Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

SEQUENTIAL IMPORTANCE SAMPLING FOR HYBRID MODEL BAYESIAN INFERENCE TO SUPPORT BIOPROCESS MECHANISM LEARNING AND ROBUST CONTROL

Wei Xie
Keqi Wang
Hua Zheng

Mechanical and Industrial Engineering
Northeastern University
360 Huntington Ave
Boston, MA 02115, USA

Ben Feng

Statistics and Actuarial Science
University of Waterloo
200 University Ave W
Waterloo, ON N2L 3G1, CANADA

ABSTRACT

Driven by the critical needs of biomanufacturing 4.0, we introduce a probabilistic knowledge graph hybrid model characterizing the risk- and science-based understanding of bioprocess mechanisms. It can faithfully capture the important properties, including nonlinear reactions, partially observed state, and nonstationary dynamics. Given very limited real process observations, we derive a posterior distribution quantifying model estimation uncertainty. To avoid the evaluation of intractable likelihoods, Approximate Bayesian Computation sampling with Sequential Monte Carlo (ABC-SMC) is utilized to approximate the posterior distribution. Under high stochastic and model uncertainties, it is computationally expensive to match output trajectories. Therefore, we create a linear Gaussian dynamic Bayesian network (LG-DBN) auxiliary likelihood-based ABC-SMC approach. Through matching the summary statistics driven through LG-DBN likelihood that can capture critical interactions and variations, the proposed algorithm can accelerate hybrid model inference, support latent state monitoring, and facilitate mechanism learning and robust control.

1 INTRODUCTION

The biopharmaceutical manufacturing industry is growing rapidly and it plays a critical role to ensure public health and support economy. *However, biomanufacturing often faces critical challenges, including high complexity, high variability, and very limited process observations.* As new biotherapeutics (e.g., cell and gene therapies) become more and more personalized, it requires more advanced manufacturing protocols. For example, the seed cells, extracted from individual patients or donors, can have different optimal culture policies. Therefore, the production process involves a complex stochastic decision process (SDP) with output trajectory dynamics and variations influenced by biological/physical/chemical (a.k.a. *biophysicochemical*) reactions occurring at molecular, cellular, and system levels.

In general, there are two main categories of biomanufacturing process modeling methodologies in the existing literature: mechanistic and data-driven approaches. The ordinary/partial differential equations (ODE/PDE) mechanistic models are developed based on biophysicochemical mechanisms. They have good interpretability and show generally higher extrapolation power than data-driven models. However, existing mechanistic models often fail to rigorously account for *uncertainties*, i.e., inherent stochasticity and model estimation uncertainty. For example, batch-to-batch variation, known as a major source of bioprocess uncertainty (Mockus et al. 2015), is ignored in deterministic mechanistic models. Therefore, mechanistic models may not fit well to the observations collected from real systems in many situations, which also limits their power in terms of mechanism learning, process monitoring, and robust control to support flexible on-demand manufacturing. On the other hand, data-driven approaches often use general

978-1-6654-7661-4/22/$31.00 ©2022 IEEE

statistical or machine learning approaches to capture process patterns observed in data. The prediction accuracy of these models largely depends on the the size of process data and their interpretability is limited.

Driven by the critical challenges of biomanufacturing and limitations of existing process modeling approaches, we developed *a probabilistic knowledge graph (KG) hybrid ("mechanistic and statistical") model* characterizing the risk- and science-based understanding of biophysicochemical reactions and bioprocess spatiotemporal causal interdependiences (Xie et al. 2022; Zheng et al. 2021; Zheng et al. 2022). It can leverage the information from existing mechanistic models within and between operation units, and facilitate mechanism learning from *heterogeneous* online and offline measurements. Zheng et al. (2021) introduced KG-based reinforcement learning (RL) to guide customized decision making. Since the proposed model-based RL scheme on the Bayesian KG, accounting for both stochastic and model uncertainties, can provide an insightful prediction on how the effect of inputs propagates through mechanism pathways, impacting on the output trajectory dynamics and variations, it can find optimal process control policies that are interpretable and robust against model uncertainty, and overcome the key challenges of biopharmaceutical manufacturing.

Zheng et al. (2022) further generalized this KG hybrid model to capture the important properties of integrated biomanufacturing processes, including nonlinear reactions, partially observed state, and non-stationary dynamics. It can faithfully represent and advance the understanding of underlying bioprocessing mechanisms. This model allows us to inference unobservable latent states and critical pathways to support process monitoring and control; for example it enables the estimation of metabolic states and cell response to environmental perturbations. Since the hybrid model involves latent state variables, nonlinear reactions, and time-varying kinetic coefficients with uncertainty (such as cell growth rate and molecular reaction rates), it is challenging to evaluate the likelihood function and derive a posterior distribution.

Approximate Bayesian Computation (ABC) is introduced in the literature to approximate posterior distributions for process models with intractable likelihoods. It bypasses the evaluation of likelihoods by simulating model parameters, generating synthetic data sets, and only accepting posterior samples when the sampled process outputs are "close" enough to real observations. For complex biomanufacturing processes with high stochastic and model uncertainties, the accept rate is very low and it is computationally challenging to generate sample trajectories close to real-world observations. Recently, there has been much interest in formalizing an auxiliary likelihood based ABC, which uses a simpler and related model to derive summary statistics as distance measure (Gleim and Pigorsch 2013; Martin et al. 2019; Sisson et al. 2018).

Following the spirit of the auxiliary likelihood-based ABC (Martin et al. 2019), we utilize a linear Gaussian dynamic Bayesian network (LG-DBN) auxiliary model to derive summary statistics as a distance measure for ABC-SMC that can support dimensional reduction and accelerate online inference on hybrid models with high fidelity characterizing complex bioprocessing mechanisms. The proposed ABC approach in conjunction with sequential importance sampling can efficiently approximate hybrid model posterior distribution. Therefore, the key contributions of this paper is: given very limited real-world data, we propose a LG-DBN auxiliary likelihood based ABC-SMC sampling approach to generate posterior samples of bioprocess hybrid model parameters quantifying model uncertainty. This simple LG-DBN auxiliary model can capture the critical dynamics and variations of bioprocess trajectory, ensure computational efficiency, and enable the inference on model and latent state, which can facilitate mechanism online learning and support robust process control. The empirical study shows that our approach can outperform the original ABC-SMC approach especially given tight computational budget.

The remainder of the paper is organized as follows. We provide the problem description and summarize the proposed framework in Section 2. Then, we present a probabilistic KG hybrid model capturing the important properties of biomanufacturing processes and describe ABC for approximating the posterior distribution of model parameters in Section 3. We derive the LG-DBN auxiliary likelihood based summary statistics to accelerate Bayesian inference on the hybrid models with high fidelity in Section 4. We conduct the empirical study on cell therapy manufacturing in Section 5 and conclude the paper in Section 6.

2 PROBLEM DESCRIPTION AND PROPOSED FRAMEWORK

Driven by the needs of biomanufacturing process online learning, monitoring, and control, we create a probabilistic KG hybrid model characterizing underlying mechanisms and causal interdependencies between critical process parameters (CPPs) and critical quality attributes (CQAs). It models how the effect of state and action at any time t, denoted by $\{\boldsymbol{s}_t, \boldsymbol{a}_t\}$, propagates through mechanism pathways impacting on the output trajectory dynamics and variations. Here we use cell culture process for illustration. The process state transition model is denoted by $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\theta})$ where $\boldsymbol{s}_t \in \mathscr{S} \subset \mathbb{R}^d$ denotes the *partially observable bioprocess state* (i.e., extra- and intra-cellular enzymes, proteins, metabolites, media), $\boldsymbol{a}_t \in \mathscr{A}$ denotes action (i.e., agitation rate, oxygen/nutrient feeding rates), $\mathscr{A}$ is a finite set of actions, and $t \in \mathscr{H} \equiv \{1, 2, \dots, H+1\}$ denotes the discrete time index. At any time t, the agent partially observes the state $\boldsymbol{s}_t$ and takes an action $\boldsymbol{a}_t$. Thus, given model parameters $\boldsymbol{\theta}$, the joint distribution of process trajectory $\boldsymbol{\tau} = (\boldsymbol{s}_1, \boldsymbol{a}_1, \dots, \boldsymbol{s}_H, \boldsymbol{a}_H, \boldsymbol{s}_{H+1})$ becomes,

$$p(\boldsymbol{\tau}|\boldsymbol{\theta}) = p(\boldsymbol{s}_1) \prod_{t=1}^{H} p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\theta}) p(\boldsymbol{a}_t).$$

Due to the nature of biopharmaceutical manufacturing, the state transition model $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\theta})$ is highly complex, non-linear, and nonstationary. The state transition $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\theta})$ is modeled by a hybrid ("mechanistic + statistical") model. Its structure takes existing mechanistic models as prior. For example, since the key factors influencing process dynamics and variability in the cell culture process are induced by cellular metabolisms (O'Brien et al. 2021), the probabilistic state transition of this KG hybrid model can incorporate cell metabolic/gene regulatory networks and account for cell-to-cell variations. *Therefore, there are key properties in biomanufacturing process, specially for personalized cell/gene therapies, including (1) partially observed state ($\boldsymbol{s}_t$) that means only limited proportion of state observable; (2) stochastic state transition model $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\theta})$ involves high inherent stochasticity; and (3) very limited and heterogeneous online and offline measurement data.*

Given limited historical observations, we focus on hybrid model Bayesian inference to support online mechanism learning, monitoring, and reliable interpretable prediction, accounting for both inherent stochasticity and model uncertainty. The posterior distribution will be derived to quantify model uncertainty.

