln 10) = w_{\text{lg}}(1/\ln 10)$

A sigmoidal function is obtained according to the above constraints (see a graph in Fig. 5).

$$w_{\text{lg}}(\bar{s}) = \frac{1.008}{1 + 9.428 \times 10^{-3} exp(10.78\bar{s})} \tag{9}$$

### 3.3. Human intervention

Human-in-the-loop learning should be adopted to correct prediction errors, bringing a priori knowledge to our model. The intervention includes three steps:

Step 1. set the initial survival ratio $\hat{s}_0$ to be one;
Step 2. clamp $\hat{s}$ between $10^{-7}$ and 1;
Step 3. exert the constraint of monotonic decreasing on $\hat{s}$.

### 3.4. Training, validation and test methods

#### 3.4.1. Performance metrics

In Fig. 5, there are two places where performance metrics are needed. One is to compare the model output after intervention with the ground truth, and the other is to compare branch outputs with the labels.

We shall discuss the model output first. For one certain group of samples, R-Squared value $R^2$ is used for accuracy quantification because it eliminates the impact of dimension compared to Mean Squared Error (MSE). $R^2$ is calculated as follows in our work.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{\mathrm{grd}} - \widehat{y}_i^{\mathrm{grd}}\right)^2}{\sum_{i=1}^{n}\left(y_i^{\mathrm{grd}} - \frac{1}{n}\sum_{j=1}^{n} y_j^{\mathrm{grd}}\right)^2}, \quad y : s \text{ or } lgs \tag{10}$$

Here only the ground truth $y^{\mathrm{grd}}$ values are involved when calculating $R^2$. This is because the augmented labels $y^{\mathrm{itp}}$ are less reliable than $y^{\mathrm{grd}}$ since they are generated by interpolation.

Branch output is compared with labels to calculate the value of the loss function $\mathscr{L}$ for the ensuing backward propagation. $R^2$ is inappropriate for $\mathscr{L}$ calculation since the denominator term may lead to numerical instability during gradient descending. In comparison, MSE is more appropriate due to its simplicity, with the following general expression,

$$\mathrm{MSE}(y, \widehat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2, \quad y : s \text{ or } lgs \tag{11}$$

Note that it is not suggested to calculate $\mathscr{L}$ with the coupled output $\widehat{s}_{\mathrm{cp}}$ or the model output $\widehat{s}$ because the former after linear combination may lead to gradient exploding, and the latter after human intervention may lead to zero gradients. See Section S3 in the supplementary material for the proof.

To weaken the influence of $s^{\mathrm{itp}}$, a new loss function Interpolated MSE (IMSE) is defined by introducing an interpolation weight coefficient $w^{\mathrm{itp}}$ with the following expression,

$$\mathrm{IMSE}(y, \widehat{y}) =^{\mathrm{def}} \frac{1}{n^{\mathrm{grd}}}\sum_{i=1}^{n^{\mathrm{grd}}}\left(y_i^{\mathrm{grd}} - \widehat{y}_i^{\mathrm{grd}}\right)^2 + \frac{w^{\mathrm{itp}}}{n^{\mathrm{itp}}}\sum_{j=1}^{n^{\mathrm{itp}}}\left(y_j^{\mathrm{itp}} - \widehat{y}_j^{\mathrm{itp}}\right)^2, \quad y : s \text{ or } lgs \tag{12}$$

where the possible values of $w^{\mathrm{itp}}$ are between zero and one.

Specifically, IMSE will be reduced to MSE if and only if $w^{\mathrm{itp}} = 1$. For the two branches, let

$$\mathscr{L}_{\mathrm{lnr}} = \mathrm{IMSE}(s, \widehat{s}_{\mathrm{lnr}}), \quad \mathscr{L}_{\mathrm{lg}} = \mathrm{IMSE}\left(lgs, lg\widehat{s}_{\mathrm{lg}}\right) \tag{13}$$

Regardless of the scale difference of $\mathscr{L}_{\mathrm{lnr}}$ and $\mathscr{L}_{\mathrm{lg}}$, it was found that the total loss converged well when $\mathscr{L}_{\mathrm{lnr}}$ and $\mathscr{L}_{\mathrm{lg}}$ were summed up directly,

$$\mathscr{L} = \mathscr{L}_{\mathrm{lnr}} + \mathscr{L}_{\mathrm{lg}} \tag{14}$$

#### 3.4.2. Hyperparameter optimization

Hyperparameters (aka algorithm parameters) are those parameters that cannot be optimized by the gradient descending method [44]. In our algorithm, nine hyperparameters are to be optimized. They are the batch size $n_{\mathrm{b}}$ (1 or 2), the learning rate $\eta$ ($2 \times 10^{-4}$ to $10^{-2}$), the weight decay coefficient $\lambda$ ($10^{-6}$ to $10^{-3}$), the flood level $b_{\mathrm{f}}$ [45] ($10^{-3}$ to $5 \times 10^{-2}$), $w^{\mathrm{itp}}$ (0 to 1), the feature number of $K$, $d_K$ (30 to 150), the feature number of $V$, $d_V$ (30 to 150), the number of heads $n_h$ (1 to 6) w.r.t. multi-head self-attention and the dropout ratio $p_{\mathrm{d}}$ (0.1 to 0.7). All of them except the above-defined $w^{\mathrm{itp}}$ are commonly used in ANN training and will not be explained any further. The optimization was carried out with the help of the Optuna framework [46].

