# Random-walk enzymes

Chi H. Mak, Phuong Pham, Samir A. Afif, and Myron F. Goodman

Phys. Rev. E **92**, 032717 — Published 17 September 2015

DOI: 10.1103/PhysRevE.92.032717

# Random-Walk Enzymes

Chi H. Mak[1,2], Phuong Pham[3], Samir A. Afif[3], Myron F. Goodman[1,3]
[1]*Department of Chemistry,* [2]*Center for Applied Mathematical Sciences and* [3]*Department of Biological Sciences,*
*University of Southern California, Los Angeles, California 90089, USA*

(Dated: September 3, 2015)

Enzymes that rely on random walk to search for substrate targets in a heterogeneously dispersed medium can leave behind complex spatial profiles of their catalyzed conversions. The catalytic signatures of these random-walk enzymes are the result of two coupled stochastic processes –scanning and catalysis. Here we develop analytical models to understand the conversion profiles produced by these enzymes, comparing an *intrusive* model, in which scanning and catalysis are tightly coupled, against a loosely-coupled *passive* model. Diagrammatic theory and path integral solutions of these models revealed clearly distinct predictions. Comparison to experimental data from catalyzed deaminations deposited on single-stranded DNA by the enzyme activation-induced deoxycytidine deaminase (AID) demonstrates that catalysis and diffusion are strongly intertwined, where the chemical conversions give rise to new stochastic trajectories that were absent if the substrate DNA was homogeneous. The C→U deamination profiles in both analytical predictions and experiments exhibit a strong contextual dependence, where the conversion rate of each target site is strongly contingent on the identities of other surrounding targets, with the intrusive model showing an excellent fit to the data. These methods can be applied to deduce sequence-dependent catalytic signatures of other DNA modification enzymes, with potential applications to cancer, gene regulation and epigenetics.

PACS numbers: 02.50.-r, 05.40.-a, 05.20.-y, 87.10.-e

Enzymes catalyze highly specific chemical transformations on their substrates. In the cell the substrates targeted by a particular enzyme are typically distributed within an inhomogeneous medium. To seek out their targets, enzymes must diffuse through this matrix to find them. Enzymatic reactions with high intrinsic turnover rates are often diffusion-limited [1, 2], where the probability of the chance encounter between enzyme and substrate controls the reaction rate. Certain enzymes, such as the endonucleases, and DNA binding proteins (e.g. lac repressor), operate with higher target location efficiency than random diffusion predicts, and models of facilitated diffusion have been advanced to explain their more rapid targeting [3–10].

While diffusion-controlled enzymatic reactions with high intrinsic turnover rates have received a great deal of attention, the question of how the diffusion of a low- or moderate-efficiency enzyme affects catalytic conversions on spatially-dispersed substrate targets has not been solved. Any agent, chemical or otherwise, that can catalyze conversions on multiple targets distributed in the underlying space it is scanning can leave behind complex spatial signatures of both its diffusive and catalytic dynamics. Our models demonstrate how the seemingly random conversions produced by such an enzyme can be guided by the interplay between its catalytic activities and motions. Permuting the targets or simply rearranging their positions can drastically alter the random outcomes. This has important implications for a number of systems.

For instance, activation-induced deoxycytidine deaminase (AID) [11] is responsible for initiating antibody diversification in B-cells, by deaminating C→U in a scanning-coupled catalytic reaction [12, 13] favoring trinucleotide WRC target motifs (W=A/T, R=A/G) [14]. This produces hypermutation in the Ig variable and switch regions, which are critical for the fitness of the immune system [15–17]. Yet even when acting on its most highly favored AAC motif, the range of catalytic efficiency is remarkably low, $\sim 1 - 7\%$ [13]. Seemingly, the combination of stochastic and inefficient catalysis has evolved to provide a highly efficient way to ensure optimal Ab diversity. Cancer genomes, on the other hand, often contain clustered mutations termed "kataegis" that are thought to be produced by AID/APOBEC dC deaminases via similar scanning-coupled enzymatic reactions [18–22]. Notably, AID, Apo3A and Apo3B appear to cause "off-target" mutations in proto-oncogenes implicated in B-cell lymphoma [23–26], breast and other cancers [18–22], which typically occur in regions of ssDNA generated perhaps during aberrant DNA replication and repair. It is here that the models have the potential to make significant impact in mutationally based disease, since the DNA sequence exerts a major influence on where mutations occur. Analogous coupled stochastic processes are also found in epigenetics where DNA methyltransferases [27] imprint methylated CpG islands at DNA sequences at or near transcription sites of genes to exert control over their expression, and in base excision repair of endogenous and exogenous DNA damage by a variety of DNA glycosylases [28, 29]. Since there are no *a priori* restrictions on nucleic acid sequence, our model applied to AID is similarly applicable to identifying coupled scanning-catalysis mechanisms for these enzymes.

In this article, we formulate two general analytical models for scanning-coupled catalysis to investigate the coupling between enzyme diffusion and catalysis. Using spatial mutational patterns measured experimentally, we deduce sequence context effects on catalysis. Employing a standard Kolmogorov equation to couple the kinetics of catalysis into the scanning motions of the enzyme, we arrive first at a "passive" model. By simple analytical arguments, we show that in this passive picture catalysis does not materially modify the statistics of the diffusion paths even though they are coupled. A more interesting alternative, an "intrusive" model, can be constructed from a path integral picture in which the catalytic action of the enzyme produces new composite paths reflecting coupled scanning-catalytic trajectories absent from the passive model. Both models have zero adjustable parameters and employ the same inputs, all of which can be determined from independent experiments. These models represent two alternative views of how catalysis might alter the scanning dynamics of an enzyme, the fundamental distinction between them resting essentially on how the paths are counted. The mathematical solutions to these models show that close coupling between scanning and catalysis generates intricate spatial relationships in the locations of the catalyzed changes in the intrusive model, and the observed catalytic efficiencies can exhibit complex contextual dependencies where the conversion of one substrate is controlled not only by its own susceptibility to catalysis, e.g., WRC hot motifs, but also by surrounding non-hot-motifs or by any DNA sequence with or without C. The passive model, on the other hand, shows no contextual dependence. More generally, we exploit a formal isomorphism between the models and quantum mechanics to interpret their predictions. Though scanning-coupled enzyme systems are classical, this quantum isomorphism suggests that the enzyme interrogates its target sites by repeatedly applying a position measurement, causing interruptions to its scanning paths. The outcomes of these effects are contingent on how the targets are arranged in space. Our results suggest that biological systems could potentially exploit contextual effects to guide the catalytic actions of random walk enzymes, which for AID could facilitate mutations in Ig variable regions that determine antibody-antigen recognition. As a practical application to biological systems, our analysis can be used to identify distinctive spatial genomic modification signatures arising from inadvertent catalysis by random-walk enzymes implicated in cancer.

## I. SCANNING, CATALYSIS, AND THE PASSIVE MODEL

When occurring separately, scanning and catalysis are described by well-known models. The scanning motions define a continuous-time random walk. Let the substrate targets $i \in \{1, 2, \cdots N\}$ be located at fixed positions $\{\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_N\}$ within the space that is being scanned by the catalyst, and each target $i$ has a different intrinsic catalytic rate $u_i$. The details of how the enzyme diffuses among the target sites can be encapsulated into a generator matrix $\mathbf{W}$, where its element $W_{ij}$ describes the transition rate of the enzyme moving from site $j$ to $i$. The transition matrix $\mathbf{W}$ allows for non-nearest-neighbor hops, and it generates a Kolmogorov equation for the scanning motions

$$\frac{dp_i}{dt} = \sum_j W_{ij} p_j \tag{1}$$

for $p_i(t)$, the time-dependent probability of finding the enzyme on target $i$. In the absence of catalysis, the solution of the Kolmogorov equation is given by the propagator matrix $\mathbf{K}_0(t) = \exp(t\mathbf{W})$, whose element $[K_0(t_2-t_1)]_{ij}$ specifies the conditional probability of finding the enzyme on target $i$ at time $t_2$ given it was on $j$ at time $t_1 < t_2$.

The Kolmogorov equation is a phenomenological equation of motion for the scanning propagator matrix $\mathbf{K}_0(t)$. The same propagator $[K_0(t_2-t_1)]_{ij}$ can be derived from a path integral over all possible stochastic trajectories $q(t)$ taken by the enzyme to diffuse from site $j$ to $i$ between time $t_1$ and $t_2$ [30],

$$[K_0(t_2 - t_1)]_{ij} = \int_j^i \mathcal{D}q(t)\mathcal{P}_0[q(t)], \tag{2}$$

where $q \in \{1, 2, \cdots N\}$, and $q(t)$ is subject to the boundary conditions $q(t_1) = j$ and $q(t_2) = i$. Along the scanning path $q(t)$, the enzyme makes transitions from one site to another. Each time a transition occurs between sites $k$ and $l$, the functional $\mathcal{P}_0$ picks up a contribution $\delta_{kl} + W_{kl}dt + o(dt)$, where $o$ is Landau's symbol and $\delta$ is Krnoecker's delta. The path integral in Eq.(2) sums over all possible transitions and over all transition times. Formally, the scanning path integral in Eq.(2) is isomorphic to the imaginary-time path integral for the Boltzmann operator [31] $\mathbf{K}_0(t) = \exp[-\beta\mathbf{H}_0]$ of a quantum particle with Hamiltonian $\mathbf{H}_0 = -\mathbf{W}$, under the mapping $t$ to the inverse temperature $\beta$ (in units where $\hbar = 1$). Under this isomorphism, the Kolmogorov equation Eq.(1) is also equivalent to the imaginary-time Schrödinger equation [31].