2.1 Hybrid Modeling for Bioprocess with Partially Observed State

At any time t, the process state is composed of observable and latent state variables, i.e., $\boldsymbol{s}_t = (\boldsymbol{x}_t, \boldsymbol{z}_t)$ with $\boldsymbol{x}_t \in \mathscr{S}_x$ and latent variables $\boldsymbol{z}_t \in \mathscr{S}_z$, where $\mathscr{S}_x \subset \mathbb{R}^{d_x}$ and $\mathscr{S}_z \subset \mathbb{R}^{d_z}$ with $\mathscr{S} = \mathscr{S}_x \times \mathscr{S}_z$ and $d = d_x + d_z$. Denote the partially observed state trajectory as $\boldsymbol{\tau}_x \equiv (\boldsymbol{x}_1, \boldsymbol{a}_1, \dots, \boldsymbol{x}_H, \boldsymbol{a}_H, \boldsymbol{x}_{H+1})$. Given model parameters $\boldsymbol{\theta}$, by integrating out latent variables $(\boldsymbol{z}_1, \dots, \boldsymbol{z}_{H+1})$, the likelihood evaluation of any observation $\boldsymbol{\tau}_x$, i.e.,

$$p(\boldsymbol{\tau}_x|\boldsymbol{\theta}) = \int \cdots \int p(\boldsymbol{\tau}|\boldsymbol{\theta}) d\boldsymbol{z}_1 \cdots d\boldsymbol{z}_{H+1},$$

is intractable especially when the dimensions of model parameters and latent states are high. This hybrid model characterizes the risk- and science-based understanding of underlying bioprocess mechanisms and quantifies spatial-temporal causal interdependencies of CPPs/CQAs. It can connect heterogeneous online and offline measures to infer unobservable state (such as metabolic state determining cell product functional behaviors and critical quality attributes), support process monitoring, and facilitate real-time release.

We model the bioprocess state transition with a hybrid ("mechanistic and statistical") model. Given the existing ODE-based mechanistic model, $\mathrm{d}\boldsymbol{s}/\mathrm{d}t = \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\phi})$, by using the finite difference approximation for derivatives, i.e., $\mathrm{d}\boldsymbol{s} \approx \Delta\boldsymbol{s}_t = \boldsymbol{s}_{t+1} - \boldsymbol{s}_t$, and $\mathrm{d}t \approx \Delta t$, we construct the hybrid model for state transition,

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \Delta t \cdot \boldsymbol{f}_x(\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{a}_t; \boldsymbol{\phi}) + \boldsymbol{e}^x_{t+1} \quad \text{and} \quad \boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \Delta t \cdot \boldsymbol{f}_z(\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{a}_t; \boldsymbol{\phi}) + \boldsymbol{e}^z_{t+1},$$

with unknown kinetic coefficients $\boldsymbol{\phi} \in \mathbb{R}^{d_\phi}$ (e.g., cell growth and inhibition rates). The function structures of $\boldsymbol{f}_x(\cdot)$ and $\boldsymbol{f}_z(\cdot)$ are the parts of $\boldsymbol{f}(\cdot)$ associated to the observable state output $\boldsymbol{x}_{t+1}$ and the latent state output

$\boldsymbol{z}_{t+1}$. By applying the central limit theorem, the residual terms, accounting for inherent stochasticity and other factors, are modeled by multivariate Gaussian distributions $\boldsymbol{e}^x_{t+1} \sim \mathscr{N}(0, V^x)$ and $\boldsymbol{e}^z_{t+1} \sim \mathscr{N}(0, V^z)$ with zero means and covariance matrices V^x and V^z. Then, the state transition distribution becomes,

$$\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{a}_t \sim \mathscr{N}\left(\boldsymbol{x}_t + \Delta t \cdot \boldsymbol{f}_x(\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{a}_t), V^x_{t+1}\right) \quad \text{and} \quad \boldsymbol{z}_{t+1}|\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{a}_t \sim \mathscr{N}\left(\boldsymbol{z}_t + \Delta t \cdot \boldsymbol{f}_z(\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{a}_t), V^z_{t+1}\right).$$

Thus, the stochastic state transition model, specified by parameters $\boldsymbol{\theta} = (\boldsymbol{\phi}, V^x, V^z)^\top$, characterizes the bioprocess inherent stochasticity, dynamics, and mechanisms (such as biophysicochemical reactions).

2.2 Challenges of Hybrid Model Inference Under High Stochasticity and Limited Data

Given limited real-world data with size m, denoted by $\mathscr{D} = \{\boldsymbol{\tau}^{(i)}_x : i = 1, 2, \ldots, m\}$, the model uncertainty is quantified by a posterior distribution derived through applying the Bayes' rule,

$$p(\boldsymbol{\theta}|\mathscr{D}) \propto p(\boldsymbol{\theta})p(\mathscr{D}|\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{i=1}^{m} p\left(\boldsymbol{\tau}^{(i)}_x \middle| \boldsymbol{\theta}\right), \tag{1}$$

where $p(\boldsymbol{\theta})$ represents the prior distribution. It is challenging to directly derive or computationally assess the posterior distribution $p(\boldsymbol{\theta}|\mathscr{D})$ in eq. (1). First, there often exist large-dimensional latent state variables $\boldsymbol{z}_t$, especially for multi-scale bioprocess model characterizing the scientific understanding of individual cell response to micro-environmental perturbation and accounting for cell-to-cell variation in metabolic/gene networks. It is computationally expensive to assess the likelihood for each observation, $p(\boldsymbol{\tau}^{(i)}_x|\boldsymbol{\theta}) = \int \cdots \int p(\boldsymbol{\tau}^{(i)}|\boldsymbol{\theta}) d\boldsymbol{z}_1 \cdots d\boldsymbol{z}_{H+1}$ with $i = 1, 2, \ldots, m$, especially for bioprocess with optical sensor online monitoring (that means the value of H is large). Second, the mechanistic model $\boldsymbol{f}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\phi})$ can be a nonlinear function of state $\boldsymbol{s}$ and parameters $\boldsymbol{\phi}$. The random kinetic coefficients often have batch-to-batch variations. For example, the kinetic coefficients (such as cell growth rate, oxygen/nutrient uptake rates, and metabolic waste excretion rate) can depend on the gene expression of seed cells and cell culture environments. Third, the amount of real-world process observations can be very limited (especially for personalized bio-drug manufacturing) even though inherent stochasticity and model uncertainty are high.

Thus, in Section 3, ABC approach is considered to approximate the posterior distribution of KG hybrid model with high fidelity that can capture the key features of biomanufacturing processes. Since it is computationally expensive especially under the situations with high stochastic and model uncertainties, LG-DBN auxiliary ABC-SMC is created to facilitate the Bayesian inference. Based on Taylor series approximation of the hybrid model, this linear auxiliary model can be accurate for biomanufacturing process with optical sensor (e.g., fluorescent probe and Raman sensors) online monitoring.

