# Custom array design

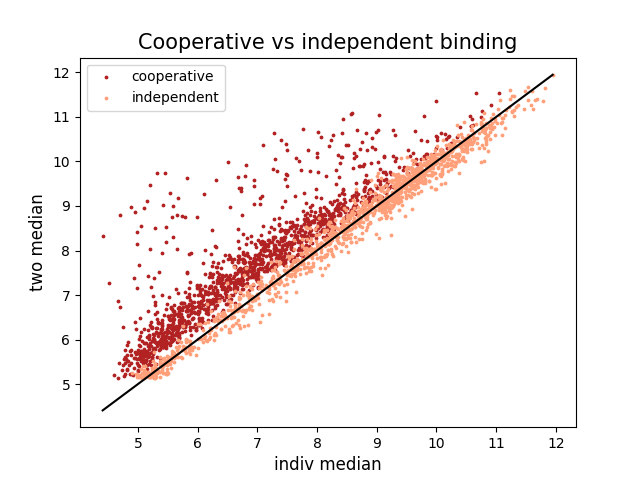
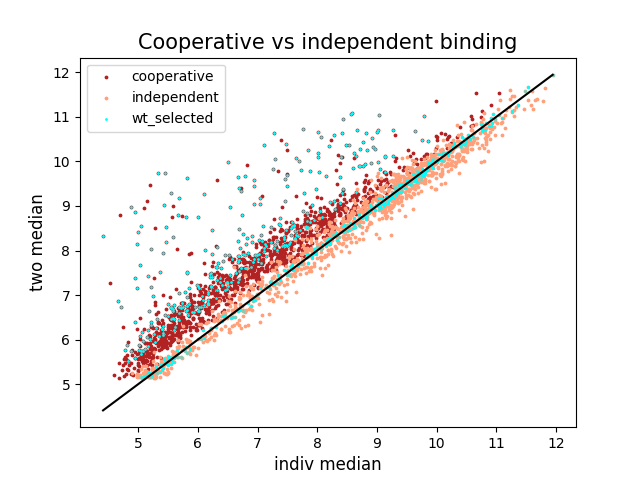
## Overview

We take 350 points with the largest diagonal distance (i.e. the most cooperative) and 350 with the smallest diagonal distance (i.e. the most independent) from the input data points.

## Ets1-Ets1

**Original plot**

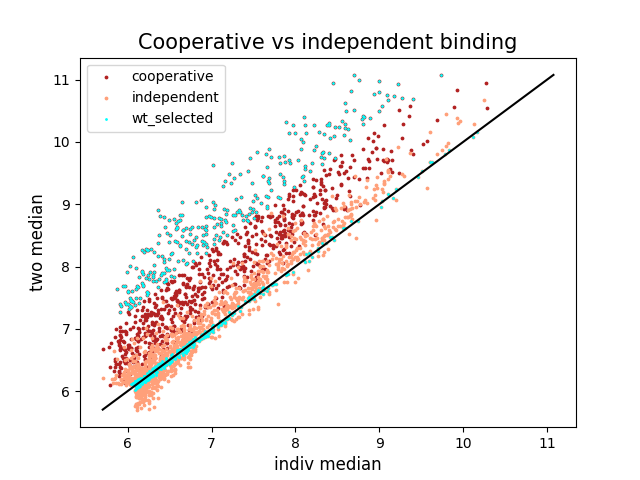
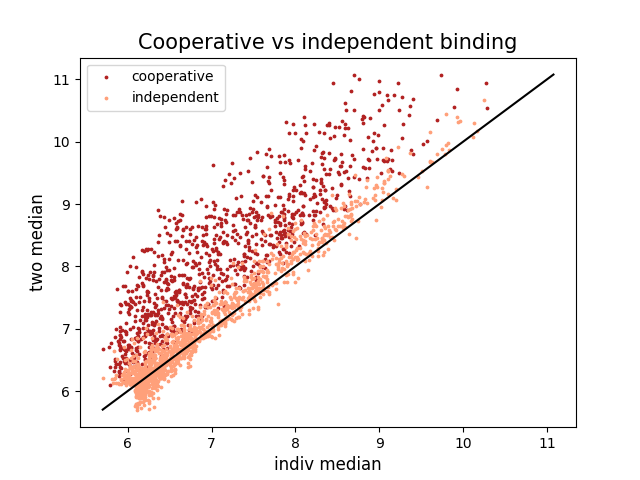
**Selected**