#### 3.4.3. Training and leave-one-out validation

With $\mathscr{D}_{\mathrm{test}}$ kept unseen during training, validation is necessary to prevent overfitting. $\mathscr{D}_{\mathrm{train+vali}}$ was further split into $\mathscr{D}_{\mathrm{train}}$ and $\mathscr{D}_{\mathrm{vali}}$. To make the most of the available data and get a more accurate model, the Leave-One-Out method [44] was used to acquire six trials of split training-validation data, i.e.,

**Trial 1** $\mathscr{D}_{\mathrm{vali},1} = \{\mathrm{G1}\}$, $\mathscr{D}_{\mathrm{train},1} = \{\mathrm{G2,G4,G6,G7,G8}\}$

…

**Trial *i*** $\mathscr{D}_{\mathrm{vali},i} = \{\mathrm{G}i\}$, $\mathscr{D}_{\mathrm{train},i} = \mathscr{D}_{\mathrm{train+vali}} - \mathscr{D}_{\mathrm{vali},i}$

…

**Trial 6** $\mathscr{D}_{\mathrm{vali},6} = \{\mathrm{G8}\}$, $\mathscr{D}_{\mathrm{train},6} = \{\mathrm{G1,G2,G4,G6,G7}\}$

During Trial $i$, with the optimum combination of hyperparameters, train and validate the CSAN simultaneously for 2000 epochs[1] with PyTorch [47]. As Fig. 5 shows, update the optimum CSAN with the current one once the weighted mean of the $R^2$ values on the two scales over $\mathscr{D}_{\mathrm{train},i}$ and $\mathscr{D}_{\mathrm{vali},i}$, i.e. $\sum_i w_i R_i^2$ where $i \in \{(\mathrm{lnr,train}), (\mathrm{lnr,vali}), (\mathrm{lg,train}), (\mathrm{lg,vali})\}$, exceeds its current maximum.

Six optimum models CSAN1, CSAN2, …, CSAN6 will be acquired through six trials. Multi-person decision is employed once again by taking the arithmetical mean of their predicted results as our final output, which is termed Bagging in ensemble learning [44]. The Bagging model will be used for the following testing and evaluation.

## 4. Results and discussion

### 4.1. Training and validation

After hyperparameter optimization, we trained each of the six CSANs for 2000 epochs. The Bagging CSAN predicted accurately over $\mathscr{D}_{\mathrm{train+vali}}$ on both scales and the lowest $R^2$ reached 0.9673.

### 4.2. Model testing and comparison

In this part, we will test the model by running it over $\mathscr{D}_{\mathrm{test}}$. At the same time, we shall make comparison by fitting and testing models reported in the literature (see Table S1.1 and Table S1.2 in the supplementary material) to demonstrate the outstanding performance of our Bagging CSAN model.

It can be seen from Fig. 6 that the shoulder and tailing in G3 and G5 are both well predicted with $R^2$ >0.98 on the linear scale and >0.99 on the log scale. Even for a different strain LGG which is not seen by our model before, we still get quite good $R^2$ values >0.94. All these good results demonstrate the effectiveness of our modeling methods and that our CSAN has learned some universal knowledge about inactivation beyond the LLC strain.

Other models have been adopted and compared. In terms of the kinetic ones, the required $k_{\mathrm{d}}$ should be calculated from the previous PCH interpolated survival curves to ensure consistency with the data preprocessing in Section 2.2.3. Also, the non-positive $k_{\mathrm{d}}$ values were discarded to acquire valid $\lg k_{\mathrm{d}}$ as what Fu et al. [19] did. For the remaining probabilistic ones, we first fitted each survival curve with the Weibull, logistic, or the modified Gompertz equation and then correlated the equation parameters with the drying conditions, i.e., $T_{\mathrm{a}}$, hot air velocity $v_{\mathrm{a}}$, the initial solid content of the carrier $w_{\mathrm{s},0}$ and the initial droplet volume $V_{\mathrm{d},0}$ by means of the multivariate linear regression.

The fitted parameter values are listed in Table S1.1 and Table S1.2, among which the parameter values of Model 8 are not given due to the failure of convergence. The R-Squared values are compared in Fig. 7. It can be concluded that:

i) It is also harder for other models to make reliable prediction on the log scale than on the linear one. Such a difference between scales demonstrates the need for a multi-task model.
ii) Our CSAN model outperforms the other models with the highest $R^2$ values under all four circumstances. The probabilistic models perform better than the kinetic ones in terms of $R^2_{\mathrm{test,\,lg}}$. An explanation is that most probabilistic models (except Model 10) have lower asymptotes but the kinetic ones do not, making the prediction of the former more reasonable in most cases.

Models 1–9 (kinetics models): $\lim_{t\to\infty} lgs_t = \lim_{t\to\infty} -\frac{1}{2.303}k_{\mathrm{d},t}t = -\infty$

Model 11 (probabilistic model): $\lim_{t\to\infty} lgs_t = \lim_{t\to\infty}\left\{\alpha +\right.$

[1] An epoch is completed when all samples in $\mathscr{D}_{\mathrm{train}}$ and $\mathscr{D}_{\mathrm{vali}}$ have been seen once.