3 SEQUENTIAL IMPORTANCE SAMPLING FOR BAYESIAN INFERENCE

When the evaluation of likelihood for each observation is computationally intractable, i.e., $p(\boldsymbol{\tau}^{(i)}_x|\boldsymbol{\theta}) = \int \cdots \int p(\boldsymbol{\tau}^{(i)}|\boldsymbol{\theta}) d\boldsymbol{z}_1 \cdots d\boldsymbol{z}_{H+1}$ for $i = 1, 2, \ldots, m$, the ABC approach is recommended to approximate the posterior distribution (Sisson et al. 2018). In the naive ABC implementation, we draw a candidate sample from the prior $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ and then generate a simulation dataset $\mathscr{D}^\star$ from the hybrid model. If the simulated dataset $\mathscr{D}^\star$ is "close" to the observed real-world observations $\mathscr{D}$, we accept the sample $\boldsymbol{\theta}$; otherwise reject it. Thus, we approximate the posterior distribution $p(\boldsymbol{\theta}|\mathscr{D})$ with $p(\boldsymbol{\theta}|d(\mathscr{D}, \mathscr{D}^\star) \leq h)$, where $d(\cdot)$ is a distance metric (e.g., Euclidean distance, likelihood distance) and h is an approximation tolerance level.

However, for any given small tolerance level h, we often face very low accept rate for complex biomanufacturing processes with high stochastic and model uncertainties. The random discrepancy between process trajectories $\mathscr{D}$ and $\mathscr{D}^\star$ can be large even when the parameter sample $\boldsymbol{\theta}$ equals to $\boldsymbol{\theta}^c$. In addition, given very limited real-world data for the complex hybrid model, the dimension of model parameters $\boldsymbol{\theta}$ is large and the model uncertainty can be high.

To increase the accept rate and ensure the computational efficient generation of samples $\boldsymbol{\theta}$ with good approximation on the critical features occurring in the real-world data, we will design the distance measure $d(\cdot)$ based on *designed* lower dimensional summary statistics, denoted by $\eta(\mathscr{D})$, in Section 4. It means that we accept samples $\boldsymbol{\theta}$ which lead to the summary statistics of simulated data, denoted by $\eta^\star = \eta(\mathscr{D}^\star)$, close to that of observations $\eta_{obs} = \eta(\mathscr{D})$. Thus, the standard ABC framework (Sisson et al. 2018) becomes

$$p_{ABC}(\boldsymbol{\theta}|\eta_{obs}) \propto \int \mathbb{1}(d(\eta^\star, \eta_{obs}) \leq h) p(\eta^\star|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\eta^\star. \tag{2}$$

As the distance tolerance h gradually decreases, we have

$$\lim_{h\to 0} p_{ABC}(\boldsymbol{\theta}|\eta_{obs}) \propto \int \delta_{\eta_{obs}}(\eta^\star) p(\eta^\star|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\eta^\star = p(\eta_{obs}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\eta_{obs}),$$

where $\delta_X(x)$ denotes the Dirac measure, defined as $\delta_X(x) = 1$ if $x = X$ and $\delta_X(x) = 0$ otherwise.

A good design of ABC summary statistics η should balance complexity v.s. informativeness. If the summary statistics η are sufficient for $\boldsymbol{\theta}$, then $p(\boldsymbol{\theta}|\eta_{obs})$ will be equivalent to $p(\boldsymbol{\theta}|\mathscr{D})$. With small threshold h, the ABC approximate $p_{ABC}(\boldsymbol{\theta}|\eta_{obs})$ in (2) can provide a good approximation of the true posterior. However, in the most situations, it is challenging to specify sufficient statistics since the KG hybrid model is built based on mechanistic models and it accounts for the key features including (1) partially observed state; (2) heterogeneous offline and online measures; (3) nonlinear mechanisms and dynamics; and (4) batch-to-batch variations on mechanistic coefficients. *Therefore, in Section 4, we project the bioprocess KG hybrid model into linear Gaussian dynamic Bayesian Network (LG-DBN) auxiliary model space that has tractable likelihood. It can capture first two moments of bioprocess dynamics and variations to support robust and optimal control.* We will use the LG-DBN likelihood to derive summary statistics accelerating the generation of critical samples $\boldsymbol{\theta}$. Our study also shows that complex KG hybrid models will asymptotically converge to a LG-DBN model as time interval Δt becomes "smaller and smaller" by applying Taylor approximation (Zheng et al. 2021). Thus, this LG-DBN approximation holds well for many cases with online sensor monitoring and bioprocess (e.g., biological state of cells) that does not change quickly.

The basic ABC generates candidate samples from the prior $p(\boldsymbol{\theta})$ and uses the accept/reject approach to retain those samples satisfying the approximation threshold requirement. This can be extremely ineffective especially for the situations using noninformative prior that has a wide sampling space. The *ABC-sequential Monte Carlo (ABC-SMC) methods* derived from the sequential importance sampling (Toni et al. 2009; Beaumont et al. 2009) can improve the sampling efficiency through generating candidate samples from updated posterior approximates. In specific, let g denote the index of ABC iterations used to improve the approximation of the posterior distribution $p(\boldsymbol{\theta}|\mathscr{D})$. We select a sequence of intermediate target distribution, denoted by $\{\pi_g\}$ for $g = 1, 2, \ldots, G$, converging to $p(\boldsymbol{\theta}|\mathscr{D})$ as we gradually reduce the tolerance level h_g,

$$\pi_g(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \mathbb{1}\left(d(\eta^\star, \eta_{obs}) \leq h_g\right). \tag{3}$$

Direct sampling from $p(\boldsymbol{\theta})$ and having the accept/reject based on the condition $\mathbb{1}\left(d(\eta^\star, \eta_{obs}) \leq h_g\right)$ in (3) is not simulation efficient. The accept rate can be low as h_g becomes smaller and smaller.

Thus, we use the *sequential importance sampling (SIS)* and select a sequence of proposal distribution, denoted by $\{\zeta_g\}$ for $g = 1, 2, \ldots, G$, to improve the sampling efficiency, i.e.,

$$\zeta_g(\boldsymbol{\theta}) = \mathbb{1}\left(\pi_g(\boldsymbol{\theta}) > 0\right) \int \pi_{g-1}(\boldsymbol{\theta}') K(\boldsymbol{\theta}', \boldsymbol{\theta}) d\boldsymbol{\theta}', \tag{4}$$

where $K(\boldsymbol{\theta}', \boldsymbol{\theta})$ is a Markov kernel. The proposal distribution $\zeta_g(\boldsymbol{\theta})$ is defined as the perturbed previous intermediate distribution π_{g-1} through the perturbation kernel K. After generating N sample particles from the proposal distribution $\boldsymbol{\theta}_n \sim \zeta_g(\boldsymbol{\theta})$ for $n = 1, 2, \ldots, N$, we weight it by $w_n^{(g)} = \pi_g(\boldsymbol{\theta}_n)/\zeta_g(\boldsymbol{\theta}_n)$. The condition, $\mathbb{1}(\pi_g(\boldsymbol{\theta}) > 0)$, in (4) is used to satisfy the importance sampling condition, i.e., $\{\boldsymbol{\theta} : \pi_g(\boldsymbol{\theta}) >$

