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.

FROM DISCOVERY TO PRODUCTION: CHALLENGES AND NOVEL METHODOLOGIES FOR NEXT GENERATION BIOMANUFACTURING

Wei Xie

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

Giulia Pedrielli

School of Computing and
Augmented Intelligence
Arizona State University
699 S Mill Avenue, Tempe, AZ 85281, USA

ABSTRACT

The increasingly pressing demand of novel bio-drugs (e.g., gene therapies) with unprecedented levels of personalization, has put a remarkable pressure on the traditionally long time required by the pharma R&D and manufacturing to go from design to production of new products. In fact, practitioners are increasingly moving away from the classical paradigm of large-scale batch production to continuous biomanufacturing with flexible and modular design, which is further supported by the recent technology advance in single-use equipment. In contrast to long design processes, low product variability (one-fits-all), and highly rigid systems, modern pharma players are answering the question: can we bring design and process control up to the speed that novel production technologies give us to quickly set up a flexible production run? In this tutorial, we present key challenges and potential solutions in terms of new bioprocess modeling and control strategies for integrated design and manufacturing of next generation bio-drugs.

1 INTRODUCTION

The rapidly growing biomanufacturing industry plays a significant role in supporting economy and ensuring public health. It has developed various innovative treatments for cancer, adult blindness, and COVID-19 among many other diseases. The industry generated more than $300 billion revenue in 2019 with about 12% annual growth rate (Langer 2020), and more than 40% of the products in the development pipeline were biopharmaceuticals. However, drug shortages have occurred at unprecedented rates, especially in the COVID-19 pandemic. It typically takes many years to discover a new bio-drug and develop the optimal production and delivery processes. The current manufacturing systems are unable to rapidly produce new drugs when needed when there is major public health issue.

Biopharmaceutical manufacturing and delivery processes face critical challenges, including high complexity, high variability, lengthy lead time, and limited process observations.

(1) **Long discovery processes.** Drug discovery is a highly experimental, expensive, complex and long process. Labs take up to a year to generate a potential design for few alternative versions of a drug (5$\sim$ 6). Algorithms can help in improving the quality and reducing the time to discovery, but substantial requirements are introduced in terms of data collection (e.g., collect information on failed experiments) to effectively adopt traditional data-driven optimization techniques.

(2) **The discovery process is separated from the production.** The process of drug design and the experiments are usually performed with processes that are completely separated from the actual manufacturing pipelines used later on. This is justified in traditional bioproductions where manufacturing systems are highly rigid, expensive and can only be used for large volume productions. But novel systems (e.g., continuous manufacturing, single use technologies) allow for low volume

2022 Winter Simulation Conference (WSC) | 978-1-6654-7661-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/WSC57314.2022.10015464

productions, so we could *move from experiments-based drug discovery to production-driven drug discovery*. This would lower the "failure at manufacturing", i.e., the scenario in which a drug passes the experimental phase but results impossible/very hard to manufacture and it is therefore rejected, thus triggering a new discovery process with consequent delays and costs.

(3) **High complexity and high uncertainty.** Biomanufacturing process consists of numerous unit operations (such as fermentation, purification, formulation, and delivery). Biotherapeutics are produced in cells (or other living organisms) whose biological processes are complex, and highly variable outputs depend on complex dynamic interactions of many factors. The upstream fermentation can impact on downstream purification cost and productivity. New biotherapeutics (e.g., cell, gene, RNA, and protein therapies) require more advanced manufacturing protocols. For example, aspirin, a classical small molecule medicine is comprised of 21 atoms, whereas many of the antibodies (mAbs) protein drug substances are comprised of greater than 25,000 atoms. The drug size is correlated to the structural complexity of biopharmaceuticals. The molecular drug structure affects its function. In addition, the dynamic interactions of hundreds of factors can impact drug quality, yield, and production cycle time. The target protein and RNA can further degrade and have modifications during manufacturing and delivery processes. *Thus, the bioprocess is the product.*

(4) **Very limited process observations.** The analytical testing time required by biopharmaceuticals of complex molecular structure is lengthy. Also, significant changes in the manufacturing process, such as new facilities, equipment, and raw materials, will typically trigger new regulatory requirements and clinical trials.

Considering all these challenges of complex biomanufacturing process design and control, human error is frequent, accounting for 80% of deviations (Cintron 2015). *It is increasingly realized by the biopharma industry that optimization, machine learning, and simulation approaches, that can incorporate physics-based and experiments supported knowledge, are a key to the next generation of products and processes.* This is because these stochastic system modeling, analytics, and optimization methodology strategies can accelerate integrated and intensified manufacturing process automation, quality-by-design (QbD), and reduce human error, thus projecting biomanufacturing in Industry 4.0.

Connecting drug discovery with manufacturing. Discovery and design of biological drugs and the associated, modular, production processes has the potential to dramatically reduce the currently prohibitive lead time from discovery to process design and manufacturing, while also increasing the quality, and reliability, of the final product.

This integration is needed more than ever. Major technological developments are already enabling future bioproductions of large quantities of small volume and highly personalized products. One of these advancements are "single use technologies", which encompass a range of products and technologies such as single-use disposable connectors, vessels, mixers, etc, which in turn enable fully automated and enclosed processes. They can support flexible and efficient manufacturing at scale and on-demand through reducing (1) sterilization and cleaning costs, (2) contamination incidents, (3) storage needs, and (4) process downtime; see Sandle (2018). Therefore, single use technologies have the potential to impact existing medium to large volume biomanufacturing processes by enabling flexible manufacturing with modular design while reducing costs. The demand for such flexibility and variability in production batches (from few liters to tens of thousands) is already testing the capacity of pharmaceutical Contract Manufacturer Organizations (CMO). For example, several research labs and startups of varying size are seeking the manufacturing capacity to produce vaccines for trials (Kyle Blankenship for FiercePharma 2020). In addition, single use technologies enable for the first time practical small volume bio-productions for personalized therapies, e.g., for cancer treatments (Mock et al. 2016; Zhu et al. 2017).

Limitations of state-of-art OR/OM approaches. Operation Research and Operations Management (OR/OM) methodologies can facilitate drug discovery, biomanufacturing system design and analysis, simulation model calibration (Lee et al. 2019; Arendt et al. 2012; Wang et al. 2019), sensitivity and uncertainty analyses (Saltelli et al. 2008; Baroni and Tarantola 2014). Mathematical programming methods

(Leachman et al. 2014), Markov decision process and stochastic optimization (Kulkarni 2015; Martagan et al. 2018) are developed to guide bioprocess decision making and optimization. *However, existing OR/OM methodologies on process modeling, analytics, and optimization are general and they often fail to consider the physics underlying the specifics of drugs as well as the mechanisms activated during bioprocesses. These limitations hinder the application of these methodologies in practice.*

In this paper, we first review the challenges and opportunities for novel algorithms in design and discovery of new biological drugs (e.g., mRNA vaccines, protein therapies) in Section 2. Specifically, we discuss the molecular structure (folding) and functionality prediction with focus on Ribonucleic acids (RNAs), proteins, and nanoparticle delivery systems. Subsequently, Section 3 presents a probabilistic knowledge graph (KG) hybrid modeling framework that can leverage the information from existing mechanistic models and facilitate learning from real-world data. Built on the hybrid model characterizing the risk- and science-based understanding on bioprocessing mechanisms, we present the risk, sensitivity, and predictive analyses to support interpretable and robust decision making. Finally, we describe the control framework that can leverage these models. The presented reinforcement learning approaches account for both inherent stochasticity and model uncertainty, to facilitate process development and control in Section 3.3. To close the tutorial, Section 4 summarizes the key challenges in the biopharmaceutical industry from discovery to production and discuss the key OR methods and tools that are needed to be developed.

