1. Introduction

In various chemical industries, a substantial amount of effort is made to scale-up the production of new chemical products from laboratory benchtop scale to commercial manufacturing capacities. These increases in scale often introduce a variety of new risks, operational and scale-sensitive, that can jeopardize the critical quality attributes of the output chemical product. This is especially a concern in the pharmaceutical industry, where the manufacturing active pharmaceutical ingredients (APIs) occurs under due regulatory scrutiny. For example, process research and development teams are held accountable for proving how processing parameters – their setpoints and deviations – will relate to important features of the drug product (e.g. yield, color, purity, identity and quantity of impurities) as well as for many intermediate stages throughout the synthetic process. Quite often, tens of millions of dollars are invested in related research activities over five to ten years with teams of dozens to hundreds of scientists assigned to the submission of a single new drug application.

Because new pharmaceutical compounds are inherently novel and unstudied, the understanding of their production processes is crucially underpinned by the generation of volumes of original laboratory data. It is thus always of critical interest to pharmaceutical process research to find methods of developing and scaling processes faster and more efficiently. A common approach is to use carefully designed laboratory experiments and data to build in-silico process models that can relieve future experimental burden and provide a statistical framework for process characterization. Although each new compound involves never-before studied chemical transformations, a case may be made to leverage prior information on *classes* of such transformations in a Bayesian framework to build models with an even lesser burden on these extremely resource-intensive experimental campaigns.

Chemical production processes are, at a high level, a sequence of steps each stringing together a reaction (combining reactants to make more complex products) and a separation (purification of desired products before entering subsequent steps). A majority of research effort in pharmaceutical development is placed on understanding the reactions themselves. Separations can be better understood a-priori because they tend to be more dependent on known properties of well-studied solvents. Models that describe chemical reactions are often "reaction kinetic models" which quantify the consumption and production rates of species over time using a network of differential equations. We desire to estimate reference "rate constants" (k_{ref}) and "activation energies" (E_A) on which the model equations depend to describe rates of change based on species concentration and temperature in the system.

2. Model Overview and Approach

To illustrate a Bayesian approach for kinetic parameter estimation, I have chosen data from a simple reaction that has been blinded to protect proprietary information. The blinded scheme can be presented as the simplified two-step sequential reaction from starting material (**R**) to product (**P**), through intermediate (**I**), as shown in the simple Equations 1 and 2.

(Eq 1.) $Reaction\ 1: R \rightarrow I\ (given\ k_1, E_{A,1})$

(Eq 2.) $Reaction\ 2: I \rightarrow P\ (given\ k_2, E_{A,2})$

The model rate equations are derived from fundamental chemical kinetic theory to provide the mass balance Equations 3, 4, and 5. The observed rate constant at any temperature is calculated from a reference rate constant and activation energy as shown in Equation 6.

(Eq 3.) $$\frac{dC_R}{dt} = -k_1 C_R$$

(Eq 4.) $$\frac{dC_I}{dt} = k_1 C_R - k_2 C_I$$

(Eq 5.) $$\frac{dC_P}{dt} = k_2 C_I$$

(Eq 6.) $$k_j = k_{ref,j} \exp\left[-\frac{E_{A,j}}{R}\left(\frac{1}{T} - \frac{1}{T_{ref}}\right)\right]$$

…where C_i is the concentration of species ***i***; t is time, k_j is the realized rate constant of reaction ***j***; R is the universal gas constant; T is the experimental temperature; and $k_{ref,j}$, $T_{ref,j}$, $E_{A,j}$ are the reference rate constants, reference temperatures, and activation energies for reaction j.

In order to evaluate model parameters (i.e. in a regression), the above rate equations are integrated in an initial-value problem using the initial concentrations of each species and the temperature of the reaction. For example, we may be interested in examining how this reaction proceeds from an initial concentration of reactant **R** at 0.5 mol/L and a temperature of 40 °C, given some set of parameters (e.g. k_{ref} = 1.0E-3 mol/L-sec at 30 °C [T_{ref}], E_A = 50 kJ/mol). The solution to the network of ODEs will give a model response that is dependent on state variables $\boldsymbol{X}$ (concentration, temperature) and the parameters of interest $\boldsymbol{\theta}$ (rate constants, activation energies) as expressed in Equation 7 for response *i* and measurement time point *j*:

(Eq 7.) $$\hat{y}_{ij}(t) = X_i(t_j|\boldsymbol{\theta})$$

We will often make the Gaussian assumption that the measured response is normally distributed around the mean $X_i(t_j|\theta)$ with some measurement error σ_y. Equation 8 thus shows the resulting parameter likelihood, with respect to our measured data points y_{ij}:

(Eq 8.) $$L(\boldsymbol{\theta}|y) = \prod_{i=1}^{n_{resp}} \prod_{j=1}^{n_t} \frac{1}{\sigma} \exp\left\{\frac{\left[y_{ij} - X_i(t_j|\boldsymbol{\theta})\right]^2}{2\sigma_y^2}\right\}$$

The conventional, frequentist approach is to maximize likelihood by minimizing the sum of squared residuals, as defined in Equation 9.

