Published in final edited form as:
J Phys Chem B. 2007 December 6; 111(48): 13600–13610. doi:10.1021/jp073708+.

Kinetic Analysis of Sequential Multi-step Reactions

Yajun Zhou[†] and **Xiaowei Zhuang**[*,†,‡,#]
Department of Chemistry and Chemical Biology, Department of Physics, Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts 02138

Abstract

Many processes in biology and chemistry involve multi-step reactions or transitions. The kinetic data associated with these reactions are manifested by superpositions of exponential decays that are often difficult to dissect. Two major challenges have hampered the kinetic analysis of multi-step chemical reactions: (1) Reliable and unbiased determination of the number of reaction steps, (2) Stable reconstruction of the distribution of kinetic rate constants. Here, we introduce two numerically stable integral transformations to solve these two challenges. The first transformation enables us to deduce the number of rate-limiting steps from kinetic measurements, even when each step has arbitrarily distributed rate constants. The second transformation allows us to reconstruct the distribution of rate constants in the multi-step reaction using the phase function approach, without fitting the data. We demonstrate the stability of the two integral transformations by both analytic proofs and numerical tests. These new methods will help providing robust and unbiased kinetic analysis for many complex chemical and biochemical reactions.

Keywords

Reaction kinetics; exponential decay; rate constant; single molecule study

1. Introduction

The multi-step reaction models are central to the mechanistic understandings of various fields of chemistry: organic,[1] atmospheric,[2] biophysical,[3–5] etc. For instance, an enzymatic reaction may be described by several sequential steps, including substrate binding, catalytic reaction, and product release;[4] an ion channel may change its conformation through multi-step allosteric transtions;[5] the movement by molecular motors requires ATP binding, ATP hydrolysis, and ADP release, with one of these steps linked to the powerstroke of a motor;[3] proteins[6,7] and RNA molecules[8–10] may fold/unfold themselves via complicated multi-stage pathways on a rugged energy landscape. A broad spectrum of experimental methods has been devised for probing into these complex chemical reactions, and these methods can be loosely classified as either ensemble measurements or single-molecule measurements. In the former case, one tracks the time course of the concentration of a certain chemical in the bulk; in the latter case, one tracks the reaction course of individual molecules one by one. The experimental data arising from these measurements often encode the kinetic information in the form of a superposition of many exponential decay modes, and the decay rate constants and the coefficients of these decay modes contain information about the underlying chemical reaction scheme.

*Author to whom correspondence should be addressed. Phone: (617) 496-9558. Fax: (617) 496-9559. Email: zhuang@chemistry.harvard.edu.
[†]Department of Chemistry and Chemical Biology, Harvard University
[‡]Department of Physics, Harvard University
[#]Howard Hughes Medical Institute, Harvard University

It is often critical to dissect the multi-step reaction kinetics quantitatively from experimental data, to determine the rate constant distribution of each step and to investigate the dependence of these rate constants on external conditions. For example, the rate constants can be affected by the concentration of substrates targeting a specific enzyme; the trans-membrane potential acting on an ion-channel; the force load on a molecular motor etc. Such dependence may elucidate the mechanism of the biological systems of interest. As individual biological molecules were brought to close scrutiny in recent single-molecule experiments, their kinetic behaviors were often found to invoke non-trivial distributions of reaction (or transition) rate constants. (See for example refs. [7–9,11–15]) These experiments suggest that biological processes may not only consist of multiple steps but a seemingly two-step reaction may also involve heterogeneous kinetics that are better described by a distribution of, instead of a single, kinetic rate constants, further complicating the analysis of biological processes that exhibit complex chemical kinetics.

The analysis of multi-step reaction kinetics is challenging in two aspects. The first aspect is to reliably determine the number of reaction steps from kinetic data. The number of reaction steps have been previously estimated by computing the "randomness parameter" [16–19] from the kinetic data, or fitting the data to a Γ-distribution function.[18,20] The "randomness parameter" approach, however, generally estimates a lower bound of the total number of rate-limiting steps, and the estimates become exact only when all reaction steps involve nearly identical rate constants.[16–18] Similarly, the validity of a Γ-distribution fit also draws on identical rate constants or a Hill function response,[21] which may not hold for general cases. The second challenge involves stable determination of the distribution of kinetic rate constants. Technically, this means to solve an inverse Laplace problem, which involves a well-known numerical instability (i.e. hypersensitivity to input noise).[22] As a general practice, a discrete distribution of rate constants has typically been deduced from Levenberg-Marquardt fitting or Hidden Markov modeling,[5,23,24] although these methods may suffer from a relatively arbitrary choice of the number of fitting parameters. Continuous distributions of rate constants have often been dealt with by stabilizing the numerical inversion of Laplace transform. The Tikhonov regularization[25,26] and maximum entropy method[27,28] are two popular stabilization methods, which stabilize the inverse Laplace transform by penalizing the curvature[26] or information entropy[27,28] of the continuous distribution profile and may distort the rate constant distribution especially when the distribution function does not exhibit uniform ruggedness. We have recently developed a phase function method to overcome the instability of the inverse Laplace problem, handling both discrete and continuous distributions of rate constants without prior knowledge input or artificial penalization.[29] Nonetheless, the phase function method derives its viability and numerical stability from a strong constraint: the combination coefficients of multiple decay modes must be all non-negative real numbers, which implies *single*-step reaction kinetics.

To address the aforementioned challenges in the kinetic analysis of multi-step reactions, we introduce two numerically stable integral transformations in this paper. The first integral transformation enables us to reliably recover the number of reaction steps from noisy kinetic data. This method applies to general multi-step reactions, in which the rate constants of individual reaction steps can take arbitrary, non-identical distributions. The second integral transformation allows us to convert a multi-step kinetic data set to a corresponding single-step kinetic data set with a distribution of rate constants. We can thus utilize the "phase function method"[29] to handle multi-step reaction kinetics as well.

The paper is organized as follows. In section 2, the mathematical formulation of data analysis is presented, with detailed descriptions of the two aforementioned integral transformations. In section 3, the numerical implementations of the theory and related discussion are presented. The paper is concluded by section 4. Although examples presented in this paper are motivated

by biophysical/biochemical processes, with a particular emphasis on interpreting single-molecule experiments which have been shown recently to reveal rich kinetics and dynamics of biological systems, we expect the numerical methods present here to be generally applicable to the analysis of multi-step reaction kinetics arising from many other chemical processes.

2. Theoretical Methods

2.1 Kinetic Data in Sequential Processes with *m* Reaction Steps

A typical *m*-step chemical reaction takes the following scheme:

$$A_1 \xrightarrow{k_1} A_2 \xrightarrow{k_2} \cdots A_m \xrightarrow{k_m} B. \tag{1}$$

According to eq 1, any molecule that starts from the "A_1 state" has to experience $(m-1)$ intermediate states A_2, …, A_m before it is turned to a "B state". The m-step reaction involves m sequential tasks in a sequel, fulfilled with a set of rate constants k_1, k_2, …, k_m. Often, the observable properties of the A states are experimentally indistinguishable, i.e. the states A_1,…, A_m may give similar fluorescence/mechanical signals in a single-molecule experiment. This degeneracy makes it difficult to fully dissect the reaction kinetics by experimental means.

An accessible quantity for a molecule that starts from the A_1 state is how long it takes to finish all the m tasks and reach the B state. Defining this as "reaction time" t (referring to the dwell time in all the A-like states), t will be the sum of m random durations $t_{dt} = t_1 + t_2 + \cdots + t_m$, where t_i is the dwell time that the molecule spends in the A_i state. Here, the random variable t_i is governed by the reaction rate constant k_i and thus follows the exponential decay law $\text{Prob}\{t_i > t\} = e^{-k_i t}$, $t \geq 0$. The probability density for t_i is thus given by $p_{t_i}(t) \equiv \lim_{\Delta t \to 0^+} \frac{\text{Prob}\{t \leq t_i < t + \Delta t\}}{\Delta t} = k_i e^{-k_i t}$, $t > 0$. The probability density for $t = t_1 + t_2 + \cdots + t_m$ then becomes a convolution of m factors:

$$p(t) = p_{t_1}(t) * \cdots * p_{t_m}(t) = k_1 e^{-k_1 t} * \cdots * k_m e^{-k_m t} \tag{2}$$

where the convolution operation "*" is defined as $f(t) * g(t) \equiv \int_0^t f(t')g(t-t')\mathrm{d}t'$. This probability density function $p(t)$ will henceforth be referred to as the "overall kinetic data" or "kinetic data", from which we will try to dissect the multi-step reaction kinetics and determine the number of reaction steps m, as well as the rate constants for each step k_1, k_2, … and k_m.

A specific case of eq 1 is the situation where $k_1 = k_2 = \cdots = k_m = k$:

$$A_1 \xrightarrow{k} A_2 \xrightarrow{k} A_3 \xrightarrow{k} \cdots \xrightarrow{k} A_m \xrightarrow{k} B. \tag{3}$$

In this case, the kinetic data $p(t)$ is simply the well-known Γ-distribution:

$$p(t) = \underbrace{p_{t_1}(t) * \cdots * p_{t_m}(t)}_{\text{convolution of } m \text{ copies}} = \frac{k^m t^{m-1}}{(m-1)!} e^{-kt}. \tag{4}$$

This functional form is widely used in homogeneous sequential kinetic models in physical chemistry. Two notable examples are photon counting process and the reaction of processive enzymes. In the latter context, the enzyme exhibiting a reaction scheme as shown in eq 3 is

often referred to as a "Poisson enzyme".[16] Hence, we will refer to the reaction scheme eq 3 as "Poisson enzyme reaction" hereafter.