2 THE MECHANICS OF DESIGN MOLECULAR FOLDING AND NANO-PARTICLES

Section 2.1 focuses on the prediction of the folding configurations. We present methods for both the secondary (2D) and tertiary (3D) RNA structure, which directly impacts on RNA drug delivery and function. We also briefly review the structure prediction for proteins associated with dominant bio-drugs in the current biopharmaceutial industry and market. In Section 2.2, we investigate predicting the stability of peptides and nano-particles. Nano-particle formulations have rapidly emerged as carriers for nucleic-acid therapies (i.e., DNA and RNA) to increase cellular uptake, support RNA delivery, and improve drug stability.

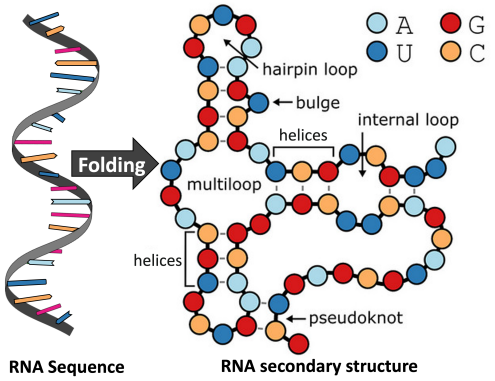

Figure 1: RNA chain folding motifs.

2.1 Algorithms for Structure Prediction

Ribonucleic acid (RNA) is a fundamental biological macromolecule, essential to all living organisms, performing a versatile array of cellular tasks including information transfer, enzymatic function, sensing, regulation, and structural function. RNA has recently emerged as a promising drug target, with new therapeutic approaches aiming to develop drugs that target RNA rather than proteins. Moreover, designed RNA molecules are used in rapidly growing fields of synthetic biology and RNA nanotechnology, with applications to diagnostics, immunotherapy, drug delivery and realization of logical operations inside cells; see for example Han et al. (2017). In Liu et al. (2022), we propose a new framework, ExpertRNA, for the automatic folding of non-pseudoknotted secondary structures for RNA molecular compounds. ExpertRNA builds upon the fortified rollout algorithm and generalizes the architecture to allow for the consideration of multiple experts that can evaluate, at each iteration, the solutions generated by the base heuristic.

RNA structure determines function. Each RNA molecule is made up of a sequence of individual units, nucleotides (bases), which are of four common types, A, U, G and C (Figure 1). Individual RNA sequences range in length from tens (tRNAs, siRNAs) to tens of thousands (viral genomes, long non-coding RNAs) and many contain further chemical modifications of the individual bases (Carell et al. 2012). *While identity is defined by sequence, the function of an RNA molecule is determined by its structure, i.e., the*

way nucleotides interact in space. Biochemists often break down RNA structure into four categories: **(1)** Primary structure refers to the sequence. **(2)** Secondary structure makes up the majority of the bonds in the structure and includes the "canonical base pairs", where A pairs with U and G pairs with C, and the "wobble base pair", where G pairs with U. This provides a 2D representation of the structure of the molecule and is the most commonly used level. **(3)** Tertiary structure defines 3D contacts via weaker, non-conical interactions. **(4)** Finally quaternary structure includes intermolecular interactions with other RNA molecules. Given the impact of structure on RNA functionality, the accurate computational prediction of the secondary and tertiary structure of RNA is an ongoing area of great interest in the computational biology community (Calonaci et al. 2020).

RNA secondary structure prediction. Most tools for secondary structure prediction (Reuter and Mathews 2010; Zadeh et al. 2011) attempt to identify the structure that minimizes the free energy (FE) associated with the RNA molecule upon pairing a subset of the nucleotides, i.e., the energy released by folding a completely unfolded RNA sequence. The underlying assumption is that the structure with the lowest free energy is also the most likely structure the RNA will adopt. Equivalently, this family of approaches relies on the basic idea that the lower the FE the more stable the RNA structure will be. A *first* challenge for this family of approaches is that it is not possible to exactly calculate the free energy due to the (i) incomplete understanding of the RNA molecular interactions, and (ii) the impractical computational cost of detailed kinetic simulation tools. As a result, several approximate models have been proposed in the literature (Andronescu et al. 2010) to estimate the free energy associated with a given secondary structure. Most of the computational savings are a result of ignoring tertiary interactions. A *second*, and possibly deeper, challenge is that this model assumes that an "optimal" structure is one that pairs nucleotides in a way that minimizes the free energy (MFE). However, RNA is known to fold cotranscriptionally (Angela et al. 2021), i.e., the simultaneous transcription of two or more genes. Equivalently, RNA molecules might adopt a kinetically-preferred structure different from the global free energy minima.

In light of these challenges, alternative approaches to structure prediction have been proposed. *Stochastic kinetic folding algorithms* (Sun et al. 2018) approximate the folding kinetics of RNA molecules as they are transcribed. Data driven approaches have also started to become popular that use machine learning to evaluate structures rather than FE or kinetic models. These include ContraFold, DMfold, and structure prediction with neural networks (Calonaci et al. 2020). Furthermore, in the attempt to achieve advantages of model or data driven approaches, methods have been proposed that attempt to aggregate multiple information sources to get more accurate secondary structure prediction. Within the data driven category, some examples of information sources are the experimentally determined SHAPE data (Lucks et al. 2011), and evolutionary covariation information (Calonaci et al. 2020). On the model driven side, statistical ensemble approaches are used to boost the solutions obtained by the different FE driven folding algorithms. To the knowledge of the authors, ensemble methods allow to mix solutions from different algorithms only upon completion, i.e., they do not enable interaction among the algorithms while they are running (Aghaeepour and Hoos 2013). A recent survey on a set of secondary structure prediction tools has reported mixed results, with data-driven approaches generally outperforming the ones based on nearest-neighbor free energy models, and with model-based ensemble approaches showing competitive results (Wayment-Steele et al. 2020).

Tertiary Structure Prediction. Concerning the tertiary structure prediction problem, fewer approaches can be found in the literature (Watkins et al. 2020). In fact, the prediction of tertiary structures is particularly challenging, and most prediction methods only work for short RNA sequences (tens of nucleotides). Data driven approaches have attracted attention also for the tertiary structure prediction. However, their accuracy remains limited due to the small number of 3D RNA structure data sets available for model training and verification.

The ExpertRNA framework for RNA structure prediction. Stemming from the observation that several folding algorithms have been proposed in the literature for secondary and, even if fewer, for tertiary structure prediction (Watkins et al. 2020), without any approach dominating the other, we propose the idea

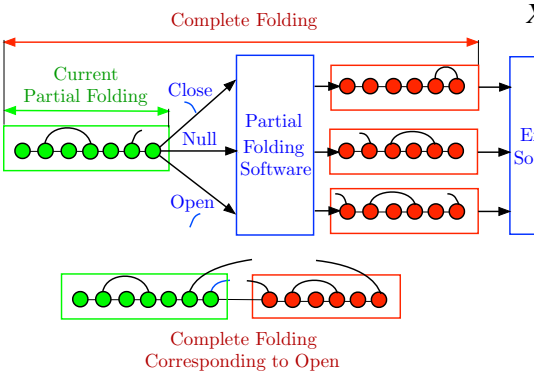

Figure 2: Overview of the ExpertRNA algorithm (source in Liu et al. (2022)).

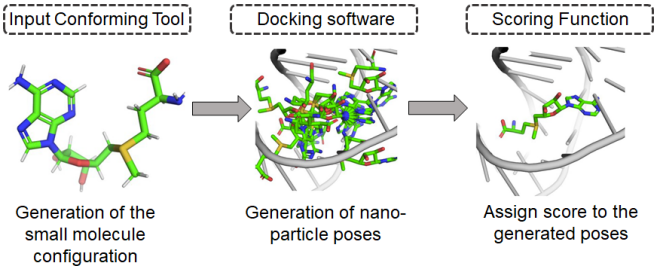

Figure 3: Components for nano-particle docking.

to build a framework that can exploit several folding tools and criteria to evaluate the quality of a folded sequence, during the algorithm execution (Liu et al. 2022). The aim of our approach is to achieve a better RNA structure prediction quality. *Figure 2 shows the structure of our ExpertRNA approach with its two main algorithmic components: (i) the partial folding; and (ii) the expert software.* To mimic interatomic interactions, the algorithm sequentially adds elements to the incomplete structure ("current partial folding" in Figure 2), which we initialize to be the empty set. The first nucleotide is chosen as the first element of the input sequence provided by the user. At each step, the subsequent nucleotide is selected, and we can choose whether to simply *sequence* it to the last assigned nucleotide ("Null" action in Figure 2) or pair it with any nucleotide in the existing structure ("Close" in Figure 2), or pair it with an element still to be assigned ("Open" in Figure 2). The definition of these actions is motivated by the physical laws that govern molecular bonding (as previously specified in feasibility determination).