Algorithm 1: DBN auxiliary based ABC-SMC for hybrid model Bayesian inference.

```
Input: the prior distribution $p(\boldsymbol{\theta})$; the number of particles $N$; process observations $\mathscr{D} = \{\boldsymbol{\tau}_x^{(i)}\}_{i=1}^m$;
       the perturbation kernel function $K(\cdot)$; the number of particles to keep at each iteration
       $N_\alpha = \lfloor \alpha N \rfloor$ with $\alpha \in [0,1]$; and the minimal acceptance rate $p_{acc_{min}}$.
Output: posterior distribution approximate $\widehat{p}(\boldsymbol{\theta}|\mathscr{D})$.
for $n = 1, \dots, N$ do
    1. Sample $\boldsymbol{\theta}_n^{(0)} \sim p(\boldsymbol{\theta})$;
    2. Generate $m \times L$ predicted trajectories $\mathscr{D}^\star = \{\boldsymbol{\tau}_x^{\star(i)}\}_{i=1}^{mL}$ using $\boldsymbol{\theta}_n^{(0)}$;
    3. Set $q_n^{(0)} = d(\eta(\mathscr{D}), \eta(\mathscr{D}^\star))$ and $w_n^{(0)} = 1$;
4. Let $h_1$ be the first $\alpha$-quantile of $q^{(0)} = \{q_n^{(0)}\}_{n=1}^N$;
5. Let $\{(\boldsymbol{\theta}_n^{(1)}, w_n^{(1)}, q_n^{(1)})\} = \{(\boldsymbol{\theta}_n^{(0)}, w_n^{(0)}, q_n^{(0)}) | q_n^{(0)} \le h_1, 1 \le n \le N\}$, $p_{acc} = 1$ and $g = 2$;
while $p_{acc} > p_{acc_{min}}$ do
    for $n = N_\alpha + 1, \dots, N$ do
        6. Sample $\boldsymbol{\theta}_n^\star$ from $\boldsymbol{\theta}_k^{(g-1)}$ with probability $\frac{w_k^{(g-1)}}{\sum_{j=1}^{N_\alpha} w_j^{(g-1)}}$, $1 \le k \le N_\alpha$;
        7. Perturb the particle to obtain $\boldsymbol{\theta}_n^{(g-1)} \sim K(\boldsymbol{\theta}|\boldsymbol{\theta}_n^\star) = \mathscr{N}(\boldsymbol{\theta}_n^\star, \Sigma)$;
        8. Generate $m \times L$ predicted trajectories $\mathscr{D}^\star = \{\boldsymbol{\tau}_x^{\star(i)}\}_{i=1}^{mL}$ using $\boldsymbol{\theta}_n^{(g-1)}$:
        9. Set $q_n^{(g-1)} = d(\eta(\mathscr{D}), \eta(\mathscr{D}^\star))$;
        10. Set $w_n^{(g-1)} = \frac{p(\boldsymbol{\theta}_n^{(g-1)}) \mathbb{1}(d(\eta(\mathscr{D}), \eta(\mathscr{D}^\star)) \le h_{g-1})}{\sum_{j=1}^{N_\alpha} \frac{w_j^{(g-1)}}{\sum_{k=1}^{N_\alpha} w_k^{(g-1)}} K(\boldsymbol{\theta}_n^{(g-1)} | \boldsymbol{\theta}_j^{(g-1)})}$;
    11. Set $p_{acc} = \frac{1}{N - N_\alpha} \sum_{k=N_\alpha+1}^N \mathbb{1}(q_k^{(g-1)} \le h_{g-1})$;
    12. Let $h_g$ be the first $\alpha$-quantile of $q^{(g-1)} = \{q_n^{(g-1)}\}_{n=1}^N$;
    13. Let $\{(\boldsymbol{\theta}_n^{(g)}, w_n^{(g)}, q_n^{(g)})\} = \{(\boldsymbol{\theta}_n^{(g-1)}, w_n^{(g-1)}, q_n^{(g-1)}) | q_n^{(g-1)} \le h_g, 1 \le n \le N\}$ and $g = g + 1$;
14. Return the approximated posterior distribution, $\widehat{p}(\boldsymbol{\theta}|\mathscr{D}) = \frac{1}{\sum_{n'=1}^{N_\alpha} w_{n'}^{(g-1)}} \sum_{n=1}^{N_\alpha} w_n^{(g-1)} \delta_{\boldsymbol{\theta}_n^{(g-1)}}(\boldsymbol{\theta})$.
```

$0\} \subset \{\boldsymbol{\theta} : \zeta_g(\boldsymbol{\theta}) > 0\}$. This can avoid the weight becoming infinite, which will lead to high variance on the SIS estimator. We set the first proposal distribution to be the prior distribution, i.e., $\zeta_1(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$.

The proposed LG-DBN auxiliary likelihood-based ABC-SMC sampling procedure is summarized in Algorithm 1. It incorporates an adaptive selection approach on the threshold h_g from Toni et al. (2009), Lenormand et al. (2013), Del Moral et al. (2006). The initial set of parameter samples $\{\boldsymbol{\theta}_n^{(0)}\}_{n=1}^N$ is generated from the prior distribution $p(\boldsymbol{\theta})$ in Step 1. The associated weights $\{w_n^{(0)}\}_{n=1}^N$ and distances $\{q_n^{(0)}\}_{n=1}^N$ are calculated in Steps 2-3. Considering the impact from stochastic uncertainty, we generate mL predicted trajectories denoted by $\mathscr{D}^\star = \{\boldsymbol{\tau}_x^{\star(i)}\}_{i=1}^{mL}$, compute the LG-DBN auxiliary based summary statistics $\eta(\mathscr{D}^\star)$, and then calculate the distance $q_n^{(0)}$. The tolerance level h_g in any g-th iteration is determined online as the α-quantile of the $\{q_n^{(g)}\}_{n=1}^N$. The particles, satisfying this tolerance denoted by $\{\boldsymbol{\theta}_n\}_{n=1}^{N_\alpha}$, constitute the weighted empirical distribution to approximate the posterior distribution in Steps 5 and 13, where $N_\alpha = \lfloor \alpha N \rfloor$. The approximation accuracy is measured by the corresponding distances $\{q_n\}_{n=1}^{N_\alpha}$. Then, $N - N_\alpha$ new particles are drawn from the proposal distribution $\zeta_g(\boldsymbol{\theta})$ in Steps 6-7. The associated weights and distances are calculated in Steps 8-10. The tolerance level h_g and the posterior distribution approximate $\pi_g(\boldsymbol{\theta})$ are updated in Steps 12-13. We repeat Steps 6-13 until the proportion of particles satisfying the tolerance level h_{g-1} among the $N - N_\alpha$ new particles is below the pre-specified threshold

$p_{acc_{min}}$. Finally, the ABC-SMC algorithm returns the weighted empirical distribution, denoted by $\widehat{p}(\boldsymbol{\theta}|\mathscr{D})$, as posterior distribution approximate in Step 14.

4 LG-DBN AUXILIARY LIKELIHOOD-BASED SUMMARY STATISTICS

Motivated by the studies (Martin et al. 2019; Gleim and Pigorsch 2013), in this section, we derive LG-DBN auxiliary likelihood-based summary statistics for ABC-SMC to capture the crucial features of the bioprocess trajectory on dynamics and variations. Given a set of observations $\mathscr{D} = \{\boldsymbol{\tau}_x^{(i)} : i = 1,2,\ldots,m\}$, we derive the MLE of LG-DBN auxiliary model, i.e., maximizing the log-likelihood $\hat{\boldsymbol{\beta}}(\mathscr{D}) = \text{argmax}_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}|\mathscr{D})$. Then we use it as the summary statistics $\eta \triangleq \hat{\boldsymbol{\beta}}$ to calculate the distance measure $q \equiv d(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}^\star)$, where $\hat{\boldsymbol{\beta}}^\star$ is the summary statistics of simulated data. In the following, we first develop the LG-DBN model with only observable state transition in Section 4.1 and then discuss the parameter estimation in Section 4.2.

4.1 The development of LG-DBN Auxiliary Model

Let $x_1^k \sim \mathscr{N}(\mu_1^{x,k},(v_1^{x,k})^2)$ with $k = 1,2\ldots,d$ model the variation in the k-th initial observed state. In practice, to ensure product quality, CPPs are strictly regulated by the specifications of ranges of values. Thus, we model $\boldsymbol{a}_t$ as a random variable, i.e., $a_t^k \sim \mathscr{N}(\lambda_t^{x,k},(\sigma_t^{x,k})^2)$ with $k = 1,2\ldots,d_a$ and $t = 1,2\ldots,H$. At any time t, the LG-DBN auxiliary model has the state transition model,

$$\boldsymbol{x}_{t+1} = \boldsymbol{\mu}_{t+1}^x + \boldsymbol{\psi}_t^x(\boldsymbol{x}_t - \boldsymbol{\mu}_t^x) + \boldsymbol{\psi}_t^a(\boldsymbol{a}_t - \boldsymbol{\mu}_t^a) + (V_{t+1}^x)^{\frac{1}{2}}\boldsymbol{\omega}, \tag{5}$$

where $\boldsymbol{\mu}_t^x = (\mu_t^1,\ldots,\mu_t^{d_x})$, $\boldsymbol{\mu}_t^a = (\lambda_t^1,\ldots,\lambda_t^{d_a})$, $\boldsymbol{\omega}$ is an d_x-dimensional standard normal random vector, and $V_{t+1}^x = \text{diag}((v_{t+1}^{x,k})^2)$ is a diagonal covariance matrix. The coefficients $\boldsymbol{\psi}_t^x$ and $\boldsymbol{\psi}_t^a$ measure the main effects of current observed state $\boldsymbol{x}_t$ and action $\boldsymbol{a}_t$ on the next observed state $\boldsymbol{x}_{t+1}$. Let $\boldsymbol{\sigma}_t = (\sigma_t^1,\ldots,\sigma_t^{d_a})$ and $\boldsymbol{v}_t^x = (v_t^{x,1},\ldots,v_t^{x,d_x})$. Thus, the LG-DBN model, specified by parameters $\boldsymbol{\beta} = (\boldsymbol{\mu}^x,\boldsymbol{\mu}^a,\boldsymbol{\psi}^x,\boldsymbol{\psi}^a,\boldsymbol{\sigma},\boldsymbol{v}^x) = \{(\boldsymbol{\mu}_t^x,\boldsymbol{\mu}_t^a,\boldsymbol{\psi}_t^x,\boldsymbol{\psi}_t^a,\boldsymbol{\sigma}_t,\boldsymbol{v}_t^x)|1 \le t \le H\}$, has the joint distribution of bioprocess trajectory: $p(\boldsymbol{\tau}_x) = p(\boldsymbol{x}_1,\boldsymbol{a}_1,\ldots,\boldsymbol{x}_H,\boldsymbol{a}_H,\boldsymbol{x}_{H+1}) = p(\boldsymbol{x}_1)\prod_{t=1}^H p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t,\boldsymbol{a}_t)p(\boldsymbol{a}_t)$.

