Final Project

1. Introduction and Problem Description:

With the COVID-19 hitting the world, the importance and urgency of drug manufacturing has grown at a massive scale. Due to the high demand of drugs across the world to treat patients, process improvement for drug manufacturing has become a topic of research in the industry.

Through our research effort, we aim to develop a simulation model to study and observe methods to reduce batch-to-batch variability in yield (protein level), quality (impurity level), and production cycle time in the biopharmaceutical production process. We focus our study on a typical bio-drug substance production process which includes an upstream cell culture process (USP) and a downstream target protein purification (DSP).

Typically, the Upstream Cell Culture Process (USP) starts with pre-culture. Our study considers a modified version of the drug substance production process given by Wang et al [1]. We model a modified process that considers only the use of external media and no internally produced media. After the main fermentation step, we move to Downstream Protein Purification Processes (USP). For simplicity, our removes cleaning stages. The modified process can be seen in Figure 1.

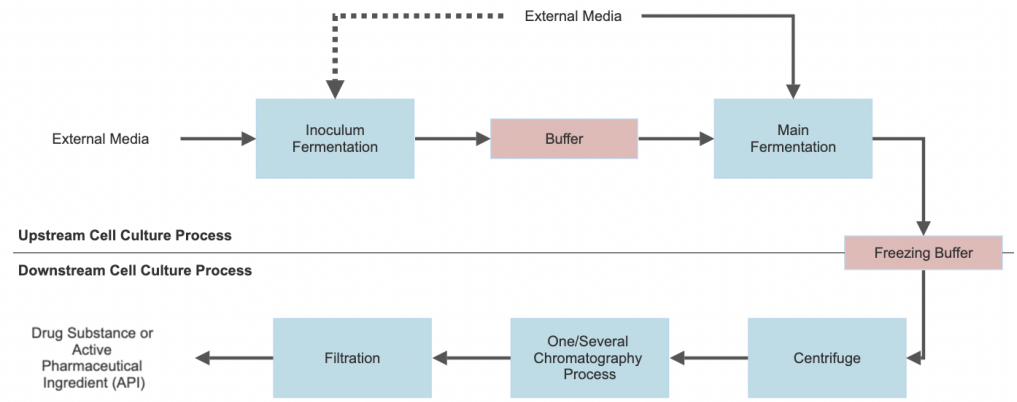

Figure 1: Modified drug substance production process

These main steps that we focus on include: (1) pre-culture and expansion, (2) fermentation and harvest, (3) centrifugation(s), (4) chromatography/purification, and (5) filtration.

2. Input Modelling:

In this section, we describe our methods for obtaining our input modelling procedure i.e., finding the best fit distribution for our given sample data. For the simulation model presented in this report, we model the interarrival times of the samples using a Poisson process with mean 2 hours. To determine the input model for the model parameters, we have

limited sample data (sample size = 20), so a parametric bootstrap with 1000 resamples and the parameter confidence interval as 95%, is run using the @Risk software. We mainly consider the following fit ranking criteria to model our inputs:

- Kolmogorov-Smirnov (K-S)
- Anderson-Darling tests (A-D)
- Q-Q Plot (the distribution is a good fit if the plot is close to 45^{o})

Upstream Cell Culture Process

1. **Inoculation Fermentation**: This is obtained using our available data for starting biomass, where we assume $X_{0,i} \sim F_{X_0}$ for $i = 1, 2, \ldots, m$. Using the given data for X_0, the following fit results are obtained:

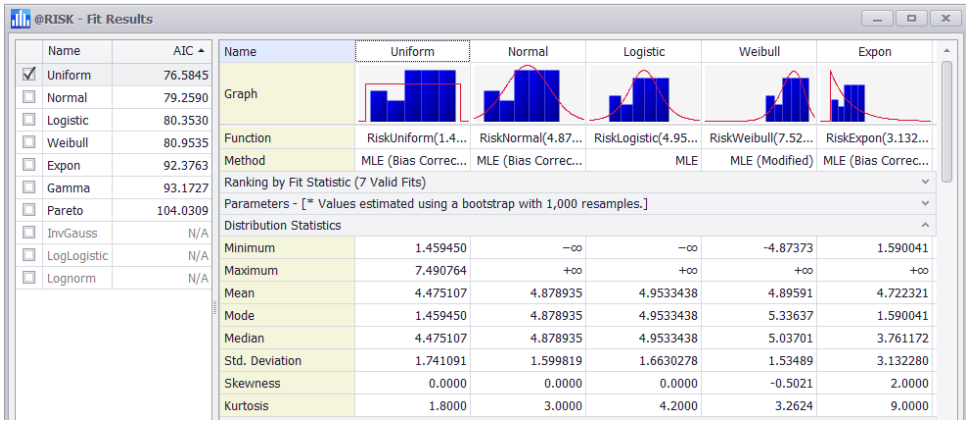

Fig 2: Distribution Statistics for the chosen distributions for the variable X_0

Information Criteria					
Akaike (AIC)	76.584485	79.259032	80.3529855	80.95352	92.376325
Bayesian (BIC)	77.870067	80.544615	81.6385677	82.44071	93.661908
Average Log-Likelihood	-1.796965	-1.863829	-1.8911776	-1.83634	-2.191761

Fig 3: The Information Criteria for X_0

Anderson-Darling Test - [* Values estimated using a bootstrap with 1,000 resamples.]					
Statistic	0.786596	0.275231	0.2770552	0.24150	2.517877
P-Value*	0.328000	0.656000	0.6220000	N/A	0.000000
Kolmogorov-Smirnov Test - [* Values estimated using a bootstrap with 1,000 resamples.]					
Statistic	0.216432	0.103244	0.0913218	0.10500	0.325328
P-Value*	0.218000	0.821000	0.8960000	N/A	0.001000

Fig 4: A-D and K-S p-values for X_0

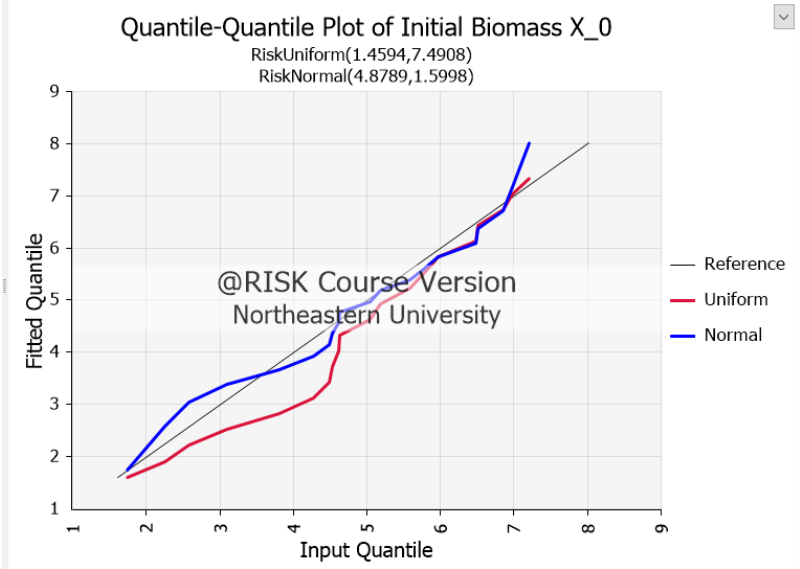

Fig 5: Q-Q Plot comparing Uniform and Normal Distribution for X_0

Based on the p-value of the K-S statistic and domain knowledge, the normal distribution is chosen to be the best fit distribution for modelling X_0 with mean 4.89, and variance $(1.60)^2$. Another aspect to investigate is the possibility of producing negative values in our distribution generation, we incorporate a truncated normal distribution during the simulation where any negative values would be automatically become zero.

2. **Main Fermentation:** The available data gave us starting biomass X0 and ending protein level Xf. We used @RISK to determine a model to fit to our empirical distribution. For our protein model and impurity, we use the following: $X_F = X_0 \cdot e^{\mu T + \varepsilon_P}$ and $I_F = X_F \cdot \alpha \cdot e^{\varepsilon_I}$, where $\varepsilon_P \sim N(0, \sigma_P^2)$ and $\varepsilon_I \sim N(0, \sigma_I^2)$ and $\alpha = 1.5$, to quantify the batch-to-batch variations in protein and impurity levels after fermentation. We take the log of this function and then fit a linear regression to obtain values for our parameters for mean and variance.

   Using the given data for X_F (Protein level post main Fermentation step) and the data of X_0, the parameters μ and ε_P can be estimated by using the log transformation of $X_F = X_0 \cdot e^{\mu T + \varepsilon_P}$. By fitting a linear regression line to the log transformation, we get that parameter $(\mu T + \varepsilon_P)$ is modelled by a normal distribution with mean 1.80 and variance $(0.27)^2$. Thus, obtaining the distribution for the parameter ε_P to be $\sim N(0, (0.27)^2)$.