Algorithms for protein structure prediction Several approaches have been proposed in the literature for protein structure prediction. Also in the case of proteins, we distinguish primary, secondary, and tertiary structure prediction. It is important to predict protein structure due to the implications in medicine as well as biotechnnology. Several algorithms have been proposed in the area with an increasing push into deep learning mainly justified by the exhaustive data sets freely available for proteins. Given different folding software to allow constraints to be provided by our method, we investigate how to embed existing tools such as PEP-FOLD and variants (Lamiable et al. 2016), AWSEM (Jin et al. 2020), Rosetta (Chaudhury et al. 2010), and compare to the Maestro software from Schrodinger LLC. Importantly, once every two years the Critical Assessment of protein Structure Prediction (CASP) experiments are held to assess the state of the art in the field in a blind fashion, by presenting predictor groups with protein sequences whose structures have been solved but have not yet been made publicly available. DeepMind's entry, AlphaFold, placed first in the Free Modeling (FM) category, which assesses methods on their ability to predict novel protein folds (the Zhang group placed first in the Template-Based Modeling (TBM) category, which assess methods on predicting proteins whose folds are related to ones already in the Protein Data Bank) (AlQuraishi 2019). DeepMind's success generated significant public interest. Their approach builds on two ideas developed in the academic community during the preceding decade: (i) the use of co-evolutionary analysis to map residue co-variation in protein sequence to physical contact in protein structure, and (ii) the application of deep neural networks to robustly identify patterns in protein sequence and co-evolutionary couplings and convert them into contact maps (Marx 2022).

2.2 Predicting Stability of Nano-particles for RNA Formulation and Delivery

When nanoparticles as RNA delivery system are of interest, a foundational question arises related to the docking of multiple molecules; see the illustration in Figures 3 and 4. In case the molecules are of the same family (e.g., peptides with peptides, RNAs with RNAs), the problem can be brought back to the folding prediction approaches in Section 2.1. However, in the case of assembly of heterogeneous bodies

(i.e., peptides with RNA), new challenges arise. In this case, we normally refer to the problem of *docking* of molecules, a molecular modeling which can estimate the preferred orientation of one molecule to a second and further predict the type of signal produced and the strength of binding affinity between two molecules using scoring functions. The challenges in docking are very different from those identified in folding. Molecular docking processes are typically composed of two steps and usually a large molecule and a small molecule are considered for binding (Sellami et al. 2021): (1) The conformation of the small molecule (e.g., ligand, peptide, lipid) is predicted together with the orientation and position with respect to the binding site of the larger molecule (e.g., protein, DNA, RNA). Such location and conformation is commonly referred to as the *pose* of the nano-particle. The quality of the pose requires assessment. Such evaluation can be performed by a scoring function. Ideally, a scoring function should be capable to recover the experimental *binding* (true) and rank it highest among all the solutions proposed by the sampling algorithm. (2) A second, but especially challenging task would be to attempt the scoring of active vs. inactive compounds. This task is rarely accomplished due to the interplay of several factors that are external to the nano-particle.

We focus on the search algorithms that have been designed to *efficiently* predict the docking pose. The process of docking a target and a small molecule falls into the class of NP-hard problems due to the non countability of the number of possible poses. Hence, search becomes the approach to identify candidate solution and improving on those. In the literature, search methods can be classified into deterministic (also referred to as systematic) and stochastic (Stanzione et al. 2021). *Systematic search methods* sample within the binding molecule search space at predefined intervals and are deterministic. Within this class, we can still differentiate between exhaustive, fragmentation or conformational ensemble methods. The main difference between them is in the approach they take to deal with the binding molecule flexibility. In exhaustive search methods, for example, the docking is performed by systematically rotating all possible rotatable bonds in the binding molecule at a given interval. The drawback of this family of approaches is computational in nature as the number of possible combinations to consider goes with the number of rotatable bonds in the the binding molecule. An common exhaustive sampling method is Glide presented in Friesner et al. (2004). Fragmentation represents an attempt to improve on the computational efficiency by incrementally forming binding over fragments that the binding molecule is divided into. An approach that relies on fragmentation is FlexX (Rarey et al. 1996). Finally, in conformational ensemble methods, the binding molecule flexibility is represented by rigidly docking an ensemble of pre-generated conformations, thus improving the approach efficiency since using this approach removes the computational cost due to the exploration of the conformational space.

In *stochastic algorithms*, the binding molecule orientations and conformations are sampled by making changes to the molecule that are informed by random score values iteratively generated by a random algorithm. The orientation, conformation change is then treated as a incumbent that is accepted or rejected according to an algorithm-dependent criterion. The advantage of stochastic algorithms is that they can generate large ensembles of molecular conformations and explore a broader range of the energy landscape increasing the probability of finding a global energy minimum. However, this comes at computational cost. Examples are genetic algorithm, Monte Carlo, ant colony optimization (ACO) and tabu search methods. GOLD in (Jones et al. 1997) uses a genetic algorithm, DockVision (Hart and Read 1992) is an example of docking program using Monte Carlo stochastic method where the probability to accept a random change is calculated by using the Boltzmann probability function. An example of ACO-based approach is PLANTS (Korb et al. 2006), while PSI-DOCK uses a tabu search (Pei et al. 2006).

Data Sets. Search algorithms usually rely on data sets to retrieve potential components of the nano-particle to be assembled. In these datasets, the crystal structures of the complexes are specified using different microscopy technologies (with potentially different resolutions). Prior to any docking study, the virtual compounds database that will be screened must be carefully selected and prepared. This compound collection, often referred to as the *virtual library*, can encompass up to millions of compounds. Already prepared virtual libraries can be used, but users can also generate their own. Commercial databases represent

an important source of compounds for virtual screening (often containing more than 1 million molecules). Suppliers commonly offer free access to the files containing molecules structures in several formats (2D or 3D). These databases undergo frequent updates, new products being added while other being either removed or out of stock. Virtual screening libraries constructed from these databases should ideally be prepared when the whole virtual screening protocol is settled and ready to be used.

A particular category of databases are *bioactivity databases* providing knowledge on biological targets and their modulation mechanisms. These data are frequently used for data-set benchmarking in the context of docking protocols design prior to prospective virtual screenings and to construct predictive models of activity (Mysinger et al. 2012). Examples of databases in this category are ZINC (Sterling and Irwin 2015) or PubChem (Kim et al. 2019).

Finally, *natural products data bases* are available. Natural products were the first drugs ever used and have always been a source of drugs. In a recent retrospective study, it has been reported that, between January 1, 1981 and September 30, 2019, 23.5% of all new approved drugs and 33.6% of new approved small molecules drugs were natural products or derivatives of natural products (Newman and Cragg 2020). The chemical space covered by natural products is quite dissimilar to the one occupied by synthetic drug-like compounds (Morrison and Hergenrother 2013) and natural products are believed to constitute promising starting point for drug discovery (Rodrigues et al. 2016). Hence, these data sets can represent a source for virtual screening libraries. Numerous databases of this type are available; please refer to Sorokina and Steinbeck (2020).