Let $\boldsymbol{\mu}_\tau = [\boldsymbol{\mu}_1^x,\boldsymbol{\mu}_1^a,\ldots,\boldsymbol{\mu}_H^x,\boldsymbol{\mu}_H^a,\boldsymbol{\mu}_{H+1}^x]^\top$. Following Murphy (2012), we rewrite (5) in the following form

$$\boldsymbol{\tau}_x - \boldsymbol{\mu}_\tau = B(\boldsymbol{\tau}_x - \boldsymbol{\mu}_\tau) + \Sigma_\tau^{\frac{1}{2}}\boldsymbol{\omega}_\tau \tag{6}$$

where $\boldsymbol{\omega}_\tau$ is an $((H+1)d_x + Hd_a)$-dimensional standard normal random vector, $\Sigma_\tau^{\frac{1}{2}} = \text{diag}(\boldsymbol{v}_1^x,\boldsymbol{\sigma}_1,\ldots,\boldsymbol{v}_H^x,\boldsymbol{\sigma}_H,\boldsymbol{v}_{H+1}^x)$ is the diagonal matrix of the conditional standard deviations of observed state and actions, and the coefficient matrix of observed trajectory is written as

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \boldsymbol{\psi}_1^x & \boldsymbol{\psi}_1^a & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{\psi}_2^x & \boldsymbol{\psi}_2^a & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \boldsymbol{\psi}_H^x & \boldsymbol{\psi}_H^a & 0 & 0 \end{bmatrix}.$$

Thus, by rearranging (6) and letting $\boldsymbol{\tau}_x - \boldsymbol{\mu}_\tau = (I-B)^{-1}\Sigma_\tau^{\frac{1}{2}}\boldsymbol{\omega}_\tau$, we have $\boldsymbol{\tau}_x \sim \mathscr{N}(\boldsymbol{\mu}_\tau,(I-B)^{-1}\Sigma_\tau(I-B)^{-\top})$ with mean $\mathbb{E}[\boldsymbol{\tau}_x] = \boldsymbol{\mu}_\tau$ and covariance matrix $\text{Cov}(\boldsymbol{\tau}_x - \boldsymbol{\mu}_\tau) = (I-B)^{-1}\Sigma_\tau(I-B)^{-\top}$.

4.2 Linear Gaussian Dynamic Bayesian Network based Summary Statistics

Let $\tilde{\boldsymbol{\tau}}_x \equiv (\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{a}}_1,\ldots,\tilde{\boldsymbol{x}}_H,\tilde{\boldsymbol{a}}_H,\tilde{\boldsymbol{x}}_{H+1}) = \boldsymbol{\tau}_x - \boldsymbol{\mu}_\tau$, where $\tilde{\boldsymbol{x}}_t$ and $\tilde{\boldsymbol{a}}_t$ denote centered observable state and decision. Given m observations $\mathscr{D} = \{\boldsymbol{\tau}_x^{(i)}\}_{i=1}^m$, the unbiased estimator $\hat{\boldsymbol{\mu}}_\tau = \frac{1}{m}\sum_{i=1}^m \boldsymbol{\tau}_x^{(i)}$ can be easily obtained by using the fact $\mathbb{E}[\boldsymbol{\tau}_x] = \boldsymbol{\mu}_\tau$. The log-likelihood of the centered trajectory observations $\{\tilde{\boldsymbol{\tau}}_x^{(i)}\}_{i=1}^m$ becomes,

$$\max_{\boldsymbol{\psi}^x,\boldsymbol{\psi}^a,V} \ell\left(\tilde{\boldsymbol{\tau}}_x^{(1)},\ldots,\tilde{\boldsymbol{\tau}}_x^{(m)};\boldsymbol{\psi}^x,\boldsymbol{\psi}^a,V\right) = \max_{\boldsymbol{\psi}^x,\boldsymbol{\psi}^a,V} \log\prod_{i=1}^{m} p\left(\tilde{\boldsymbol{\tau}}_x^{(i)}\right)$$

$$= \max_{V_1} \sum_{i=1}^{m} \log p(\tilde{\boldsymbol{x}}_1^{(i)}) \left[\sum_{t=1}^{H} \max_{\sigma_t} \sum_{i=1}^{m} \log p(\tilde{\boldsymbol{a}}_t^{(i)})\right] \left[\sum_{t=1}^{H} \max_{\boldsymbol{\psi}_t^x,\boldsymbol{\psi}_t^a,\boldsymbol{v}_{t+1}^x} \sum_{i=1}^{m} \log p(\tilde{\boldsymbol{x}}_{t+1}^{(i)}|\tilde{\boldsymbol{x}}_t^{(i)},\tilde{\boldsymbol{a}}_t^{(i)})\right].$$

Since both initial state $\tilde{\boldsymbol{x}}_1$ and actions $\tilde{\boldsymbol{a}}_t$ for $t=1,\ldots,H$ are normally distributed with mean zero, the MLEs of their variance are sample covariances: $\hat{v}_1^{x,k} = \frac{1}{m}\sum_{i=1}^{m}(\tilde{x}_1^{k(i)})^2$ with $k=1,2,\ldots,d_x$ and $\hat{\sigma}_t^k = \frac{1}{m}\sum_{i=1}^{m}(\tilde{a}_t^{k(i)})^2$ with $k=1,2,\ldots,d_a$. In addition, at any time t, we have the log-likelihood of a sample $\tilde{\boldsymbol{\tau}}_x^{(i)}$

$$\log p(\tilde{\boldsymbol{x}}_{t+1}^{(i)}|\tilde{\boldsymbol{x}}_t^{(i)},\tilde{\boldsymbol{a}}_t^{(i)}) \propto -\frac{m}{2}\log|V_{t+1}^x| - \frac{1}{2}\left(\tilde{\boldsymbol{x}}_{t+1}^{(i)} - \boldsymbol{\psi}_t^x\tilde{\boldsymbol{x}}_t^{(i)} - \boldsymbol{\psi}_t^a\tilde{\boldsymbol{a}}_t^{(i)}\right)^\top V_{t+1}^x \left(\tilde{\boldsymbol{x}}_{t+1}^{(i)} - \boldsymbol{\psi}_t^x\tilde{\boldsymbol{x}}_t^{(i)} - \boldsymbol{\psi}_t^a\tilde{\boldsymbol{a}}_t^{(i)}\right).$$

Let $\tilde{\boldsymbol{x}}_{t+1}^{(i)}$ and $(\tilde{\boldsymbol{x}}_t^{(i)},\tilde{\boldsymbol{a}}_t^{(i)})$ denote the i-th rows of output matrix Y and input matrix X. Let $B_t = (\boldsymbol{\psi}_t^x,\boldsymbol{\psi}_t^a)^\top$ denote the coefficient vector. As a result, the MLEs of $\boldsymbol{\psi}_t^x$ and $\boldsymbol{\psi}_t^a$ are

$$(\hat{\boldsymbol{\psi}}_t^x,\hat{\boldsymbol{\psi}}_t^a)^\top = \hat{B}_t = \arg\max_{B_t} -\frac{1}{2}(Y-XB_t)^\top (V_{t+1}^x)^{-1}(Y-XB_t) = (X^\top(V_{t+1}^x)^{-1}X)^{-1}X^\top(V_{t+1}^x)^{-1}Y.$$

The MLE of each standard deviation can be computed by $\hat{v}_t^{x,k} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\tilde{x}_t^{k(i)}\right)^2}$ (Fuller and Rao 1978). In sum, given observations $\mathscr{D}$, the MLE of LG-DBN auxiliary model is $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\mu}}^x,\hat{\boldsymbol{\mu}}^a,\hat{\boldsymbol{\psi}}^x,\hat{\boldsymbol{\psi}}^a,\hat{\boldsymbol{\sigma}},\hat{\boldsymbol{v}}^x)$.