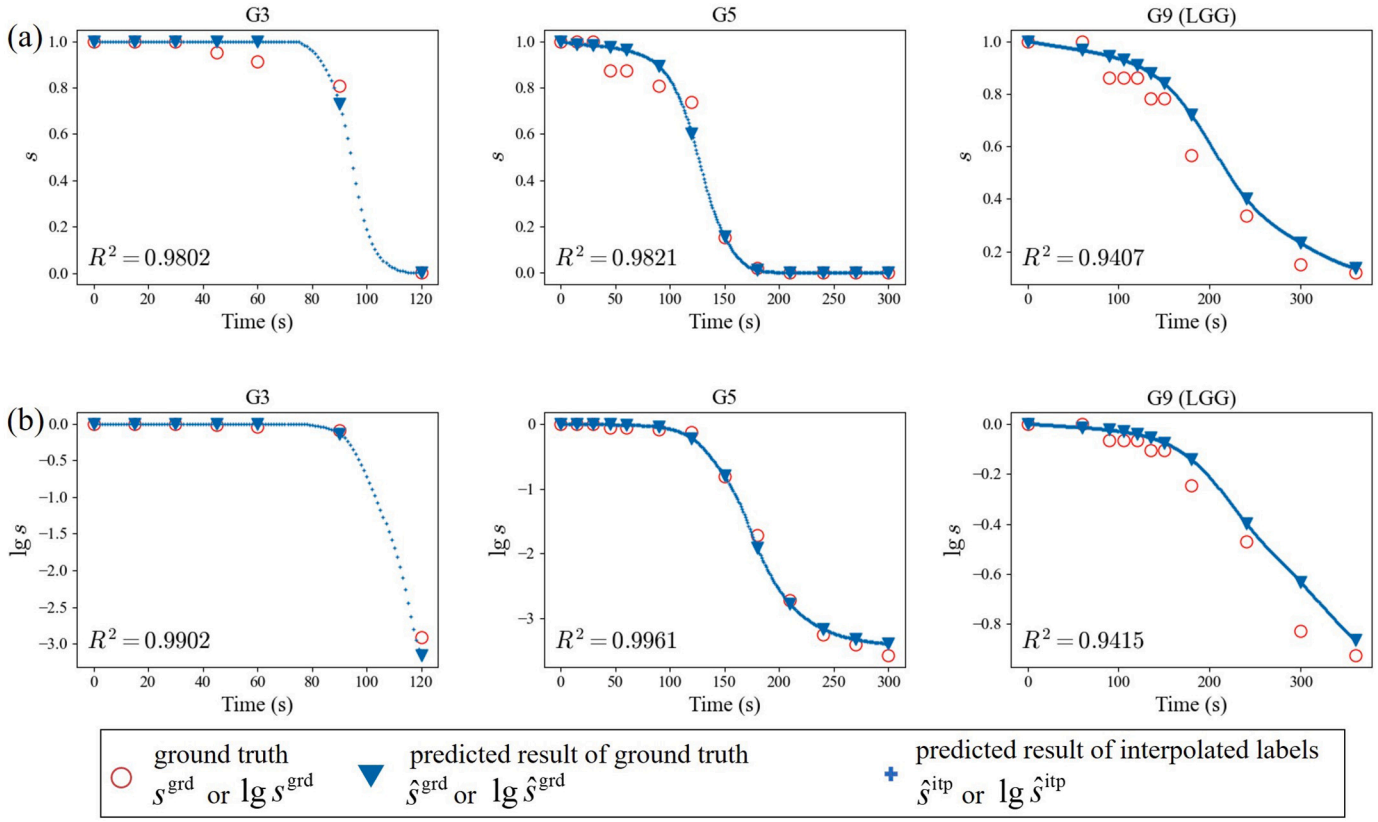



**Fig. 6.** The predicted results of the Bagging model over $\mathscr{D}_{test}$ on (a) the linear scale and (b) the log scale.

