# Leveraging noisy qualitative data for effective parameter inference in biophysical models

Jinghao Chen*[1,2] and William S. Hlavacek†[3]

[1]Institute for the Advanced Study of Human Biology, Kyoto University, Japan
[2]Department of Mathematics, University of California, Irvine, USA
[3]Theoretical Division, Los Alamos National Laboratory, USA

**Abstract**

Current mathematical models often emphasize high-precision quantitative data, overlooking the potential of noisy qualitative data, which is frequently generated in biological experiments but underutilized in modeling efforts. In this study, we reanalyzed DNA sequences from massively parallel experiments that characterize gene expression levels, encountering noisy data grouped into finite categories. To further address the challenges in the resolution of the data, only a subset of such data was selected and was further downscaled into 2-bin binary classes intentionally. Despite this deliberate simplification, the detailed protein-DNA binding energy distribution was successfully reconstructed through parameterizing the biological model using a straightforward negative logistic likelihood with stochastic gradient descent. Uncertainty quantification was performed using bootstraps. Our approach demonstrated consistency with outcomes from more complex models using refined data, precisely matching all known consensus sequences. Furthermore, other crucial parameters were jointly inferred, some of which were previously unattainable through existing models. We also introduced alternative parameterization methods, including a hinge loss function with a soft margin penalty and an extension of the negative logistic loss into a multi-category scenario, both replicating similar results. This research highlights the potential of noisy qualitative biological data and demonstrates that, with appropriate model parameterization approaches, we can extract meaningful insights into the underlying mechanisms of gene expression regulation.

*chen.jinghao.3n@kyoto-u.ac.jp, jinghc2@uci.edu
†wish@lanl.gov

# Contents

# 1 Introduction

In the current landscape of mathematical modeling, there is a predominant focus on high-precision quantitative data, often overlooking the inherent noise and complexity of biological data. Our research, however, seeks to address this gap by leveraging high-throughput, low-resolution qualitative data for the parameterization of biological models, with promising results.

Our primary emphasis lies in the parameterization of a biophysical model, detailed in Appendix Section 6.1 Eq. 18, pertaining to the regulation of the lac operon in E. coli, which was originally presented by Kuhlman *et al.* [1]. To achieve this, we undertook a reanalysis of high-throughput transcription regulatory sequences (TRS) data, as generated by Kinney *et al.* [2]. It is worth noting that Kinney *et al.* had previously performed model parameter inference using a method based on mutual information. However, this method had inherent limitations, particularly in its inability to infer certain model parameters. Moreover, it relied on an artificial "pseudo likelihood," rendering it statistically unreliable for uncertainty quantification.

Our research endeavor aims to surmount these limitations and, in the process, demonstrate the efficacy of inferring model parameters from low-quality datasets characterized by less data and lower resolution.

The primary objectives of this project encompass the following:

- Reanalysis of DNA sequence data originally presented by Kinney *et al.* [2]. This involves downsampling the data into binary labels linked to only two fluorescence levels.
- Parameterization of the biophysical model initially proposed by Kuhlman *et al.* [1]. This parameterization is achieved using statistically interpretable techniques, such as negative logistic likelihood. This approach is adopted to address the limitations observed in the mutual information method employed by Kinney *et al.* [2].
- The development of alternative parameterization methods and the extension of existing approaches from binary classifiers to more versatile and robust multi-category classifiers.
- A comprehensive assessment of the results and the implementation of uncertainty quantification through bootstrapping.

# 2 Model parameterization

## 2.1 Data processing

We have a dataset consisting of $M_0$ sequences, denoted as $\{\sigma_k\}_{k=1}^{M_0}$, along with their corresponding fluorescence levels, denoted as $\{\mu_k\}_{k=1}^{M_0}$. This dataset combines batches from 1 to 9, where for any $k = 1, 2, ..., M_0$:

$$\mu_k \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\} \tag{1}$$

To simplify the fluorescence levels, we use the bin-5 fluorescence as the reference level. By dropping bin-5 data, we can represent the fluorescence levels as binary classes, denoted as $\{y_k\}_{k=1}^{M}$:

$$y_k = \begin{cases} 1 & , \mu_k > 5 \\ 0 & , \mu_k < 5 \end{cases} \tag{2}$$

which is a reduced dataset, where the bin-5 category has been excluded. The data resolution has been minimized, with the transition from 9 classes in Eq. 1 to binary in Eq. 2 (the lowest possible resolution).

