update 10/10/2018

BEFORE WE START...

Unfortunately, we cannot use Koen's results directly, because the files on the PNAS website were labelled incorrectly. In reality, the association data is taken using 1nM dCas9, while the dissociation (and occupation) data used 10nM dCas9.

In essence, where Koen's project showed the data to be of 'class I' (see below), we will now show the data to really be of class II in stead.

Moreover - as we will show in the end - Ilya's data dictates that the free-energy landscape drops between locations 9-12. Also, he has data using blocks of mismatches (pos 2-20, 3-20, 4-20, etc.) that indicates that the length of the mismatch block has little to no effect. We will show his data in the final section, but bare in mind that this has partially been our motivation to try different fits.

A. TWO CLASSES OF FIT RESULTS

We get very similar on-target landscapes every time around. However, they are not as Koen got (not because of Koen, but because of PNAS!).

For our interpretation we group the fit results based on the mismatch penalties. As you will see below, this changes 'what are the important stable states'.

Another important concept we will speak of here is the 'aproximate' free-energy landscape of the on-target:

$$\Delta F_n = -\log\left(\sum_{i=0}^{n} e^{-E_i}\right) \tag{1}$$

with $E_i = \sum_{j=o}^{i} \epsilon_C(j)$ representing the energy landscape as we have been talking about it so far.
Note that we thus construct this free-energy 'landscape' by combining systems in which everything is mismatched from position n onwards. In this manner- with sufficient mismatch penalties - no state from n onwards will have a significant contribution to the free-energy and you will be probing the on-target landscape.
As will be explained below, Ilya's measurements are essentially designed to return an estimate of this bound free-energy.

I. class I: non-constant ϵ_I: Full R-loop still most stable state with multiple mismatches.

- **Long story short: Took a while, but we realise this is actually not the relevant regime. You may just skip to next section.**
- This is (more or less) the regime Koen's results are in.
- With these values, even after placing a second mismatch the system's most stable state is when a complete R-loop has been formed.
- For as far as the dissociation experiment goes, any mismatched system can basically be seen to have two stable states: One with the completely formed R-loop and the other right before the mismatch.
- Using some toy-modelling - not shown here as this eventually is not the important regime anyways - a good fit to both single and multiple mismatch dissociation rates - at the very least setting the ratio between the maximum values (see the fit) - nesseciates that $\epsilon_I \approx 1.5 k_B T$.
- However, in order to achieve suffiently low association rates and occupancies for PAM proximal mismatches, the ϵ_I needs to be larger.

- As a result, the fit basically used the on-target landscape only to set the overal stability (mostly) and essentially uses the mismatch penalties to create all contrasts between PAM distal and proximal mismatch configurations.

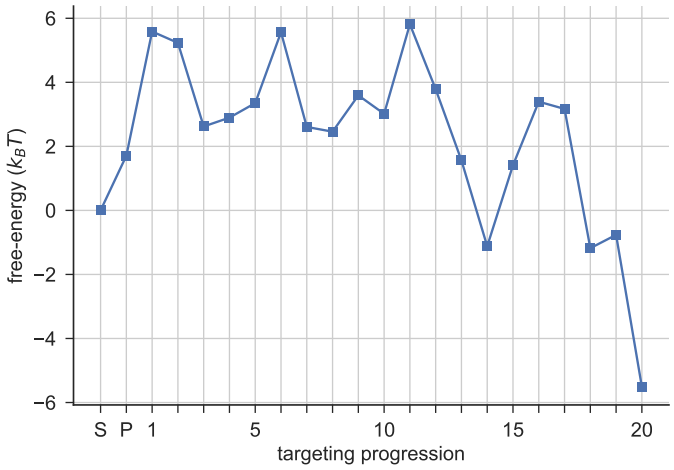
- Further fluctuations in the heatmaps exactly follow the pattern of the mismatch penalties.

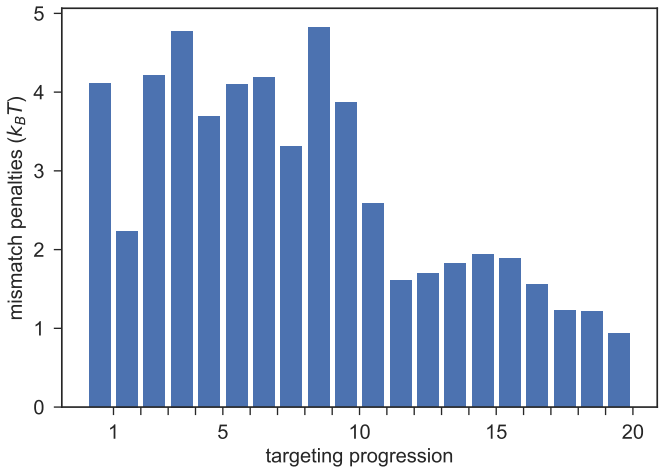
- In this regime, placing more mismatches will decrease the association rate and occupancy, and increase the dissociation rate.

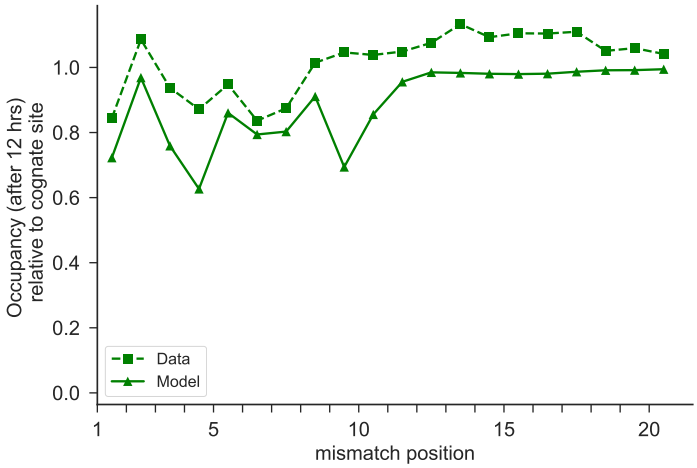
- Not yet a good fit to the single-mismatch association data

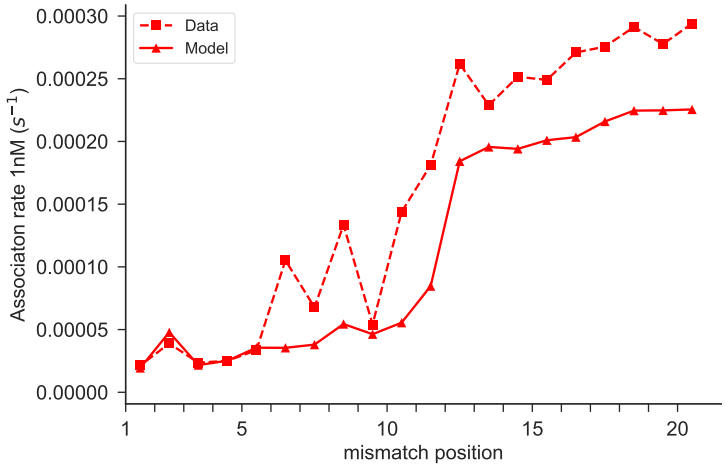
- This is surely inconsistent with Ilya's data. Given the system still has stable states after multiple mismatches, it clearly matters how many will be placed for the effective association and dissociation rates (see the figure with blocks of mismatches in final section).

