Narrow Operating Space Based on the Inversion of Latent Structures Model for Glycosylation Process

JING WANG1,2, (Member, IEEE), XIAOLU CHEN2, (Graduate Student Member, IEEE), YUHAN NAN2, JINGLIN ZHOU2, (Member, IEEE), AND TONGLAI XUE1

1School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China
2College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

Corresponding authors: Jing Wang (jwang@ncut.edu.cn) and Tonglai Xue (xuelt@ncut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61973023 and Grant 61573050, and in part by the Beijing Natural Science Foundation under Grant 4202052.

ABSTRACT N-linked glycan distribution plays a significant role in the generation of therapeutic proteins. It is challenging to determine the operating conditions when developing a new biopharmaceutical product with the desired glycan distributions. The glycosylation is a highly complex nonlinear system, and it is difficult to develop a reliable first-principle model that heavily relies on experimentation. Our goal is to develop a nonlinear data-driven model and find an appropriate operating space included kinds of input combination from process variables based on this model to ensure the desired product quality. A methodology is proposed based on the inversion of a nonlinear latent-variable model (locality preserving projection to latent structures, LPPLS) to identify a subspace of the knowledge space. The normal operating points of the input variables are designed based on the LPPLS inversion, and the range of operating conditions are expanded around the normal operation points through the prediction uncertainty analysis of forward and inversion model simultaneously. Finally, the designated operation space from LPPLS inversion is applied in a benchmark glycosylation model.

INDEX TERMS Locality preserving projection to latent structures, model inversion, uncertainty analysis, glycosylation.

I. INTRODUCTION

The monoclonal antibodies play an important role in biomedical science as a glycosylation end products, since the products of the glycosylation process are directly associated with the therapeutic effects of related proteins. Existing literatures demonstrated that effectiveness in vivo activity and immunogenicity of therapeutic proteins are affected by the structure of the end products [1]–[3]. Glycosylation reaction actually is an intricate and nonlinear process described by plenty of differential equations [4]–[7]. Obtaining desired glycan distributions via proper operating conditions are challenging for the biopharmaceutical manufacturers.

The US Food and Drug Administration (FDA) has initiated Quality by Design (QbD), a new method to drug development and production based on the mathematical tools [8]. Finding a suitable range of operating condition is necessary for assurance of the critical quality attributes through an intensive understanding of the relationship between the product quality and the input variables. In order to obtain the desired end products, it is crucial to guide the glycosylation process development by insight into the tight interconnection between manufacturing process conditions and product quality [9]–[11].

The space of operating conditions can be determined through the appropriate model between the process variable and product quality. First-principles model is an effective way to predict the quality or performance adequately in some chemical reactions [12]. It gives the kinetic order and reaction rate constant (prospective reaction conditions) and optimize the desired outcome with appropriate understanding of physical significance. Pantelides et al. pointed out that the first-principles model is a reliable method for describing the complex and nonlinear systems and is very useful for calculating acceptable ranges based on the assessment [13]. Design of experiment (DOE) studies are conducted to gain physiological insight and determine the acceptable input variable ranges [14]. Rathore and Winkle used DOE method to identify the parameters for process characterization [15]. Chatzizacharia and Hatzivramidis did a comparative
experiment in which several different methods were used to find the operating conditions for various data features [16]. Schoberer assess systematically the controllability of different types of specific glycoforms and glycans in glycosylation process through a statistical design of experiments scheme and Analysis of Variance (ANOVA) [17].

Mathematical modeling has been encouraged by the regulatory agencies to support the implementation of QbD paradigms [18]. The first-principles model provides a clear description of the process mechanism. However, it relies heavily on the experiments, which makes the development of reliable first-principles models difficult for the pharmaceutical industry. What’s more, the experimental effort raises significantly according to the fractional factorial design if the input number is large. Therefore, the trend of modeling in QbD is data-based rather than experience-based or experiments-based. It benefits from all kinds of data modeling methods such as multivariate statistical analysis, machine learning and artificial intelligence, who systematically extract much useful information from data of the developed products.

Latent variable model is a specifically designed tool to analyze massive amounts of correlated data, including principal component analysis (PCA) [19] and projection to latent structures (PLS) [20] and so on. The underlying factors acting on a system are extracted optimally to describe the relationship between the quality outputs and inputs. Latent variable model inversion is firstly presented to develop the operating conditions by Jaeckle and MacGregor [21]. Then the general framework for latent variable model inversion for the design and manufacturing of new products is given [22]. By inverting a latent model, a set of inputs enables a process to obtain an assigned output. If the model is sufficiently accurate to be unaffected by uncertainty, the model can be reversed directly to find the appropriate operating point without changing the quality of the products. However, since all the latent-variable models are subject to the uncertainty, the uncertainty inevitably is back-propagated to the designed inputs and the designated operating space when the models are inverted [23], [24]. For example, Pierantonio and Filippo raised the PLS inversion model to locate the operating space inside the given space under the uncertainty of the PLS model predictions [25].

Unlike other manufacturing industries, the glycosylation is a non-template driven cellular process that is exceedingly high-dimensional, complex, and non-linear. There is no sufficient mechanistic knowledge to develop high-fidelity mathematical model of glycosylation, so data-based model is a good choice for the development of new drugs. The traditional PLS is a linear modeling technique that has limited representation if process variables are relative in a strongly non-linear way. There are several nonlinear PLS methods, such as Kernel Partial Least Squares (KPLS) [26], Quadratic Partial Least Squares (QPLS) [27] and Neural Network Partial Least Squares (NNPLS) [28]. These nonlinear PLS approaches more and less suffer from the high computational complexity. The principle of locally weighted regression is good solution for dealing with the nonlinearity, which goes back to seminal work of Cleveland W. ( [29], [30]. This approach was elaborated into many other forms. For example, [30] proposed an incremental form of locally weighted projection regression, where PLS projection was incorporated on locally weighted inputs. Similarly, motivated by the distinct advantage of locality preserving projection (LPP) in nonlinear dimensionality reduction, two kinds of nonlinear PLS techniques are given, e.g., Global and Local Partial Least Squares (QGPLS) [31] and Locality Preserving Partial Least Squares (LPPLS) [32]. Especially, LPPLS utilizes the LPP technique instead of the extraction role of principal components in PLS, and can be more simple to build nonlinear PLS model and easily to apply in actual process.