While accounting for the enzyme's scanning motions is simple via diffusion paths, how to incorporate catalytic rates into the stochastic model is not as clear. In ordinary diffusion-reaction systems [32, 33], the species that is diffusing is also the reactant. This naturally gives rise to source or sink terms in its diffusion equation. But in the type of scanning-coupled catalytic systems describing random-walk enzymes, it is the catalyst that is executing the random walk while the reactions are occurring on immobile targets. Since the concentration of the catalyst itself does not change with time, there is no *a priori* reason to include the reactions into the diffusion propagator. In fact, there is no obvious mechanism for how the chemical conversions should impact the enzyme's scanning paths at all.

The kinetics of enzyme catalysis are typically described by Poisson statistics [33, 34]. As the enzyme randomly scans its target sites, catalytic conversions occur as a secondary stochastic Poisson process along each scanning path $q(t)$, where the intrinsic catalytic susceptibility $u_i$ at each target $i$ may be site-dependent. Using a spin-1/2 variable $m_i \in \{-\frac{1}{2}, +\frac{1}{2}\}$ to denote the chemical state of each site $i$ where $-\frac{1}{2}$ represents an unconverted target and $+\frac{1}{2}$ a converted one, the state of the system can be specified by a vector $\mathbf{x} = (q, m_1, m_2, \ldots, m_N)$, where $q$ gives the location of the enzyme and $m_i$ the current chemical state of each site $i$. Using a straightforward extension of the scanning model in Eq.(1), we can incorporate catalytic activities into a scanning-mutation Kolmogorov equation to describe the time-evolution of the probability $P_{\mathbf{x}}(t)$ of each state $\mathbf{x}$:

$$\frac{dP_{\mathbf{x}}}{dt} = \sum_{\mathbf{x}'} \Omega_{\mathbf{x}\mathbf{x}'} P_{\mathbf{x}'}. \tag{3}$$

The transition matrix elements $\Omega_{\mathbf{x}\mathbf{x}'}$ in Eq.(3) between states $\mathbf{x}' = (q', m'_1, m'_2, \ldots, m'_N)$ and $\mathbf{x} = (q, m_1, m_2, \ldots, m_N)$ are given for $\mathbf{x} \neq \mathbf{x}'$ by:

$$\Omega_{\mathbf{x}\mathbf{x}'} = \begin{cases} W_{ij}, & \text{if } q = i \neq j, q' = j \text{ and } m_k = m'_k \ \forall\, k, \\ u_i, & \text{if } q = q' = i, m_i = m'_i + 1 \text{ and} \\ & \qquad m_k = m'_k \ \forall\, k \neq i, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

and $\Omega_{\mathbf{x}\mathbf{x}} \equiv -\sum_{\mathbf{x}' \neq \mathbf{x}} \Omega_{\mathbf{x}\mathbf{x}'}$. The first condition in Eq.(4) describes transitions due to scanning, where $W_{ij}$ are the same scanning transition rates in the scanning-only model Eq.(1). The second describes transitions due to catalytic conversions, which can occur on site $i$ with rate $u_i$ only when the enzyme is also at $q = i$. Initially all targets start in the $-\frac{1}{2}$ state. A target may be converted only when the enzyme visits it and may be converted no more than once even if the enzyme revisits it.

While scanning and catalysis are manifestly coupled through the transition matrix elements Eq.(4), it is easy to show that using these transition rules in Eq.(3) does not materially alter the statistics of the scanning trajectories compared to those in a scanning-only system described by Eq.(1). We can deduce this by reducing Eq.(3), grouping states $\mathbf{x}$ into sets according to the location of the enzyme $q$. Let $\tilde{i} = \{\mathbf{x} : q = i\}$ be the set of all states in which the enzyme is at position $q = i$, regardless of the current chemical state of the targets. The sum $\tilde{P}_i \equiv \sum_{\mathbf{x} \in \tilde{i}} P_{\mathbf{x}}$ then denotes the total probability over this set. Using this reduction, Eq.(3) can be re-expressed as:

$$\frac{d\tilde{P}_i}{dt} = \sum_j W_{ij} \tilde{P}_j. \tag{5}$$

This reduction demonstrates that the composite probability $\tilde{P}_i$ over each set $\tilde{i}$ evolves in time solely under the influence of scanning alone. This is true because while there are interconversions among states within each set $\tilde{i}$ due to the conversion rates $u_i$, $\tilde{P}_i$ within each set is conserved under catalysis. Only scanning modifies the composite $\tilde{P}_i$. Equation (5) is identical to the scanning-only Kolmogorov equation (1). According to the phenomenological equation of motion Eq.(3), the scanning motions of the enzyme are therefore decoupled from its enzymatic activities.

We will refer to the Kolmogorov equation (3) as the *passive model* for scanning-coupled catalytic processes. The complete decoupling of scanning dynamics from catalysis in the passive model is significant in several ways. First, the phenomenological stochastic model Eq.(3) suggests that the scanning paths are not perturbed in any way by the chemical activities of the enzyme. Second, as long as the scanning transition matrix $\mathbf{W}$ is isotropic, a large ensemble of the enzyme's scanning paths should cover the entire target space uniformly, simply because the initial binding site at the beginning of each path is random. Third, and most importantly, since the scanning paths access the target space uniformly, the observed conversion probability of each target site $i$ should be a function only of its intrinsic susceptibility $u_i$ and completely independent of its neighbors'.

The catalytic signature of a random-walk enzyme operating under the passive model is trivial. The characteristics of the random-walk have no bearing on the observed conversion probabilities of the targets as long as the scanning paths cover the target space uniformly under the action of $\mathbf{W}$. The target conversion profile in the passive model has no contextual dependence at all.

## II. THE INTRUSIVE MODEL

In the passive model, the mutations are purely ancillary – they do not modify the random walk of the enzyme and they occur stochastically along the same diffusion paths the enzyme would have taken if there were no mutations. This perspective seems intuitive, because scanning is random and the enzyme would not have known whether a target site is mutable until it has actually reached it. If this passive picture is correct, scanning should completely decouple from catalysis and the problem becomes trivially solvable.

To formulate an alternate perspective, we turn to the path integral picture in Eq.(2). We consider composite scanning-mutation paths in which both catalysis and scanning characterize the overall time evolution of the system and use diagrams to deduce the probability of each path. The mutation variables $\{m_i\}$, which specify the chemical states of the targets, are denoted by the vector $\mathbf{m}$, and their time evolution is described by a path $\mathbf{m}(t)$ while $q(t)$ describes the scanning path of the enzyme. The composite path is therefore $\mathbf{x}(t) = (q(t), \mathbf{m}(t))$.

The paths that make up the intrusive model are illustrated diagrammatically in Fig. 1 in the form of a perturbation series. For notational simplicity, a single variable $m(t) = \sum_i m_i(t)$ is used to describe the mutations. Every time a conversion is made $m(t)$ is incremented by 1, and using the instantaneous position of the enzyme

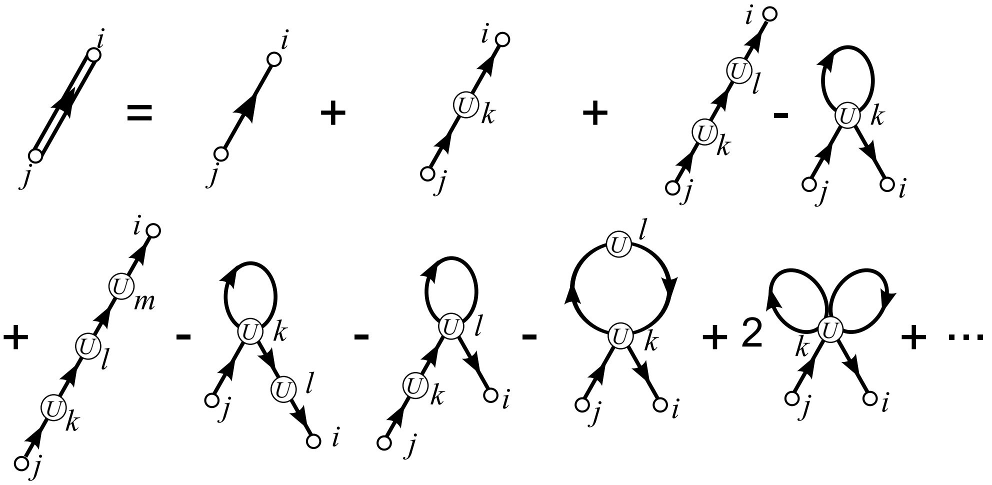

FIG. 1. Diagrammatic representation of the perturbation series for scanning-coupled catalysis. Single lines represent scanning trajectories. Circles labeled U represent mutation events. Loop diagrams correct for paths with multiple mutations on the same motif. The full propagator, represented by the double line, is the sum over all possible paths.

the location of the converted target can be ascertained. From the perspective of the intrusive picture, two composite paths $(q(t), m(t))$ and $(q'(t), m'(t))$ are considered distinct if $m(t) \neq m'(t)$ even if $q(t) = q'(t)$. To sum all possible scanning and mutation trajectories that the system may take from time 0 to $t$, we turn on the mutation rates and count the additional paths being generated.