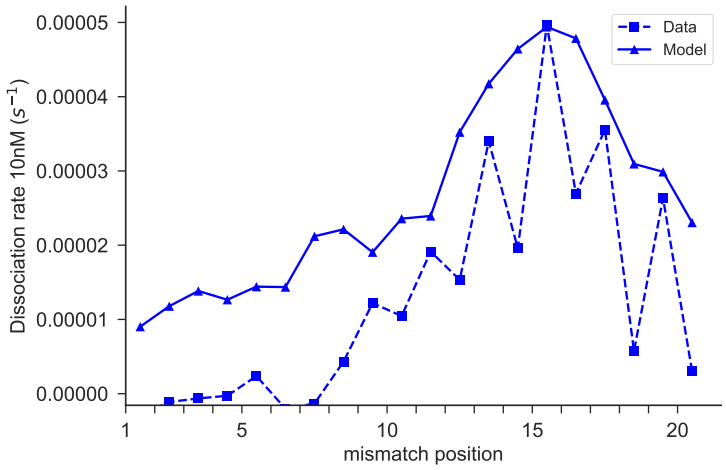
- Also the resulting free-energy landscape for the cognate site does not match that of Ilya's.

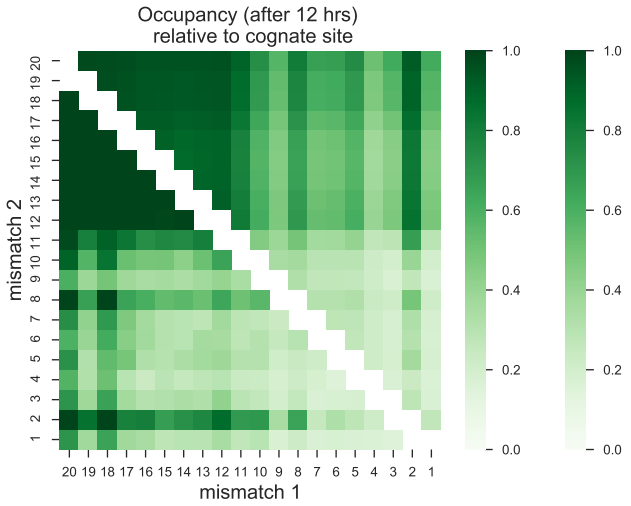


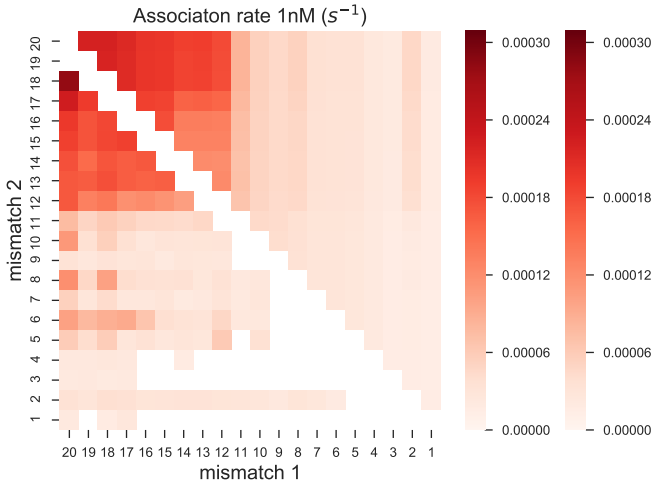


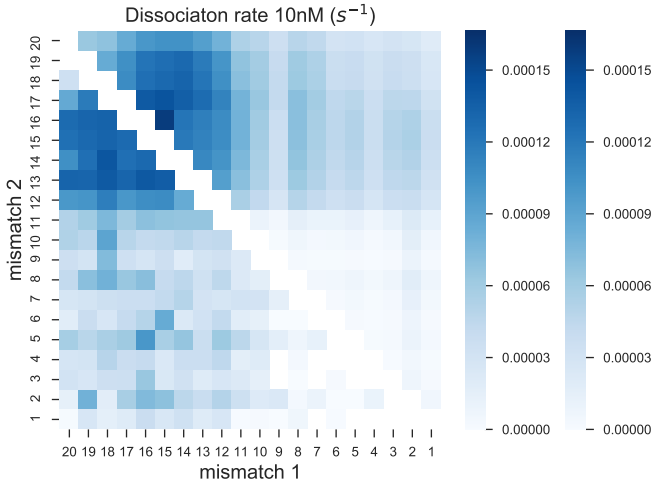

Occupancy (after 12 hrs)
relative to cognate site
1.0
0.8
0.6
0.4
0.2
0.0
Data
Model
1
5
10
15
20
mismatch position


Associaton rate 1nM (s^{-1})
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
Data
Model
1
5
10
15
20
mismatch position


Dissociation rate 10nM (s^{-1})
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
Data
Model
1
5
10
15
20
mismatch position


Occupancy (after 12 hrs)
relative to cognate site
mismatch 2
mismatch 1
20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
1.0
0.8
0.6
0.4
0.2
0.0


Associaton rate 1nM (s^{-1})
mismatch 2
mismatch 1
20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
0.00030
0.00024
0.00018
0.00012
0.00006
0.00000


Dissociaton rate 10nM (s^{-1})
mismatch 2
mismatch 1
20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
0.00015
0.00012
0.00009
0.00006
0.00003
0.00000

II. class II: (near) constant ϵ_I: No more statble states after second mismatch

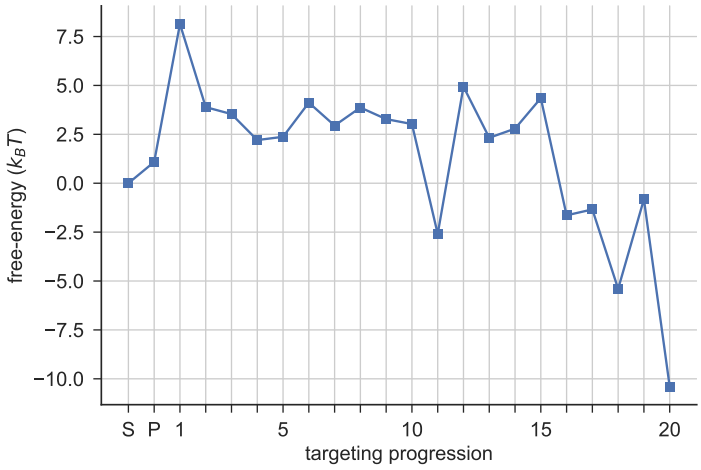
- If all datasets are included as training no fit directly resulted in something of this class. However, it consistently appears when we enforce a constant mismatch penalty (or when we ommit the dissociation data as will be made clear below).

