update 10/10/2018

FIT TO DATA FROM BOYLE ET AL.

- used the Association experiments as training data.
- used ony single and double mismatches and used the median over different nucleotide sequences.
- The parameters we fit (and the ranges we allowed):
  - 20x $\epsilon_C(n) \in [-10, 10]k_BT$: the gains for matches
  - 20x $\epsilon_I(n) \in [0, 10]k_BT$: The amount of energy added to the landscape for a mismatch at position n
  - $\epsilon(0) = \epsilon_{PAM} \in [0, 10]k_BT$, the energy of the PAM-bound state (to speed things up, we already know it should be unstable comparred to solution)
  - The (log of the) rate from solution to PAM: $\log_{10}(k) \in [3, -7]$
  - The (log of the) rate from PAM to the first R-loop state: $\log_{10}(k) \in [3, -7]$
  - Internal forward rates (between any set of R-loop states): $\log_{10}(k) \in [3, -7]$

  NOTE: We tested the fit with wider boundaries, unfortunately the solutions either do not converge or lead to significant numerical errors for to large rates. The latter is due to the way we solve the MAster equation using a numerical exponentiation of a matrix. NOTE: However, the results do not go the the boundaries
- Repeated the fit 110 times
- selected 64 fits with lowest χ^2. There is a somewhat arbitrary threshold applied. I chose the threshhold such that all selected fits perform good - by eye - on both single and double mismatches. There are much less datapoints for the single-mismatches. Hence, every now and then the SA algorithm settles for a mediocre fit to the single mismatches and still fits the double mismatches perfectly. Luckily I learned how to discriminate them based on the (final) χ^2.
- The coarse-grained free-energy - in which we treat all bound states as one collective (macro-)state - comes out exactly the same every time around
- When viewed in the 21-state system, we see that the data is to some degree indiscriminate.
- We used the median free-energy landscape (the microscopic one) to further test this hypothesis.
- Indeed, the median landscape performs as good as any of the individual fits. Note that the indidual fits selected - if I am not mistaken - differ by less then 10% in χ^2.
- After averaging the energies, we still need to set the rates.
- The individual values of the 3 rates tend to differ quite some between different rates. So how did it still ensure a good fit to the measured association rate ?
- We know that in order to get a proper fit to the two-mismatch heatmap the solution satisfies:
  1. $2\epsilon_I > \sum_{i=0}^{N} \epsilon_C[i]$
  2. The coarse-grained bound free-energy is well determined: $F_n \approx -\log(\sum_{i=0}^{n} e^{-\epsilon_C[i]})$ (note: with multiple mismatches no more stable states are available after them)
  3. R-loop propagation is much faster than initiation of the R-loop. The on-target has a measurable/slow association rate, while seed mismatches are subjected to rejection events that are faster than detectable. What is kept constant is the total rate for completing the R-loop of the on-target. Assuming R-loop initiation is rate limmiting allows us to only keep track of the probability of completing the R-loop and the rate for starting the R-loop. Essentially, the system needs time to initiate the R-loop after which there is a probability to not (immediately) reject the substrate. If the R-loop gets completed, the protein is essentially so stably bound that we may ignore any further rejections. Hence, denoting Solution, PAM and R-loop by 'S', 'P', and 'R' and using an effective rate of $k_{PR} \times P_{n=1}$ - with $P_{n=1}$ the fixation probability of the BD-process starting from the first R-loop state - the effective rate to initiate and thereby complete R-loop becomes:

$$k_{OT} = \frac{k_{SP}k_{PR}P_{n=1}}{k_{SP} + k_{PS} + k_{PR}P_{n=1}}$$

Assuming equillibrium between the PAM bound and unbound states ($k_{SP} + k_{PS} >> k_{RP}$):

$$k_{OT} \approx \frac{k_{PR}P_{n=1}}{1 + K_D}$$

with

$$K_D = \frac{k_{PS}}{k_{SP}} = e^{+\epsilon_{PAM}}$$

Given an energy landscape, K_D is fixed through ϵ_{PAM}. Furthermore, assuming partial equillibration, k_{SP} is irrelevant. We end up with two free parameters, the internal forward rate (k_f) and the rate into the R-loop from the PAM (k_{PR}) that will determine the value of k_{init} - which is fixed by the association rate for the on-target. Re-writing in terms of k_{PR}:

$$k_{PR} = k_{OT}(1 + K_D)(1 - \frac{e^{\epsilon_1}}{k_f P_{n=2}})^{-1}$$

Our plan is to repeat the SA-optimisation multiple times and use the averaged energy landscape as our candidate. Can we then find a value for the internal forward rate, such that the fit is still good? Moreover, by sweeping the internal rate, we can find the bounds it must satisfy in order for the fit to be good enough.