In this study, we restrict our attention to the intracellular process of glycosylation at the micro-scale, before tackling the practical problem of finding appropriate macro-scale process variables to control the process of glycosylation [33], [34]. Here, LPPLS method is used to provide insight into the internal nonlinear interaction of the inputs (bio-enzymes) and outputs (various glycan classes or specific glycoforms) for glycosylation process. One can determine the appropriated process operating conditions that enables one to obtain an assigned pharmaceutical product by inverting the LPPLS model. The operating space is found only under the uncertainty from the forward model output prediction and the null space in the existing literatures. Different from the current approaches, a narrow operating space is located inside the given knowledge space under the prediction uncertainty both in the forward and inverse LPPLS models. As we known when the uncertainty back-propagates to the calculated inputs in the inverse model, it will double the prediction error and relax the experiment space. The input uncertainty directly from the inverse model can bracket the operating space than the null space approach.

The highlights of our work include,

(1) A methodology is presented to determine the operating space for a new biopharmaceutical product manufactured from the protein glycosylation process in vivo. The methodology relies on the latent variable model explored from the historical databases on the developed products that are similar to the new ones. The existing approaches for glycosylation are based on the first-principles model and analysis of variance methods.

(2) A new application of PLS and LPPLS model inversion is proposed to interpret the nonlinear relationship between the critical product attributes and the critical operation variables, and LPPLS model inversion is derived for the operating space. Since glycosylation is a cellular process with the exceedingly high-dimension, complexity, and non-linearity, LPPLS model shows higher prediction precision than the traditional PLS model.

(3) Operating space is predetermined under the prediction uncertainty both in the forward and inverse LPPLS models. The input uncertainty directly from the inverse model can crumple the operating space based on the traditional null
space approach since the proposed method avoids the uncertainty back-propagation to the calculated inputs in the inverse model.

The proposed LPPLS inversion methodology is applied to an illustrative glycosylation model that is faithfully represented by the Krambeck and Betenbaugh (KB2005 model) [5]. The designed protein qualities are characterized by eight classes of glycan as well as three specific glycoforms that typically found in recombinant biologics produced with Chinese Hamster Ovary (CHO) cells. Specific nine kinds of glycosylation enzymes are selected as the input variables since they have the direct/indirect and interactive effects on glycosylation enzymes are selected as the input variables since they have the direct/indirect and interactive effects on glycan outputs in the glycosylation process.

II. MATHEMATICAL BACKGROUND

PLS is an effective statistical tool to analyze highly correlated data by reducing the vast information into a few meaningful quantities and interpreting the implicit relationship. However, PLS extracts the maximum correlation between the quality variables and process variables (the global structural information), so it shows poor capability in some strong local nonlinear system. LPP is a newly developed dimensionality reduction algorithm [35], which makes a global nonlinear problem converted into the combination of a plurality of local linear problem by introducing local structure information. A novel combination statistical model of LPP and PLS, named as LPPLS, is given to utilize LPP technique instead of the role of PCA theory in PLS [32]. LPPLS shows excellent nonlinear modeling features, so it is suitable to analyze the highly nonlinear and intricate glycosylation process. Now we use LPPLS to extract the maximum relevant information between glycosylation enzymes and glycan outputs in glycosylation process.

A. THE LPPLS MODEL

Normalization is the transformation of the original data linearization method into the range of [0 1] using a linear function. The normalization formula is as $X_{norm} = (X - X_{min})/(X_{max} - X_{min})$. $X_{norm}$ is the normalized data set, $X$ is the original data, and $X_{max}$ and $X_{min}$ are the maximum and minimum values of the original data set. This method achieves an equal scaling of the original data.

Consider two normalized data sets, $X = [x_1^T, \ldots, x_n^T]^T \in R^{n \times m}$ and $Y = [y_1^T, \ldots, y_n^T]^T \in R^{n \times k}$, LPP seeks transformation matrix $W = [w_1, \ldots, w_d] \in R^{m \times d}$ and $C = [c_1, \ldots, c_d] \in R^{k \times d}$ that maps $X$ and $Y$ to a low-dimensional subspace $\Phi = [\phi_1^T, \ldots, \phi_n^T]^T \in R^{n \times d}(d \ll m)$ and $\Psi = [\psi_1^T, \ldots, \psi_n^T]^T \in R^{k \times d}(d \ll k)$, respectively. Considering the problem of mapping $X$ and $Y$ to the subspace vector by the transformation.

$$\phi_i = x_i W \quad (i = 1, 2, \ldots, k)$$
$$\psi_i = y_i C \quad (i = 1, 2, \ldots, n)$$

LPPLS is a multivariate regression technique used to correlate an input and response. It can be used to solve nonlinear problem by local linear technology. The LPPLS model is based on the performance of LPP to preserve the local structure. It also maintain the properties of the PLS to maximize the extraction of relevant information.

Extract the local feature from the normalized $X$ and $Y$

$$X_L = S_1^{1/2} X$$
$$Y_L = S_2^{1/2} Y$$

The principal components of the input space $X$ and output space $Y$ are extracted by LPP simultaneously to preserve local characteristics. Both $S_1$ and $S_2$ are $n \times n$ dimensional matrix, and $S_0$ is an adjacency weighting matrix denoting the neighborhood relationship between $x_i$ and $x_j$ (or $y_i$ and $y_j$),

$$S_0 = \begin{cases} 
    e^{-|x_i-x_j|^2/\delta_s}, & x_i, x_j \in \Xi (x_i; x_j) \\
    0, & \text{otherwise} 
\end{cases}$$

where $\delta_s$ is the neighbor parameter.

A number of latent variables $T = (t_1, \ldots, t_k)$ are defined to represent a low dimensional space through the LPPLS locality preserving projects. The neighboring mapping of scaled and mean centered $X_L$ and $Y_L$ are decomposed as follows

$$X_L = \sum_{i=1}^{d} t_i p_i^T + X = TP^T + \bar{X}$$
$$Y_L = \sum_{i=1}^{d} t_i q_i^T + Y = TQ^T + \bar{Y}$$

where $T = (t_1, \ldots, t_k)$ indicates the latent variables $P = [p_1, p_2, \ldots, p_A]$ and $Q = [q_1, q_2, \ldots, q_A]$ are loading matrix for $X_L$ and $Y_L$ respectively.