Despite the apparent difference between the reaction schemes eq 1 and eq 3 (and their respective kinetic data eq 2 and eq 4), they are intrinsically connected. eq 2 is related to eq 4 through a weighted superposition:

$$k_1 e^{-k_1 t} * k_2 e^{-k_1 t} \cdots * k_m e^{-k_m t} = \frac{t^{m-1}}{(m-1)!} \int k^m e^{-kt} \mathrm{d}\mu_{k_1 k_2 \cdots k_m}(k), \tag{5}$$

where $\mu_{k_1 k_2 \cdots k_m}(k)$ is a non-decreasing function of k, bounded by 0 and 1. Physically speaking, this means that the prototypical m-step reaction $A_1 \xrightarrow{k_1} A_2 \xrightarrow{k_2} \cdots A_m \xrightarrow{k_m} B$ can always be realized by superposing many "Poisson enzyme reactions" $A_1 \xrightarrow{k} A_2 \xrightarrow{k} \cdots A_m \xrightarrow{k} B$, with a weighted distribution of k.

We defer the rigorous derivation of eq 5 and the explicit form of $\mu_{k_1 k_2 \cdots k_m}(k)$ to Supporting Information A.1. Here, we will only illustrate the validity of eq 5 for the specific case $m = 2$, where we have the following result:

$$\begin{aligned} p_{k_1 k_2}(t) &= k_1 e^{-k_1 t} * k_2 e^{-k_2 t} \\ &= \begin{cases} \frac{k_1 k_2}{k_2 - k_1}(e^{-k_1 t} - e^{-k_2 t}), & k_1 \neq k_2 \\ (k*)^2 t e^{-k* t}, & k_1 = k_2 = k* \end{cases} \end{aligned} . \tag{6}$$

Noting that the identity $e^{-k_1 t} - e^{-k_2 t} = t \int_{k_1}^{k_2} e^{-kt} \mathrm{d}k$ is valid for both $k_1 = k_2$ and $k_1 \neq k_2$, we can cast $p_{k_1 k_2}(t)$ in eq 6 into the unified form as

$$p_{k_1 k_2}(t) = t \int k^2 e^{-kt} \mathrm{d}\mu_{k_1 k_2}(k), \text{ with } \mu_{k_1 k_2}(k) = \mathrm{Prob}\{\xi/k_1 + (1-\xi)/k_2 > 1/k\}. \tag{7}$$

Here, ξ is a random variable uniformly distributed on the closed interval [0,1].

We can generalize eq 5 to reaction schemes that are more complex than shown in eq 1. For instance, we may allow the rate constant k_i of the ith step to sort from a random distribution, instead of taking a fixed value. In the reaction scheme

$$A_1 \xrightarrow[F_1(k)]{k_1} A_2 \xrightarrow[F_2(k)]{k_2} \cdots A_m \xrightarrow[F_m(k)]{k_m} B,$$

the random variable k_i obeys the cumulative distribution $F_j(k) \equiv \mathrm{Prob}\{k_j < k\}$. In this case, the corresponding kinetic data will take the form:

$$p(t) = [\int k e^{-kt} \mathrm{d}F_1(k)] * \cdots * [\int k e^{-kt} \mathrm{d}F_m(k)] = \frac{t^{m-1}}{(m-1)!} \int k^m e^{-kt} \mathrm{d}\mu(k), \tag{8}$$

Here, $\mu(k) = \int \mathrm{d}F_1(k_1) \cdots \int \mathrm{d}F_m(k_m) \mu_{k_1 \cdots k_m}(k)$ is still a non-decreasing function of k, bounded by 0 and 1. Therefore, the "Poisson enzyme reactions" $A_1 \xrightarrow{k} A_2 \xrightarrow{k} \cdots A_m \xrightarrow{k} B$ are elementary building blocks of arbitrarily complicated m-step sequential reactions.

From eq 8, we make two observations:

- **i.** The analysis of m-step sequential reactions boils down to the reconstruction of a bounded and non-decreasing function $\mu(k)$.
- **ii.** The number of reaction steps m is imbedded in the $t \to 0^+$ behavior of $p(t)$: for any non-decreasing function $\mu(k)$, we have $p(t) \propto t^{m-1}$ as $t \to 0^+$.

Motivated by these observations, we propose to analyze the sequential kinetic data with four sequential steps:

1. Performing a "heat capacity transform" $\hat{C}$ to the kinetic data $p(t)$, and deducing the number of reaction steps m by exploiting the asymptotic behavior $p(t) \propto t^{m-1}$, $t \to 0^+$;
2. Applying an "m-to-1 transformation" $\hat{J}_m$ to the kinetic data, and casting it into the form $\mathcal{P}(t) = (\hat{J}_m\, p)(t) = \int k e^{-kt}\, \mathrm{d}\mu(k)$;
3. Reconstructing the non-decreasing function $\mu(k)$ from the 1-step kinetic data $\mathcal{P}(t) = \int k e^{-kt}\, \mathrm{d}\mu(k)$ using the "phase function approach";[29]
4. Extracting kinetic information from the reconstructed $\mu(k)$.

2.2 Deriving the Number of Reaction Steps through the "Heat Capacity Transform"

In the following, we will develop a method to deduce the number of reaction steps m from kinetic data as shown in eq 8.

The function $p(t)$ has a universal asymptotic behavior $p(t) \propto t^{m-1}$, $t \to 0^+$, so one may naively guess that m can be deduced by computing the limiting behavior of $d \ln p(t)/d \ln t$, as $t \to 0^+$. In reality, however, the estimate of $d \ln p(t)/d \ln t$ can encounter the "ln0" instability when noise is superimposed on the kinetic data $p(t)$. Here, we will suppress such instability by applying an integral transform to the kinetic data $p(t)$.

Before presenting the integral transform, we first quantify the noise source that affects the detection of $p(t)$. Taking single molecule experiments as examples, the kinetic data derived from single-molecule experiments are often described as a histogram, where the histogram count $n_\tau(t)$ in the bin $[t, t+\tau)$ describes the number of molecules that reach the final state B in the time interval $[t, t+\tau)$. The count $n_\tau(t)$ asymptotically scales with the probability density $p(t)$:

$$n_\tau(t) \sim N \tau p(t), \text{for } N \ggg 1. \tag{9}$$

Here, $N = \Sigma^{(\tau)} n_\tau(t)$ is the total number of molecules counted and "$\Sigma^{(\tau)}$" denotes summation over all the time bins (i.e. over the entire observation time window). The uncertainty in $p(t)$ arises from the Poisson counting noise and thus the variance of $n_\tau(t)$ is equal to its mean:

$$\langle [\delta n_\tau(t)]^2 \rangle = \langle n_t(t) \rangle \approx N \tau p(t). \tag{10}$$

From this, it is clear that we cannot naively estimate $m \sim \mathrm{d} \ln p(t)/\mathrm{d} \ln t$, $t \to 0^+$, due to the numerical uncertainty $|\delta \ln p(t)| = |\frac{\delta n_\tau(t)}{n_\tau(t)}| \sim 1/\sqrt{n_\tau(t)} \to +\infty$ (Note: $n_\tau(t) \propto p(t) \propto t^{m-1} \to 0$, as $t \to 0^+$ for multi-step reactions where $m > 1$).

To overcome this numerical instability, we exploit the following transform:

$$(\widehat{C}n_\tau)(k) \equiv \sum^{(\tau)} n_\tau(t) \frac{(kt)^2 e^{kt}}{(e^{kt}-1)^2}, \tag{11}$$

which is the discretized version of the integral transform for $p(t)$,

$(\widehat{C}p)(k) \approx \int_0^{+\infty} p(t) \frac{(kt)^2 e^{kt}}{(e^{kt}-1)^2} \mathrm{d}t$. The integral transform $\hat{C}$ significantly suppresses the noise in $\delta n_\tau(t)$ and the asymptotic behavior of $(\hat{C}n_\tau)(k)$ is given by

$$(\widehat{C}n_\tau)(k) \approx N\tau \int_0^{+\infty} p(t) \frac{(kt)^2 e^{kt}}{(e^{kt}-1)^2} \mathrm{d}t \propto \frac{1}{k^m}, \quad k \to +\infty. \tag{12}$$

(Supporting Information B provides an analytic derivation of eq 12.) Therefore, we can derive the number of reaction steps, m, from a numerical estimate of the following limit:

$$\lim_{k \to +\infty} \frac{\mathrm{d}\ln(\widehat{C}n_\tau)(k)}{\mathrm{d}\ln(1/k)}. \tag{13}$$

Due to Poisson noise $\langle [\delta n_\tau(t)]^2 \rangle = \langle n_t(t) \rangle \approx N\tau p(t)$, the uncertainty in $(\hat{C}n_\tau)(k)$ is given by:

$$|\delta(\widehat{C}n_\tau)(k)| \le \min\left\{ \sqrt{N}, 7.21 \sqrt{\int_0^{+\infty} \frac{N(t)}{t} e^{-kt} \mathrm{d}t} \right\}. \tag{14}$$

(See Supporting Information B for a proof of eq 14.) Here, $N(t) \equiv \sum_{t'<t}^{(\tau)} n_\tau(t')$ is the cumulative count for molecules with reaction time smaller than t.