The first diagram on the right of the equality, the zero-order term, represents all scanning paths with no mutations, from target $j$ to $i$ over time $t$, which is just $\mathbf{K}_0(t) = \exp(t\mathbf{W})$, shown as a single arrow in Fig. 1. The next diagram depicts propagation from site $j$ to $i$ at time $t_1$ with the propagator $\mathbf{K}_0(t_1)$, where a mutation occurs with probability $u_k dt_1$, and $m(t)$ is incremented by 1 at $t_1$. This is followed by propagation from $k$ to $i$ to the final time $t$ with the propagator $\mathbf{K}_0(t - t_1)$. This first-order term accounts for the additional paths that are spawned as the result of one mutation event during the scanning trajectory. The third diagram represents second-order paths with two mutations on target $k$ at time $t_1$ and then $l$ at $t_2$, with arrows representing the bare propagators. Since mutations cannot occur on the same target twice, the fourth diagram represents second-order terms with $k = l$ which must be removed from the path sum, denoted by a minus sign (see Appendix A for details). The second row in Fig. 1 represents all third-order diagrams. All higher-order diagrams (not shown) are constructed in the same way. For each, a time-ordered integration over the intermediates times $t_1 < t_2 < \cdots$ as well as a sum over all intermediate targets $k, l, \ldots$ are required. The white circles at the termini of each diagram indicate that a sum over $i$ and $j$ are also needed since the starting and ending positions of the enzyme should be uniformly distributed across all the targets. The sum over all diagrams then yields the full propagator $\mathbf{K}(t)$ shown as the double line on the left (Fig. 1).

The perturbation approach treats the problem as a branching process, where the mutation events spawn new paths on top of the scanning trajectories. Notice that the interaction between scanning and catalysis modifies neither the intrinsic scanning transition matrix $\mathbf{W}$ nor the intrinsic mutation rates $\{u_i\}$. The same $\mathbf{W}$ and $\{u_i\}$ are also assumed in the passive picture. The only difference between the two models is in how the paths are counted. Appendix A provides more mathematical details on the diagrams and the paths.

Perturbation series like those in Fig. 1 are familiar in quantum and statistical physics. Accurate approximations to the series can be developed by selecting appropriate partial sums. The series can alternatively be solved by rearranging the diagrams to yield an integral equation [34]. If there are no non-Markovian effects, diagrams can often be summed by Laplace transforms or by diagonalization [35].

If the catalytic rate is not too high and the duration of the paths is not too long, repeat conversion attempts on any target should be rare, and a reasonable approximation is to ignore all loop diagrams in Fig. 1. Resumming the remaining terms then yields the Markovian approximation $\mathbf{K}_1(t) = \exp[t(\mathbf{W} + \mathbf{U})]$, where $\mathbf{U}$ is a diagonal matrix with elements $U_{ii} = u_i$, containing the site-dependent intrinsic mutation rates. This approximate propagator $\mathbf{K}_1(t) = \exp[t(\mathbf{W} + \mathbf{U})]$ turns out to be isomorphic to the Boltzmann operator $\exp[-\beta\mathbf{H}_1]$ for a quantum particle on a discrete lattice under the mapping $t$ to the inverse temperature $\beta$ (in units where $\hbar = 1$), and $-(\mathbf{W} + \mathbf{U})$ to the Hamiltonian $\mathbf{H}_1$. Isomorphisms of this type are well-known [31, 36], and for scanning-coupled catalysis, $-\mathbf{W}$ maps to the translational Hamiltonian of the quantum particle and $-\mathbf{U}$ to the potential energy. Exploiting the quantum analogy, the propagator $\mathbf{K}_1(t)$ and all observables can be computed analytically by using the eigenfunctions and eigenvalues from the solution of the Schrödinger equation $\mathbf{H}_1\Psi = E\Psi$. (See

Appendix B for details.)

Notice that the only inputs into the intrusive model, $\mathbf{W}$ and $\{u_i\}$, are the same used for the passive model via the transition rates defined by Eq.(4). However, the outcomes of the intrusive model are markedly different. In the passive model, Eq.(5) suggests that the mutations do not materially modify the statistics of the scanning paths compared to a scanning-only system. In the intrusive model, the mutations when considered together with the scanning trajectories create new paths that the passive model did not consider as distinct. While in the passive model the probabilities of the scanning trajectories are unaltered regardless of what mutation events might occur along them, the intrusive model counts a scanning path $q(t)$ that is associated with a mutation path $m(t)$ as distinct from another composite path $(q(t), m'(t))$ where $m(t) \neq m'(t)$.

## III. COMPARING INTRUSIVE AND PASSIVE MODEL PREDICTIONS WITH AID MUTATIONAL PATTERNS

To demonstrate the salient features of the passive versus intrusive pictures and to make contact with experiments, we apply the solution $\mathbf{K}_1(t)$ to AID-catalyzed C→U transitions observed in ssDNA mutant libraries. The quantities calculated are the expected observed mutation probability of each target, which are obtained from the propagator by solving the Schrödinger equation. Mathematical details are given in Appendix B.

The enzyme AID binds and scans ssDNA sequences processively [13, 14], deaminating C nucleotides to U preferentially at trinucleotide WR$\underline{\text{C}}$ hot motifs (W=A/T, R=A/G) over SY$\underline{\text{C}}$ cold motifs (S=C/G, Y=C/T) [14]. The mutations are random, but in a large library of mutant clones we can measure the mutation probability on each motif. Varying the sequence and composition of the hot/cold motifs on the DNA allows us to study how scanning and catalysis interact with each other. Modifying the DNA sequence does not significantly alter the intrinsic scanning transition matrix $\mathbf{W}$ for AID, which has been measured experimentally for a homogeneous ssDNA substrate [12]. By inserting different DNA sequences or “cassettes” into the substrate, we can rigorously test the analytical models.

Both the intrusive and the passive models are parameter-free. The only inputs are the $\mathbf{W}$ matrix and the intrinsic mutation susceptibilities $\{u_i\}$, and these can be independently and experimentally determined using homogeneous substrates. Predictions from the two models are illustrated in Fig. 2 for a test sequence designed to provide direct comparison with experiments. This test cassette consists of 63 motifs. On the left side are 30 alternating hot (AA$\underline{\text{C}}$) and hot′ (AG$\underline{\text{C}}$) motifs. On the right are 30 alternating hot (AA$\underline{\text{C}}$) and cold (GT$\underline{\text{C}}$) motifs, with a 3-silent-motif spacer between them. This 63-motif cassette is embedded between two extended silent

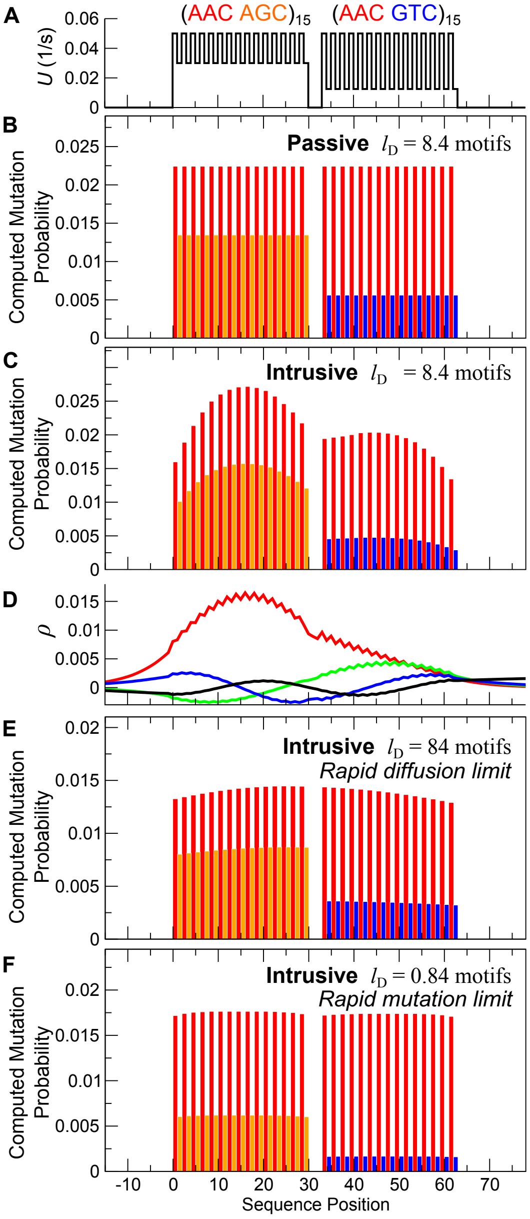

FIG. 2. Computed mutation probabilities on a hot-hot′/hot-cold test cassette. (A) Intrinsic mutations rates $u_i$ on the 63-motif hot-hot′/hot-cold cassette for AID-catalyzed C→U mutations on ssDNA. (B) Mutation probabilities predicted by the passive model as a function of sequence position are directly proportional to the intrinsic mutation rates. (C) Mutation probabilities predicted by the intrusive model shows contextual dependence, where the computed mutation probability of a site is influenced by surrounding motifs. (D) Probability $\rho_i$ of finding the enzyme on the cassette corresponding to C, decomposed into contributions from the four lowest eigenfunctions in the order red, green, blue and black. (E) Mutation probabilities predicted by the intrusive picture in the rapid diffusion limit where the system reduces to the passive picture. (F) Mutation probabilities predicted by the intrusive picture in the slow diffusion limit, which also reduces to the passive picture.

sequences on both the 5′ and 3′ ends. Figure 2A shows the site-by-site mutation susceptibilities $u_i$ along this test sequence, which define the $\mathbf{U}$ matrix used in our calculations. Experiments, described below, have been carried out on the same cassette sequence. The predicted mutation profiles in Figs. 2 B, C, E and F were computed for $t$ = 45s, corresponding to the length of these experiments.