(Eq 9.) $$SSR(\theta) \propto \sum_{i=1}^{n_{resp}} \sum_{j=1}^{n_t} \left[y_{ij} - X_i(t_j|\boldsymbol{\theta})\right]^2$$

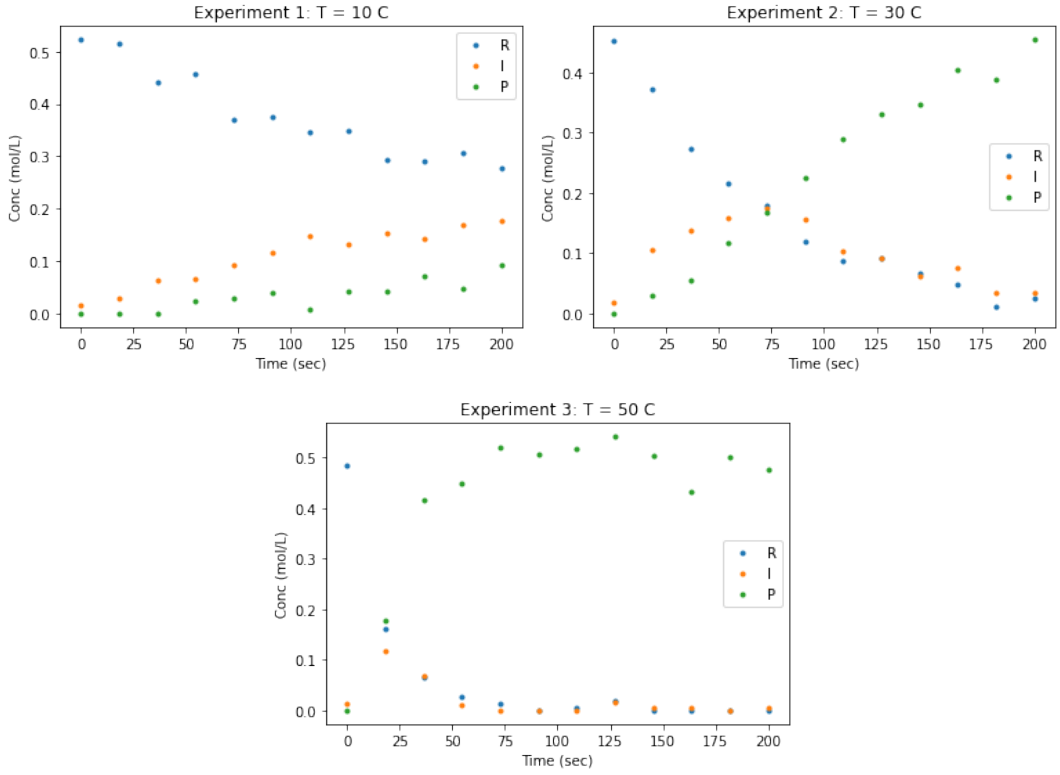
We may instead consider a Bayesian approach, whereby we may infer the parameter set, perhaps more efficiently, by assigning priors which are combined with our observation likelihood to generate a posterior distribution. This proportionality is expressed in Equation 10.

(Eq 10.) $\pi(y|\theta) \propto L(\theta|y) * \pi_0(\theta)$

Since chemical reactions can be classified, we may rely on expert chemist opinion to assign a prior distribution for parameters based on knowledge of similar reactions from the past. This potentially allows to incorporate such experiential information as "the intermediate formation is fast, but its conversion to product is slower," "the half-life of intermediate formation is approximately 14 hours," "the reaction tends to double in rate with a 5 °C increase in temperature," et cetera. These are potentially valuable knowledge which are commonly available in practice. A Bayesian approach allows us to tap into this. For the purposes of this demonstration, we will assume various values for the prior distribution as detailed in Tables 2 and 4 alongside the following discussion.

3. Data

I have selected three representative "reaction profiles" from the available dataset which illustrate how the stepwise conversion of **R** to **P** through **I** proceeds over time at three different temperatures (10 °C = 283 K, 30 °C = 303 K, 50 °C = 323 K).



(Fig. 1) Reaction profiling data, concentration over time, for experiments at three temperature setpoints

There is a noticeable amount of noise in the data, but there is enough sensitivity in the profiles (conversion over time and difference in rates of conversion versus temperature) to provide reasonable estimates for the four parameters of interest. For simplification, we will

assume that the Gaussian measurement error (σ_y) is a constant value and equal for each species. This is practically appropriate since all three species are measured with the exact same methodology (at-line sample capture and dilution; separation and quantification by UPLC-UV).

4. Frequentist vs. Bayesian Approach

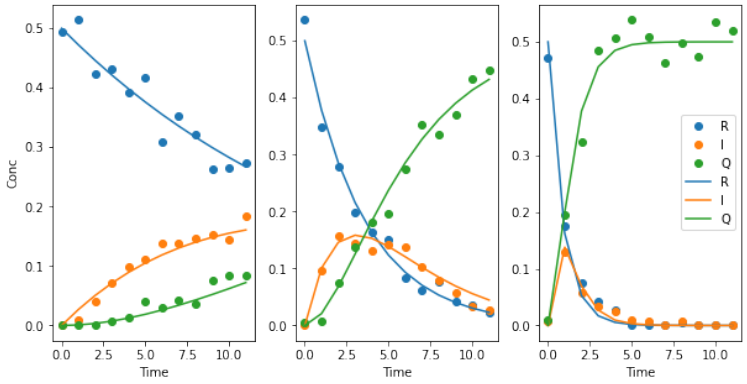
4A. Maximum Likelihood Estimation

The conventional, frequentist approach to estimating the kinetic parameters is available in several software packages but has been replicated simply using Python in the attached code. A function is written to calculate the residuals, and the SciPy *Optimize.Least_Squares* function uses this to find the best-fit parameters that minimize the sum of squared residuals, as introduced in Section 2.