The score matrix $T$ can be computed directly

$$T = X_L R = S_1^{1/2} XR$$
$$r_i = \prod_{j=1}^{i-1} (t_m - w_j p_j^T) w_i$$

where $R = [r_1, \ldots, r_d]$. From eq.(7), we obtain

$$X = S_1^{-1/2} \left( TP^T + \bar{X} \right) = T_0 p^T + X'$$
$$Y = S_2^{-1/2} \left( S_1^{1/2} T_0 Q + \bar{Y} \right)$$

where $T_0$ is the core matrix of PLS method, $X'$ is $S_2^{-1/2} \bar{X}$, $\bar{X}$ and $\bar{Y}$ represent the residual space of input and output respectively.

B. FORWARD MODEL UNCERTAINTY

There is mismatch on predictions in the LPPLS model due to model calibration. The mean square error prediction method is used to characterize the prediction uncertainty.

The confidence interval (CI) on the observed output $y_{obs}$ based the LPPLS forward model is defined as:

$$CI = \hat{y}_{obs} + t_{b/2, N-d} s$$

where $t_{b/2, N-d}$ is the Student’s $t$ distribution for $b$ degrees of freedom.
where \( \hat{y}_{\text{obs}} \) is the estimation output \( y_{\text{obs}} \), \( N \) is the number of the LPPLS model calibration samples, \( d \) represents the number of degrees of model freedom, and \( \delta \) is the significance level for the CI. The \( s \) is standard deviation, which can be calculated as:

\[
s = SE \sqrt{1 + h_{\text{obs}} + \frac{1}{N}}
\]

(12)

\( SE \) is the standard error of calibration and \( h_{\text{obs}} \) is the leverage of the observation:

\[
SE = \sqrt{\frac{\sum_{n=1}^{N} (y_n - \hat{y}_n)^2}{N - d}}
\]

(13)

\[
h_{\text{obs}} = \frac{r_{\text{obs}}^2}{N - 1}
\]

(14)

where \( y_n \) is the \( n \)-th measured output and \( \hat{y}_n \) represents the \( n \)-th estimated output of the model calibration dataset.

### III. NARROWING THE OPERATING SPACE BASED ON LPPLS INVERSION

#### A. LPPLS MODEL INVERSION

Process design based on the latent variable models was introduced by Jaeckle and MacGregor [21, 37]. In general, latent variable models are used to predict the output of a model using the relationships established between variables [36]. Conversely, the latent variable regression model inversion provides helpful guidance for developing new products. It is built from the existing data and developed to assist product and process design [21]. In its inverse form, the unknown combination \( X_{\text{new}} \) of inputs can be calculated and a feasible target region for the operating conditions can be specified by suggesting to obtain a desired output \( y_{\text{des}} \).

The LPPLS relationship between input and output can be written as following,

\[
X_L = B_L Y_L
\]

(15)

\[
X = BY
\]

(16)

where \( B_L = Q (Q^T Q)^{-1} P^T \) and \( B = S_x^{-1/2} S_{y_des}^{1/2} Q (Q^T Q)^{-1} P^T \) represents the correlation matrix between input and output.

Here, products with a single quality attribute are used for analysis. Suppose the relationship between \( y \) and \( X \) in the LPPLS model is given in Eq. (9) and Eq. (10). We assume the case that desire response \( y_{\text{des}} \) is completely defined, then the LPPLS model inversion can be applied to calculate the operation conditions \( X_{\text{new}} \) corresponding to the desired quality \( y_{\text{des}} \).

Consider the following two cases:

- The dimension of new operation conditions \( X_{\text{new}} \) is same as that of \( y \). Then it is shown as:

\[
X_{\text{new}} = S_x^{-1/2} Q (Q^T Q)^{-1} P^T y_{\text{des}}
\]

(17)

where

\[
l_{\text{des}} = F_0 L_{\text{des}} Q (Q^T Q)^{-1}
\]

\[
= S_y^{1/2} y_{\text{des}} Q (Q^T Q)^{-1}
\]

(18)

- The dimension of \( X_{\text{des}} \) is larger than that of \( y \). The result would be located in an underdetermined equation system, which may cause an infinite number of solutions. We denote a \( \text{rank}(X) - \text{rank}(y) \) dimension of null space as

\[
X_{\text{null}} = \lambda^T G S_A P^T
\]

(19)

\[
Y = T Q^T = U S_A V
\]

(20)

where \( G \) are the left singular vectors of \( S_A \) that in Eq. (20) associated with the \( \text{rank}(X) - \text{rank}(y) \) dimensional zero singular values. A set of possible conditions within the range of past operating conditions is provided by selecting \( \lambda \).

#### B. UNCERTAINTY OF LPPLS MODEL INVERSION

The LPPLS inverse model give the possible input space \( X_{\text{new}} \), which may differ from the actual value \( X \) corresponding to \( y_{\text{des}} \) in the actual process. The difference comes from the uncertainty in the model. The prediction uncertainty of the forward model was explored by estimating the standard deviation of the prediction error \( s \) [23]. There are two steps, the first step is to obtain an estimate of the standard deviation of the prediction error \( s \). Then, assuming that the estimated error follows the t-statistic, a confidence interval can be established.

However, it is unreasonable to focus solely on the uncertainty of the output \( y \) instead of the input space \( X \). The modeling uncertainty is doubled and the error is increased after back-propagation to the inputs. So calculating input uncertainty directly from the inverse model is necessary.

The inputs variance can be calculated from Eq. (17), and it is noticed that the main sources of \( y_{\text{des}} \) uncertainty is the parameter \( S_x \), \( S_y \), \( P \) and \( Q \) in the model calibration,

\[
\text{var}(X_{\text{new}}) = \text{var} \left( S_x^{-1/2} S_{y_des}^{1/2} Q (Q^T Q)^{-1} P^T y_{\text{des}} \right)
\]

(21)

To facilitate the calculation of the variance in Eq. (21), the coefficient of variation \( CV[x] \) is defined for solving.

\[
E[CV(x)] = \sqrt{\text{var}(x)}
\]

(22)

where \( E(x) \) is the expectation. \( CV(x) \) is the standard deviation of a random variable with a unit mean.