## 2.2 Negative logistic likelihood

Let $\bar{\tau}$ be the reference rate of transcription associate with bin-5 fluorescence level. Hence,

$$\begin{cases} \tau_\theta(\sigma_k) > \bar{\tau}, & \text{if } y_k = 1 \\ \tau_\theta(\sigma_k) < \bar{\tau}, & \text{if } y_k = 0 \end{cases} \quad \text{for } k = 1, 2, ..., M \tag{3}$$

Now we can construct the negative logistic likelihood (NLL) loss:

$$J = -\frac{1}{M}\sum_{k=1}^{M}[y_k \cdot \log(S(\tau_k - \bar{\tau})) + (1 - y_k) \cdot \log(1 - S(\tau_k - \bar{\tau}))], \tag{4}$$

where $S$ is the standard sigmoid function

$$S(x) = \frac{1}{1 + e^{-x}}. \tag{5}$$

For the model of rate of transcription please see Appendix Section 6.1 Eq. 6.1. The binding energies of CRP and RNAP are assumed to be the sum of energies through all the binding positions, detailed in Appendix Section 6.2. It is important to note that $\tau_k$ contains a scaling factor $\tau_{\max}$, making it challenging to directly infer the exact values of both the maximal transcription rate $\tau_{\max}$ and reference transcription rate $\bar{\tau}$. Nevertheless, we can deduce their ratio $\bar{\tau}/\tau_{\max}$ by applying

$$\frac{\tau_k - \bar{\tau}}{\tau_{\max}} = \frac{C_r e^{-\varepsilon_r/RT} + C_c C_r e^{-(\varepsilon_c+\varepsilon_r+\varepsilon_i)/RT}}{1 + C_c e^{-\varepsilon_c/RT} + C_r e^{-\varepsilon_r/RT} + C_c C_r e^{-(\varepsilon_c+\varepsilon_r+\varepsilon_i)/RT}} - \frac{\bar{\tau}}{\tau_{\max}} \tag{6}$$

in the loss functions (Eq. 4). Hence, we aim to infer a set of parameters denoted as:

$$\hat{\theta} = \left(\frac{\bar{\tau}}{\tau_{\max}}, C_c, C_r, \varepsilon_i, [\text{CRP energy matrix}], [\text{RNAP energy matrix}]\right) \tag{7}$$

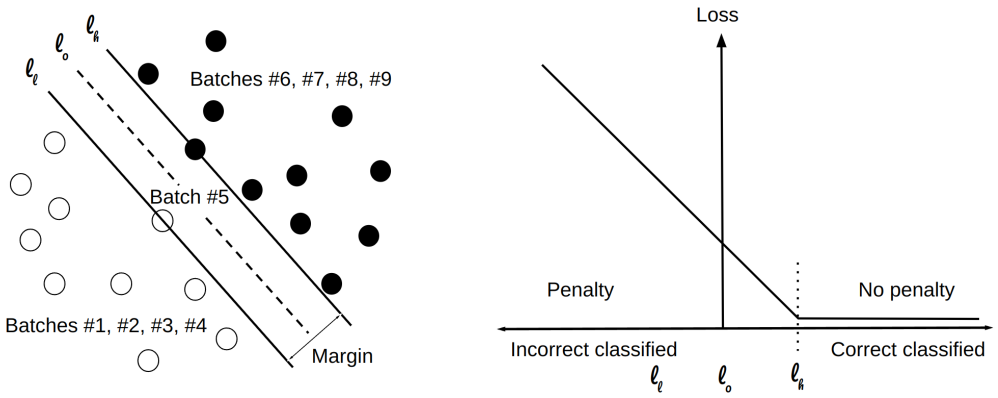
where these parameters are obtained by minimizing the loss function:

$$\hat{\theta} = \arg\min_\theta J. \tag{8}$$

## 2.3 Alternative model: hinge loss with soft margin

We now aim to incorporate the static penalty introduced by Mitra *et al.* [3] to devise a more straightforward yet equally effective loss function. Since the entire bin-5 has been dropped, it is natural to consider a margin separating two classes of data instead of just a single layer hyperplane (Fig. 1a). That is, for the class labeled by $y_k = 1$,

$$\exists \Delta\tau > 0 \ \text{ s.t. } \ \tau_k > \bar{\tau} + \Delta\tau \tag{9}$$




(a) Margin. (b) Hinge.

