A General Theoretical Framework to Design Base Editors with Reduced Bystander Effects

Qian Wang
University of Science and Technology of China
Jie Yang
Rice University
Zhicheng Zhong
University of Science and technology of China
Xue Gao
Rice University https://orcid.org/0000-0003-3213-9704
Anatoly Kolomeisky (tolya@rice.edu)
Rice University https://orcid.org/0000-0001-5677-6690

Article

Keywords: Base editors, bystander editing, chemical-kinetic modeling, molecular dynamic simulations

Posted Date: February 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-210065/v1

Version of Record: A version of this preprint was published at Nature Communications on November 11th, 2021. See the published version at https://doi.org/10.1038/s41467-021-26789-5.

A General Theoretical Framework to Design Base Editors with Reduced Bystander Effects

Qian Wang[1,2,#,*], Jie Yang[3,#], Zhicheng Zhong[1], Xue Gao[3,4,5*], Anatoly B. Kolomeisky[2,3,5,6,*]

1. Hefei National Laboratory for Physical Sciences at the Microscale and Department of Physics, University of Science and Technology of China, Hefei, Anhui 230026, China
2. Center for Theoretical Biological Physics, Rice University, Houston, TX 77005, USA
3. Department of Chemical and Biomolecular Engineering, Rice University, Houston, TX 77005, USA
4. Department of Bioengineering, Rice University, Houston, TX 77005, USA
5. Department of Chemistry, Rice University, Houston, TX 77005, USA
6. Department of Physics and Astronomy, Rice University, Houston, TX 77005, USA

# These authors contributed equally to this work

Corresponding authors: Qian Wang (wqq@ustc.edu.cn); Xue Gao (xue.gao@rice.edu) and Anatoly B. Kolomeisky (tolya@rice.edu)

Keywords: Base editors, bystander editing, chemical-kinetic modeling, molecular dynamic simulations

Abstract
Base editors (BEs) hold great potential for gene therapy. However, high precision base editing requires BEs that can discriminate between the target base and multiple bystander bases within a narrow active window (4 - 10 nucleotides). To assist in the design of these optimized editors, we propose a discrete-state stochastic approach to build an analytical model that describes the probabilities of editing the target base and bystanders. Combined with all-atom molecular dynamic simulations, our model well reproduces the experimental data of A3A-BE3 and its variants for target and bystander editing. Building upon this model, we propose several general principles that can guide the design of BEs with a reduced bystander effect. We used these principles to improve the A3G-BEs with high precision and verified their base-editing activities experimentally. In summary, our study provides a computational-aided platform to assist in designing BEs with reduced bystander effects.

Introduction

The development of genome editing tools associated with the clustered regularly interspaced short palindromic repeat (CRISPR) systems has revolutionized biomedical studies. Holding a great potential for the treatment of genetic diseases, diverse precise genome editing tools based on CRISPR-Cas9 have been developed, such as the homology-directed repair (HDR) based systems, as well as cytosine and adenosine base editors (BE)(1-3). While the HDR method requires double-stranded DNA breaks (DSBs) and causes unpredictable editing outcomes, BEs use nickase Cas9 (nCas9), enabling more precise modifications without generating DSBs (4-6). For example, cytosine BEs (CBEs) are constructed by fusion of a cytosine deaminase domain with nCas9. This fusion protein forms a complex with the guided RNA and performs site-specific deamination to convert cytosine (C) to uracil (U) in the deaminase activity window. This U gets subsequently replaced to thymine (T) by the endogenous cellular repairing machinery, resulting in an overall C-to-T substitution at the defined genomic site. Since point mutations are responsible for more than half of human disease-associated genetic variants (2), BEs are superior due to their higher editing efficiency in the correction of pathogenic single nucleotide polymorphisms (7), avoiding unwanted DSBs, and preventing the formation of insertions and deletions (4, 5).

While some BE variants were engineered to improve the product purity and overall editing efficiency (8, 9), one of the major challenges in base editing is to discriminate among multiple identical bases located within the deaminase active window (2) of 4-10 nucleotides. As a result, the target base and also other bystander bases will all be modified, impacting negatively on the precision of genome editing outcomes. To address this issue, further engineered BEs have been developed by introducing beneficial mutations to deaminase (10-12). For example, compared to the wild-type APOBEC3A (A3A)-BE3, an engineered A3A CBE with the mutation N57G maintained high editing activity at the target C in the T<u>C</u>R motif with greatly reduced activity against bystanders (11). Also, followed by several rounds of screening and validation of rational mutagenesis, we previously engineered an APOBEC3G (A3G)-CBE that preferentially edits the second C in the "C<u>C</u>" motif with 6,000-fold improvement in perfectly modified alleles compared to the original BE4max (12). Despite these successes, a general theoretical framework to guide the design of mutations that can lead to high editing activity at the target base and low activity at bystanders (defined as BE high editing selectivity) is still missing. Mutation selections in these previous studies were mostly guided by structural considerations: starting from the identification of key residues in the deaminase near the DNA binding motif and then mutating those residues to form a candidate library for experimental validation. This design process could be greatly accelerated with a comprehensive theoretical model that could quantitatively predict the effect of specific mutations on editing activity at the target base and bystanders. In addition, such theoretical model would also improve our fundamental understanding of the biochemical and biophysical processes that take place during base editing.

Molecular dynamic (MD) simulations have been used to study the activity of BE complexes and the role of beneficial mutations to enhance overall editing activity (both at target and bystanders) (13, 14). Herein we present a comprehensive multi-scale theoretical approach to describe the molecular processes taking place during BE editing, explaining at the microscopic level the role of beneficial mutations in selecting target base over bystanders. To fulfill this goal, we have built a general theoretical framework combining a discrete-state stochastic (chemical-kinetic) model and MD simulations, explicitly calculating the base editing probability at both target base and bystanders. In our model, we include an important parameter, ΔE_m, the binding affinity between deaminases and ssDNA. This parameter could be modulated by introducing various mutations into BE and the value was measured through MD simulations. This framework helps establish a relationship between mutations and BE editing selectivity. We then propose a theoretical principle that BE selectivity is non-monotonically dependent on ΔE_m, which therefore must be properly modulated to obtain the highest BE selectivity. In addition, other relevant

kinetic parameters are included in the model, such as the binding rate between Cas9 and ssDNA and the deamination rate of BE, *etc*, allowing us to discuss how ΔE_m cooperatively interacts with those parameters to affect BEs editing selectivity. Our model successfully explained how mutations influence the editing selectivity of A3A-BE3. Finally, we were also able to design new mutations to further improve the selectivity of the A3G-BE system and verified the improved editing selectivity experimentally. Thus, the framework we propose opens multiple opportunities for future engineering of base editors using theory-driven methods.

Results

Kinetic model of base editing

We have developed a discrete-state stochastic model to describe the dynamics of target and bystander editing. This is a minimal chemical-kinetic approach that considers the most relevant features of base editing. For convenience, unless noted otherwise, we will use the A3A-BE3 editing the EGFP site 1 as an example. In this theoretical model (Fig. 1), it is assumed that the Cas9 domain of CBE can bind to ssDNA with a rate u_0 , initiating the base editing. Alternatively, the protein complex can go to an unproductive state where editing cannot take place, with a rate of u_4. Next, either the Cas9 domain dissociates from DNA with a rate w_0 or the target cytidine binds to the deaminase catalytic site with a rate u_1. Then the cytidine can either dissociate from the site with a rate w_1 without being edited, or it can be chemically transformed to uridine with a rate u_3 (state 5). Similarly, the bystander cytidine may bind to the deaminase with a rate u_2, and subsequently, it can either unbind with a rate w_2 without being edited, or it can be chemically transformed with a rate u_3 (state 6). After that, while Cas9 is still bound to DNA, the deaminase can continue editing other cytidines in this region with the same sequence of events (state 9-12). Instead, if Cas9 dissociates from DNA, uridine will be quickly transformed to thymidine through DNA repair (states 7, 13 or states 8, 14). This U-to-T editing decreases the re-binding rate of Cas9 to ssDNA (state 7→5) because the very fast endogenous DNA repair and replication machinery has changed the DNA sequence from G:C pair to A:T pair. In this case, it does not perfectly match the spacer sequence of sgRNA. Therefore, the rebinding rate is modeled as $m*u_0$ with $0 < m < 1$, where m reflects the lower rebinding ability of the BE complex. Note that the kinetic network in Fig. 1 is a minimal description of the complex chemical processes that take place during base editing.

To evaluate the dynamics of base editing, we explored the first-passage probabilities method successfully used in various problems in chemistry, physics and biology (15-18). In the case of EGFP site 1 editing by A3A-BE3 there are four possible products as shown in Fig. 1: CTC (state 1, failed editing), CTT (state 13, only the target base is edited), TTC (state 14, only the bystander is edited) and TTT (state 12, both the target base and the bystander are edited). The explicit solution for the probability of each product outcome is given below (see derivations in the SI Appendix):

$$P_{CTT} = P_5 \cdot \frac{u_4 w_0 (u_3+w_2)}{(u_2+w_0)(u_3+w_2)(u_4+mu_0)-u_2 w_2 (u_4+mu_0)-mu_0 w_0 (u_3+w_2)} \quad [1]$$

$$P_{TTC} = P_6 \cdot \frac{u_4 w_0 (u_3+w_1)}{(u_1+w_0)(u_3+w_1)(u_4+mu_0)-u_1 w_1 (u_4+mu_0)-mu_0 w_0 (u_3+w_1)} \quad [2]$$

$$P_{TTT} = P_5 \cdot \left[1 - \frac{u_4 w_0 (u_3+w_2)}{(u_2+w_0)(u_3+w_2)(u_4+mu_0)-u_2 w_2 (u_4+mu_0)-mu_0 w_0 (u_3+w_2)}\right] + P_6 \cdot \left[1 - \frac{u_4 w_0 (u_3+w_1)}{(u_1+w_0)(u_3+w_1)(u_4+mu_0)-u_1 w_1 (u_4+mu_0)-mu_0 w_0 (u_3+w_1)}\right] \quad [3]$$

P_{CTC} can be calculated as one minus the sum of the above three probabilities. In equations [1-3], P_5 and P_6 are two intermediate parameters satisfying:

$$P_5 = \frac{u_0 u_1 u_3 (u_3+w_2)}{(u_1+u_2+w_0)(u_3+w_1)(u_3+w_2)(u_0+u_4)-u_1 w_1 (u_3+w_2)(u_0+u_4)-u_2 w_2 (u_3+w_1)(u_0+u_4)-u_0 w_0 (u_3+w_1)(u_3+w_2)} \quad [4]$$

$$P_6 = \frac{u_0 u_1 u_3 (u_3+w_1)}{(u_1+u_2+w_0)(u_3+w_1)(u_3+w_2)(u_0+u_4)-u_1 w_1 (u_3+w_2)(u_0+u_4)-u_2 w_2 (u_3+w_1)(u_0+u_4)-u_0 w_0 (u_3+w_1)(u_3+w_2)} \quad [5]$$