Figure 2B shows the mutation probability of each motif on this hot-hot′/hot-cold cassette as determined using the passive model. In the passive model, the expected mutation probability is simply related to the intrinsic catalytic susceptibility of each site individually, which is clear comparing Figs. 2B and A. On the other hand, Fig. 2C, which shows contrasting predictions from the intrusive model, reveals more complex behaviors. In the intrusive picture, mutation probabilities computed using the approximate propagator $\mathbf{K}_1$ from the perturbation series (see Appendix B for computational details) are not straightforwardly related to the intrinsic mutation susceptibilities. Instead, they exhibit a nontrivial contextual dependence where the observed hit rate of each site appears to be contingent on the identities of the surrounding sites. In particular, the mutation probabilities are lower on the left and right edges of the cassette closest to the silent motifs on the two ends. The hot motifs (red) are predicted to be converted with a higher observed hit rate when they are among other hot′ motifs (orange) but lower when they are among cold ones (blue). The intrusive model therefore suggests that the sequence can exert substantial control over the mutation rate of individual target sites on the DNA. Figures 2D, E and F, respectively, show the probability of finding the enzyme on this cassette projected onto its eigenfunctions as well as predictions from the intrusive model in two separate limits, rapid diffusion or rapid mutation, and these will be discussed below.

The corresponding experiments are summarized in Fig. 3. Libraries of mutant clones on a number of *inhomogeneous* ssDNA sequences with mixed motifs were analyzed. Figure 3A shows the experimental setup (see Appendix F for experimental details), with the previously reported [12] scanning transition matrix elements $W_{ij}$ plotted in Fig. 3B as a function of distance $i-j$. This $\mathbf{W}$ [12] was used as input to our calculations along with the average intrinsic AID-catalyzed deamination rates measured for each trinucleotide motif [14]. These inputs were defined using independent experiments on homogeneous substrates [12] and the models have zero adjustable parameters.

Figure 3 shows experimentally determined and calculated mutation signatures on two different ssDNA cassettes. Results shown on the left panels in Fig. 3 (C, D and E) correspond to the sequence $(\mathrm{AA\underline{C}\ AG\underline{C}})_{15}$-sss-$(\mathrm{AG\underline{C}\ GT\underline{C}})_{15}$, where sss is a 9-nt spacer, flanked by two extended sequences of silent motifs on the 5′ and 3′ ends. The intrinsic deamination rates along this sequence, shown in Fig. 3C, have ratios of roughly 5:3:1 for $\mathrm{AA\underline{C}{:}AG\underline{C}{:}GT\underline{C}}$ (this cassette is the same as the one in Fig. 2). The observed mutation probabilities shown in Fig. 3D from a batch of 814 clones are compared against predictions from the intrusive model shown in Fig. 3E. The experiments clearly exhibit similar contextual effects as the intrusive model predicts. Not plotted explicitly in Fig. 3 are full predictions from the passive model. In the passive model, the mutational probabilities are simply proportional to the intrinsic site-by-site catalytic susceptibilities $u_i$, resulting in a mutational profile that would have the same appearance as Fig. 3C. Some site-to-site variations in the observed target conversion probabilities are due to statistics related to sample sizes. The experimental uncertainties in the mutation are approximately square root of the observed counts, typically $< 5$. While the predicted spectra are smooth, the observed ones are noisy, but the agreement between experiment and predictions are quantifiably significant, as we will show below.

Results for a second ssDNA test sequence are shown on the right panels in Fig. 3 (F, G and H). This consists of a $(\mathrm{AA\underline{C}\ AG\underline{C}})_{15}$-sss-$(\mathrm{AA\underline{C}\ GA\underline{C}})_{15}$-sss-$(\mathrm{AA\underline{C}\ GA\underline{C}})_{15}$ cassette flanked by two silent sequences. The intrinsic deamination rates of $\mathrm{AA\underline{C}{:}AG\underline{C}{:}GA\underline{C}}$(shown in Fig. 3F) are roughly 5:3:0.5, and this cassette corresponds to a mixed sequence of hot-hot′ motifs on the left and alternating hot-frigid motifs in both the center and the right of the substrate. The experimentally observed mutation profile is shown in Fig. 3G, with the intrusive model prediction in Fig. 3H. Again, passive model results, which should be identical in appearance to the intrinsic catalytic susceptibilities for this cassette (Fig. 3F), are not shown explicitly.

For both cassettes, experimental data in Fig. 3D and G show that the hot motifs in the leftmost hot-hot′ region have a higher mutation frequency than the hot motifs in other regions. In contrast, the hot motifs in the hot-frigid regions (Fig. 3G, center and right spectra) are colder than those in the hot-cold region (Fig. 3D, right spectrum). In Fig. 3D, the total hot-spot mutation count on the left hot-hot′ region on the cassette is 643, compared to 434 on the right hot-cold region. Similarly in Fig. 3G, the total hot-spot mutations on the left hot-hot′ region is 572, compared to 230 and 243 in the center and right hot-frigid regions, respectively. On the edges of both cassettes in Fig. 3D and G, there are noticeable depletions in the mutation counts transitioning into the silent regions. The observed variations in the mutation probabilities across both cassettes are significantly stronger than fluctuations coming from experimental variability. The experimental variability, $\sim$(number of mutations)$^{1/2}$, is typically $< 5$ over the entire cassette, whereas the sequence-dependent effects on motif deamination efficiencies are generally $>$ 20. Both sets of experimental data also exhibit contextual signatures consistent with each other. The experimental mutation profiles in Figs. 3E and G clearly corroborate predictions from the intrusive model in Figs. 3E and H. Contrasting this, the passive model predicts that the hot motif mutation probabilities should have no sequence dependence. It can be soundly rejected with a $p$-value