One may get more physical insights into the transform $\hat{C}$ by comparing it to Debye's formula that links the phonon spectrum $D(\nu)$ of a solid to its heat capacity C_V:

$$C_V = k_B \int_0^{+\infty} \mathrm{d}\nu D(\nu) \frac{(h\nu/k_B T)^2 e^{h\nu/k_B T}}{(e^{h\nu/k_B T}-1)^2}$$

where h is the Planck constant, ν is the phonon frequency, k_B is the Boltzmann constant, and T is the absolute temperature. For an m-dimensional solid, the phonon spectrum is $D(\nu) \propto \nu^{m-1}$ in the low frequency limit $\nu \to 0^+$ (which is analogous to $p(t) \propto t^{m-1}$ as $t \to 0^+$), and accordingly the heat capacity goes as $C_V \propto T^m$ in the low temperature limit $T \to 0^+$ (which is analogous to the asymptotic behavior $(\hat{C}p)(k)$: $(\hat{C}n_\tau)(k) \propto 1\ k^m$ as $k \to +\infty$). Just like the observation that the often rugged phonon spectrum is transformed into a smooth heat capacity curve as a function of temperature, the relatively noisy $p(t)$ will be transformed into a relative smooth $(\hat{C}n_\tau)(k)$, due to the low-pass filtering property of the transformation[30]. Hence, we referred to the integral transform $\hat{C}$ as the "heat capacity transform".

2.3 Reducing Multi-Step Kinetics to 1-Step Kinetics by the Integral Transform $\hat{J}_m$

Once the number of reaction step m is deduced, we will further extract the kinetic information for a multi-step reaction by reconstituting the non-decreasing distribution function $\mu(k)$, which relates to the kinetic data $p(t)$ through eq 8:

$$p(t)=\frac{t^{m-1}}{(m-1)!}\int k^m e^{-kt}\mathrm{d}\mu(k).$$

The distribution function $\mu(k)$ can further provide chemical information after we combine it with additional inputs about sequential reaction schemes. For the $m = 1$ case, one can stably reconstruct $\mu(k)$ from $p(t)$ using the phase function method.[29] To reconstruct $\mu(k)$ in the $m > 1$ case, we need to transform $p(t)$ to the form of "1-step kinetics".

Such an "m-to-1 transform" $\hat{J}_m$ can be realized as follows:

$$\widehat{J}_m p(t) \equiv (m-1)!\int_t^{+\infty}\mathrm{d}t_{m-1}\cdots\int_3^{+\infty}\mathrm{d}t_2\int_2^{+\infty}\mathrm{d}t_1\frac{p(t_1)}{t_1^{m-1}}. \tag{15}$$

This linear transform convert $p(t)$ to $\mathcal{P}(t) = (\hat{J}_m\, p)(t) \equiv \int k e^{-kt}\, \mathrm{d}\mu(k)$. The integral transform $\hat{J}_m$ is a numerically stable transformation, i.e. any small perturbation $f(t) = \delta p(t)$ to the kinetic data $p(t)$ will remain small after being transformed to $(\hat{J}_m\, f)(t)$. Quantitatively, one can verify the following inequality (See derivations in Supporting Information C): g

$$\int_0^{+\infty}[(\widehat{J}_m f)(t)]^2\mathrm{d}t \le c_m^2\int_0^{+\infty}[f(t)]^2\mathrm{d}t$$

and the noise amplification ratio c_m satisfies

$$c_m < \sqrt{\pi\left(m-\frac{1}{2}\right)\sqrt{\frac{m-\frac{1}{2}}{m+\frac{1}{2}}}\left(1-\frac{1}{4m^2}\right)^{-\frac{m}{2}}} \sim \sqrt{\pi\left(m-\frac{1}{2}\right)}. \tag{16}$$

We note that the reconstruction of $\mu(k)$ from the 1-step kinetic data also has a finite amplification ratio $2+\sqrt{3}$ (See Ref. 29). Thus, the mapping from the m-step kinetic data to the non-decreasing function $\mu(k)$ is numerically stable, the overall noise amplification ratio being $(2+\sqrt{3})c_m < +\infty$.

2.4 Reconstructing $\mu(k)$ from the Kinetic Data $p(t)$ Using the "Phase Function Approach"

We have previously developed a "phase function method" to reconstruct a non-decreasing $\mu(k)$ from 1-step reaction kinetics $\mathcal{P}(t) = (\hat{J}_m\, p)(t) \equiv \int k e^{-kt}\, \mathrm{d}\mu(k)$.[29] Now that we have reduced the m-step reaction kinetics to 1-step reaction kinetics by the "m-to-1 transform" $\hat{J}_m$, we can use the phase function approach to deduce $\mu(k)$. The main idea was to regard exponential decay e^{-kt} as an oscillation $e^{i\omega t}$ in purely imaginary frequency $\omega = ik$. We can thus decompose the kinetic data as superposition of oscillation modes by using Fourier transform and then converts information on the Reω-axis to the Imω-axis (k-axis).[29]

The "phase function approach" reconstructs a non-decreasing $\mu(k)$ in three steps as outlined below:[29]

1. Perform a Fourier transform to the data $\wp(t)$ for complex-valued frequency $\omega = \mathrm{Re}\omega + i\,\mathrm{Im}\omega$ satisfying $\mathrm{Im}\omega \le 0$, and determine the phase function

$$\varphi(\omega) = \arg\left[\int_0^{+\infty} \wp(t) e^{-i\omega t} \mathrm{d}t\right];$$

2. Since $\varphi(\omega) = \mathrm{Im}\ln\left[\int_0^{+\infty} \wp(t) e^{-i\omega t} \mathrm{d}t\right]$ is the imaginary part of an analytic function $\ln\left[\int_0^{+\infty} \wp(t) e^{-i\omega t} \mathrm{d}t\right]$, it must satisfy the two-dimensional Laplace equation $\partial^2\varphi(\omega)/\partial(\mathrm{Re}\omega)^2 + \partial^2\varphi(\omega)/\partial(\mathrm{Im}\omega)^2 = 0$. We can thus numerically solve this Laplace equation to deduce $\varphi(\omega)$ for all $\mathrm{Re}\omega \neq 0$, which satisfies[29]

$$0 \le \varphi(\omega) < \pi, \quad \text{for} \mathrm{Re}\omega < 0; \tag{17}$$

3. We can then obtain the phase function on the imaginary axis $\varphi(ik) \equiv \lim_{\varepsilon \to 0^+} \varphi(ik - \varepsilon)$, which contains all the information necessary for the reconstruction of $\mu(k)$ and allows $\mu(k)$ to be derived using the following transform $\hat{R}$:

$$\mu(k) = (\widehat{R}\varphi)(k) \equiv \lim_{\varepsilon \to 0^+} \int_0^k \mathrm{d}k' \left\{ \frac{\sin\varphi(ik' - \varepsilon)}{\pi k'} \exp\left[-\int \ln|1 - \frac{k' + i\varepsilon}{k''}| \frac{\mathrm{d}\varphi(ik'')}{\pi} \right] \right\}. \tag{18}$$

The non-decreasing property of $\mu(k)$ and the boundedness of the associated phase function $0 \le \varphi(ik) \le \pi$ ensure the numerical stability of this algorithm against noise in the time domain data.[29] The uncertainty in $\mu(k) = (\hat{R}\varphi)(k)$ is estimated by:

$$|\delta\mu(k)| \sim (2 + \sqrt{3})\varepsilon^{\#} \sqrt{\mu(k)}, \tag{19}$$

where $\varepsilon^{\#}$ is the noise level (t-domain relative error) in the kinetic data $\wp(t)$.

2.5 From $\mu(k)$ to Kinetic Information of Interest

From $\mu(k)$, we can compute a frequency spectrum of the kinetic data $p(t)$ as:

$$\tilde{p}(\omega) = (\widehat{\Omega}_m \mu)(\omega) \equiv \int \left(1 + \frac{i\omega}{k}\right)^{-m} \mathrm{d}\mu(k), \ \mathrm{Re}\omega \neq 0. \tag{20}$$

The function $\tilde{p}(\omega)$ is referred to as the "frequency spectrum" because it coincides with the Fourier transform of $p(t) = \frac{1}{(m-1)!} \int k^m t^{m-1} e^{-kt} \mathrm{d}\mu(k)$:

$$\tilde{p}(\omega) = (\widehat{\Omega}_m \mu)(\omega) = \int_0^{+\infty} p(t) e^{-i\omega t} \mathrm{d}t, \ \mathrm{Im}\omega \le 0.$$

This spectral representation $\tilde{p}(\omega) = (\Omega \wedge_m \mu)(\omega)$ can help reveal kinetic information that is implicitly imbedded in the non-decreasing function $\mu(k)$, as shown below.