## Ets1-Runx1

**Original plot**

**Selected**



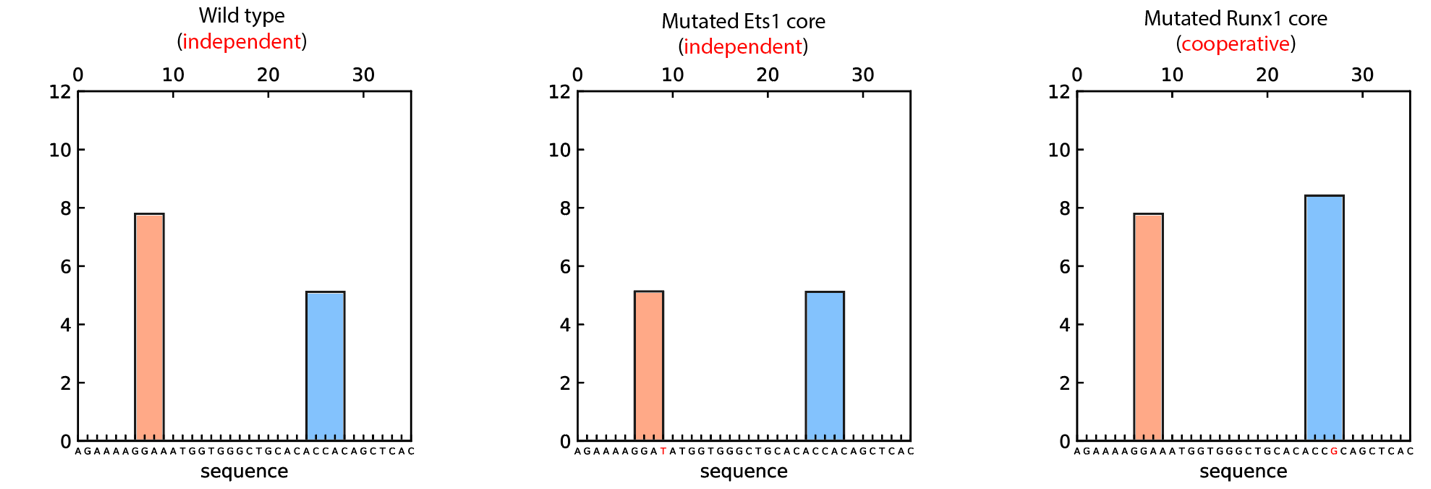
# Making custom sequences (i.e. making mutations)

We aim for mutations that change one feature with little to no change to the other features. Furthermore, for every mutation, we made sure that no new sites are created. We eliminated the mutants if they affect the other site or create new sites.

