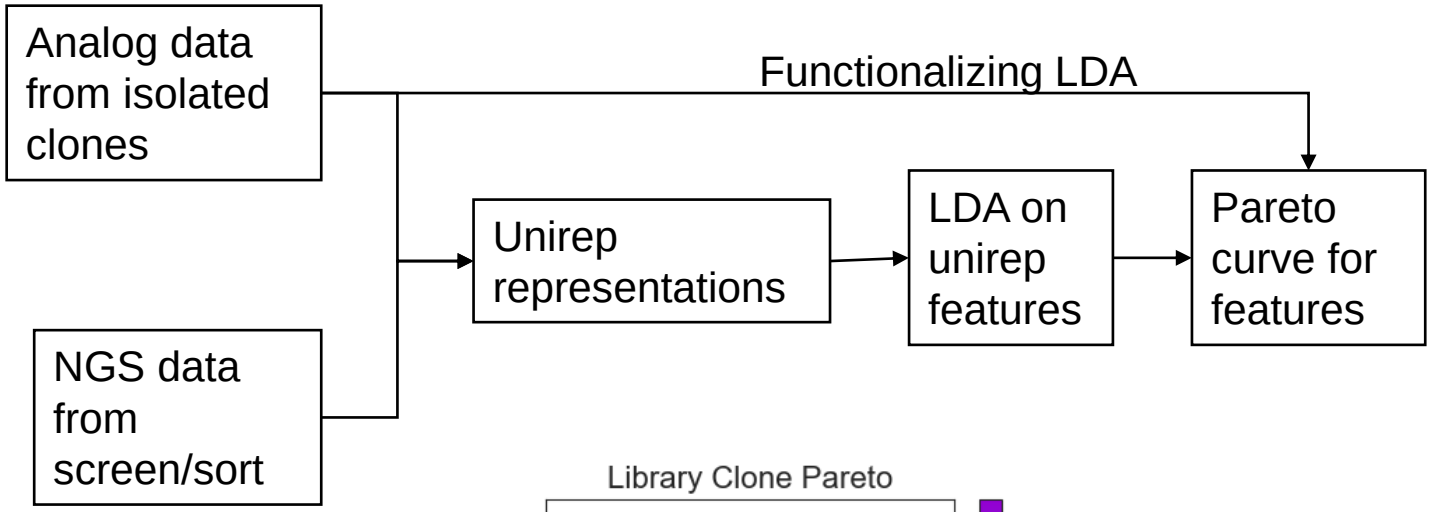
8.19.20 Update

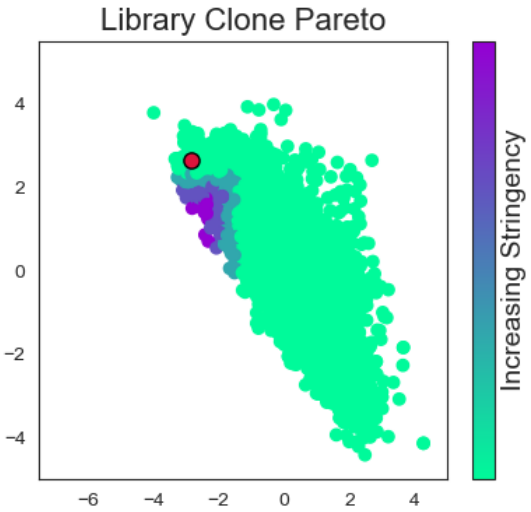
Emily Makowski

Patrick Kinnunen

Emi Specificity

Process flow chart – Pareto optmization





Next steps: Proposing new clones

- Simplest way: start from some of the best isolated sequences, and try all possible single mutations
- Computationally tractable, fits with Emily's leave-one-out analysis

LHQWY…	Base sequence
AHQWY…	Mutation 1 at site 1
BHQWY…	
…	
LAQWY…	Mutation 1 at site 2
LBQWY…	
…	
LHQWV…	Mutation 19 at site N

Constraints:

- No new cysteine
- Only new mutations in CDR H2 and CDR H3 were considered
- WT + 6 different isolated clones were used as base sequences

$$[\mathbf{18}\ new\ AAs] * [\mathbf{25}\ CDR23\ sites] * [\mathbf{7}\ base\ sequences] = \mathbf{3150}\ \boldsymbol{new\ sequences}$$

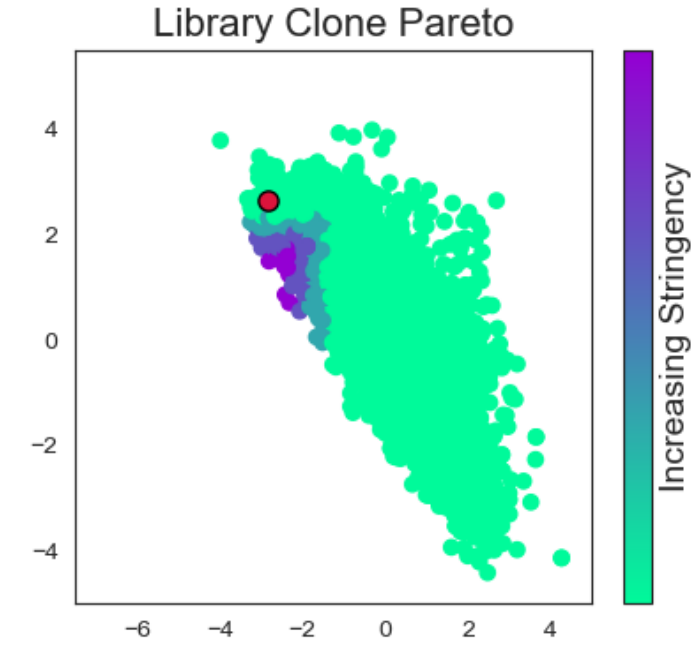
- Comparing pairwise mutations: ~56k sequences

Comparing new and old clones

LHQWY…	Base sequence
BHQWY…	Mutation 1 at site 1
CHQWY…	
…	
LAQWY…	Mutation 1 at site 2
LBQWY…	
…	
LBQWV…	Mutation 19 at site N

