A Bayesian Analysis of Steady–State Enzyme Data leads to Estimates of Rate Constants and Uncertainties in a Multi-Step Reaction, along with Free Energy Profiles.†

Ian Barr*,‡,¶

‡*Dominican University of California, San Rafael, CA 94901.*

¶*Current affiliation: Metagenomi, Inc., Emeryville, CA 94608*

E-mail: ian.barr.lab@gmail.com

Phone: +1 (805)709 0549

Abstract

The microscopic rate constants that govern an enzymatic reaction are only directly measured under certain experimental set-ups, such as stopped flow, quenched flow, or temperature-jump assays; the majority of enzymology proceeds from steady state conditions which leads to a set of more easily–observable parameters such as k_{cat}, K_M, and observed Kinetic Isotope Effects $\left({}^{D}k_{cat}\right)$. This paper further develops a model from Toney (2013) to estimate microscopic rate constants from steady-state data for a set of reversible, four–step reactions. This paper uses the Bayesian modeling software `Stan`, and demonstrates the benefits of Bayesian data analysis in the estimation of these rate constants. In contrast to the optimization methods employed often in the estimation of kinetic constants, a Bayesian treatment is more equipped to estimate the uncertainties of each parameter; sampling from the posterior distribution using Hamiltonian Monte Carlo immediately gives parameter estimates as mean or median of the posterior, and also confidence intervals that express the uncertainty of each parameter.

†This article is a preprint, and has not yet undergone peer review.

1 Introduction

Estimation of the rate constants associated with each step of an enzymatic mechanisms is rarely straightforward, due to complexity of the reactions and lack of an ability to observe each intermediate species during the course of a reaction. The two enzymes under study here are alanine racemase (AR, EC 5.1.1.1), which catalyzes the reversible conversion of L-alanine to D-alanine, and triosephosphate isomerase (TIM, EC 5.3.1.1), which functions in glycolysis to convert dihydroxyacetone phosphate into D-glyceraldehyde 3-phosphate. Both are classified as isomerases, and take a single substate in both the forward and reverse directions. The general reaction scheme for AR and TIM is given in Scheme 1. In order to fully characterize these reactions, kinetically,

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{EZ} \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} \mathrm{EP} \underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} \mathrm{E} + \mathrm{P}$$

Scheme 1: The Reaction scheme for a reversible, 4-step reaction.[2]

we would like to estimate the rate constants for every step. In addition, if certain rate constants are isotopically sensitive, there will be additional values to estimate. For an enzymatic reaction scheme with four reversible steps, that leaves us with 8 microscopic rate constants to determine. In Scheme 1, k_1 and k_{-4} are second–order rate constants, and all others are first order. EZ is an intermediate that reacts rapidly in both directions. The substrates are taken to be L-alanine for AR and dihydroxyacetone phosphate for TIM, though the reactions are reversible. First order rate constants k_2 and k_{-3} are isotopically sensitive, with primary kinetic isotope effects ${}^{D}k_2$ and ${}^{D}k_{-3}$.

Since we cannot directly measure k_1, k_2, etc., we have to rely on indirect methods of determining those values. Ref. 1, which is the starting point for this work, uses a series of measurements done under steady state condition, each of which can be related to the microscopic rate constants mathematically (Eqs. (2)– (12)). Through incorporation of sufficient experimental data, it is possible in principle to determine each of the microscopic rate constants. In Ref. 1, global fitting is used to extract individual rate constants from steady-state reaction data. Global fitting in this case refers to the use of a target function containing contributions from all of the experimental data, from which are estimated a set of parameters consistent with the entire data set through non-linear regression. The earlier work used standard non-linear optimization algorithms to minimize the relative squared error of a set of data points. The target function used was

$$f(\boldsymbol{\theta}, \boldsymbol{D}) = \sum_i \left(\frac{\hat{D}(\boldsymbol{\theta}) - D_i}{D_i} \right)^2 \tag{1}$$

Where $\boldsymbol{\theta}$ is a vector of parameters to be estimated, $\boldsymbol{D}$ is a vector of experimental values, and $\hat{D}(\boldsymbol{\theta})$ is the function relating the parameters to the experimental value. This function leads to a minimization of the relative standard deviation (RSD), which is preferred because the experimental values are different orders of magnitude so they must be scaled to avoid bias. Ref. 1 showed that convergence was achievable using non-linear optimization, and that the method was reasonably robust. The fact that an optimization algorithm converges on a set of parameter values is not in itself useful, unless we have some confidence in those numbers. Ref. 1 wisely uses a method whereby a set of randomly generated values with the same mean and standard deviation as the experimental data are fed into the optimization algorithm, and parameters are re-calculated for each set, allowing an estimation of parameter uncertainty. Other non-linear methods would employ the Hessian matrix or bootstrapping to the same effect.[3,4] These methods fall under the rubric of frequentist analysis, which is often faster and equally as accurate as Bayesian methods are, given plentiful, high quality data. However, when the number of parameters to be estimated is nearly equal to the number of data points, as in the current case, Bayesian methods can provide invaluable information about the most likely parameter values, given all available data, and the uncertainty the estimates of each parameter.[5] Here I show that a Bayesian modeling of the same system gives robust and useful estimates of the rate constants and their associated uncertainties. In addition, a Bayesian treatment is able to handle cases of possible experimental error, at the cost of greater uncertainty in the parameter posterior distributions.[6]

1.1 Incorporation of the Equilibrium Constant

We have here introduced a new data value to improve the estimate, the equilibrium constant K_{eq} (Eq. 13). This is equal to the product of the forward rate constants divided by the reverse rate constants, and can be determined experimentally by measuring the concentrations of reactant and product at equilibrium, or indirectly from the forward and reverse k_{cat}/K_m values using the Haldane relationship:[11,12]

$$k_{cat,f} = \frac{k_2}{1 + k_2/k_4 + ((k_2 + k_{-2})/k_3)(1 + k_{-3}/k_4)} \tag{2}$$

$$k_{cat,r} = \frac{k_{-3}}{1 + k_{-3}/k_{-1} + ((k_3 + k_{-3})/k_{-2})(1 + k_2/k_{-1})} \tag{3}$$

$$K_{m,f} = \frac{k_{-1}}{k_1} \cdot \frac{1 + k_2/k_{-1} + (k_{-2}/k_3)(1 + k_{-3}/k_4)}{1 + k_2/k_4 + ((k_2 + k_{-2})/k_3)(1 + k_{-3}/k_4)} \tag{4}$$

$$K_{m,r} = \frac{k_4}{k_{-4}} \cdot \frac{1 + k_{-3}/k_4 + (k_3/k_{-2})(1 + k_2/k_{-1})}{1 + k_{-3}/k_{-1} + ((k_3 + k_{-3})/k_{-2})(1 + k_2/k_{-1})} \tag{5}$$

$$K_{EZ} = \left(\frac{k_{-2}}{k_2} + \frac{k_3}{k_{-3}} + 1\right)^{-1} \tag{6}$$

$$\theta = \frac{k_{-2}(1 + k_{-3}/k_4)}{k_3(1 + k_2/k_{-1})} \tag{7}$$

$$\text{slope} = \frac{k_{-1}k_{-2}k_{-3} + k_2k_3k_4}{k_{-1}k_{-2}k_{-3} + k_2k_3k_4 + k_{-1}k_3k_4 + k_{-1}k_{-2}k_4} \tag{8}$$

$${}^{D}k_{cat,f} = \frac{{}^{D}k_2 + (k_2/k_3)(1 + k_{-3}/k_4) + k_2/k_4 + {}^{D}k_2(k_{-2}/k_3)(1 + k_{-3}/k_4)}{1 + (k_2/k_3)(1 + k_{-3}/k_4) + k_2/k_4 + (k_{-2}/k_3)(1 + k_{-3}/k_4)} \tag{9}$$

$${}^{D}k_{cat,r} = \frac{{}^{D}k_{-3} + (k_{-3}/k_{-2})(1 + k_2/k_{-1}) + k_{-3}/k_{-1} + {}^{D}k_{-3}(k_3/k_{-2})(1 + k_2/k_{-1})}{1 + (k_{-3}/k_{-2})(1 + k_2/k_{-1}) + k_{-3}/k_{-1} + (k_3/k_{-2})(1 + k_2/k_{-1})} \tag{10}$$

$${}^{D}\left(\frac{k_{cat}}{K_m}\right)_f = \frac{{}^{D}k_2 + k_2/k_{-1} + {}^{D}k_2(k_{-2}/k_3)(1 + k_{-3}/k_4)}{1 + (k_2/k_{-1}) + (k_{-2}/k_3)(1 + k_{-3}/k_4)} \tag{11}$$

$${}^{D}\left(\frac{k_{cat}}{K_m}\right)_r = \frac{{}^{D}k_{-3} + (k_{-3}/k_4) + {}^{D}k_{-3}(k_3/k_{-2})(1 + k_2/k_{-1})}{1 + (k_{-3}/k_4) + (k_3/k_{-2})(1 + k_2/k_{-1})} \tag{12}$$

$$K_{eq} = \frac{k_1k_2k_3k_4}{k_{-1}k_{-2}k_{-3}k_{-4}} \tag{13}$$

Figure 1: Equations used in the analysis of data for TIM, AR and the simulated data. Eqs. (2)– (12) are used in Ref. 1, and Eq. (13) is included in this analysis.

