|  |
| --- |
| Analysis of Modeling WISP1 Binding to Collagen I Using a Markov Chain Monte Carlo Method |

|  |
| --- |
| Michael Fouts  12-9-2022 |

# Table of Contents

[Table of Contents 1](#_Toc123042304)

[Table of Figures 1](#_Toc123042305)

[Table of Tables 1](#_Toc123042306)

[Table of Equations 1](#_Toc123042307)

[Introduction 3](#_Toc123042308)

[Methodology 4](#_Toc123042309)

[Model Derivations 4](#_Toc123042310)

[Markov Chain Monte Carlo Evaluation 6](#_Toc123042311)

[Results and Discussion 9](#_Toc123042312)

[Evaluation of Bayes Factors 9](#_Toc123042313)

[Specific Binding Model Evaluation 10](#_Toc123042314)

[Non-Specific Binding Model Evaluation 13](#_Toc123042315)

[Combined Model Evaluation 16](#_Toc123042316)

[Conclusion 21](#_Toc123042317)

[References 22](#_Toc123042318)

## Table of Figures

[Figure 1: Data of Delay of Fibrogenesis [3] 6](#_Toc121501589)

[Figure 2: Model Comparison with Data 9](#_Toc121501590)

[Figure 3: Natural Log of Alpha vs MCMC Step (Specific Binding Model) 10](#_Toc121501591)

[Figure 4: Natural Log of KSB vs MCMC Step (Specific Binding Model) 11](#_Toc121501592)

[Figure 5: Gelman-Rubin Statistic for Specific Binding Model 12](#_Toc121501593)

[Figure 6: Acceptance Rate vs. MCMC Step (Specific Binding) 12](#_Toc121501594)

[Figure 7: Pairs Plot (Specific Binding Model) 13](#_Toc121501595)

[Figure 8: Natural Log of KNSB\*α vs MCMC Step (Non-Specific Binding Model) 14](#_Toc121501596)

[Figure 9: Gelman-Rubin Statistic vs MCMC Step (Non-Specific Binding Model) 14](#_Toc121501597)

[Figure 10: Acceptance Rate vs. MCMC Step (Non-Specific Binding) 15](#_Toc121501598)

[Figure 11: Natural Log Likelihood vs Natural Log KNSB\*α 16](#_Toc121501599)

[Figure 12: Natural Log Alpha vs MCMC Step (Combined Model) 17](#_Toc121501600)

[Figure 13: Natural Log KSB vs MCMC Step (Combined Model) 17](#_Toc121501601)

[Figure 14: Natural Log KNSB vs MCMC Step (Combined Model) 18](#_Toc121501602)

[Figure 15: Gelman-Rubin Statistic (Combined Model) 19](#_Toc121501603)

[Figure 16: Acceptance Rate vs. MCMC Step (Combined Model) 19](#_Toc121501604)

[Figure 17: Pairs Plot (Combined Model) 20](#_Toc121501605)

## Table of Tables

[Table 1: Bayes Factor Results 9](#_Toc121501606)

[Table 2: Maximum Likelihood Parameter Values (Specific Binding) 10](#_Toc121501607)

[Table 3: Maximum Likelihood Parameter Values (Combined Model) 16](#_Toc121501608)

## Table of Equations

[Equation 1: Rate of Fibril Formation 4](#_Toc121501609)

[Equation 2: Rate of Change in Monomer Concentration 4](#_Toc121501610)

[Equation 3: Rate of Change in Complex Concentration (Specific Binding Model) 4](#_Toc121501611)

[Equation 4: Pseudo-Steady State Specific Binding Model 4](#_Toc121501612)

[Equation 5: Monomer Balance 5](#_Toc121501613)

[Equation 6: Overall Specific Binding Model 5](#_Toc121501614)

[Equation 7: Rate of Change in Complex Concentration (Non-Specific Binding Model) 5](#_Toc121501615)

[Equation 8: Overall Non-Specific Binding Model 5](#_Toc121501616)

[Equation 9: Overall Combined Model 5](#_Toc121501617)

[Equation 10: Interpretation of Complex to Monomer Ratio to Time Delay in Reaction 6](#_Toc121501618)

[Equation 11: Final Specific Binding Model 6](#_Toc121501619)

[Equation 12: Final Non-Specific Binding Model 6](#_Toc121501620)