   As we consider $\beta_0 = 1$, where $\beta_0 = \sigma_I / \sigma_P$ to study the impact of relative impurities variation on the output, as well as to determine our σ_I given our σ_P values. The distribution for ε_I *is the same as* ε_P i.e., $\sim N(0, (0.27)^2)$.

Downstream Protein Purification Process

3. **Centrifuge:** We use (X_C, I_C) to denote protein and impurity levels after the centrifuge, respectively. We assume that it does not change the protein and simply removes a random proportion of impurity, $I_C = Q \cdot I_F$. Given existing literature [1], we allow Q to follow a Uniform distribution given by $Q \sim$ Unif(0.4, 0.5).
4. **Chromatography:** We use (X_P, I_P) to identify protein and impurity levels of our antigen after chromatography. Each chromatography step follows the $X_P = Q_P \cdot X_C$ and $I_P = Q_I \cdot I_C$; where the

distribution of Q_P and Q_I can be obtained by fitting a distribution using the @Risk Software with the given data of Q_P and Q_I for the chromatography step.

The Fit results obtained for Q_P:

@RISK - Fit Results

Name	AIC
Uniform	-76.2273
Normal	-70.9708
Weibull	-69.7118
Logistic	-68.7325
LogLogistic	-66.4466
Expon	-58.4443
Pareto	-56.9357
Gamma	N/A
Lognorm	N/A

Name	Uniform	Normal	Weibull	Logistic	LogLogistic
Graph					
Function	RiskUniform(0.6...	RiskNormal(0.79...	RiskWeibull(2.54...	RiskLogistic(0.78...	RiskLogLogistic(0...
Method	MLE (Bias Correc...	MLE (Bias Correc...	MLE (Modified)	MLE	MLE (Modified)
Ranking by Fit Statistic (7 Valid Fits)					
Parameters - [* Values estimated using a bootstrap with 1,000 resamples.]					
Distribution Statistics					
Minimum	0.6571625	−∞	0.6293662	−∞	0.2482751
Maximum	0.9178249	+∞	+∞	+∞	+∞
Mean	0.7874937	0.79031654	0.7905195	0.78885685	0.7915038
Mode	0.6571625	0.79031654	0.7785413	0.78885685	0.7800600
Median	0.7874937	0.79031654	0.7865506	0.78885685	0.7863092
Std. Deviation	0.0752467	0.07000470	0.0679056	0.07482627	0.0761140
Skewness	0.0000	0.0000	0.3396	0.0000	0.6930
Kurtosis	1.8000	3.0000	2.8387	4.2000	5.4272
Percentiles					
Information Criteria					
Akaike (AIC)	-76.2273247	-70.97081892	-69.7118093	-68.73251102	-66.4465705
Bayesian (BIC)	-73.8693744	-68.61286860	-66.4312940	-66.37456070	-63.1660552
Average Log-Likelihood	1.3445295	1.25692106	1.2772481	1.21961592	1.2228275

Fig 6: *Distribution Statistics and Information Criteria for the chosen distributions for the variable* Q_P

Anderson-Darling Test - [* Values estimated using a bootstrap with 1,000 resamples.]					
Statistic	0.3899263	0.36015583	0.3266652	0.42242554	0.3768423
P-Value*	0.7600000	0.43600000	N/A	0.24400000	N/A
Kolmogorov-Smirnov Test - [* Values estimated using a bootstrap with 1,000 resamples.]					
Statistic	0.1319801	0.13143149	0.1097184	0.12670537	0.1119350
P-Value*	0.5690000	0.19300000	N/A	0.10700000	N/A