Scoring Functions. There are three important applications of scoring functions in molecular docking (Huang et al. 2010). The *first* of these is the determination of the binding mode and site of small molecule to its target. Specifically, once a target has been defined, molecular docking generates hundreds of thousands of possible binding orientations/conformations for the small molecule (e.g., ligand, peptide) at the active site around the target molecule. The scoring function is used to rank such small molecule orientations/conformations by evaluating the binding tightness of each of the candidate complexes. Ideally, the scoring function would rank the highest the experimentally determined binding mode. Given the determined binding mode, scientists would be able to gain a deep understanding of the molecular binding mechanism and to further design an efficient drug by modifying either the target or the small molecule. It is important to highlight that, instead of scoring functions, other computational methodologies based on molecular dynamics or Monte Carlo simulations could be used to model the dynamics of the binding process thus resulting in a more accurate prediction of binding affinity. However, these models result in computationally prohibitive free energy calculations, ultimately resulting impractical for the evaluation of large numbers of molecular complexes. As a result the application of high fidelity simulation techniques is generally reduced to predicting binding affinity in small nano-particles (Ain et al. 2015).

The *second application* of a scoring function, related to the first, is the prediction of the *absolute binding affinity* between the active compounds. This is particularly important in lead optimization, i.e., the process to improve the tightness of binding for low-affinity hits or lead compounds that have been identified. In this phase, an accurate scoring function can greatly increase the optimization efficiency and save costs by computationally predicting the binding affinities between the conformed small molecule and the target before the much more expensive synthesis and experimental steps.

The *third application*, is related to the foundational task of structure-based design, that fundamentally attempts to identify the potential drug hits/leads for a given target by searching a large compound database, this is commonly referred to as *virtual database screening*. A reliable scoring function should be able to associate higher rank to known binders following their binding scores during database screening. In fact, due to the expensive cost of experimental screening and sometimes unavailability of high-throughput assays, virtual database screening has played an increasingly important role in drug discovery.

Classical scoring functions can be classified into three groups: forcefield, knowledge-based and empirical (Ballester and Mitchell 2010). An alternative to the classical approach to the design of scoring functions, a non-parametric machine-learning approach can be taken to implicitly capture the binding interactions that

are hard to explicitly model by classical approaches in a computationally efficient way. By not imposing a particular functional form for the function, the collective effect of intermolecular inter-actions in binding can be directly inferred from experimental data, which should lead to scoring mechanisms that are characterized by greater generality and, consequently, prediction accuracy.

3 BIOMANUFACTURING HYBRID MODELING AND ANALYTICS

Driven by the critical challenges and needs of biopharmaceutical manufacturing, we propose the probabilistic knowledge graph (KG) hybrid model characterizing the spatial-temporal causal interdependencies of critical process parameters (CPPs) and critical quality attributes (CQAs) (Zheng et al. 2022; Xie et al. 2022). To faithfully represent underlying bioprocessing mechanisms, this hybrid model can capture the important properties, including nonlinear reactions, partially observed state, and nonstationary dynamics. It has time-varying kinetic coefficients (such as molecular reaction rates) with uncertainty representing batch-to-batch variation. To avoid the evaluation of intractable likelihood, we further investigate a computational Bayesian inference approach, called Approximate Bayesian Computation (ABC), to efficiently approximate the posterior distribution. Then, assisted by the Bayesian KG, accounting for inherent stochasticity and model uncertainty, we present *interpretable* risk, sensitivity, and predictive analyses. This study can support: (1) biomanufacturing stochastic decision process (SDP) mechanism online learning; (2) bioprocess soft sensor monitoring (such as tracking latent metabolic state associated with therapeutic cell function and product quality); and (3) optimal and robust process control.

Illustration Example: mRNA lipid nanoparticle formulation process. Here we use mRNA lipid nanoparticle (mRNA-LPN) formulation process as an illustration example; see Figure 4. Lipid-based formulations have rapidly emerged as carriers in nucleic-acid therapies (i.e., DNA and RNA) to increase cellular uptake, support RNA delivery, and improve drug stability. The mRNA lipid nanoparticle (LNP) formulation utilizes microfluidic or T-junction mixing to rapidly combine an ethanol phase containing hydrophobic lipids and an aqueous phase containing mRNA in a buffer. Then, the self-assembly of LNPs with mRNA is driven by the hydrophobic and electrostatic force field that is influenced by the design of lipids, the selection of LNP formulation and CPPs, as well as the phases of mRNA-LNP complex. The pH-responsive ionizable cationic lipids have the surface charge modulated, controlling the efficient binding with the oppositely charged polynucleotides, which will support self-assembly, influence mRNA-LNP thermodynamic stability, prolong the circulation time of mRNA-LNP complexes, facilitate endosomal escape and mRNA release, and increase their therapeutic. Therefore, the dynamics and variations of mRNA-LNP formulation process output trajectory depends on complex interactions of (1) the design of lipids and (2) CPPs (e.g., flow rate ratio, total flow rate, temperature, lipid choices, buffer choices), which impacts on the quality of LNP delivery system specified CQAs including a) RNA integrity level; b) species composition/concentrations/phases, particle size distribution, zeta potential, surface charge; and c) mRNA-LNP, bound/unbound mRNA.

In Section 3.1, we present the macro-phases or operation units involved in the manufacturing of a bioproduct, while Sections 3.2 and 3.3 illustrate novel methods developed for the efficient modeling, analytics, and control of such process when targeting novel synthetic products (e.g., variations of the mRNA-LPN).

3.1 Operation Unites of a Biomanufacturing process

Biopharmaceutical manufacturing process is crucially important to determine product quality and productivity. It typically includes the main unit operations listed below. (1) **Fermentation and Drug Substance Synthesis**: Living organisms (e.g., cells, yeasts) are mixed with appropriate medium and enzymes under carefully controlled conditions to grow and synthesize the target drug substance. The byproducts or unwanted impurities are also generated at the meantime, which impacts downstream purification operation and cost. During different growth and production phases of cell and yeast life cycle, different media compositions and

feeding strategies are used to improve the productivity and reduce the waste generation. (2) **Centrifugation(s)**: The centrifuge device is used for separation of particles, e.g., cells, subcellular organelles, viruses, large molecules such as proteins, from a solution according to their size, shape, density, viscosity of the medium and rotor speed, during which bulk of impurities would be removed. (3) **Chromatography and Purification**: As the mixture of solutes flows through a packed resin bed, specific solutes are separated as they are bound. Chromatography serves as the most critical part for purification, and it usually determines the purity of product. Since removing more impurities often results in removing more drug substance during chromatography step, there is often a trade-off between productivity and purity. (4) **Filtration**: It is applied at several stages for capture (i.e., concentrate the product), intermediate purification, and polishing (i.e., eliminates trace contaminants and impurities) purpose. (5) **Formulation and Filling**: To maintain the safety and efficacy of the drug substance during the storage, delivery, and facilitate the patient absorption of active pharmaceutical ingredient (API), the purified drug substance is usually formulated with carefully selected excipients and nanoparticle carriers into stable drug products and filled into dose containers. (6) **Freeze Drying**: It is used to stabilize bio-drugs through removing water or other solvents from the frozen matrix and converting the water directly from solid phase to vapor phase through sublimation. Freeze drying is critical for immobilizing the bio-drug product in storage and delivery, as the kinetics of most chemical and physical degradation reactions are significantly decreased. (7) **Quality Assurance/Control (QA/QC)**: QA and QC are performed to ensure the quality of the raw material selection, the production process, and the final bio-drug product. In sum, Step (1) belongs to upstream fermentation and drug substance synthesis, Steps (2)–(4) belong to downstream purification, and Steps (5)–(7) are for finished drug filling/formulation, freeze drying, and product quality control testing.

There are interactions of hundreds of factors at different productions steps impacting drug quality, yield and production cycle time. These factors can be divided into critical process parameters (CPPs) and critical quality attributes (CQAs) in general; see the definitions of CPPs/CQAs in ICH-Q8R2 (Guideline, ICH Harmonised Tripartite and others 2009).

CPP: At each process unit operation, CPPs are defined as critical process parameters whose *variability* impacts on product CQAs, and therefore should be monitored and controlled to ensure the process produces the desired quality.

CQA: A physical, chemical, biological, or microbiological property that should be within an appropriate limit, range, or distribution to ensure the desired product quality.