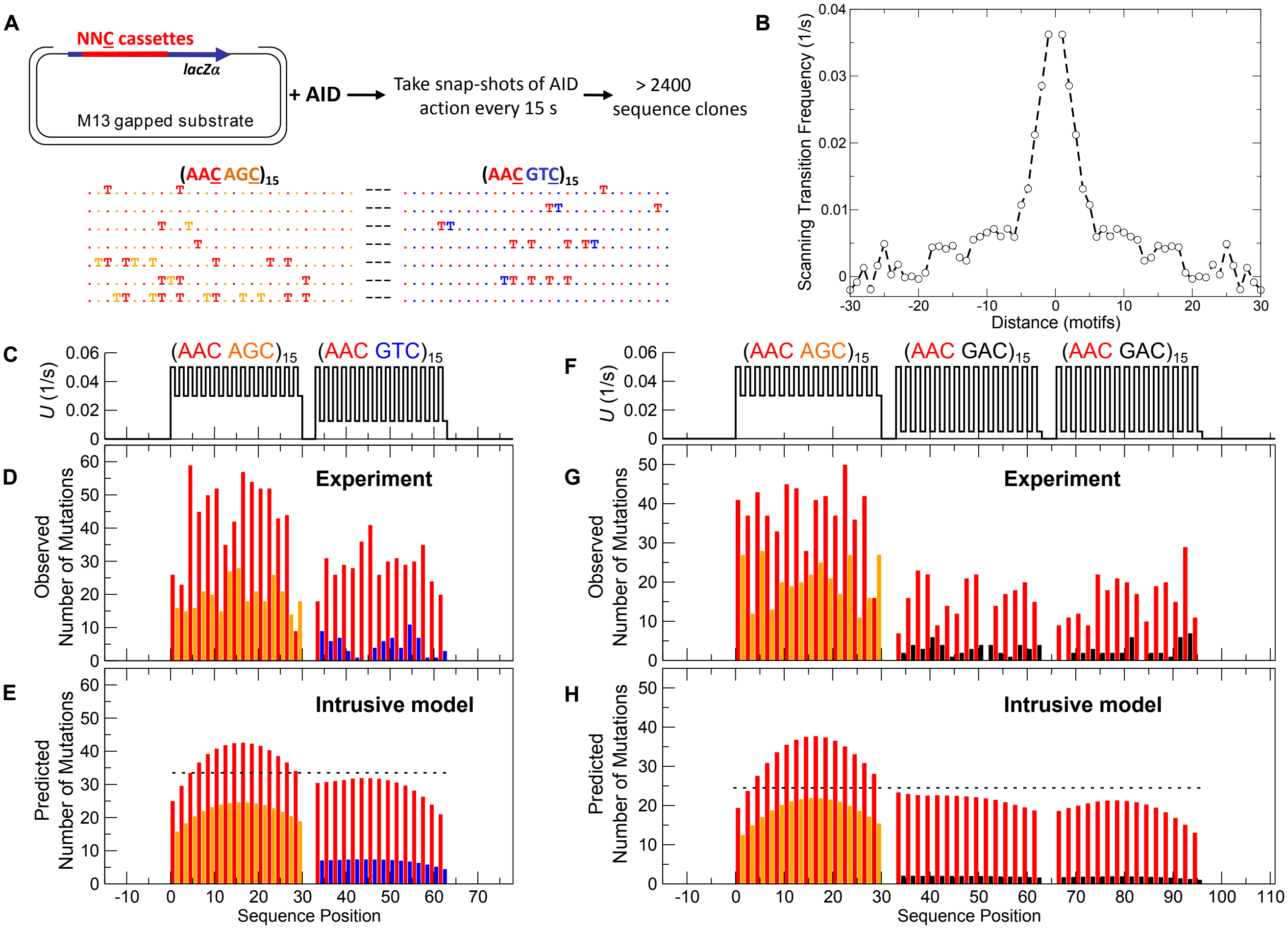

FIG. 3. Experimental setup and results. (A) Deamination assay reports AID-catalyzed deaminations on target cassettes with multiple trinucleotide motifs NNC̲ embedded in *lacZα* reporter gene. Examples of deaminated mutant clones with C→U deaminations shown as Ts. (B) Elements of the scanning transition matrix $W_{ij}$ as a function of distance $i - j$ derived from mutation correlation analysis on a homogeneous (AGC̲)$_{56}$ cassette [12]. (C) Motif-dependent intrinsic deamination rates along the DNA sequence for an inhomogeneous hot-hot′/hot-cold cassette, and (D) the experimentally observed mutation frequencies showing sequence-dependent deamination probabilities, (E) the mutation frequencies predicted by the intrusive model after $t$ = 45s, the length of the experiments. (F) Intrinsic deamination rates along the DNA sequence for a hot-hot′/hot-frigid/hot-frigid cassette, and (G) the observed mutation frequencies, (H) the frequencies predicted by the intrusive model after $t$ = 60s. Experimental variability in (D) and (G), ∼(number of mutations)$^{1/2}$, are typically $< 5$, whereas the sequence-dependent effects are generally $> 20$. Dashed lines in (E) and (H) indicate expected hot-motif mutation counts from the passive model which exhibit no sequence context dependence. (Other details of the passive model results are not shown.)

$< 0.0001$ based on the experimentally observed differentials in the hot-motif mutation counts in the hot-hot′ region compared to the hot-cold and hot-frigid regions on the two cassettes. Mutation counts from the passive model, which are uniform across the entire sequence regardless of sequence context, are shown as dashed lines in Figs. 3E and H for comparison.

The intrusive model explains the contextual dependence in the mutation profiles observed in the experiments. Biological systems could potentially exploit these contextual effects to guide the catalytic actions of random-walk enzymes. It may also be possible to control the target conversion efficiencies of scanning enzymes by re-engineering the substrate sequence. While the experiments studied here used regularly repeating target sequences, similar contextual effects occur for random sequences, though the details of the contextual signatures will depend on the interplay between the length scales of the scanning versus the inhomogeneity of the target sequence discussed in the next section.

## IV. ISOMORPHISM BETWEEN SEQUENCE-DEPENDENT CATALYTIC COUPLING AND QUANTUM DELOCALIZATION

Despite the fact that this is a classical system, the origin of the contextual influence or "spillover" derived from the sequence neighborhood predicted by the intrusive model and observed in the experiments can best be understood via the quantum isomorphism. This quantum isomorphism suggests that the enzyme interrogates its target sites by repeatedly applying a position operator (see Eq.(6) below), causing interruptions to its scanning paths. The outcomes of these effects are contingent on how the targets are arranged in space, thereby producing the observed contextual dependence in the target conversion spectrum.

Any quantum particle, due to its translational Hamiltonian, has an intrinsic dispersion which characterizes the extent of its delocalization. The same dispersion effect, in the equivalent scanning-coupled catalysis system, enables the enzyme to communicate information from surrounding motifs across a distance. The length scale of the spillover predicted by the intrusive model is controlled by the characteristic diffusion length of the enzyme, which is the equivalence of quantum delocalization. But while diffusion induces dispersion, the heterogeneity of the intrinsic mutation rates on the substrate targets also competes against it to attempt to contain the spillover. Formally, this competition is analogous to quantum confinement. The motif-dependent intrinsic mutation rates $u_i$ produce a potential $\mathbf{U}$ in the quantum analog, and hot motifs map onto low-energy sites on the potential energy surface. These hot motifs behave like attractors for the scanning paths, which visit these hot sites with a disproportionately higher frequency. Figure 2D shows the probability of finding the enzyme $\rho_i$ as a function of position on the sequence, decomposed into individual eigenfunctions of $\mathbf{H}_1$. As expected, the ground state (red) contributes most significantly to the overall $\rho_i$ and it is predominately localized in the hot-hot$'$ domain. The next eigenfunction (green) is localized in the hot-cold domain. These two eigenfunctions make up most of the contributions to $\rho_i$. By permuting or rearranging the hot, hot$'$ and cold motifs, the eigenfunctions can be shifted. It is therefore possible to produce different dispersion structures by engineering the sequence. When the variation of the mutation rates on a heterogeneous substrate sequence is comparable to the diffusion length of the enzyme, confinement sets in. The diffusion length of the enzyme $l_D$, which is the typical distance travelled by the enzyme between mutations, can be estimated from the diffusion coefficient $D_0$ associated with the bare propagator $\mathbf{K}_0(t)$ according to $\sqrt{D_0/\bar{u}}$, where $\bar{u}$ is the average intrinsic mutation rate across all sites. If $l_D$ is smaller than the length scale of the spatial heterogeneity in the site-dependent mutation rates, the paths will be trapped. However, when the substrate is replaced by a homogeneous repeating sequence with motifs of the same kind, the potential surface in the quantum analog becomes flat. This causes scanning to uncouple from catalysis and reduces a homogeneous substrate in the intrusive picture to the passive model, as we have previously shown mathematically and experimentally [12].

While the length scales of diffusion versus motif heterogeneity "interact" to generate marked sequence-coupled mutation rates, the timescales of the mutations versus scanning must also match in order for contextual spillover to be significant. In the limit where one is much faster than the other, contextual dependence disappears. First, in the limit where diffusion is fast, the enzyme will be able to scan all targets between mutations. Consequently, conversion on each site should simply occur proportionately to its intrinsic mutation susceptibility, and in this limit the intrusive model reverts to the passive model. This is illustrated in Fig. 2E for the same cassette as Fig. 2C, and the spillover effect is now gone. Second, if mutations occur much faster than diffusion the enzyme would be almost stationary between mutations, and since the initial binding of the enzyme has no site-preference the observed mutation probabilities should be simply proportional to the intrinsic mutation rates. Therefore, the rapid mutation limit also reduces the intrusive picture to the passive picture, and this is illustrated in Fig. 2F. Enzymes optimized for high target-seeking efficiency, such as the endonucleases, operate in this regime because of their fast catalytic rates. In this limit the Markovian approximation overestimates the conversion probabilities because it allows multiple hits on the same motif, and Fig. 2F shows that hit rates on the hotter motifs are exaggerated. However, including non-Markovian effects will not alter the fact that spillover is absent in the limit diffusion is very slow. Because of these two opposing limits, rapid diffusion versus rapid mutation, nontrivial contextual effects in the mutation probabilities will only manifest themselves inside a special parameter regime where the length scales as well as the timescales of both scanning and catalysis become comparable. The particular combination of scanning and mutations characteristics of AID-catalyzed mutations on ssDNA places it right in the center of this nontrivial parameter regime. Appendices C and D discuss other equivalent models and how they are related in the intrusive picture.

## V. NON-MARKOVIAN MONTE CARLO ANALYSIS TO ELIMINATE SAME-SITE CATALYTIC EVENTS

To capture non-Markovian effects left out in the approximate propagator $\mathbf{K}_1(t)$, we exploit the quantum isomorphism further to construct a Hamiltonian that is fully equivalent to scanning-coupled catalysis in which multiple conversions of the same target are prohibited. The imaginary-time quantum system that emerges is analogous to a magnetic encoder moving over a onedimensional spin-1/2 lattice with the Hamiltonian:

$$\mathbf{H}_2 = -\mathbf{W} - \sum_{i=1}^{N} u_i \mathbf{Q}_i [\hat{\sigma}_x]_i - b \sum_{i=1}^{N} [\hat{\sigma}_z]_i. \qquad (6)$$

In Eq.(6), $\mathbf{W}$ and $\mathbf{Q}$ operate on the scanning degree of freedom $q$. $[\hat{\sigma}_x]_i$ and $[\hat{\sigma}_z]_i$ are $2 \times 2$ Pauli matrices operating on the spin degree of freedom on each site $i$. Each spin on the lattice corresponds to a mutable motif and begins in its $\downarrow$-state representing an unmutated site. $-\mathbf{W}$ maps to the translational Hamiltonian. The diagonal matrix $\mathbf{Q}_i$ measures whether the encoder is on site $i$, and if it is, it can rewrite the spin state of that site by flipping it from $\downarrow$ to $\uparrow$ via the operator $[\hat{\sigma}_x]_i$, and the coupling $u_i$ is site-dependent. Once flipped, the operator $[\hat{\sigma}_z]_i$ provides a bias to stabilize the $\uparrow$-state, inhibiting motifs from receiving multiple mutations. The Boltzmann operator $\mathbf{K}_2(\beta) = \exp[-\beta \mathbf{H}_2]$ for this system is again completely equivalent to the propagator for scanning-coupled catalysis when $\beta$ is mapped to time $t$. (See Appendix E for more details.) The Hamiltonian (6) cannot be solved analytically or by numerical diagonalization (the size of the basis set being $2^N \times N$). Figure 4B shows results from large-scale path integral Monte Carlo simulations [37, 38] for the $\uparrow$-spin (i.e. converted motif) profile in the isomorphic quantum system corresponding to the mutation probabilities on the 96-motif hot-hot′/hot-frigid/hot-frigid cassette employed in the experiments shown in Fig. 3G, and Fig. 4C the corresponding probability of finding the enzyme as a function of sequence position. For AID catalysis, the reaction rates are slow ($\leq 0.05$ s$^{-1}$) [12, 13]. Given the typical during of the experiments (30 s $\leq t \leq$ 2 min), repeat events on each motif are expected to be rare. Comparing Figs. 4B and D, it is clear that while non-Markovian effects are present, they do not significantly alter the qualitative signatures of the spillover effects. Analogous to the hot-hot′/hot-cold cassette, the computed mutation probabilities in Fig. 4B corroborate the experimental observations in Fig. 3G. In contrast with the hot-hot′/hot-cold cassette (Fig. 3D and E), the mutation profile of the hot-hot′/hot-frigid/hot-frigid cassette in the center domain does not exhibit the same rounding as the two edges. Instead, the center domain interpolates between the 5′ hot-hot′ and the 3′ hot-frigid edges, lending further support to the hypothesis that the spillover effects are due to the dispersion inherent in the $\mathbf{W}$ matrix ($l_D \sim$ 8.4 motifs).