Fig 7: A-D and K-S p-values for Q_P

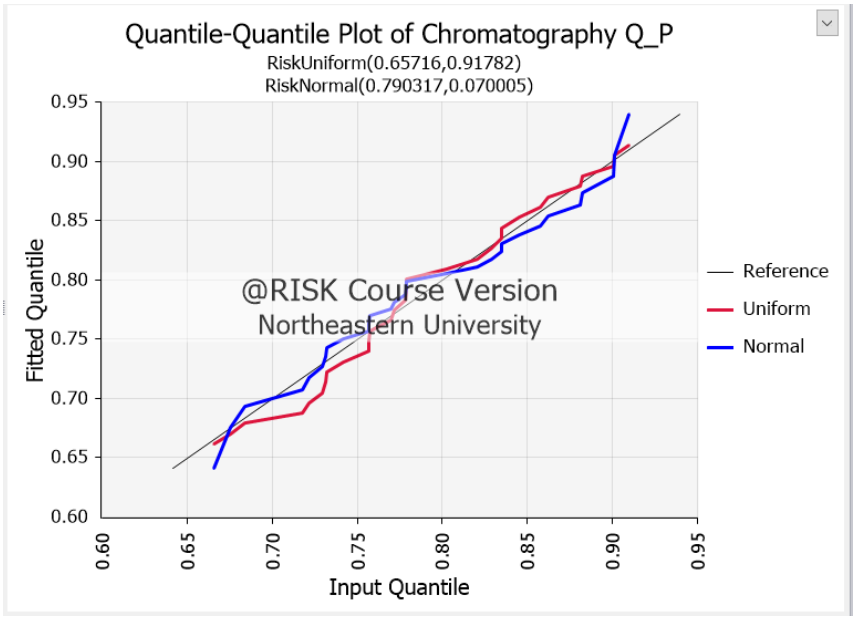

Fig 8: Q-Q Plot comparing Uniform and Normal Distributions for Q_P

Based on the p-value of the AIC Ranking, p-value in the A-D statistic, domain knowledge and the higher inclination of the Uniform distribution Q-Q plot towards 45-degree angle as compared to the Normal distribution, the Uniform distribution is chosen to be the best fit distribution for modelling Q_P with minimum 0.66, and maximum 0.92.
Similarly, the Fit results obtained for Q_I:

@RISK - Fit Results

Name	AIC
☑ Uniform	-47.3026
☐ Weibull	-40.3999
☐ Normal	-40.3188
☐ Gamma	-38.9412
☐ Lognorm	-38.5735
☐ Logistic	-38.2159
☐ LogLogistic	-36.3853
☐ Expon	-35.4034
☐ Pareto	-28.0838

Name	Uniform	Weibull	Normal	Gamma	Lognorm
Graph					
Parameters - [* Values estimated using a bootstrap with 1,000 resamples.]					
Distribution Statistics					
Minimum	0.2424452	0.2273287	−∞	0.15465394	-0.1325976
Maximum	0.6645715	+∞	+∞	+∞	+∞
Mean	0.4535084	0.4389138	0.4392097	0.43920966	0.4393011
Mode	0.2424452	0.3884518	0.4392097	0.38873348	0.4052463
Median	0.4535084	0.4239971	0.4392097	0.42257270	0.4277165
Std. Deviation	0.1218574	0.1156970	0.1166793	0.11984693	0.1168945
Skewness	0.0000	0.6994	0.0000	0.8423	0.6217
Kurtosis	1.8000	3.3805	3.0000	4.0643	3.6951
Percentiles					
Information Criteria					
Akaike (AIC)	-47.3026057	-40.3999449	-40.3188208	-38.94124120	-38.5735407
Bayesian (BIC)	-44.9446553	-37.1194297	-37.9608705	-35.66072597	-35.2930255
Average Log-Likelihood	0.8624508	0.7887170	0.7460544	0.76440530	0.7582770

Fig 9: *Distribution Statistics and Information Criteria for the chosen distributions for the variable* Q_I

Anderson-Darling Test - [* Values estimated using a bootstrap with 1,000 resamples.]					
Statistic	0.4579402	0.3159834	0.3653869	0.32487911	0.3235846
P-Value*	0.6680000	N/A	0.4200000	N/A	N/A
Kolmogorov-Smirnov Test - [* Values estimated using a bootstrap with 1,000 resamples.]					
Statistic	0.1492037	0.0955137	0.1023152	0.10165318	0.0874842
P-Value*	0.4200000	N/A	0.5510000	N/A	N/A