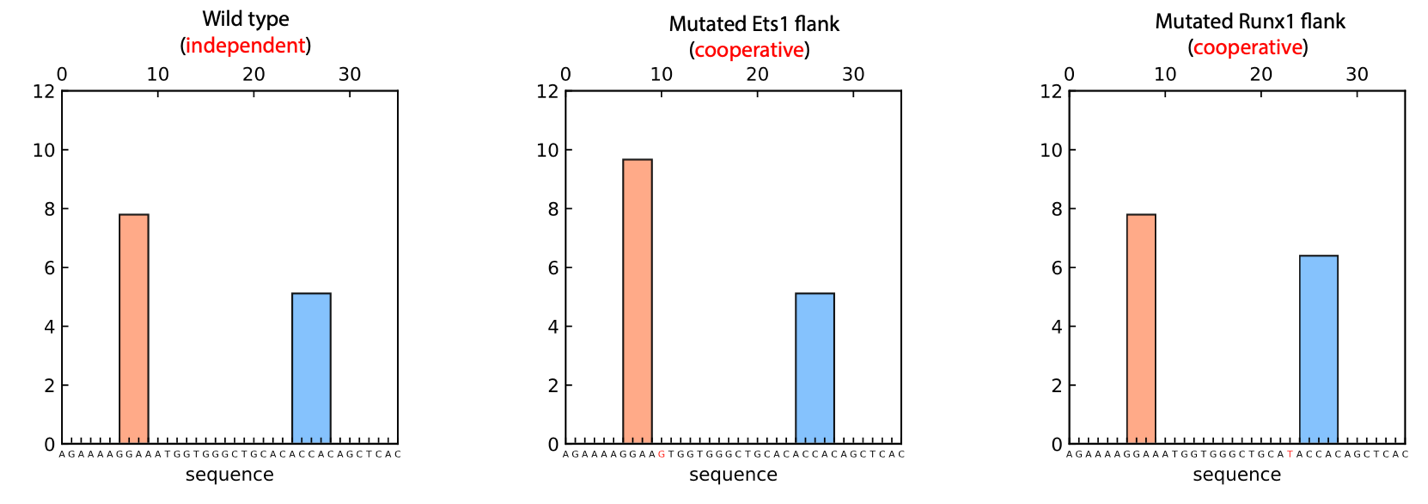
Note: In the Ets1-Runx1 analyses, the first site is always Ets1. For the plot, Ets1 is denoted with red and Runx1 with blue