The results from this least-squares regression are summarized in Table 1. As expected, the lower sensitivity to reaction 2 (a consequence of the sequential reaction scheme) provides a lower confidence in the parameters for reaction 2 than 1. However, the results are still overall reasonably confident despite this and the noticeable measurement noise.

(Table 1) Results of MLE regression

Parameter	**log10($k_{ref,1}$)**	**log10($k_{ref,1}$)**	**$E_{A,1}$**	**$E_{A,2}$**
Max. Likelihood Estimate	-1.813	-1.690	56.67	62.31
Variance	9.798E-5	2.603E-4	1.324	7.095
(Linearized) Relative 95% Conf. Int.	±1.07%	±1.87%	±3.98%	±8.38%



(Fig. 2) Overlay of best-fit MLE regression with experimental data

4B. Bayesian Inference with Informative Priors

The Bayesian approach to the same problem relies on an identical calculation of squared residuals, leading to a value which is proportional to the likelihood function as follows based on the description in section 2. Note that we are treating measurement error

as a constant, so that this parameter simplifies out of the proportional likelihood function (Equation 10) that is necessary for sampling the posterior.

$$\textit{(Eq 10.)} \quad L(\theta) \propto \exp\left[\frac{(y-\hat{y}(\theta))^2}{\sigma_y}\right] \propto \exp\left[-SSR(\theta)\right]$$

The discriminating feature, of course, is the introduction of prior distributions to each of the parameters. Based on expert opinion from historical experience with similar reactions, informative prior distributions were established for this demonstration. These distributions are described in the table below.

(Table 2) Example prior distributions on kinetic parameters

Parameter	Distribution Family	Mean	Standard Dev.	Units
$\log_{10}(k_{ref})$	Normal	-1.0	1.0	1/sec
E_A	Normal	50	25	kJ/mol

Note that each class of parameter is set to share the same prior distribution, as expert opinion is unlikely to provide valuable information at a more granular level. This type of information might be inferred from chemical structures (e.g. **I** might be less reactive than **R** for steric or electronic reasons) but is not usually available with any certainty.

Also, based on chemistry expertise, we suppose that the prior should include no co-dependence of parameters. Fundamentally, these parameters are independent by definition, and any apparent covariance is expected only to be a feature of experimental design, estimation technique, or other modeling choice. In light of this, the joint prior distribution can be summarized with the following mean vector and covariance matrix:

$$\boldsymbol{\mu_\theta} = [-1.0, -1.0, 50, 50]$$

$$\boldsymbol{\Sigma_\theta} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 625 & 0 \\ 0 & 0 & 0 & 625 \end{bmatrix}$$

Unfortunately, there is an absence of software for Bayesian inference that can conveniently handle chemical kinetic data. It is possible to handle differential equations in the Python package *PyStan*; however, it is extremely difficult or impossible to implement complex functions (i.e. multiple co-dependent ODEs returning a grid of solutions) within the model specification framework. Because of this, I opted to use a built-from-scratch basic Metropolis-Hastings random walk algorithm to sample the posterior distribution.

It is known for Metropolis-Hastings algorithms that tuning the proposal distribution for optimally efficient sampling can become difficult especially in multiple dimensions. It is difficult to detect convergence problems, and my initial attempts involved trial and error to produce acceptable results. However, these were very dependent on the inputs, and ultimately I implemented an adaptive method to update the proposal distribution. It is

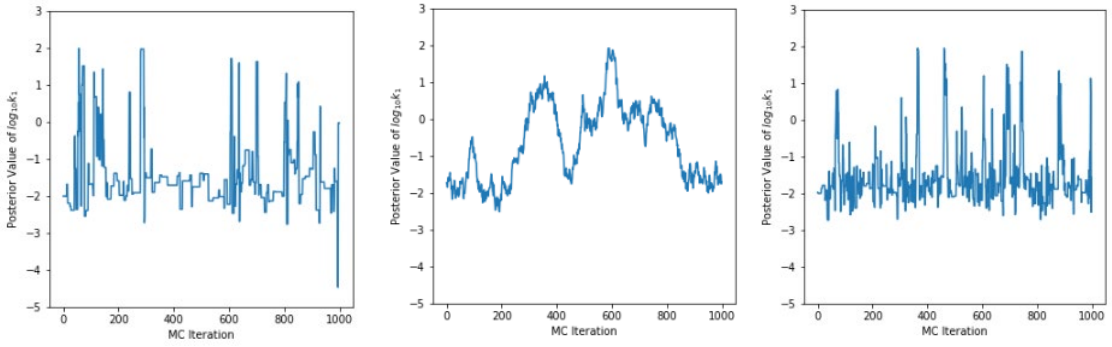
widely published that the optimal scale of a proposal distribution, sampling from a Gaussian target, can be expressed as a function of sample variance and dimensionality according to Equations 11 and 12[1, 2, 3]. A demonstration of sub-optimal and optimal scales of proposal distribution is shown in Figure 3.