3.2 KG Hybrid Model for Biomanufacturing Stochastic Decision Process

The probabilistic KG hybrid model proposed in Zheng et al. (2022) and Xie et al. (2022) can provide the risk- and science-based understanding of underlying stochastic decision process (SDP) mechanisms for controlled production processes. The input-output relationship in each step is modeled by *a hybrid ("mechanistic+statistical") model* that can leverage the prior knowledge on biophysicochemical mechanisms from existing mechanistic models and further advance scientific learning from process data. Specifically, at any time t, the process state $\boldsymbol{s}_t$ (such as CQAs) composed of observable and latent state variables, i.e., $\boldsymbol{s}_t = (\boldsymbol{x}_t, \boldsymbol{z}_t)$ (e.g., particle size distribution, RNA sequence and integrity level), and CPPs action $\boldsymbol{a}_t$ (e.g., temperature, mixing flow rate, pH) interactively influence on the dynamics and variations of output trajectories (e.g., mRNA-LNP formulation and self-assembling processes). Given the existing nonlinear ODE/PDE-based mechanistic model (such as biomolecular dynamics, thermodynamics, molecular interactions of mRNA and lipids which can affect the RNA integrity), represented by $\mathrm{d}\boldsymbol{s}/\mathrm{d}t = \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\beta})$, by applying the finite difference approximations on derivatives, we construct the hybrid model for state transition,

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{x}_t + \Delta t \cdot \boldsymbol{f}_x(\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{a}_t; \boldsymbol{\beta}_t) + \boldsymbol{e}^x_{t+1}, \\ \boldsymbol{z}_{t+1} &= \boldsymbol{z}_t + \Delta t \cdot \boldsymbol{f}_z(\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{a}_t; \boldsymbol{\beta}_t) + \boldsymbol{e}^z_{t+1}, \end{aligned}$$

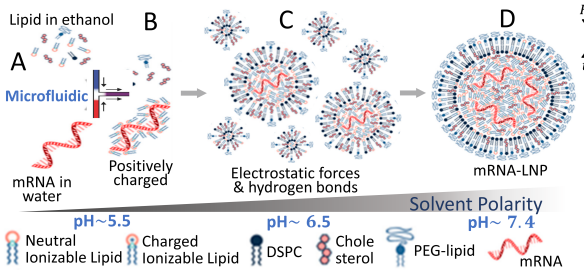

Figure 4: An illustration of mRNA lipid nanoparticle formulation and self-assembly process (figure adapted from Buschmann et al. (2021)).

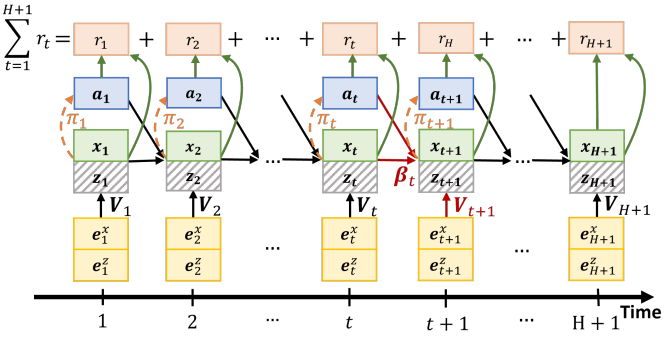

Figure 5: Policy-augmented KG network for SDP.

with unknown random kinetic coefficients $\boldsymbol{\beta}_t \in \mathbb{R}^{d_\beta}$ (e.g., particle clustering rates) accounting for batch-to-batch variation. The function structures of $\boldsymbol{f}_x(\cdot)$ and $\boldsymbol{f}_z(\cdot)$ are derived from $\boldsymbol{f}(\cdot)$ in the mechanistic models. The residual terms are modeled by multivariate Gaussian distributions $\boldsymbol{e}^x_{t+1} \sim \mathcal{N}(0, V^x_{t+1})$ and $\boldsymbol{e}^z_{t+1} \sim \mathcal{N}(0, V^z_{t+1})$ with zero means and covariance matrices V^x_{t+1} and V^z_{t+1} by applying CLT. The kinetic coefficients $\boldsymbol{\beta}_t$ can change cross different phases of bioprocess accounting for the fact that the process dynamics can be time-varying. The statistical residual terms $\boldsymbol{e}_t = (\boldsymbol{e}^x_t, \boldsymbol{e}^z_t)$ allow us to account for the impact from bioprocess noises, raw material variations, ignored CPPs, sensor measurement errors, and other uncontrollable factors (e.g., contamination) occurring at any time step t.

The probabilistic KG of integrated biomanufacturing process can be visualized by a directed network as shown in Figure 5. The observed state variables $\boldsymbol{x}_t$ and the latent state variables $\boldsymbol{z}_t$ are represented by solid and shaded nodes respectively. The directed edges represent causal interactions. At any time period $t+1$, the process state output node $\boldsymbol{s}_{t+1} = (\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1})$ depends on its parent nodes: $\boldsymbol{s}_{t+1} = \boldsymbol{f}(Pa(\boldsymbol{s}_{t+1}); \boldsymbol{\theta}_t)$ with $Pa(\boldsymbol{s}_{t+1}) = (\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{e}_{t+1})$ and model parameters denoted by $\boldsymbol{\theta}_t$. To represent the underlying controlled SDP, we create a *policy augmented KG network* by including additional edges: 1) connecting state $\boldsymbol{s}_t$ to action $\boldsymbol{a}_t$ representing the causal effect of the policy, $\boldsymbol{a}_t = \pi_t(\boldsymbol{s}_t|\boldsymbol{\phi})$ specified by parameters $\boldsymbol{\phi}$; and 2) connecting actions and states to the immediate reward $r_t(\boldsymbol{s}_t, \boldsymbol{a}_t)$ (e.g., cost and RNA production). *This KG network models how the effect of current state and action, $\{\boldsymbol{s}_t, \boldsymbol{a}_t\}$, propagates through mechanism pathways impacting on the output trajectory and the accumulated reward.*

Bayesian KG representing risk- and science-based mechanism understanding. Leveraging the information from existing mechanistic models and heterogeneous online/offline measurements, the probabilistic KG hybrid model represents the understanding of bioprocess mechanisms. It allows us to inference the latent state (e.g., RNA and nanoparticle binding strength, mRNA-LNP folding structure), which can support biomanufacturing online monitoring and real-time release. Correctly quantifying all sources of uncertainty can facilitate optimal learning, guide risk reduction, and support robust control. *Therefore, the Bayesian KG, accounting for inherent stochasticity and model uncertainty, can be created and used to support integrated bioprocess risk, sensitivity, and predictive analyses.*

Given finite real-world data of the partially observed bioprocess trajectory with size m, denoted by $\mathscr{D}_m = \{\boldsymbol{\tau}^{(i)}_x : i = 1, 2, \ldots, m\}$ with $\boldsymbol{\tau}_x \equiv (\boldsymbol{x}_1, \boldsymbol{a}_1, \ldots, \boldsymbol{x}_H, \boldsymbol{a}_H, \boldsymbol{x}_{H+1})$, the model uncertainty is quantified by a posterior distribution,

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

where $p(\boldsymbol{\theta})$ represents the prior distribution. Since the likelihood evaluation of the KG hybrid model, with high fidelity to capture the critical properties of bioprocessing, is intractable, i.e., $p(\boldsymbol{\tau}_x|\boldsymbol{\theta}) = \int \cdots \int p(\boldsymbol{\tau}|\boldsymbol{\theta}) d\boldsymbol{z}_1 \cdots d\boldsymbol{z}_{H+1}$, approximate Bayesian computation sampling with Sequential Monte Carlo (ABC-SMC) is developed to approximate the posterior distribution (Zheng et al. 2022; Xie et al. 2022).

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 $\mathscr{D}_m$, we accept the sample $\boldsymbol{\theta}$; otherwise reject it. The accept rate can be very low since it is very computationally expensive to match random trajectories from complex bioprocesses especially under the situations with high stochastic and model uncertainties.