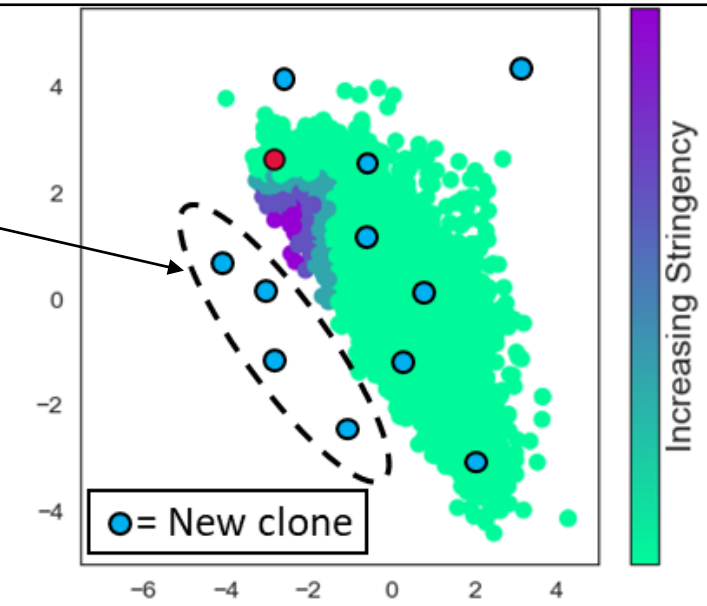
↓

Get ~3500 unirep representations



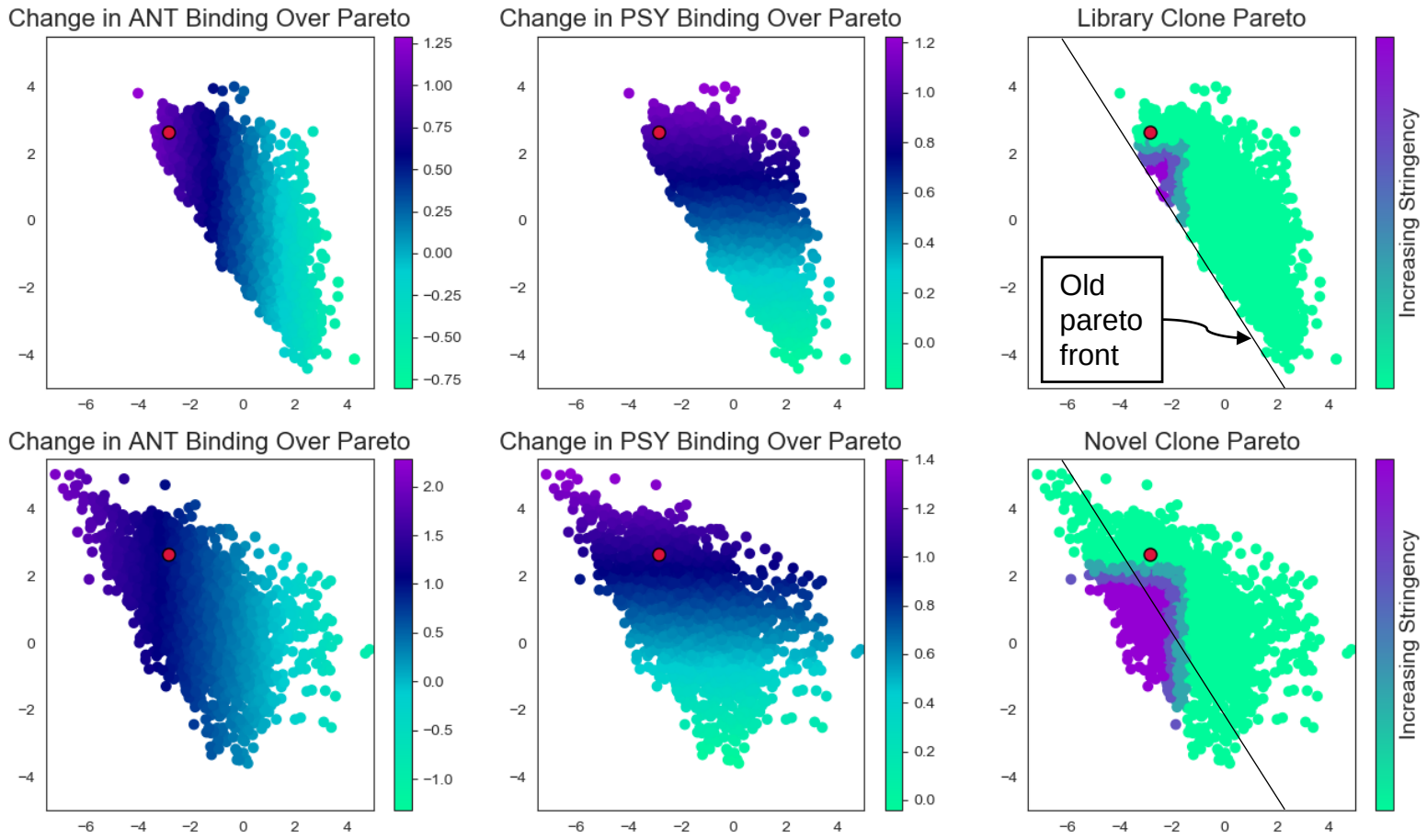
With **hypothetical** new clones

Good new clones

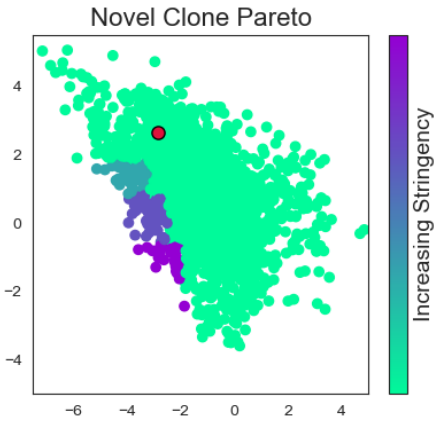


Proposed clones move the pareto front

By picking new mutations, we get many different clones that are past the original pareto front