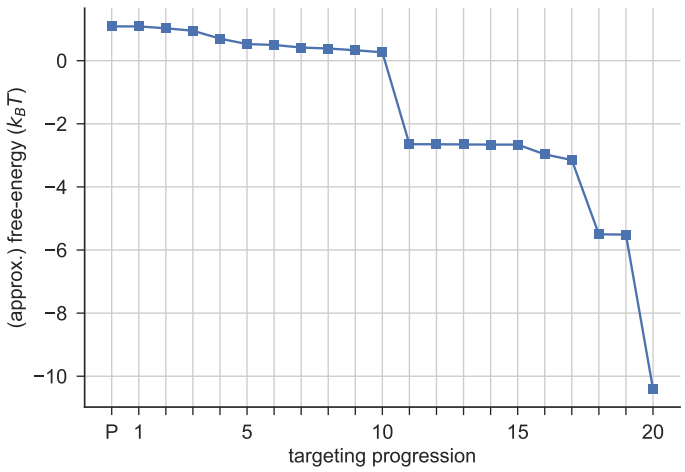
- With these values, placing two (or more) mismatches destabilises the last state in the energy landscape. In other words: forming a full R-loop does not make the system more stable than the solution state.

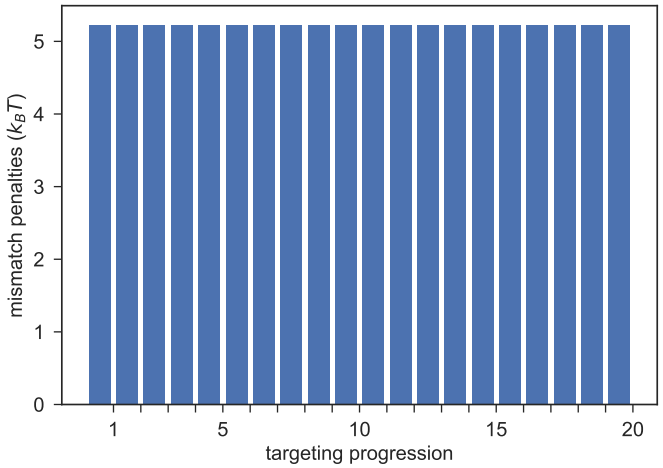
- To fix the dissociation data, the on-target landscape exhibits a sharp dip around position 10. As we will show below, the dissociation data nessiciates that basically the entire gain in stability is placed in 1 step (see the graph of the bound free-energy/ (approx.) free-energy).

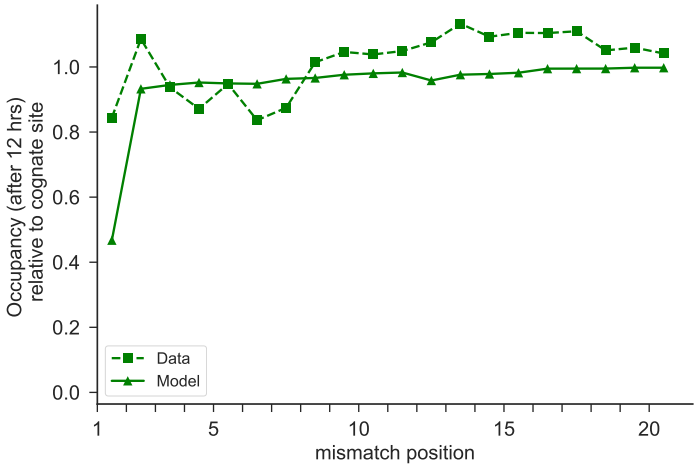
- The resulting free-energy landscape is not consistent with Ilya's data and does not even give such a good fit to Boyle's data.

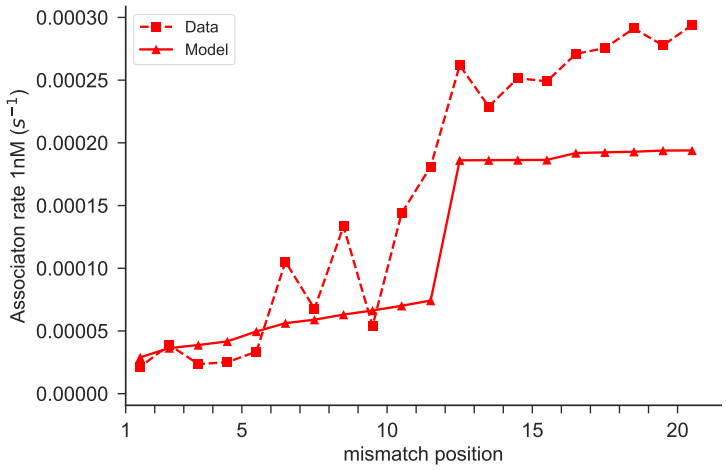
- Still no good fit to the single-mismatch association data