## VI. SUMMARY

The intrusive model predicts nontrivial context dependencies in the mutation probabilities on the substrate targets in AID-catalyzed C→U mutations deposited on single-stranded DNAs due to the coupling between scanning and catalysis. Experiments confirm these predictions. How do the catalyzed conversions impact the scanning paths so significantly if they are not being transported with the enzyme? Within the formal mathematical model we have described, the origin of how mutation events and scanning become entangled is related to what is commonly referred to as measurement theory in quantum mechanics. When a measurement is applied to a quantum particle, it is thrown into an eigenstate of the measurement operator. In the quantum isomorphic system described by $\mathbf{H}_2$, mutations on each site $i$ are generated by the term $\mathbf{Q}_i[\hat{\sigma}_x]_i$. $\mathbf{Q}_i$ measures the position of the enzyme and $[\hat{\sigma}_x]_i$ performs the conversion. Because of the coupling between $\mathbf{Q}_i$ and $[\hat{\sigma}_x]_i$, even though mutations may not be directly modifying the diffusive behavior of the enzyme, they effectively perform a measurement on the enzyme's position repeatedly. Because of this, the catalyzed conversions materially interact with the enzyme's scanning paths in a nontrivial manner, which requires that they must be treated explicitly in the path integrals to properly describe the time-evolution of the composite system. Direct support for this may have been observed in recent single-molecule studies of Apo3G on ssDNA [39] and p53 on dsDNA [40], both showing evidence for quasi-localized substrate scanning.

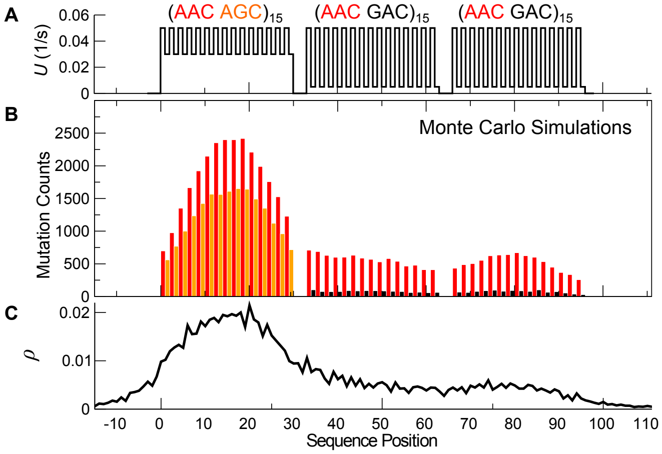

FIG. 4. Computed mutation probabilities on a hot-hot′/hot-frigid/hot-frigid test cassette. (A) The site-dependent intrinsic mutation rates $u_i$ on this 96-motif cassette, which corresponds to the experiment in Fig. 3G. (B) Mutation probabilities computed by Monte Carlo simulations. (C) Probability of finding the enzyme $\rho_i$ as a function of sequence position on the substrate from the Monte Carlo simulations.

The intrusive model describes the contextual signature expressed in the mutational probabilities quite well. Mathematically, the passive and intrusive pictures investigated in this paper represent the only parameter-free minimal analytical models that could be invoked to explain the stochastic dynamics of scanning-coupled enzymatic processes. The intrusive picture apparently contains sufficient physics to explain the key experimental observables. Incorporating additional biological or molecular mechanisms into the models requires adding more parameters. Indeed, it will be interesting to try to experimentally ascertain precise biophysical mechanisms underlying the coupling between scanning and catalysis implicated by the intrusive model.

## ACKNOWLEDGMENTS

This work was supported by the National Science Foundation CHE-0713981 (CHM) and by National Institutes of Health grants ES13192 (MFG) and GM21422 (MFG).

## Appendix A: Perturbation Series

Each diagram in Fig. 1 represents a term in the perturbation expansion of the propagator in orders of $\mathbf{U}$, the diagonal matrix containing the site-dependent intrinsic mutation susceptibilities $\{u_i\}$ on the substrate. For example, the third and fourth diagrams on the right of the equation in Fig. 1 correspond to the second-order terms:

$$\int_0^t dt_1 \int_0^{t_1} dt_2 \sum_{i,j=1}^{N} \sum_{k,l=1}^{N} \left[e^{t_1 \mathbf{W}}\right]_{jk} u_k \left[e^{(t_2-t_1)\mathbf{W}}\right]_{kl} \times u_l \left[e^{(t-t_2)\mathbf{W}}\right]_{li} (1-\delta_{kl}), \qquad \text{(A1)}$$

where $\delta$ is Kronecker's delta and $\mathbf{W}$ is the scanning transition rate matrix. Coefficient of each of the terms corresponding to the loops diagrams can be found in [41]. For example, coefficients of all third-order diagrams on the second line of Fig. 1 may be obtained by expanding the product $(1-\delta_{kl})(1-\delta_{lm})(1-\delta_{km})$.

Non-loop diagrams in the perturbation series may be resummed using a Laplace transform in time. The resulting approximate propagator is:

$$\mathbf{K}_1(s) = \int_0^\infty dt e^{-st} \mathbf{K}_1(t) = [s - \mathbf{W} - \mathbf{U}]^{-1}. \qquad \text{(A2)}$$

This leads to a closed-form solution for the approximate propagator $\mathbf{K}_1(t) = \exp[t(\mathbf{W}+\mathbf{U})]$ in the time domain. This propagator satisfies the Block equation, which is the imaginary-time equivalent of the time-dependent Schrödinger equation:

$$\partial \mathbf{K}_1(t)/\partial t = -\mathbf{H}_1 \mathbf{K}_1(t) = [\mathbf{W}+\mathbf{U}]\mathbf{K}_1(t), \qquad \text{(A3)}$$

where $\mathbf{H}_1 = -(\mathbf{W}+\mathbf{U})$. The propagator $\mathbf{K}_1(t)$ can be computed by using the eigenfunctions and eigenvalues of the corresponding time-independent Schrödinger equation $\mathbf{H}_1\Psi = E\Psi$. In the quantum isomorphism, $-\mathbf{W}$ is equivalent to the translational Hamiltonian and $-\mathbf{U}$ to the potential. In quantum statistical mechanics, the matrix exponential $\mathbf{K}_1(t) = \exp[-\beta\mathbf{H}_1]$ corresponds to the Boltzmann operator, which is often referred to as the imaginary-time propagator.

Whereas in imaginary-time quantum mechanics it is usually the trace of $\exp[-\beta\mathbf{H}_1]$ that is of interest, in scanning-coupled catalysis, the path integral corresponds to the grand sum over all elements of the propagator matrix $\exp[t(\mathbf{W}+\mathbf{U}]$, since a trajectory may start from any motif $j$ and end on any $i$. Except for this difference, the two problems are mathematically equivalent. Due to this isomorphism, the scanning-coupled catalysis of a random-walk enzyme is formally identical to the problem of a quantum particle with a translational Hamiltonian $-\mathbf{W}$ subject to the site-dependent potential specified by the elements of the diagonal matrix $-\mathbf{U}$. The isomorphic quantum particle has an intrinsic delocalization length under the action of $-\mathbf{W}$, but the potential $-\mathbf{U}$ tampers this delocalization. The interplay between delocalization and confinement is manifested in the scanning-coupled catalysis by AID on ssDNAs in interesting ways. Notice that since motifs with higher intrinsic deamination rates map to sites with low potential energies on the isomorphic quantum lattice, hot motifs behave like attractors. In the special case of a homogeneous substrate which maps to a constant potential, the eigenfunctions of the isomorphic quantum particle are delocalized over the entire substrate in the form of Fourier waves. Counteracting this is the tendency of the potential $-\mathbf{U}$ to confine the eigenfunctions, which happens when $-\mathbf{U}$ is no longer constant on substrates with inhomogeneous motif sequences.