In experiments, a common way to quantify editing efficiency is to measure the overall probability of editing the target cytidine, P_t (11, 12). To compare these predictions with experimental results, P_t was calculated as:

$$P_t = P_{CTT} + P_{TTT} \quad [6]$$

Similarly, the overall probability of editing the bystander cytidine, P_b, was calculated as:

$$P_b = P_{TTC} + P_{TTT} \quad [7]$$

Our goal was to parameterize the model by reproducing experimentally measured probabilities, P_t and P_b. Here, we assume that the binding between the cytidine (both target and bystander) and the deaminase is mainly a diffusion controlled process. Therefore, considering that target and bystander cytidine are chemically identical and are very close, we added an additional approximation:

$$u_2 = u_1 \quad [8]$$

$$w_2 = w_1 e^{\Delta\Delta E_0/kT} = w_1 e^{[\Delta E_0(bystander)-\Delta E_0(target)]/k_B T} \quad [9]$$

The physical meaning of these expressions is the following: the binding rate to the target or the bystander takes place at the same rate, but the unbinding is governed by the strength of the interactions between the DNA substrate and the protein complex. In eqn. [9], the term ΔE_0 represents the binding free energy between the ssDNA binding motif and the deaminase. $\Delta\Delta E_0$ represents the difference in ΔE_0 between target editing and bystander editing. This difference arises from the sequence shift in the binding interface. An example is shown in Fig. 1, where the sequence of ssDNA binding motif changes from "**T**$_{-1}$C$_0$" in the case of target editing, to "**G**$_{-1}$C$_0$" in the case of bystander editing. This change can be formalized by a mutation from thymine to guanine at position -1, which perturbs the binding free energy and further influences the unbinding rate *w* of the cytidine from the catalytic site. Note that this approximation can also be explained using thermodynamic arguments, since the ratio between rates of binding and unbinding is related to the free energy difference between two states: the state where the protein-RNA complex is bound to the DNA chain and the state where both DNA and protein complex are free, $\frac{u_1}{w_1} = e^{-\Delta E_0(target)/k_B T}$, $\frac{u_2}{w_2} = e^{-\Delta E_0(bystander)/k_B T}$. Using eqn. [8] one can derive the result in eqn. [9].

Similarly, any deaminase mutation can be represented as a perturbation in binding free energy relative to the wild type,

$$w_{1,mutation} = w_{1,WT} e^{\Delta\Delta E_m/k_B T} = w_{1,WT} e^{[\Delta E_0(mutation)-\Delta E_0(WT)]/k_B T} \quad [10]$$

$\Delta\Delta E_m$ represents the difference in free energy due to mutations.

Substituting eqns. [8-10] into eqns. [6-7], we obtain:

$$P_t = \frac{\left(\gamma_1+m+\gamma_1\gamma_3+\gamma_1\gamma_2\gamma_3 e^{\frac{\Delta\Delta E_m}{k_B T}}\right)\left(1+\gamma_2 e^{\frac{\Delta\Delta E_0+\Delta\Delta E_m}{k_B T}}\right)+(\gamma_1+m)(1+\gamma_2 e^{\frac{\Delta\Delta E_m}{k_B T}})}{\left(\gamma_1+m+\gamma_1\gamma_3+\gamma_1\gamma_2\gamma_3 e^{\frac{\Delta\Delta E_m}{k_B T}}\right)\cdot[(2+2\gamma_1+\gamma_1\gamma_3)\left(1+\gamma_2 e^{\frac{\Delta\Delta E_m}{k_B T}}\right)\left(1+\gamma_2 e^{\frac{\Delta\Delta E_0+\Delta\Delta E_m}{k_B T}}\right)-\gamma_2 e^{\frac{\Delta\Delta E_m}{k_B T}}(1+\gamma_1)\left(1+2\gamma_2 e^{\frac{\Delta\Delta E_0+\Delta\Delta E_m}{k_B T}}+e^{\frac{\Delta\Delta E_0}{k_B T}}\right)]} \quad [11]$$

$$P_b = \frac{\left(\gamma_1+m+\gamma_1\gamma_3+\gamma_1\gamma_2\gamma_3 e^{\frac{\Delta\Delta E_0+\Delta\Delta E_m}{k_BT}}\right)\left(1+\gamma_2 e^{\frac{\Delta\Delta E_m}{k_BT}}\right)+(\gamma_1+m)(1+\gamma_2 e^{\frac{\Delta\Delta E_0+\Delta\Delta E_m}{k_BT}})}{\left(\gamma_1+m+\gamma_1\gamma_3+\gamma_1\gamma_2\gamma_3 e^{\frac{\Delta\Delta E_0+\Delta\Delta E_m}{k_BT}}\right)\cdot[(2+2\gamma_1+\gamma_1\gamma_3)\left(1+\gamma_2 e^{\frac{\Delta\Delta E_m}{k_BT}}\right)\left(1+\gamma_2 e^{\frac{\Delta\Delta E_0+\Delta\Delta E_m}{k_BT}}\right)-\gamma_2 e^{\frac{\Delta\Delta E_m}{k_BT}}(1+\gamma_1)\left(1+2\gamma_2 e^{\frac{\Delta\Delta E_0+\Delta\Delta E_m}{k_BT}}+e^{\frac{\Delta\Delta E_0}{k_BT}}\right)]}$$

[12]

$$\gamma_1 = u_4/u_0 \quad [13]$$

$$\gamma_2 = w_{1,WT}/u_3 \quad [14]$$

$$\gamma_3 = w_0/u_1 \quad [15]$$

Eqns. [11-15] give the full analytical expression that can be used to calculate editing probability. There are six free parameters to describe the base editing process (γ_1, γ_2, γ_3, m, $\Delta\Delta E_0$, $\Delta\Delta E_m$) but this number can be reduced using additional information. For example, previous binding experiments (19) have indicated that A3A binds to ssDNA with $K_d = 57\mu M, K_M = 62\mu M$ and $k_{cat} = 1.1/s$. From these values one can infer that $w_{1,WT} = 12.54/s$ and $u_3 = 1.1/s$. Therefore, after the cytidine binds to the catalytic site, the relative probability between unbinding and the chemical transformation step, γ_2, is 11.4. Next, if the changed ssDNA sequence no longer perfectly matches the sgRNA sequence, we assume that successful editing prevents rebinding of Cas9 to ssDNA, therefore m=0. Nevertheless, we show below that this assumption only has a minor effect on the final results. Lastly, we performed all-atom computational simulations to estimate $\Delta\Delta E_0$ and $\Delta\Delta E_m$, as shown in the next section. As a result, only two free parameters remain in the model, γ_1 and γ_3 (eqns. 13 and 15), both of which were parameterized by reproducing experimental values of P_t and P_b .

Computational estimates of binding free energy changes

We chose four CBEs developed by the Joung group (11): A3A(S99A), A3A(Y130F), A3A(N57Q), and A3A(N57A) to calculate the binding free energy changes between ssDNA and A3A. These CBE variants reduce the bystander effect to different extents while maintaining a high probability of on-target editing. The binding interface in the wild type A3A-ssDNA binding complex is shown in the crystal structure (PDB ID: 5KEG) (Fig. 2A). The carbonyl oxygen of Ser99 forms a hydrogen bond with the N4 atom of the cytidine in the catalytic site (dC_0). The hydroxyl group of Tyr130 forms a hydrogen bond with the 5'-phosphate of dC_0. Lastly, the nitrogen atom in the sidechain of Asp57 forms a hydrogen bond with the O3' atom of dC_0. Therefore, all four CBE variants appear to destabilize the binding between A3A and ssDNA ($\Delta\Delta E_m$ >0) by breaking this hydrogen-bonding network. In addition, since A3A recognizes the $\underline{\mathbf{T}}_{-1}C_0$ instead of the $\underline{\mathbf{G}}_{-1}C_0$ motif, the binding free energy to the deaminase should be higher (more repulsive) for the bystander cytidine than for the target cytidine ($\Delta\Delta E_0$ >0). To quantitatively calculate $\Delta\Delta E_0$ and $\Delta\Delta E_m$, we utilized the so-called "alchemical free-energy calculations" based on molecular dynamic simulations (20, 21). A thermodynamic cycle was constructed to convert $\Delta\Delta E_0$ and $\Delta\Delta E_m$ (Fig. 2B, $\Delta G_3 - \Delta G_1$) to the difference between two slow alchemical transitions (Fig. 2B, $\Delta G_2 - \Delta G_4$). One transition is the free energy change for the A3A-ssDNA complex due to mutations (Fig. 2B, ΔG_2) whereas the other is the free energy change for A3A alone due to mutations (Fig. 2B, ΔG_4). Calculated values indeed show that mutations cause an apparent increase in the deaminase-ssDNA binding free energy (Fig. 2C), consistent with predictions based on the structural data.

The rationale for A3A mutants that reduce the bystander effect

To check whether this model can reproduce the experimentally measured on-target and bystander editing probability, we substituted $\Delta\Delta E_0$ and $\Delta\Delta E_m$ calculated above into eqns. [11-15] and adjusted γ_1 and γ_3. The resulting theoretical prediction is in very good agreement with the experimental measurements (Fig. 3A), with values $\gamma_1 = \frac{u_4}{u_0} = 2.4$ and $\gamma_3 = \frac{w_0}{u_1} = 9.56 * 10^{-6}$. The $\gamma_1$value indicates that there is a significant fraction of BEs failing to initiate editing, whereas γ_3 indicates that the residence time of Cas9 on ssDNA is sufficient for the deaminase to function. We note here that the choice of *m*, which quantifies the effect of sgRNA mismatch on the rebinding rate of Cas9 and ssDNA, does not significantly affect the result (Fig. S1).