Fig 10: A-D and K-S p-values for Q_I

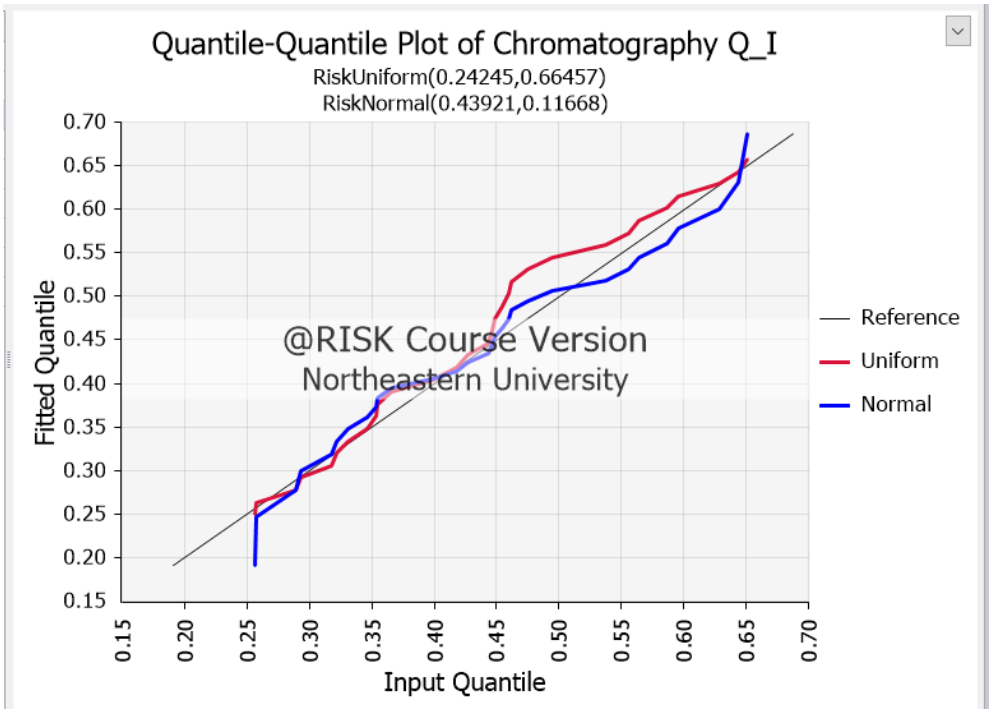

Fig 11: Q-Q Plot comparing Uniform and Normal Distributions for Q_I

Based on the p-value of the A-D statistic, domain knowledge and the higher inclination of the Uniform distribution Q-Q plot towards 45^0 angle as compared to the Normal distribution, the Uniform distribution is chosen to be the best fit distribution for modelling Q_I with minimum 0.25, and maximum 0.67.

5. **Filtration:** Protein levels are not affected during the filtration step, only impurity levels are slightly affected. $Q_{fr} \sim$ Unif(0.99, 1) is used for our Q ratio. Where, $I_{Fr} = Q_{Fr}.I_P$

3. Simulation Modelling:

We created a simulation model to model the changes in biopharmaceutical batch attributes during production, which can help manage production risks and inform operational decisions.

The resources of bio-drug substance production system include the following equipment: inoculum tank [5], main vessel [5], centrifuge equipment [2], chromatography [5] and filtration equipment(s) [2], where the number in the bracket [·] gives the capacity or the number of each equipment(s).

We start with the biomass attribute, X_0, and consider the inoculum fermentation as a warm-up step. After the main fermentation, the batch becomes the drug product, with protein level X and impurity level I. We quantify the uncertainty in the production system to study batch-to-batch variation. After that, we use bio-chemical knowledge to model the stochastic input-output relationship for each process unit operation, including main fermentation, centrifuge, chromatography, and filtration. Our simulation models the probability distributions for each process unit operation to capture cell-level dynamics and the input-output stochastic relationship for each batch or entity. For our model, we ignore the no-wait constraint and the FDA impurity percentage constraints for simplification.

There are many assumptions required to accurately model our simulation. The first of which is that all processes follow a fixed processing time as per Table 1.

Time (hours)	Antigen A
Inoculum Fermentation	24
Main Fermentation	72
Centrifuge	2.5
Chromatography	8.0
Filtration	2.0

Table1: Fixed Processing time for each step

First, we assume that there is always a sufficient inventory of external media. There is no warmup period as we are looking at batch-to-batch variation. For our model, we ignore the no-wait constraint and the FDA impurity percentage constraints for simplification. We also assume that equipment is immediately switched with clean equipment after each step in the process and it does not affect the simulation model.

4. Input Uncertainty Quantification:

As our input models are estimated using 20 real world data, we considered a non-parametric bootstrap approach to quantify our input model estimation uncertainty using the @Risk software. Following this we run a parametric bootstrap using the given data in our simulation model and obtain the batch-to-batch means of protein/impurity level.