The uncertainty of \( X \)-space comes from model parameter which can be calculate as

\[
\frac{E[\sqrt{\text{var}(X_{\text{new}})}]}{E(X_{\text{new}})} = \frac{E[\sqrt{\text{var}(S_x^{-1/2} S_{y_des}^{1/2} Q (Q^T Q)^{-1} P^T y_{\text{des}}))]}{E(S_x^{-1/2} S_{y_des}^{1/2} Q (Q^T Q)^{-1} P^T y_{\text{des}}))}
\]

\[
= \frac{E[\sqrt{\text{var}(S_x^{-1/2} S_{y_des}^{1/2} Q (Q^T Q)^{-1} P^T y_{\text{des}}))]}{E(S_x^{-1/2} S_{y_des}^{1/2} Q (Q^T Q)^{-1} P^T y_{\text{des}}))}
\]
calculated in Eq. (48) [39]:

\[
E[\sqrt{\text{var}(X_{\text{new}})}] = E(X_{\text{new}}) \frac{E[\sqrt{\text{var}(X/y)}]}{E(S_x^{-1/2}S_y^{-1/2}Q(Q^TQ)^{-1}P^T)}
\]

where the \(X\) and \(y\) both are training data. Finally, we get

\[
E[\sqrt{\text{var}(X_{\text{new}})}] = E(X_{\text{new}}) \frac{E[\sqrt{\text{var}(X/y)}]}{E(S_x^{-1/2}S_y^{-1/2}Q(Q^TQ)^{-1}P^T)}
\] (24)

The standard deviation \(s\) of the prediction error is defined by the error of the model parameters. According to the uncertainty theory proposed in [38], the estimated error follows the t-statistic, 100 \(\times (1 - \delta)\)% confidence interval (CI) on \(X_{\text{new}}\) is calculated in Eq. (48) [39]:

\[
CI = \bar{X}_{\text{new}} \pm t_{\frac{\alpha}{2}}s
\]

\[
s = E[\sqrt{\text{var}(X_{\text{new}})}]
\] (26)

C. DETERMINE THE OPERATING SPACE

We suppose that \(X_{\text{new}}\) is a series of inputs that can be used to achieve the desired quality of \(Y_{\text{des}}\) in complex pharmaceutical process, where \(X_{\text{new}}\) could be several different input combinations. Assuming that the model is accurate enough and there is no uncertainty, the score vector \(T_{\text{new}}\) obtained by the model inversion projection will not affect the product quality in the null space range. In this case, the operating space corresponding to the product \(Y_{\text{des}}\) is identified as null space. However, the latent variable models are affected by prediction uncertainty, which affect the null space calculation itself as well.

In some way, the experimental effort could be significantly reduced if the experimental domain were restricted to the region of subspace within which the operating space is likely to lie. Pierantoni [25] considered that the prediction uncertainty directly from forward model is back-propagated to the calculated inputs which would affect operating space by increasing the region of null space. Though the operating space is conveniently narrower than the knowledge space which is the entire space of the knowledge from the available historical dataset, there is a huge amount of computation and possibility of input value combination.

We propose a new calculating uncertainty method of inputs directly from the inverse model. When considering uncertainty of inputs, we utilize the same \(Y_{\text{des}}\) and project on \(T_{\text{new}}\) that in the null space. This methodology would narrow the operating space into the segment of null space related to \(Y_{\text{des}}\), and be able to improve experimental efficiency.

The narrowed operating space by following the steps which is adapted from reference [25].

(1) Evaluate the uncertainty of \(y_{\text{des}}\) according to the Facco P’s [25] method. The significance level is \(\delta = 0.05\), which corresponds to a confidence level of 95%. The probability density function of the \(t\) distribution centered on \(y_{\text{des}}\) shown as Figure 1(a).

(2) The score \(t_{\text{new}}\) corresponding to \(y_{\text{des}}\) projected in the space of the first two latent variables are represented with a small triangle. The null spaces are calculated and expanded into the red region by 95% CI for \(y_{\text{des}}\) in Figure 1(b). The 95% CI for \(y_{\text{des}}\) is highlighted.

(3) The 95% confidence interval on \(X_{\text{new}}\) are described in Figure 1(c) and confidence interval of inputs is lying along the null space.

(4) The narrowed operating space located in the intersection between the null spaces calculated by the prediction uncertainty for \(y_{\text{des}}\) 95% CI (who obtained from the forward model) and null space uncertainty for \(X_{\text{new}}\) 95% CI directly (who obtained from the inverse model).

IV. CASE STUDY

A. MATHEMATICAL MODEL OF GLYCOSYLATION PROCESS

Glycosylation is a post-translational modification process both at the cellular and protein level in which a carbohydrate glycan is added to protein in the endoplasmic reticulum and Golgi apparatus of a cell. The glycoproteins migrate through the four compartments of Golgi in sequence and encounter various enzymes involved to generate different kinds of specific glycoforms. Significant efforts have been made in understanding the roles [41] and the mathematical models of glycosylation [4]–[7]. Glycoprotein surface connects kinds of oligosaccharide precursor to the glycosylation sites by the action of glycosylation enzymes. The sequential enzymatic reactions of the glycosylation process are described via a reaction network with a set of reaction rules. The mathematical model for testing our method are previously developed by Krambeck and Betenbaugh (KB2005 model). This model gives the effects of 11 known enzymes and 4 donor substrate concentrations are studied. The whole reaction network KB2005 model consists of 20 reaction rules and generates 22,871 glycosylation reactions and 7,565 oligosaccharide structures [5]. The first few generations of the reaction network are illustrated in Figure 2 but the complete network would be unwieldy to project graphically. To improve the calculation efficiency of the modeling and data analysis in section 3, the KB2005 model is simplified by deleting several unnecessary enzymes. Finally, we have 365 reaction networks and 1460 enzyme reactions. The calibration and validation matrices (so-called history data) are simulated in MATLAB. The historical data set includes 100 observations of nine input glycosylation enzymes (ManI, ManII, FucT, GnTI, GnTII, GnTIV, GnTV, GalT, SiaT) and of one response variable (one of the glycan class output S0, S1, S3, G1, G2, G3, G4, F1 or one of specific glycoforms, A1F, A2G1F, A2G2S1F).