Table 1: Experimental Values used to estimate Rate Constants

	Alanine Racemase		Triosephosphate Isomerase	
	Mean (SD)	Ref.	Mean (SD)	Ref.
$k_{cat,f}$	1740 (10)	1	750 (50)	1
$k_{cat,r}$	1280 (12)	1	8350 (350)	1
$K_{m,f}$	5.4 (0.1)	1	1.35 (0.15)	1
$K_{m,r}$	4.0 (0.1)	1	0.05 (0.01)	1
K_{EZ}	1.6×10^{-4} (0.4×10^{-4})	1	≤ 0.05	1
θ	0.5 (0.1)	1	3 (1)	1
slope	0.015 (0.015)	1	0.8 (0.1)	1
${}^{D}k_{cat,f}$	1.5 (0.1)	1	3.4 (0.1)	1
${}^{D}k_{cat,r}$	1.4 (0.03)	1	1.6 (0.1)	1
${}^{D}\left(\frac{k_{cat}}{K_m}\right)_f$	1.6 (0.1)	1	3.4 (0.1)	1
${}^{D}\left(\frac{k_{cat}}{K_m}\right)_r$	1.3 (0.1)	1	1.6 (0.1)	1
K_{eq}	1.0 (0.05)	7	0.0035 (0.0001)	8–10

$$K_{eq} = \frac{k_{cat,f} K_{m,r}}{k_{cat,r} K_{m,f}}$$

Direct measurement of K_{eq} is to be preferred, since use of the Haldane relationship utilizes the k_{cat} and K_m values that are already incorporated into the model, so using these values again tends to bias the estimates. For the same reason, the values of $\left({}^{D}(k_{cat}/K_m) - 1\right)/\left({}^{D}k_{cat} - 1\right)$ used in Ref. 1 are not used here, because they represented re-use of data that is already incorporated as ${}^{D}k_{cat}$ and ${}^{D}(k_{cat}/K_m)$. However, in some cases the K_{eq} might be hard to measure directly, and the Haldane relationship may be used (with caution). More reliable estimates for K_{eq} might be possible using the Haldane relationship if there exists high quality data for homologues of the enzyme, or point mutants, because the K_{eq} values calculated by the Haldane relationship should theoretically be the same for all active versions of an enzyme, as long as the temperature and buffer composition are similar. In this case we can average the values from several sources to obtain a more reliable estimate for K_{eq}.

An additional reason for using the value of K_{eq} is that the expression contains k_1 and k_{-4}, which each only appear in one other equation (for K_m, forward and reverse). This means that we are dependent on accurate measurement of K_m to get reliable values for k_1 and k_{-4}, in the absence of any further information. For enzymes such as AR, which converts L-alanine to D-alanine, the

K_{eq} is theoretically exactly 1, since there is no reason L-alanine would have a higher or lower free energy than D-alanine in a mostly achiral aqueous solution. For TIM, there is no direct measurement of the K_{eq} available in the literature, possibly because both dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate are themselves in equilibrium with their catalytically-inactive hydrated forms.[8] So in order to obtain the K_{eq} for the unhydrated forms, I averaged 4 literature values for K_{eq}, derived from the Haldane relationship.[8–10]

2 Considerations for Accurate Parameter estimation

2.1 General Limitation of the Bayesian Method

Every form of parameter estimation rests on a set of assumptions about the data and a model; this case is no different. Stan, as with other Bayesian modeling software, requires these assumptions be made explicit. Each parameter needs a prior distribution, which can affect the final result. The form of the model partly determines the results, and an incorrect model will lead to unhelpful results.

2.2 Choice of Priors for ks.

One aspect of a Bayesian analysis that differs from the function minimization procedures used in Ref. 1 is the requirement to specify a prior distribution for each of the parameters. This is information that is incorporated into the model according to the modified version of Bayes' Law:[13]

$$p(\Theta|D) \propto p(D|\Theta)p(\Theta)$$

Here, the posterior distribution of the parameters $p(\Theta|D)$, the output of our simulation, is the product of the likelihood function $p(D|\Theta)$ and the prior distribution for the parameters $p(\Theta)$. I have chosen a uninformative prior $k \sim \text{Exponential}\left(\beta\right)$ for each of the ks, based on the following assumptions:

1. The value of k is necessarily > 0, so an exponential distribution has the same domain.

2. The exponential distribution is often seen in physically relevant phenomena.[14,15]

3. Setting $k \sim \text{Exponential}(\beta)$ with $\beta << 1$ gives a broad distribution that covers the region from 1 to 1×10^9, typical values for microscopic rate constants.

4. Nonetheless, the prior is not too restrictive, because we have poor prior information about which values are typical for a rate constant.

This last point is especially important, as too restrictive a prior can end up determining the shape of the posterior distribution in the absence of sufficient experimental data.

The prior for k_2 (and the other ks) is implemented as follows in Stan:

```
 transformed data{
  real<lower=0> kfp;
  ...
  kfp = kf[1]; // the lower limit for all forward rate constants is kcat,f.
  ...
}
 parameters {
   ...
  real<lower=kfp, upper=1e12> k2; // rate of bond vibration. Fersht(1999)
   ...
  real<lower=0> bet; // Hyperprior parameter for all ks
}
 k2 ~ exponential(bet); // Prior ditribution for k2
 ...
 bet ~ gamma(1,1); // Hyperprior; Gamma is conjugate prior for Exp.
```

Here, we utilize a hyperprior β; the prior distribution for k_2 depends on the parameter β, which is also estimated over the course of the simulation. This allows a great deal of flexibility while keeping the mathematical form of the priors constant. The hyperprior for β is set as $\beta \sim \text{Gamma}(1, 1)$, a relatively uninformative prior with most of the mass below 1.

2.3 Choice of Priors for Intrinsic KIEs $\left({}^{D}\mathbf{k_i}\right)$.

Kinetic isotope effects are strictly positive quantities, and for the comparison between deuterium and protium the intrinsic KIE of step i is

$$ {}^{D}k_i = \frac{k_{i,P}}{k_{i,D}} $$

where $k_{i,P}$ and $k_{i,D}$ are the rate constants of the reaction with protonated and deuterated substrate. Common ranges for primary KIEs are 1.5 – 3, in the absence of quantum-mechanical tunneling.[16] Rarely, inverse KIEs are observed where ${}^{D}k_i < 1$. Given these constraints, I set the prior as

$$ {}^{D}k_i \sim \text{Lognormal}\,(1, 0.5) $$

Figure 10 graphs the prior used for KIEs. We see that most of the mass is between 1 and 4, but the density extends to infinity in the positive direction. I limit the value of KIEs to less than 500, based on the fact that the largest measured enzymatic ${}^{D}k_{cat}$ is around 500.[17] Any KIE greater than 6 is likely to be due to quantum mechanical effects, and in cases where this is suspected (e.g. hydride transfer) the prior could be adjusted to reflect the expected ranges of values.

2.4 The Problem constants – k_1 and k_{-4}.

In Ref. 1 and here, there are difficulties in accurately determining k_1 and k_{-4} for both TIM and AR. Significantly, in Ref. 1 k_1 and k_{-4} each only appear in one equation, the one for $K_{m,f}$ (Eq. 4) and $K_{m,f}$ (Eq. 5).

$$ \begin{aligned} K_{m,f} &= \frac{k_{-1}}{k_1} \cdot \frac{1 + k_2/k_{-1} + (k_{-2}/k_3)(1 + k_{-3}/k_4)}{1 + k_2/k_4 + ((k_2 + k_{-2})/k_3)(1 + k_{-3}/k_4)} \\ K_{m,r} &= \frac{k_4}{k_{-4}} \cdot \frac{1 + k_{-3}/k_4 + (k_3/k_{-2})(1 + k_2/k_{-1})}{1 + k_{-3}/k_{-1} + ((k_3 + k_{-3})/k_2)(1 + k_2/k_{-1})} \end{aligned} $$

The intuitive effect of this is that each of the experimental values *besides* $K_{m,f}$ and $K_{m,r}$ only indirectly provide information as to the true value of k_1 and k_{-4}, by helping to determine the values of the other parameters. But an interesting effect of this can be seen in Figure 7, which shows correlation between parameters during the course of the simulation as the posterior distribution is explored.