[Equation 13: Final Combined Model 6](#_Toc121501621)

[Equation 14: Likelihood of a Generic Normal Distribution 7](#_Toc121501622)

[Equation 15: Wishart Prior for Variance-Covariance Matrix 7](#_Toc121501623)

[Equation 16: Likelihood for a Wishart Distribution 7](#_Toc121501624)

[Equation 17: Bayes Factor 8](#_Toc121501625)

[Equation 18: Likelihood for Approximate Bayesian Computation 8](#_Toc121501626)

# Introduction

Understanding how proteins interact with their surroundings is of paramount importance for revealing their mechanisms and the underlying health implications these interactions cause. Collagen is an incredibly important protein, as it represents about 25% of the total mass in mammals and is found in all animals [1]. Monomers of collogen commonly form chains called fibrils that are then assembled into larger fibers that make up the basis of several tissues in the body. Specifically, the collagen I monomer (also called type I collagen), makes up the majority of the most abundant collagen fibrils, making it an important type of collagen to understand. Studies have indicated that collagen I has binding sites, a key piece of information in understanding how it functions. [2]

WISP1 is a matricellular protein that is secreted by tumor cells and alters the signaling of other biological reactions. WISP1 has been seen to bind with collagen I, causing collagen I to become more linear and consequently promoting breast cancer metastasis. Studies have shown that higher WISP1 expression in tumors correlates with faster progression to metastasis and worse patient outcomes. [3]

The purpose of this project is to investigate the mechanisms of WISP1 binding onto collagen I. By understanding the mechanisms behind the interactions between WISP1 and collagen I, additional therapies and intervention methods can be theorized and researched to help improve patient prognosis. Three specific models are proposed in this analysis: 1) WISP1 specifically binding to the collagen I sites. 2) WISP 1 non-specifically binding to collagen I. 3) There is a combination of WISP1 specifically binding and non-specifically binding to collagen I.

# Methodology

## Model Derivations

The fibril formation of collagen I is an entropy-driven self-assembly [1] and the change in concentration can be generally described by Equation 1. This equation details that the rate of fibril formation is both dependent on the concentration of the collagen I monomer (now referred to as monomer) and already existing fibril.

Equation : Rate of Fibril Formation

The rate of change in the collagen I monomer concentration can be generally described as having two components: 1) the rate at which fibrils are formed which consumes monomers and 2) the rate at which WISP1 interacts with collagen I to form a complex. The second component includes both the rate of WISP1 binding to collagen and the rate of disassociation of the complex into the original substrates. This relation can be shown in Equation 2.

Equation : Rate of Change in Monomer Concentration

Looking at data for the formation of fibrils of collagen I, it is proposed that the formation of complex happens at a much faster rate than the formation of fibrils. This leads to a critical assumption in our analysis that the formation of fibrils is at a dynamic equilibrium and does not significantly affect the change of concentration of collagen I monomer when dealing with the time scales associated with the formation of the WISP/collagen I complex (now referred to as simply, complex). With this assumption in mind, the three models can now be proposed to postulate the interaction between WISP1 and collagen I.

Model 1 details a model representative of WISP1 binding to sites on collagen I, referred to as the Specific Binding Model. For this model, collagen I is considered to be the limiting component in the reaction and WISP1 in excess for all concentrations. We can describe the formation of complex with the Specific Binding Model using Equation 3.

Equation : Rate of Change in Complex Concentration (Specific Binding Model)

An additional assumption is proposed for this interaction. Since the complex is an intermediate in the larger reaction of the formation of the fibrils and the complex forms quickly, it is chosen to assume that the complex concentration stays at a steady-state level through the fibril formation (known as a pseudo-steady state assumption). This allows the rate of change of complex formation with respect to time to be assumed as zero as shown in Equation 4.

Equation : Pseudo-Steady State Specific Binding Model

With the experimental setup, it is difficult to directly monitor the amount of monomer free in reaction at any point in time; however, a collagen balance can be performed to relate the free monomer to more easily measured quantities, shown in Equation 5.

Equation : Monomer Balance

Combining Equation 4 and Equation 5, rearranging to solve for the ratio of complex to total monomer, and combining related rate parameters into an overall rate parameter K­­­­SB­­, Equation 6 shows the overall model used to represent specific binding.