200 Clones on the new pareto front



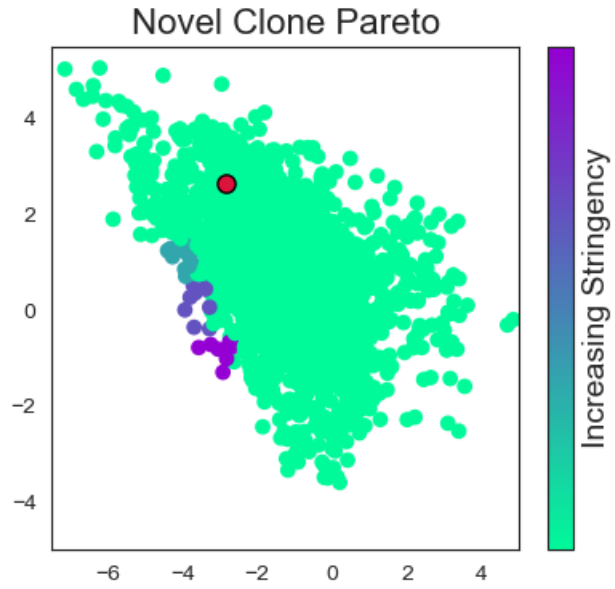
- WT can be made better by adding W to HCDR2
- W > F > M > L are the most common residues added
- 52A(P) > 60(N) > 57(T) > 58(T) are the most common residues mutated away
- EM8 > EM45-005 are the most common base sequences found on the pareto frontier

	33	50	54	55	56	93	97	102
WT	Y	R	R	R	G	A	W	Y
EM45-005	Y	R	T	A	G	A	L	Y
EM8	Y	R	A	G	G	A	A	F

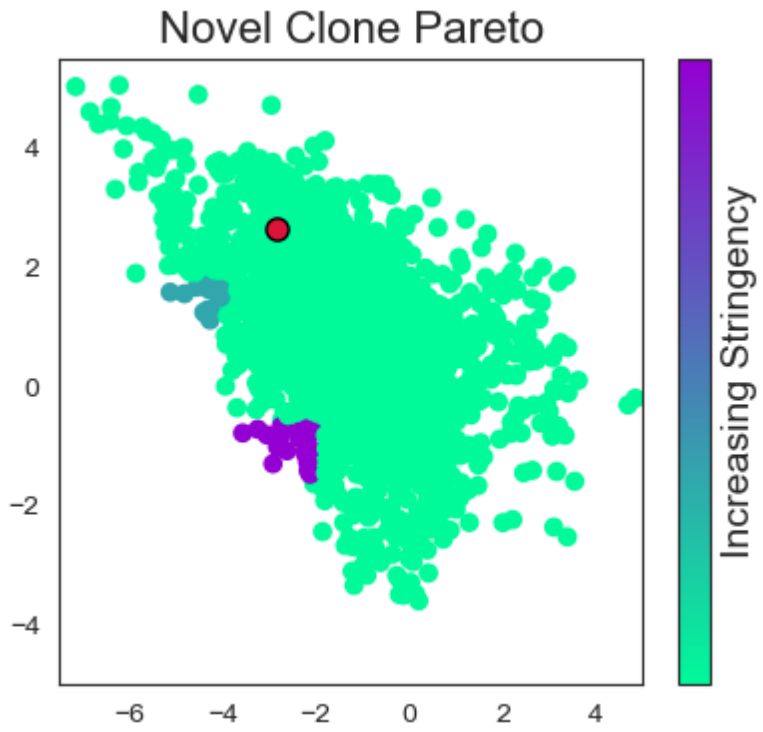
44 clones on the pareto front

- W > L = N = E = F are the most common residues added
  - Mostly W
- 52A(P) is the most common residue mutated away from
- EM45-005 > EM8 are the most common base sequences found on the pareto frontier

- We could clone these for evaluation
  - After



41 clones to make and evaluate



- 2 populations – focused on either antigen binding or specificity
  - Evaluate the tunability of the analysis
  - Evaluate the quality of analysis for different properties
    - Currently we think it's stronger for antigen binding than polyspecificty binding