In row 5, column 1, we see that the values of k_1 and k_{-1} are linearly correlated, as are the values of k_4 and k_{-4} in row 8, column 4. Looking at the equations for K_m, we see that this is largely due to the fact that each contains the factor k_1/k_{-1} or k_4/k_{-4}, and since this is the sole place that k_1 and k_{-4} appear in this model, ambiguity in k_{-1} is passed along to k_1, etc. Adding the data for the K_{eq} doesn't alter this, as the expression for K_{eq} also contains k_1/k_{-1} and k_4/k_{-4}. This tells us that k_1 and k_{-4} can't be considered separately from k_{-1} and k_4; all this model can give us, in the absence of strong prior information about k_1 and k_{-4}, is the ratios k_1/k_{-1} and k_4/k_{-4}, i.e. the equilibrium constants for the first and fourth steps. Thus in my Stan code I have replaced k_1/k_{-1} and k_4/k_{-4}, where they appear, with K_1 and K_4. This slightly simplifies the calculations, and for a reversible reaction such as these it is reasonable to assume that the forward and reverse constants are within three orders of magnitude of each other, so we can limit the value of K_1 and K_4 during the simulation to between 0 and 10^3. Indeed, in both TIM and AR the values determined are approximately equal to $K_{m,f}$ and $1/K_{m,r}$, though this is not necessarily true in general as K_ms can be greater than, less than, or equal to the association equilibrium constant (e.g. K_1) in the case of a multi-step reaction.[18]

3 Results and discussion

3.1 Application to Simulated Data

We base our estimates on a set of 12 equations, and we estimate 11 parameters from these data points and their uncertainties, following the general rule of thumb that one can estimate at best $n - 1$ unknown parameters from n data points. However, this is only best–case; experimental error and the structure of the model can limit our ability to estimate parameters effectively. The primary difficulty here is one of structural identifiability;[19] can we, even with ideal data, estimate the parameters given the model we have?

To test the ability of our model to accurately determine rate constants, I simulated a data set with a fixed relative standard deviation (RSD) for all experimental values. I chose values for ks in the range of 10^3 to 10^8, and two isotope effect values in the classical range (1-6). With a RSD of 0.01, representing ideal experimental conditions, the modeled mean values are all within 10% of the true value, and the 90% confidence intervals contain the true value. Repeating this with other

simulated values gives equally accurate results. The $\hat{R}$ statistic[13,20] measures the average divergence between MCMC chains during a simulation; in ideal data the value is 1.0 exactly, indicating that all the chains in the simulation have converged on the same posterior distribution. I have used 4 independent chains in each analysis. The $\hat{R}$ for all of the parameters in this investigation is less than 1.1, as prescribed by Ref. 13.

Increasing the RSD to 0.1, a much more realistic value, shows the model beginning to drift away from the true values and an increase in uncertainty. Nonetheless only two of the parameters is off by more than 50% – k_4 and k_{-3}. These two parameters are highly correlated, and in the absence of stronger prior information are likely to deviate from their simulated values. Notably, the ratio of k_4 to k_{-3} is simulated as 0.5 and the fit shows 0.375, suggesting than an increase in prior information for either k_4 or k_{-3} would greatly improve the estimate of both. While Toney validated his model with ideal datasets, his test data didn't include experimental error and is analogous to my dataset with RSD = 0.01. The results of the modeling show that the current method is able to accurately determine rate constants under ideal conditions, and that experimental error begins to affect this at higher levels, as expected. I conclude from this that the model is structurally identifiable, with some parameters such as the intrinsic isotope effects determined with better precision than others.

Table 2: Statistical summary of the Stan output for the Simulated Data. $\hat{R}$ is the Gelman–Rubin statistic.[20] 90% CI is the 90% confidence interval for the posterior of each parameter. *RSD* is Relative Standard Deviation.

Parameter	Sim. Value	Mean (RSD = 0.01)	90% CI	$\hat{R}$	Mean (RSD = 0.1)	90% CI	$\hat{R}$
k_1	2×10^4	1.97×10^4	$(1.86 - 2.09) \times 10^4$	1.00	1.58×10^4	$(1.39 - 1.80) \times 10^4$	1.00
k_2	1×10^3	9.93×10^2	$(9.56 - 10.3) \times 10^2$	1.00	7.51×10^2	$(4.32 - 11.2) \times 10^2$	1.00
k_3	1×10^7	1.00×10^7	$(9.51 - 10.5) \times 10^6$	1.00	1.40×10^7	$(7.07 - 27.2) \times 10^6$	1.00
k_4	1×10^4	1.06×10^4	$(8.61 - 13.1) \times 10^3$	1.00	$\mathbf{3.45 \times 10^6}$	$\mathbf{(9.22 - 75.4) \times 10^5}$	1.00
k_{-1}	5×10^3	4.94×10^3	$(4.68 - 5.23) \times 10^3$	1.00	3.99×10^3	$(3.91 - 4.10) \times 10^3$	1.00
k_{-2}	1×10^8	9.81×10^7	$(9.43 - 10.2) \times 10^7$	1.00	6.69×10^7	$(3.81 - 10.1) \times 10^7$	1.00
k_{-3}	2×10^4	2.15×10^4	$(1.73 - 2.66) \times 10^4$	1.00	$\mathbf{9.19 \times 10^6}$	$\mathbf{(2.05 - 20.4) \times 10^6}$	1.00
k_{-4}	1×10^5	9.92×10^4	$(9.67 - 10.2) \times 10^4$	1.00	8.96×10^4	$(7.22 - 11.3) \times 10^4$	1.00
${}^D k_{-3}$	3.5	3.48	(3.44 – 3.52)	1.00	3.26	(2.86 – 3.66)	1.00
${}^D k_{-3}$	1.5	1.53	(1.48 – 1.58)	1.00	1.95	(1.35 – 2.81)	1.00

Table 3: Statistical summary of the Stan output for the AR Data. $\hat{R}$ is the Gelman–Rubin statistic.[20] 90% CI is the 90% confidence interval for the posterior of each parameter.

Parameter	Mean	Median	90% CI	$\hat{R}$	Num. Opt. (Toney, 2013)
k_1 ($mM^{-1}s^{-1}$)	1.4×10^4	1.3×10^4	$(7.2 \times 10^3, 2.6 \times 10^4)$	1.00	$> 10^5$
k_2 (s^{-1})	2.8×10^3	2.8×10^3	$(2.5 \times 10^3, 3.0 \times 10^3)$	1.00	2600(200)
k_3 (s^{-1})	8.3×10^6	8.2×10^6	$(6.3 \times 10^6, 1.1 \times 10^7)$	1.00	$1.4 \times 10^7 (0.3 \times 10^7)$
k_4 (s^{-1})	9.0×10^5	7.2×10^5	$(2.5 \times 10^5, 2.1 \times 10^6)$	1.00	$9 \times 10^6 (8 \times 10^7)$
k_{-1} (s^{-1})	7.5×10^4	6.7×10^4	$(3.7 \times 10^4, 1.4 \times 10^5)$	1.00	$> 10^6$
k_{-2} (s^{-1})	4.7×10^6	4.6×10^6	$(6.7 \times 10^5, 5.9 \times 10^6)$	1.00	$6.8 \times 10^6 (0.7 \times 10^6)$
k_{-3} (s^{-1})	3.8×10^3	3.7×10^3	$(3.2 \times 10^3, 4.5 \times 10^3)$	1.00	4000(700)
k_{-4} ($mM^{-1}s^{-1}$)	2.1×10^2	1.7×10^2	$(5.6 \times 10^1, 4.9 \times 10^2)$	1.00	$2 \times 10^4 (2 \times 10^5)$
${}^D k_2$	1.56	1.56	(1.44, 1.67)	1.00	1.55(0.11)
${}^D k_{-3}$	1.41	1.41	(1.36, 1.46)	1.00	1.35(0.09)

3.2 Alanine Racemase

Table 3 shows the output of the Stan modeling for Alanine Racemase. We see that all of the parameters have converged well, as shown by the $\hat{R}$ values lying close to unity. From the table, and from the graphs in Figure 11, the intrinsic KIEs (${}^D k_2$ and ${}^D k_{-3}$) are in good agreement with the analysis from Ref. 1. The intrinsic KIEs for AR are especially well-defined, as shown by Figure 11. The prior and posterior distributions of both ${}^D k_2$ and ${}^D k_{-3}$ are shown, and the prior is much broader than the posterior, showing that the experimental data have been instumental in determining the mean and confidence intervals for the KIEs through the likelihood function. Others show some disagreement, especially the values of k_3 and k_{-2} which differ by > 10-fold. The reason for this is not entirely clear, but is likely due to the effects of experimental uncertainty, as both in Ref. 1 and the present work the models are shown to give accurate results with ideal data. In the case of real world data, the difference between the models and algorithms becomes more important as error increases, as do the assumptions behind each. In every case where the present results and those of Ref. 1 disagree significantly (> 10-fold), the latter parameters show a great deal of uncertainty. In the case of AR, these are k_3, k_{-2} and k_{-3}. For k_3 and k_{-2} we only have lower bounds in Ref. 1, and the SD of k_{-3} is 4 times the mean. In the present work, Stan is using the log of the joint likelihood function to estimate the shape and position of the posterior distribution, under the influence of a prior distribution; in Ref. 1 the program is trying to minimize a cost function (Eq. 1). Function minimization in the absence of a prior distribution can behave

similarly to doing a Bayesian analysis with a Uniform prior on all parameters. In a case where the domain of each parameter is on the order of 10^{12}, for a Bayesian analysis this gives us a prior where the parameter is nine times as likely to be in the range 10^{11} to 10^{12} as between 0 and 10^{11}. This is part of the motivation for the use of the exponential distribution as a prior for all ks, to correct this bias towards larger numbers. In minimizing the function, if the experimental data are subject to error the parameters can vary freely over large ranges during the search, and converge to a wide range of values. This search is unbiased by a prior distribution, so might be preferred as long as the uncertainty in parameter estimates can be contained and there is little prior information. However in cases where parameters cannot be defined to within even an order of magnitude by function minimization, a Bayesian analysis such as the one shown here should be considered. I also note that the values that I estimate for all the parameters are consistent with the experimental data; the experimental means and theoretical mean values agree to within 1% in all cases. It may be that due to experimental uncertainty more than one set of parameters is consistent with the data (multimodality). In this case, an improvement of the experimental data or a more informative prior distribution might be necessary to resolve the problem.