Consider the m-step sequential reaction scheme $A_1 \xrightarrow[F_1(k)]{k_1} A_2 \xrightarrow[F_2(k)]{k_2} \cdots A_m \xrightarrow[F_m(k)]{k_m} B$, where the rate constant k_j obeys the probability distribution $F_j(k) = \mathrm{Prob}\{k_j < k\}$ for $j = 1,2,\ldots, m$. The scheme invokes no further branched pathways. Chemically speaking, in this scenario, we assume that the molecule has to challenge an energy barrier with distributed barrier height at each reaction

step, but the distributions of barrier heights in distinct reaction steps are independent. From the kinetic data in the *t*-domain $p(t) = [\int ke^{-kt}\, dF_1(k)] * \cdots * [\int ke^{-kt}\, dF_m(k)]$, which is a convolution of *m* factors, one can get the frequency spectrum in the ω-domain as the product of *m* single-step reaction frequency spectra:

$$\tilde{p}(\omega) = \left[\int \left(1+\frac{i\omega}{k}\right)^{-1} dF_1(k)\right] \times \cdots \times \left[\int \left(1+\frac{i\omega}{k}\right)^{-1} dF_m(k)\right] = \prod_{j=1}^{m} \int \left(1+\frac{i\omega}{k}\right)^{-1} dF_j(k). \tag{21}$$

Because the function $\tilde{p}(\omega)$ given in eq 21 satisfies $0 < |\tilde{p}(\omega)| < +\infty$ for $\mathrm{Re}\omega < 0$, it is possible to define a *single-valued* phase function $\psi(\omega) = \mathrm{Im}\ln \tilde{p}(\omega)$ as

$$\psi(\omega) \equiv \mathrm{Im}\int_0^{\omega} d\ln \tilde{p}(\omega') = \arg \tilde{p}(\omega) = \arg \int \left(1+\frac{i\omega}{k}\right)^{-m} d\mu(k),\ \mathrm{Re}\omega < 0. \tag{22}$$

(See Supporting Information A.2 for a proof of the above statement.) According to eq 21, the phase $\psi(\omega)$ further decomposes to the sum of *m* terms:

$$\psi(\omega) = \arg \tilde{p}(\omega) = \sum_{j=1}^{m} \varphi_j(\omega),\ \mathrm{Re}\omega < 0,\ \text{where}$$
$$0 \le \varphi_j(\omega) = \arg\left[\int \left(1+\frac{i\omega}{k}\right)^{-1} dF_j(k)\right] \le \arg \frac{1}{i\omega} \le \pi.$$

As we bring $\psi(\omega)$ to the limit $\mathrm{Re}\omega \to 0^-$, we obtain:

$$\psi(ik) = \varphi_1(ik) + \cdots + \varphi_m(ik) = \sum_{j=1}^{m} \varphi_j(ik),\ 0 \le \varphi_j(ik) \le \pi,\ \text{for}\ j = 1,2,\ldots,m.$$

Evidently, this algebraic sum in the *k*-domain is a mathematically more convenient form than the original *m*-term convolution in the *t*-domain: $p(t) = [\int ke^{-kt}\, dF_1(k)] * \cdots * [\int ke^{-kt}\, dF_m(k)]$. Here, we note that the crucial bridge between *k*- and *t*-domains builds on the condition $0 < |\tilde{p}(\omega)| < +\infty$ for $\mathrm{Re}\omega \neq 0$, but such a bound on $|\tilde{p}(\omega)|$, while being rigorously true for sequential reactions following the form $A_1 \xrightarrow[F_1(k)]{k_1} A_2 \xrightarrow[F_2(k)]{k_2} \cdots A_m \xrightarrow[F_m(k)]{k_m} B$, may not necessarily hold for branched reaction schemes. For a more general reaction, it is possible that $\tilde{p}(\omega)$ vanishes for some $\mathrm{Re}\omega \neq 0$, leading to singularities with $\ln \tilde{p}(\omega) = \ln 0 = \infty$ and ill-defined phase functions.

Naturally, the next question is how to determine the distributions of rate constants $F_j(k) = \mathrm{Prob}\{k_j < k\}$ ($j = 1,2,\ldots, m$) from the phase function decomposition $\psi(ik) = \sum_{j=1}^{m} \varphi_j(ik)$? Clearly, once we know each individual $\varphi_j(ik)$, we can unambiguously reconstruct $F_j(k) = (\hat{R}\varphi_j)(k)$ (using eq 18). However, with the mere knowledge of one kinetic data set $p(t)$, the best one can obtain is just the sum of *m* phase functions $\varphi_j(ik)$, $j = 1,2, \ldots, m$, and the problem of recovering the *m* individual summands is *underdetermined*. To fully characterize the kinetics, we usually need additional inputs.

Here, we will discuss two often encountered circumstances in which additional knowledge can help us fully dissect the multi-step reaction kinetics using $\psi(ik)$.

Situation 1: The distributions of k_j for each step are identical—This is often assumed for the movement kinetics of motor proteins (such as kinesins[31] and myosins[14]) and nucleic acid-translocating enzymes (such as polymerases[15] and helicases[32]) where the elementary motion steps adopt approximately identical rate constants. Under this assumption, we can model the kinetic data $p(t)$ by the convolution of m identical copies of single-step reaction kinetics:

$$p(t)=\underbrace{\left[\int ke^{-kt}\mathrm{d}F(k)\right]*\cdots*\left[\int ke^{-kt}\mathrm{d}F(k)\right]}_{\text{convolutions of } m \text{ copies in total}}.$$

Accordingly, the phase function of the overall sequential reaction is given by $\psi(ik)=\sum_{j=1}^{m}\varphi_j(ik)$, with all individual phase functions to be identical: $\varphi_j(ik) = \psi(ik)/m$. Therefore, we can determine $F(k)$ using the reconstruction formula eq 18: $F(k) = (\hat{R}\varphi_j)(k)$.

Situation 2: The distributions of k_j scale independently with external experimental conditions—In many cases, the distribution of rate constants of each step of a multi-step process can be independently tuned by external conditions. A simple example is that the rate constant of enzyme-substrate binding process scales linearly with respect to concentration of the substrate. In these cases, we are not only able to determine a single phase function

$$\psi(ik)=\varphi_1(ik)+\cdots+\varphi_m(ik)$$

but can also determine the scaled phase function

$$\psi^{(s_1,\cdots,s_m)}(ik)=\varphi_1(ik/s_1)+\cdots+\varphi_m(ik/s_m)$$

from experimentally measured $p(t)$ functions at the scaled experimental conditions. Here, s_1, …, s_m are sets of scaling factors for these conditions (substrate concentration etc.). By choosing a sufficient number of different sets of (s_1, …, s_m) (i.e. a sufficient number of independent kinetic experiments), we can uniquely determine the functional forms of individual phase functions φ_1, …,φ_m and thus the rate constant distributions of $F_j(k) = (\hat{R}\varphi_j)(k)$.

To flesh out this idea, we illustrate the case $m = 2$ with the enzymatic reaction:

$$\mathrm{E+S}\xrightarrow{k_1=\kappa_1[S]}\mathrm{ES}\xrightarrow{k_2}\mathrm{E+P}.$$

Here, the enzyme (E) binds a substrate (S) to form an enzyme-substrate complex (ES) with a distributed binding rate constant $k_1 = \kappa_1[\mathrm{S}]$, where [S] is the concentration of the substrate. The enzyme-substrate complex (ES) then undergoes a catalytic step, executed with a distributed rate constant k_2 ($F_2(k) = \mathrm{Prob}\{k_2 < k\}$) to release the product (P). Suppose that we conduct two experiments: the first for $[\mathrm{S}] = [\mathrm{S}]_0$, and the second for $[\mathrm{S}] = s[\mathrm{S}]_0$, where $s > 1$ is a scaling factor. The resulting kinetic data sets are $p^{[1]}(t)$ and $p^{[2]}(t)$, respectively. If the distribution of $k_1^{[1]}=\kappa_1[\mathrm{S}]_0$ in the first experiment is denoted by $F_1^{[1]}(k)=\mathrm{Prob}\{k_1^{[1]}<k\}$, then the distribution of $k_1^{[2]}=\kappa_1 s[\mathrm{S}]_0$ in the second experiment is given by $F_1^{[2]}(k)=\mathrm{Prob}\{k_1^{[2]}<k\}=F_1^{[1]}(k/s)$. Now we can reconstruct, from $p^{[1]}(t)$ and $p^{[2]}(t)$, two phase functions

$$\psi^{[1]}(ik)=\varphi_1(ik)+\varphi_2(ik)$$
$$\psi^{[2]}(ik)=\varphi_1(ik/s)+\varphi_2(ik),$$

respectively. We can compute the difference between these two phase functions to get a function

$$g(k) \equiv \psi^{[1]}(ik) - \psi^{[2]}(ik)=\varphi_1(ik) - \varphi_1(ik/s),$$

from which we can fully recover the phase function $\varphi_1(ik)$ as

$$\varphi_1(ik)=[\varphi_1(ik) - \varphi_1(ik/s)]+[\varphi_1(ik/s) - \varphi_1(ik/s^2)]+\cdots=\sum_{q=0}^{\infty} g(k/s^q). \quad (23)$$

Such a series actually terminates after finite terms, because when $k < k$ (k is the slowest detectable rate constants due to the finite time duration of the experiment), $\varphi_1(ik) = 0$ and thus $g(k)=0$. Therefore, we can fully determine φ_1 (ik) and φ_2 (ik), from which we can directly determine the rate constant distributions governing the two steps of the enzymatic reaction using $F_1^{[1]}=\widehat{R}\varphi_1$, $F_2 = \hat{R}\varphi_2$ where $\hat{R}$ is again given by eq 18. This $m = 2$ example can be generalized to any reaction form $A_1 \xrightarrow[F_1(k)]{k_1} A_2 \xrightarrow[F_2(k)]{k_2} \cdots A_m \xrightarrow[F_m(k)]{k_m} B$ with $m > 2$ as long as the rate constant at all but one step can be scaled independently (for example, by tuning the concentration of (m−1) distinct substrates that bind to an enzyme in a sequential manner).

3. Numerical Results and Discussion

Here, we shall numerically illustrate the above theories and address several practical problems in the analysis of multi-step reaction kinetics:

1. Determine the number of reaction steps from noisy kinetic data;
2. Deduce the probability distribution of rate constants in each reaction step based on the assumption that the distribution is identical for all steps;
3. Determine the probability distributions of individual reaction steps in “enzymatic reactions” where the rate constant of each step can be independently scaled.