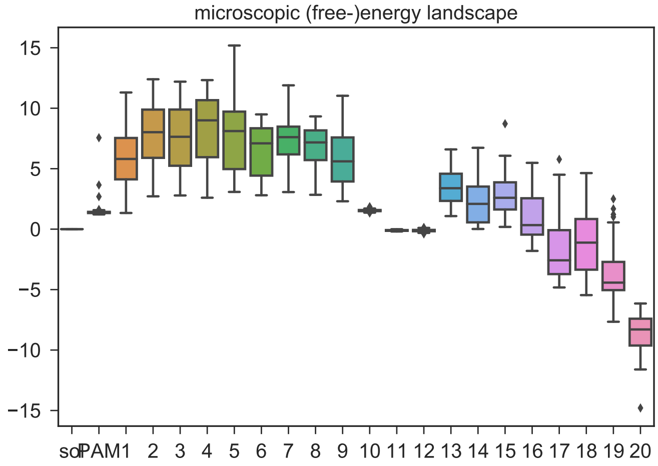

FIG. 1: *Microscopic free-energy landscapes from all 64 selected fits in box-plot format.*

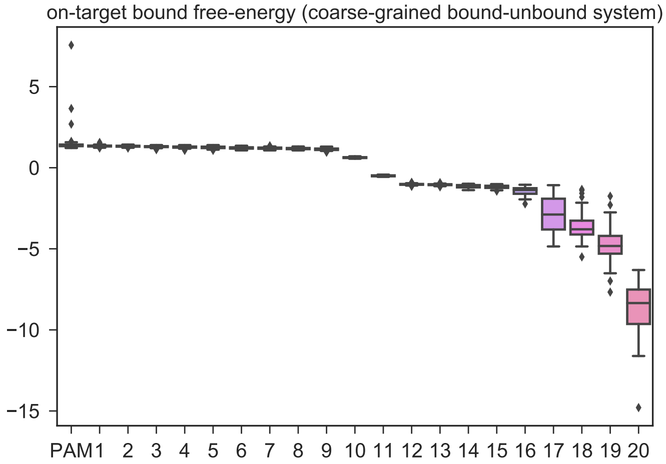

FIG. 2: *Coarse-Grained free-energy diagrams from all 64 selected fits in box-plot format.*

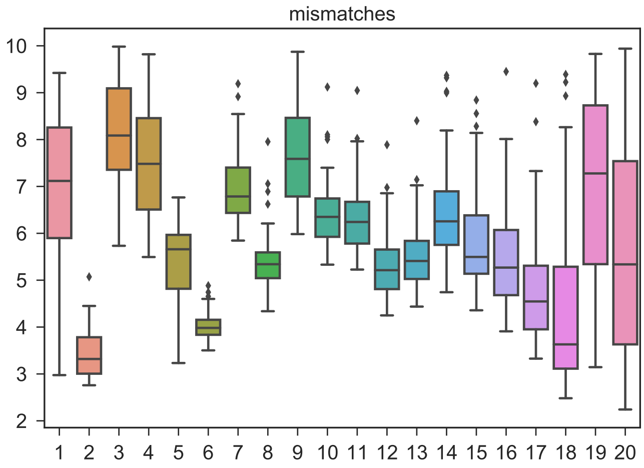

FIG. 3: *Mismatch penalties from all 64 selected fits in box-plot format.*

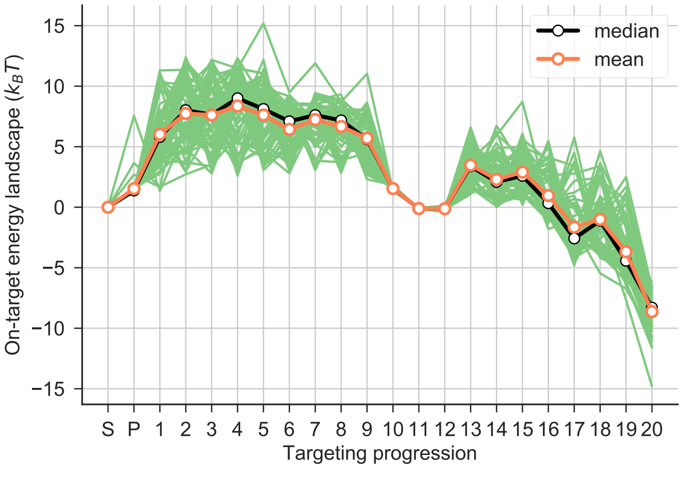

FIG. 4: *All microscopic free-energy landscapes (green). Further analysis/plots are made using the median solution shown in black. The averaged solution (orange) naturally yields similar results.*