Using our calculated batch-to-batch means of protein/impurity level, we record the 95% CI lower/upper bounds and bootstrap variance as follows:

Output	*Mean Protein*	*Mean Impurity*	*Mean Cycle Time*	*SD protein*	*SD Impurity*	*SD Cycle Time*
CI Lower	48.98	12.62	1309.59	53.13	14.57	714.13
CI Upper	51.51	13.28	1314.03	59.95	16.38	716.59
Variance	83.56	5.59	255.86	606.37	42.63	79.35

Table 2: Impact of Input estimation uncertainty on simulation output at USP/DSP = (5,2)

Output	*Mean Protein*	*Mean Impurity*	*Cycle Time*	*SD protein*	*SD Impurity*	*SD Cycle Time*
CI Lower	49.85	12.85	772.05	53.46	14.71	402.58
CI Upper	51.86	13.36	776.07	58.15	15.97	404.84
Variance	52.68	3.43	210.24	286.35	20.89	66.63

Table 3: Impact of Input estimation uncertainty on simulation output at USP/DSP = (8,4)

USP/DSP = (5,2) refers to the releasing policies for upstream and downstream, respectively. Running the sample again for (8,4), we get lower and upper bounds for our confidence intervals of {772.05, 776.07}

5. Model Validation

We validate our model by recreating our model under simplified conditions. As our model is already a simplified version of the system proposed in Wang el. Al (1), we have already ignored cleaning times. And so, we create a Jackson Network that modifies the inter-arrival times and processing times following exponential distributions. Our simplified simulation experiments show that our total analytical mean cycle time [772.05,776.07] matches the estimated total cycle time 95% confidence interval [690.576, 805.72] for USP/DSP = (8,4) and the total analytical mean cycle time of [1309.59, 1314.03] just barely lies outside our 95% confidence interval [1084.91, 1269.25] for USP/DSP = (5,2).

6. Sensitivity Analysis:

In order to perform sensitivity analysis of the system performance, we simulated our model at various levels of β_0, at 0.5, 1.0, and 1.5. We estimated the batch-to-batch Protein, Impurity levels and the cycle time. We ran the simulation with 200 bootstrapped samples and for 100 replications with a run length of 200 batches.

The simulation results clearly indicate that with increasing β_0, we see that impurity variation increases due to $\beta_0 = \sigma_I/\sigma_P$. Because of this, many more batches would be getting dropped at various steps due to not meeting our passing requirements and as a result, our total amount of protein being passed is also decreased. We also will get a lower utilization for our chromatography servers.

Beta Value	**0.5**	**1.0**	**1.5**
Mean Protein	43.58 ± 0.79	46.61 ± 0.75	51.34 ± 0.92
Mean Impurity	11.24 ± 0.19	12.02 ± 0.20	13.23 ± 0.24
Mean Cycle Time	774.06 ± 2.01	774.49 ± 1.45	774.59 ± 1.22
SD Protein	36.77 ± 1.43	46.53 ± 1.81	61.42 ± 2.88

SD Impurity	10.31 ± 0.41	12.88 ± 0.51	16.86 ± 0.80
SD Cycle Time	403.71 ± 1.13	404.02 ± 0.83	404.18 ± 0.70

Table 4: Sensitivity Analysis of System Performance at β_0 = 0.5, 1.0, 1.5

7. Output Analysis:

Comparing our results to Wang et. Al (1), we know that our final levels of protein and impurities are much higher. These errors in our system are accounted for since we make many assumptions and remove certain constraints that have allowed us to greatly simplify our model.

8. Discussion

Observed in our sensitivity analysis, we can see that changing β_0 has a significant impact on batch-to-batch variability, however, this is something that we can't observe as we do not model any Quality Control (QC) threshold γ. Though this is the case, we can assume, since lower β_0 values translates to having more consistency in our processes, that we would also be more likely to pass γ thresholds.

In our study, we see the effects that equipment variability has on our batch-to-batch for the drug substance protein purification process. Looking towards the future, we can observe more factors that may introduce more sources of variability such as:

- *Starting material variability*: The quality of the starting material, such as the protein source, can vary depending on factors such as seasonality, supplier, or production method. This can lead to differences in protein yield, purity, and activity.
- *Process variability*: Variations in the purification process, such as changes in pH, temperature, or buffer composition, can affect the yield and quality of the protein product.
- *Operator variability*: Variations in the skills and experience of operators can impact the quality of the protein product, especially in steps that require manual manipulation.
- *Environmental variability*: Environmental factors, such as temperature, humidity, or air quality, can impact the stability and quality of the protein product.
- *Analytical variability:* Variations in the accuracy and precision of analytical methods used to monitor the purification process can affect the assessment of protein yield and purity.

9. References

[1]. "Stochastic simulation model development for biopharmaceutical production process risk analysis and stability control", B Wang, W Xie, T Martagan, A Akcay, CG Corlu - 2019 winter simulation conference (wsc), 2019.