3.1 Finding the Number of Reaction Steps

To simulate kinetic data with realistic experimental noise (such as the noise level carried by single-molecule experiments), we used Monte Carlo simulations to construct the histograms of overall reaction times for a variety of reaction schemes in the form of $A_1 \xrightarrow[F_1(k)]{k_1} A_2 \xrightarrow[F_2(k)]{k_2} \cdots A_m \xrightarrow[F_m(k)]{k_m} B$ (m = 2, 3, 4 and 5) (Figures 1*a*, 1*c*, 1*e* and 1*g*). Here the overall reaction time of each molecule is defined as the dwell time in all the *A* states (“reactant” and “intermediate” states) before the molecule reach the “product” state *B*. Each reaction step may take a fixed rate constant (Figures 1*a*, 1*e* and 1*g*) or sort from a non-trivial distribution (Figure 1*c*). The distributions of rate constants may be identical (Figures 1*a* and 1*c*) or distinct (Figures 1*e* and 1*g*) among the different reaction steps.

In more details, for a given reaction scheme $A_1 \xrightarrow[F_1(k)]{k_1} A_2 \xrightarrow[F_2(k)]{k_2} \cdots A_m \xrightarrow[F_m(k)]{k_m} B$, a random reaction time was simulated as the sum of m random integers $R = R_1 + R_2 + \cdots + R_m$, where R_j follows the distribution $\text{Prob}\{R_j > t\} = \int e^{-kt}\, dF_j(k)$. A total of $N = 8000$ such random integers R are simulated for each reaction scheme. The 8000 reaction times are then binned to construct a histogram, $n_\tau(t)$ (Figures 1*a*, 1*c*, 1*e* and 1*g*), with the bin size equal to the time resolution τ, which is taken as $\tau = 1$ in all examples shown in this paper. Such a Monte Carlo approach was designed to reflect realistic noise arising from finite time resolution and finite total counts present in single molecule experiments.

We then tested the "heat capacity transform" approach with these simulated kinetic data (Figures 1*b*, 1*d*, 1*f*, and 1*h*). First we performed the transform $\hat{C}$ on the simulated histogram $n_\tau(t)$:

$$(\widehat{C}n_\tau)(k) \equiv \sum_{t/\tau=0,1,2,\ldots} n_\tau(t)\frac{(kt)^2 e^{kt}}{(e^{kt}-1)^2}.$$

As shown in Figures 1*b*, 1*d*, 1*f* and 1*h*, the transform yields $(\hat{C}n_\tau)(k) \approx N = 8000$ in the small k limit as expected. To obtain a numerically stable estimate of the number of reaction steps using the large k limit, $m = \lim_{k\to+\infty} \frac{d\ln(\widehat{C}n_\tau)(k)}{d\ln(1/k)}$, we took the following procedure to estimate the derivative and limit: (1) We estimated the derivative with a dynamic window size $\varepsilon_k = \ln|(\hat{C}n_\tau)(k) + |\delta(\hat{C}n_\tau)(k)|| - \ln(\hat{C}n_\tau)(k)$ where $|\delta(\hat{C}n_\tau)(k)|$ is noise level in $(\hat{C}n_\tau)(k)$ as given in eq 14. As a result,

$$m(k) \equiv \frac{d\ln[(\widehat{C}n_\tau)(k)]}{d\ln(1/k)} \approx \frac{\ln[(\widehat{C}n_\tau)(ke^{-\varepsilon_k})] - \ln[(\widehat{C}n_\tau)(k)]}{\varepsilon_k}. \quad (24)$$

(2) To properly estimate the limit at large k, we note that the rate constant $k > 1/\tau$ cannot be determined by an experiment with time resolution of τ and that the estimate of $\ln[(\hat{C}n_\tau)(k)]$ is reliable only when $(\hat{C}n_\tau)(k) \pm |\delta(\hat{C}n_\tau)(k)| > 0$. Bearing these two facts in mind, we set a maximum value for k in the evaluation of $\lim_{k\to+\infty}[d\ln(\widehat{C}n_\tau)(k)/d\ln(1/k)]$, i.e.

$$k_{\max} \equiv \min\{1/\tau, k_{\text{cut-off}}\}, \quad (25)$$

where $k_{\text{cut-off}} \equiv \max\{k: |\delta(\hat{C}n_\tau)(k)| \le (\hat{C}n_\tau)(k)\}$. Rather than literally pursuing $k \to +\infty$, we calculate the number of steps as the nearest integer to $m^* = \lim_{k\to k_{\max}} m(k)$.

As shown in Figure 1, the m^* values recover the correct numbers of reaction steps to within an error of ± 10%. This is not only true for multi-step reactions with identical rate constant distribution at each step (Figures 1*a*–1*d*) but also true for reactions with disparate rate constants at different steps (Figures 1*e*–1*h*). This method applies to reactions with a single rate constant in each step (Figures 1*a*, 1*b*, and 1*e*–1*h*) as well as reactions with a broad distribution of rate constants (Figures 1*c* and 1*d*). The number of reaction steps as high as $m = 5$ can be faithfully recovered (Figures 1*g* and 1*h*) when the noise level for the t-domain kinetic data is ~1% (Poisson counting noise corresponding to $N = 8000$). We expect that with lower noise in the

kinetic data, even high numbers of the reaction steps can be accurately deduced (See Supporting Information D). In comparison, the "randomness parameter" defined by

$$r=\frac{\langle t^2\rangle-\langle t\rangle^2}{\langle t\rangle^2}=\frac{\langle(t-\langle t\rangle)^2\rangle}{\langle t\rangle^2} \tag{26}$$

gives r = 0.51, 0.75, 0.34 and 0.35 for the kinetic data shown in Figures 1*a*, 1*c*, 1*e* and 1*g*, respectively. Here $\langle\cdot\rangle$ denotes an average weighted by the reaction time histogram $n_\tau(t)$. The lower bound of the number of reaction steps is then given by $m = 1/r$, corresponding to ~2, 1.3, 3 and 3 steps for the 4 cases. Indeed, when each reaction step has a single and identical rate constant k, as shown in Fig. 1a, this lower bound accurately describes the actual number of reactions steps as expected. However, for the other cases, $1/r$ significantly deviated from the number of reactions steps as well as rate-limiting steps.

The "heat capacity transform" thus provides a more general and precise approach to derive the number of reaction steps from kinetic data. However, for fixed data precision in the time domain, the "heat capacity transform" may also fail when the number of reaction steps m or the dispersion of rate constants is too large (See Supporting Information D for more quantitative discussion).

3.2 Reconstructing Identically Distributed Rate Constants for Sequential Multi-Step Reactions

The major concern of this subsection is to fully determine the reaction scheme $A_1\xrightarrow[F_1(k)]{k_1}A_2\xrightarrow[F_2(k)]{k_2}\cdots A_m\xrightarrow[F_m(k)]{k_m}B$ from kinetic data $p(t)$, provided that $k_1, \cdots, k_m$ are identically distributed: $F_1(k) = \cdots = F_m(k)$. This type of kinetic scheme can approximate many processive biomolecular processes. For instance, we may wish to characterize the kinetic profile of an RNA polymerase molecule that transcribes a piece of DNA or a helicase molecule that unwinds a DNA or RNA duplex. The typical kinetic data often provides information about how long it takes the polymerase molecule to transcribe DNA with a certain number of nucleotides or how long it takes the helicase to unwind a duplex of a certain number of base-pairs without directly revealing the actual translocation step-sizes of these molecules. If we neglect the sequence dependence, the kinetics of these processive biochemical reactions can be approximated by the scheme $A_1\xrightarrow[F(k)]{k_1}A_2\xrightarrow[F(k)]{k_2}\cdots A_m\xrightarrow[F(k)]{k_m}B$. Using the "heat capacity transform" $\hat{C}$, we may deduce the number of reaction steps m required for translocating across n nucleotides. Subsequently, the step-size of the molecule of interest can be computed as n/m and then the probability distribution of rate constants for an elementary translocation step can be deduced.

In Figure 2, we simulated the situation for m = 3 or 5 with dispersed kinetic rate constants. The distributions of the rate constants were identical but independent of each other. The reaction times in Figure 2 were simulated as the sum of random integers $R = R_1 + \cdots + R_m$, m = 3 or 5. For the m = 3 case (Figure 2*a*), the rate constants were allowed to take two discrete values with equal probability, mimicking the scenario that the molecule has two distinct conformation each with a distinct reaction rate (as depicted by the black curve in Figure 2*b*). For the m = 5 case (Figure 2*c*), the reaction barriers for each step (i.e. $\ln k$) were taken to be uniformly distributed on an interval (as depicted by the black curve in Figure 2*d*).

We used four steps to deduce the number of reaction steps and the underlying $F(k)$ for each step from the simulated overall kinetic data given in Figures 2*a* and 2*c*.

i. We deduced the number of reaction steps by applying the "heat capacity transform" $\hat{C}$ to the simulated histogram $n_\tau(t)$. Taking the limit of $m^* = \lim_{k\to k_{\max}} m(k)$ with $m(k)$ given by eq 24, the simulated histogram shown in Figure 2*a* leads to an estimate $m^* = 3.21$, which rounds off to $m = 3$, agreeing quantitatively with the preset number of reaction steps. Similarly, when data in Figure 2*c* were analyzed by transform $\hat{C}$, the result is $m^* = 4.61$, which rounds off to $m = 5$, also agreeing with the preset number of reaction steps.

ii. We applied the "*m*-to-1" transformation $\hat{J}_m$ to the raw data $n_\tau(t)$. After $\hat{J}_m$ transformation, the kinetic data took the form $\int ke^{-kt}d\mu(k)$, where the non-decreasing function $\mu(k)$ can be reconstructed via the "phase function approach" (data not shown).

iii. We computed the frequency spectrum $\tilde{p}(\omega) = (\hat{\Omega}_m\mu)(\omega) \equiv \int(1+i\omega/k)^{-m}\, d\mu(k)$ from $\mu(k)$, and deduced the phase function. The phase function for individual reaction steps were then deduced from $\varphi(ik) = \psi(ik)/m$.

iv. We reconstructed the rate constant distributions $F(k)$ from $\varphi(ik)$ using the $\hat{R}$ transform as shown in eq 18.