Process flow chart – Pareto optimization

Propose new mutations

Unirep Representations

Analog data from isolated clones

Functionalizing LDA

Unirep representations

LDA on unirep features

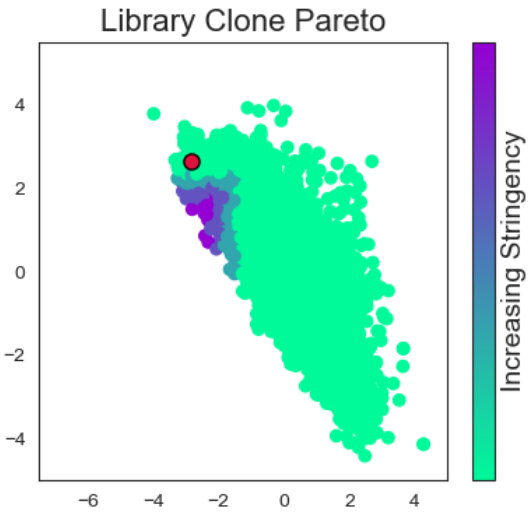
Pareto curve for features

Plot new mutations on pareto curve

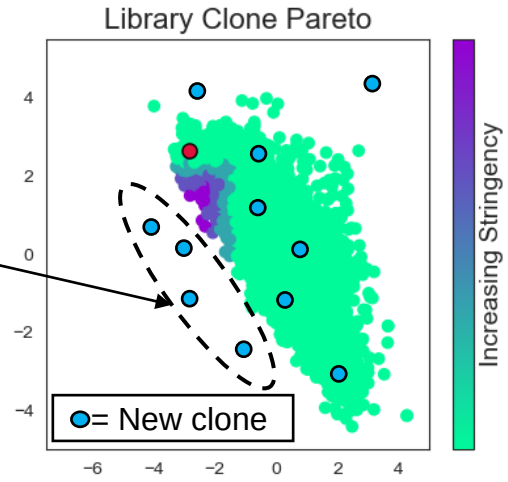
Analyze suggested “good” new clones

New suggested mutations/sites

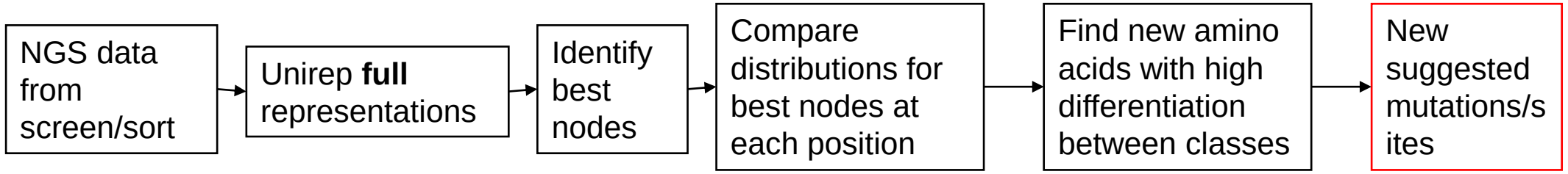
NGS data from screen/sort

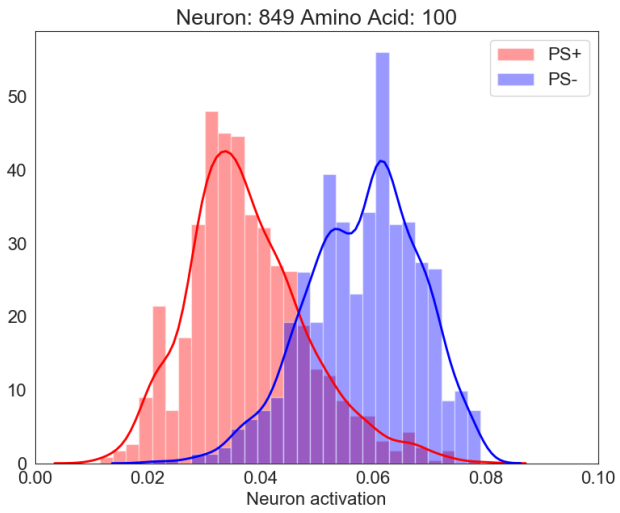


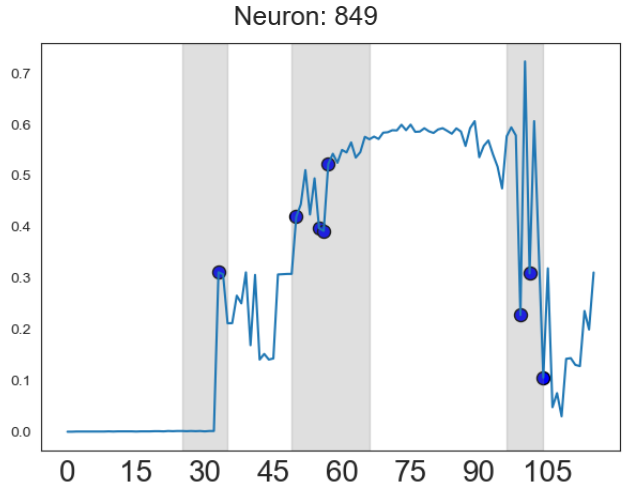
Good new clones



Process flow chart – Feature analysis

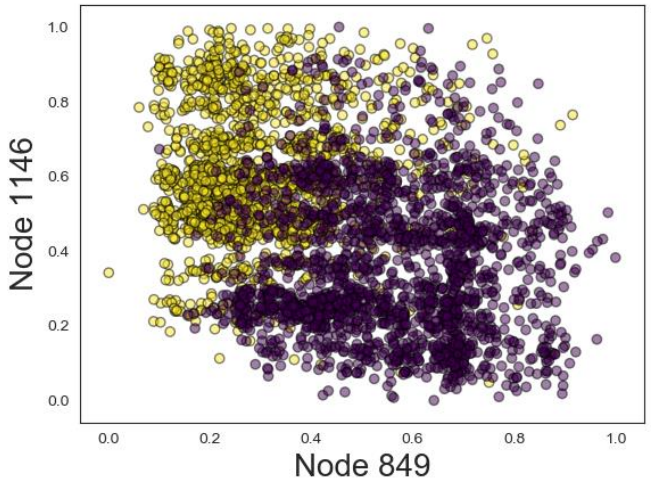


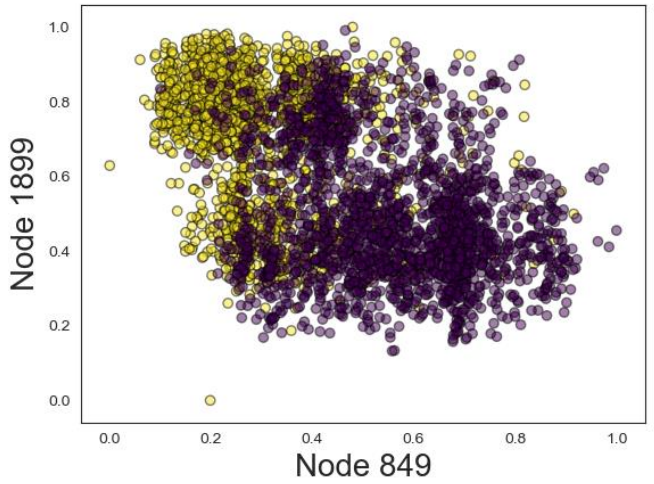


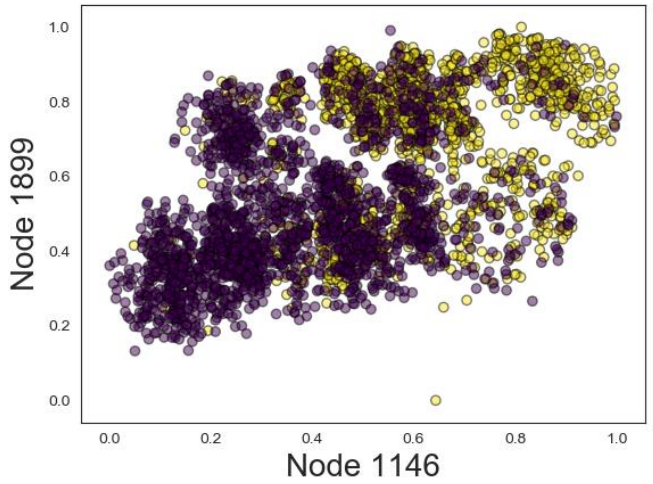


3 nodes that classify best

- 849, 1146, 1899 – determined by seeing which features(nodes) best classified dataset with SVM with high penalty on feature coefficients
  - Small coefficients become zero
  - If only three features are non-zero, then features 849, 1146, and 1899 are the best features