This theoretical model can be used to explain why the single mutation N57G greatly improves the editing selectivity of A3A-BE3. First, the ratio between the probabilities of having the target cytidine edited before the bystander (Fig. 1, transition state $2 \to 3 \to 5 \to \cdots$) and that of the reversed events (Fig. 1, transition state $2 \to 4 \to 6 \to \cdots$) can be calculated as:

$$\frac{P(state\ 2\to3\to\cdots)}{P(state\ 2\to4\to\cdots)} = \frac{\frac{u_3}{u_3+w_1}}{\frac{u_3}{u_3+w_2}} = \frac{1+\gamma_2 e^{\frac{\Delta\Delta E_0+\Delta\Delta E_m}{k_BT}}}{1+\gamma_2 e^{\frac{\Delta\Delta E_m}{k_BT}}} \qquad [16]$$

with $\gamma_2 = 11.4$. In the case of A3A, the ratio can be approximated as $e^{\frac{\Delta\Delta E_0}{k_BT}}$. As A3A significantly prefers the TC motif to the GC motif ($\Delta\Delta E_0 \sim 6\ k_BT$) this ratio is larger than 400. As a result, for both A3A(WT) and A3A(N57G), the probability of having only the bystander edited is very low (Fig. 3B, blue line, almost zero). After the target cytidine is edited (Fig. 1, state 5), the system has the choice of getting released with the product CTT (Fig. 1, state 13) or to continue editing the bystander, leading to the product TTT (Fig. 1, state 12). The outcome is largely influenced by the ratio between w_2 and u_3. If w_2 is significantly larger than u_3, bystander editing will be blocked because the residence time for the bystander cytidine in the catalytic site is too short to complete the transition to thymidine. Analytically, after the target cytidine gets edited, the probability ratio between CTT and TTT can be calculated as:

$$\frac{P(state\ 5\to13)}{P(state\ 5\to12)} = \frac{w_0}{u_2}\left(1+\frac{w_2}{u_3}\right) = \gamma_3\left(1+\gamma_2 e^{\frac{\Delta\Delta E_0+\Delta\Delta E_m}{k_BT}}\right) \qquad [17]$$

For A3A(WT), this ratio is 0.056, meaning that the dominant product is TTT (Fig. 3B, purple square at $\Delta\Delta E_m = 0$). This explains that for the wild-type A3A the editing efficiency of the target cytidine is similar to that of the bystander. In sharp comparison, as $\Delta\Delta E_m$ increases by 5.8 k_BT for A3A(N57G), the dominant edited product changes to CTT (Fig. 3B, green line at $\Delta\Delta E_m$= 5.8 k_BT). In this case, A3A(N57G) minimizes the bystander effect while maintaining a high probability of editing the target base.

Similar arguments can be presented for the other three A3A mutants (S99A, Y130F, and N57Q). However, it is critical to note that to gain high editing selectivity, mutated residues have a non-monotonic effect in the deaminase-ssDNA interface. Here, selectivity is defined as the difference in probabilities between editing the target and editing the bystander. Weakening the binding interface up to 5-7 k_BT (depending on the system) greatly improves selectivity (Fig. 3A), but selectivity drops when $\Delta\Delta E_m$ continues increasing. This result can be explained using physical considerations. Increasing $\Delta\Delta E_m$ leads to faster-unbinding rates between cytidine and deaminase. At moderate values of $\Delta\Delta E_m$, target editing is less affected (Fig. 1, state 2→5) but bystander editing is blocked (Fig. 1, state 5→12). However, for very large values of $\Delta\Delta E_m$, both editing pathways are essentially blocked and the system prefers to go into the inactive state (Fig. 1, state 1). Therefore, proper modulation of the binding interface is the key to optimize base editing selectivity. We further prove this point in the next section.

Optimizing CTD A3G system

In this section, we employ our theoretical method to analyze and optimize the editing of *EMX1* site 1 by another BE, CTD A3G (the C-terminal domain of A3G). The target cytidine is C_6 and the bystander is C_5 (see the definitions of C_5 and C_6 in Fig. 4). Our goal is to increase the editing probability at C_6 while keeping that at C_5 as low as possible. Theoretical calculations (Fig. 4B) show that this can be done by properly decreasing $\Delta\Delta E_m$, i.e., stabilizing the binding interface between CTD A3G and ssDNA. In our previous work (12) we designed three mutation sets for A3G for this purpose. However, although mutation sets A+B+C (Fig. 4A, the pink bar) improved the editing probability at C_6, they also significantly increased the editing probability at the bystander C_5 from 8% to 41%. Based on the theoretical calculation, this is due to excessive reduction of $\Delta\Delta E_m$ (Fig. 4B). A simple solution is to slightly increase $\Delta\Delta E_m$ by removing mutation set A. The latter includes P247K and Q318K, which overly attract ssDNA. This has been verified experimentally: CTD A3G with mutation sets B+C (Fig. 4B, the blue bar) increased the editing probability at C_6 from 38.9% to 44.4% while the editing probability at C_5 is still low. These results support the validity of our theoretical method.

Discussion

In this work, we have developed a theoretical framework to understand the process of base editing. The presented approach suggests several general rules to design BEs with improved editing selectivity. This goal is fulfilled by modulating the binding affinity between deaminase and ssDNA using mutagenesis ($\Delta\Delta E_m$). The principle is to guarantee that the residence time of deaminase on ssDNA is sufficiently long to complete the editing of the first on-target site, while being too short for editing the second (bystander) site. Our theoretical method predicts optimal values for $\Delta\Delta E_m$. Away from these optimal values, selectivity decreases. Therefore, instead of testing experimentally a set of candidate BE mutants, one can instead set up a computational pre-screening process by estimating the $\Delta\Delta E_m$ of those variants, and only candidates near the optimal value can then be tested experimentally. Herein, we used alchemical free-energy calculations to estimate $\Delta\Delta E_m$. The accuracy of this method has been validated in the A3A and A3G system (Figs. 3 and 4). Future work will develop carefully parameterized scoring functions for ssDNA-protein interactions, so that the prediction of $\Delta\Delta E_m$ is accelerated. In addition, while our model is built based on the chemical and physical nature of editing systems, an alternative model (22) based on machine learning (ML) techniques has been proposed to predict target and bystander editing from sequence information. Combining these methods to a physics-constrained ML may improve the prediction ability.

Eqns. [11-15] indicate that for a given system the editing probability is regulated by two parameters, γ_1 and γ_3, in addition to $\Delta\Delta E_m$. Therefore, we plotted the editing probability for different values of γ_1 (Fig. 5A) and γ_3 (Fig. 5B). We first reduced the parameter γ_1 (Fig. 5A) which can be achieved by increasing the on-rate of Cas9 to ssDNA. It turns out that the editing selectivity for the WT system is not affected by γ_1 (Fig. 5A, solid blue line vs dashed blue line at $\Delta\Delta E_m = 0$) as the efficiencies of both target and bystander editing increase synchronously. However, the selectivity greatly improves when $\Delta\Delta E_m$ is 5-7 k_BT, meaning that γ_1 amplifies the deaminase mutation regulation effect. This suggests an effective combination strategy in the design of highly selective BE: optimization of $\Delta\Delta E_m$ first, then reducing γ_1 to amplify this effect. We then reduced the parameter γ_3 (Fig. 5B). Our calculation indicates that reducing γ_3 does not change the maximum editing selectivity but induces a right shift in the editing profile, i.e., a larger $\Delta\Delta E_m$ value is required to achieve the maximum editing selectivity.

Another interesting question is whether a general mutation to all BEs homologs exists that optimizes editing. Unfortunately, we found that a mutation working perfectly for one type

of deaminase may fail for another type, even when they are homologs. For example, A3A mutations N57G and Y315 greatly reduce the bystander effect while maintaining a high probability of target editing, but A3G N244G almost loses the base editing ability (A3A N57 aligns with A3G N244) (Fig. S2A). The theoretical model developed above explains this negative outcome, since $\Delta\Delta E_m$ for N244G turns out to be much larger than its optimal value (Fig. S2B). Our model shows that an energy shift of only few k_BT can dramatically change the selectivity of a BE, from its optimal value to zero. This is due to subtle differences between homologs: the on-rate of A3G binding to ssDNA is lower that for A3A, as the former can form a large dimer that interferes with binding. In addition, A3G has lower $\Delta\Delta E_0$ (see Eqn. [9]) because the sequence of the ssDNA binding motif changes from "**C**$_{-1}$C$_0$" in the target editing to "**T**$_{-1}$C$_0$" in the bystander editing. The energy perturbation from **C**$_{-1}$ to **T**$_{-1}$ is smaller than that from **T**$_{-1}$ to **G**$_{-1}$ in the case of A3A. These differences cause A3A and A3G to have distinct behavior under equivalent mutations. In fact, even for the same BE, editing different loci can change the binding free energy by a few k_BT because at different loci the neighboring bases of the bystander may vary. Therefore, each BE may require a unique optimization to achieve high editing selectivity. Nevertheless, our approach gives quantifiable parameters that can be used to accelerate the search for best editors. In summary, a general design strategy would be a) employing the chemical-kinetic model to determine the binding free energy changes required to achieve the maximum editing selectivity, ΔE_{peak}; b) designing mutations in the deaminase, estimating $\Delta\Delta E_m$ and selecting the ones near ΔE_{peak}; c) keeping mutations picked in the previous step and designing extra mutations that increase the binding rate of Cas9 to DNA substrate; d) experimental validation of these changes.

Method

Free Energy Calculations by Molecular Dynamic Simulations

We utilized molecular dynamics based on alchemical free-energy calculations (20, 21) (Fig. 2) to estimate the binding free energy changes under various mutations (11). All simulations were carried out using the Gromacs package (23). Amber99sb*ILDN force field was used (24). The integration time step was set to 2 fs. The initial states of the A3A-ssDNA (25) and the A3G-ssDNA (26) binding complex were taken from their crystal structures (PDB ID: 5keg and 6bux, respectively). Then we used the pmx webserver (27, 28) to generate hybrid structures and topologies representing mutations. Each system was solvated in a cubic box with TIP3P water molecules and 0.1M NaCl. The dimension of the box is 9 nm. The temperature was maintained at 300K by the Berendsen thermostat (29) while the pressure was maintained at 1.0 atm by using the Parrinello-Rahman barostat (30). Electrostatic interactions were calculated by the Particle Mesh Ewald method (31). The soft-core function was used for the nonbonded interactions during the alchemical transitions (32). For each system, energy minimization was first performed, followed by 1-ns NVT and 1-ns NPT equilibration with the protein configuration restrained. Then the system was further equilibrated for 5 ns without any restrain. The last snapshot of the trajectory served as the starting configuration for the following alchemical transitions.

The alchemical transition ($\lambda = 0 \rightarrow \lambda = 1$) was divided into 21 consecutive windows with the bin size of 0.05. For each window i, λ was first increased from 0 to λ_i ($\lambda_i = 0, 0.05, 0.1, 0.15 \ldots$) with a slow rate 10^{-8} /step, then was fixed to λ_i for 40-ns production run. dH/dl values were recorded every 100 steps. The free energy and error bar were estimated by Bennett's acceptance ratio method (33).

Experiment

Mammalian cell culture

HEK293T cells (American Type Culture Collection, CRL-3216) were cultured in GlutaMAX™ high-glucose Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 100 U/mL penicillin-streptomycin. Cells were maintained in a humidity atmosphere at 37°C with 5% CO_2 and passaged at a ratio of 1:4 when reaching 90% confluency by using TrypLE Express. Above mentioned cell, culture reagents are all from Thermo Fisher. Mycoplasma testing was performed monthly using a mycoplasma PCR detection kit (abm).