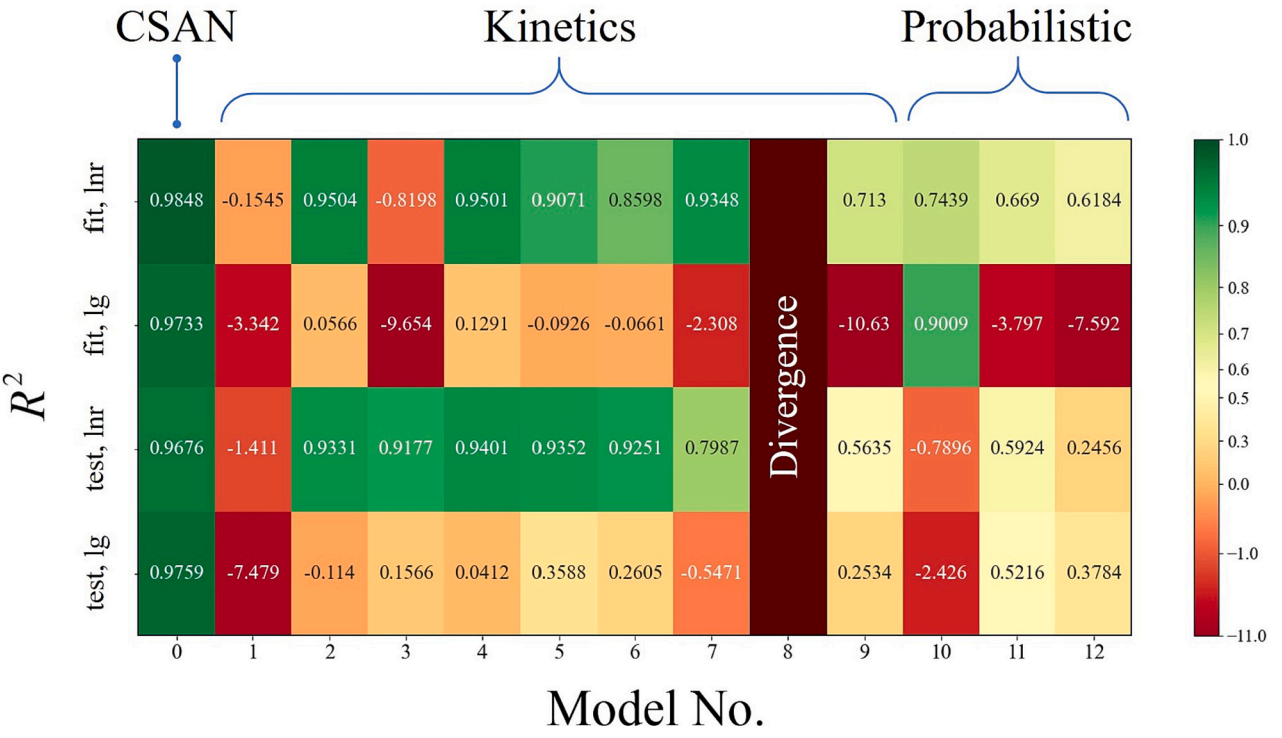



**Fig. 7.** R-Squared values of different models over $\mathscr{D}_{train+vali}$ and $\mathscr{D}_{test}$ on two scales. Model 0 represents the CSAN model developed in this work. The $R^2$ here is the arithmetical mean of the six (G1, G2, G4, G6, G7 and G8 of $\mathscr{D}_{train+vali}$) or three (G3 and G5 of $\mathscr{D}_{test}$) $R^2$ values.

$\left.\frac{\omega-\alpha}{1+exp[4\sigma(\tau-lgt)/(\omega-\alpha)]}\right\} = \omega$

Model 12 (probabilistic model): $\lim_{t\to\infty} lgs_t = \lim_{t\to\infty}\{Cexp[-exp(A+Bt)]-Cexp[-exp(A)]\} = -Cexp[-exp(A)]$

Model 0 (CSAN): $s_t \geq 10^{-7}\ (\forall t \geq 0) \quad \Rightarrow \quad \lg s_t \geq -7\ (\forall t \geq 0)$

Notwithstanding the lower asymptotes, the probabilistic models

cannot predict well in terms of $R^2_{\text{fit, lg}}$. This is because the unfitted $\omega$ of Model 11 and the badly fitted $A$ and $C$ of Model 12 ($R^2 < 0.52$) which all decide the lower asymptotes.

An exception is Model 10 with no lower asymptote but the second highest $R^2_{\text{fit, lg}}$, demonstrating the applicability of the Weibull model to describing the survival curve. Meanwhile, the much lower $R^2_{\text{test, lg}}$ shows that it has overfitted the data. It implies that the model parameters, $\alpha$ and $\beta$, were not properly correlated with the drying conditions.

In summary, the listed kinetic models fail to describe the tailing, and most listed probabilistic ones are not suitable for such multi-task prediction as well.

### 4.3. Quantitative analysis of the historical influence

The attention matrix is natural to show the correlation between any two states, or rather, the influence of State$_j$ on State$_i$, which is exactly the value of $A_{i,j}$. $A_{i,j}$ is normalized to the range of [0,1] so that an $A_{i,j}$ nearer to one implies a greater influence. Here two attention matrices are visualized in Fig. 8, one from $\mathscr{D}_{\text{train+vali}}$ and the other from $\mathscr{D}_{\text{test}}$, to explore what our CSANs have learned about the inactivation history.

Generally, the places nearer to the diagonal appear lighter, with the diagonal being the lightest and the left bottom corner being the darkest. From this result, it is deemed by our CSANs that the nearer a past state is, the stronger influence it will have on the current state. This is in line with common sense that the past influence may diminish with time. It is also to be noted that the colors do not always become lighter from the left bottom to the right top, which is partly because the initial degree and the decay rate of the influence from each state may vary from one another. From this perspective, our CSANs did learn some intrinsic knowledge about the inactivation process. Also, influences of the past states, especially those near past ones, should never be neglected.

### 4.4. Optimization of SDD towards enhanced cell viability and solvent removal

The developed CSAN model was utilized to optimize the single droplet drying process. It was formulated as a bi-objective optimization problem with an expectation to simultaneously improve cell viability and solvent removal level. The design variables include hot air temperature ($T_a$), hot air velocity ($v_a$), initial solid content ($w_{s,0}$), initial droplet volume ($V_{d,0}$), and duration of drying ($t$). Many models have been proposed to predict the drying kinetics of a single droplet [40], among which the Reaction Engineering Approach (REA) [48,49] was chosen due to its power of prediction. The predicted drying kinetics, i.e., the dynamic evolution of $T_d$ and $X$, were then fed into CSAN to predict the inactivation dynamics. The terminal solid content was calculated from the terminal moisture content as,

$$w_s = \frac{1}{X+1} \tag{15}$$

Coupling the two models in series gives an REA-CSAN model, which can predict the terminal solid content and survival ratio with specified drying conditions.