The ABC-sequential Monte Carlo (ABC-SMC) methods (Toni et al. 2009; Martin et al. 2019) can improve the sampling efficiency through: (1) generating candidate samples from updated posterior approximates by using sequential importance sampling (SIS); and (2) matching "designed" summary statistics, denoted by $\eta(\mathscr{D})$, instead trajectory observations. Following the spirit of the auxiliary likelihood-based ABC (Martin et al. 2019), we create a linear Gaussian dynamic Bayesian network (LG-DBN) auxiliary model and derive summary statistics for ABC-SMC that can accelerate online Bayesian inference on KG hybrid models (Xie et al. 2022). This simple LG-DBN auxiliary model, in conjunction with SIS, can capture critical biophysicochemical interactions and variations of bioprocess trajectory, ensure the computational efficiency, and enable high quality of inference. *Therefore, the proposed LG-DBN auxiliary likelihood-based ABC-SMC approach can support process soft sensor monitoring, facilitate mechanism online learning, and guide robust process control.*

Interpretable prediction and sensitivity analysis. Given process model parameters and policy parameters, denoted by $(\boldsymbol{\theta}, \boldsymbol{\phi})$, the spatial-temporal interdependencies of the bioprocess SDP trajectory $\boldsymbol{\tau} = (\boldsymbol{s}_1, \boldsymbol{a}_1, \ldots, \boldsymbol{s}_H, \boldsymbol{a}_H, \boldsymbol{s}_{H+1})$ is quantified by the joint distribution, $p(\boldsymbol{\tau}|\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{s}_1) \prod_{t=1}^{H} p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\theta}) \pi_{\boldsymbol{\phi}}(\boldsymbol{a}_t|\boldsymbol{s}_t)$, which depends on underlying bioprocess mechanisms, sensor network design, and data collection strategies. For each batch of production, given inputs denoted by $\mathbf{X}$ (e.g., mRNA sequence, lipid design), we can predict any intermediate or final outputs, denoted by $\mathbf{Y}$ (e.g., CQAs of mRNA-LNP) by using Bayesian KG. The *prediction risk* can be quantified by the posterior predictive distribution, $P(\mathbf{Y}|\mathbf{X}) = \int P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathscr{D}_m) d\boldsymbol{\theta}$, accounting for both stochastic and model uncertainties.

We create a Shapley value (SV)-based sensitivity analysis scheme on the Bayesian KG, called "KG-SV", to support backward root cause analysis and forward interpretable predictive analysis (Xie et al. 2022; Zheng et al. 2021) . Since the proposed Bayesian KG-SV can faithfully account for bioprocess causal interdependencies and biophysicochemical interactions, it can correctly assess the effect from each input variation (such as RNA virus mutation), as well as the impact of each source of stochastic and model uncertainties on the prediction risks. The criticality assessment of input factors is based on the estimated values and estimation uncertainties of interpretable KG hybrid model parameters – such as mechanism pathways in the bioprocess KG from inputs to output (i.e., biomolecular reaction rates, mRNA lipid nanoparticle clustering kinetic parameters). Since model uncertainty can be efficiently reduced by most "informative" data collection and SDP inherent stochasticity can be controlled by decision making, the Bayesian KG based risk, sensitivity, and predictive analyses can identify bottlenecks, guide optimal learning based data collection, and enhance biomanufacturing process CPPs/CQAs specifications for QbD.

3.3 Reinforcement Learning for Bioprocess Design and Control

The proposed Bayesian KG built in conjunction with reinforcement learning (RL) can support long-term prediction and guide interpretable, robust, and optimal decision making. Given any feasible policy specified by parameters $\boldsymbol{\phi} \in \mathbb{C}$, i.e., $\boldsymbol{a}_t = \pi_t(\boldsymbol{s}_t|\boldsymbol{\phi})$, the optimization of the policy π is to maximize the expected accumulated reward,

$$J(\boldsymbol{\phi}) \equiv \mathrm{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathscr{D})} \left[\mathrm{E}_{\boldsymbol{\tau} \sim p(\boldsymbol{\tau}|\boldsymbol{\theta})} \left[\sum_{t=1}^{H+1} r_t(\boldsymbol{s}_t, \boldsymbol{a}_t) \middle| \boldsymbol{\pi}, \boldsymbol{\theta} \right] \right],$$

accounting for bioprocess inherent stochasticity and model uncertainty, where $\mathbb{C}$ is a feasible region. At any k-th iteration, we can use the policy gradient to solve the optimization,

$$\boldsymbol{\phi}_{k+1} = \Pi_{\mathbb{C}} \left(\boldsymbol{\phi}_k + \eta_k \nabla J(\boldsymbol{\phi}_k) \right),$$

where η_k is a suitable stepsize and $\Pi_{\mathbb{C}}$ is a projection onto $\mathbb{C}$.

Reinforcement learning scheme on Bayesian KG. We propose model-based RL scheme on the Bayesian KG (Zheng et al. 2021), which can provide an insightful prediction on how the effect of input factor propagates through bioprocess mechanism pathways and impacts on the outputs. It can find control policies that are interpretable and robust against heterogeneous model uncertainty, and overcome the key challenges of biopharmaceutical manufacturing, i.e., high complexity, high uncertainty, and very limited process data. To support real-time control for complex biomanufacturing processes, we provide a provably convergent stochastic policy gradient optimization and it is computationally efficient through reusing computations associated with similar input-output mechanism pathways.

Hybrid model likelihood ratio based historical observation reuse. Since each experiment run is very computationally expensive especially for multi-scale bioprocess hybrid model, we propose KG assisted multiple important sampling ("KG-MIS") to accelerate policy gradient optimization (Zheng et al. 2021). Basically, we can select and reuse the "most relevant" historical trajectories, improve policy gradient estimation, and accelerate the search for the optimal robust policy. For high dimensional SDP, this study can selectively reuse historical trajectories having similar underlying distributions with that of target SDP and improve the estimation of policy gradient.

In classical policy gradient (PG) approach, at any k-th iteration, the sample average approximation (SAA) is used to estimate the gradient based on n new trajectories generated, $\nabla \widehat{J}^{PG}(\boldsymbol{\theta}_k) = \frac{1}{n}\sum_{j=1}^{n} g\left(\boldsymbol{\tau}^{(k,j)}|\boldsymbol{\theta}^{(k,j)},\boldsymbol{\phi}_k\right)$ with the scenario gradient $g(\boldsymbol{\tau}|\boldsymbol{\theta},\boldsymbol{\phi}) = \nabla_{\boldsymbol{\theta}} \mathrm{E}_{\boldsymbol{\tau}}\left[\sum_{t=1}^{H} r_t(\boldsymbol{s}_t,\boldsymbol{a}_t)\middle|\boldsymbol{\theta},\boldsymbol{\phi}\right]$, where $\boldsymbol{\theta}^{(k,j)} \sim p(\boldsymbol{\theta}|\mathscr{D}_m)$ and $\boldsymbol{\tau}^{(k,j)} \sim p(\boldsymbol{\tau}|\boldsymbol{\theta}^{(k,j)},\boldsymbol{\phi}_k)$. The target SDP mixture distribution $p_k(\boldsymbol{\tau}) = \frac{1}{n}\sum_{j=1}^{n} p(\boldsymbol{\tau}|\boldsymbol{\theta}^{(k,j)},\boldsymbol{\phi}_k)$ accounts for both process stochastic and model uncertainties. Motivated by the studies on multiple important sampling (MIS) (Dong et al. 2018; Feng and Staum 2017), we create a KG-MIS policy gradient unbiased estimator,

$$\nabla \widehat{J}^{MIS}(\boldsymbol{\phi}_k) = \frac{1}{n|U_k|}\sum_{i\in U_k}\sum_{j=1}^{n} f_k\left(\boldsymbol{\tau}^{(i,j)}\right) g\left(\boldsymbol{\tau}^{(i,j)}\middle|\boldsymbol{\theta}^{(k,j)},\boldsymbol{\phi}_k\right) \quad \text{with} \quad f_k(\boldsymbol{\tau}) = \frac{p_k(\boldsymbol{\tau})}{\sum_{i\in U_k} p(\boldsymbol{\tau}|\boldsymbol{\theta}_i,\boldsymbol{\phi}_i)/|U_k|}.$$

Since an inappropriate selection of reuse set U_k can lead to the inflated estimation variance of policy gradient, we propose a variance reduction based experience replay criteria (Zheng et al. 2021), which can automatically select the most relevant historical trajectories generated under different decision policies $\boldsymbol{\phi}$ and model parameters $\boldsymbol{\theta}$ from different posterior distributions resulting from online learning and process control. We prove that the proposed approach is asymptotically convergent and show it significantly outperforms classical policy gradient approach. Furthermore, we extend the proposed KG-MIS framework so that it can select and reuse the most relevant partial trajectories from historical observations (Zheng and Xie 2022), i.e., the reuse unit is defined based on state-action transition $(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')$. This study can allow us to flexibly integrate and leverage the relevant information collected from different production lines and facilitate personalized bio-drug manufacturing.