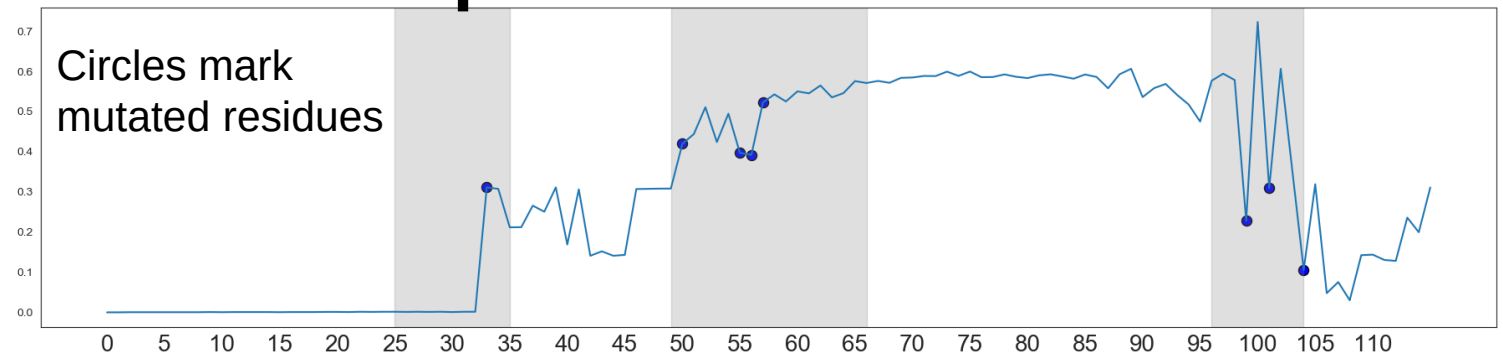




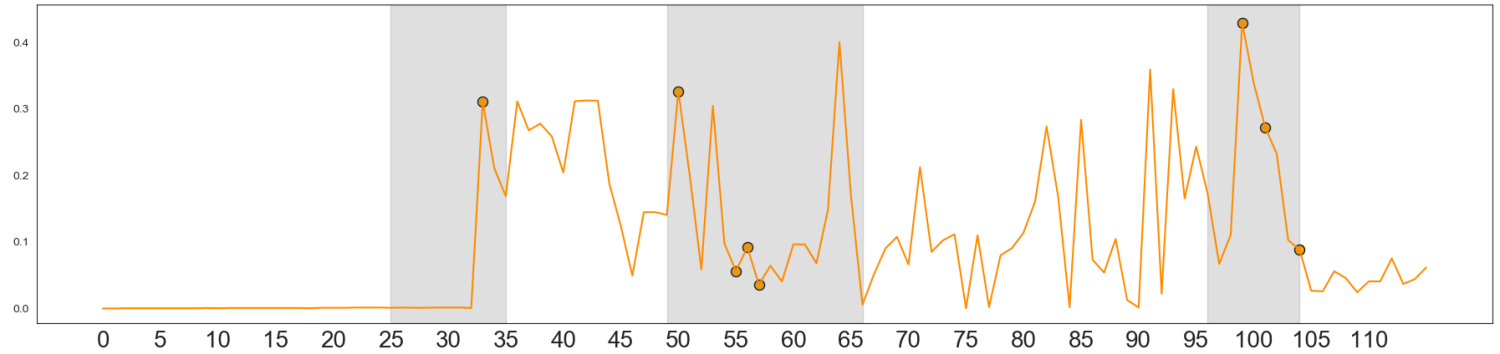
Quantifying difference between specific and nonspecific clone weight values

- Jansen Shannon entropy – symmetric divergence between distributions
  - Quantifies distance between distributions

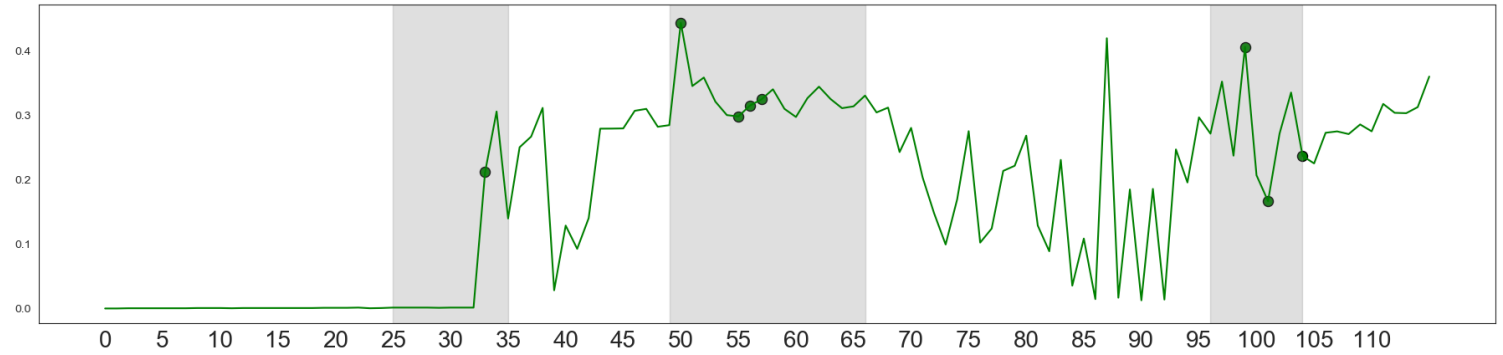
Difference between specific and nonspecific clones