3.3 Triosephosphate Isomerase

Table 4: Statistical summary of the Stan output for the TIM Data. $\hat{R}$ is the Gelman–Rubin statistic.[20] 90% CI is the 90% confidence interval for the posterior of each parameter.

Parameter	Mean	Median	90% CI	$\hat{R}$	Num. Opt. (Ref. 1)
k_1 ($\mathrm{mM^{-1}s^{-1}}$)	1.4×10^4	1.3×10^4	$(9.7 \times 10^3, 2.2 \times 10^4)$	1.01	8400(1500)
k_2 ($\mathrm{s^{-1}}$)	4.4×10^3	4.2×10^3	$(2.9 \times 10^3, 6.3 \times 10^3)$	1.00	4000(1600)
k_3 ($\mathrm{s^{-1}}$)	9.7×10^4	8.3×10^4	$(4.1 \times 10^4, 1.9 \times 10^5)$	1.00	$> 10^9$
k_4 ($\mathrm{s^{-1}}$)	1.3×10^4	1.3×10^4	$(9.5 \times 10^3, 1.8 \times 10^4)$	1.00	$1.2 \times 10^4 (0.3 \times 10^4)$
k_{-1} ($\mathrm{s^{-1}}$)	2.0×10^4	1.9×10^4	$(1.4 \times 10^4, 3.1 \times 10^5)$	1.00	$1.1 \times 10^4 (0.2 \times 10^4)$
k_{-2} ($\mathrm{s^{-1}}$)	7.9×10^4	7.2×10^6	$(4.0 \times 10^4, 1.4 \times 10^5)$	1.00	$> 10^8$
k_{-3} ($\mathrm{s^{-1}}$)	5.3×10^4	4.8×10^4	$(2.8 \times 10^4, 9.8 \times 10^4)$	1.00	$5 \times 10^5 (20 \times 10^5)$
k_{-4} ($\mathrm{mM^{-1}s^{-1}}$)	4.0×10^4	3.8×10^4	$(2.2 \times 10^4, 6.3 \times 10^4)$	1.00	$2.3 \times 10^5 (0.3 \times 10^5)$
${}^D k_2$	3.57	3.57	$(3.45, 3.71)$	1.00	3.56(0.08)
${}^D k_{-3}$	2.64	2.61	$(2.10, 3.31)$	1.00	3.3(0.5)

For TIM, similar issues arise as for AR. There is an assumption that the uncertainties in each parameter are distributed normally. Looking at the estimates for the TIM dataset, Ref. 1 estimated

k_{-3} as 5×10^5 with a SD (20×10^5) ; if you assume normalcy that would mean a full 40% of the confidence interval lies below zero, where it is impossible for the value to be. Likewise for k_{-2}, the optimization results give wide ranges for the parameter confidence interval.

Nickbarg and Knowles[9] also calculated the ratios of the forward and reverse rate constants for yeast TIM. Table 5 shows a comparison of the results from the current study compared with Ref. 1 and Nickbarg and Knowles (1988). There is agreement between the ratios of the forward and reverse constants for the first and fourth steps (k_1/k_{-1} and k_4/k_{-4}). The current work and Nickbarg and Knowles (1988) give essentially identical estimates for k_2/k_{-2} and k_3/k_{-3}, with Ref. 1 differing by $\approx 10^3$ in both cases. At stake is the question of whether the complex of TIM with the enediol intermediate (EZ, in Scheme 1) is significantly higher in energy than the other enzyme forms. Higher energy would destabilize EZ, leading to higher k_{-2} and k_3 values and would therefore lead to k_2/k_{-2} approaching zero and k_3/k_{-3} much greater than one. While it is not my intention to wade into this debate, the results presented here are not consistent with a high-energy intermediate.

Table 5: Comparison of forward and reverse rate-constant ratios from the present work, Nickbarg and Knowles (1988), and Toney (2013). Results in the two rightmost columns are given as Mean (SD).

	Mean	90% CI	Nickbarg and Knowles (1988)	Toney (2013)
k_1/k_{-1}	0.7	(0.6, 0.9)	0.6 (0.3)	0.8 (0.2)
k_2/k_{-2}	0.05	(0.02, 0.09)	0.1 (0.5)	$< 4 \times 10^{-5}$
k_3/k_{-3}	1.8	(0.8, 3.6)	2 (10)	> 2000
k_4/k_{-4}	0.06	(0.04, 0.09)	0.04 (0.02)	0.05 (0.01)

3.4 Free Energy Profiles for TIM and AR

Starting with the MCMC draws for the eight rate constants, we can construct a series of FEPs and overlay them in an attempt to map the most likely profile for each of the enzyme. In Figures 8 and 9, I have randomly chosen 500 of the 4000 draws for each MCMC run, and calculated the FEP for that particular draw using the Eyring equation with $\kappa = 1$ and $T = 298K$. The FEPs show a great deal of overlap when plotted on top of each other, and species with uncertain energies show up as a series of lines, as in the EZ intermediate of AR (Figure 8). In Figures 8 and 9, I have also calculated the averrage energy of each species, shown as a red horizontal line.

4 Methods

4.1 Incorporation of Experimental Error.

The data[1] is expressed as mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$), as is commonly done in biochemical studies. The data are incorporated into the model as follows:

$$\mu = f\left(k_1, k_2, \ldots, {}^{D}k_{-3}\right)$$
$$\hat{\mu} \sim \text{Normal}(\mu, \hat{\sigma})$$

This is represented in `Stan` as follows, for $k_{cat,f}$ (ignoring all other data values):

```
data { // data are entered as vectors of length 2.
  vector[2] kf; // kf[1] is mean, kf[2] is sd.
  ... }
transformed parameters{
  real<lower=0> mukf; // "true" value of kf, function of k1, k2,... etc.
  mukf = k2/(1 + k2/k4 + ((k2+k_2)/k3)*(1 + k_3/k4)) ;
  ... }
model {
  kf[1] ~ normal(mukf, kf[2]); // Likelihood function.
  ... }
```

Where the true value of the experimentally-determined parameter (e.g. k_{cat}) is assumed to be drawn from a distribution with mean μ, and the standard deviation is set equal to the experimentally-determined uncertainty in the value. The model will then incorporate the mean value ($\hat{\mu}$) as data, and estimate a true value for it as well (μ), based on the global fitting. While one could use a distribution other than Normal to model the error, most published results use models that assume Normally-distributed error so when incorporating results from others it is important to follow this assumption. Ideally, the model would incorporate the raw data instead and proceed from there to estimates of the rate constants; however, the raw data is not available, as is often the

case with biochemical data. Nonetheless it is still possible to get estimates of the rate constants from published results.

4.2 MCMC Analysis

To estimate the posterior distribution of each of the eight parameters, I used `cmdstanr` 0.3.0 running under R 3.5, which is built on `cmdstan` 2.26.1.[21] For each analysis, 5000 iterations of the sampler were run on 4 parallel chains. The first 4000 of each were 'warm-up' samples, in which the parameters and step sizes are tuned by `Stan's` NUTS algorithm. This value, higher than the default value of 1000 warm-up samples, was necessary to ensure that the sampling distribution was stable, but had minimal effect on the runtime of the program. Runtimes on MacOS using a 3 GHz quad-core processor and 4 parallel chains ranged from 2 - 60s, without diagnostic errors after sampling. Figures were generated using `ggplot2` and the `bayesplot` package, except Figure 10 which was plotted with `gnuplot`.

Acknowledgement

This work was partially supported by a Summer Research Grant from Dominican University of California.

Supporting Information Available

The following files are available free of charge.

- simulate.stan: Stan model file
- enrg.R: R script file to process Toney (2013) data.
- simulate.R: R script file to simulate and process data.

References

(1) Toney, M. D. Common enzymological experiments allow free energy profile determination. *Biochemistry* **2013**, *52*, 5952–65.

(2) Cornish-Bowden, A. *Fundamentals of Enzyme Kinetics*, 3rd ed.; Portland Press Limited: London, 2004.

(3) James, F.; Roos, M. Minuit - a system for function minimization and analysis of the parameter errors and correlations. *Computer Physics Communications* **1975**, *10*, 343–367.

(4) Gallant, A. R. Nonlinear Regression. *The American Statistician* **1975**, *29*, 73–81.

(5) McNeish, D. On Using Bayesian Methods to Address Small Sample Problems. *Structural Equation Modeling: A Multidisciplinary Journal* **2016**, *23*, 750–773.