Figure 1: An alternative model based on Support Vector Machine (SVM).

and $\tau_k < \bar{\tau} - \Delta\tau$ for the class labeled as $y_k = 0$. Without loss of generality, these non-dimensional rate of transcription can be rescaled to yield a unit $\Delta\tau = 1$, hence the hinge loss can be formulated from static penalties accordingly:

$$J = \frac{1}{M}\sum_{k=1}^{M} \max\left(0, 1 + (2y_k - 1)\cdot(\tau_k - \bar{\tau})\right). \tag{10}$$

When $y_k = 0$, the linear penalty is applied as $\tau_k$ exceeds $\bar{\tau} - 1$, and its magnitude is proportional to the amount by which it surpasses this threshold. Similarly, when $y_k = 1$, the linear penalty is applied as $\tau_k$ falls below $\bar{\tau} + 1$, and its magnitude is proportional to the deviation from this bound (Fig. 1b). This implies that the range $[\bar{\tau} - 1, \bar{\tau} + 1]$ represents the (non-dimensional) reference rate of transcription span for bin-5 sequences.

The hinge loss defined in Eq. 10 serves as a soft margin support vector machine (SVM) model, effectively penalizing misclassifications. This insight prompts us to explore the margin width, which can be quantitatively expressed as a function of the energy matrix:

$$\frac{2}{\|(\tau_k)_{k=1}^{M}\|} \tag{11}$$

where $\|\cdot\|$ takes the $L^2$ norm

$$\|(\tau_k)_{k=1}^{M}\| = \sqrt{\sum_{k=1}^{M} \tau_k^2}. \tag{12}$$

## 2.4 Model extension: multi-category scenario

In contrast to binary classification ($y_k = 0$ or 1), our model is being extended to a multi-category scenario. In binary classification, the probability of a fluorescence level being higher than

the reference level ($y_k = 1$) conditioned on the sequence-based parameter $\theta$ can be interpreted as the cumulative probability function of parameter $\theta$ [3]. In other words,

$$\mathbb{P}(y_k = 1|\theta) = \mathbb{P}(\tau_k > \bar{\tau}|\theta) = cdf(\theta, \sigma_k, \bar{\tau}) = S(\tau_\theta(\sigma_k) - \bar{\tau}), \tag{13}$$

while conversely,

$$\mathbb{P}(y_k = 0|\theta) = 1 - \mathbb{P}(y_k = 1|\theta). \tag{14}$$

Analogously, we can define conditional probabilities for the multi-category case. Instead of a single threshold $\bar{\tau}$, we now require two thresholds, $\bar{\tau}_1$ and $\bar{\tau}_2$, for three classes:

$$\begin{aligned} \mathbb{P}(\tau_k < \bar{\tau}_1|\theta) &= 1 - cdf(\theta, \sigma_k, \bar{\tau}_1) \\ \mathbb{P}(\bar{\tau}_1 < \tau_k < \bar{\tau}_2|\theta) &= cdf(\theta, \sigma_k, \bar{\tau}_1) - cdf(\theta, \sigma_k, \bar{\tau}_2) \\ \mathbb{P}(\tau_k > \bar{\tau}_2|\theta) &= cdf(\theta, \sigma_k, \bar{\tau}_2) \end{aligned} \tag{15}$$

Assuming that $\bar{\tau}_1$ and $\bar{\tau}_2$ are sufficiently separated, we can simplify the probability $\mathbb{P}(\bar{\tau}_1 < \tau_k < \bar{\tau}_2|\theta)$ as follows:

$$\mathbb{P}(\bar{\tau}_1 < \tau_k < \bar{\tau}_2|\theta) = cdf(\theta, \sigma_k, \bar{\tau}_1)(1 - cdf(\theta, \sigma_k, \bar{\tau}_2)) \tag{16}$$

Consequently, the loss function for the multi-category scenario can be expressed as:

$$J = -\frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{2}[y_k \cdot \log(S(\tau_\theta(\sigma_k) - \bar{\tau}_i)) + (1 - y_k) \cdot \log(1 - S(\tau\theta(\sigma_k) - \bar{\tau}_i))] \tag{17}$$

This loss function accommodates multiple categories for an improved understanding of the data.