4 CHALLENGES AND OPPORTUNITIES FOR OPERATIONS RESEARCH

Increased flexibility, that is necessary to achieve personalized products and manufacturing, should be considered early on as integral part of product design. In fact, for achieving a "full circle", not only the manufacturing technology needs to be flexible, but also the drug design and the process control need to support it. Novel operations research approaches and simulation platforms can substantially improve the performance of CMO allowing for larger variability of products with potentially small volume per variant capitalizing upon single use/disposable technologies.

Drug discovery is positively impacted by optimization methods. These should embed scarce data and low fidelity physical models characterizing the existing understanding of bioprocess mechanisms. In fact, gray-box search methods are a very active field of research and we believe drug design represents a leading

opportunity for further development. The expert-based framework is an example of such approaches, but more efforts are necessary.

Novel simulation methodologies are necessary for analyzing end-to-end biomanufacturing processes and supporting interoperability. Among new upcoming techniques are hybrid modeling, data integration, risk management, and interpretable robust process control. However these are an example and more development in the ares is required.

Acknowledgements

We acknowledge Hua Zheng and Keqi Wang (Northeastern University) for their help on the paper preparation. This work was partially supported by the grants NSF#2046588-2007861 (PI Pedrielli) and NIST#70NANB21H086 (PI Xie).

REFERENCES

Aghaeepour, N., and H. H. Hoos. 2013. "Ensemble-based Prediction of RNA Secondary Structures". *BMC Bioinformatics* 14 (1): 139.

Ain, Q. U., A. Aleksandrova, F. D. Roessler, and P. J. Ballester. 2015. "Machine-Learning Scoring Functions to Improve Structure-based Binding Affinity Prediction and Virtual Screening". *Wiley Interdisciplinary Reviews: Computational Molecular Science* 5 (6): 405–424.

AlQuraishi, M. 2019. "AlphaFold at CASP13". *Bioinformatics* 35 (22): 4862–4865.

Andronescu, M., A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy. 2010. "Computational Approaches for RNA Energy Parameter Estimation". *RNA* 16 (12): 2304–2318.

Angela, M. Y., P. M. Gasper, L. Cheng, L. B. Lai, S. Kaur, V. Gopalan, A. A. Chen, and J. B. Lucks. 2021. "Computationally Reconstructing Cotranscriptional RNA Folding from Experimental Data Reveals Rearrangement of Non-Native Folding Intermediates". *Molecular Cell* 81 (4): 870–883.

Arendt, P. D., D. W. Apley, and W. Chen. 2012. "Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability". *Journal of Mechanical Design* 134 (10).

Ballester, P. J., and J. B. Mitchell. 2010. "A Machine Learning Approach to Predicting Protein–Ligand Binding Affinity with Applications to Molecular Docking". *Bioinformatics* 26 (9): 1169–1175.

Baroni, G., and S. Tarantola. 2014. "A General Probabilistic Framework for Uncertainty and Global Sensitivity Analysis of Deterministic Models: A Hydrological Case Study". *Environmental Modelling & Software* 51:26–34.

Buschmann, M. D., M. J. Carrasco, S. Alishetty, M. Paige, M. G. Alameh, and D. Weissman. 2021. "Nanomaterial Delivery Systems for mRNA Vaccines". *Vaccines* 9 (1): 65.

Calonaci, N., A. Jones, F. Cuturello, M. Sattler, and G. Bussi. 2020. "Machine Learning a Model for RNA Structure Prediction". *NAR Genomics and Bioinformatics* 2 (4): lqaa090.

Carell, T., C. Brandmayr, A. Hienzsch, M. Müller, D. Pearson, V. Reiter, I. Thoma, P. Thumbs, and M. Wagner. 2012. "Structure and Function of Noncanonical Nucleobases". *Angewandte Chemie International Edition* 51 (29): 7110–7131.

Chaudhury, S., S. Lyskov, and J. J. Gray. 2010. "PyRosetta: A Script-based Interface for Implementing Molecular Modeling Algorithms Using Rosetta". *Bioinformatics* 26 (5): 689–691.

Cintron, R. 2015. *Human Factors Analysis and Classification System Interrater Reliability for Biopharmaceutical Manufacturing Investigations*. Ph. D. thesis, Walden University. https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1193&context=dissertations, accessed 24th September 2022.

Dong, J., M. B. Feng, and B. L. Nelson. 2018. "Unbiased Metamodeling via Likelihood Ratios". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1778–1789. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Feng, M., and J. Staum. 2017, October. "Green Simulation: Reusing the Output of Repeated Experiments". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 27 (4): 23:1–23:28.

Kyle Blankenship for FiercePharma 2020. "Samsung Scores $362M Deal to Help Vir Scale Up COVID-19 Antibody Production.". https://www.fiercepharma.com/manufacturing/vir-inks-deal-samsung-to-scale-up-covid-19-antibody-manufacturing, accessed 24th September 2022.

Friesner, R. A., J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry et al. 2004. "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy". *Journal of Medicinal Chemistry* 47 (7): 1739–1749.

Guideline, ICH Harmonised Tripartite and others 2009. "Pharmaceutical Development Q8 (R2)". https://vnras.com/wp-content/uploads/2017/05/Q8R2_PHARMACEUTICAL-DEVELOPMENT.pdf, accessed 24th September 2022.

Han, D., X. Qi, C. Myhrvold, B. Wang, M. Dai, S. Jiang, M. Bates, Y. Liu, B. An, F. Zhang et al. 2017. "Single-Stranded DNA and RNA Origami". *Science* 358 (6369): eaao2648.

Hart, T. N., and R. J. Read. 1992. "A Multiple-start Monte Carlo Docking Method". *Proteins: Structure, Function, and Bioinformatics* 13 (3): 206–222.

Huang, S.-Y., S. Z. Grinter, and X. Zou. 2010. "Scoring Functions and Their Evaluation Methods for Protein–Ligand Docking: Recent Advances and Future Directions". *Physical Chemistry Chemical Physics* 12 (40): 12899–12908.

Jin, S., V. G. Contessoto, M. Chen, N. P. Schafer, W. Lu, X. Chen, C. Bueno, A. Hajitaheri, B. J. Sirovetz, A. Davtyan et al. 2020. "AWSEM-Suite: A Protein Structure Prediction Server Based on Template-Guided, Coevolutionary-Enhanced Optimized Folding Landscapes". *Nucleic Acids Research* 48 (W1): W25–W30.

Jones, G., P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. 1997. "Development and Validation of A Genetic Algorithm for Flexible Docking". *Journal of Molecular Biology* 267 (3): 727–748.

Kim, S., J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu et al. 2019. "PubChem 2019 Update: Improved Access to Chemical Data". *Nucleic Acids Research* 47 (D1): D1102–D1109.

Korb, O., T. Stützle, and T. E. Exner. 2006. "PLANTS: Application of Ant Colony Optimization to Structure-based Drug Design". In *International Workshop on Ant Colony optimization and swarm intelligence*, edited by M. Dorigo, L. M. Gambardella, M. Birattari, A. Martinoli, R. Poli, and T. Stützle, 247–258. Berlin, Heidelberg: Springer.