5 EMPIRICAL STUDY

In this section, we use the erythroblast cell therapy manufacturing example presented in Glen et al. (2018) to assess the performance of the proposed LG-DBN auxiliary likelihood-based ABC-SMC approach.

5.1 Hybrid Modeling for Cell Therapy Manufacturing Process

The cell culture process of erythroblast exhibits two phases: a relatively uninhibited growth phase followed by an inhibited phase. Glen et al. (2018) identified that this reversible inhibition is caused by an unknown cell-driven factor rather than commonly known mass transfer or metabolic limitations. They developed an ODE-based mechanistic model describing the dynamics of an unidentified autocrine growth inhibitor accumulation and its impact on the erythroblast cell production process, i.e.,

$$\frac{\mathrm{d}\rho_t}{\mathrm{d}t} = r_g\rho_t\left(1-\left(1+e^{k_s(k_c-I_t)}\right)^{-1}\right) \quad \text{and} \quad \frac{\mathrm{d}I_t}{\mathrm{d}t} = \frac{\mathrm{d}\rho_t}{\mathrm{d}t} - r_d I_t,$$

where ρ_t and I_t represent the cell density and the inhibitor concentration (i.e., latent state) at time t. The kinetic coefficients $\boldsymbol{\phi} = \{r_g,k_s,k_c,r_d\}$ denote the cell growth rate, the inhibitor sensitivity, the inhibitor threshold, and the inhibitor decay. Then, we construct the hybrid model, i.e.,

$$\rho_{t+1} = \rho_t + \Delta t\cdot r_g\rho_t\left(1-\left(1+e^{k_s(k_c-I_t)}\right)^{-1}\right) + e_t^\rho \quad \text{and} \quad I_{t+1} = I_t + \Delta t\cdot\left(\frac{\rho_{t+1}-\rho_t}{\Delta t} - r_d I_t\right) + e_t^I, \tag{7}$$

where the residuals follow the normal distributions $e_t^\rho \sim \mathscr{N}(0,v_\rho^2)$ and $e_t^I \sim \mathscr{N}(0,v_I^2)$ by applying CLT. Therefore, the hybrid model is specified by parameters $\boldsymbol{\theta} = (r_g,k_s,k_c,r_d,v_\rho,v_I)$. The prediction is made on the interval of three hours $\Delta t = 3$ from 0 to 30 hours (corresponding to time step $t = 1,2,\ldots,11$).

We denote the "true" hybrid model with underlying parameters $\boldsymbol{\theta}^c$. Following Glen et al. (2018), we specify the true mechanistic parameter values as $\boldsymbol{\phi}^c = \{r_g,k_s,k_c,r_d\} = \{0.057,3.4,2.6,0.005\}$. We set the bioprocess noise level $v = v_\rho = v_I$, the initial cell density 3 $\times 10^6$ cells/mL (i.e., $\rho_1 = 3$), and no initial inhibition (i.e., $I_1 = 0$). Based on the simulation data generated by the true hybrid model, we assess the performance of the proposed LG-DBN auxiliary ABC-SMC algorithm under different levels of bioprocess noise $v = \{0.1,0.2\}$ and model uncertainty induced with the different data size, i.e., $m = 3,6,20$ batches.

5.2 LG-DBN Auxiliary Sequential Importance Sampling Performance Assessment

We compare the performance of LG-DBN auxiliary ABC-SMC with naive ABC-SMC in terms of: (1) prediction accuracy, (2) computation time, and (3) posterior concentration. The distance metrics of naive ABC-SMC is $d(\mathscr{D},\mathscr{D}^\star)$. The results are estimated based 30 macro-replications. We set the number of particles $N = 400$, the ratio $\alpha = 0.5$, the number of replications $L = 60$, and the minimal accept rate $P_{acc_{min}} = 0.15$. The prior distributions of model parameters are set as: $r_g \sim U(0,0.5)$, $k_s \sim U(0,5)$, $k_c \sim U(0,5)$, $r_d \sim U(0,0.05)$, $v_\rho \sim U(0,0.2)$, and $v_I \sim U(0,0.2)$.

One of the major benefits induced by the LG-DBN auxiliary likelihood is that it provides an efficient way to measure the distance between simulated and observed samples, which quickly leads to posterior samples fitting well on dynamics and variations. To show the advantage of LG-DBN auxiliary ABC-SMC, we first study its computational efficiency. For each r-th macro replication, let $T_w^{(r)}$ and $T_{wo}^{(r)}$ represent the computation cost of the ABC-SMC algorithm with and without LG-DBN auxiliary. The computational efficiency improvement is evaluated as the time consuming ratio defined as $C^{(r)} = T_{wo}^{(r)}/T_w^{(r)}$. We record the 95% confidence interval (CI) for improvement, denoted by $\bar{C} \pm 1.96 \times S_C/\sqrt{30}$, where $\bar{C} = \frac{1}{30}\sum_{r=1}^{30} T_{wo}^{(r)}/T_w^{(r)}$ and $S_C = [\sum_{r=1}^{30}(T_{wo}^{(r)}/T_w^{(r)} - \bar{C})^2/29]^{1/2}$; see the results in Table 1. With the LG-DBN auxiliary, the ABC-SMC algorithm shows significant improvement in computational efficiency. In all different settings, the mean computation cost of naive ABC-SMC is higher than the LG-DBN auxiliary based ABC-SMC by 27% (at low variance and small sample size) to 163% (at high variance and relative larger sample size).

Table 1: Computational efficiency improvement ratio.

Process Noise	m = 3	m = 6	m = 20
$v = 0.1$	1.27 ± 0.11	1.43 ± 0.11	2.44 ± 0.15
$v = 0.2$	1.39 ± 0.08	1.52 ± 0.17	2.63 ± 0.20

Then, we compare the prediction accuracy of the posterior predictive distribution obtained from ABC-SMC with and without LG-DBN auxiliary. We estimate the parameters $\boldsymbol{\theta} = (r_g,k_s,k_c,r_d,v_\rho,v_I)$. Specifically, in each macro replication, we generate posterior samples $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{N_\alpha}$ by LG-DBN auxiliary and naive ABC-SMC approaches to approximate the posterior predictive distribution,