Equation : Overall Specific Binding Model

Model 2 represents non-specific binding and it assumes that the rate of complex formation is not dependent on the sites available on the monomers but instead, just the total amount of monomer present. This can be represented with Equation 7.

Equation : Rate of Change in Complex Concentration (Non-Specific Binding Model)

Applying the pseudo steady state assumption and consolidating the rate constants into a single rate constant, the non-specific binding model can be derived and is shown in Equation 8.

Equation : Overall Non-Specific Binding Model

The third model, known as the combined model, integrates the components of the specific and non-specific binding models by adding their terms together and can be seen in Equation 9.

Equation : Overall Combined Model

One issue that arises with these models is that the data provided does not explicitly give the ratio of complex to total monomer present so an analog must be proposed in order to evaluate the three models derived. Data was given in relation to fibril formation from Collagen I monomers using optical density as an analog for the percentage of initial monomer used in the reaction with respect to time. Varying amounts of WISP1 were added to these solutions and the time delays were recorded. [3] The data can be seen in Figure 1: Data of Delay of Fibrogenesis [3].

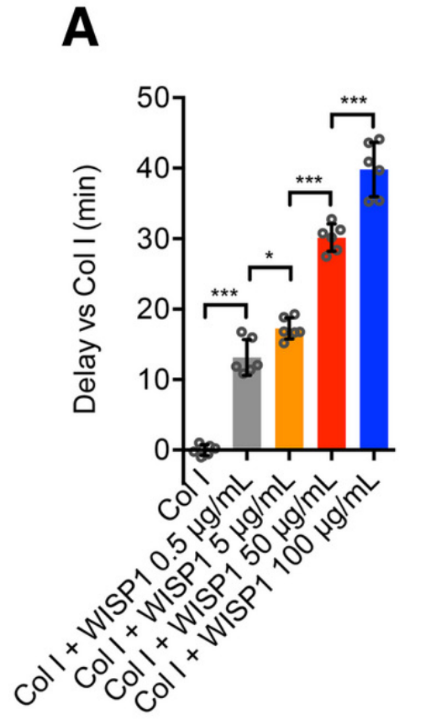


Figure : Data of Delay of Fibrogenesis [3]

The assumption made for this analysis was that there is a linear relationship between the time delay of the reaction and the ratio of complex to the total monomer. This relation can be seen in Equation 10 using a parameter α.

Equation : Interpretation of Complex to Monomer Ratio to Time Delay in Reaction

With this relation, the final equations used in this analysis can be derived and are shown in Equation 11: Final Specific Binding Model, Equation 12: Final Non-Specific Binding Model, Equation 13: Final Combined Model.

Equation : Final Specific Binding Model

Equation : Final Non-Specific Binding Model

Equation : Final Combined Model

## Markov Chain Monte Carlo Evaluation

To evaluate the parameters of the three models, a Bayesian approach was selected because it allows for model comparisons based on calculations for probability of each hypothesized model given the measured data. This is more helpful for our analysis than traditional frequentist approaches which calculate a confidence range for parameters and simply accepts or rejects null hypotheses because this analysis is more interested in the evidence comparisons between models. A Markov Chain Monte Carlo (MCMC) method was chosen to evaluate each of the models using an Adaptive Metropolis-Hastings algorithm for the proposals because this allows for the sampling from the posterior distribution without having to normalize by the evidence, one of the most difficult aspects of Bayesian inference.

This method generates proposals based on adding a random step size generated from a normal distribution to the last accepted parameter values of evaluation to generate a conditional likelihood of the data collected given our model and the proposed parameter set. If the likelihood is greater than that of the previous parameter set, the parameters are always accepted. If the likelihood is less than that of the previous parameter set, the parameters may be accepted with a probability related to the ratio of the current parameter sets likelihood to the previous parameter sets likelihood. This logic is then repeated for a predetermined number of steps. This method is adaptive because it aims to have a predetermined acceptance rate and if the acceptance rate is higher than the target, it will generate larger step sizes, and if the acceptance rate is lower than the target, it will take smaller step sizes. To ensure optimum efficiency of the Metropolis-Hastings proposal, an acceptance rate of 20% was selected when running these chains. This approach makes sure the posterior distribution is appropriately being explored.

For this problem, it was assumed that the measurement errors have a normal distribution with a mean of 0. Traditionally, this leads to a likelihood that is equal to Equation 14, where σ is the standard deviation of the observations and µ is the results from proposed models given a parameter set.