# 3 Results

Due to the large size of our dataset, we employed the stochastic gradient descent method with 500 epochs to search for the optimal energy matrices (Fig. 2). Other model parameters in the biophysical model (Eq. 18) can be jointly inferred with energy matrices. To assess the degree of uncertainty, we executed bootstrapping, with the outcomes illustrated in Fig. 3.

In Table 1, we present a comparative analysis of our results with those of Kinney *et al.* This table highlights our ability not just to replicate the energy matrices (see Fig. 2) but also to derive crucial model parameters, notably $C_r$ and $\bar{\tau}/\tau_{\max}$. It is worth noting that these parameters were proved to be unattainable using mutual information approach [2].

| model parameter | J Kinney *et al.* 2010 | J Chen *et al.* 2023 |
|---|---|---|
| $\bar{\tau}/\tau_{\max}$ | unknown | $0.467 \pm 0.005$ |
| $C_c$ | $10^{-1.2\pm0.2}$ | $0.027 \pm 0.002$ |
| $C_r$ | unknown | $0.053 \pm 0.002$ |
| $\varepsilon_i$ | $-3.26 \pm 0.41$ kcal/mol | $-4.855 \pm 0.077$ kcal/mol |

Table 1: Inferred parameters for normalized reference transcription rate $\bar{\tau}/\tau_{\max}$, CRP concentration $C_c$, RNAP concentration $C_r$ and CRP-RNAP interaction energy $\varepsilon_i$.

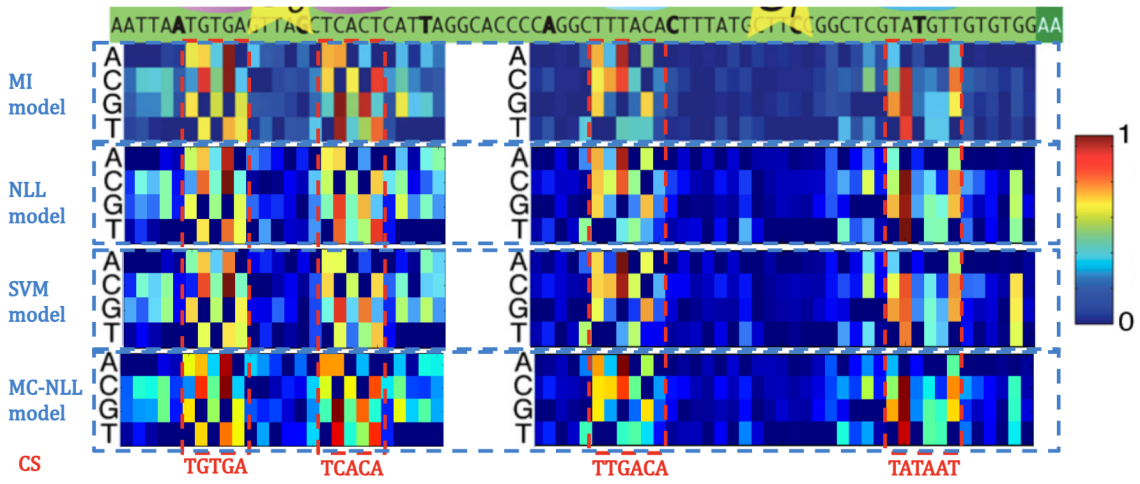



Figure 2: Energy matrices trained using various models. From top to bottom: MI model by Kinney *et al.* [2] based on mutual information; NLL model and SVM model as binary classifiers developed in our study, and MC-NLL model representing the multi-category NLL model introduced in the model extension section. Consensus sequences (CS) are highlighted in red at the bottom, corresponding to entries with the lowest binding energy (shown in the darkest blue color) for each position in the energy matrices.