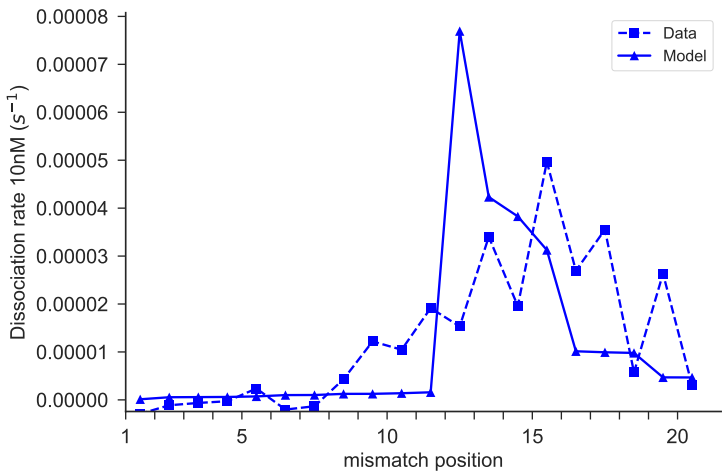


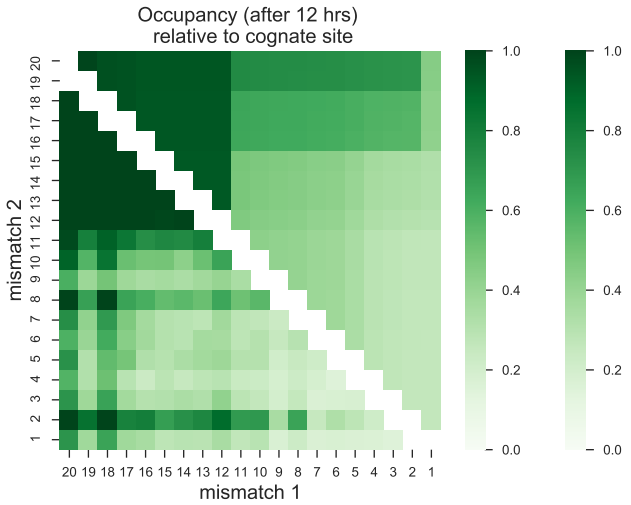


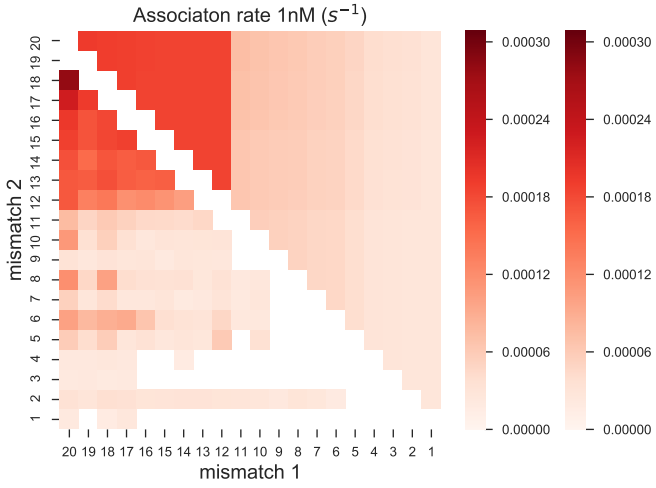

mismatch penalties (k_BT)
0
1
2
3
4
5
1
5
10
15
20
targeting progression

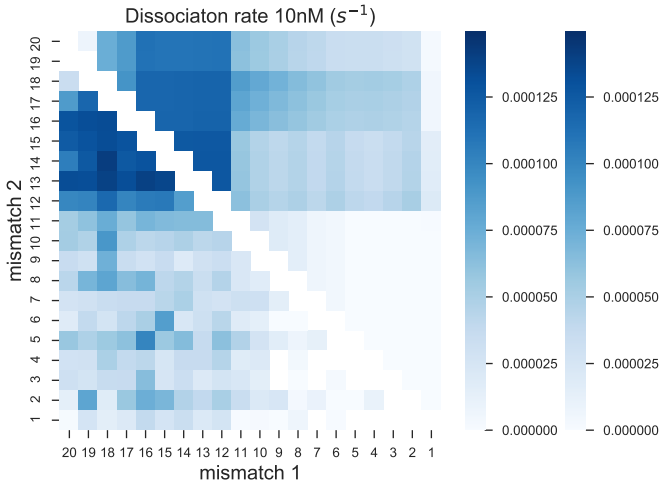

Occupancy (after 12 hrs)
relative to cognate site
0.0
0.2
0.4
0.6
0.8
1.0
Data
Model
1
5
10
15
20
mismatch position


Associaton rate 1nM (s^{-1})
0.00000
0.00005
0.00010
0.00015
0.00020
0.00025
0.00030
Data
Model
1
5
10
15
20
mismatch position


Data
Model
Dissociation rate 10nM (s^{-1})
0.00008
0.00007
0.00006
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
1
5
10
15
20
mismatch position

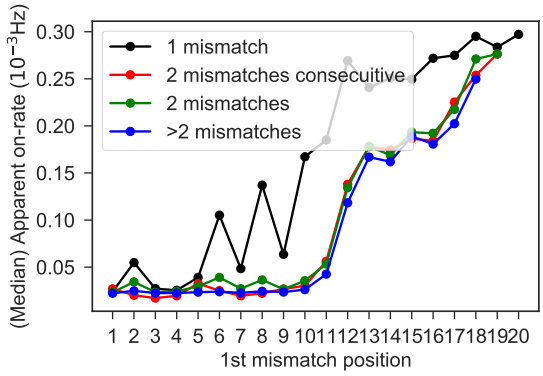

Occupancy (after 12 hrs)
relative to cognate site
mismatch 2
mismatch 1
1.0
0.8
0.6
0.4
0.2
0.0

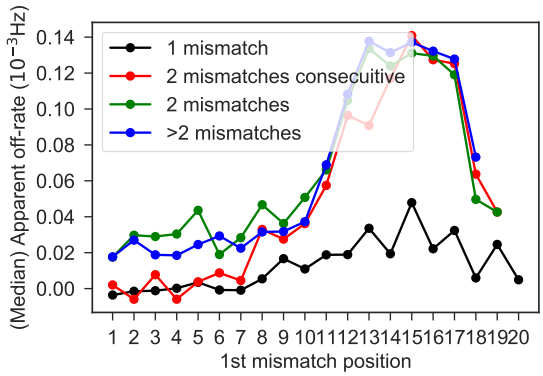

Associaton rate 1nM (s^{-1})
mismatch 2
mismatch 1
0.00030
0.00024
0.00018
0.00012
0.00006
0.00000

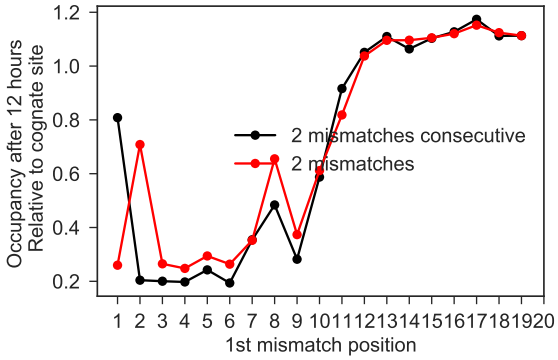


B. HINTS FROM DATA: "TWO-STATE" APPROXIMATION

- Clearly, placing more than two mismatches does not alter the rates and occupancy further (see figures further down in this section).
- Also, the data is similar to that of the mismatches placed after each other.
- Explanation: an effective two-state system when more than one mismatch is placed.
- We imagine all mismatches to be placed straight after each other.
- Also, the system has no stable states from the (second) mismatch onwards. In other words, from two mismatches on, we can imagine there to be mismatches all the way through the hybrid from the first mismatches' position.
- The R-loop will form until the mismatches (before the block)
- Two-states: Solution (or PAM) and the matching part of the R-loop







