# Project 1: Use ABA data as training to determine free-energy landscape. [Bsc project]

* Can we use CHAMP (ABA) data to train our physical model?
  + Get free-energy landscapes for both on-target and off-target at concentration x and for guide sequence y.
  + If not possible, suggest additional experiments.
* Landscape (d)Cas12 vs (d)Cas9
* Comparing (d)Cas9 variants (engineered)
* Can we use the landscape from target D + sequence dependency and retrieve the landscape of target E?
* Do indels behave as mismatches? [**not immediate**]

## Data needed:

* dCas9 + dCas12 + dCas9 engineered variants
* For dCas9: Target E + Target D
* mismatches + indels
* for every sequence. The bound fraction for each used concentration:   
  seq | concentration 1 | concentration 2 | …..   
  ------------------------------------------------------------

AATCGG | (value1, error1) | (value2 , error2) | …

* 1 value per sequence (cluster) and concentration is perfect (median for instance).
* For improved fitting, the error (standard deviation) for each sequence (cluster)

# Project 2: Predicting cleavage propensity from binding [Bsc project]

* Can we translate the landscape of dCas9 and predict cleavage behavior of Cas9?
* How can we modify the landscape of (d)Cas9 to capture that of the (d)Cas9 variants?
* Possible quick look at sequence dependency?
* (Genome-wide) off-target prediction:
  + PR curves, PR curves and more PR curves…. (different guides/targets, compare to existing prediction tools, different genomes perhaps…)
  + Use other prediction tools to predict NucleaSeq data
  + Make code more efficient were needed

## Data needed:

See projects 1 and 3

# Project 3: Use NucleaSeq data to determine free-energy landscape directly [Msc project]

* Can we use the NucleaSeq data (cleavage) as training to get the landscape?
* If this works: Compare landscapes Cas9, engineered Cas9 and Cas12
* Did project 2 actually work? (This is perhaps the best way to find out if we got all information out of the data that it contains)

## Data needed:

* Cas9 + Cas12 + Cas9 engineered variants Primarily target E needed. Target D will be used later for sure
* NucleaSeq data [Si]t/[Si]0 versus time for (mis)matched sequences   
  seq | time 1 | time 2 | …..   
  ------------------------------------------------------------

AATCGG | (value1, error1) | (value2 , error2) | …

* 1 number per sequence per time point is good.
* Error (standard deviation) for each point if possible
* For comparison: Fitted rates for every sequence. Do you have error bars here (say, by using bootstrapping and re-fitting the exponential)?   
  seq | fitted rate| error  
  ------------------------------------------------------------

AATCGG | rate | error