# Immediate goals

1. Use currently predicted (d)Cas9 free-energy landscape to predict CHAMP data
2. Construct a pipeline to get the free-energy landscape (model parameters) from CHAMP data.
   1. With this we should be able to compare different (d)Cas9 variants, (d)Cas9 to (d)Cas12 etc.
3. get the free-energy landscape for active Cas9 (variants) from (d)Cas9 data.
   1. First step: Can we morph the landscape of WT Cas9 into something that explains the engineered Cas9’s behaviours?
4. Construct pipeline to get the free-energy landscape by training against NucleaSeq data.
   1. Another route towards answering above questions: Comparing Cas9 variants, Cas12, etc. on physical grounds
   2. Serves as a means of directly obtaining free-energy landscape engineered cas9 without knowledge of wildtype or deactivated protein.

# Future steps

1. Incorporate sequence dependency.
2. Differences in guide-sequence: Are these explainable using sequence dependency? (e.g. target E vs target D)
3. Can insertion/deletions be modelled just like sequence mismatches?

# Data needed

## CHAMP:

* 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)

## NucleaSeq:

* 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