(6) Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo. arXiv, 1701.02434. 2018.

(7) Goldberg, R. N.; Tewari, Y. B. Thermodynamics of Enzyme-Catalyzed Reactions: Part 5. Isomerases and Ligases. *Journal of Physical and Chemical Reference Data* **1995**, *24*, 1765–1801.

(8) Albery, W. J.; Knowles, J. R. Free-energy profile of the reaction catalyzed by triosephosphate isomerase. *Biochemistry* **1976**, *15*, 5627–31.

(9) Nickbarg, E. B.; Knowles, J. R. Triosephosphate isomerase: energetics of the reaction catalyzed by the yeast enzyme expressed in Escherichia coli. *Biochemistry* **1988**, *27*, 5939–47.

(10) Sampson, N. S.; Knowles, J. R. Segmental motion in catalysis: investigation of a hydrogen bond critical for loop closure in the reaction of triosephosphate isomerase. *Biochemistry* **1992**, *31*, 8488–94.

(11) Cleland, W. W. An analysis of Haldane Relationships. *Methods Enzymol* **1982**, *87*, 366–369.

(12) Fersht, A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*; W.H. Freeman: New York, 1999.

(13) Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; Rubin, D. B. *Bayesian Data Analysis*, third edition ed.; Chapman & Hall/CRC texts in statistical science; 2013.

(14) Clauset, A.; Shalizi, C. R.; Newman, M. E. J. Power-Law Distributions in Empirical Data. *SIAM Review* **2009**, *51*, 661–703.

(15) Mitzenmacher, M. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics* **2004**, *1*, 226–251.

(16) Klinman, J. P. The role of tunneling in enzyme catalysis of C-H activation. *Biochim Biophys Acta* **2006**, *1757*, 981–7.

(17) Hu, S.; Sharma, S. C.; Scouras, A. D.; Soudackov, A. V.; Carr, C. A. M.; Hammes-Schiffer, S.; Alber, T.; Klinman, J. P. Extremely elevated room-temperature kinetic isotope effects quantify the critical role of barrier width in enzymatic C-H activation. *J Am Chem Soc* **2014**, *136*, 8157–8160.

(18) Dalziel, K. Physical significance of Michaelis constants. *Nature* **1962**, *196*, 1203–5.

(19) Bellman, R.; Åström, K. On structural identifiability. *Mathematical Biosciences* **1970**, *7*, 329–339.

(20) Gelman, A.; Rubin, D. B. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **1992**, *7*, 457–472.

(21) Stan Development Team, Stan Modeling Language Users Guide and Reference Manual, version 2.27. *mc-stan.org* **2019**, .

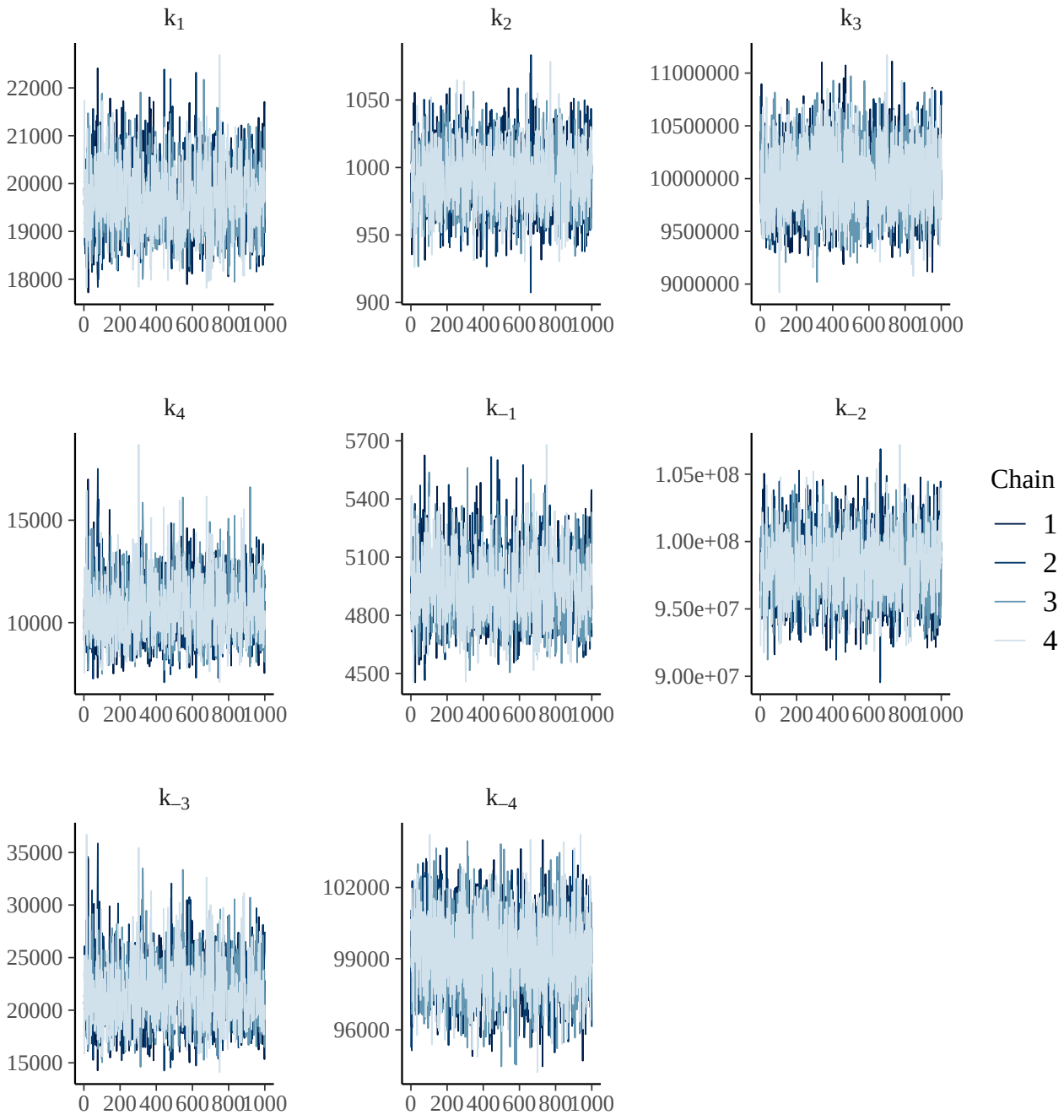

Figure 2: Traceplots of the modeled rate constants for the simulated data set, with experimental SD set to 0.01 for all quantities. The y-axis is the parameter value at each draw, and the x-axis are the sample numbers, post warm-up.