## Mutating strength

1. Change binding site cores

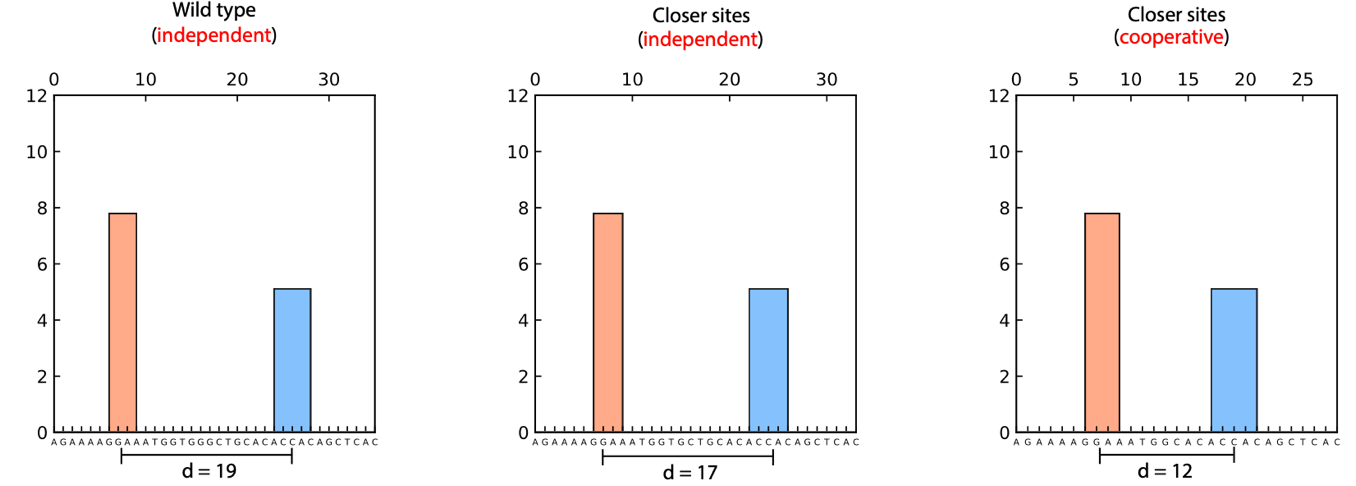


1. Change the flanking regions



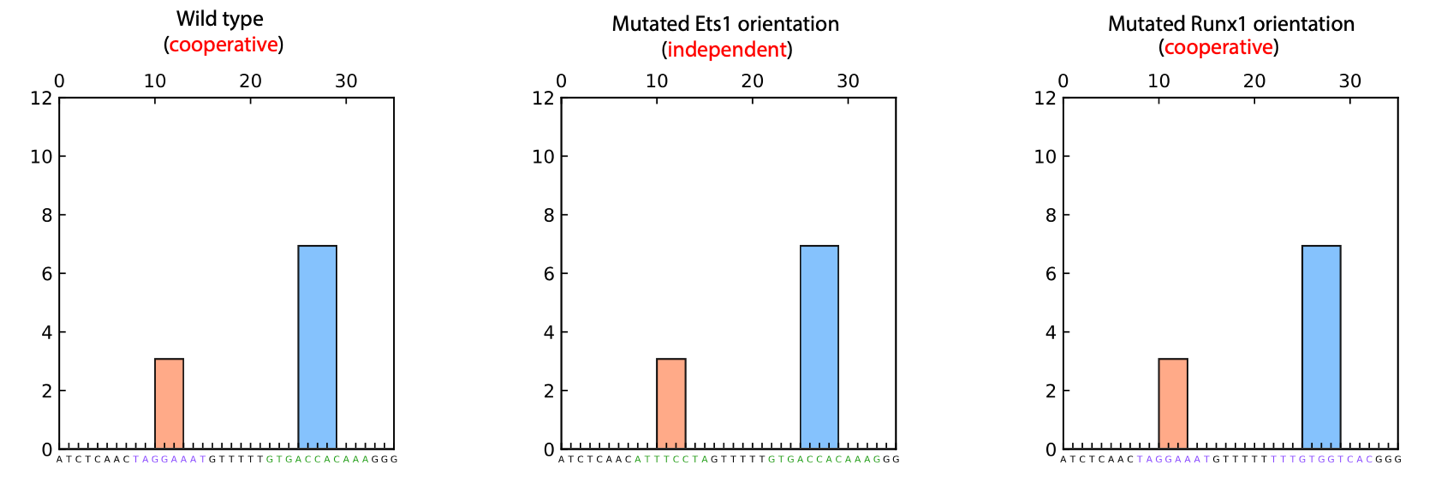
## Mutating distance

Make the sites closer through deletion of the middle bases. The mutants are shorter because we are not appending. We keep making the sites closer until the sites are 5bp apart.