Equation : Likelihood of a Generic Normal Distribution

However, since the primary consideration of this analysis is to compare the models against each other and σ is unknown, this adds extra complexity to the analysis. If σ is assumed to some constant, it allows Equation 14 to be simplified to a proportionality to the sum squared error; however, this leads to possible issues with very similar likelihoods for very small values of sum squared error and can create issues with convergence. Therefore, for this analysis, it was chosen to use a likelihood relationship in the form of the Wishart distribution because it has a higher contrast in likelihoods at small sum squared errors and this contrast increases with more data. This allows MCMC to search more optimally in the parameter space. For this, a prior is assumed for the variance-covariance matrix as shown in Equation 15: Wishart Prior for Variance-Covariance Matrix where k is the number of observed variables in the data [4]. All other priors are to be considered uninformed priors. The proportionality for the Likelihood of the Wishart Distribution used in this analysis can be seen in Equation 16: Likelihood for a Wishart Distribution.

Equation : Wishart Prior for Variance-Covariance Matrix

Equation : Likelihood for a Wishart Distribution

The Gelman-Rubin statistic was chosen as a metric for the convergence of the MCMC chains. This approach compares the variance of the individual chains with the variance between the chains to determine when the chains are evaluating the same parameter space. Four chains are to be used in this analysis and a value of 1.1 or lower was used as the criteria for convergence.

To compare the models against one and other, a Bayes Factor was used to compute which model is favored and by how much. This is a ratio of a Monte Carlo integration of the posterior distribution. This is convenient for an MCMC approach as the likelihoods calculated are samples from the posterior distribution. The Bayes Factor comparing a generic set of models i and j can be seen in Equation 17: Bayes Factor.

Equation : Bayes Factor

When comparing the posterior distributions, the Monte Carlo integration can be calculated by taking the average of the likelihoods from the steps post burn in. While the number of points may vary depending on the application, typically 5,000 to 20,000 points is suggested. [5] The last 10,000 points of a 40,000 step chain will be selected from each chain for the integration. Consideration was also given to using samples from non-consecutive steps (for example, every third step) post convergence; however, the previous approach was selected due to its simplicity in comparison to this.

For the final Bayes Factor likelihoods, a more conservative approach was used. Using the Wishart Distribution with a large number of data points, the difference in likelihoods calculated can be very large for small parameter changes, which leads to a more focused search. However, this creates very exaggerated Bayes Factors. A more conservative estimate is the Approximate Bayesian Computation, which is proportional to the inverse of the sum squared error as shown in Equation 18: Likelihood for Approximate Bayesian Computation and is not dependent on the number of data points. Because of this, the Approximate Bayesian Computation likelihood was used in evaluating the Bayes Factors.

Equation : Likelihood for Approximate Bayesian Computation

# Results and Discussion

## Evaluation of Bayes Factors

The results of the analysis can be found in Table 1: Bayes Factor Results.

Table : Bayes Factor Results

|  |  |  |
| --- | --- | --- |
| **Model Comparison** | **Bayes Factor** | **Result** |
| Specific Binding / Non-Specific Binding | 3.2 | Positive evidence favoring Specific Binding |
| Combined / Specific Binding | 5.7 | Positive evidence favoring Combined |
| Combined / Non-Specific Binding | 18.3 | Positive evidence favoring Combined |

The results show positive evidence supporting the combined model as the best given the data. The maximum likelihood values were found for each model and the parameter set was then put into the models and the results are shown on Figure 2: Model Comparison with Data. In the figure, there is a sharp increase in time delay at low concentrations of WISP1, however at concentrations above ~5 µg/mL, the time delay to WISP1 concentration becomes linear and the combined model is the only one that properly identifies this relationship. The Specific Binding Model shows a leveling off of time delay at high WISP1 concentrations and the Non-Specific Binding Model shows a linear relationship between the WISP1 concentrations, neither of which fully reflects the data.

Chart, line chart

Description automatically generated

Figure : Model Comparison with Data

From a biological perspective, this suggests that at low concentrations, WISP1 chooses to bind to the sites located on Collagen I, representing a specific binding behavior; however, once the sites fill up on the Collagen I monomers, WISP continues to pile onto the surface, representing a Non-Specific Binding behavior.