The Golgi network were modeled as a well-mixed reactor. Ultimately, the glycoforms satisfy the following equation at a steady state [5]:

\[
P_{ij} = P_{ij-1} + \tau_i r_{ij}
\]

where \(P_{ij}\) is the concentration of glycoform, and \(i\) indicate species in the Golgi compartment \(j = 4\), \(\tau_i\) is the residence time of compartment \(j\). For the simplified model, \(i = 365\), \(r_{ij}\)
is the reaction rate of producting glycoform $i$,

$$r_{ij} = \frac{k_f [E_t] \left( [UDP-S][P_i] - \frac{1}{K_m'} \left[P_{i+1}\right][UDP] \right)}{K_m (K_{md} + [UDP-S]) \left(1 + \sum_j \left[P_j\right]/K_m\right)}$$

(28)

$K_m$, $k_f$ and $K_{md}$ are the constants, which represent enzyme kinetic parameters. UDP-S is the concentration of nucleotide sugar donors, $E_t$ is the enzyme concentration. These nonlinear equations are solved to obtain the concentrations of each glycoform $i$ in the four Golgi compartments.

It is a challenge to design operating conditions due to the complex nature of glycosylation. Krambeck FJ, Betenbaugh (2005) [5] developed the first-principle models to get better understanding of the whole glycosylation process, yet the
B. PLS AND LPPLS MODEL COMPARISON

The calibration data sets are used to build a latent variable model. The model evaluates the relationship between input and output. The number of latent variables is determined by a screen test, and the selection principle is to include enough information in the inputs and outputs as much as possible [40]. Finally, two potential variables were selected and the model created includes 96.1% for the y variance.

Latent variable models are first built using the calibration data sets and are developed to evaluate the relationship between input and output. The number of latent variables is determined by the screen test to explain a sufficiently large fraction of the variance not only of the product quality but also of the input variables. We using 2 latent variables the model explains 96.1% of the variance of y.

For LPPLS model, the parameter $\delta$ is for adjusting adjacent relationship matrix $S_x$ and $S_y$. The adjustment of $S_x$ and $S_y$ will have an impact on the output predictive ability of LPPLS model. Here, the parameter $\delta = 0.1$ is determined from the history data by trial and error. The comparison of relative error can be calculated using the formula. The relative error $= \frac{(y - y_{pre})}{y}$. $y_{pre}$ means the predictive value of y.

Table 1 shows the performance of PLS and LPPLS model on the calibration and validation data in order to testing the operation variables x effect on response of glycosylation. We calculate the mean SSE value of these output, the mean SSE of validation is reduced from 11.64 to 9.56, which indicates calibration and validation accuracy of LPPLS is higher than that of PLS. It represent that LPPLS method has good fitting for the nonlinear relationship between the inputs and outputs with lower SSE value of nearly each glycan or glycoform response.

Table 1 shows that the predictive ability of PLS and LPPLS models for glycan response S0 of glycosylation process in Figure 3 and Figure 4. All the comparison of shows that LPPLS have better generalization ability than that of PLS model dealing with nonlinear system.

C. MODEL INVERSION AND RESULT

In this section, the results of the operation conditions based on inverse model uncertainty are shown. We utilize model inversions to find the $X_{new}$ related to desire output, and compare the relative error between the value of $y_{des}$ and the value of new output $y_{new}$ (the simulation result in mechanism model of glycosylation process corresponding to $X_{new}$) to observe the effect of two different methods. Table 2 gives the results of desired glycosylation $y_{des}$ output and the value of $y_{new}$ corresponding to operating conditions calculated from PLS model and LPPLS model respectively. And the comparison of relative error is calculated using the formula $error = \frac{(y_{new} - y_{des})}{y_{des}}$. It clearly demonstrates both value of $y_{new}$ within operating condition $X_{new}$ of PLS and LPPLS are similar to the expected value $y_{des}$, and the $y_{new}$ of LPPLS is closer to the desire output $y_{des}$ than that with PLS. The result indicate that LPPLS is more precise to search a proper
operating conditions. PLS method may be limited when process variables are associated in a strongly nonlinear manner. The method of using LPPLS to implement glycosylation reaction process appeared more suitable for its local preserving quality. For the advantages of LPPLS modeling and the performance comparison with other methods, our previous work explores the superiority of LPPLS for nonlinear problem processing [32].

D. THE OPERATING SPACE BASED ON PLS MODEL AND LPPLS MODEL

In order to obtain the operating conditions corresponding to the desired product quality, we design the space of the operation conditions directly based on model uncertainty and confidence interval is set to be $\delta = 0.05$ (significance level).

Table 3 gives the corresponding operation conditions and its confidence intervals for desired product quality. It can be
TABLE 2. The $y_{new}$ of PLS and LPPLS within $X_{new}$.

| RESPONSE | $Y_{des}$ | $Y_{new}(PLS)$ | RE(PLS) | $Y_{new}(LPPLS)$ | RE(LPPLS) |
|----------|-----------|----------------|---------|-----------------|-----------|
| S0       | 23        | 20.48          | 0.11    | 20.55           | 0.10      |
| S1       | 28        | 27.49          | 0.01    | 27.99           | 0.00      |
| S2       | 40        | 43.98          | 0.10    | 36.67           | 0.08      |
| S3       | 8         | 7.96           | 0.01    | 7.76            | 0.03      |
| G0       | 0.25      | 0.11           | 0.56    | 0.36            | 0.44      |
| G1       | 5.6       | 5.54           | 0.01    | 5.54            | 0.01      |
| G2       | 35        | 33.06          | 0.06    | 32.19           | 0.08      |
| G3       | 40        | 41.49          | 0.04    | 41.00           | 0.03      |
| G4       | 19        | 21.53          | 0.14    | 21.50           | 0.13      |
| F1       | 92        | 96.01          | 0.04    | 96.02           | 0.04      |
| A1F      | 0.12      | 0.06           | 0.50    | 0.093           | 0.03      |
| A2G1F    | 1.82      | 1.60           | 0.12    | 1.62            | 0.11      |
| A2G2S1F  | 9.95      | 10.93          | 0.10    | 10.44           | 0.05      |

TABLE 3. Operation conditions and CI for part of desired glycan and glycoform.