Plasmid construction

The full-length human codon-optimized wild-type A3G with setA mutations (P200A + N236A + P247K + Q318K + Q322K) were synthesized as gBlock (IDT) and inserted into the BE4max construct (Addgene #112093) to replace the rAPOBEC1 region, resulting in hA3G-BE. To do so, both insertion and vectors were amplified using primers with overhangs containing Esp3I recognition sites, which would generate compatible complementary sticky ends after cutting. Then the one-pot Golden Gate assembly was employed to cut and ligate two amplified pieces by using Esp3I and T4 DNA ligase (NEB). Likewise, the Y315F and N244G variants were respectively generated by designing an extra pair of primers containing the indicated mutations and performing a 3-piece assembly. As previously described (12), set A, set B (H248N + K249L + H250L + G251C + F252G + L253F + E254Y), and set C (L234K + F310K + C243A + C321A + C356A) mutations were introduced to the C-terminal domain (CTD) of A3G (previously known as A3G-BE4.4) (12) to generate CTD A3G (Set A+B+C)-BE (previously A3G-BE5.5), and CTD A3G (Set B+C)-BE.

HEK293T transfection, genomic DNA extraction, amplicon sequencing, and analysis

Cell transfection was performed as previously described with slight modifications (12). Briefly, HEK293T cells were seeded into a poly-D-lysine–coated 48-well plate (Corning) at a density of 4.5×10^4 cells per well in 250 μL antibiotic-free culture medium supplemented with 10% FBS. In about 12 – 16 hours, upon reaching 70% confluency, cells of each well were transfected with 750 ng BE plasmids and 250 ng sgRNA plasmids using 1.5 μl Lipofectamine 2000 (Thermo Fisher Scientific) dispersed in 25 μL Opti-MEM according to the manufacturer's instructions. Three days later, removing the medium, washing the cells gently with PBS (Thermo Fisher Scientific), and lysing the cells at 37°C for 1–2 h with 100 μl of lysis buffer per well containing 10 mM Tris-HCl (pH 7.5), 0.05% SDS, and 25 μg/mL proteinase K (Fisher BioReagents). The cell lysates were then subjected to heat inactivation at 80°C for 0.5–1 h.

A total of 100 ng genomic DNA was then amplified at the *EMX1* target site by using primers attached with the partial Illumina adapters and 8-bp compatible and nucleotide-balanced indices on both 5' and 3' end. The forward and reverse primers are as follows. For: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNTGTGGTTCCAGAACCGGAG-3'; Rev: 5'-GACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNCTCTGCCCTCGTGGGTTT-3'. The protospacer sequence is 5'-GAGTCCGAGCAGAAGAAGAA-3'. The amplicon sequence is 5'-TGTGGTTCCAGAACCGGAGGACAAAGTACAAACGGCAGAAGCTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGAAGCAGGCCAATGGGGAGGACATCGATGTCACCTCCAATGACTAGGGTGGGCAACCACAAACCCACGAGGGCAGAG-3'.

Amplicons were pooled, column purified (Qiagen), recovered in nuclease-free water (Invitrogen), and quantified by the Qubit dsDNA HS assay (Life Technologies). A volume of 25 μL sample with the final concentration adjusted to 20 ng/μL was submitted for Amplicon-EZ sequencing (Genewiz). Fastq files were then downloaded from Genewiz Ftp server and analyzed by using CRISPResso2 (https://github.com/pinellolab/CRISPResso2) to align reads and quantify the base editing efficiency and frequency (34).

Statistical analysis

All experiments were performed with at least three independent biological replicates. Means, standard error of the mean, and p values were calculated by using GraphPad (Prism 8). P values were analyzed using two-tailed Student's t-test, with a statistical significance level denoted by ns (not significant), *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, and ****$p < 0.0001$.

Acknowledgements

QW acknowledges the funding support from "USTC Research Funds of the Double First-Class Initiative" (YD2030002006), and from the Center for Theoretical Biological Physics sponsored by the NSF (PHY-2019745). ABK acknowledges the support from the Welch Foundation (C-1559), and from the NSF (CHE-1953453 and MCB-1941106); XG acknowledges the funding support from the NIH (1R35GM138207) and the Rice University Creative Ventures Fund. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of the University of Science and Technology of China and the Supercomputing Center of Rice University.

Figure Legends

Figure 1. Chemical-kinetic model of A3A-BE3 editing the EGFP site 1. Deaminase, Cas9, target and bystander base are represented by blue, orange, red and yellow squares, respectively. "C" represents cytosine, "T" represents thymine and "X" represents uridine or thymine. Editing is modeled as a multiple-step chemical reaction where Cas9 first binds to ssDNA, cytidine then binds to the catalytic site of the deaminase and is chemically converted to thymidine. Here, u_j and w_j (j = 0 to 4) represent the chemical-kinetic rates for various transitions. The model has a total of 15 states and can produce four outcomes (dashed squares): CTC (failed editing), CTT (only the target base is edited), TTC (only the bystander is edited), TTT (both target and bystander bases are edited).

Figure 2. Calculation of binding free energy changes between ssDNA binding motif and A3A deaminase under various mutations. (A) Hydrogen bonding network between cytidine dC_0 and A3A residues 57, 99 and 130; (B) thermodynamic cycle used to calculate the binding free energy changes: ssDNA (black line), A3A (blue balloons). A mutation at the binding interface is shown by a transition from small orange circle to orange triangle; (C) computationally estimated changes in binding free energy for A3A deaminase mutations S99A, Y130F, N57Q, N57G and binding free energy change between A3A binding to target and bystander cytidines. The energy unit is k_BT and T = 300K.

Figure 3. Theoretical model of the A3A-BE3 editing system. (A) Comparison between theoretical calculations (solid lines) and experimental data (exp, circles). $\Delta\Delta E_m$ represents the perturbation in binding free energy due to the different mutations; k_BT is the unit of energy; P_t and P_b represent the overall probability of editing target and bystander cytidine, respectively; (B) calculated editing probabilities for products CTT, TTC and TTT. $P_t = P_{CTT} + P_{TTT}$; $P_b = P_{TTC} + P_{TTT}$.

Figure 4. Engineering of CTD A3G-BEs that edit the *EMX1* site 1. (A) Experimental measurements; (B) comparison between theoretical calculations (solid lines) and experimental measurement (circle). Please see the definition of set A, B and C in the method section.

Figure 5. Base editing pattern of A3A-BE3 regulated by (A) γ_1 and (B) γ_3. The definition of $\Delta\Delta E_m$, γ_1 and γ_3 can be found in eqns. [10,13-15]. P_t and P_b are the overall probabilities of editing the target and bystander cytidine, respectively. The difference between P_t and P_b is shown in blue. The setting with original parameters is represented by solid lines (case 1) whereas variants are represented by dashed lines (case 2: γ_1 divided by five; case 3: γ_3 divided by five).

Figure 1.

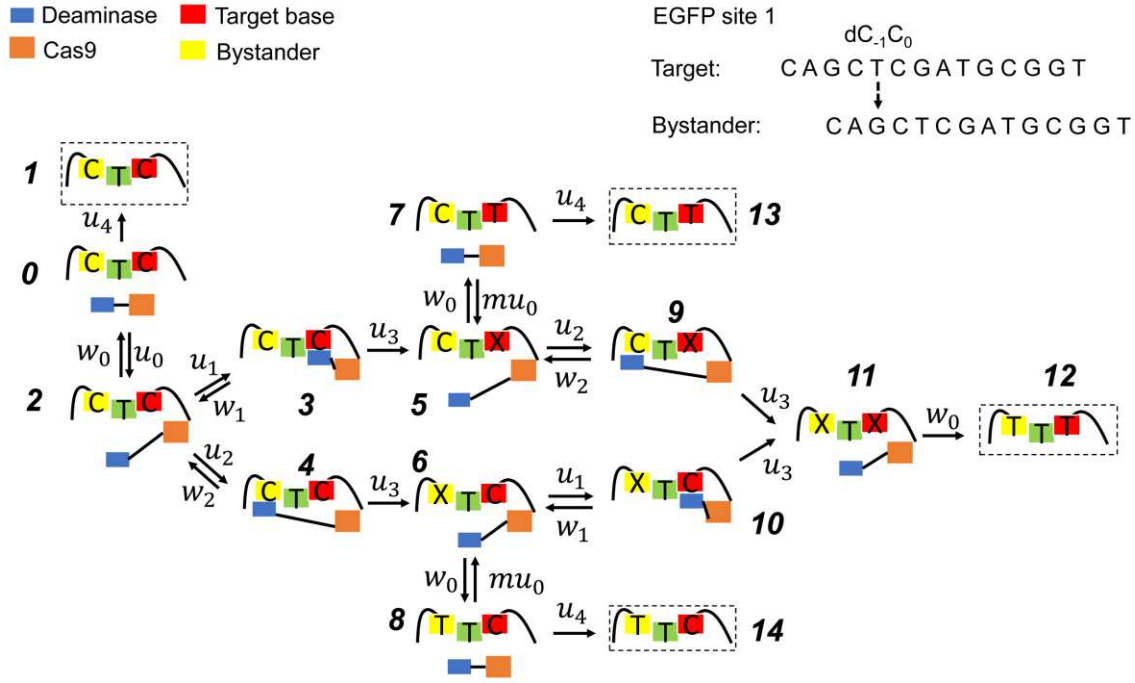

Deaminase
Target base
Cas9
Bystander
EGFP site 1
dC-1C0
Target: C A G C T C G A T G C G G T
Bystander: C A G C T C G A T G C G G T
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
u4
w0
u0
u1
w1
u2
w2
u3
mu0

Figure 2.