## Specific Binding Model Evaluation

The Specific Binding Model had two parameters to consider, α and KSB. The maximum likelihood values can be seen in Table 2: Maximum Likelihood Parameter Values.

Table : Maximum Likelihood Parameter Values (Specific Binding)

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| α | 36.68 minutes |
| KSB | 3.918 µg/mL |

The MCMC search for the parameter values can be seen in Figure 3: Natural Log of Alpha vs MCMC Step (Specific Binding Model) and Figure 4: Natural Log of KSB­ vs MCMC Step (Specific Binding Model). The variance of the MCMC search with regards to α is much smaller than KSB. This shows that α has a larger effect on the likelihood than KSB and is the primary parameter of consideration.

Chart

Description automatically generated with medium confidence

Figure : Natural Log of Alpha vs MCMC Step (Specific Binding Model)

Chart

Description automatically generated

Figure : Natural Log of KSB­ vs MCMC Step (Specific Binding Model)

The Gelman-Rubin statistic can be seen in Figure 5: Gelman-Rubin Statistic for Specific Binding Model. Using the criteria that convergence happens at a value less than 1.1 (marked with the green line on the figure), convergence for this model happens before 5,000 steps, creating ~35,000 data points to be used to evaluate the likelihood integration for the Bayes Factor. However, Figure 6: Acceptance Rate vs. MCMC Step (Specific Binding) shows that the Metropolis-Hastings Proposal acceptance rate reached ideal acceptance rate of 0.2 for all chains around step 25,000, meaning the posterior distribution is being sampled more conservatively than desired before then.

Chart

Description automatically generated

Figure : Gelman-Rubin Statistic for Specific Binding Model

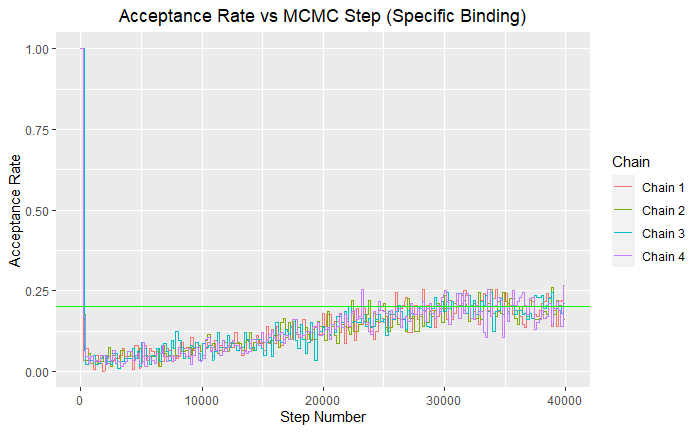


Figure : Acceptance Rate vs. MCMC Step (Specific Binding)

Evaluating the parameters, the data from the last 10,000 samples of each chain form a normal distribution, showing that the space is well explored. There is some correlation between KSB and α, suggesting that the relation between α and the time delay may interfere with calculating KSB and should actual data of complex to monomer total appear, the KSBvalues may be different. These insights can be seen in Figure 7: Pairs Plot (Specific Binding Model).

Chart, line chart

Description automatically generated

Figure : Pairs Plot (Specific Binding Model)

## Non-Specific Binding Model Evaluation

Due to the linear dependance of α and KNSB in the Non-Specific Binding Model, the variables cannot be evaluated independently and must be evaluated in combination. After evaluation, the maximum likelihood value for KNSBα is 0.4460 min·mL/µg. Figure 8: Natural Log of KNSB\*α vs MCMC Step (Non-Specific Binding Model) and Figure 9: Gelman-Rubin Statistic vs MCMC Step (Non-Specific Binding Model) both show the parameter search quickly converges. Additionally, Figure 10: Acceptance Rate vs. MCMC Step (Non-Specific Binding) shows that the acceptance rate reached 20% and that the chain was optimally exploring the space.

Chart

Description automatically generated with low confidence

Figure : Natural Log of KNSB\*α vs MCMC Step (Non-Specific Binding Model)

Chart

Description automatically generated

Figure : Gelman-Rubin Statistic vs MCMC Step (Non-Specific Binding Model)

Chart

Description automatically generated

Figure : Acceptance Rate vs. MCMC Step (Non-Specific Binding)

Because there is only one parameter used in this model, a proportional posterior likelihood can be viewed in two dimensions. Figure 11: Natural Log Likelihood vs Natural Log KNSB\*α shows the concentrated area of maximum likelihood for this model and gives insight that the likelihood is very steep in this case, mostly due to the large number of data points used that increase the slope of the likelihood function from the Wishart Distribution.