C. TOY MODEL: TWO-STATE SYSTEM

- Clearly, the data says we need the parameter regime of 'class II'.
- However, we keep failing at getting a consistent fit to all available data.
- We therefore constructed a little two-state 'Toy-Model' to see - given class II - which datasets are actually numerically comptable with one another.
- Solving Master Equation for two-state system:

$$P_{on}(t) = \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} \left(1 - e^{-(k_{\text{on}}+k_{\text{off}})t}\right) \tag{2}$$

I. Predict based on association data

- Assume that the on-rate cannot be sequence dependent. Therefore k_{on} can be estimated based on the measured association rate for the cognate target.
- The association rate is measured by fitting a line through the points $P(t = 500)$, $P(t = 1000)$ and $P(t = 1500)$ and enforcing it to go through the origin. If $P(t)$ is based on the above equation, the association rate equals $k_{\text{two-state}}$, a function of k_{on} - which is known - and k_{off}.
- For every measured association rate. Invert - numerically solve for the roots of -

$$k_{\text{asso}} - k_{\text{two-state}} = 0 \tag{3}$$

- this gives an estimated off-rate, k_{off}, for every association rate
- use the k_{off} values to estimate the dissociation rate (as measured):

$$P_{on}(t) \approx e^{-k_{\text{off}}t} \tag{4}$$

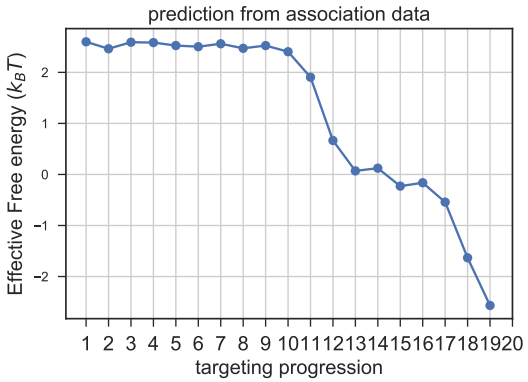
  assuming the two-state approximation. The dissociation rate is again the product of a linear fit to three time points.

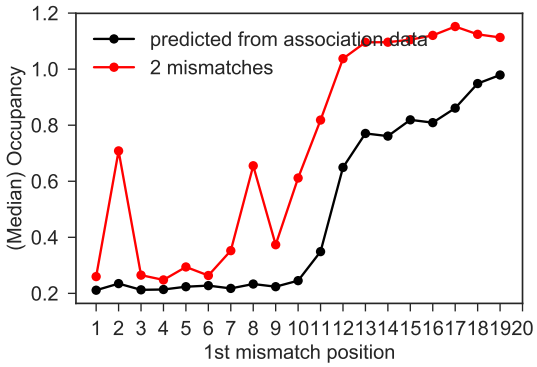
- use the k_{off} values to estimate the free-energy landscape of the on-target. Given the block of mismatches destabilizes the system after it, the free-energy landscape can be constructed by combining the truncated systems. In other words: first mismatch pos 2-20, then 3-20, 4-20, etc.

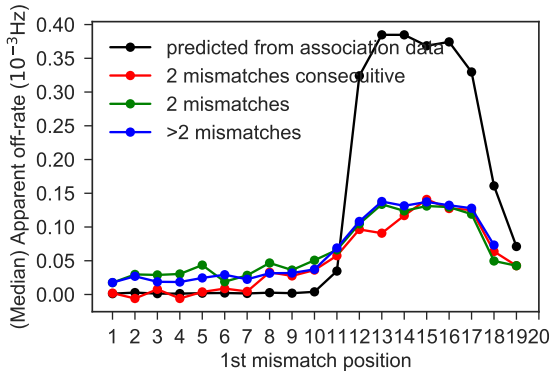
$$\Delta F_n \approx \log(k_{\text{off}}(n)) \tag{5}$$

  with $k_{\text{off}}(n)$ the off-rate from a target with mismatches from position n onwards.
- use the k_{off} values to predict the occupancy after T =12hrs:

$$P_{on}(12hrs) = \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} \left(1 - e^{-(k_{\text{on}}+k_{\text{off}})12*3600}\right) \tag{6}$$



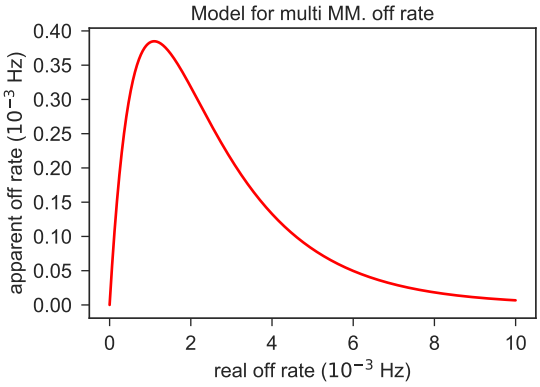


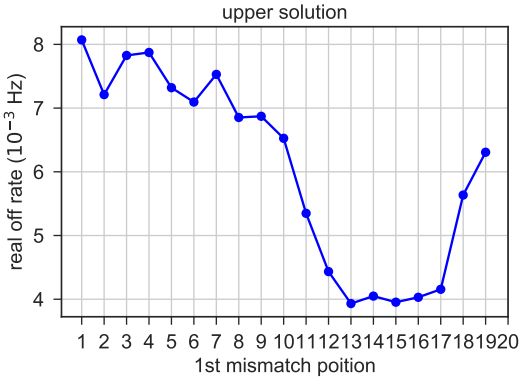


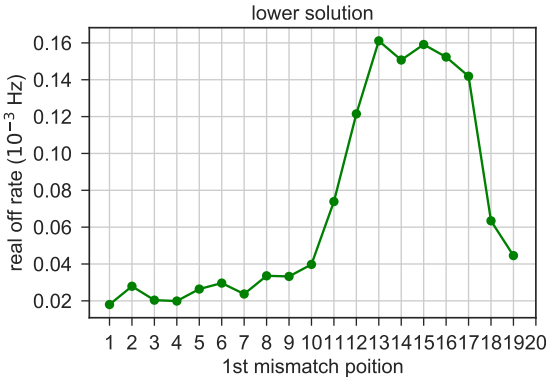
Conclusion: Although no quantitative agreement is achieved, qualitatively the association rate is a decent predictor of the other data.