$$\textit{(Eq 11.)} \quad \epsilon \sim MVN(0, \Sigma_P)$$

$$\textit{(Eq 12.)} \quad \Sigma_P = \frac{2.38^2}{d}\Sigma_\theta$$

The adaptive random-walk MH algorithm I have written is summarized as follows:

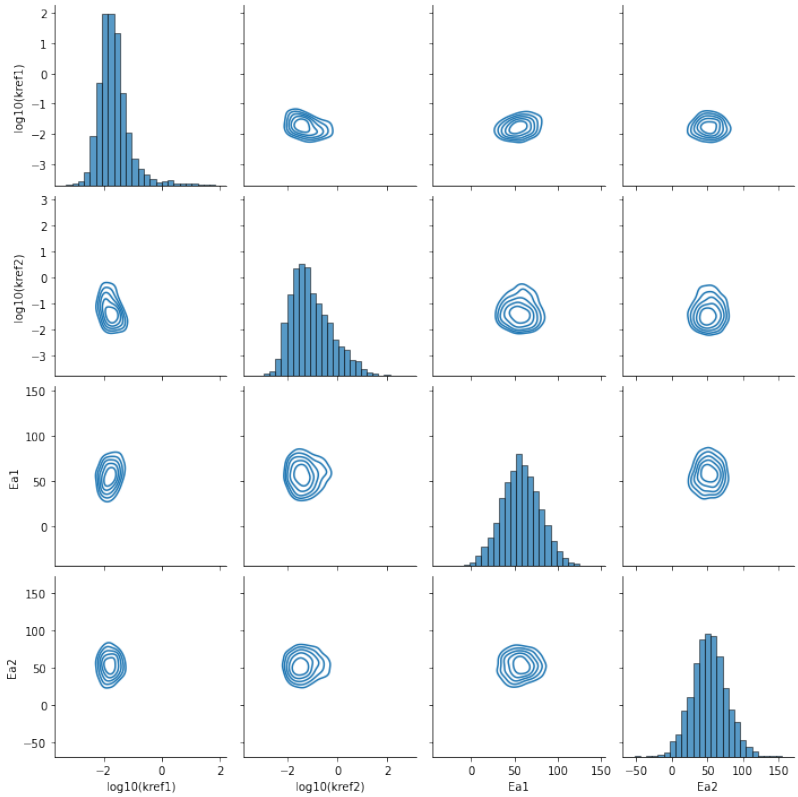
1. **Initialize with reasonable parameter and covariance estimates**
   a. $\boldsymbol{\theta_{n=0} = [-1, -1, 50, 50]}$
   b. $\boldsymbol{\Sigma_\theta = [1, 1, 625, 625]}$
2. **Repeat for B burn-in/warm-up iterations**
   a. **Generate multivariate random walk using the optimal scale of the parameter covariance estimate**
      i. $\boldsymbol{\epsilon \sim MVN\left(0, \frac{2.38^2}{d}\Sigma_\theta\right)}$
   b. **Propose new parameter values from random walk** $\boldsymbol{\theta_n^* = \theta_{n-1} + \epsilon}$
   c. **Calculate posterior proportional values for the current parameters and proposed parameters**
      i. $\boldsymbol{\pi_n^*(\theta_n^*|y) \propto L(\theta_n^*|y) * \pi_0(\theta_n^*)}$
      ii. $\boldsymbol{\pi_n(\theta_n|y) \propto L(\theta_n|y) * \pi_0(\theta_n)}$
   d. **Calculate the acceptance ratio**
      i. $\boldsymbol{\rho = \frac{\pi_n^*(\theta_n^*|y)}{\pi_n(\theta_n|y)} \wedge 1}$
   e. **Accept proposal** $\boldsymbol{\theta_n^*}$ **with probability ρ, else retain** $\boldsymbol{\theta_n = \theta_{n-1}}$
   f. **Adapt the proposal distribution with an updated estimate of parameter covariance using accepted steps of current chain**
3. **Retain the adapted proposal distribution from the burn-in samples B, then discard that chain**
4. **Repeat steps 2a – 2e for M final iterations**



(Fig. 3) Example of under-sampling (left, acceptance = 16%, $\Sigma_P \approx (8^2/d)\,\Sigma_\theta$), over-sampling (middle, acceptance = 96%, $\Sigma_P \approx (0.1^2/d)\,\Sigma_\theta$), and optimized sampling (right, acceptance = 42%, $\Sigma_P \approx (2.38^2/d)\,\Sigma_\theta$)

With this optimized algorithm, I was able to achieve the expected efficiency of 25-40% acceptance (a function of dimensionality). For all simulations, I ran at least 10K burn-in

iterations and 100K final iterations in a practicable amount of time to produce well-behaved posterior distributions.