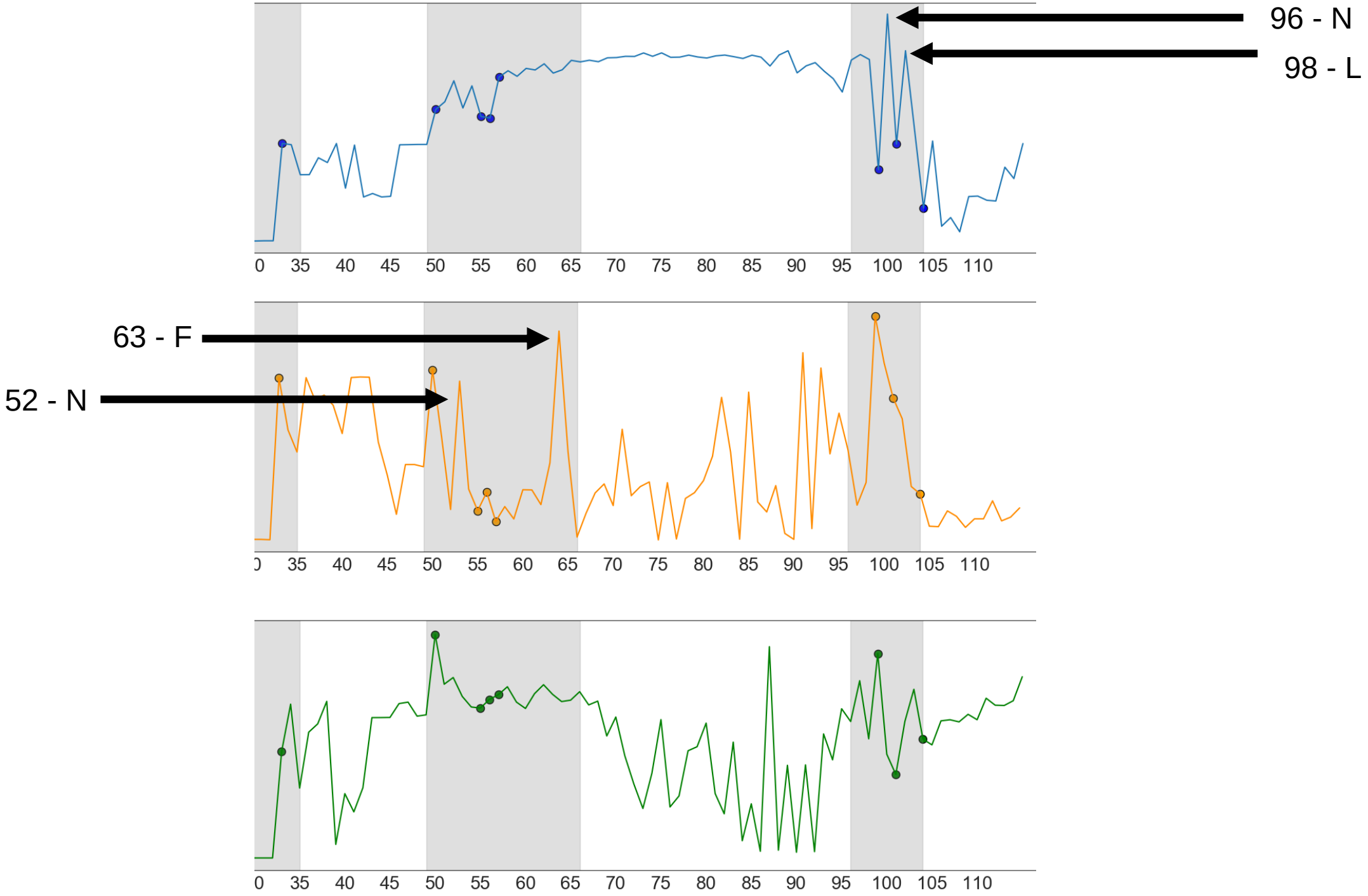
849



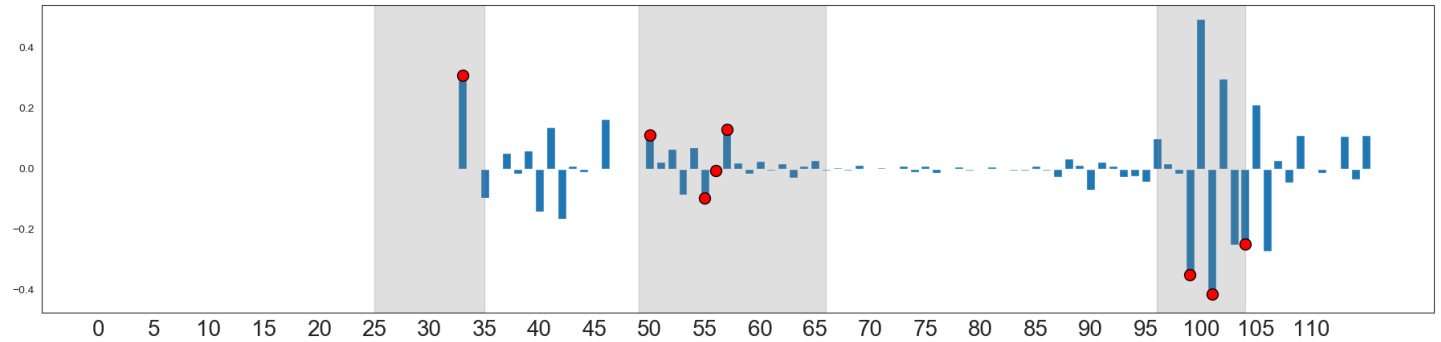
1146



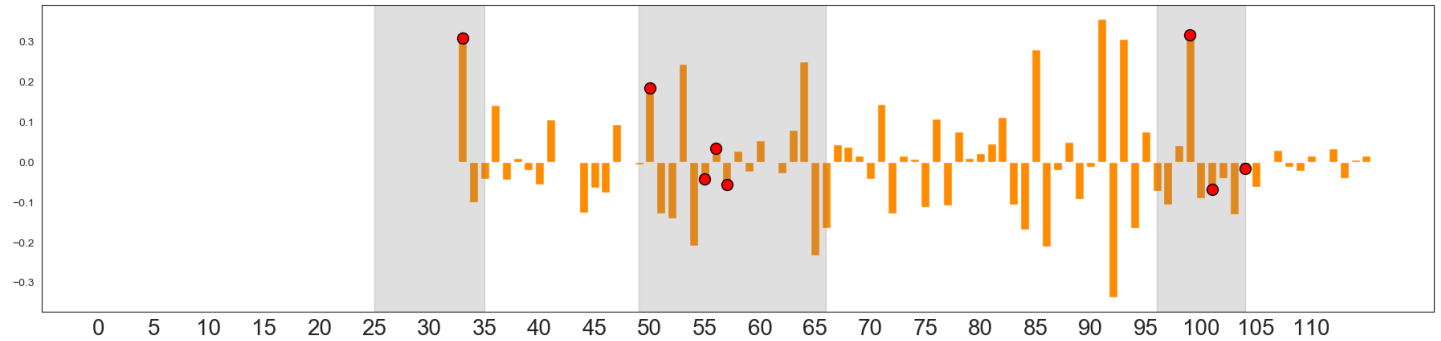
1899


96 - N
98 - L
63 - F
52 - N
0 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
0 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
0 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110