The design space $\mathscr{S}$ was specified as $\{(T_a, v_a, w_{s,0}, V_{d,0}, t) \mid 50 \le T_a$ (°C) $\le 90, 0 \le v_a$ (m/s) $\le 2, 5 \le w_{s,0}$ (wt%) $\le 30, 0.5 \le V_{d,0}$ (μL) $\le 3, 50 \le t$ (s) $\le 300\}$. Other drying conditions were fixed: the air humidity was 0.0001 kg/kg and the chamber pressure was 101,325 Pa. $\mathscr{S}$ was thoroughly explored by Monte Carlo sampling using a uniform distribution. As a result, a total of 50,000 in-silico experiments were carried out. For each experiment, the corresponding terminal solid content and cell survival ratio (after drying) can be plotted as a data point in Fig. 9 (see green squares). After imposing minimum values of two objectives (i.e., 0.65 survival ratio and 90 wt% solid content), one can identify two optimal solutions from this plot (see red stars in the figure). The values of decision variables corresponding to these two optimal solutions are

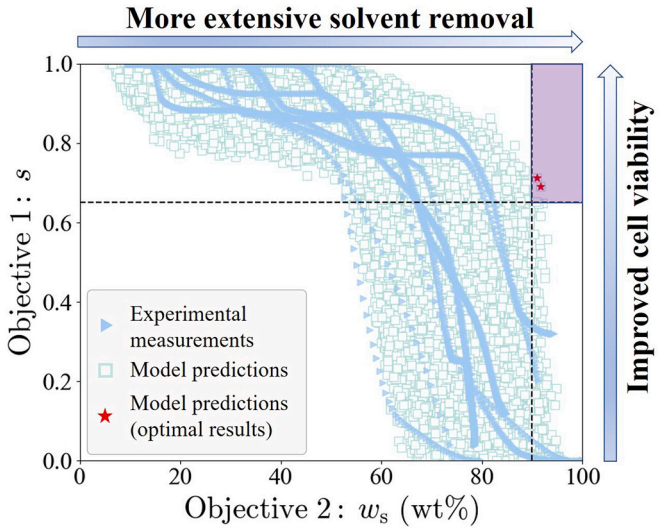



**Fig. 9.** Solution candidates and optimal solutions of the bi-objective optimization problem. Each data point in this plot shows the values of two objectives (i.e. the survival ratio and the terminal solid content) of one specific experiment, which can be either in silico or physical. The shaded area shows a space for desirable solutions.

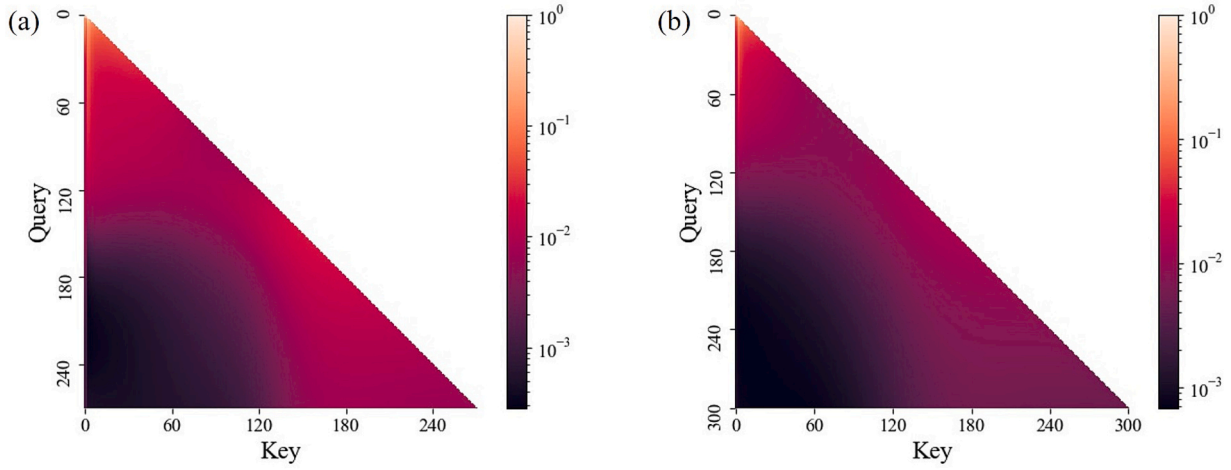



**Fig. 8.** Heatmaps of the lower triangular attention matrices of the CSALs: (a) G2 from $\mathscr{D}_{\text{train+vali}}$; (b) G5 from $\mathscr{D}_{\text{test}}$.

listed in Table 2. Experimental measurements are also plotted in the same figure as blue triangles to show the superiority of the identified optimal solutions.

Note that, each data point in this figure only shows values of two objectives of a specific experiment. The corresponding values of five decision variables can be found in a user-interactive interface in the HTML file uploaded as a supplementary material, where people can hover the mouse over a data point to check decision variable values.