## Appendix B: Computing Mutation Probabilities from Eigensolutions to the Schrödinger Equation

Once the solution for the propagator $\mathbf{K}_1(t)$ is known, the mutation probability $P_1(m_i = +\frac{1}{2}; t)$ on any motif $i$ on the sequence may be computed as follows. $P_1(m_i = +\frac{1}{2}; t)$ is proportional to the sum over all paths with any deamination on $i$. Instead of computing the sum over this subset of paths, it is actually easier to sum the paths complementary to this set, i.e. those that have *no* deaminations on $i$ at all. This is easily done by calculating $\exp[t(\mathbf{W}+\overline{\mathbf{U}})]$ where $\overline{\mathbf{U}}$ is identical to $\mathbf{U}$ except one element $\overline{U}_{ii}$ has been set to 0. Subtracting $\exp[t(\mathbf{W}+\mathbf{U})]$ from the full propagator $\exp[t(\mathbf{W}+\mathbf{U})]$ then yields a sum over all paths with mutations at $i$. In this way, the mutation probability profile across the entire substrate can be computed by zeroing out the mutation rate on each motif one by one and repeating the diagonalization for each.

The motif-dependent mutation probabilities $P_1(m_i = +\frac{1}{2}; t)$ are single-site reduced probabilities of $P_n$. $P_n(m_i = +\frac{1}{2}, m_j = +\frac{1}{2}, \ldots; t)$ is the joint probability of finding $n$ mutations on sites $i$, $j$, etc.. While we have focused exclusively on the single-site mutation probabilities in this paper, higher-order correlations among mutations on two or more motifs are related to the joint probabilities $P_2$, $P_3$, etc. These correlations contain additional information regarding the coupling between scanning and mutations. They also control how the mutation are clustered. These multi-point mutation correlations can be calculated easily with a method similar to that used for the single-site mutation probabilities: By suspending deaminations on more than one motif in the cassette at a time, the Hamiltonian can be re-diagonalized and the number of paths with mutations simultaneously on two or more sites can thus be computed.

The diagonalization of the Hamiltonian was performed numerically in Linpack. Typically, a long silent sequence with 60 to 90 motifs having zero intrinsic mutation susceptibility was appended to both the $5'$ and $3'$ ends of the cassette. Periodic boundary condition was used for the scanning transition rate matrix, and we verified that the boundary effects were negligible by varying the lengths of the silent end caps.

## Appendix C: Equivalence to Diffusion with Source Terms

The Block equation Eq.(A3) has another alternative interpretation. Equation (A3) is formally equivalent to a diffusion equation with a site-dependent source term $\mathbf{U}$. This correspondence implies that in the intrusive picture the problem of scanning-coupled catalysis where the chemical conversions are deposited on a stationary substrate instead of being transported with the diffusing species can actually be modeled by a diffusion equation with spatially-distributed source terms. While this result is a direct consequence of the intrusive picture, as we have discussed in the main text, there is no obvious *a priori* basis for incorporating the mutations into the diffusion equation of the enzyme. This equivalence would not have been obvious without the analytical models presented in the main text.

## Appendix D: Path Integrals and Other Equivalent Systems

In the perturbation series, each diffusion path of the scanning enzyme is also coupled to how many mutations occur along its trajectory and the times and positions at which they are deposited. If the proper measure is assigned to these paths according to the prescription in Sect. I, the propagator may be expressed in terms of a path integral over all possible scanning trajectories $q(t)$ [31]:

$$[K(t)]_{qq'} = \int_q^{q'} \mathcal{D}q(t)\mathcal{P}_0[q(t)] \int \mathcal{D}m(t) e^{I[q(t),m(t)]}, \tag{D1}$$

where the functional $\mathcal{P}_0[q(t)]$ represents the intrinsic weight of the scanning path $q(t)$ coming from the bare propagator $\mathbf{K}_0(t) = \exp[t\mathbf{W}]$, whose matrix elements are $[\mathbf{K}_0(t)]_{qq'} = \int_q^{q'} \mathcal{D}q(t)\mathcal{P}_0[q(t)]$, $\int \mathcal{D}m(t)$ denotes an integral over all possible mutation paths $m(t)$, and the functional $I[q(t), m(t)]$ describes the interaction between mutations and scanning.

Within the passive model, the scanning paths $q(t)$ and mutation paths $m(t)$ have no interaction with each other, equivalent to setting $I[q(t), m(t)] = 0$. On the other hand, in the intrusive picture it is easy to show from the perturbation series that the Markovian approximation to the propagator $\mathbf{K}_1(t)$ corresponds to the path integral in Eq.(D1) with $\int \mathcal{D}m(t) \exp(I[q(t), m(t)]) \rightarrow \exp(\int_0^t u_{q(\tau)} d\tau)$. This result can be interpreted as a temporally-nonhomogeneous Poisson process occurring along different scanning paths, each with a time-dependent mutation rate $u_{q(t)}$. In terms of this, the approximate propagator becomes:

$$[K_1(t)]_{qq'} = \int_q^{q'} \mathcal{D}q(t)\mathcal{P}_0[q(t)] \exp\left[\int_0^t u_{q(\tau} d\tau\right], \tag{D2}$$

where for each scanning path $q(t)$, the factor $\exp(\int_0^t u_{q(\tau})d\tau$ reflects the total measure of all possible mutation paths that may occur along $q(t)$. This exponential factor ascribes higher preference to scanning paths that frequent the hot spots, and the effect of this is reflected in Fig. 2D, which shows that the probability of finding the enzyme is higher at the hot motifs compared to the cold. As the scanning paths are drawn to the hot motifs, their characteristic dispersions are controlled by the diffusion rate of the enzyme which limits how far the enzyme may travel over time. The combination of these two factors causes the probability of finding the enzyme at sites closer to the hot motifs to be disproportionately higher, and this is manifested as the spillover effects observed in the mutational probability profiles.

In the discussions surrounding the Hamiltonian $\mathbf{H}_1$ of the quantum isomorphic system, we have argued that a homogeneous substrate with motifs having a constant mutation rate $u$ across all sites corresponds to a quantum system with a flat potential energy surface, for which scanning should uncouple from the mutations. In the path integral picture, we can also interpret this uncoupling as a result of the integral $\int_0^t u_{q(\tau)} d\tau \rightarrow ut$, which becomes identical for all scanning paths $q(t)$. In the special case of a homogeneous substrate, every scanning path has identical weight in the intrusive picture, and this reduces the intrusive model to a passive one.

The path integral (D2) suggests an isomorphism to yet another quantum system. In this isomorphic system, the motion of a magnetic encoder is coupled to a single spin-1/2 system, with the Hamiltonian

$$\mathbf{H}_3 = -(\mathbf{W} + \mathbf{U}\hat{\sigma}_x). \tag{D3}$$

This Hamiltonian is similar to $\mathbf{H}_1 = -(\mathbf{W} + \mathbf{U})$, except the spin degree of freedom has been rendered explicit. Each time $\mathbf{U}\hat{\sigma}_x$ acts, it measures the mutation rate at the location of the encoder $q$ and simultaneously flips the spin $m$ from $\downarrow$ to $\uparrow$ or vice versa, with the number of spin flips along the dual path $(q(t), m(t))$ representing the total number of mutations. In the Hamiltonian

(D3), the number of spin flips is unconstrained, and this corresponds to the Markovian limit of the perturbation series in Fig. 1. The Hamiltonian (D3) can be easily solved by using the (unnormalized) symmetric and antisymmetric spin superpositions $\{|\downarrow\rangle + |\uparrow\rangle, |\downarrow\rangle - |\uparrow\rangle\}$, arriving at the same formal result given in Eq.(D2) for the time-nonhomogeneous Poisson process.

## Appendix E: Spin-1/2 Lattice Model

The spin-1/2 lattice quantum system describe by the Hamiltonian (6) is a multi-spin generalization of Hamiltonian (D3), where the mutation on each motif $i$ is represented by an individual spin variable $m_i \in \{\downarrow, \uparrow\}$. Ascribing a separate spin to each motif allows the mutations to be counted individually. With $b > 0$, the additional term $-b\sum_{i=1}^{N}\{\hat{\sigma}_z\}_i$ in Eq.(6) stabilizes the ↑ state of each site once it has been flipped, preventing multiple mutations from being deposited on the same motif. The Hamiltonian Eq.(6) thus captures all non-Markovian terms in the perturbation series in Fig. 1 as well. When the bias $b = 0$, the system would revert to fully Markovian, and in this limit the Hamiltonian Eq.(6) can be solved by using the symmetric and antisymmetric spin superpositions $\{|\downarrow\rangle_i + |\uparrow\rangle_i, |\downarrow\rangle_i - |\uparrow\rangle_i\}$ for each spin, arriving at formally the same result as Eq.(D2) for the one-spin Hamiltonian (D3). When the bias $b \neq 0$, the system can no longer be solved exactly. This complexity comes from including non-Markovian effects.

A discretized path integral Monte Carlo simulation was constructed for the Hamiltonian in Eq.(6) based on standard methods using second-order accurate short-time propagators [38]. The convergence of the discretized path integrals were slow. Typically, a discretization time between 0.1 to 0.05s were used for the MC simulations to generate scanning and mutation paths of 45 to 60s in duration. The ergodicity of the simulations was also rather weak, requiring approximately 0.5 trillion Monte Carlo passes in total to generate the results shown in Fig. 4B.