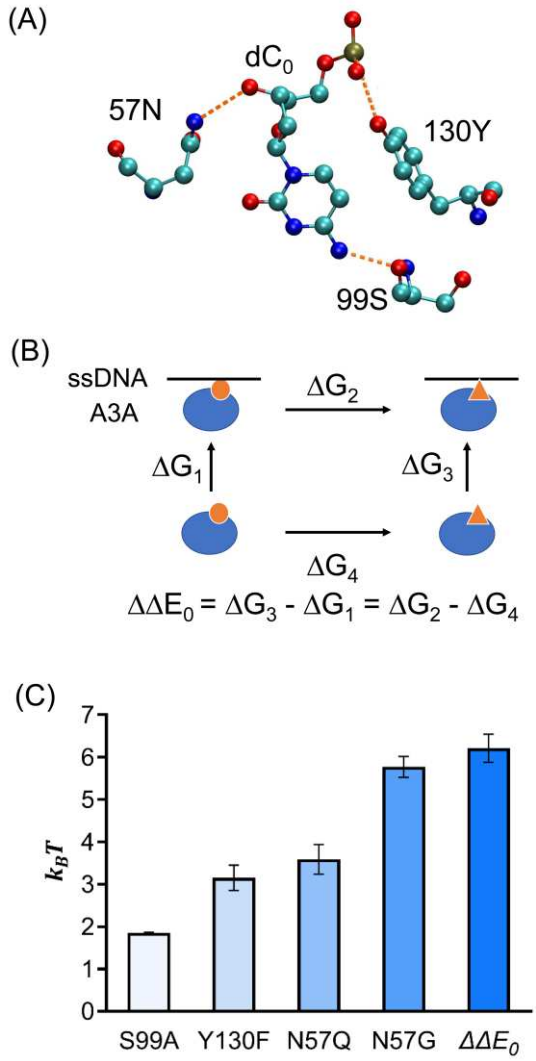

(A)
dC0
57N
130Y
99S
(B)
ssDNA
A3A
ΔG2
ΔG1
ΔG3
ΔG4
ΔΔE0 = ΔG3 - ΔG1 = ΔG2 - ΔG4
(C)
kBT
0
1
2
3
4
5
6
7
S99A
Y130F
N57Q
N57G
ΔΔE0

Figure 3.

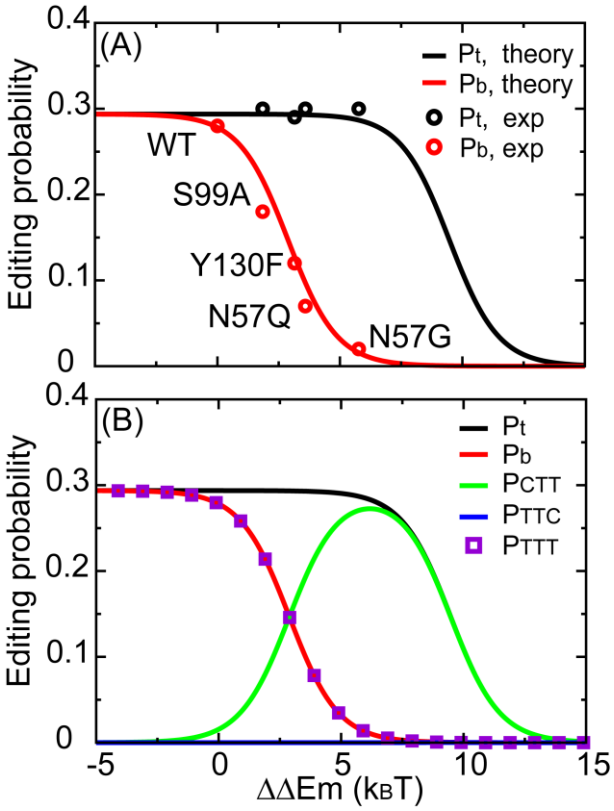

(A)
Pt, theory
Pb, theory
Pt, exp
Pb, exp
WT
S99A
Y130F
N57Q
N57G
Editing probability
0.4
0.3
0.2
0.1
0
(B)
Pt
Pb
PCTT
PTTC
PTTT
Editing probability
-5
0
5
10
15
ΔΔEm (kBT)

Figure 4.

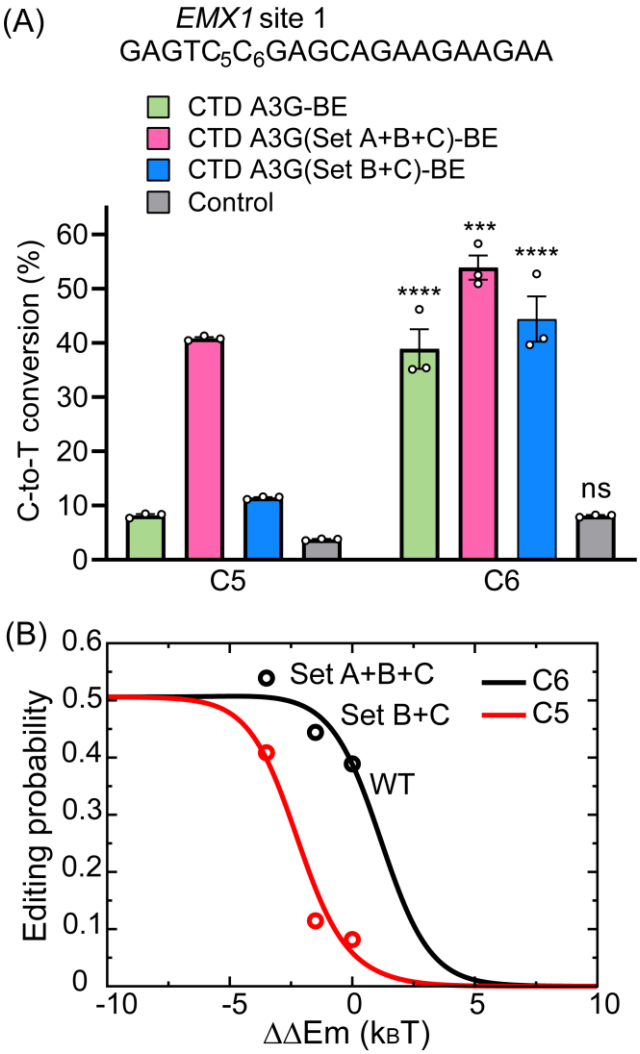
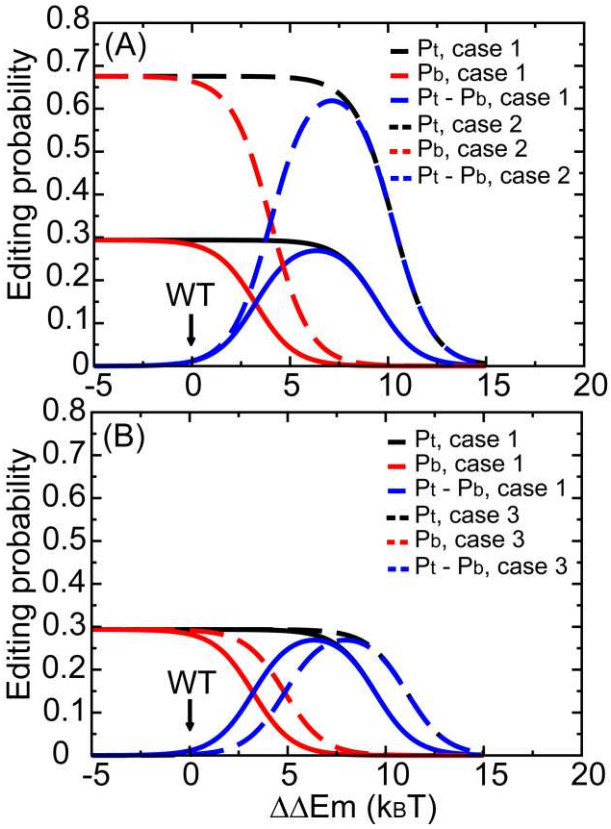

(A)
EMX1 site 1
GAGTC5C6GAGCAGAAGAAGAA
CTD A3G-BE
CTD A3G(Set A+B+C)-BE
CTD A3G(Set B+C)-BE
Control
C-to-T conversion (%)
0
10
20
30
40
50
60
****
***
****
ns
C5
C6
(B)
Editing probability
0
0.1
0.2
0.3
0.4
0.5
0.6
Set A+B+C
Set B+C
WT
C6
C5
-10
-5
0
5
10
ΔΔEm (kBT)

Figure 5.


(A)
Pt, case 1
Pb, case 1
Pt - Pb, case 1
Pt, case 2
Pb, case 2
Pt - Pb, case 2
WT
Editing probability
(B)
Pt, case 1
Pb, case 1
Pt - Pb, case 1
Pt, case 3
Pb, case 3
Pt - Pb, case 3
WT
Editing probability
ΔΔEm (kBT)

Reference

1. Hsu PD, Lander ES, & Zhang F (2014) Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell* 157(6):1262-1278.
2. Rees HA & Liu DR (2018) Base editing: precision chemistry on the genome and transcriptome of living cells. *Nature Reviews Genetics* 19(12):770-788.
3. Anzalone AV, Koblan LW, & Liu DR (2020) Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* 38(7):824-844.
4. Gaudelli NM*, et al.* (2017) Programmable base editing of A.T to G.C in genomic DNA without DNA cleavage. *Nature* 551(7681):464-471.
5. Komor AC, Kim YB, Packer MS, Zuris JA, & Liu DR (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533(7603):420-424.
6. Nishida K*, et al.* (2016) Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* 353(6305):8.
7. Komor AC, Badran AH, & Liu DR (2017) CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. *Cell* 168(1-2):20-36.
8. Komor AC*, et al.* (2017) Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T: A base editors with higher efficiency and product purity. *Sci. Adv.* 3(8):9.
9. Koblan LW*, et al.* (2018) Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* 36(9):843-846.
10. Kim YB*, et al.* (2017) Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat. Biotechnol.* 35(4):371-376.
11. Gehrke JM*, et al.* (2018) An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities. *Nat. Biotechnol.* 36(10):977-982.
12. Lee SS*, et al.* (2020) Single C-to-T substitution using engineered APOBEC3G-nCas9 base editors with minimum genome- and transcriptome-wide off-target effects. *Sci. Adv.* 6(29):12.
13. Rallapalli KL, Komor AC, & Paesani F (2020) Computer simulations explain mutation-induced effects on the DNA editing by adenine base editors. *Sci. Adv.* 6(10):11.
14. Rallapalli KL, Ranzou BL, Ganapathy KR, Komor AC, & Paesani F (2020) Retracing the evolutionary trajectory of adenine base editors using theoretical approaches. *bioRxiv preprint*.
15. Kolomeisky AB (2015) *Motor Proteins and Molecular Motors* (CRC Press).
16. Shvets AA & Kolomeisky AB (2017) Mechanism of Genome Interrogation: How CRISPR RNA-Guided Cas9 Proteins Locate Specific Targets on DNA. *Biophysical Journal* 113(7):1416-1424.
17. Wang Q*, et al.* (2017) Molecular origin of the weak susceptibility of kinesin velocity to loads and its relation to the collective behavior of kinesins. *Proceedings of the National Academy of Sciences of the United States of America* 114(41):E8611-E8617.
18. Kampen NGV (2007) *Stochastic Processes in Physics and Chemistry* (North Holland).
19. Byeon I-JL*, et al.* (2013) NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. *Nature Communications* 4:1890.
20. Gapsys V, Michielssens S, Seeliger D, & de Groot BL (2016) Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. *Angewandte Chemie-International Edition* 55(26):7364-7368.