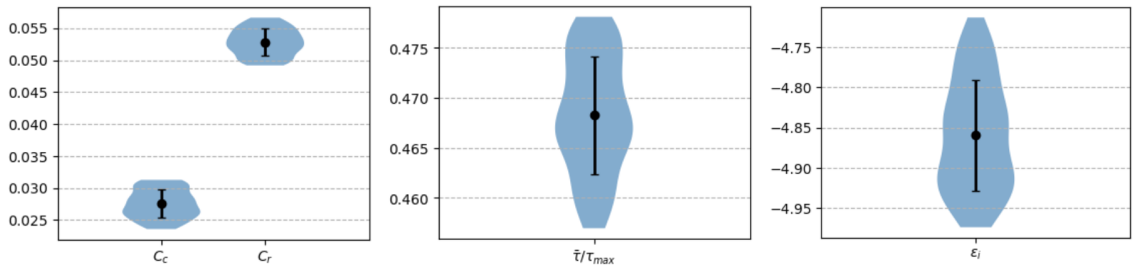



Figure 3: Bootstrapping results.

Finally, we can assess the accuracy of our binary classifier, as shown in Figure 4. In this example, we use the NLL model, and similar observations apply to the SVM model. It's noteworthy that accuracy is relatively higher when the data is far from the threshold (bin-5), which is to be expected. Overall, the accuracies across the batches are consistently high.

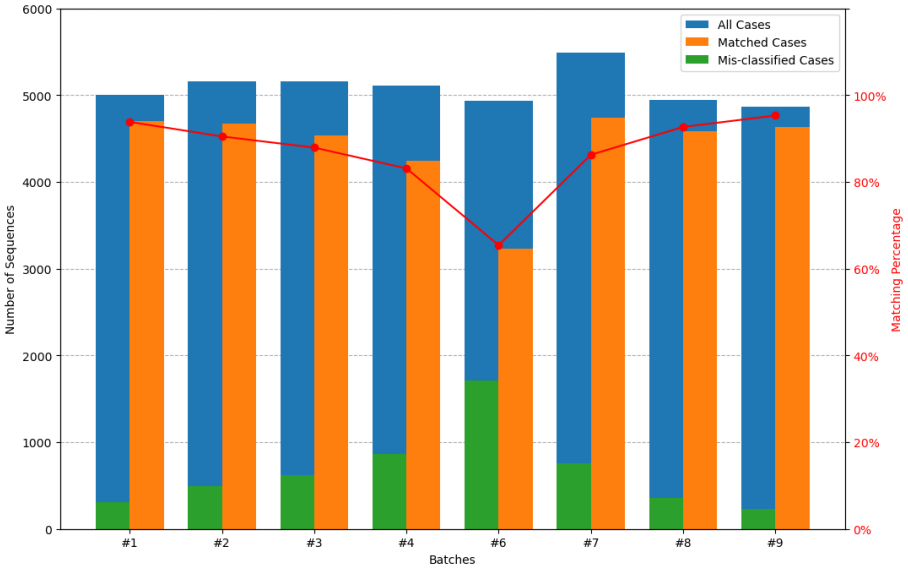



Figure 4: Accuracy quantification.

# 4 Discussion

# 5 Acknowledgement

This research was supported in part by an appointment with the National Science Foundation (NSF) Mathematical Sciences Graduate Internship (MSGI) Program. This program is administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and NSF. ORISE is managed for DOE by ORAU. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of NSF, ORAU/ORISE, or DOE.

# References


[1] Kuhlman T, Zhang Z, Saier MH Jr, and Hwa T. "Combinatorial transcriptional control of the lactose operon of Escherichia coli". In: *Proc Natl Acad Sci* 104.14 (2007), pp. 6043–6048. DOI: 10.1073/pnas.0606717104.

[2] Kinney JB, Murugan A, Callan CG Jr, and Cox EC. "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence". In: *Proc Natl Acad Sci* 107.20 (2010), pp. 9158–9163. DOI: 10.1073/pnas.1004290107.

[3] Mitra ED, Dias R, Posner RG, and Hlavacek WS. "Using both qualitative and quantitative data in parameter identification for systems biology models". In: *Nat Commun* 9.1 (2018), p. 3901. DOI: 10.1038/s41467-018-06439-z.