II. Predict based on dissociation data

- This time we have the measured dissocation rates from which we can - again by root solving - obtain the k_{off} values
- from here we can convert it to assocation rates and occupancies.
- Only thing is, there is a catch: There are two roots (k_{off} values) for every dissocation rate:

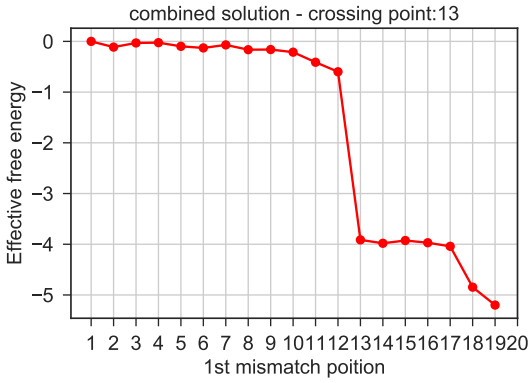






- Hence, after finding both roots we must have a way of deciding which root to use when constructing the free-energy landscape.
- The free-energy is a monotonically decreasing function. Therefore, so must the k_{off} - enforcing us to (at least once) switch between the two possible roots.
- Starting from the upper solution, we see that there are essentially two choices we can make. Either switch around 18 or around 13.

- no mater what choice - switch at 18 not shown - the largest drop in free energy will be whenever this switch is made.
- Long story short... The dissociation data can only be reproduced (quantitatively) if there is 1 single large drop in (free-)energy.
- As the free-energy after this drop remains rather constant, the ϵ_C values should actually be high again: This gives that "zig-zag-like" pattern in the Simulated Annealing fits.



Conclusion: Clearly, the dissociation rate cannot be used as a predictor of the association data and occupancy data. The numbers are simply incompatible.

III. Predict based on occupancy data

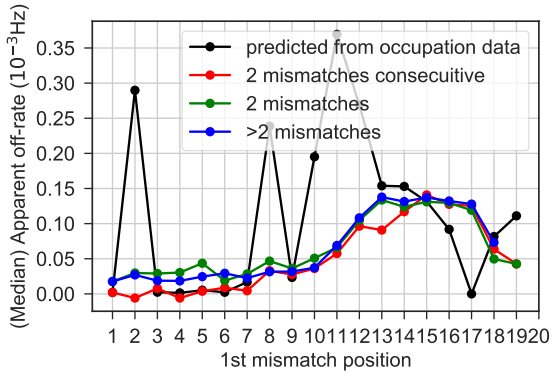
- using the occupation data we obtain the free-energy using (assuming equilibration):

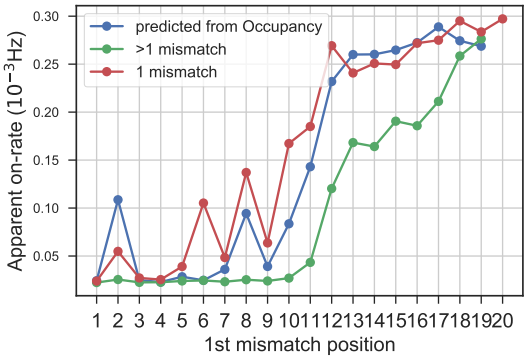
$$P(\Delta F) = \frac{1}{1 + e^{-\Delta F}} \tag{7}$$

  with the solution having an energy of zero.
- From here on, we can convert it to assocation and dissociation rates. First we obtain the k_{off} values.

$$\frac{k_{\text{on}}^{OT}}{k_{\text{off}}} = \frac{P_{off}}{P_{on}} = e^{-\Delta F} \tag{8}$$

- Then, we can estimate the association rate from the k_{off} values - by once again performing the proper linear fit.

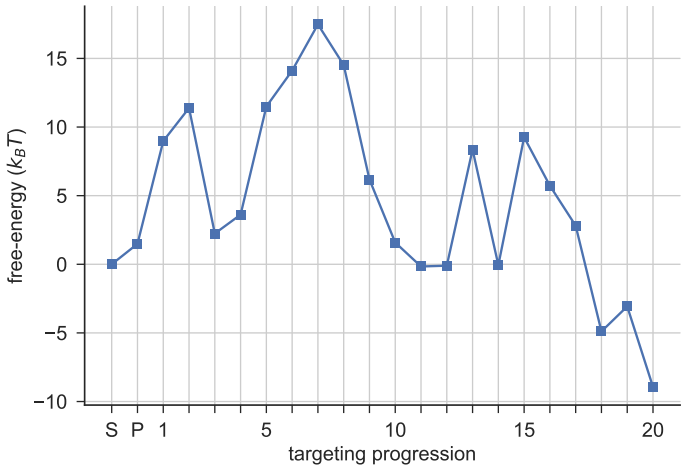


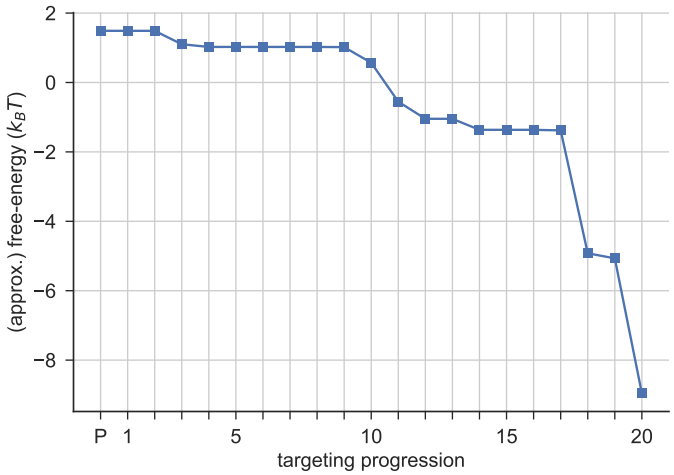


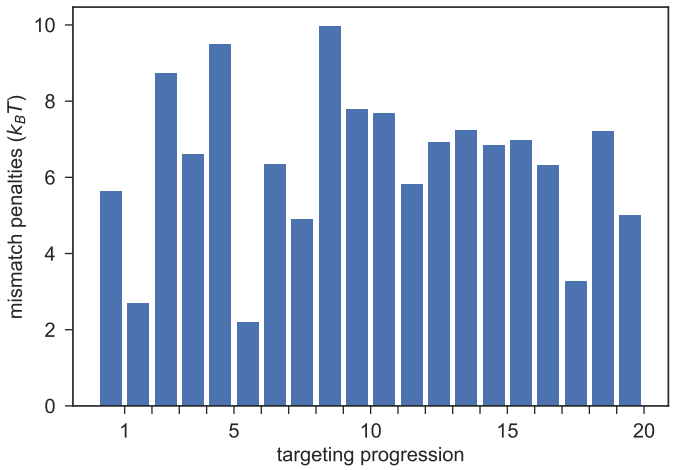
Conclusion: The occupancy data does not contain temporal information. Hence, these numbers cannot be used to accurately predict the association and dissociation data.