The reconstructed rate constant distributions are shown as the red curves in Figures 2*b* and 2*d*, with the vertical error bars given by $|\delta F(k)| \sim (2+\sqrt{3})\varepsilon^{\#}\sqrt{F(k)}$ (eq 19). The reconstruction result compares quantitatively well (within error) with the preset distribution of rate constants (black curves) in both cases. The horizontal error bars on the preset $F(k)$ (depicted by the grey zones) arise from finite sampling error in the simulated data (i.e. at the noise level (~1%) corresponding to the Poisson counting error for $N = 8000$, a single exponential decay cannot be distinguished from the superposition of exponential decay modes with a range of rate constants defined by $\Delta\ln k = \pm 0.4$.)

3.3 Reconstructing Michaelis-Menten Kinetics with Distributed Rate Constants

In this subsection, we dissect a 2-step enzymatic reaction $\mathrm{E+S} \xrightarrow{k_1=\kappa_1[S]} \mathrm{ES} \xrightarrow{k_2} \mathrm{E+P}$. The rate of the substrate-binding step is tunable by substrate concentration [S]. We aim to reconstruct the probability distribution of rate constants in both the substrate-binding step and the catalysis step.

In Figure 3 we simulated two enzymatic reactions:

$$\mathrm{E+S} \xrightarrow{k_1=\kappa_1[S]_0} \mathrm{ES} \xrightarrow{k_2} \mathrm{E+P}\ (\text{Figure } 3a)$$

$$\mathrm{E+S} \xrightarrow{k_1=10\kappa_1[S]_0} \mathrm{ES} \xrightarrow{k_2} \mathrm{E+P}\ (\text{Figure } 3b)$$

(Here, the unidirectional reaction schemes are approximations to the Michaelis-Menten scheme that involves one reversible step:

$$\mathrm{E+S} \underset{\kappa_{-1}}{\overset{\kappa_1[S]}{\rightleftarrows}} \mathrm{ES} \xrightarrow{k_2} \mathrm{E+P},$$

under the conditions $\kappa_{-1} \ll k_2$ and $\kappa_{-1} \ll \kappa_1[S]$. The unidirectional approximation is often valid as $\kappa_{-1} \ll k_2$ is satisfied for a wide class of enzymes and it is also possible to achieve $\kappa_{-1} \ll \kappa_1[S]$ by properly tuning the substrate concentration [S].) Two independent numerical simulations for the two conditions $[S] = [S]_0$ and $[S] = 10[S]_0$ were conducted to obtain the

reaction time histograms in Figures 3*a* and 3*b*, with the preset rate constant distributions for $k_1 = \kappa_1[S]_0$ and k_2 as shown in Figures 3*c* and 3*d* (black curves).

The numerical reconstruction of $F_1(k) = \text{Prob}\{k_1 = \kappa_1[S]_0 < k\}$ and $F_2(k) = \text{Prob}\{k_2 < k\}$ was achieved in the following four steps.

- **i.** We applied the "2-to-1 transformation" $\hat{J}_2$ to both histograms in Figures 3*a* and 3*b*, and reconstructed two non-decreasing functions $\mu^{[1]}(k)$ and $\mu^{[2]}(k)$ (not shown) using the phase function method.
- **ii.** We computed the phase functions $\psi^{[1]}(ik) = \arg[(\hat{\Omega}_2\mu^{[1]})(ik)]$ and $\psi^{[2]}(ik) = \arg[(\hat{\Omega}_2\mu^{[2]})(ik)]$ using eqs 20 and 22.

  From $\psi^{[1]}(ik) = \varphi_1(ik) + \varphi_2(ik)$, and $\psi^{[2]}(ik) = \varphi_1(ik/10) + \varphi_2(ik)$, we deduced the phase function corresponding to the each reaction step, $\varphi_1(ik)$ and $\varphi_2(ik)$, using eq 23.
- **iii.** We reconstructed the distribution of rate constants for each step as $F_1(k) = (\hat{R}\varphi_1)(k)$ and $F_2(k) = (\hat{R}\varphi_2)(k)$ using eq 18.

The reconstructed cumulative distributions $F_1(k) = \text{Prob}\{\kappa_1[S]_0 < k\}$ and $F_2(k) = \text{Prob}\{k_2 < k\}$ (red curves in Figures 3*c* and 3*d*) both agree well with the preset distribution (black curves) within the finite sampling error depicted by the grey zones in the figures, arising from the finite number reaction times simulated ($N = 8000$). Here, the rate constant distributions in a 2-step reaction were uniquely determined by simultaneous analysis of two experimental data sets and did not invoke fitting the raw data to subjective models of the distributions. In contrast, typical fitting methods are not adequate to determine such arbitrary distributions of rate constants as shown in Figures 3*c* and 3*d*.

4. Conclusions

It is a challenging task to dissect the kinetics of a multi-step chemical reaction. This is especially true if the totally number of reaction steps are unknown *a priori* or if some of the reaction steps cannot be described by a single rate constant. In this work, we introduced two numerically stable integral transforms, the "heat capacity transform" $\hat{C}$ and the "*m*-to-1 transform" $\hat{J}_m$, to help analyzing the general *m*-step reaction kinetics $A_1 \xrightarrow[F_1(k)]{k_1} A_2 \xrightarrow[F_2(k)]{k_2} \cdots A_m \xrightarrow[F_m(k)]{k_m} B$ using the phase function approach.

The "heat capacity transform" $\hat{C}$ maps the overall kinetic data, which typically follows the form $p(t)=\frac{1}{(m-1)!}\int k^m t^{m-1}e^{-kt}\mathrm{d}\mu(k)$, to $(\widehat{C}p)(k')=\int_0^{+\infty} p(t)(k't)^2 e^{k't}/(e^{k't}-1)^2\mathrm{d}t$. It allows us to reliably infer the number of reaction steps *m* from noisy kinetic data, by tracking the asymptotic behavior $(\hat{C}p)(k') \propto 1/k'^m$, $k' \to +\infty$. The "heat capacity transform" method is applicable to general *m*-step reactions in which the rate constants can follow arbitrary distributions and/or are different among distinct steps. It should be noted, however, that the performance of the "heat capacity transform" is still subject to the numerical quality of the input data, i.e., it cannot resolve two reaction models, this distinction of which lie within the experimental noise. For example, while a single-molecule experiment with $N = 2000$ events may be sufficient to distinguish the 4-step and 5-step reaction models, experiments with $N = 200$ events could barely resolve 2-step versus 3-step reactions. (See Supporting Information D for a quantitative argument for this fundamental limitation.)

The "*m*-to-1 transform" $\hat{J}_m$ maps the overall kinetic data, $p(t)$ to $(m-1)!\int_t^{+\infty}\mathrm{d}t_{m-1}\cdots\int_3^{+\infty}\mathrm{d}t_2\int_2^{+\infty}\frac{p(t_1)\mathrm{d}t_1}{t_1^{m-1}}$. It effectively converts an *m*-step kinetic data set to a 1-

step kinetic data set $\mathcal{P}(t) = \int k e^{-kt} d\mu(k)$, allowing the "effective" rate constant distribution $\mu(k)$ to be stably reconstructed using the "phase function approach", a method we recently developed for stabilizing the inverse Laplace transform.[29] The breakdown of a unique multi-step reaction scheme is, however, usually an underdetermined problem with a single kinetic data set $p(t)$ in spite of a full determination of $\mu(k)$. As we have demonstrated in this paper, the "effective" rate constant distribution $\mu(k)$, when combined with additional input of kinetic information offers a powerful and straight forward strategy to fully dissect the m-step reaction kinetics $A_1 \xrightarrow[F_1(k)]{k_1} A_2 \xrightarrow[F_2(k)]{k_2} \cdots A_m \xrightarrow[F_m(k)]{k_m} B$.

Although we have presented our method in the context of sequential multi-step reactions with all irreversible steps, the same quantitative analysis applies to many non-sequential, partially reversible reactions as well. The classical Michaelis-Menten mechanism offers a telling example of how a partly reversible reaction is isomorphic to a sequential reaction with all irreversible steps. The reaction kinetics for

$$\mathrm{E+S} \underset{\kappa_{-1}}{\overset{\kappa_1[\mathrm{S}]}{\rightleftarrows}} \mathrm{ES} \xrightarrow{k_2} \mathrm{E+P}$$

is identical to the 2-step reaction $A_1 \xrightarrow{k_1'} A_2 \xrightarrow{k_2'} B$, where $k_1'+k_2'=\kappa_1[\mathrm{S}]+\kappa_2+\kappa_{-1}$ and $k_1'k_2'=\kappa_1\kappa_2[\mathrm{S}]$.[3,19] For the general kinetic correspondence between partly-reversible and totally-irreversible reaction schemes, we refer the readers to ref [33], where an algebraic approach was introduced for "reducing" complex reaction schemes (that involve reversible steps and loops) to more familiar paradigms.