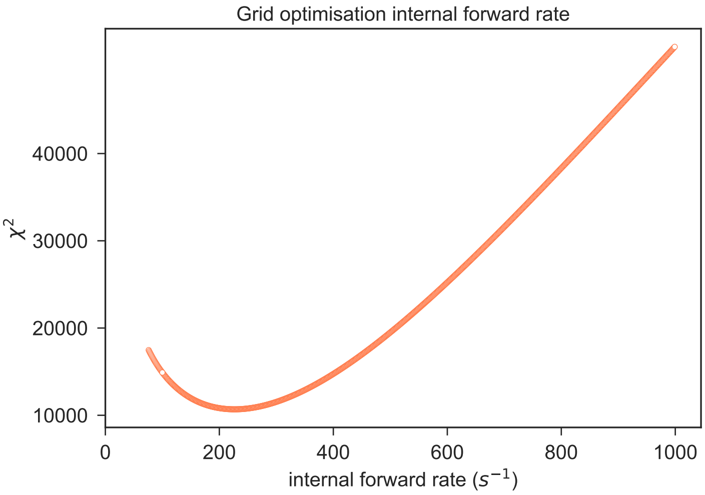

FIG. 5: *After fixing the energies, we use a quick grid optimisation to determine the internal forward rate by fitting only to the single-mismatches.*

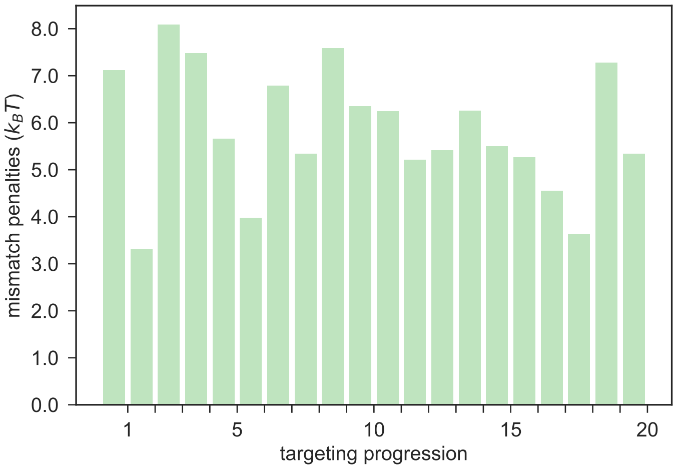

FIG. 6: *Median result for the mismatch penalties.*

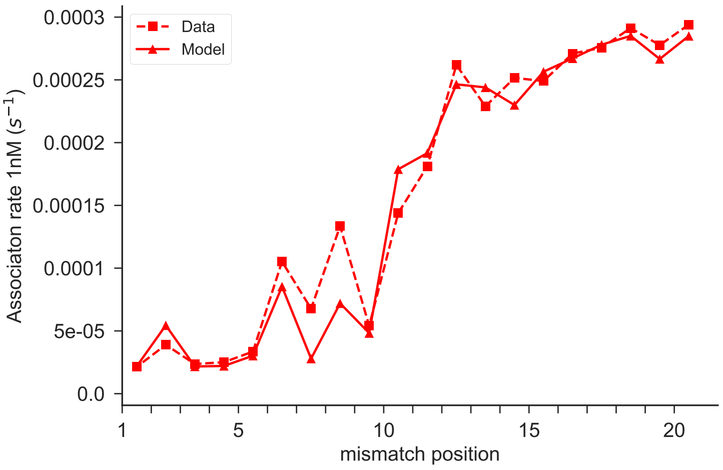

FIG. 7: *"Fit" result for the single-mismatch on-rate using the median landscape. We can show this is an identical result to any of the individual fits.*