## Appendix F: Experimental Methods

The library of *lacZα* clones containing AID-catalyzed C→U deaminations in inhomogeneous cassettes of trinucleotide motifs were generated experimentally as follows. Gapped DNA substrates containing either 60 trinucleotide motifs (AAC AGC)$_{15}$-sss-(AAC GTC)$_{15}$ or 90 trinucleotide motifs (AAC AGC)$_{15}$-sss-(AAC GAC)$_{15}$-sss-(AAC GAC)$_{15}$ embedded in *lacZα* (see Fig. 3A, sketch) were constructed as described in [13], where sss represents a 9-nt silent spacer. The gapped DNA were incubated with AID and deamination reactions were quenched at 15, 30, 45, 60, 120, 300 and 600s. C→U deaminations in trinucleotide NN$\underline{\text{C}}$motifs create stop codons within the *lacZα* reading frame that result in mutant M13 phage clones. Mutant M13 phage DNA was isolated, and the inserted cassettes and the *lacZα* portion on the 3′ side of the cassette were sequenced. C→U deaminations were detected as C→T transition mutations [12, 13]. To ensure that virtually all deaminations on individual substrates were caused by a single AID molecule, AID and gapped DNA concentrations were chosen so that the fractions of mutated clones were always less than about 2%, as prescribed by Poisson statistics [12, 13]. The mutation probabilities shown in Fig. 3D for the hot-hot′/hot-cold cassette were obtained from clones with an incubation time of 45s. Those shown in Fig. 3G for the hot-hot′/hot-frigid/hot-frigid cassette were collected from a number of experiments with various incubation times averaging approximately 60s.

[1] R. A. Alberty and G. G. Hammes, J. Phys. Chem. **62**, 154 (1958).

[2] P. H. Richter and M. Eigen, Biophys. Chem. **2**, 255 (1974).

[3] A. D. Riggs, Bourgeoi.S, and M. Cohn, J. Mol. Biol. **53**, 401 (1970).

[4] O. G. Berg and C. Blomberg, Biophys. Chem. **4**, 367 (1976).

[5] R. B. Winter, O. G. Berg, and P. H. von Hippel, Biochemistry **20**, 6961 (1981).

[6] S. E. Halford and J. F. Marko, Nucleic Acids Res. **32**, 3040 (2004).

[7] M. Coppey, O. Benichou, R. Voituriez, and M. Moreau, Biophys. J. **87**, 1640 (2004).

[8] O. Benichou, M. Coppey, M. Moreau, P. H. Suet, and R. Voituriez, Phys. Rev. Lett. **94**, 198101 (2005).

[9] L. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J. Leith, and A. Kosmrlj, J Phys Math Theor **42**, 434013 (2009).

[10] A. Tafvizi, F. Huang, J. S. Leith, A. R. Fersht, L. A. Mirny, and A. M. van Oijen, Biophys. J. **95**, L1 (2008).

[11] M. Muramatsu, K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, and T. Honjo, Cell **102**, 553 (2000).

[12] C. H. Mak, P. Pham, S. A. Afif, and M. F. Goodman, J. Biol. Chem. **288**, 29786 (2013).

[13] P. Pham, P. Calabrese, S. J. Park, and M. F. Goodman, J. Biol. Chem. **286**, 24931 (2011).

[14] P. Pham, R. Bransteitter, J. Petruska, and M. F. Goodman, Nature **424**, 103 (2003).

[15] S. G. Conticello, Genome Biol **9**, 229 (2008).

[16] J. U. Peled, F. L. Kuang, M. D. Iglesias-Ussel, S. Roa, S. L. Kalis, M. F. Goodman, and M. D. Scharff, Annu. Rev. Immunol. **26**, 481 (2008).

[17] M. Jaszczur, J. G. Bertram, P. Pham, M. D. Scharff, and M. F. Goodman, Cell. Mol. Life Sci. **70**, 3089 (2013).

[18] S. Nik-Zainal, P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, A. Shlien, S. L. Cooke, J. Hinton, A. Menzies, L. A. Stebbings, C. Leroy, M. Jia, R. Rance, L. J. Mudie, S. J. Gamble, P. J. Stephens, S. McLaren, P. S. Tarpey, E. Papaemmanuil, H. R.

Davies, I. Varela, D. J. McBride, G. R. Bignell, K. Leung, A. P. Butler, J. W. Teague, S. Martin, G. Jonsson, O. Mariani, S. Boyault, P. Miron, A. Fatima, A. Langerod, S. A. Aparicio, A. Tutt, A. M. Sieuwerts, A. Borg, G. Thomas, A. V. Salomon, A. L. Richardson, A. L. Borresen-Dale, P. A. Futreal, M. R. Stratton, and P. J. Campbell, Cell **149**, 994 (2012).

[19] S. A. Roberts and D. A. Gordenin, BioEssays, DOI: 10.1002/bies.201300140(2014).

[20] S. A. Roberts, M. S. Lawrence, L. J. Klimczak, S. A. Grimm, D. Fargo, P. Stojanov, A. Kiezun, G. V. Kryukov, S. L. Carter, G. Saksena, S. Harris, R. R. Shah, M. A. Resnick, G. Getz, and D. A. Gordenin, Nat. Genet. **45**, 970 (2013).

[21] S. A. Roberts, J. Sterling, C. Thompson, S. Harris, D. Mav, R. Shah, L. J. Klimczak, G. V. Kryukov, E. Malc, P. A. Mieczkowski, M. A. Resnick, and D. A. Gordenin, Mol. Cell **46**, 424 (2012).

[22] M. B. Burns, L. Lackey, M. A. Carpenter, A. Rathore, A. M. Land, B. Leonard, E. W. Refsland, D. Kotandeniya, N. Tretyakova, J. B. Nikas, D. Yee, N. A. Temiz, D. E. Donohue, R. M. McDougle, W. L. Brown, E. K. Law, and R. S. Harris, Nature **494**, 366 (2013).

[23] L. Pasqualucci, P. Neumeister, T. Goossens, G. Nanjangud, R. S. Chaganti, R. Kuppers, and R. Dalla-Favera, Nature **412**, 341 (2001).

[24] G. Gaidano, L. Pasqualucci, D. Capello, E. Berra, C. Deambrogi, D. Rossi, L. Maria Larocca, A. Gloghini, A. Carbone, and R. Dalla-Favera, Blood **102**, 1833 (2003).

[25] M. Montesinos-Rongen, R. Schmitz, C. Courts, W. Stenzel, D. Bechtel, G. Niedobitek, I. Blumcke, G. Reifenberger, A. von Deimling, B. Jungnickel, O. D. Wiestler, R. Kuppers, and M. Deckert, Am. J. Pathol. **166**, 1773 (2005).

[26] D. Rossi, E. Berra, M. Cerri, C. Deambrogi, C. Barbieri, S. Franceschetti, M. Lunghi, A. Conconi, M. Paulli, A. Matolcsy, L. Pasqualucci, D. Capello, and G. Gaidano, Haematologica **91**, 1405 (2006).

[27] A. Jeltsch and R. Z. Jurkowska, Trends Biochem. Sci. **39**, 310 (2014).

[28] P. C. Blainey, G. Luo, S. C. Kou, W. F. Mangel, G. L. Verdine, B. Bagchi, and X. S. Xie, Nat Struct Mol Biol **16**, 1224 (2009).

[29] R. H. Porecha and J. T. Stivers, Proc Natl Acad Sci U S A **105**, 10791 (2008).

[30] M. Chaichian and A. P. Demichev, *Path integrals in physics* (Institute of Physics, Bristol ; Philadelphia, 2001).

[31] R. P. Feynman and A. R. Hibbs, *Quantum mechanics and path integrals*, International series in pure and applied physics (McGraw-Hill, New York,, 1965) pp. xiv, 365 p.

[32] H. S. Carslaw and J. C. Jaeger, *Conduction of heat in solids*, 2nd ed. (Clarendon Press ; Oxford University Press, Oxford Oxfordshire New York, 1986) pp. viii, 510 p.

[33] C. W. Gardiner, *Handbook of stochastic methods for physics, chemistry, and the natural sciences*, 3rd ed., Springer series in synergetics, (Springer-Verlag, Berlin ; New York, 2004) pp. xvii, 415 p.

[34] L. J. S. Allen, *An introduction to stochastic processes with applications to biology*, 2nd ed. (Chapman and Hall/CRC, Boca Raton, FL, 2011) pp. xxiv, 466 p.

[35] G. Grimmett and D. Stirzaker, *Probability and random processes*, 3rd ed. (Oxford University Press, Oxford ; New York, 2001) pp. xii, 596 p.

[36] H. Kleinert, *Path integrals in quantum mechanics, statistics, polymer physics, and financial markets*, 4th ed. (World Scientific, Hackensack, NJ, 2006) pp. xliii, 1547 p.

[37] D. Chandler and P. G. Wolynes, J. Chem. Phys. **74**, 4078 (1981).

[38] K. S. Schweizer, R. M. Stratt, D. Chandler, and P. G. Wolynes, J. Chem. Phys. **75**, 1347 (1981).

[39] G. Senavirathne, M. Jaszczur, P. A. Auerbach, T. G. Upton, L. Chelico, M. F. Goodman, and D. Rueda, J. Biol. Chem. **287**, 15826 (2012).

[40] J. S. Leith, A. Tafvizi, F. Huang, W. E. Uspal, P. S. Doyle, A. R. Fersht, L. A. Mirny, and A. M. van Oijen, P Natl Acad Sci USA **109**, 16552 (2012).

[41] G. W. Ford and G. E. Uhlenbeck, P Natl Acad Sci USA **42**, 122 (1956).