# 6 Appendix

## 6.1 Model of transcription rate

In accordance with Kuhlman *et al.* [1], we made an assumption that the transcription rate $\tau$ at the *lac* promoter is directly related to the presence of RNA polymerase (RNAP) at its binding site in a state of thermal equilibrium. This relationship is quantitatively described as follows:

$$\tau = \tau_{\max} \frac{C_r e^{-\varepsilon_r/RT} + C_c C_r e^{-(\varepsilon_c+\varepsilon_r+\varepsilon_i)/RT}}{1 + C_c e^{-\varepsilon_c/RT} + C_r e^{-\varepsilon_r/RT} + C_c C_r e^{-(\varepsilon_c+\varepsilon_r+\varepsilon_i)/RT}}. \tag{18}$$

The model given above involves several parameters: $\tau_{\max}$, the transcription rate with full RNAP occupancy; $C_c$ and $C_r$, the concentrations of CRP and RNAP; $R = 1.98 \times 10^{-3}$ kcal/mol°K, the gas constant; and $T = 310$°K, the cell-induced temperature. The key parameter is $\varepsilon_i$, representing the interaction energy between CRP and RNAP. In short, these model parameters can be summarized in the following Table 2.

| model parameter | meaning |
|---|---|
| $\varepsilon_c$ | CRP binding energy |
| $\varepsilon_r$ | RNAP binding energy |
| $\varepsilon_i$ | CRP-RNAP interaction energy |
| $C_c$ | CRP concentration |
| $C_r$ | RNAP concentration |
| $\tau_{\max}$ | maximal transcription rate |
| $R$ | gas constant ($1.98 \times 10^{-3}$ kcal/mol°K) |
| $T$ | cell-induced temperature (310°K) |

Table 2: Model parameters for Eq. 18.

It is noteworthy that $\varepsilon_c$ and $\varepsilon_r$ are the only sequence-dependent parameters in this model, while all other parameters remain constant for different sequence inputs $\sigma_k$. Importantly, the transcription rate $\tau$ exhibits global monotonicity (monotonic increasing) with respect to both $\varepsilon_c$ and $\varepsilon_r$. Hence, we can write the transcription rate of sequence $\sigma_k$ as $\tau_k := \tau(\epsilon_{c,k}, \epsilon_{r,k}) = \tau(\sigma_k)$, and set the reference transcription rate $\bar{\tau}$ to be the transcription rate of sequences in bin-5.

## 6.2 Energy matrix

Recall that the energy matrix models: each base within a protein's binding site was assumed to contribute additively to the overall binding energy [2]. Under this hypothesis, the binding energy of CRP $\varepsilon_c$ can be decomposed as

$$\varepsilon_c = \sum_{i=-74}^{-49} \varepsilon_c^i \tag{19}$$

where $\varepsilon_c^i$ is the binding energy at the position $i$, which depends on the nucleotide $b$ at position $i$:

$$\varepsilon_c^i(b) = \varepsilon_c^{i,A} \cdot \mathbb{1}_A(b) + \varepsilon_c^{i,C} \cdot \mathbb{1}_C(b) + \varepsilon_c^{i,G} \cdot \mathbb{1}_G(b) + \varepsilon_c^{i,T} \cdot \mathbb{1}_T(b) \tag{20}$$

Here, $\mathbb{1}_A(b)$ is the indicator function (characteristic function) defined as

$$\mathbb{1}_A(b) = \begin{cases} 1 & , b = A \\ 0 & , \text{otherwise} \end{cases} \tag{21}$$

and analogously for $\mathbb{1}_C(b)$, $\mathbb{1}_G(b)$ and $\mathbb{1}_T(b)$.

By collecting the energy weights $\varepsilon_c^{i,b}$ for all possible bases $b \in \{A, C, G, T\}$ and binding positions $i \in [-74, -49] \cap \mathbb{Z}$, we can construct the energy matrix of CRP, denoted as $\theta_c = (\varepsilon_c^{i,b})$, which has 4 rows and 26 columns. Now, the total binding energy of CRP on a sequence $\sigma = (b_i)_{i=-74}^{-49}$ in Eq. 19 can be rewritten as

$$\varepsilon_c(\sigma) = \sum_{i=-74}^{-49} \varepsilon_c^i(b_i). \tag{22}$$

## 6.3 Energy shift

The energy matrices have been adjusted for convention. Following the energy matrix model assumption, where each *lac* promoter site contributes independently and additively to the total binding energy, we set the minimal entry at each individual column to 0, representing the wild-type *lac* promoter site having zero energy. Furthermore, the results obtained from the model are non-dimensional, allowing us to rescale the values to a range between 0 and 1 by dividing each entry by the maximal value in the entire matrix. This rescaling is convenient since the dimensional matrix is determined only up to a multiplicative constant, and it ensures that the energy matrix remains comparable and interpretable across different experiments or datasets.