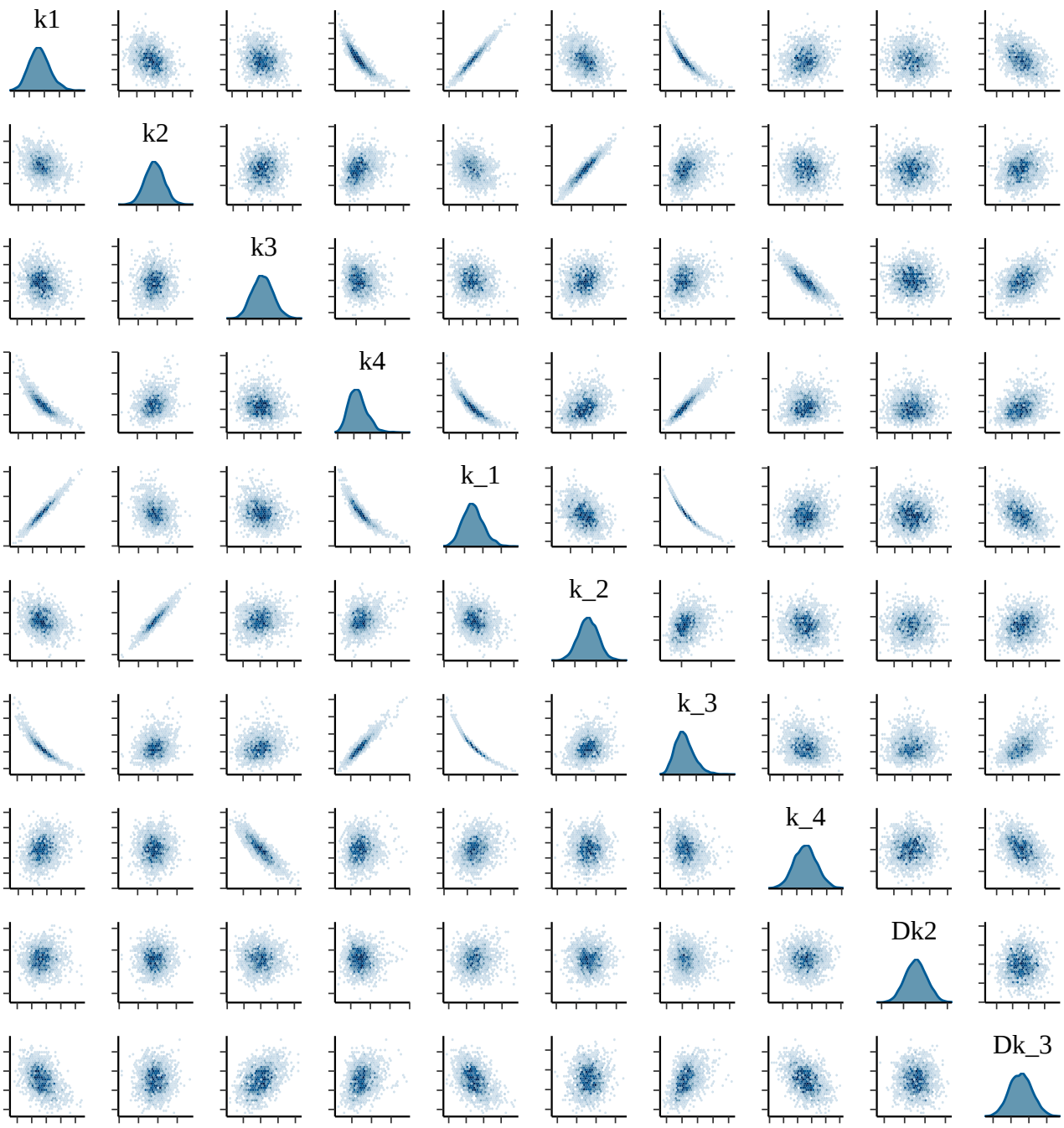

Figure 3: Pairwise comparison of the MCMC Draws for simulated rate constants, showing correlations between parameters.

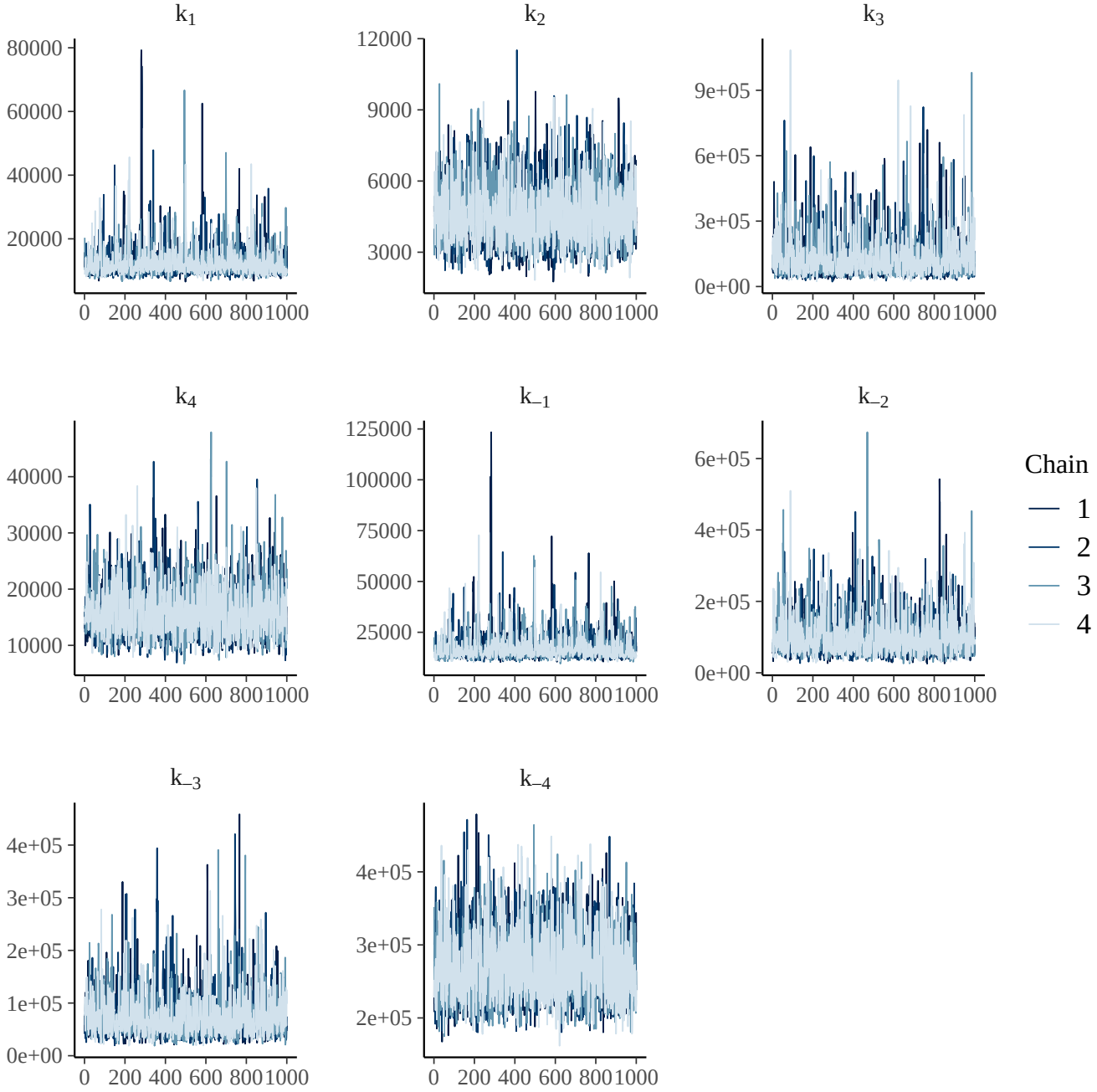

Figure 4: Traceplots of the rate constants for TIM.

Figure 5: Pairwise comparison of the MCMC Draws for TIM rate constants, showing correlation between k_1 and k_{-1}

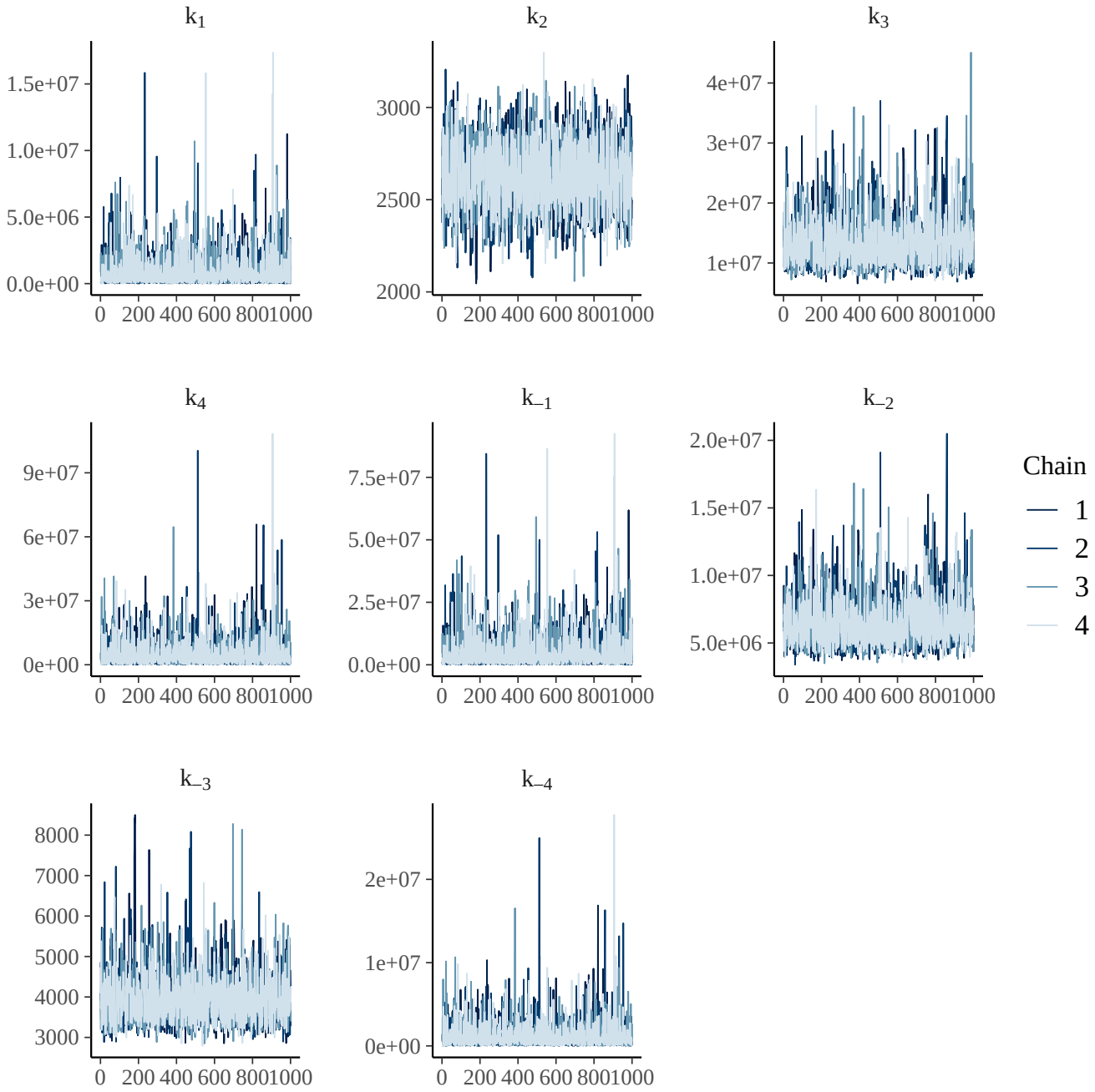

Figure 6: Traceplots of the rate constants for AR.