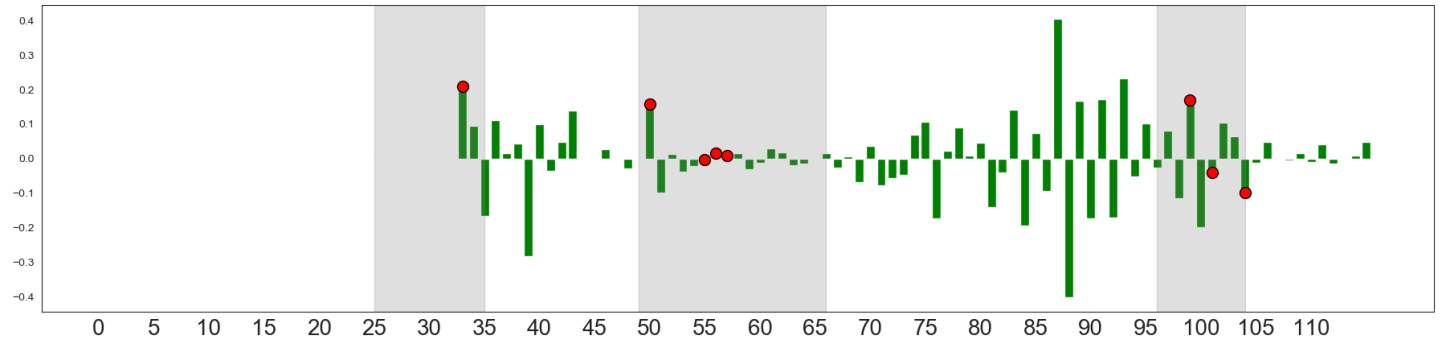
Change in the difference between specific and nonspecific clones



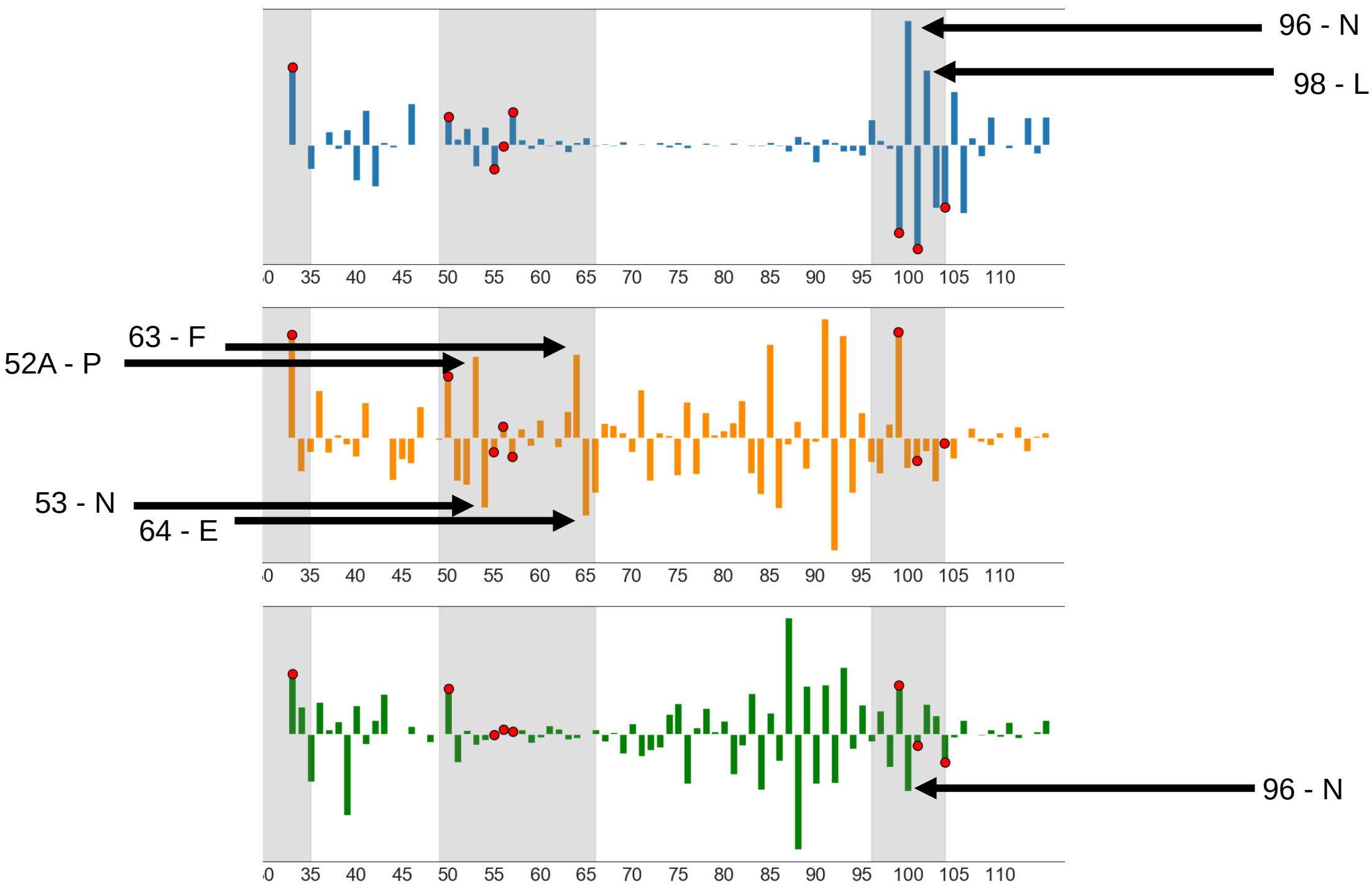
849



1146



1899


96 - N
98 - L
63 - F
52A - P
53 - N
64 - E
96 - N
0 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110