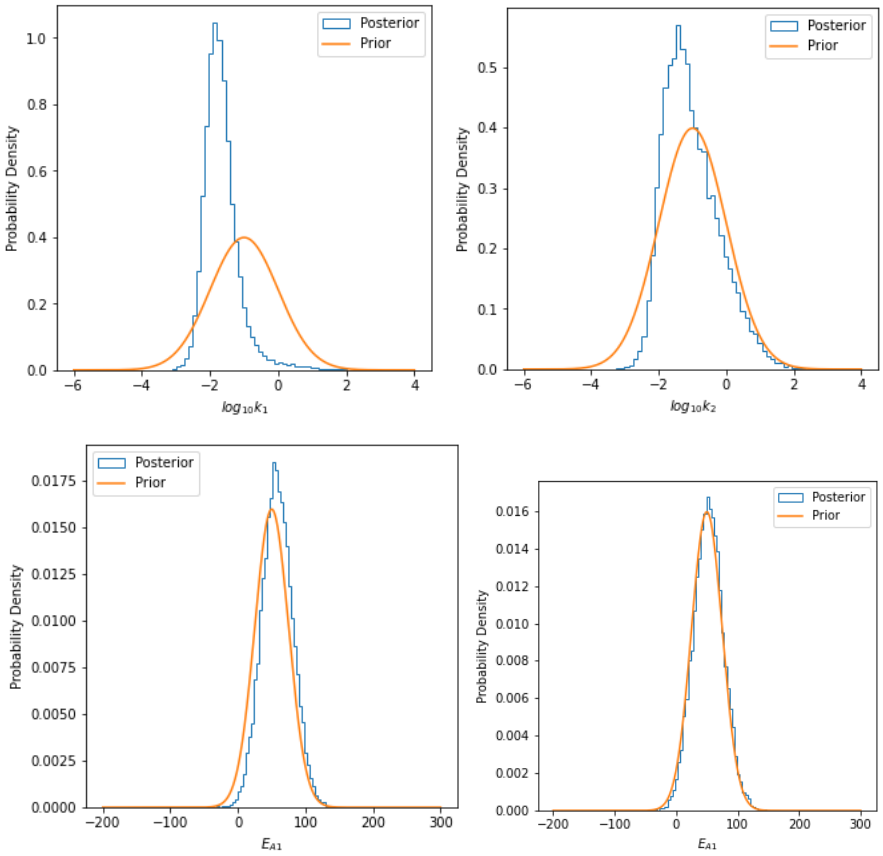
(Fig. 4) Univariate histograms and bivariate density contour plots of the 4-D joint posterior, constructed from normal priors

In Figure 4, I have plotted the 1-D histograms and 2-way joint density contours from the resulting accepted posterior samples. Summary statistics and equitailed credible sets are available in Table 3. The approximate maximum a-posteriori (MAP) estimate is found by binning the distributions into 50 bins then selecting the bin of highest frequency. This is expected to suffice as an estimate of the posterior mode and provide the best agreement to the maximum likelihood estimate.

(Table 3) Results of Bayesian inference, using normal priors

Parameter	**log10($k_{ref,1}$)**	**log10($k_{ref,2}$)**	**$E_{A,1}$**	**$E_{A,2}$**
Posterior Mean	-1.667	-1.039	57.59	52.03
Prior Variance	1	1	625	625
Posterior Variance	0.291	0.818	513	603
95% Equitailed Credible Set	-2.444 to -0.259	-2.280 to 0.818	14.01 to 102.79	4.53 to 100.57
Posterior Pseudo-MAP Estimate	-1.910 to -1.792	-1.491 to -1.364	50.99 to 54.91	48.65 to 52.85
MLE Estimate	-1.813	-1.690	56.67	62.31

As expected, the Bayesian MAP estimates and Frequentist MLE estimates are in reasonable agreement. The discrepancy is attributed to the informative prior, which has its own maximum that distorts the MAP estimate from the position of pure maximum likelihood. The mean of the distribution is a distinctly different value, especially for the distributions which exhibit skew. Notably, we can see a major disagreement in the estimates of parameter variance from the two approaches. It is quite likely that the frequentist approach is extremely underestimating this value, as the variance estimate comes from a linear approximation of the covariance matrix that is calculated only locally using the gradient at the best-fit values. Since the fit is in fact clearly very nonlinear, this approximation proves to be a bad one. On the other hand, the much wider Bayesian estimate of variance makes more practical sense and is easily interpreted visually. The Bayesian update to the prior are visually represented in Figure 5.