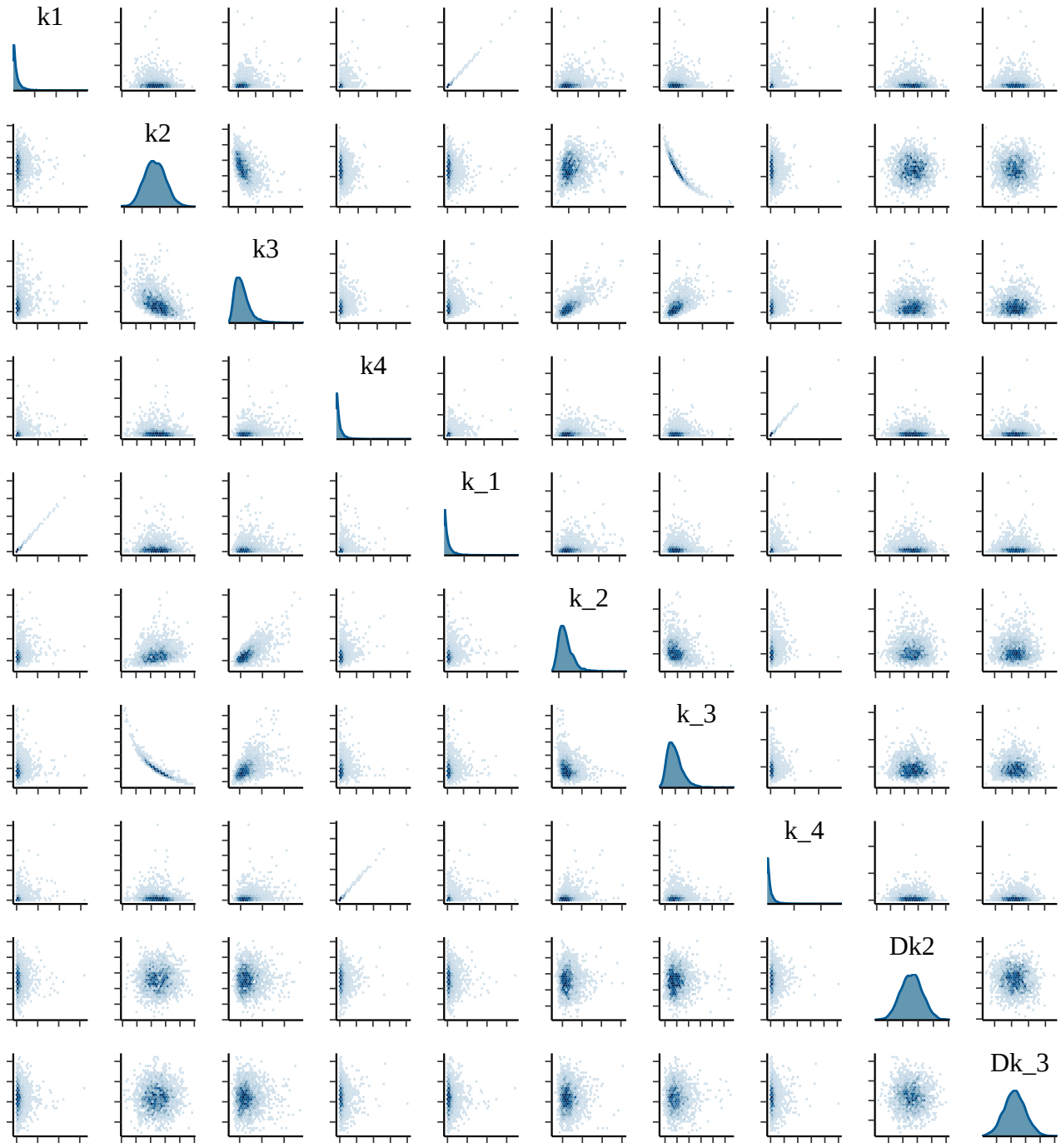

Figure 7: Pairwise comparison of the MCMC Draws for AR rate constants, showing correlation between k_1 and k_{-1}

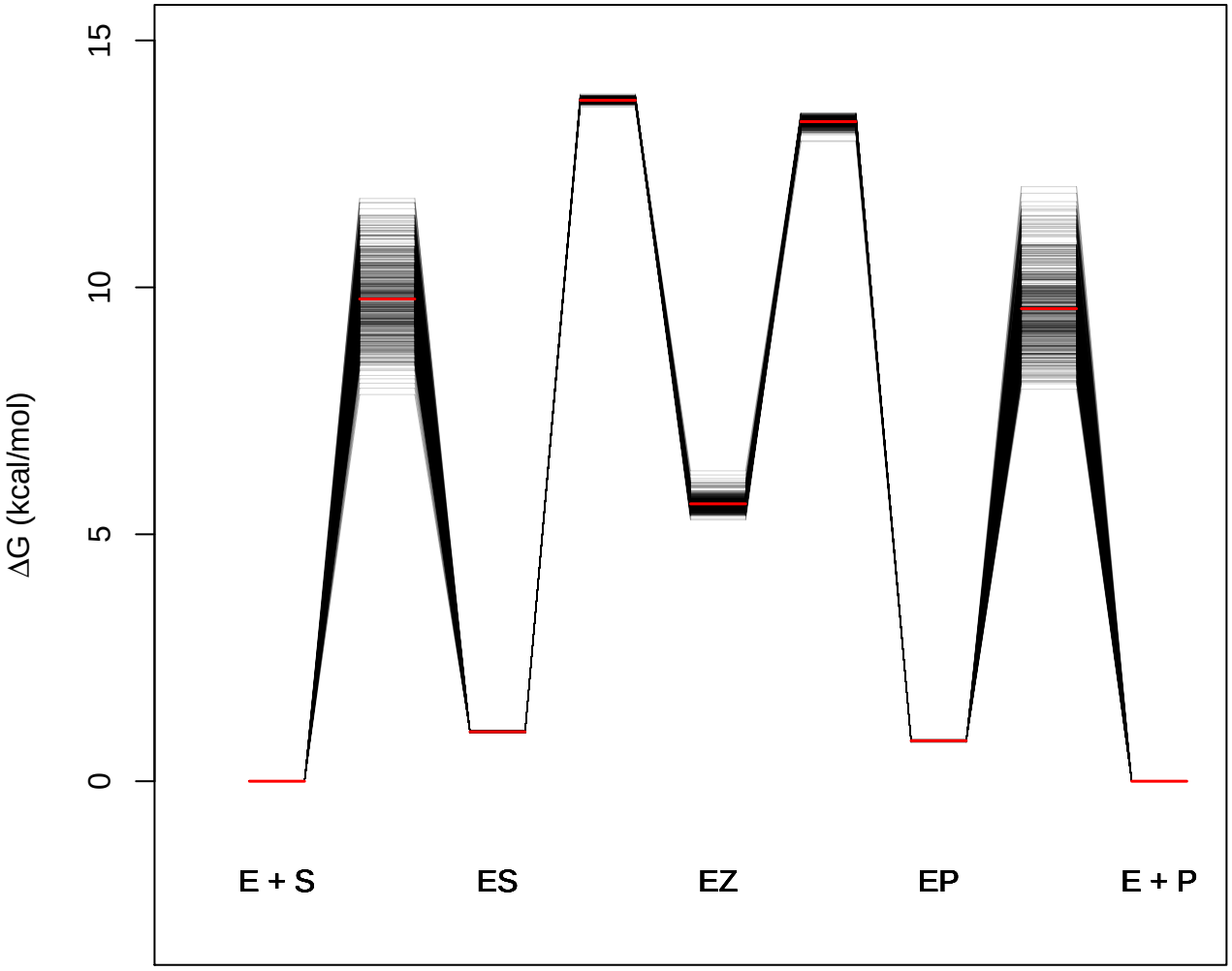

Figure 8: The Free Energy Profile for the AR reaction, relative to the starting substrate. The red lines are the means, for a given species.