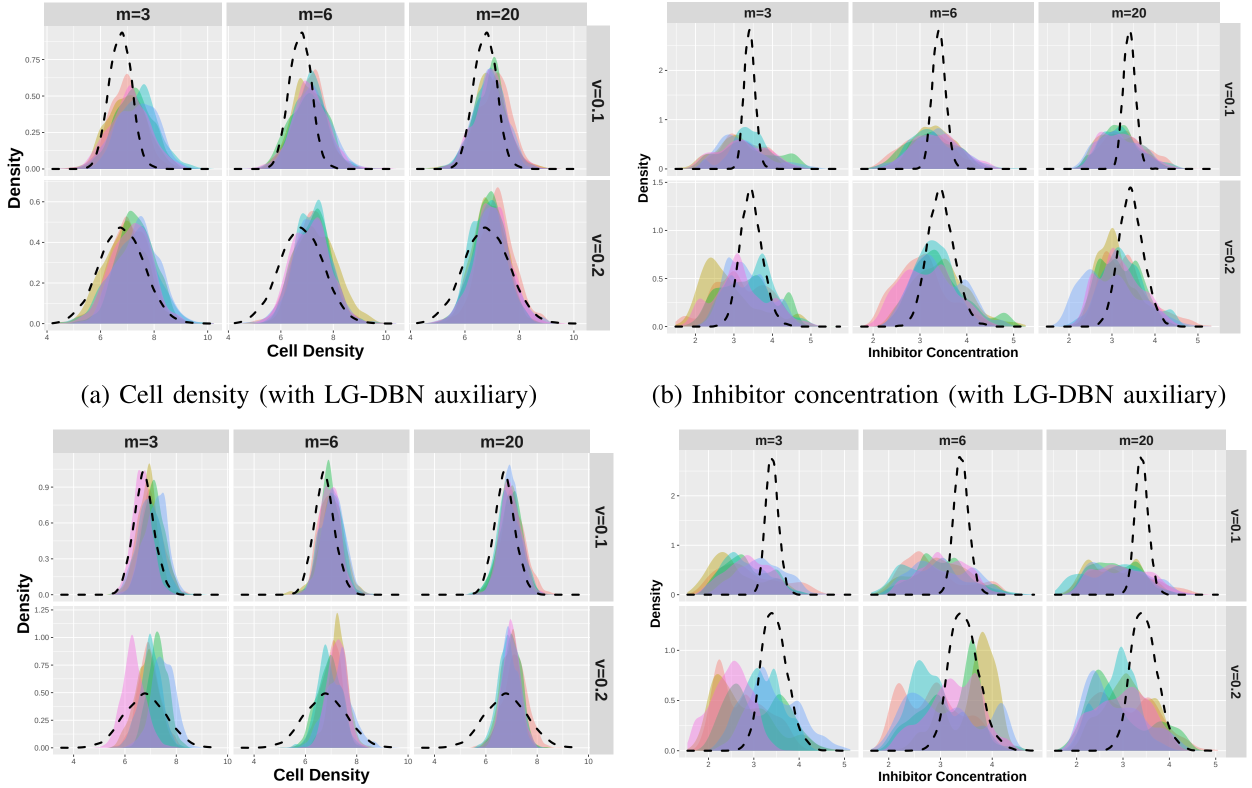
$$p(\rho_t,I_t|\rho_1,I_1,\mathscr{D}) = \int p(\rho_t,I_t|\boldsymbol{\theta},\rho_1,I_1)p(\boldsymbol{\theta}|\mathscr{D})d\boldsymbol{\theta} = \frac{1}{N_\alpha}\sum_{i=1}^{N_\alpha} p\left(\rho_t,I_t|\rho_1,I_1,\boldsymbol{\theta}^{(i)}\right),$$

where the probability density $p(\rho_t,I_t|\rho_1,I_1,\boldsymbol{\theta}^{(i)})$ is computed by the hybrid model (7) for $i = 1,2,\ldots,N_\alpha$. Given the "true" model parameters $\boldsymbol{\theta}^c$, we can also construct the predictive distribution $p(\rho_t,I_t|\rho_1,I_1,\boldsymbol{\theta}^c)$ from the model (7). Figure 1 shows posterior predictive distributions of cell density and inhibitor concentration at the 30-th hour or timestep $t = 11$ given a fixed initial state $(\rho_1,I_1) = (3,0)$. The black dashed line represents the predictive distribution of "true" model $p(\rho_{11},I_{11}|\rho_1,I_1,\boldsymbol{\theta}^c)$.

Table 2: The K-S statistics of cell density and inhibitor accumulation at the 30-th hour (i.e., $t = 11$).

		ABC-SMC with LG-DBN auxiliary			ABC-SMC without LG-DBN auxiliary		
State	Process Noise	$m = 3$	$m = 6$	$m = 20$	$m = 3$	$m = 6$	$m = 20$
ρ_t	$v = 0.1$	0.34 ± 0.04	0.31 ± 0.03	0.25 ± 0.02	0.26 ± 0.05	0.24 ± 0.04	0.23 ± 0.03
	$v = 0.2$	0.25 ± 0.05	0.22 ± 0.04	0.19 ± 0.02	0.36 ± 0.04	0.32 ± 0.03	0.28 ± 0.02
I_t	$v = 0.1$	0.45 ± 0.04	0.46 ± 0.03	0.44 ± 0.02	0.68 ± 0.04	0.69 ± 0.03	0.67 ± 0.02
	$v = 0.2$	0.38 ± 0.05	0.37 ± 0.05	0.36 ± 0.04	0.53 ± 0.07	0.55 ± 0.06	0.56 ± 0.04

By comparing Figure 1(a)-(b) to Figure 1(c)-(d), we observe that LG-DBN auxiliary ABC-SMC shows more robust performance across macro-replications and the posterior predictive distributions are generally closer to the "true" predictive distribution than naive ABC-SMC. We further investigate Panel (a) and (c).



(a) Cell density (with LG-DBN auxiliary)

(b) Inhibitor concentration (with LG-DBN auxiliary)

(c) Cell density (without LG-DBN auxiliary)

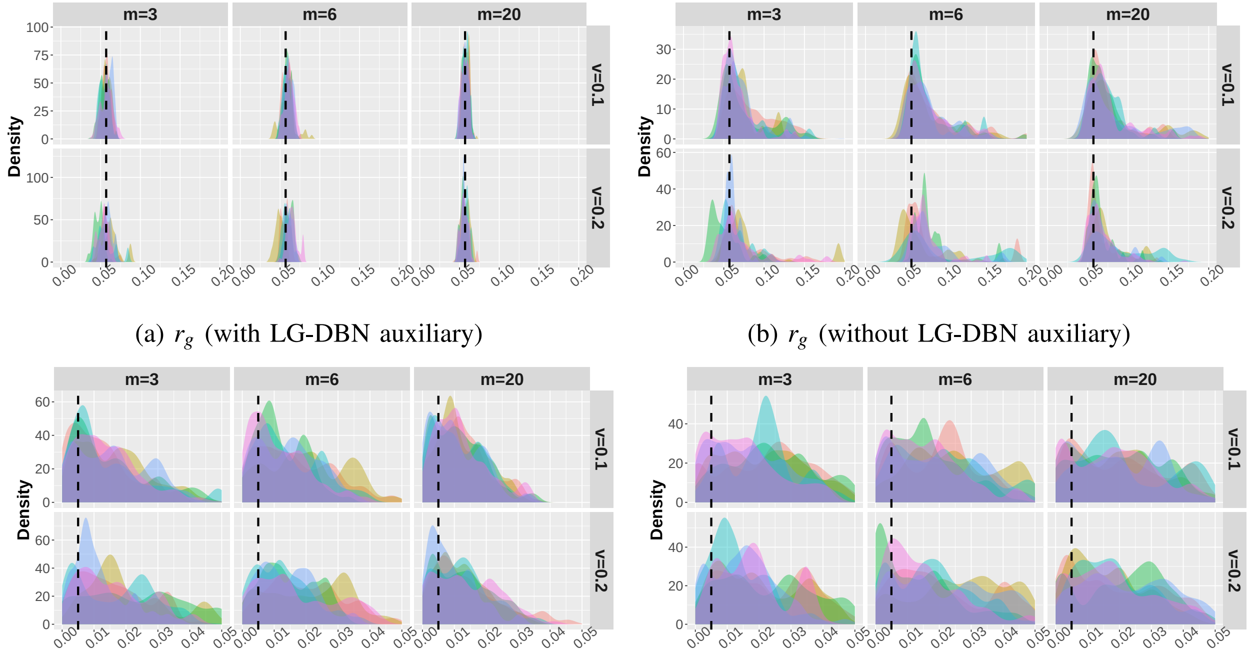
(d) Inhibitor concentration (without LG-DBN auxiliary)

Figure 1: Posterior predictive distributions of cell density and inhibitor concentration at the 30-th hour ($t = 11$) $p(p_t, I_t|\rho_1, I_1)$ obtained from 6 macro-replications (simulated with common random numbers). The color filled areas under the probability density curve represent estimated posterior predictive distributions from different macro-replications. The black dashed line represents the predictive distribution of the "true" model, i.e. $p(\rho_t, I_t|\rho_1, I_1, \boldsymbol{\theta}^c)$. The rows of each panel are related to noise levels (i.e. $v = 0.1, 0.2$) while the columns of each panel are sample sizes of observations (i.e., $m = 3, 6, 20$).

In low noise level $v = 0.1$, the auxiliary based ABC-SMC tends to overestimate the variance v_ρ causing the estimated posterior predictive distributions more flat than the "true" predictive distribution. However, in high noise level, the posterior predictive distribution of LG-DBN auxiliary ABC-SMC is more accurate than that from naive ABC-SMC which consistently underestimates the variance v_ρ. The LG-DBN auxiliary ABC-SMC consistently shows better prediction on inhibitor concentration; see Figure 1(b) and 1(d).