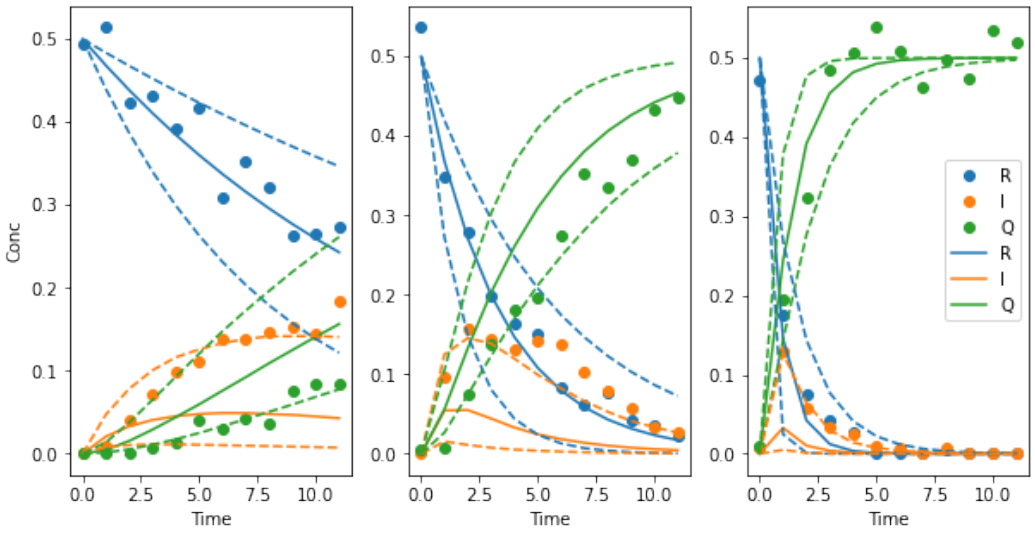


(Fig. 5) Univariate histograms of the 4-D joint posterior overlaid with the initial normal priors

Also quite interesting is the relative movement of each parameter from its priorThe parameters which had the lowest MLE variance saw the most movement from prior mean to posterior mean, as well as for the scaling of variance (ranging from 30 to 96% of prior variance). Of course, this makes sense as the posterior precision can be calculated as a

weighted sum of the prior and likelihood precision. Colloquially speaking, the evidence for the parameters of reaction 1 is much more compelling than that for reaction 2, and the rate is more easily inferred than activation energy from the kinetic profile. The former is the mathematical result of serial reactions, whereby sequential reactions are dependent on their predecessors and often involve lower-concentration species with thus lower magnitude sensitivity to change. Since we have more evidence for reaction 1, Bayes theorem applies a stronger update to the prior information.

It is worth noting that more accurate confidence boundaries can also be calculated for the frequentist case using estimates of measurement error and χ^2 tests, and I would expect these to better agree with the presented Bayesian approach, since ultimately they rely on an identical calculation of residuals. However, this is not a default feature of any common chemical kinetics estimation software. As such, although we can employ probabilistic features to make the conventional frequentist approach more informative and accurate, it is important to note that this is a default, intrinsic feature of every Bayesian approach.



(Fig. 6) Overlay of Bayesian prediction, using median and 50% equitailed credible sets, with experimental data

Using the posterior parameter distribution estimated in this exercise, the posterior predictive distribution was calculated for each species in these three experiments. The results, using the 25-50-75 percentiles, are shown in Figure 6. Unlike the case with the MLE prediction, we do not see an excellent overlay of the data with the predictions. This is an expected result for the Bayesian estimation, for which we are jointly capturing both data and prior rather than just maximizing to the likelihood of the data. With a surplus of data with low variance, the prior would eventually exert less influence, and the Bayesian prediction would approach that of the MLE. Still, this is an important aspect of the Bayesian approach which is important to capture and might cause some discomfort in this field where scientists are typically focused on the prediction of data rather than the estimation of parameters.

4C. Bayesian Inference with Non-Informative Priors

As a final exercise in the Bayesian approach, I chose to try non-informative (uniform) priors. The distribution hyperparameters are summarized in Table 4. The results from this exercise are presented in Table 5 and Figure 7.

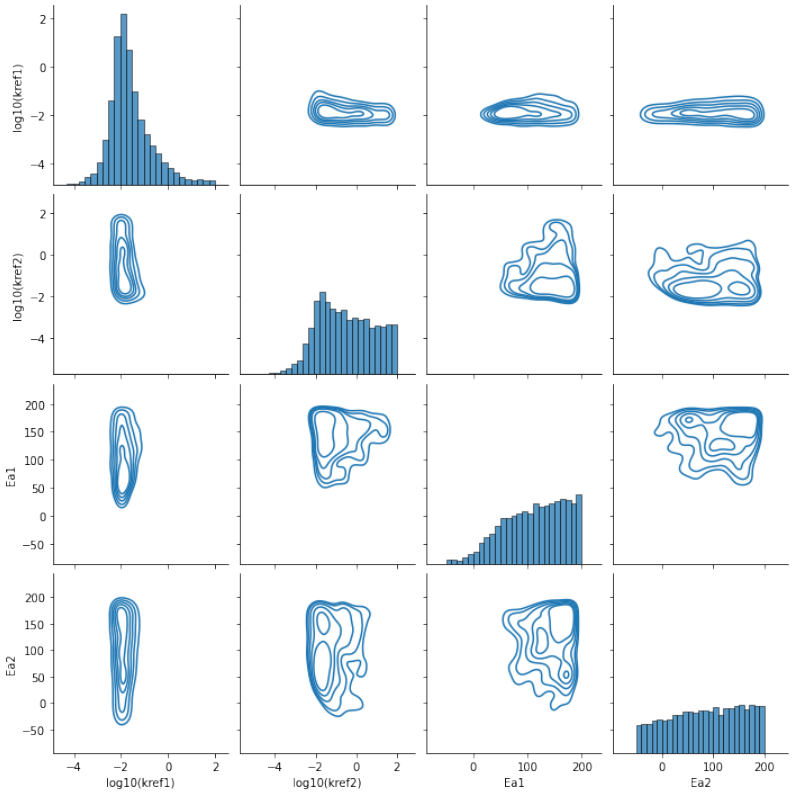
(Table 4) Example non-informative prior distributions on kinetic parameters