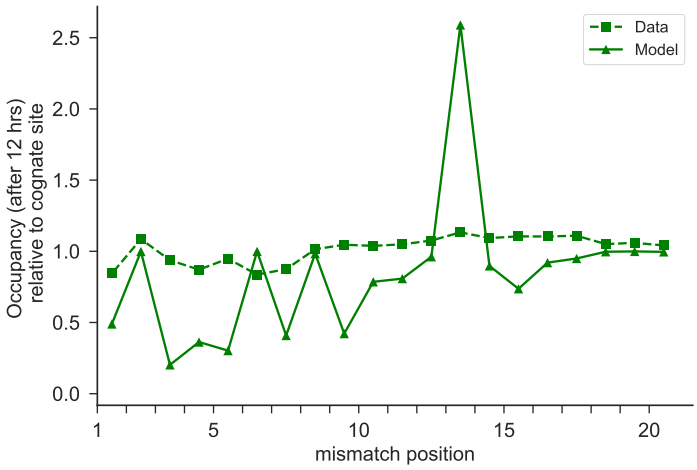
D. SIMMULATED ANNEALING: FIT TO ASSOCIATION DATA CONFIRMS TWO-STATE APPROXIMATION HOLDS

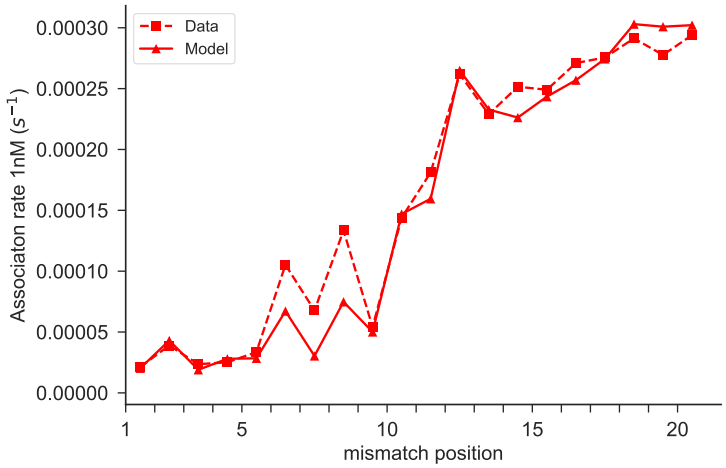
- Fit only to association data (above analysis indicated that this should at least give correct shape of landscape).
- We even get a consistent fit to both single and double (hence multiple) mismatch association rate data.
- Results from the fit look, qualitatively, similar to those from the Toy-model analysis shown above.
- Note that only the association rate is fitted. All other plots actually are predictions.
- FYI: We had quite loose bounds for the rates and the fit decided to use such large numbers that sometimes the occupancy spikes above 1 due to numerical errors in the algorithm for the matrix exponential. We started a fit with tighter bounds.. no worries :).

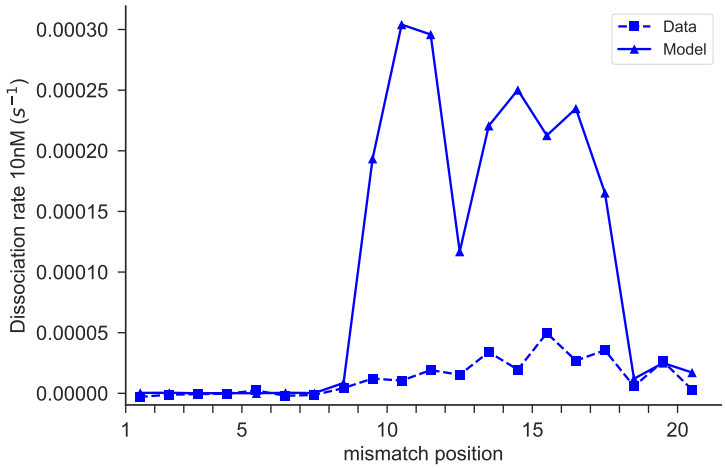


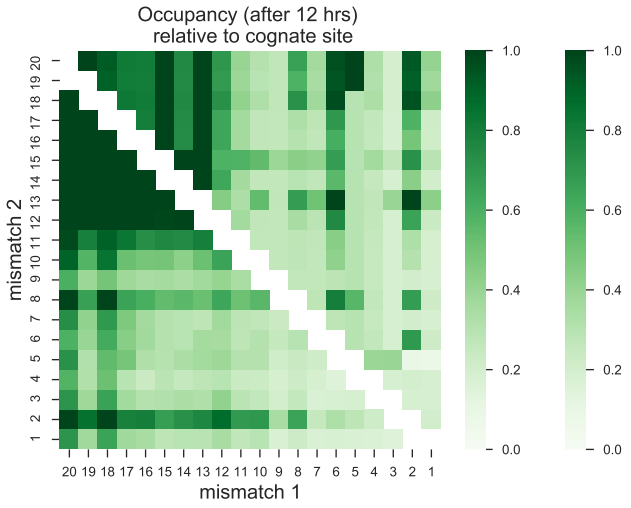

2
0
−2
−4
−6
−8
(approx.) free-energy (k_BT)
P
1
5
10
15
20
targeting progression