We further use the Kolmogorov–Smirnov(K-S) statistics to assess the performance of LG-DBN auxiliary ABC-SMC and naive ABC-SMC. The K-S statistics quantifies the distance between posterior predictive distribution and predictive distribution of "true" model. The K-S statistics is $D = \sup_s |F^c(s) - F^p(s)|$ for $s \in \{\rho, I\}$, where $F^c(s)$ and $F^p(s)$ are the empirical distribution functions of the samples from predictive distribution of "true" model and posterior predictive distribution respectively. The smaller value of K-S statistic means better approximation performance of posterior predictive distribution. The number of samples used to construct the empirical distribution is $K = 2000$ in each macro-replication. We summarize 95% CIs of distances for both cell density and inhibitor accumulation at the 30-th hour, denoted by $\bar{D} \pm 1.96 \times S_D/\sqrt{30}$ in Table 2, where $\bar{D} = \frac{1}{30}\sum_{r=1}^{30} D^{(r)}$ and $S_D = [\sum_{r=1}^{30}(D^{(r)} - \bar{D})^2/29]^{1/2}$.

As shown in Table 2, the LG-DBN auxiliary ABC-SMC algorithm has better performance in inhibitor concentration prediction – latent state estimation – at all levels of model estimation uncertainty and stochastic uncertainty. It also provides better prediction on cell density under high stochastic uncertainty.



(a) r_g (with LG-DBN auxiliary)

(b) r_g (without LG-DBN auxiliary)

(c) r_d (with LG-DBN auxiliary)

(d) r_d (without LG-DBN auxiliary)

Figure 2: Posterior distributions of r_g and r_d of 6 macro-replications. The posterior distributions estimated by auxiliary based ABC-SMC are shown in Panels (a), (c). The posterior distributions estimated by naive ABC-SMC are shown in Panels (b), (d). The black dashed lines represent the “true” value of parameters.

The results are consistent with the observations obtained from Figure 1. The performance improvement can be further observed from the estimated posterior distribution of hybrid model parameters; see the representative plots of cell growth rate r_g and inhibitor decay rate r_d in Figure 2. The posterior distribution estimated by the LG-DBN auxiliary ABC-SMC has better concentration, defined as the posterior mass around the true parameter (Ho et al. 2020), than naive ABC-SMC in all noise levels and sample sizes.

Notice that due to the structure of the kinetic model in (7) and a small value $r_d^c = 0.005$, the observable state ρ_t is not so sensitive to the changes in the inhibitor decay rate r_d and the inhibitor concentration I_t. Even thought it is more challenging to estimate the latent state I_t and its mechanistic model parameter r_d, the LG-DBN auxiliary ABC-SMC tends to perform better.

In sum, compared with naive ABC-SMC, the proposed LG-DBN auxiliary ABC-SMC algorithm tends to have better prediction accuracy and computational efficiency especially under the situations with high stochastic and model uncertainties. This can benefit bioprocess mechanism learning and robust control.

6 CONCLUSION

To leverage the information from existing mechanistic models and facilitate learning from real-world data, we develop a probabilistic knowledge graph (KG) hybrid model that can faithfully capture the important properties of bioprocesses, including nonlinear reactions, partially observed state, and nonstationary dynamics. Since the likelihood is intractable, approximate Bayesian computation (ABC) sampling strategy is used to generate samples to approximate the posterior distribution. For complex biomanufacturing processes with high stochastic and model uncertainties, it is computationally challenging to generate simulated trajectories close to real-world observations. Therefore, in this paper, we utilize a simple linear Gaussian dynamic Bayesian network (LG-DBN) auxiliary model to design summary statistics for ABC-SMC, which can accelerate Bayesian inference on the probabilistic KG hybrid model with high fidelity characterizing complex bioprocessing mechanisms. The empirical study demonstrates that the proposed LG-DBN auxiliary

ABC-SMC can improve computational efficiency and prediction accuracy. In the future research, we will extend this research to multi-scale bioprocess hybrid model in order to facilitate underlying mechanism learning, support process monitoring, and guide robust control at both cellular and system levels.

REFERENCES

Beaumont, M. A., J.-M. Cornuet, J.-M. Marin, and C. P. Robert. 2009. "Adaptive Approximate Bayesian Computation". *Biometrika* 96(4):983–990.

Del Moral, P., A. Doucet, and A. Jasra. 2006. "Sequential Monte Carlo Samplers". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(3):411–436.

Fuller, W. A., and J. Rao. 1978. "Estimation for a Linear Regression Model with Unknown Diagonal Covariance Matrix". *The Annals of Statistics* 6(5):1149–1158.

Gleim, A., and C. Pigorsch. 2013. "Approximate Bayesian Computation with Indirect Summary Statistics". Technical report, University of Bonn.

Glen, K. E., E. A. Cheeseman, A. J. Stacey, and R. J. Thomas. 2018. "A Mechanistic Model of Erythroblast Growth Inhibition Providing a Framework for Optimisation of Cell Therapy Manufacturing". *Biochemical Engineering Journal* 133:28–38.

Ho, L. S. T., B. T. Nguyen, V. Dinh, and D. Nguyen. 2020. "Posterior Concentration and Fast Convergence Rates for Generalized Bayesian Learning". *Information Sciences* 538:372–383.

Lenormand, M., F. Jabot, and G. Deffuant. 2013. "Adaptive Approximate Bayesian Computation for Complex Models". *Computational Statistics* 28(6):2777–2796.

Martin, G. M., B. P. McCabe, D. T. Frazier, W. Maneesoonthorn, and C. P. Robert. 2019. "Auxiliary Likelihood-based Approximate Bayesian Computation in State Space Models". *Journal of Computational and Graphical Statistics* 28(3):508–522.

Mockus, L., J. J. Peterson, J. M. Lainez, and G. V. Reklaitis. 2015. "Batch-to-batch Variation: A Key Component for Modeling Chemical Manufacturing Processes". *Organic Process Research & Development* 19(8):908–914.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

O'Brien, C. M., Q. Zhang, P. Daoutidis, and W.-S. Hu. 2021. "A Hybrid Mechanistic-empirical Model for in silico Mammalian Cell Bioprocess Simulation". *Metabolic Engineering* 66:31–40.

Sisson, S. A., Y. Fan, and M. Beaumont. 2018. *Handbook of Approximate Bayesian Computation*. Boca Raton, FL: Chapman & Hall/CRC Press.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf. 2009. "Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamical Systems". *Journal of the Royal Society Interface* 6(31):187–202.

Xie, W., B. Wang, C. Li, D. Xie, and J. Auclair. 2022. "Interpretable Biomanufacturing Process Risk and Sensitivity Analyses for Quality-by-Design and Stability Control". *Naval Research Logistics* 69(3):461–483.

Zheng, H., W. Xie, I. O. Ryzhov, and D. Xie. 2021. "Policy Optimization in Bayesian Network Hybrid Models of Biomanufacturing Processes". *arXiv preprint arXiv:2105.06543*. https://arxiv.org/abs/2105.06543, accessed 24th September 2022.

Zheng, H., W. Xie, K. Wang, and Z. Li. 2022. "Opportunities of Hybrid Model-based Reinforcement Learning for Cell Therapy Manufacturing Process Development and Control". *arXiv preprint arXiv:2201.03116*. https://arxiv.org/abs/2201.03116, accessed 24th September 2022.

AUTHOR BIOGRAPHIES

WEI XIE is an assistant professor in the Department of Mechanical and Industrial Engineering (MIE) at Northeastern University. Her research interests include interpretable AI/ML, computer simulation, data analytics, and stochastic optimization for cyber-physical system risk management, learning, and automation. Her email address is w.xie@northeastern.edu. Her website is http://www1.coe.neu.edu/∼wxie/

KEQI WANG is Ph.D. candidate in MIE at Northeastern University. His research interests include machine learning, data analytics, and computer simulation. His email address is wang.keq@northeastern.edu.

HUA ZHENG is Ph.D. candidate in MIE at Northeastern University. His research interests include machine learning, data analytics, computer simulation and stochastic optimization. His email address is zheng.hua1@northeastern.edu. His website is https://zhenghuazx.github.io/hua.zheng/

BEN FENG is an Assistant Professor in actuarial science at the University of Waterloo. His research interests include stochastic simulation design and analysis, optimization via simulation, nonlinear optimization, and financial and actuarial applications of simulation and optimization methodologies. His e-mail address is ben.feng@uwaterloo.ca. His website is http://www.math.uwaterloo.ca/∼mbfeng/.