Chart, line chart

Description automatically generated

Figure : Natural Log Likelihood vs Natural Log KNSB\*α

## Combined Model Evaluation

The Combined Model had three parameters to consider, α, KSB, and KNSB. The maximum likelihood values can be seen in Figure 11: Natural Log Likelihood vs Natural Log KNSB\*α.

Table : Maximum Likelihood Parameter Values (Combined Model)

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| α | 17.37 minutes |
| KSB | 0.1700 µg/mL |
| KNSB | 0.01331 mL/µg |

The variance of the converged chains shows the significance of the parameters to the overall likelihood integration of the Model. Based on this, in order from most important to least important parameter to the contribution to the likelihood, it goes α, KNSB, and lastly KSB. These details can be seen in Figure 12: Natural Log Alpha vs MCMC Step (Combined Model), Figure 13: Natural Log KSB­ vs MCMC Step (Combined Model), and Figure 14: Natural Log KNSB vs MCMC Step (Combined Model).

Chart

Description automatically generated

Figure : Natural Log Alpha vs MCMC Step (Combined Model)

Chart, histogram

Description automatically generated

Figure : Natural Log KSB­ vs MCMC Step (Combined Model)

Chart

Description automatically generated

Figure : Natural Log KNSB vs MCMC Step (Combined Model)

With the most parameters, the combined model takes the longest to converge due to the exponential nature of dimensionality. Based on the Gelman-Rubin Statistic seen in Figure 15: Gelman-Rubin Statistic (Combined Model), the chain doesn’t converge until step ~30,000, meaning 10,000 steps can be used to integrate the likelihood for the calculation of the Bayes Factor. Additionally, Figure 16: Acceptance Rate vs. MCMC Step (Combined Model) shows that the acceptance rate does converge around 20% and is optimally exploring the posterior distribution.

Chart, histogram

Description automatically generated

Figure : Gelman-Rubin Statistic (Combined Model)

Chart

Description automatically generated

Figure : Acceptance Rate vs. MCMC Step (Combined Model)

Evaluating the parameters, the samples show a normal distribution, again, showing that the space is well explored. There is a very strong correlation between KNSB and α, which plays into the same issue seen in the Non-Specific Binding model where the parameters cannot be calculated dependent of one and other. There are smaller correlations between the other parameters, but as a rule of thumb for Pearson’s Correlation Coefficients, anything below an absolute value of 0.7 is not highly correlated. All these evaluations can be seen in Figure 17: Pairs Plot (Combined Model).

Chart

Description automatically generated

Figure : Pairs Plot (Combined Model)

# Conclusion

The evidence suggests that the Combined Model is the best of the three considered. This is because the data suggests a dual Specific and Non-Specific Binding nature of WISP1 to Collagen I, depending on the WISP1 concentration present. This information may be used to help with intervention methods in breast cancer metastasis by understanding that the dominant binding mechanisms differ depending on the concentration of WISP1 secreted from tumors.

# References

1. Klinke DJ. Project: Collagen Fibrinogenesis (Project Statement)
2. Prockop D.J., Fertala A. Inhibition of the self-assembly of collagen I into fibrils with synthetic peptides. Demonstration that assembly is driven by specific binding sites on the monomers. J Biol Chem. 1998;273:1559815604
3. Jia H, Janjanam J, Wu SC, Wang R, Pano G, Celestine M, Martinot O, Breeze-Jones H, Clayton G, Garcin C, Shirinifard A, Zaske AM, Finkelstein D, Labelle M. The tumor cell-secreted matricellular protein WISP1 drives pro-metastatic collagen linearization. The EMBO Journal (2019)38:e101302
4. Box, G. E. P. & Draper, N. R. The bayesian estimation of common parameters from several responses. *Biometrika* **52,** 355–365 (1965).
5. Quinlan, Brenda. “Dimensional Analysis: How Many Monte Carlo Simulations Should I Run? Part 2.” Engineering Talk - DCS's Blog about Tolerance Analysis, Dimensional Control Systems, Inc., 26 Jan. 2015, https://blog.3dcs.com/dimensional-analysis-how-many-monte-carlo-simulations-should-i-run.