| RESPONSE  | $X_{new}$ | PLS | CI($X_{new}$) | S | $X_{new}$ | LPPLS | CI($X_{new}$) | S |
|-----------|-----------|-----|---------------|---|-----------|-------|---------------|---|
| $S0(y_{des} = 23)$ | 2.548 | 2.38-2.72 | 0.09 | 2.552 | 1.96-3.15 | 0.30 |
|           | 0.698   | 0.56-0.84 | 0.07 | 0.606 | 0.29-0.92 | 0.16 |
|           | 2.952   | 2.86-3.04 | 0.05 | 2.967 | 2.77-3.17 | 0.10 |
|           | 1.266   | 1.13-1.40 | 0.07 | 1.254 | 1.19-1.31 | 0.03 |
|           | 3.463   | 3.38-3.55 | 0.04 | 3.451 | 3.39-3.51 | 0.03 |
|           | 0.400   | 0.03-0.77 | 0.18 | 0.391 | 0.25-0.53 | 0.07 |
|           | 3.559   | 3.37-3.75 | 0.09 | 3.547 | 3.33-3.76 | 0.11 |
|           | 1.240   | 1.10-1.38 | 0.07 | 1.255 | 1.16-1.35 | 0.05 |
|           | 0.951   | 0.91-0.99 | 0.03 | 0.889 | 0.85-0.93 | 0.02 |
| $A1F(y_{des} = 0.12)$ | 2.550 | 2.29 - 2.81 | 0.13 | 2.555 | 2.38 - 2.74 | 0.09 |
|           | 0.670   | 0.57 - 0.77 | 0.05 | 0.693 | 0.59 - 0.79 | 0.05 |
|           | 2.937   | 2.80 - 3.01 | 0.07 | 2.931 | 2.85 - 3.01 | 0.04 |
|           | 1.252   | 1.19 - 1.31 | 0.03 | 1.250 | 1.13 - 1.37 | 0.06 |
|           | 3.495   | 3.15 - 3.84 | 0.17 | 3.494 | 3.09 - 3.89 | 0.20 |
|           | 0.411   | 0.35 - 0.47 | 0.03 | 0.416 | 0.36 - 0.48 | 0.03 |
|           | 3.538   | 3.32 - 3.76 | 0.11 | 3.535 | 3.08 - 4.00 | 0.23 |
|           | 1.236   | 1.14 - 1.34 | 0.05 | 1.230 | 0.95 - 1.51 | 0.14 |
|           | 0.931   | 0.83 - 1.03 | 0.05 | 0.929 | 0.69 - 1.17 | 0.12 |

clearly seen that the operating space is a narrow and effective region of the original knowledge space from KB2005 model $X_{Low} = [1.25, 0.33, 1.52, 0.64, 1.81, 0.2, 2.67, 0.66, 0.5]$, $X_{High} = [3.75, 0.99, 4.57, 1.93, 5.42, 0.6, 4.54, 1.89, 1.5]$ which brackets the entire fraction of the glycosylation experiment space. The designated space can greatly reduce the cost of the experiment.

As discussed previously, LPPLS seems to be a more appropriate way to find operating conditions. Comparing the average standard deviation by applying PLS and LPPLS in Table 3, it’s worth noting that values of the average standard deviation $s$ are related to the mechanism model of glycosylation process which reflect different enzymes have different degree of influence on the different products. The smaller the value of $s$, the greater correlation with the products. The average standard deviation of SiaT which is tight bound up with output (glycan class S0) should lie in a low range value 0.03, and that of LPPLS (0.02) is a bit lower for its high modeling accuracy. SiaT only provides sialic acid residues, which is the most important part of the S glycans during glycosylation. It is consistent with the glycosylation machinery. Low value of the $s$ indicates that SiaT is a sensitive input enzyme associate with S0. Furthermore, the value $s$ of enzyme unnecessary for production of S0 like Fuct and Galt are also improved in LPPLS via expanding CI to larger regions from 2.38 - 2.72 and 0.56 - 0.84 to 1.96 - 3.15 and 0.29 - 0.92.

The dataset contains 100 observations of 9 input enzymes and one response variable S0. Using 2 latent variables, the model explains 96.1% of the variance of $y$ in PLS.
and LPPLS. Confidence limits of historical data is considered in the latent space of the scores in the shape of an ellipsoid as well. Within $A = 2 > k = 1$, a null space exists.

Figure 5 shows a graphical representation of the one-dimensional null space. It shows the fractional space of the first two latent variables. Here, circles represent historical data, and dashed ovals are 95% confidence space. We assume value of $\lambda$ is from $-0.4$ to $0.4$ and then the null spaces associated to each score vector are calculated resulting in the blue line of Figure 4. The green little triangle indicates the operating conditions vector $T_{new}$. Process operating conditions $X_{real} = X_{new} + X_{null}$ produce the product $y_{des}$ with same quality characteristics since null space not affecting the desire output. Engineering judgement would be used to select a reasonable conditions from by different value of $\lambda$ to determine the most economical operation.

In figure 4, the red line stands for the $T_{new}$ designate space which is bracketed by model uncertainty given in section 3.4 at the significance level $\delta = 0.05$. The operating space is lying within the null space and narrow the entire region of it. In addition, the boundary of operating space $X$ by applying the LPPLS inverse model are shown in Table 3. Compared with null space in Figure 4, the operating space significantly reduces the required experimental effort to achieve the desire output.

The narrowed design space could be found in the intersection area between the null spaces calculated by the prediction uncertainty on $y_{des}$ and null space uncertainty for $X_{new}$, when both the inversion certainty of input the prediction uncertainty of output are analyzed simultaneously. The uncertainty on $y_{des}$ is estimated at $\delta = 0.05$ which are described in Reference [24] Figure 6, panel (a) illustrates that the bracket design space of PLS associated with each score are located in red area. The design space of LPPLS (red region in panel (b)) is obviously narrowed.

Figure 7 illustrates the comparison of predict value of $y$ within PLS and LPPLS inversion forms respectively. Each $\bar{y}_{pred}$ is mean values of $y_{pred}$ calculate by a series $X_{new}$ selected from operating space of PLS and LPPLS. Product quality $y_{pred}$ is obtained with different value from several operating conditions $X_{new}$ corresponding to their own $y_{des}$. Therefore, multiple sets of operating conditions are selected to reach a statistically significant results. In this experiment, we chose 4 set of product quality of S0 ($y_{des} = 22, 23, 24, 25$) and 10 set of operating conditions for each product quality. For example, $X_{new}$ related to S0--des = 21 is calculated...
by model inversion forms and expanded to operating space under model uncertainty, we chose 10 $X_{new}$ in operating space and obtain 10 different value of $y_{pred}$. We see that the $y_{pred}$ obtained from LPPLS inversion are much closer to desired values than that from PLS inversion form. The results showed that LPPLS is more accurate compared with PLS, which means the nonlinear method of searching glycosylation process operating space is effective.