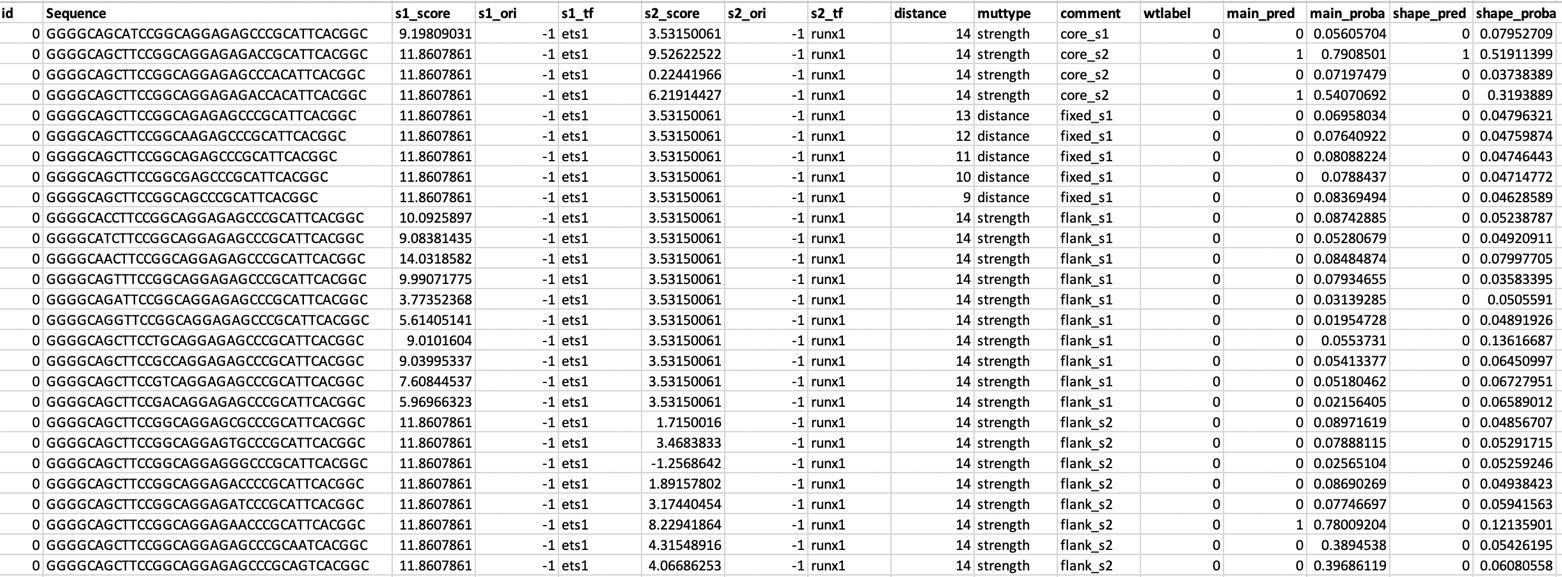
## Mutating orientation

We mutated the sites by taking the reverse complement of the core and its flanking regions.



## Result

The resulting table consists of all possible mutations that fulfill the constraints, i.e. they do not create / eliminate sites and have minimum to no change to the other site). Along with the predictions of the random forest models.

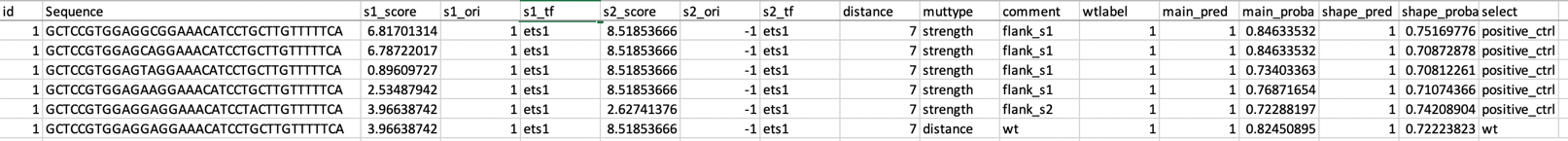


# Selecting custom sequences

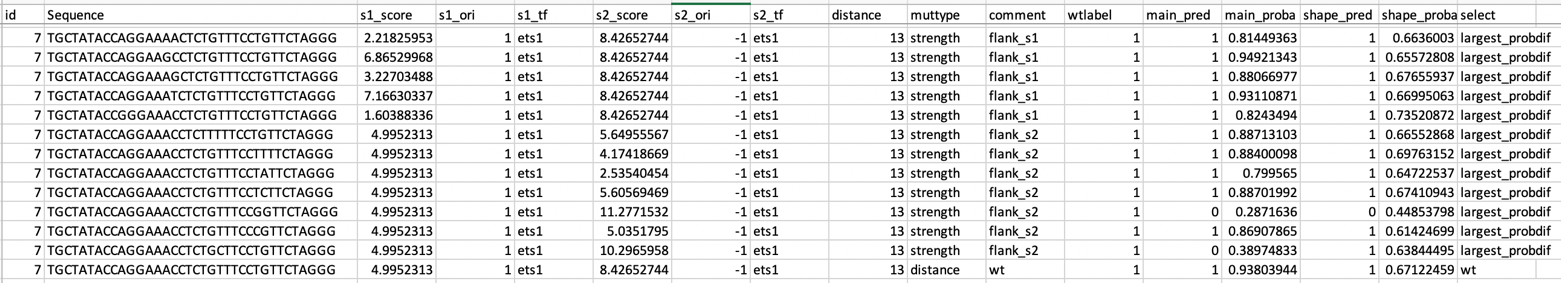
From all possible mutants, we select sequences where there are changes in the predictions among the mutant groups.