21. Gapsys V & de Groot BL (2017) Alchemical Free Energy Calculations for Nucleotide Mutations in Protein-DNA Complexes. *Journal of Chemical Theory and Computation* 13(12):6275-6289.
22. Arbab M, *et al.* (2020) Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning. *Cell* 182(2):463-480.
23. Pronk S, *et al.* (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29(7):845-854.
24. Lindorff-Larsen K, *et al.* (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins-Structure Function and Bioinformatics* 78(8):1950-1958.
25. Kouno T, *et al.* (2017) Crystal structure of APOBEC3A bound to single-stranded DNA reveals structural basis for cytidine deamination and specificity. *Nature Communications* 8:8.
26. Maiti A, *et al.* (2018) Crystal structure of the catalytic domain of HIV-1 restriction factor APOBEC3G in complex with ssDNA. *Nature Communications* 9:11.
27. Gapsyse V & de Groot BL (2017) pmx Webserver: A User Friendly Interface for Alchemistry. *Journal of Chemical Information and Modeling* 57(2):109-114.
28. Gapsys V, Michielssens S, Seeliger D, & de Groot BL (2015) pmx: Automated Protein Structure and Topology Generation for Alchemical Perturbations. *Journal of Computational Chemistry* 36(5):348-354.
29. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, & Haak JR (1984) Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* 81(8):3684-3690.
30. Parrinello M & Rahman A (1981) Polymorphic transitions in single-crystals - a new molecular-dynamics method. *J. Appl. Phys.* 52(12):7182-7190.
31. Essmann U, *et al.* (1995) A smooth particle mesh ewald method. *J. Chem. Phys.* 103(19):8577-8593.
32. Gapsys V, Seeliger D, & de Groot BL (2012) New Soft-Core Potential Function for Molecular Dynamics Based Alchemical Free Energy Calculations. *Journal of Chemical Theory and Computation* 8(7):2373-2382.
33. Bennett CH (1976) Efficient estimation of free-energy differences from monte-carlo data. *J. Comput. Phys.* 22(2):245-268.
34. Clement K, *et al.* (2019) CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* 37(3):224-226.

SI Appendix

To obtain the analytical solution of the editing probability for the kinetic model in Fig. 1, we introduced a first-passage probability density function $F_i(t)$, representing the probability to finish the editing process at time *t*, starting from the state *i*. Then the editing probability can be calculated as

$$P = \int_0^\infty F_0(t)dt \quad [S1]$$

The temporal evolution of $F_i(t)$ is controlled by the backward master equations:

$$\frac{dF_0(t)}{dt} = u_0F_2(t) + u_4F_1(t) - (u_0 + u_4)F_0(t) \quad [S2]$$

$$\frac{dF_2(t)}{dt} = u_1F_3(t) + u_2F_4(t) + w_0F_0(t) - (u_1 + u_2 + w_0)F_2(t) \quad [S3]$$

$$\frac{dF_3(t)}{dt} = u_3F_5(t) + w_1F_2(t) - (u_3 + w_1)F_3(t) \quad [S4]$$

$$\frac{dF_4(t)}{dt} = u_3F_6(t) + w_2F_2(t) - (u_3 + w_2)F_4(t) \quad [S5]$$

$$\frac{dF_5(t)}{dt} = u_2F_9(t) + w_0F_7(t) - (u_2 + w_0)F_5(t) \quad [S6]$$

$$\frac{dF_6(t)}{dt} = u_1F_{10}(t) + w_0F_8(t) - (u_1 + w_0)F_6(t) \quad [S7]$$

$$\frac{dF_7(t)}{dt} = u_4F_{13}(t) + mu_0F_5(t) - (u_4 + mu_0)F_7(t) \quad [S8]$$

$$\frac{dF_8(t)}{dt} = u_4F_{14}(t) + mu_0F_6(t) - (u_4 + mu_0)F_8(t) \quad [S9]$$

$$\frac{dF_9(t)}{dt} = u_3F_{11}(t) + w_2F_5(t) - (u_3 + w_2)F_9(t) \quad [S10]$$

$$\frac{dF_{10}(t)}{dt} = u_3F_{11}(t) + w_1F_6(t) - (u_3 + w_1)F_{10}(t) \quad [S11]$$

$$\frac{dF_{11}(t)}{dt} = w_0F_{12}(t) - w_0F_{11}(t) \quad [S12]$$

Eqns. [S1-S12] can be solved by utilizing Laplace transformations:

$$F_i(s) = \int_0^\infty F_i(t)e^{-st}dt \quad [S13]$$

Then eqns. [S1-S12] can be transformed to linear equations:

$$(s + u_0 + u_4)F_0(s) = u_0F_2(s) + u_4F_1(s) \quad [S14]$$

$$(s + u_1 + u_2 + w_0)F_2(s) = u_1F_3(s) + u_2F_4(s) + w_0F_0(s) \quad [S15]$$

$$(s + u_3 + w_1)F_3(s) = u_3F_5(s) + w_1F_2(s) \quad [S16]$$

$$(s + u_3 + w_2)F_4(s) = u_3F_6(s) + w_2F_2(s) \quad [S17]$$

$$(s + u_2 + w_0)F_5(s) = u_2F_9(s) + w_0F_7(s) \quad [S18]$$

$$(s + u_1 + w_0)F_6(s) = u_1F_{10}(s) + w_0F_8(s) \quad [S19]$$

$$(s + u_4 + mu_0)F_7(s) = u_4F_{13}(s) + mu_0F_5(s) \quad [S20]$$

$$(s + u_4 + mu_0)F_8(s) = u_4F_{14}(s) + mu_0F_6(s) \quad [S21]$$

$$(s + u_3 + w_2)F_9(s) = u_3F_{11}(s) + w_2F_5(s) \quad [S22]$$

$$(s + u_3 + w_1)F_{10}(s) = u_3F_{11}(s) + w_1F_6(s) \quad [S23]$$

$$(s + w_0)F_{11}(s) = w_0F_{12}(s) \quad [S24]$$

In addition, the editing probability can be written as:

$$P = F_0(s)|_{s=0} \quad [S25]$$

Solving the editing probability at different states requires different boundary conditions. As detailed in the main text, there are a total of four editing possibilities: CTC (Fig. 1, state 1, failed editing), CTT (state 13, only the target base is edited), TTC (state 14, only the is bystander edited) and TTT (state 12, both the target and the bystander base are edited).

For CTC, the boundary conditions are:

$$F_1(s) = 1;\ F_{12}(s) = 0;\ F_{13}(s) = 0;\ F_{14}(s) = 0 \quad [S26]$$

For CTT:

$$F_1(s) = 0;\ F_{12}(s) = 0;\ F_{13}(s) = 1;\ F_{14}(s) = 0 \quad [S27]$$

For TTC:

$$F_1(s) = 0;\ F_{12}(s) = 0;\ F_{13}(s) = 0;\ F_{14}(s) = 1 \quad [S28]$$

For TTT:

$$F_1(s) = 0;\ F_{12}(s) = 1;\ F_{13}(s) = 0;\ F_{14}(s) = 0 \quad [S29]$$

Solving P from eqns. [S14-S29] gives the analytical solutions of the editing probability (eqns. [1-5] in the main text).

We note that due to lack of experimental data, our model simplifies the influence from residue mutations. Strictly speaking, the energy perturbation due to mutations also mildly influences the on-rate besides the off-rate. So that eqns. [8-9] in the main text generally should be written as

$$u_2 = u_1 e^{\theta \cdot \Delta\Delta E_0/kT} \quad [S30]$$

$$w_2 = w_1 e^{(1-\theta) \cdot \Delta\Delta E_0/kT} \quad [S31]$$

while θ is a distribution factor ($0 < \theta < 1$). However, this approximation will not qualitatively change the main conclusions of this work. A more detailed theoretical model can be easily implemented when more experimental data are available. In addition, it is important to add that our theoretical method could also evaluate the reaction times for the editing, which can be viewed as mean first-passage times in our approach. The editing time can be calculated as:

$$T = -dF_0(s)/ds|_{s=0}/P \quad [S32]$$

With future advances in experimental methods this might also serve as another way of determining the parameters of the system.

Figure S1. Calculated editing probability is affected by parameter m ($0<m<1$). $m = 0$ means that succesful editing abolishes re-binding of Cas9 to ssDNA due to sgRNA mismatch. $m = 1$ means that succesful editing has no effect on re-binding of Cas9. $\Delta\Delta E_m$ represents the binding free energy perturbations due to different mutations. The unit of energy is k_BT. P_t *and* P_b are the overall probabilities of editing the target and bystander cytidine, respectively.

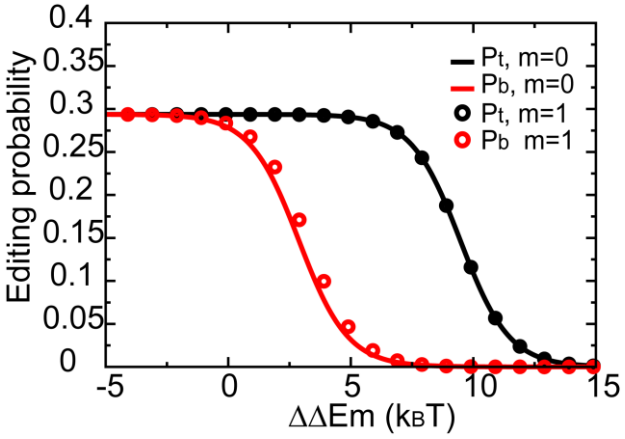
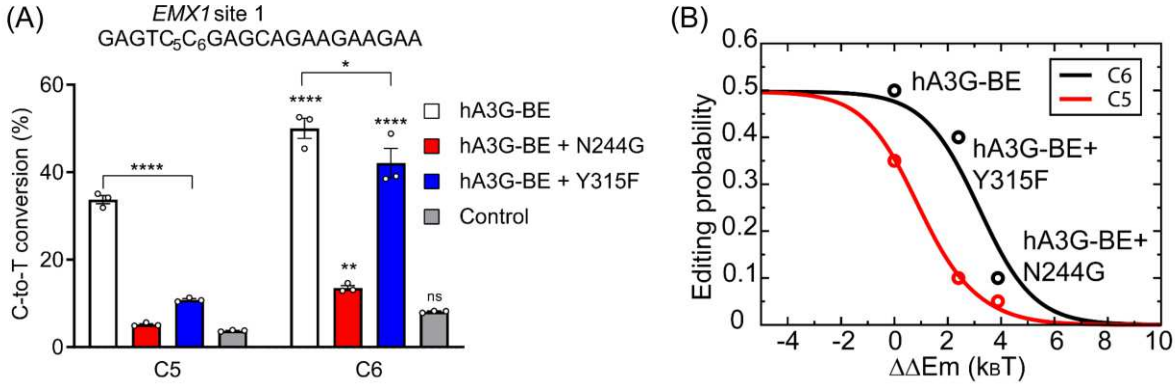


Figure S2. A3G mutants editing the *EMX1* site 1. (A) Experimental measurements and (B) theoretical calculations.