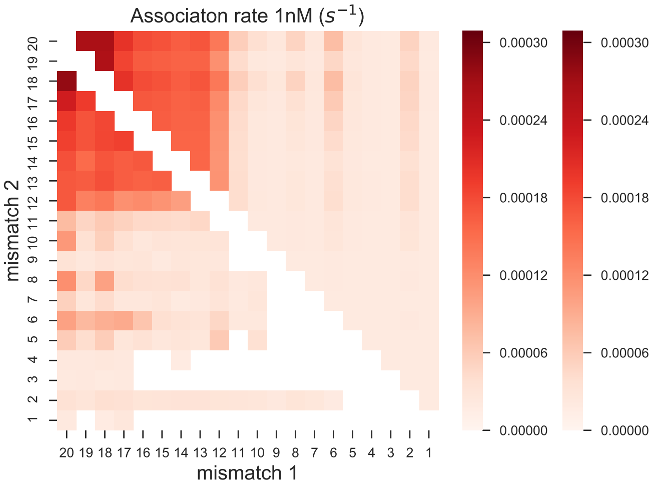

FIG. 8: *Result for double-mismatches. Again a 'perfect' fit*

COMPARE TO ILYA'S DATA

- Using the parameters shown above. Simulated the experiment as done by Ilya's group to determine the exact same observable (ΔABA).
- A zero parameter fit to the data from Ilya Finkelstein.
- The dataset below used the exact same guide sequence as Boyle et al. did

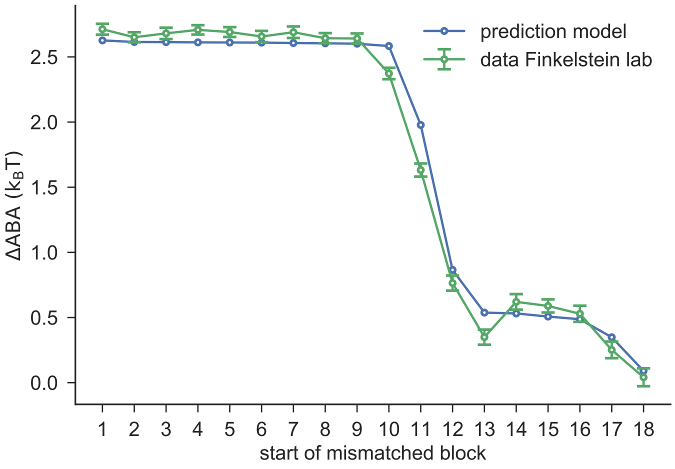

FIG. 9: *Blocks of mismatches. Took median of starting point. Comparisson with entire heatmaps for all starting points and lengths of blocks also available. It makes sense that this quantity very much resembles the coarse-grained free-energy mentioned previously.*

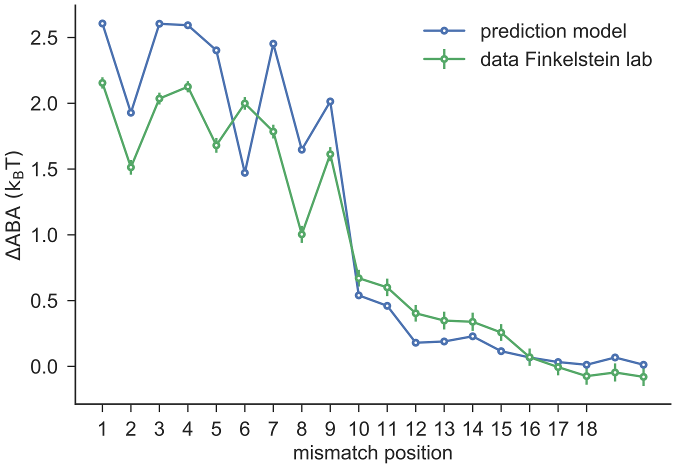

FIG. 10: *Comparisson to single-mismatches*

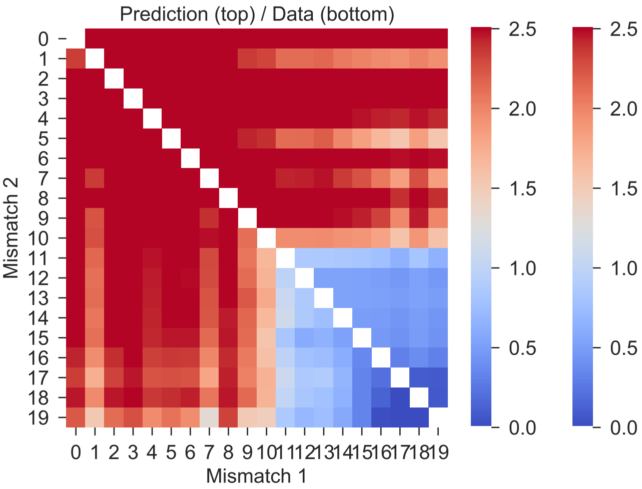

FIG. 11: *Comparisson to double mismatches. The two colorbars correspond to the prediction and the data separately: Just to confirm they actually produced the same values.*