## 5. Conclusion

A multi-task CSAN model was developed to predict the probiotic survival dynamics during SDD. It was found that the CSALs can capture and utilize effectively the historical drying information, whose influences should not be neglected. Compared with kinetics and probabilistic models, our model performed well on both linear- and log-scales predictions. It successfully captured the shoulder and tailing features commonly demonstrated by the experimental survival curves. It has been shown to be generalized well even for a new strain LGG whose data have not been used for training. This capability implies that our model has learned some intrinsic knowledge about cell inactivation. By thorough exploration of the design space with the developed model, two optimal conditions of SDD have been found, under which the terminal solid content and the cell survival ratio can reach up to 91 wt% and 0.71 respectively. We want to point out that the focus and unique contribution of this work is the characterization of the inactivation of probiotics. The particle morphology, which is for sure important, has not been studied here. It will be addressed in our future work. Moreover, in future, coupled with CFD models for spray dryers, such a neural network model is expected to be able to guide industrial-scale production of powders containing active probiotics through spray drying.

## Notation

*Latin letter*

| | |
|---|---|
| $T$ | temperature (°C) |
| $t$ | exposure time (s) |
| $k_\mathrm{d}$ | inactivation rate constant ($s^{-1}$) |
| $x$ | a certain property of the cell surroundings (the unit is decided by the quantity it denotes) |
| $Q$ | query matrix (−) |
| $\mathbf{q}$ | query vector (−) |
| $A$ | attention matrix (−) |
| $W$ | weight matrix (−) |
| $d$ | number of channels (−) |
| $w_\mathrm{lg}$ | weight function of the log branch |
| $y$ | a general term to denote $s$ or lg$s$ (−) |
| $b_\mathrm{f}$ | flood level (−) |
| $p_\mathrm{d}$ | dropout ratio (−) |
| $w$ | content (wt%) |
| $V$ | volume (μL) |
| $X$ | average moisture content on a dry basis (kg/kg) |
| $s$ | survival ratio (−) |
| State | state of a cell (−) |
| $N$ | cell count (cfu/mL) |
| $K$ | key matrix (−) |
| $\mathbf{k}$ | key vector (−) |
| $\mathbf{X}$ | input matrix (−) |
| $\mathbf{b}$ | bias vector (−) |
| $\bar{s}$ | arithmetic mean of $\hat{s}_\mathrm{lnr}$ and $\hat{s}_\mathrm{lg}$ (−) |
| $R^2$ | R-Squared value (−) |
| $\mathscr{L}$ | loss function (−) |
| $n_\mathrm{b}$ | batch size (−) |
| $n_\mathrm{h}$ | number of heads regarding the attention mechanism (−) |

**Table 2**
The values of objectives and decision variables corresponding to the two optimal solutions.

| $w_\mathrm{s}$ (wt%) | $s$ | $T_\mathrm{a}$ (°C) | $v_\mathrm{a}$ (m/s) | $w_\mathrm{s,0}$ (wt%) | $V_\mathrm{d,0}$ (μL) | $t$ (s) |
|---|---|---|---|---|---|---|
| 91.0 | 0.71 | 50 | 1.93 | 5.2 | 0.7 | 278 |
| 91.7 | 0.69 | 51 | 1.67 | 5.4 | 0.6 | 265 |

| | |
|---|---|
| $v$ | velocity (m/s) |
| $\mathscr{S}$ | design space |

*Greek letter*

| | |
|---|---|
| $\mu$ | mean of $x$ (the same as that of $x$) |
| $\phi$ | convolutional kernel (−) |
| $\lambda$ | coefficient of weight decay (−) |
| $\sigma$ | standard deviation (the same as that of $x$) |
| $\eta$ | learning rate (−) |

*Superscript*

| | |
|---|---|
| $\sim$ | preprocessed |
| grd | ground truth |
| $\hat{}$ | predicted |
| itp | interpolated |

*Subscript*

| | |
|---|---|
| d | droplet |
| 0 | time zero |
| vali | validation |
| $t[i]$ | time instant of the $i$-th sampling point |
| $\star$ | all the elements along one dimension |
| lnr | linear scale |
| cp | coupling |
| s | solid |
| $t$ | time instant $t$ |
| train | training |
| test | testing |
| c | a certain channel |
| $i-1:i+1$ | the $i-1$-th to $i+1$-th elements along one dimension |
| lg | log scale |
| a | ambient; air |

## Declaration of Competing Interest

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

## Data availability

The data that has been used is confidential.

## Acknowledgements

This study was supported by the National Natural Science Foundation of China (21978184, 22178239), the “Jiangsu Innovation and Entrepreneurship (Shuang Chuang) Program”, the “Jiangsu Specially-Appointed Professors Program”, and the “Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions”.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.powtec.2022.118042.

## References


[1] J.H. Crowe, F.A. Hoekstra, L.M. Crowe, Anhydrobiosis, Annu. Rev. Physiol. 54 (1992) 579–599, https://doi.org/10.1146/annurev.ph.54.030192.003051.

[2] C. Santivarangkna, U. Kulozik, P. Foerst, Alternative drying processes for the industrial preservation of lactic acid starter cultures, Biotechnol. Prog. 23 (2007) 302–315, https://doi.org/10.1021/bp060268f.

[3] S.H. Peighambardoust, A. Golshan Tafti, J. Hesari, Application of spray drying for preservation of lactic acid starter cultures: a review, Trends Food Sci. Technol. 22 (2011) 215–224, https://doi.org/10.1016/j.tifs.2011.01.009.

[4] X.C. Meng, C. Stanton, G.F. Fitzgerald, C. Daly, R.P. Ross, Anhydrobiotics: the challenges of drying probiotic cultures, Food Chem. 106 (2008) 1406–1416, https://doi.org/10.1016/j.foodchem.2007.04.076.