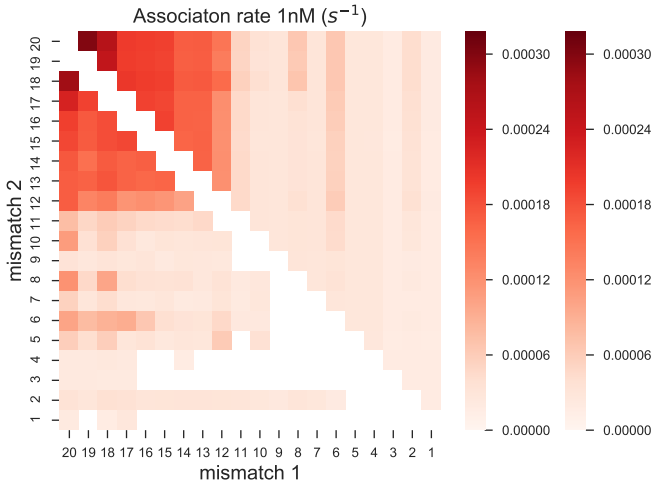

10
8
6
4
2
0
mismatch penalties (k_BT)
1
5
10
15
20
targeting progression

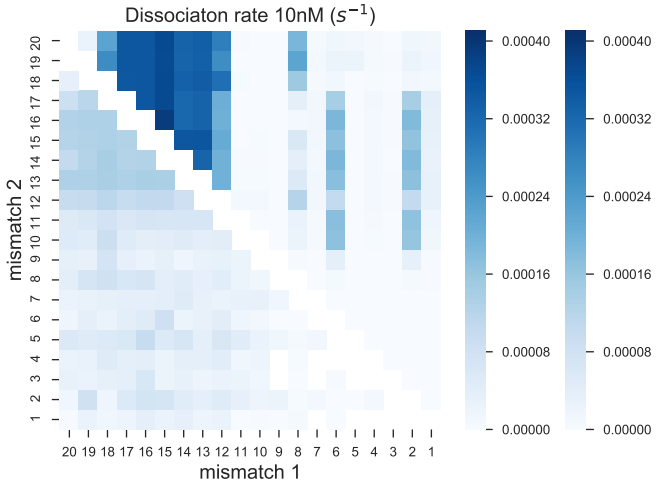

Data
Model
2.5
2.0
1.5
1.0
0.5
0.0
Occupancy (after 12 hrs)
relative to cognate site
1
5
10
15
20
mismatch position


Data
Model
Associaton rate 1nM (s^{-1})
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
1
5
10
15
20
mismatch position


Data
Model
Dissociation rate 10nM (s^{-1})
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
1
5
10
15
20
mismatch position

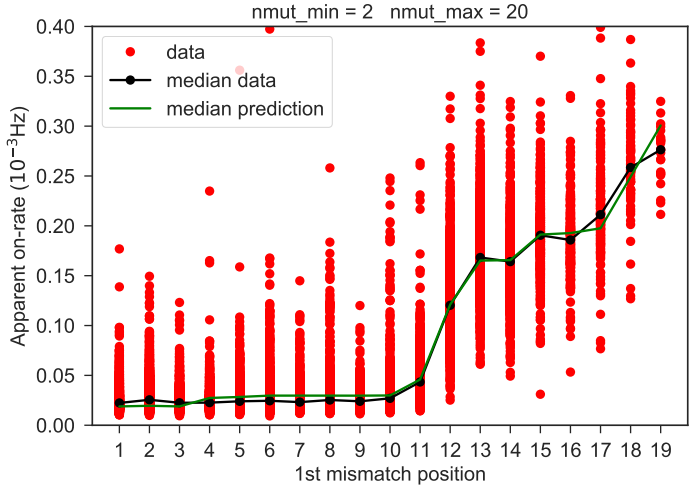

Occupancy (after 12 hrs)
relative to cognate site
mismatch 2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
mismatch 1
1.0
0.8
0.6
0.4
0.2
0.0

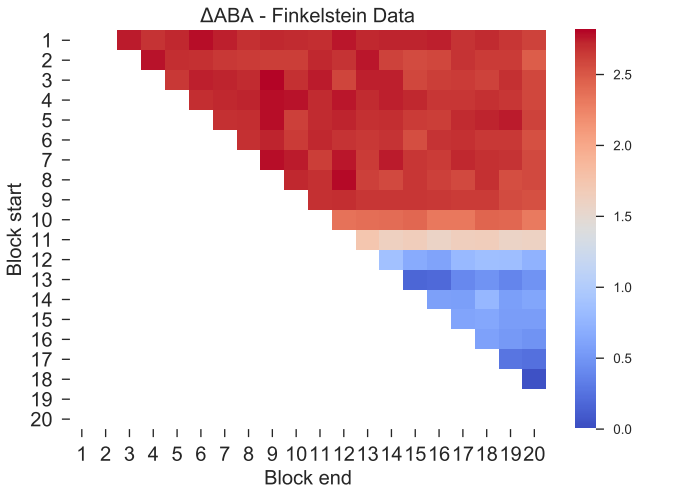
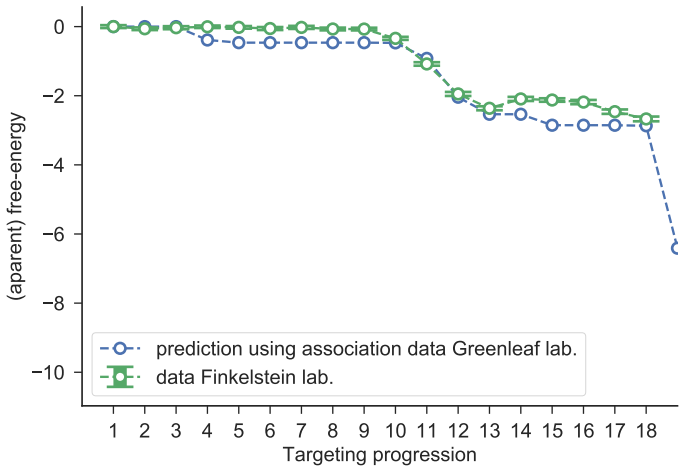




E. PREDICTING MULTIPLE MISMATCHES

We can predict the association rate for multiple mismatches. Note that these have not been provided as training.



F. RESULTING FREE-ENERGY LANDSCAPE IS CONSISTENT WITH ILYA'S DATA

- They assume equillibration and - using the Hill equation- estimate the dissociation constant.
- The log of this dissociation constant approximates the bound free-energy. This is the reason we have been plotting them as done in the previous sections.
- This Aparant Binding Affinity (ABA) is then compared to the value of the cognate site. Hence, the on-target has free-energy - or ΔABA - of zero.
- Not used here, but we have a code in place that mimics his experiment and does not need to assume the system to be equillibrated. Indeed, when looking at his figures - even those published - the fit to the Hill Equation is not impressive. A more accurate K_D estimate can be obtained by solving for the occupation after 10 minutes (the time at which they took their data).
- One particularly nice experiment for us is where they analized targets with blocks of mismatches. Due to their high-throughput measurements, they have data for blocks of all lengths and starting positions.
- Interestingly, also Ilya's data shows that placing more than 2 mismatches does not further destabilize the system significantly. Again hinting towards 'class II'.
- the figure superimposes the free-energy landscape obtained from Simulated Annealing - only using Boyle's association data - and Ilya's experimental estimate of the same.
- **Conclusion: Without any imput from Ilya's data - no parameter estimation - we actually nearly manage to reproduce his signal based on Boyle's data.**