Parameter	Distribution Family	Lower Limit	Upper Limit	Units
$\log_{10}(k_{ref})$	Uniform	-5.0	2.0	1/sec
E_A	Uniform	-50	200	kJ/mol

(Table 5) Results of Bayesian inference, using uniform priors

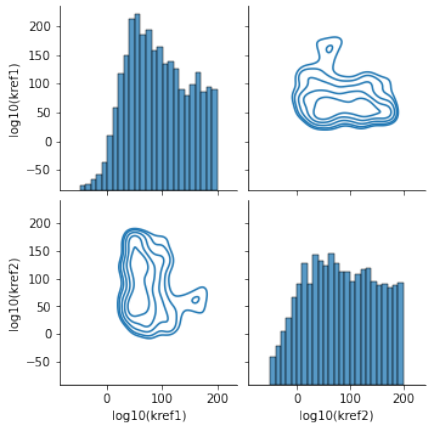
Parameter	$\log10(k_{ref,1})$	$\log10(k_{ref,2})$	$E_{A,1}$	$E_{A,2}$
Posterior Mean	-1.617	-0.515	114.02	84.67
Prior Variance	4.083	4.083	5,208	5,208
Posterior Variance	0.817	1.903	3,165	4,864
95% Equitailed Credible Set	-3.148 to 0.684	-2.958 to 1.848	-1.76 to 196.09	-40.23 to 193.95
Posterior Pseudo-MAP Estimate	-2.025 to -1.886	-1.768 to -1.629	189.98 to 194.98	174.99 to 179.99
MLE Estimate	-1.813	-1.690	56.67	62.31

Similar to the case with normal priors, the MAP estimates match well with the MLE estimates while the posterior means are offset by the skewness of the distributions. However, there is the noteworthy exception of the final two parameters (E_{A1} and E_{A2}). When either of these parameters are sampled jointly with the rest of the set, we find that the maximum density is severely distorted by covariance of parameters. This phenomenon is very clear in the contour plots of Figure 7, where one can observe these contours that closely follow the corners of the 2D "box" imposed by the square of uniform priors from both parameters. The very limited evidence for these particular parameters, at least in combination with other parameters' distributions, does not substantially change the non-informative nature of the priors. The end result is that these parameters cannot be estimated easily with Bayesian methods if done jointly with the full parameter set. The non-informative approach thus makes this estimation task more difficult than with the conventional MLE approach.



(Fig. 7) Histograms and bivariate KDE contour plots of the joint posterior, constructed from uniform priors

As a point of illustration, if only these two least-confident parameters are sampled together, the issue is mitigated as the distributions tighten enough so that we can reasonably select a posterior mode which is not at the end of the domain and agrees better with the MLE estimate as expected. This can be seen in Figure 8 with further statistics in Table 6. Of course, this is impractical because we would need to perform this sampling while fixing the other two parameters, which would be seriously inadvisable if we truly had no informative prior experience.



(Fig. 8) Histograms and contours of E_{A1} and E_{A2} sampled together from uniform priors

An alternative approach could be to initially sample the full set, fix k_{ref1} and k_{ref2} with an appropriate Bayesian estimator, then sample the remaining parameters. This is actually a relatively common approach in conventional MLE fitting of chemical kinetics. It is often the case that the rate constant k_{ref} and activation energy E_A will have high covariance for sparse, noisy, or poorly designed datasets so that a stepwise fitting is occasionally admissible. However, presenting a conditional fit (i.e. $E_{Ai} \mid k_{ref,i}$) is not ideal and can be difficult to communicate effectively.

(Table 6) Results of Bayesian inference, E_{A1} and E_{A2} sampled together from uniform priors, given fixed estimates of k_{ref1} and k_{ref2}

Parameter	**$E_{A,1}$**	**$E_{A,2}$**
Posterior Mean	94.97	86.80
Prior Variance	5,208	5,208
Posterior Variance	3142	4266
95% Equitailed Credible Set	-1.69 to 194.08	-30.74 to 194.15
Posterior Pseudo-MAP Estimate	66.40 to 71.35	35.12 to 40.12
MLE Estimate	56.67	62.31

Fortunately, by the time a reaction reaches the point in its development for a model to be proposed, it is quite likely that a chemistry team will have enough prior experience to express at least a mildly informative prior. Thus, although a non-informative prior can clearly create problems for Bayesian approach to certain datasets, it is unlikely that this would ever be a necessary approach.

5. Conclusions

The conventional methodology for maximum-likelihood estimation of reaction kinetic parameters was compared with a Bayesian approach. The former is orders of magnitude faster and is accessible in a variety of commercial software but tends to lead to imprecise estimates of parameter variance. This may be addressed with probabilistic augmentations although these are rarely performed in practice. On the other hand, the Bayesian approach provides an inherent and far more realistic treatment of probability at the high cost of end-user expertise in statistics and programming.

Apart from the mathematical advantages of a Bayesian approach, there is the elegance of jointly capturing expert (prior) opinion and new data. Even in the industry of pharmaceutical process development, where original data is paramount to novel research, there is yet some untapped value in the careful selection of informative priors. In such a way, Bayesian modeling may further the R&D interests of maximizing knowledge capture while minimizing resource expenditure.