Kulkarni, N. S. 2015. "A Modular Approach for Modeling Active Pharmaceutical Ingredient Manufacturing Plant: A Case Study". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 2260–2271. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lamiable, A., P. Thévenet, J. Rey, M. Vavrusa, P. Derreumaux, and P. Tufféry. 2016. "PEP-FOLD3: Faster De Novo Structure Prediction for Linear Peptides in Solution and in Complex". *Nucleic Acids Research* 44 (W1): W449–W454.

Langer, Eric 2020. "Trends and Opportunities in Biopharmaceutical Product Development and Manufacturing". https://www.manufacturingchemist.com/news/article_page/Trends_and_opportunities_in_biopharmaceutical_product_development_and_manufacturing/170557, accessed 24th September 2022.

Leachman, R. C., L. Johnston, S. Li, and Z.-J. Shen. 2014. "An Automated Planning Engine for Biopharmaceutical Production". *European Journal of Operational Research* 238 (1): 327–338.

Lee, G., W. Kim, H. Oh, B. D. Youn, and N. H. Kim. 2019. "Review of Statistical Model Calibration and Validation—from the Perspective of Uncertainty Structures". *Structural and Multidisciplinary Optimization*:1–26.

Liu, M., E. Poppleton, G. Pedrielli, P. Šulc, and D. P. Bertsekas. 2022. "ExpertRNA: A New Framework for RNA Secondary Structure Prediction". *INFORMS Journal on Computing*.

Lucks, J. B., S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin. 2011. "Multiplexed RNA Structure Characterization with Selective 2'-Hydroxyl Acylation Analyzed by Primer Extension Sequencing (SHAPE-Seq)". *Proceedings of the National Academy of Sciences* 108 (27): 11063–11068.

Martagan, T., A. Krishnamurthy, P. A. Leland, and C. T. Maravelias. 2018. "Performance Guarantees and Optimal Purification Decisions for Engineered Proteins". *Operations Research* 66 (1): 18–41.

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.

Marx, V. 2022. "Method of the Year: Protein Structure Prediction". *Nature Methods* 19 (1): 5–10.

Mock, U., L. Nickolay, B. Philip, G. W.-K. Cheung, H. Zhan, I. C. Johnston, A. D. Kaiser, K. Peggs, M. Pule, A. J. Thrasher et al. 2016. "Automated Manufacturing of Chimeric Antigen Receptor T Cells for Adoptive Immunotherapy Using CliniMACS Prodigy". *Cytotherapy* 18 (8): 1002–1011.

Morrison, K. C., and P. J. Hergenrother. 2013. "Natural Products as Starting Points for the Synthesis of Complex and Diverse Compounds". *Natural Product Reports* 31 (1): 6–14.

Mysinger, M. M., M. Carchia, J. J. Irwin, and B. K. Shoichet. 2012. "Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking". *Journal of Medicinal Chemistry* 55 (14): 6582–6594.

Newman, D. J., and G. M. Cragg. 2020. "Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019". *Journal of Natural Products* 83 (3): 770–803.

Pei, J., Q. Wang, Z. Liu, Q. Li, K. Yang, and L. Lai. 2006. "PSI-DOCK: Towards Highly Efficient and Accurate Flexible Ligand Docking". *Proteins: Structure, Function, and Bioinformatics* 62 (4): 934–946.

Rarey, M., B. Kramer, T. Lengauer, and G. Klebe. 1996. "A Fast Flexible Docking Method Using an Incremental Construction Algorithm". *Journal of Molecular Biology* 261 (3): 470–489.

Reuter, J. S., and D. H. Mathews. 2010. "RNAstructure: Software for RNA Secondary Structure Prediction and Analysis". *BMC Bioinformatics* 11 (1): 129.

Rodrigues, T., D. Reker, P. Schneider, and G. Schneider. 2016. "Counting on Natural Products for Drug Design". *Nature Chemistry* 8 (6): 531–541.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2008. *Global Sensitivity Analysis: The Primer*. The Atrium, Southern Gate, Chichester, England: John Wiley & Sons.

Sandle, T. 2018. "Strategy for the Adoption of Single-Use Technology". *European Pharmaceutical Review*.

Sellami, A., M. Réau, F. Langenfeld, N. Lagarde, and M. Montes. 2021. "Virtual Libraries for Docking Methods: Guidelines for the Selection and the Preparation". In *Molecular Docking for Computer-Aided Drug Design*, 99–117. Elsevier.

Sorokina, M., and C. Steinbeck. 2020. "Review on Natural Products Databases: Where to Find Data in 2020". *Journal of Cheminformatics* 12 (1): 1–51.

Stanzione, F., I. Giangreco, and J. C. Cole. 2021. "Use of Molecular Docking Computational Tools in Drug Discovery". *Progress in Medicinal Chemistry* 60:273–343.

Sterling, T., and J. J. Irwin. 2015. "ZINC 15–Ligand Discovery for Everyone". *Journal of Chemical Information and Modeling* 55 (11): 2324–2337.

Sun, T.-t., C. Zhao, and S.-J. Chen. 2018. "Predicting Cotranscriptional Folding Kinetics for Riboswitch". *The Journal of Physical Chemistry B* 122 (30): 7484–7496.

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.

Wang, B., Q. Zhang, and W. Xie. 2019. "Bayesian Sequential Data Collection for Stochastic Simulation Calibration". *European Journal of Operational Research* 277 (1): 300–316.

Watkins, A. M., R. Rangan, and R. Das. 2020. "FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds". *Structure*.

Wayment-Steele, H. K., W. Kladwang, E. Participants, and R. Das. 2020. "RNA Secondary Structure Packages Ranked and Improved by High-throughput Experiments". *BioRxiv*. https://www.biorxiv.org/content/10.1101/2020.05.29.124511v2, accessed 24th September 2022.

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.

Xie, W., K. Wang, H. Zheng, and B. Feng. 2022. "Dynamic Bayesian Network Auxiliary ABC-SMC for Hybrid Model Bayesian Inference to Accelerate Biomanufacturing Process Mechanism Learning and Robust Control". *arXiv preprint arXiv:2205.02410*. https://arxiv.org/abs/2205.02410, accessed 24th September 2022.

Zadeh, J. N., C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce. 2011. "NUPACK: Analysis and Design of Nucleic Acid Systems". *Journal of Computational Chemistry* 32 (1): 170–173.

Zheng, H., and W. Xie. 2022. "Variance Reduction based Partial Trajectory Reuse to Accelerate Policy Gradient Optimization". In *Proceedings of the 2022 Winter Simulation Conference*, edited by B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. Corlu, L. Lee, E. Chew, T. Roeder, and P. Lenderman. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zheng, H., W. Xie, and M. B. Feng. 2021. "Variance Reduction based Experience Replay for Policy Optimization". *arXiv preprint arXiv:2110.08902*. https://arxiv.org/abs/2110.08902, accessed 24th September 2022.

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*. accepted by INFORMS Journal on Computing. 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.

Zhu, F., N. Shah, H. Xu, D. Schneider, R. Orentas, B. Dropulic, P. Hari, and C. A. Keever-Taylor. 2017. "Closed-System Manufacturing of CD19 and Dual-targeted CD20/19 Chimeric Antigen Receptor T cells Using the CliniMACS Prodigy Device at an Academic Medical Center". *Cytotherapy*.

AUTHOR BIOGRAPHIES

WEI XIE is an assistant professor in MIE at Northeastern University. Her research interests include interpretable Artificial Intelligence (AI), machine learning (ML), computer simulation, digital twin, 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/

GIULIA PEDRIELLI is Assistant Professor for the School of Computing and Augmented Intelligence at Arizona State University. Her expertise is in the design and analysis of sampling based algorithms, Monte Carlo methods, reinforcement learning, and dynamic simulation. Applications range from molecular design, process design and control, testing and verification of Cyber Physical Systems. Her email address is giulia.pedrielli@asu.edu, her webpage can be found at https://www.gpedriel.com/.