[5] Y.C. Wei, M.W. Woo, C. Selomulya, W.D. Wu, J. Xiao, X.D. Chen, Numerical simulation of mono-disperse droplet spray dryer under the influence of nozzle motion, Powder Technol. 355 (2019) 93–105, https://doi.org/10.1016/j.powtec.2019.07.017.

[6] J. Xiao, S.J. Yang, O.A. George, A. Putranto, W.D. Wu, X.D. Chen, Numerical simulation of mono-disperse droplet spray dryer: coupling distinctively different sized chambers, Chem. Eng. Sci. 200 (2019) 12–26, https://doi.org/10.1016/j.ces.2019.01.030.

[7] H. Jubaer, S. Afshar, G. Le Maout, S. Mejean, C. Selomulya, J. Xiao, X.D. Chen, R. Jeantet, M.W. Woo, The impact of self-sustained oscillations on particle residence time in a commercial scale spray dryer, Powder Technol. 360 (2020) 1177–1191, https://doi.org/10.1016/j.powtec.2019.11.023.

[8] N. Fu, J. Xiao, M.W. Woo, X.D. Chen, Frontiers in Spray Drying, 1st ed., CRC Press, Florida, 2020. https://www.routledge.com/Frontiers-in-Spray-Drying/Fu-Xiao-Woo-Dong-Chen/p/book/9781138364738 (accessed September 14, 2021).

[9] N. Fu, X.D. Chen, Towards a maximal cell survival in convective thermal drying processes, Food Res. Int. 44 (2011) 1127–1149, https://doi.org/10.1016/j.foodres.2011.03.053.

[10] O. Rahn, Physical methods of sterilization of microörganisms, Bacteriol. Rev. 9 (1945) 1–47. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC440886/ (accessed April 30, 2022).

[11] J.H. Crowe, J.F. Carpenter, L.M. Crowe, T.J. Anchordoguy, Are freezing and dehydration similar stress vectors? A comparison of modes of interaction of stabilizing solutes with biomolecules, Cryobiology. 27 (1990) 219–231, https://doi.org/10.1016/0011-2240(90)90023-W.

[12] J.H. Crowe, L.M. Crowe, F.A. Hoekstra, Phase transitions and permeability changes in dry membranes during rehydration, J. Bioenerg. Biomembr. 21 (1989) 77–91, https://doi.org/10.1007/BF00762213.

[13] H. Chick, The process of disinfection by chemical agencies and hot water, Epidemiol. Infect. 10 (1910) 237–286, https://doi.org/10.1017/S0022172400042959.

[14] J.H. Watkins, C.E. Winslow, Factors determining the rate of mortality of bacteria exposed to alkalinity and heat, J. Bacteriol. 24 (1932) 243–265, https://doi.org/10.1128/jb.24.3.243-265.1932.

[15] Institute of Food Technologists (IFT), Overarching principles: kinetics and pathogens of concern for all technologies, J. Food Sci. 65 (2000) 16–31, https://doi.org/10.1111/j.1750-3841.2000.tb00615.x.

[16] M. van Boekel, On the use of the Weibull model to describe thermal inactivation of microbial vegetative cells, Int. J. Food Microbiol. 74 (2002) 139–159, https://doi.org/10.1016/S0168-1605(01)00742-5.

[17] L. Humpheson, M.R. Adams, W.A. Anderson, M.B. Cole, Biphasic thermal inactivation kinetics in *salmonella enteritidis* PT4, Appl. Environ. Microbiol. 64 (1998) 459–464, https://doi.org/10.1128/AEM.64.2.459-464.1998.

[18] J.P.P.M. Smelt, J.C. Hellemons, P.C. Wouters, S.J.C. van Gerwen, Physiological and mathematical aspects in setting criteria for decontamination of foods by physical means, Int. J. Food Microbiol. 78 (2002) 57–77, https://doi.org/10.1016/S0168-1605(02)00242-8.

[19] N. Fu, M.W. Woo, C. Selomulya, X.D. Chen, Inactivation of *Lactococcus lactis* ssp. *cremoris* cells in a droplet during convective drying, Biochem. Eng. J. 79 (2013) 46–56, https://doi.org/10.1016/j.bej.2013.06.015.

[20] H. Huang, M.S.-L. Brooks, H.-J. Huang, X.D. Chen, Inactivation kinetics of yeast cells during infrared drying, Dry. Technol. 27 (2009) 1060–1068, https://doi.org/10.1080/07373930903218453.

[21] X.M. Li, S.X.Q. Lin, X.D. Chen, L.Z. Chen, D. Pearce, Inactivation kinetics of probiotic bacteria during the drying of single milk droplets, Dry. Technol. 24 (2006) 695–701, https://doi.org/10.1080/07373930600684890.

[22] G. Meerdink, K. Van't Riet, Prediction of product quality during spray drying, Food Bioprod. Process. 73 (1995) 165–170.

[23] L. Zhang, M.A.I. Schutyser, R.M. Boom, X.D. Chen, Kinetic study of the thermal inactivation of *lactobacillus plantarum* during bread baking, Dry. Technol. 37 (2019) 1277–1289, https://doi.org/10.1080/07373937.2018.1495647.