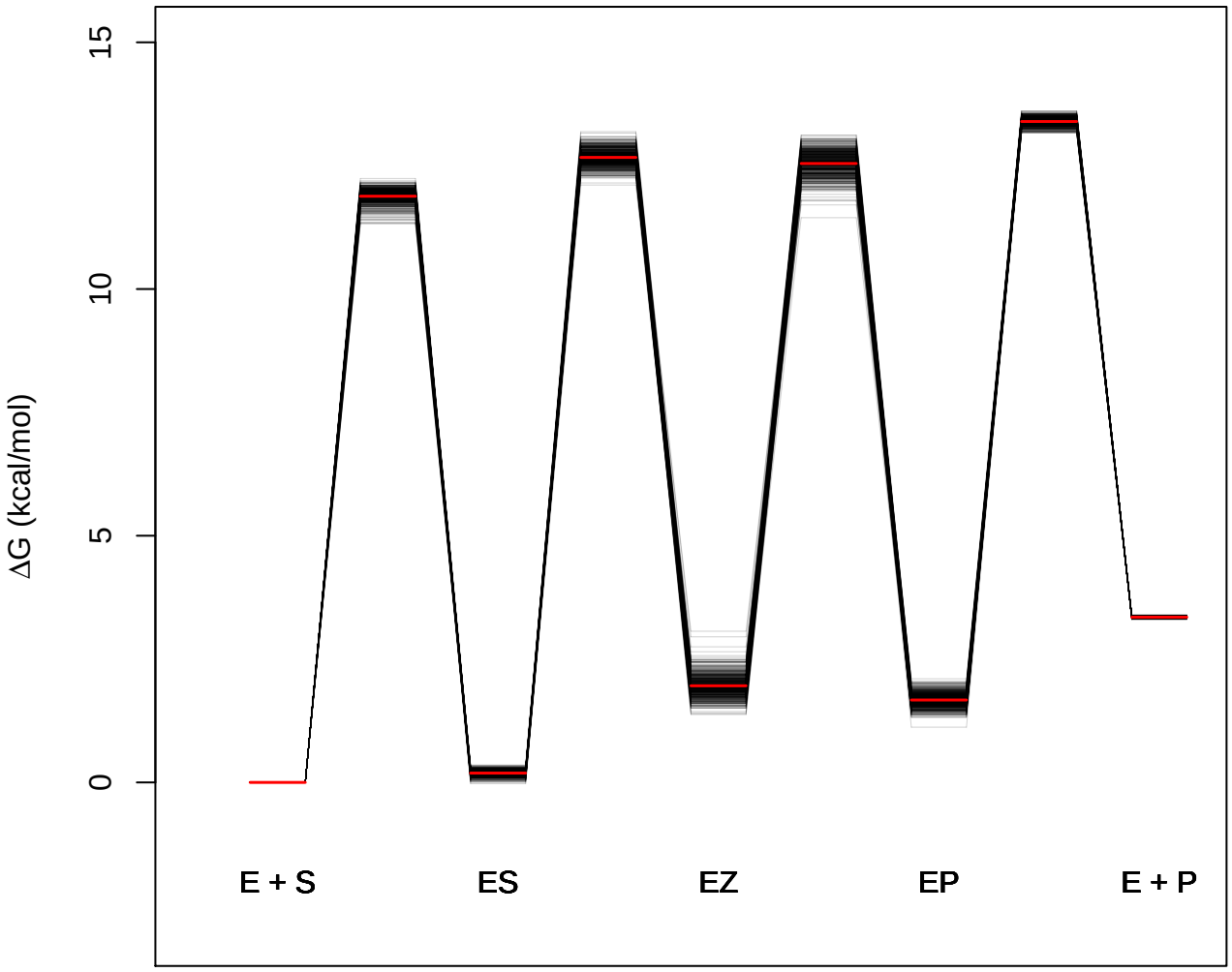

Figure 9: The Free Energy Profile for the TIM reaction, relative to the starting substrate. The red lines are the means, for a given species.

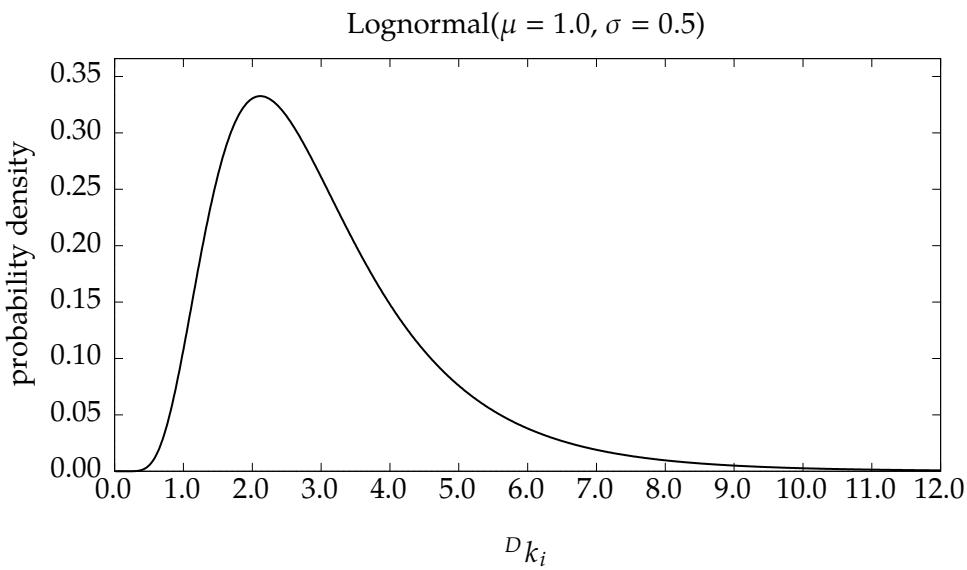

Figure 10: The prior distribution used for intrinsic KIEs $\left({}^{D}k_i\right)$.

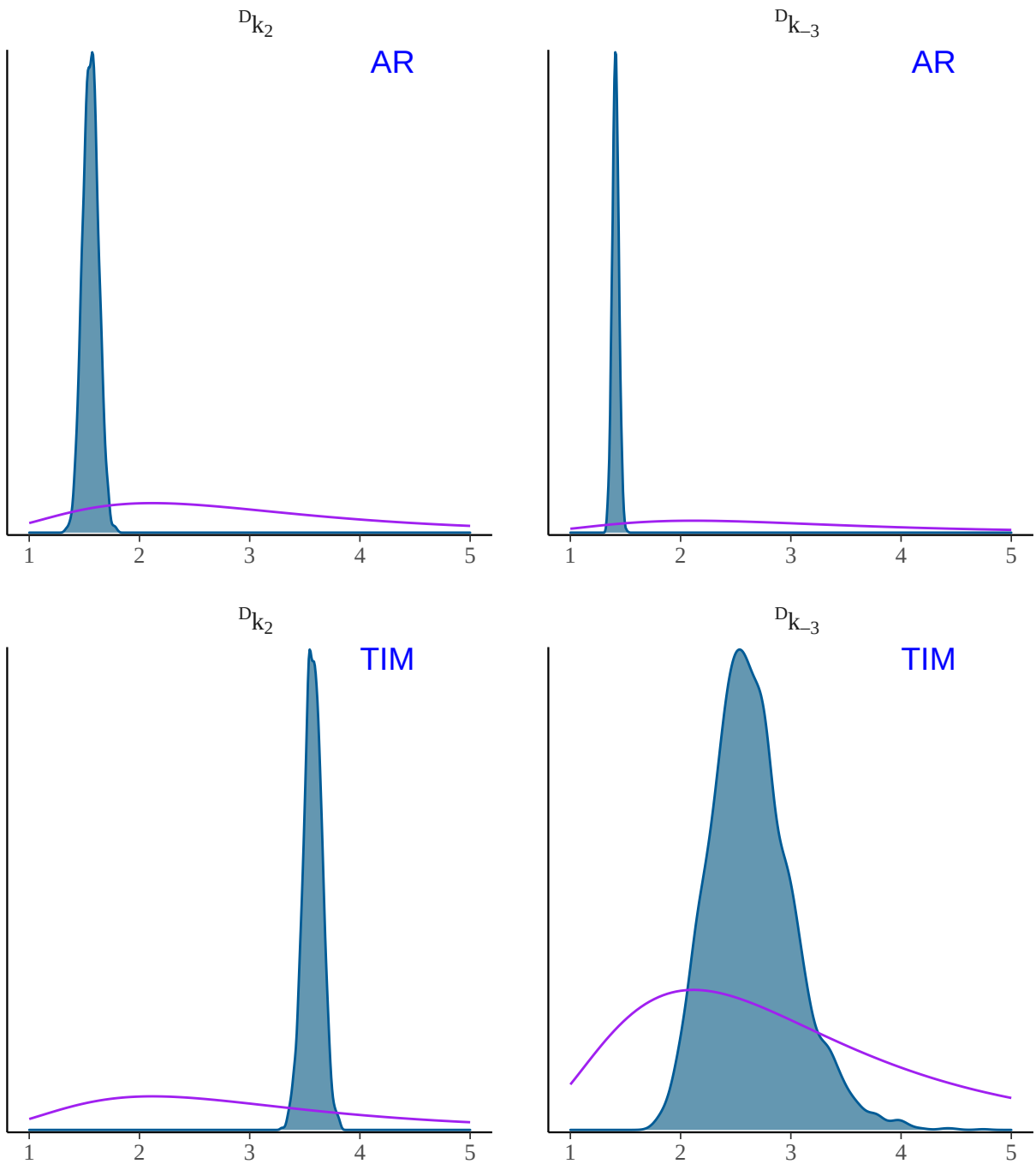

Figure 11: Posterior and prior distributions for the intrinsic KIEs $({}^{D}k_i)$ of TIM and AR. The prior distribution is shown in purple, and the posterior is shown in blue, filled.