Figures

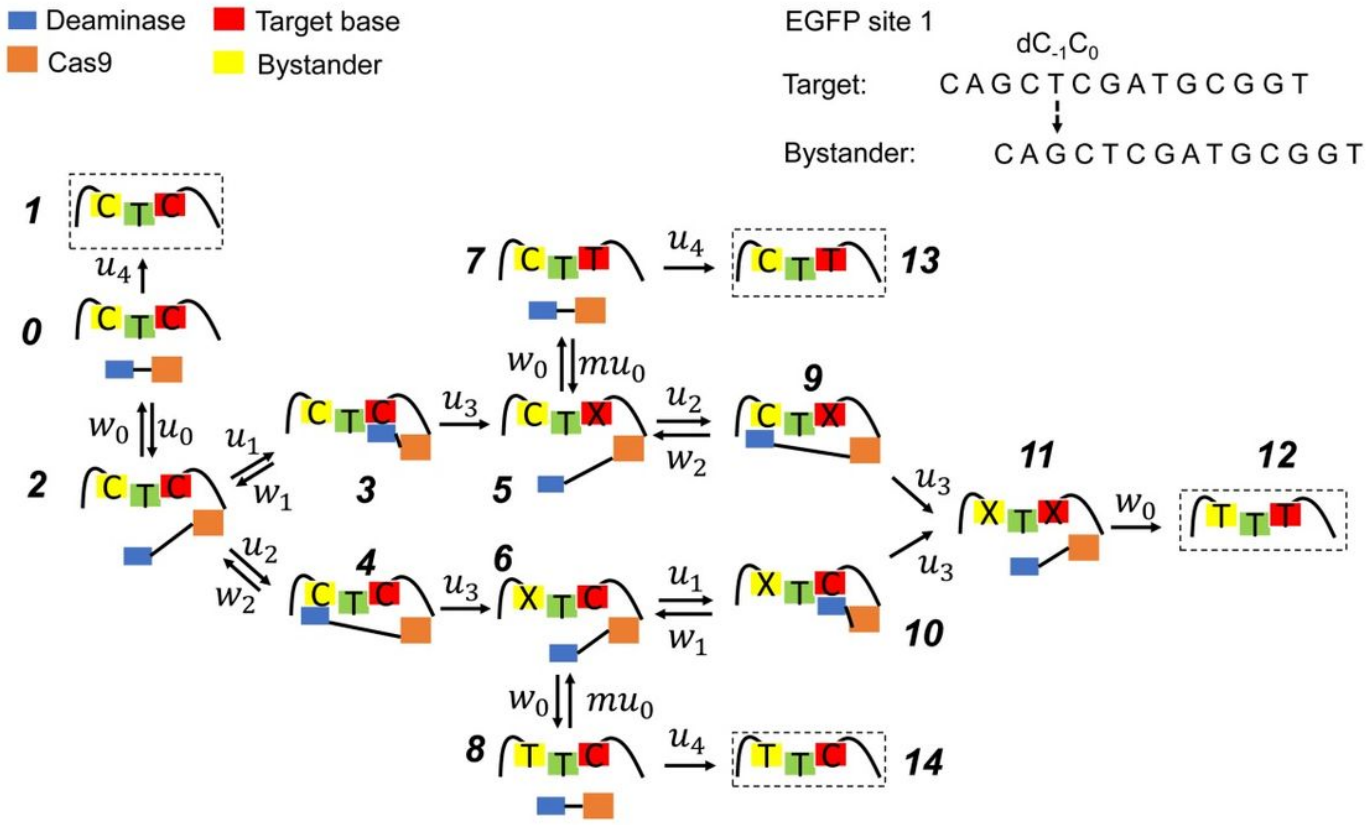
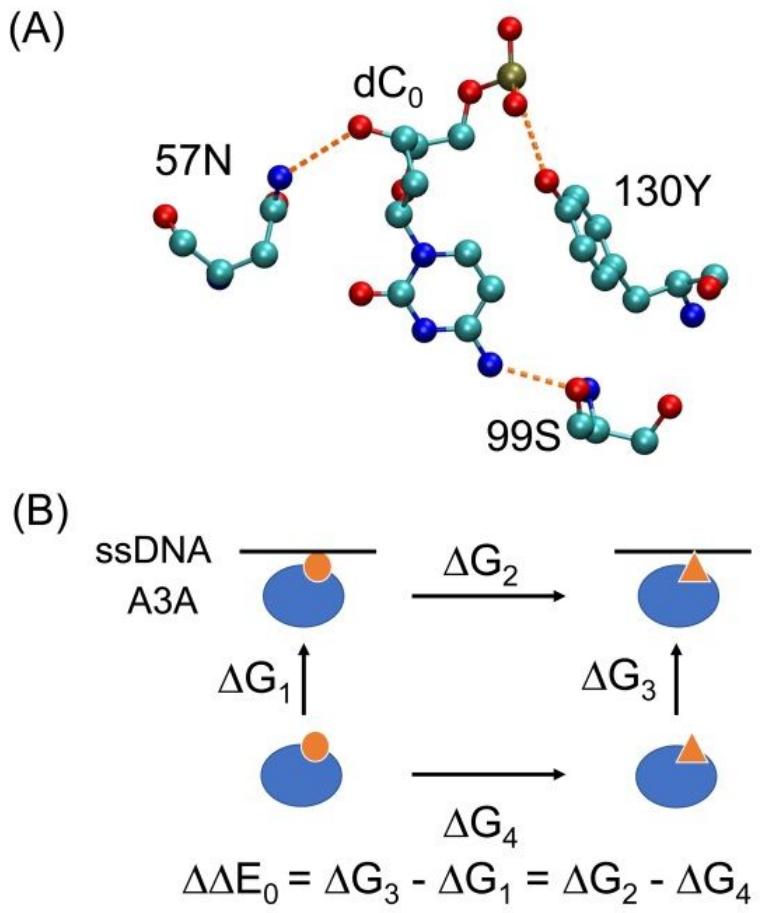


Figure 1

Chemical-kinetic model of A3A-BE3 editing the EGFP site 1. Deaminase, Cas9, target and bystander base are represented by blue, orange, red and yellow squares, respectively. “C” represents cytosine, “T” represents thymine and “X” represents uridine or thymine. Editing is modeled as a multiple-step chemical reaction where Cas9 first binds to ssDNA, cytidine then binds to the catalytic site of the deaminase and is chemically converted to thymidine. Here, uj and wj (j = 0 to 4) represent the chemical-kinetic rates for various transitions. The model has a total of 15 states and can produce four outcomes (dashed squares): CTC (failed editing), CTT (only the target base is edited), TTC (only the bystander is edited), TTT (both target and bystander bases are edited).



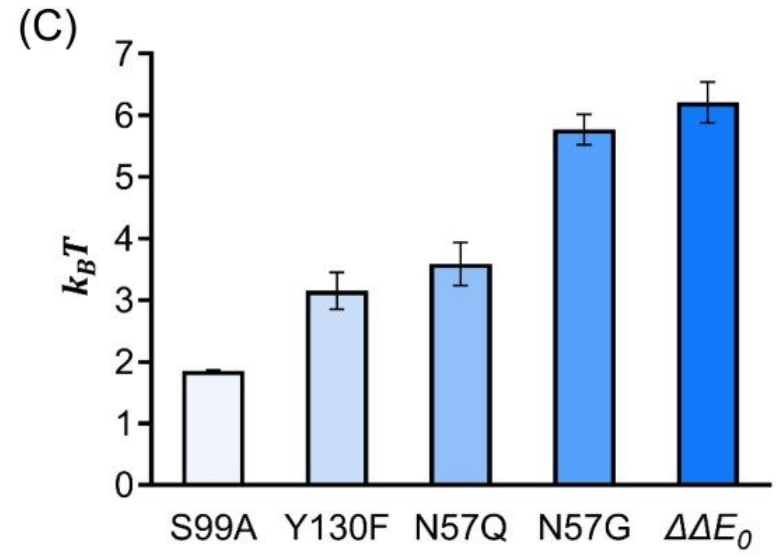


Figure 2

Calculation of binding free energy changes between ssDNA binding motif and A3A deaminase under various mutations. (A) Hydrogen bonding network between cytidine dC0 and A3A residues 57, 99 and 130; (B) thermodynamic cycle used to calculate the binding free energy changes: ssDNA (black line), A3A (blue balloons). A mutation at the binding interface is shown by a transition from small orange circle to orange triangle; (C) computationally estimated changes in binding free energy for A3A deaminase

mutations S99A, Y130F, N57Q, N57G and binding free energy change between A3A binding to target and bystander cytidines. The energy unit is kBT and T = 300K.