V. CONCLUSION

In this article, we have developed a methodology based on local preserving projections latent variable in inversion form to determine the process operating conditions of desired quality specifications. The confidence interval based on the inverse model is used to design the space of the manipulated quality specifications. The local preserving projections latent variable in inversion form is an area for further research.

Furthermore, this study has only considered the case that desire class glycan determined by a single quality attribute. Extending this methodology to the multivariate glycan groups is an area for further research.

REFERENCES

[1] I. J. del Val, A. Constantinou, A. Dell, S. Haslam, K. M. Polizzi, and C. Kontoravdi, “A quantitative and mechanistic model for monoclonal antibody glycosylation as a function of nutrient availability during cell culture,” *BMC Prog.*, vol. 7, no. Suppl 6, p. 010, 2013.
[2] A. Helenius, “Intracellular functions of N-Linked glycans,” *Science*, vol. 291, no. 5512, pp. 2364–2369, Mar. 2001.
[3] A. M. Sinclair and S. Elliott, “Glycoengineering: The effect of glycosylation on the properties of therapeutic proteins,” *J. Pharmaceutical Sci.*, vol. 94, no. 8, pp. 1626–1635, Aug. 2005.
[4] P. Umapathi and J. E. Bailey, “A mathematical model of N-linked glycoform biosynthesis,” *Biotechnol. Bioeng.*, vol. 55, no. 6, pp. 890–908, Sep. 1997.
[5] F. J. Krambeck and M. J. Betenbaugh, “A mathematical model of N-linked glycosylation,” *Biotechnol. Bioeng.*, vol. 92, no. 6, pp. 711–728, Dec. 2005.
[6] G. Liu, D. D. Marathe, K. L. Matta, and S. Neelamegham, “Systems-level modeling of cellular glycosylation reaction networks: O-linked glycan formation on natural selectin ligands,” *Bioinformatics*, vol. 24, no. 23, pp. 2740–2747, Dec. 2008.
[7] I. Jimenez del Val, J. M. Nagy, and C. Kontoravdi, “A dynamic mathematical model for monoclonal antibody N-linked glycosylation and nucleotide sugar donor transport within a maturing golgi apparatus,” *Biotechnol. Prog.*, vol. 27, no. 6, pp. 1730–1743, Nov. 2011.
[8] S. T. Elliott, “Pharmaceutical CGMPs for the 21st century—A risk based approach. Final report. U.S. Department of Health and Human Services,” *U.S. Food Drug Admin.*, Sep. 2004. [Online]. Available: https://www.fda.gov/media/77391/download
[9] M. Sokolov, J. Ritscher, N. MacKinnon, J.-M. Bielsser, D. Brühlmann, D. Rothenhäusler, G. Thanei, M. Soos, M. Stettler, J. Souquet, H. Broly, M. Morbidelli, and A. Butté, “Robust factor selection in early cell culture process development for the production of a biosimilar monoclonal antibody,” *Biotechnol. Prog.*, vol. 33, no. 1, pp. 181–191, Jan. 2017.
[10] M. Sokolov, M. Morbidelli, A. Butté, J. Souquet, and H. Broly, “Sequential multivariate cell culture modeling at multiple scales supports systematic shaping of a monoclonal antibody toward a quality target,” *Biotechnol. J.*, vol. 13, no. 4, Apr. 2018, Art. no. 1700461.
[11] M. Sokolov, J. Ritscher, N. MacKinnon, J. Souquet, H. Broly, M. Morbidelli, and A. Butté, “Enhanced process understanding and multivariate prediction of the relationship between cell culture process and monoclonal antibody quality,” *Biotechnol. Prog.*, vol. 33, no. 5, pp. 1368–1380, Sep. 2017.
[12] S. Muanan, B. Dumitrescu, and C. D. Giurcaneanu, “Maximum entropy expectation-maximization algorithm for fitting latent-variable graphical models to multivariate time series,” *Entropy*, vol. 20, no. 1, pp. 76–96, 2018.
[13] J. J. Peterson, “A Bayesian approach to the ICH q8 definition of design space,” *J. Biopharmaceutical Statist.*, vol. 18, no. 5, pp. 959–975, Sep. 2008.
[14] D. Brühlmann, M. Sokolov, A. Butté, M. Sauer, J. Hemberger, J. Souquet, H. Broly, and M. Jordan, “Parallel experimental design and multivariate analysis provides efficient screening of cell culture media supplements to improve biosimilar product quality,” *Biotechnol. Bioeng.*, vol. 114, no. 7, pp. 1448–1458, Jul. 2017.
[15] M. Sato, Y. Ochiai, S. Kijima, N. Nagai, Y. Ando, M. Shikano, and Y. Nomura, “Quantitative modeling and simulation in PMDA: A Japanese regulatory perspective,” *CPT, Pharmacometrics Syst. Pharmacol.*, vol. 6, no. 7, pp. 413–415, Jul. 2017.
[16] K. A. Chatzizacharia and D. T. Hatziavramidis, “DesignSpace approach for pharmaceutical tablet development,” *Ind. Eng. Chem. Res.*, vol. 53, no. 30, pp. 12003–12009, 2014.
[17] J. Schoberer, Y. J. Shin, U. Vavra, C. Veit, and R. Strasser, “Analytical protein glycosylation in the ER,” *Methods Mol. Biol.*, vol. 1691, pp. 205–222, Oct. 2018.
[18] ICH Quality Implementation Working Group, Points to Consider (R2), ICH endorsed guide for ICH Q8/Q9/Q10 Implementation, ICH, Geneva, Switzerland, 2011.
[19] M. Sokolov, M. Soos, B. Neunstoecklin, M. Morbidelli, A. Butté, R. Lecra, T. Solacroup, M. Stettler, and H. Broly, “Fingerprint detection and process prediction by multivariate analysis of fed-batch monoclonal antibody cell culture data,” *Biotechnol. Prog.*, vol. 31, no. 6, pp. 1633–1644, Nov. 2015.
[20] P. Geladi and B. R. Kowalski, “Partial least-squares regression: A tutorial,” *Analytica Chim. Acta*, vol. 185, pp. 1–17, Jan. 1986.
[21] C. M. Jackle and J. F. Macgregor, “Product design through multivariate statistical analysis of process data,” *AICHE J.*, vol. 44, no. 5, pp. 1105–1118, May 1998.
[22] W. Li and C. Zhao, “Latent variable based concurrent multi-trends analysis method for monitoring batch processes with irregular and limited batches,” *Can. J. Chem. Eng.*, vol. 95, no. 9, pp. 1817–1829, Sep. 2017.
[23] Y. A. Shardt, S. Mehrkanoon, K. Zhang, X. Yang, J. Suykens, S. X. Ding, and K. Peng, “Modelling the strip thickness in hot steel rolling mills using least-squares support vector machines,” *Can. J. Chem. Eng.*, vol. 96, no. 1, pp. 171–178, 2017.
[24] K. Faber and B. R. Kowalski, “Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares,” *J. Chemometrics*, vol. 11, no. 3, pp. 181–238, May 1997.
[25] P. Faccio, F. Dal Pastro, N. Meneghetti, F. Bezzo, and M. Barolo, “Bracketing the design space within the knowledge space in pharmaceutical product development,” *Ind. Eng. Chem. Res.*, vol. 54, no. 18, pp. 5128–5138, May 2015.
[26] R. Rosipal and L. J. Trejo, “Kernel partial least squares regression in reproducing kernel Hilbert space,” *J. Mach. Learn. Res.*, vol. 2, pp. 97–123, Dec. 2001.
[27] L. Wang, “Enhanced fault detection for nonlinear processes using modified kernel partial least squares and the statistical local approach,” *Can. J. Chem. Eng.*, vol. 96, no. 5, pp. 1116–1126, May 2018.
[28] R. Jia, Z. Mao, and F. Wang, “Combining just-in-time modelling and batch-wise unfolded PLS model for the derivative-free batch-to-batch optimization,” *Can. J. Chem. Eng.*, vol. 96, no. 5, pp. 1156–1167, May 2018.
[29] W. S. Cleveland, “Robust locally weighted regression and smoothing scatterplots,” J. Amer. Stat. Assoc., vol. 74, no. 368, pp. 829–836, Dec. 1979.