We select mutations that:

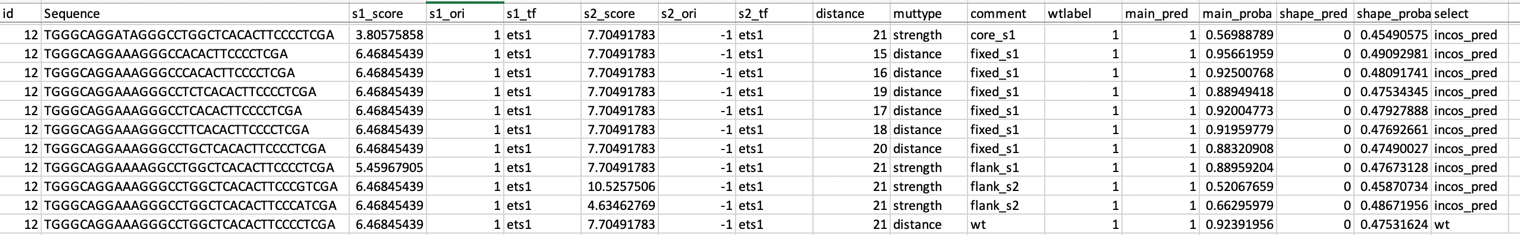
1. Have high probability (i.e. cooperative) or low probability (i.e. independent), also agree between the two models 🡪 positive control



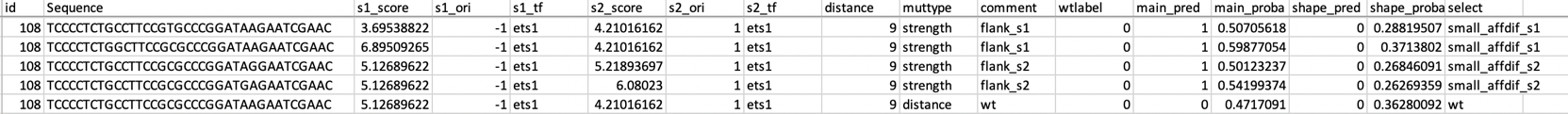
1. Have large probability difference



1. Inconsistent between main and shape models



1. Small strength difference that change the main predictions (small: <= 2 difference in PWM score)



## Results

**Filter by pwm vs imads delta:**

**Ets1-Runx1**

Number sequences 7520

Rsq Ets1 delta 0.8774148649431964

Rsq Runx1 delta 0.6313683610377285

**Ets1-Ets1**

Number sequences 15557

Rsq Ets1 delta s1 0.875032364302305

Rsq Ets1 delta s2 0.8733125875512251

**Ets1-Runx1**

Number of correct predicted cooperative wt, main\_pred: 249/250

Number of correct predicted independent wt, main\_pred: 187/190

Number of correct predicted cooperative wt, shape\_pred: 248/250

Number of correct predicted independent wt, shape\_pred: 185/190

**Ets1-Ets1**

Number of correct predicted cooperative wt, main\_pred: 293/335

Number of correct predicted independent wt, main\_pred: 308/338

Number of correct predicted cooperative wt, shape\_pred: 213/335

Number of correct predicted independent wt, shape\_pred: 309/338

**Final array count:**

neglen ets 171

neglen runx 253

84654 spots for er

88824 spots for ee

Ets1-Ets1 count 3701

Seqlen after joined 3701

strength 2249

distance 1461

orientation 501

literature 4

Ets1-Runx1 count 4703

strength 4612

distance 1069

orientation 668

literature 8

**Probes from literatures:**

Wotton94,GATCCACAACAGGATGTGGTTTGACATTTA,er

Wotton94,ACCGGAAGACGGTTGAACCACATT,er

Wotton94,CACTTCCGCTAAAATACCGCATT,er

Wotton94,ACCGGAAGACGGTTGAACCACATT,er

Wotton94,CACTTCCGCTAAAATACCGCATT,er

Wotton94,ACCGGATGTGTGGTAT,er

Shrivasta14,AGAGGATGTGGCTTCC,er

Tcra,CAGAGGATGTGGCTTC,er

Baillat02\_wt,ACCAAGACAGGAAGCACTTCCTGGAGATTA,ee

Baillat02\_dr,ACCAAGACAGGAAGCCAGGAAGTGAGATTA,ee

Baillat02\_ip,ACCAAGAGCTTCCTGCAGGAAGTGAGATTA,ee

Baillat02\_wt4,ACCAAGACAGGAAGCACGCACTTCCTGGAGATTA,ee

Lamber08,TAGACAGGAAGCACTTCCTGGAG,ee