In this paper, we mainly discussed the analysis of dwell time histograms that involve complex superpositions of exponential decays and introduced a model-independent method for analyzing these reaction kinetics. Many other statistical descriptions for kinetics in complex systems also manifest themselves as multi-exponential decays, especially in systems with a wide spectrum of time-scales. Notably, the decays of reactant concentration in bulk measurements (such as the drug metabolism and toxin degradation in pharmacokinetics[34]), the stochastic switching kinetics between molecular states (such as the random blinking of a quantum dot[35]), and the fluctuations of certain physiological responses (such as the polarization and depolarization of a nerve impulse[36]) all fall into this category. We anticipate that the method presented in this work would be helpful toward quantitative analysis of the complex kinetics of these processes. As we aimed at extracting maximum information from least assumptions, our model-independent method does not adapt itself to additional knowledge inputs automatically. In the case where such input is available (such as connectivity between multiple nodes in a reaction scheme[37,38]), our method may be complemented by other kinetics analysis approaches, such as the hidden Markov modeling[24,38] to provide a fuller picture of the reaction kinetics.

Supporting Information Available

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work is supported in part by the National Institutes of Health and the David and Lucile Packard Foundation. X. Z. is a Howard Hughes Medical Institute investigator.

References

1. Sykes, P. A Guidebook to Mechanism in Organic Chemistry. Vol. 6. Pearson Education; Harlow, England: 1986.
2. Petrucci, RH.; Harwood, WS.; Herring, FG. General Chemistry - Principles and Modern Applications. Vol. 8. Prentice Hall; Upper Saddle River, NJ: 2002.
3. Schnitzer MJ, Block SM. Cold Spring Harb Symp Quant Biol 1995;60:793. [PubMed: 8824454]
4. Nelson, DL.; Cox, MM. Lehninger Principles of Biochemistry. Vol. 3. Worth Publishers; Gordonsville, VA: 2000.
5. Qin F, Li L. Biophys J 2004;87:1657. [PubMed: 15345545]
6. Astumian RD. Appl Phys A 2002;75
7. Yang H, Luo G, Karnchanaphanurach P, Louie TM, Rech I, Cova S, Xun L, Xie XS. Science 2003;302:262. [PubMed: 14551431]
8. Zhuang X, Bartley LE, Babcock HP, Russell R, Ha T, Herschlag D, Chu S. Science 2000;288:2048. [PubMed: 10856219]
9. Tan E, Wilson TJ, Nahas MK, Clegg RM, Lilley DMJ, Ha T. Proc Nat Acad Sci USA 2003;100:9308. [PubMed: 12883002]
10. Liu S, Bokinsky GE, Walter NG, Zhuang X. Proc Nat Acad Sci USA 2007;104:12634. [PubMed: 17496145]
11. Lu HP, Xun L, Xie XS. Science 1998;282:1877. [PubMed: 9836635]
12. Zhuang X, Kim H, Pereira MJB, Babcock HP, Walter NG, Chu S. Science 2002;296:1473. [PubMed: 12029135]
13. Hong MK, Harbron EJ, O'Connor DB, Guo J, Barbara PF, Levin JG, Musier-Forsyth K. J Mol Biol 2003;325:1. [PubMed: 12473448]
14. Yildiz A, Forkey JN, McKinney SA, Ha T, Goldman YE, Selvin PR. Science 2003;300:2061. [PubMed: 12791999]
15. Shaevitz JW, Abbondanzieri EA, Landick R, Block SM. Nature 2003;426:684. [PubMed: 14634670]
16. Svoboda K, Mitra PP, Block SM. Proc Nat Acad Sci USA 1994;91:11782. [PubMed: 7991536]
17. Schnitzer MJ, Block SM. Nature 1997;388:386. [PubMed: 9237757]
18. Xie S. Single Mol 2001;2:229.
19. English BP, Min W, van Oijen AM, Lee KT, Luo G, Sun H, Cherayil BJ, Kou SC, Xie XS. Nat Chem Bio 2006;2:87. [PubMed: 16415859]
20. Cai L, Friedman N, Xie XS. Nature 2006;440:358. [PubMed: 16541077]
21. Friedman N, Cai L, Xie XS. Phys Rev Lett 2006;97:168302. [PubMed: 17155441]
22. McWhirter JG, Pike ER. J Phys A: Math Gen 1978;11:1729.
23. Venkataramanan L, Sigworth FJ. Biophys J 2002;82:1930. [PubMed: 11916851]
24. McKinney SA, Joo C, Ha T. Biophys J 2006;91:1941. [PubMed: 16766620]
25. Tikhonov, AN.; Arsenin, VY. Solution of Ill-posed Problems. John Wiley and Sons; New York, NY: 1977.
26. Provencher SW. Comput Phys Comm 1982;27:213.
27. Livesey AK, Brochon JC. Biophys J 1987;52:693.
28. Steinbach PJ, Ionescu R, Matthews CR. Biophys J 2002;82:2244. [PubMed: 11916879]
29. Zhou Y, Zhuang X. Biophys J 2006;91:4045. [PubMed: 16980370]
30. Istratov AA, Vyvenko OF. Rev Sci Instr 1999;70:1233.
31. Yildiz A, Tomishige M, Vale RD, Selvin PR. Science 2004;303:676. [PubMed: 14684828]
32. Dohoney KM, Gelles J. Nature 2001;409:370. [PubMed: 11201749]
33. Shaevitz JW, Block SM, Schnitzer MJ. Biophys J 2005;89:2277. [PubMed: 16040752]
34. Lunn, DJ. Bayesian Analysis of Population Pharmacokinetic/Pharmacodynamic Models. In: Husmeier, D.; Dybowski, R.; Roberts, S., editors. Probabilistic Modeling in Bioinformatics and Medical Informatics. Springer-Verlag; London: 2005.
35. Verberk R, van Oijen AM, Orrit M. Phys Rev B 2002;66:233202.

36. Hodgkin AL, Huxley AF. J Physiol 1952;116:449. [PubMed: 14946713]
37. Flomenbom O, Klafter J, Szabo A. Biophys J 2005;88:3780. [PubMed: 15764653]
38. Andrec M, Levy RM, Talaga DS. J Phys Chem A 2003;107

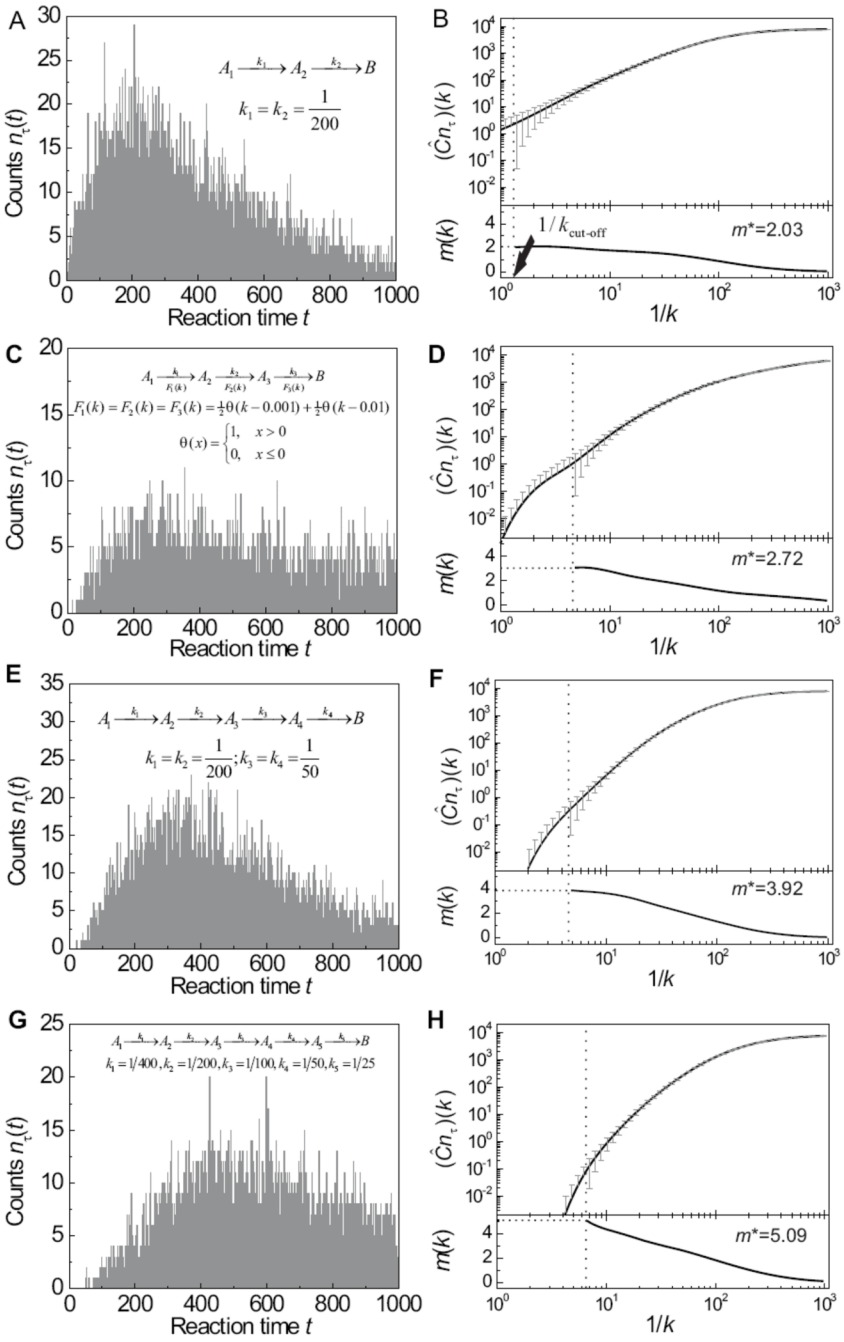

Figure 1.