[30] W. S. Cleveland and S. J. Devlin, “Locally weighted regression: An approach to regression analysis by local fitting,” J. Amer. Stat. Assoc., vol. 83, no. 403, pp. 596–610, Sep. 1988.

[31] B. Zhong, J. Wang, J. Zhou, H. Wu, and Q. Jin, “Quality-related statistical process monitoring method based on global and local partial least-squares projection,” Ind. Eng. Chem. Res., vol. 55, no. 6, pp. 1609–1622, Feb. 2016.

[32] J. Wang, B. Zhong, and J. L. Zhou, “Quality-relevant fault monitoring based on locality-preserving partial least-squares statistical models,” Ind. Eng. Chem. Res., vol. 56, no. 24, pp. 7009–7020, Jun. 2017.

[33] M. M. St. Amand, D. Radhakrishnan, A. S. Robinson, and B. A. Ogunnaike, “Identification of manipulated variables for a glycosylation control strategy,” Biotechnol. Bioeng., vol. 111, no. 10, pp. 1957–1970, Oct. 2014.

[34] M. M. St. Amand, K. Tran, D. Radhakrishnan, A. S. Robinson, and B. A. Ogunnaike, “Controllability analysis of protein glycosylation in cho cells,” PLoS ONE, vol. 9, no. 2, Feb. 2014, Art. no. e87973.

[35] X. F. He and P. T. Niyogi, “Locality preserving projections,” in Proc. Conf. Adv. Neural Inf. Process. Syst., Vancouver, BC, Canada, 2004, pp. 153–160.

[36] E. Tomba, P. Facco, F. Bezzo, and M. Barolo, “Latent variable modeling to assist the implementation of quality-by-design paradigms in pharmaceutical development and manufacturing: A review,” Int. J. Pharmaceutics, vol. 457, no. 1, pp. 283–297, Nov. 2013.

[37] C. M. Jaeckle and J. F. MacGregor, “Industrial applications of product design through the inversion of latent variable models,” Chemometric Intell. Lab. Syst., vol. 50, no. 2, pp. 199–210, Mar. 2000.

[38] H. Martens and M. Martens, “Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR),” Food Qual. Preference, vol. 11, nos. 1–2, pp. 5–16, Jan. 2000.

[39] L. Zhang and S. Garcia-Munoz, “A comparison of different methods to estimate prediction uncertainty using partial least squares (PLS): A practitioner’s perspective,” Chemometric Intell. Lab. Syst., vol. 97, no. 2, pp. 152–158, Jul. 2009.

[40] M. E. Ismail and D. Stanton, “Classical orthogonal polynomials as moments,” Can. J. Math., vol. 49, no. 3, pp. 601–616, 2018.

[41] P. Hossler, S. F. Khattak, and Z. J. Li, “Optimal and consistent protein glycosylation in mammalian cell culture,” Glycobiology, vol. 19, no. 9, pp. 936–949, Sep. 2009.

**JING WANG** (Member, IEEE) received the B.Eng. degree in industry automation and the Ph.D. degree in control theory and control engineering from Northeastern University, in 1994 and 1998, respectively.

She was a Professor with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China, from 1999 to 2020. She is currently a Professor with the School of Electrical and Control Engineering, North China University of Technology, Beijing. Her research interests include advanced control, process monitoring, fault detection and diagnosis, and their applications.

**XIAOLU CHEN** (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree in control science and engineering with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China. Her research interests include the modeling and fault diagnosis of complex industrial processes, data causality analysis, and intelligent learning algorithm.

**JINGLIN ZHOU** (Member, IEEE) received the B.Eng. degree from the Daqing Petroleum Institute, in 1999, the M.Sc. degree from Hunan University, in 2002, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2005.

He is currently a Professor with the College of Information Science and Technology, Beijing University of Chemical Technology. His research interests include stochastic distribution control, fault detection and diagnosis, variable structure control, and their applications.

**TONGLAI XUE** received the Ph.D. degree in environmental science and engineering from the Beijing University of Technology, in 2016.

He is currently a Senior Engineer with the School of Electrical and Control Engineering, North China University of Technology, Beijing, China. His research interests include advanced control, process control, and data mining.

**YUHAN NAN** received the B.Eng. degree in automation and the M.Sc. degree in control science and engineering from the Beijing University of Chemical Technology, in 2015 and 2018, respectively. Her research interests include statistical analysis and data mining.