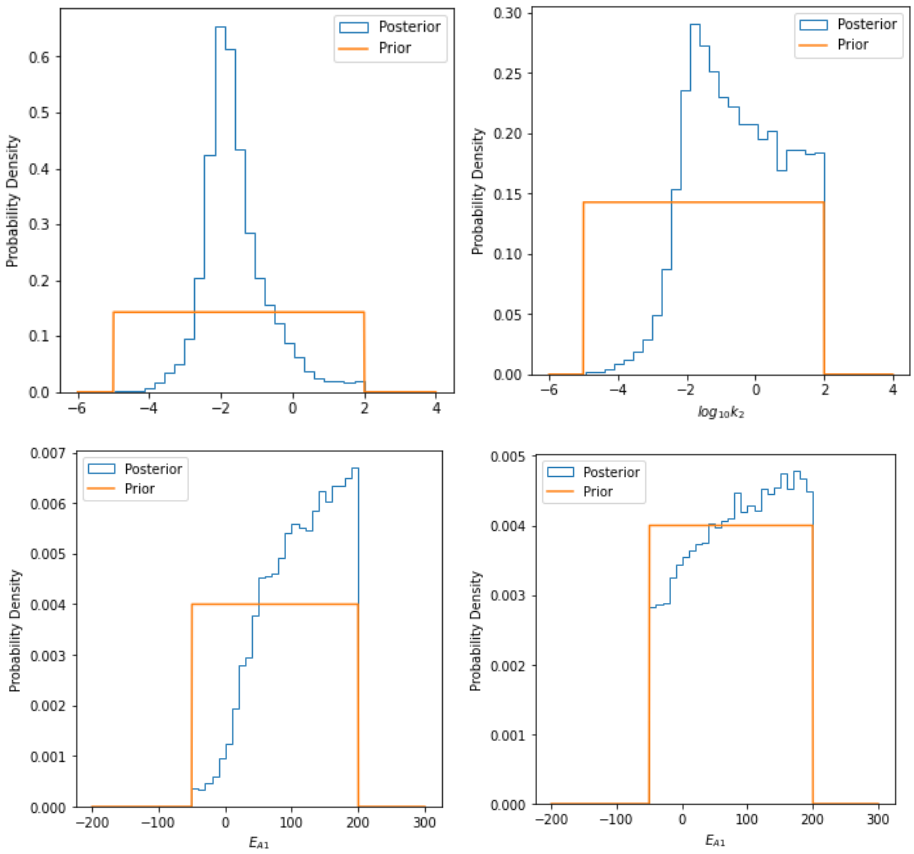
These conclusions unfortunately highlight the gap which exists for Bayesian process modeling involving differential equations and complex systems. As the expanding field of data science increasingly brings to light Bayesian approaches with faster computation and more efficient samplers, hopefully the closing of this gap will come to bear.

In the meantime, the analysis of such systems can be conducted with relative ease with typical samplers. Future work will focus on integrating this workflow with more advanced samplers such as the state-of-the-art No U-Turn Sampler (NUTS) and other Hamiltonian or adaptive algorithms.

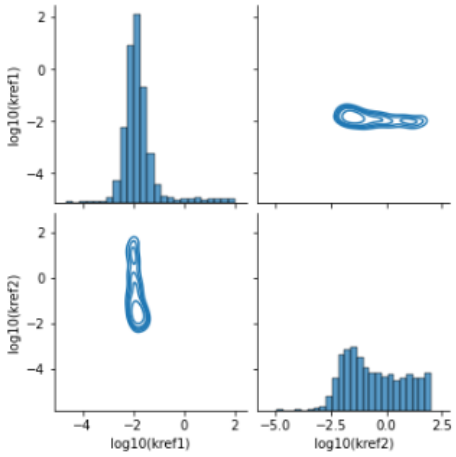
Appendix A: References

1. Gelman, A.; Roberts, G. O.; Gilks, W. R (1996). Efficient Metropolis Jumping Rules. Bayesian Statistics 5, Oxford University Press: pp. 599-607.
2. Spencer, S.E. (2021). Accelerating adaptation in the adaptive Metropolis–Hastings random walk algorithm. Aust. N. Z. J. Stat., 63: 468-484.
3. Sejdinovic, D.; Strathmann, H.; Garcia, M. L.; Andrieu, C.; Gretton, A (2014). Kernel Adaptive Metropolis Hastings. Proceedings of the 31st International Conference on Machin eLearning, Beijing, China. JMLR: W&CP Volume 32.

Appendix B: Supplementary Figures and Tables



(Fig. B1) Univariate histograms of the 4-D joint posterior overlaid with the initial uniform priors



(Fig. B2) Histograms and contours of k_{ref1} and k_{ref2} sampled together from uniform priors

(Table B1) Results of Bayesian inference, k_{ref1} and k_{ref} sampled together from uniform priors

Parameter	**log10($k_{ref,1}$)**	**log10($k_{ref,2}$)**
Posterior Mean	-1.775	-0.515
Prior Variance	0.502	1.756
Posterior Variance	4.083	4.083
95% Eqt. Credible Set	-2.643 to 0.666	-2.512 to 1.877
Posterior MAP Estimate	-2.045 to -1.906	-1.749 to -1.610
MLE Estimate	-1.813	-1.690