After obtaining the dimensional energy matrix, we can calculate the energy shift in units of kcal per mol by summing up all the individual shifting amounts at each site. According to Kinney *et al.* [2], this energy shift is approximately -7 kcal/mol for CRP and around -8.3 kcal/mol for RNAP.

Let's assume that at site $i$, the energy shift is represented by $\delta_i$ (to avoid abusing the notation since $\varepsilon_i$ is the interaction energy between CRP and RNAP), and the binding energy is denoted as $\hat{\varepsilon}_c^i(b)$. As a reference, the binding energy at position $i$ without energy shifting, denoted by $\varepsilon_c^i(b)$, is shown in Eq. 20. Now, the modified $\hat{\varepsilon}_c^i(b)$ can be expressed in the following form:

$$\begin{aligned} \hat{\varepsilon}_c^i(b) =& (\varepsilon_c^{i,A} + \delta_i) \cdot \mathbb{1}_A(b) + (\varepsilon_c^{i,C} + \delta_i) \cdot \mathbb{1}_C(b) \\ &+ (\varepsilon_c^{i,G} + \delta_i) \cdot \mathbb{1}_G(b) + (\varepsilon_c^{i,T} + \delta_i) \cdot \mathbb{1}_T(b) \\ =& \varepsilon_c^{i,A} \cdot \mathbb{1}_A(b) + \varepsilon_c^{i,C} \cdot \mathbb{1}_C(b) + \varepsilon_c^{i,G} \cdot \mathbb{1}_G(b) + \varepsilon_c^{i,T} \cdot \mathbb{1}_T(b) \\ &+ \delta_i \cdot (\mathbb{1}_A(b) + \mathbb{1}_C(b) + \mathbb{1}_G(b) + \mathbb{1}_T(b)) \\ =& \varepsilon_c^i(b) + \delta_i \cdot (\mathbb{1}_A(b) + \mathbb{1}_C(b) + \mathbb{1}_G(b) + \mathbb{1}_T(b)). \end{aligned} \tag{23}$$

Since $b \in \{A, C, G, T\}$, we have

$$\mathbb{1}_A(b) + \mathbb{1}_C(b) + \mathbb{1}_G(b) + \mathbb{1}_T(b) \equiv 1, \tag{24}$$

therefore Eq. 23 goes to

$$\hat{\varepsilon}_c^i(b) = \varepsilon_c^i(b) + \delta_i. \tag{25}$$

Now with Eq. 25, the total binding energy of CRP (the form without energy shifting is Eq. 22) becomes

$$\begin{aligned}\hat{\varepsilon}_c(\sigma) &= \sum_{i=-74}^{-49} (\varepsilon_c^i(b_i) + \delta_i) \\ &= \sum_{i=-74}^{-49} \varepsilon_c^i(b_i) + \sum_{i=-74}^{-49} \delta_i \\ &= \varepsilon_c(\sigma) + \delta_c\end{aligned} \tag{26}$$

where the total shift energy $\delta_c$ for CRP is defined in the following form:

$$\delta_c := \sum_{i=-74}^{-49} \delta_i. \tag{27}$$

This implies that each site can have its individual energy shift without requiring simultaneous adjustments for all sites. The overall total energy shift is simply the sum of these shifts at each site. Consequently, setting the lowest energy at each site to zero becomes feasible, facilitating easier comparison of the entire energy matrix. The only additional factor to consider is a constant term $\delta_c$ in Eq. 27, the total energy shift.

## 6.4 Stochastic gradient descent

Recall that Eq. 8 represents an optimization problem that can be solved using stochastic gradient descent (SGD) algorithm:

*Initialize* parameters $\theta$
*Given* initial step size $h$, stopping criterion (e.g., 50 epochs)
**while** (not done)
  **for** $k = 1, ..., M$:
    *Compute* transcription rate $\tau = \tau_\theta(\sigma_k)$ from $\sigma_k$
    *Compute* loss $J = J_\theta(\sigma_k)$ from such $\tau$
    *Update* parameters $\theta \leftarrow \theta - h\nabla_\theta J$
  **end**
**end**
*Return* $\theta$