The "heat capacity transform" $\hat{C}$ determines the number of reaction steps. Multi-step processes (m = 2, 3, 4, and 5) with various reaction schemes are analyzed by the transform $\hat{C}$ (eq 11). Panels *a*, *c*, *e* and *g* show Monte Carlo simulations of the reaction time histograms $n_\tau(t)$ from multi-step reactions (schematically shown in insets). The time resolution is set to $\tau = 1$. The total number of events simulated is 8000. Panels *b*, *d*, *f* and *h* show numerical analyses of the histograms in *a*, *c*, *e* and *g*, respectively. The upper panels show the numerical computation of the "heat capacity transform" $(\hat{C}n_\tau)(k)$ with vertical error bars $\delta(\hat{C}n_\tau)(k)$ estimated by eq 14. Lower panel plots $m(k)$, where $m(k)$ is the estimate of the derivative $\mathrm{d}\ln(\hat{C}n_\tau)(k)/\mathrm{d}\ln(1/k)$ using eq 24. The limit $m^* = \lim_{k\to k_{\max}} m(k)$ ($k_{\max}$ is determined from eq 25 and marked by the dashed

vertical line) provides quantitatively good estimates the number of reactions steps in all five cases.

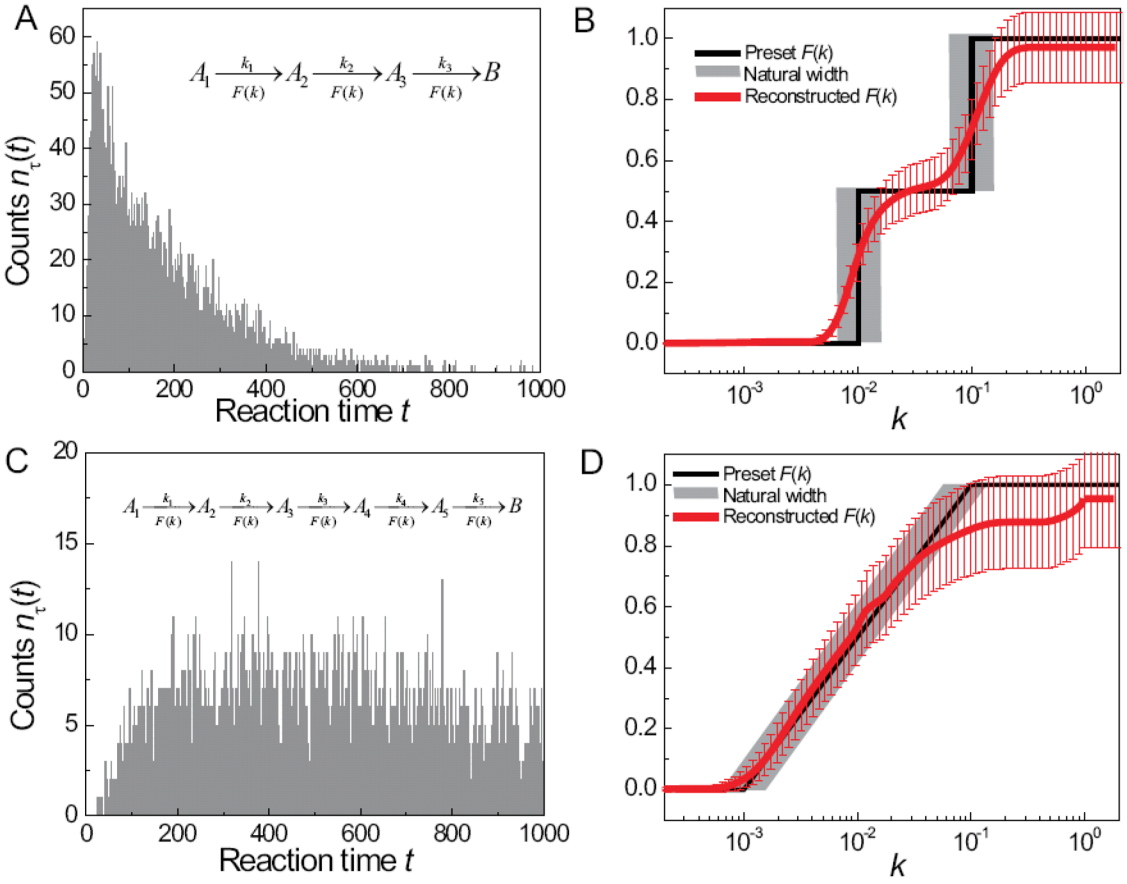

Figure 2.

Analysis of multiple-step reaction kinetics that involve identically distributed rate constants in each step. Panels *a* and *c* show the simulated reaction time distributions $n_\tau(t)$ with the reaction schemes shown in insets. The total number of counts in each histogram is 8000. The preset rate constant distributions $F(k)$ used to simulate the $n_\tau(t)$ in panels *a* and *c* are shown as black curves in panels *b* and *d*, respectively. The rate constant distributions $F(k)$ reconstructed from $n_\tau(t)$ using the phase function approach are shown as red curves. The reconstructions of $F(k)$ are achieved by computing $F(k) = (\hat{R}\varphi)(k)$ using eq 18, where the phase function $\varphi(ik) = \psi(ik)/m$ is determined from $\mu(k)$, which in turn is deduced from the corresponding *t*-domain kinetic data $n_\tau(t)$ in panel *a* and *c*. The vertical error bars are estimated by eq 19. The uncertainties in the horizontal direction ("natural width" due to finite sampling effect) are marked as grey zones in panels *b* and *d*.

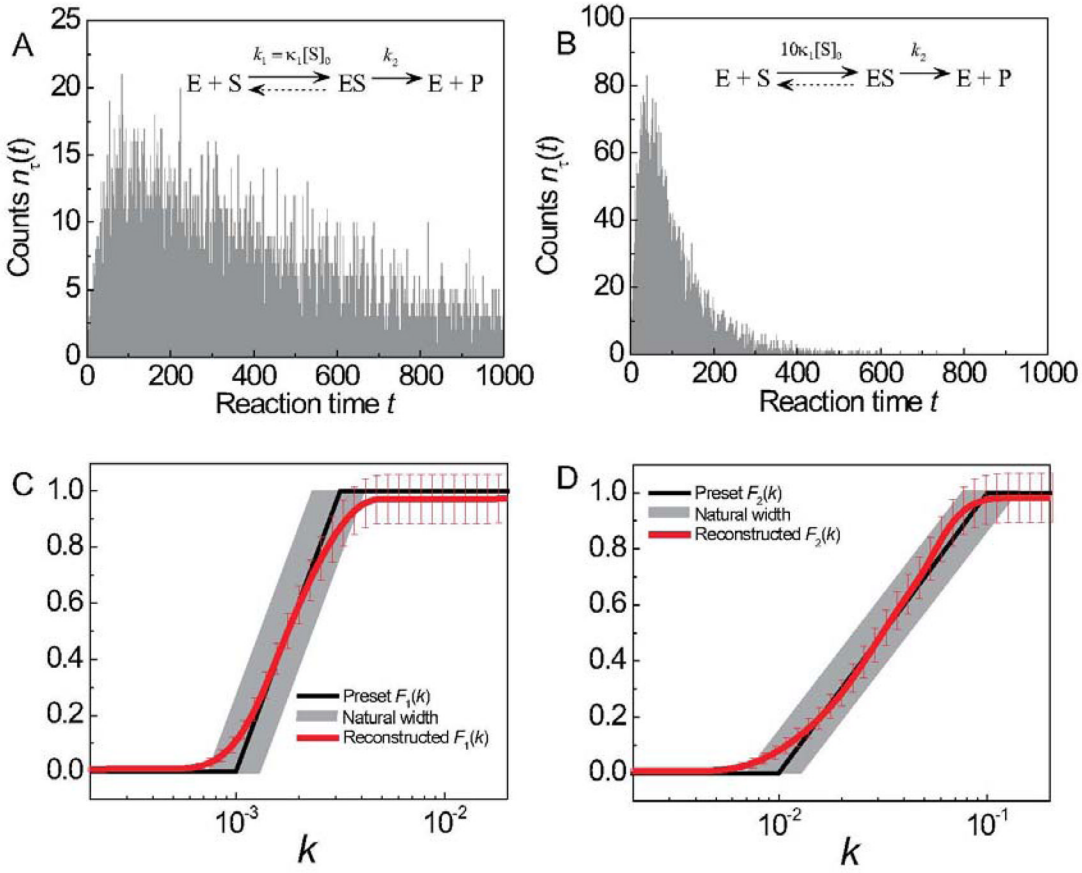

Figure 3.

Analysis of Michaelis-Menten enzymatic kinetics with dispersed kinetic rate constants. Panels *a* and *b* show the simulated data for 2-step enzymatic reactions $n_\tau(t)$ with two different substrate concentrations $[S]_0$ and $10[S]_0$. The kinetic data $n_\tau(t)$ were simulated using the preset rate constant distributions $F_1(k)$ and $F_2(k)$ shown as black curves in panels *c* and *d*. Reconstructed rate constant distributions shown in red curves compare quantitatively with the preset distributions within error. The vertical error bars were estimated by eq 19 and the horizontal errors bars marked in grey zones arise from finite sampling effect. The total number of counts per substrate concentration is 8000.