[24] M.B. Cole, K.W. Davies, G. Munro, C.D. Holyoak, D.C. Kilsby, A vitalistic model to describe the thermal inactivation of *listeria monocytogenes*, J. Ind. Microbiol. 12 (1993) 232–239, https://doi.org/10.1007/BF01584195.

[25] R.H. Linton, W.H. Carter, M.D. Pierson, C.R. Hackney, Use of a modified Gompertz equation to model nonlinear survival curves for *listeria monocytogenes* Scott a, J. Food Prot. 58 (1995) 946–954, https://doi.org/10.4315/0362-028X-58.9.946.

[26] P.A. Marechal, I.M. de Marnañón, I. Poirier, P. Gervais, The importance of the kinetics of application of physical stresses on the viability of microorganisms: significance for minimal food processing, Trends Food Sci. Technol. 10 (1999) 15–20, https://doi.org/10.1016/S0924-2244(99)00012-6.

[27] V. Venkatasubramanian, The promise of artificial intelligence in chemical engineering: is it here, finally? AICHE J. 65 (2019) 466–478, https://doi.org/10.1002/aic.16489.

[28] G. Cybenkot, Approximation by superpositions of a sigmoidal function, Math. Control Signals Syst. 2 (1989) 303–314, https://doi.org/10.1007/BF02551274.

[29] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (1989) 359–366, https://doi.org/10.1016/0893-6080(89)90020-8.

[30] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (1958) 386–408, https://doi.org/10.1037/h0042519.

[31] M.N. Hajmeer, I.A. Basheer, Y.M. Najjar, Computational neural networks for predictive microbiology II. Application to microbial growth, Int. J. Food Microbiol. 34 (1997) 51–66, https://doi.org/10.1016/S0168-1605(96)01169-5.

[32] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K.J. Lang, Phoneme recognition using time-delay neural networks, IEEE Trans. Acoust. Speech Signal Process. 37 (1989) 328–339, https://doi.org/10.1109/29.21701.

[33] J.L. Elman, Finding structure in time, Cogn. Sci. 14 (1990) 179–211, https://doi.org/10.1207/s15516709cog1402_1.

[34] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[35] H. Kim, M. Park, C.W. Kim, D. Shin, Source localization for hazardous material release in an outdoor chemical plant via a combination of LSTM-RNN and CFD simulation, Comput. Chem. Eng. 125 (2019) 476–489, https://doi.org/10.1016/j.compchemeng.2019.03.012.

[36] H.F. Fang, Y.Y. Liu, W.B. Zhang, Biomass moisture content prediction in fluidized bed dryer based on LSTM neural network, CIESC J. 71 (2020) 307–314, https://doi.org/10.11949/0438-1157.20190692.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is all you Need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 6000–6010.

[38] S.Y. Li, X.Y. Jin, Y. Xuan, X.Y. Zhou, W.H. Chen, Y.-X. Wang, X.F. Yan, Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Curran Associates, Inc, 2019, pp. 1–11. https://papers.nips.cc/paper/2019/hash/6775a0635c302542da2c32aa19d86be0-Abstract.html (accessed September 10, 2021).

[39] L. Solomon, U. Maurice, M. Mezhericher, A. Levy, Single-droplet drying processes at varying environmental conditions, Dry. Technol. 38 (2020) 27–37, https://doi.org/10.1080/07373937.2019.1599904.

[40] M. Mezhericher, A. Levy, I. Borde, Theoretical models of single droplet drying kinetics: a review, Dry. Technol. 28 (2010) 278–293, https://doi.org/10.1080/07373930903530337.

[41] N. Fu, M.W. Woo, X.D. Chen, Single droplet drying technique to study drying kinetics measurement and particle functionality: a review, Dry. Technol. 30 (2012) 1771–1785, https://doi.org/10.1080/07373937.2012.708002.

[42] H. Mao, X.D. Chen, N. Fu, Exploring the integrity of cellular membrane and resistance to digestive juices of dehydrated lactic acid bacteria as influenced by drying kinetics, Food Res. Int. 157 (2022), 111395, https://doi.org/10.1016/j.foodres.2022.111395.

[43] S. Bai, J.Z. Kolter, V. Koltun, An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv:1803.01271, 2018. http://arxiv.org/abs/1803.01271. accessed August 19, 2021.

[44] Z.H. Zhou, Machine Learning, 1st ed, Springer, Singapore, 2021, https://doi.org/10.1007/978-981-15-1967-3.

[45] T. Ishida, I. Yamane, T. Sakai, G. Niu, M. Sugiyama, Do we need zero training loss after achieving zero training error?, in: Proceedings of the 37th International Conference on Machine Learning, International Machine Learning Society, 2020, pp. 4554–4564. https://arxiv.org/pdf/2002.08709.pdf.

[46] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, Anchorage AK USA, 2019, pp. 2623–2631, https://doi.org/10.1145/3292500.3330701.

[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: an imperative style, high-performance deep learning library, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2019, pp. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[48] X.D. Chen, G.Z. Xie, Fingerprints of the drying behaviour of particulate or thin layer food materials established using a reaction engineering model, Food Bioprod. Process. 75 (1997) 213–222, https://doi.org/10.1205/096030897531612.

[49] M. Yu, Influence of the Solubility of Solids on Droplet Drying Kinetics and Solute Migration Behaviour, Master, Soochow University, 2018, https://kns.cnki.net/kns8/manage/export?filename=1018246113.nh&dbname=CMFD201901 (accessed January 24, 2022).