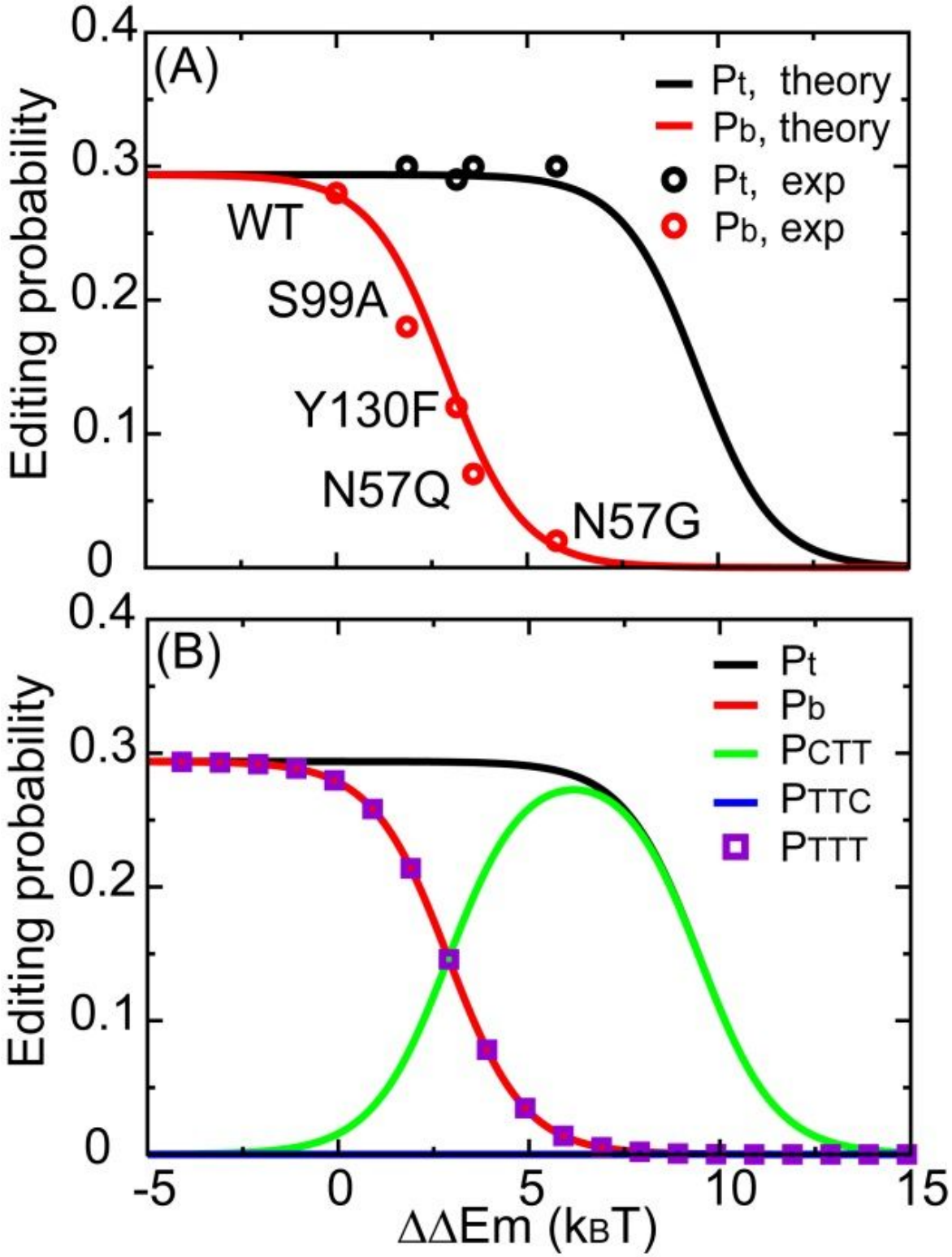


Figure 3

Theoretical model of the A3A-BE3 editing system. (A) Comparison between theoretical calculations (solid lines) and experimental data (exp, circles). ΔΔ□□ represents the perturbation in binding free energy due to the different mutations; kBT is the unit of energy; □□ and □□ represent the overall probability of editing

target and bystander cytidine, respectively; (B) calculated editing probabilities for products CTT, TTC and TTT. □□=□□□□+□□□□; □□=□□□□+□□□□.

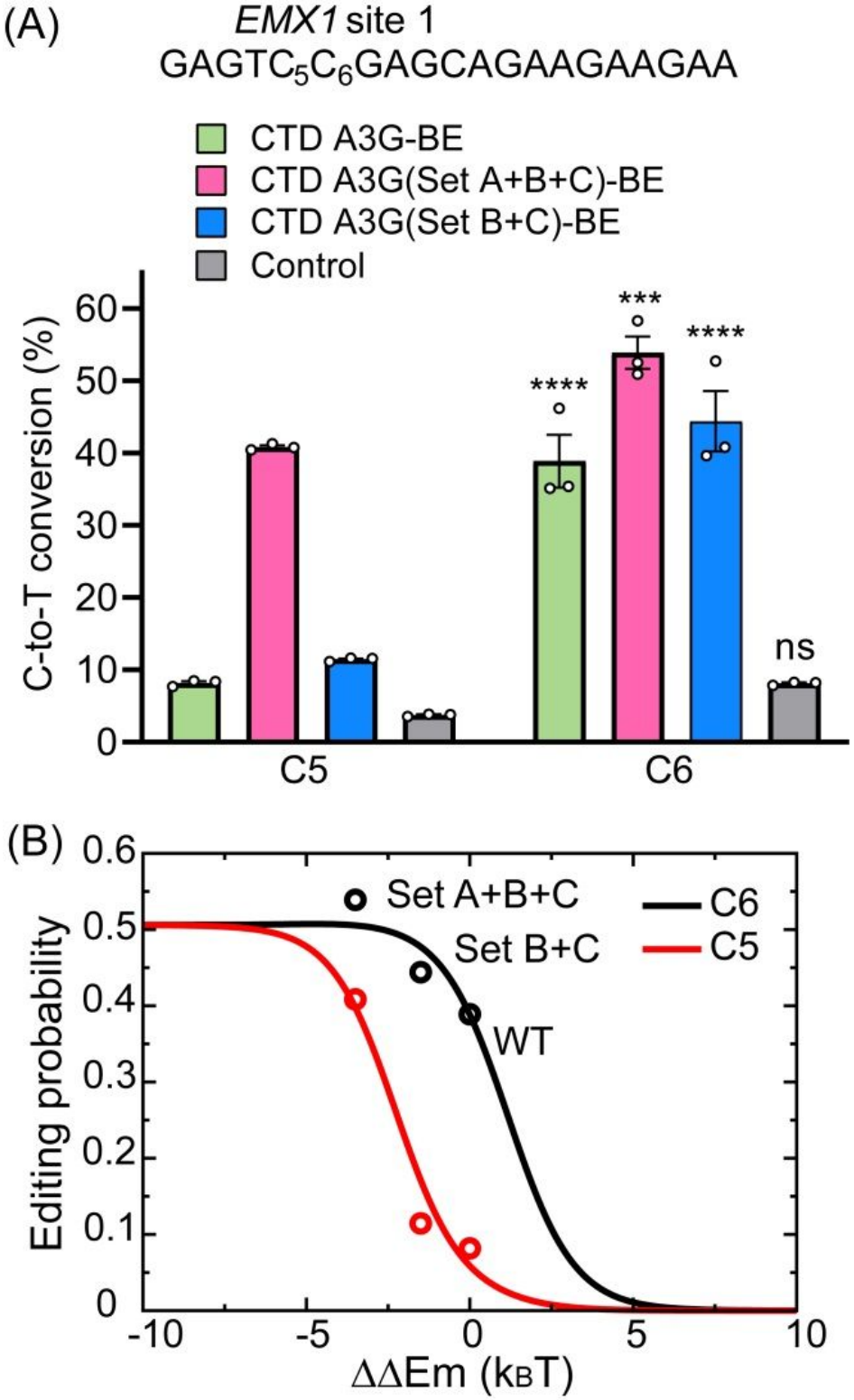


Figure 4

Engineering of CTD A3G-BEs that edit the EMX1 site 1. (A) Experimental measurements; (B) comparison between theoretical calculations (solid lines) and experimental measurement (circle). Please see the definition of set A, B and C in the method section.

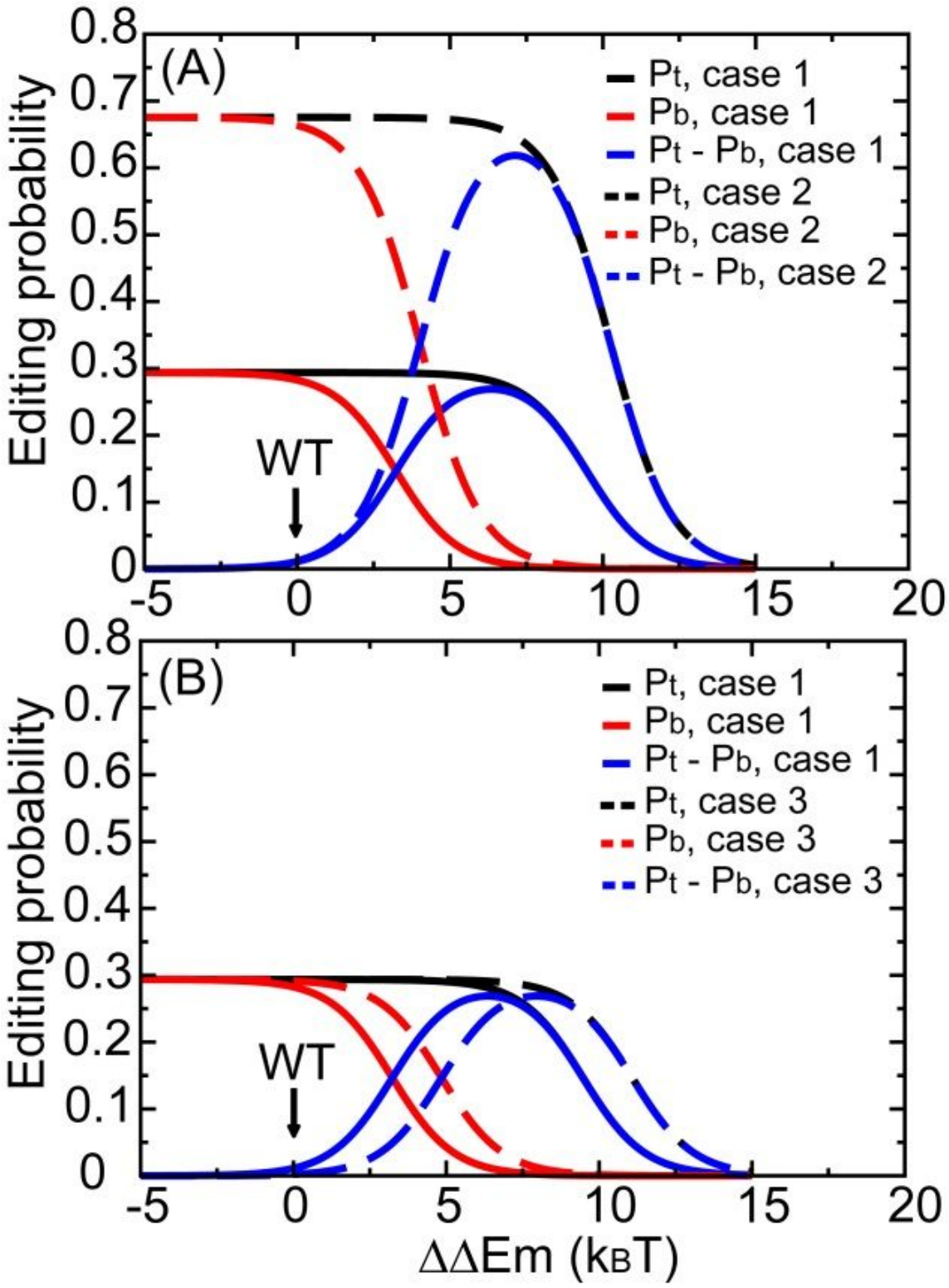


Figure 5

Base editing pattern of A3A-BE3 regulated by (A) 1 and (B) 3. The definition of ΔΔ, 1 and 3 can be found in eqns. [10,13-15]. and are the overall probabilities of editing the target and bystander cytidine, respectively. The difference between and is shown in blue. The setting with original parameters is represented by solid lines (case 1) whereas variants are represented by dashed lines (case 2: 1 divided by five; case 3: 3 divided by five).