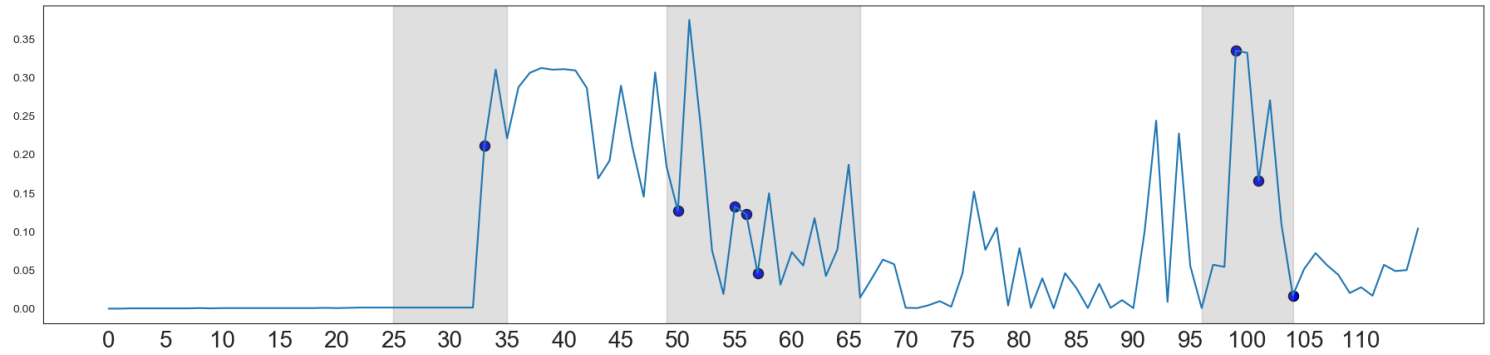
Residues suggested to be interesting for specificity

- Mutations suggested by the mutational scan (200 seqs on pareto frontier – optimized for both psy and ant)
  - 51, 52, 52A, 53, 54, 55, 56, 57, 58, 60, 62, 98

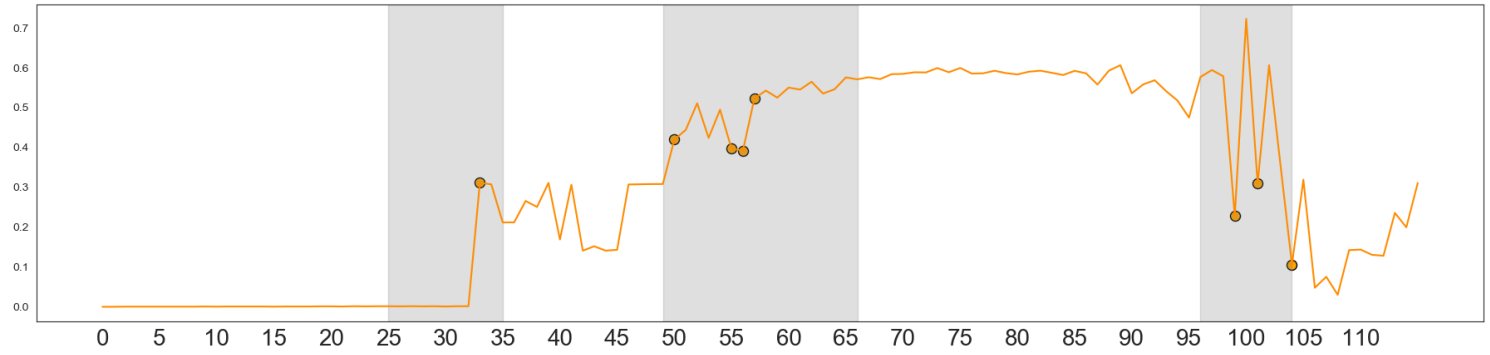
- Mutations suggested by the mutational scan (either high divergence between classes or high change in divergence between classes for specificity only)
  - 52, 52A, 53, 63, 64, 96, 98

Most frequently suggested
Commonly suggested

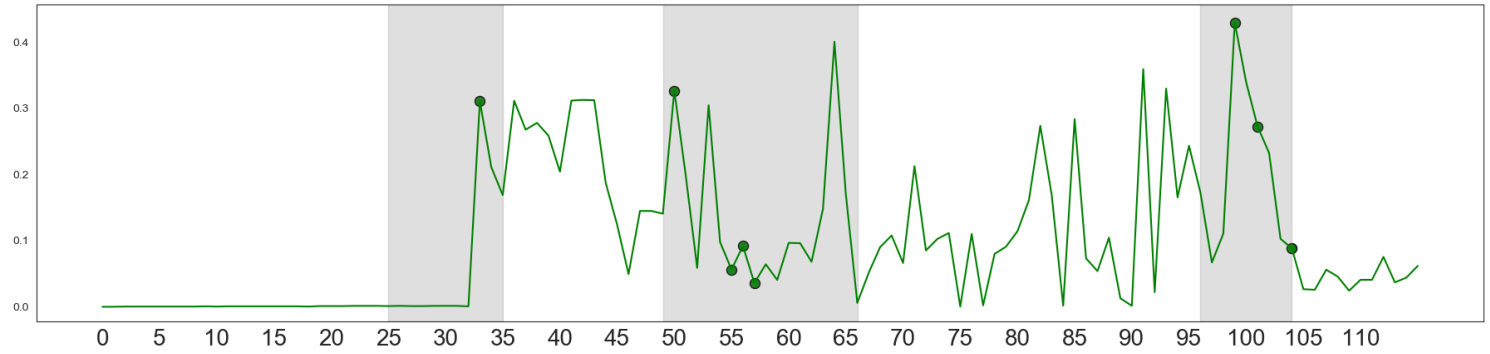
Change in the difference between antigen binding and non-antigen binding clones



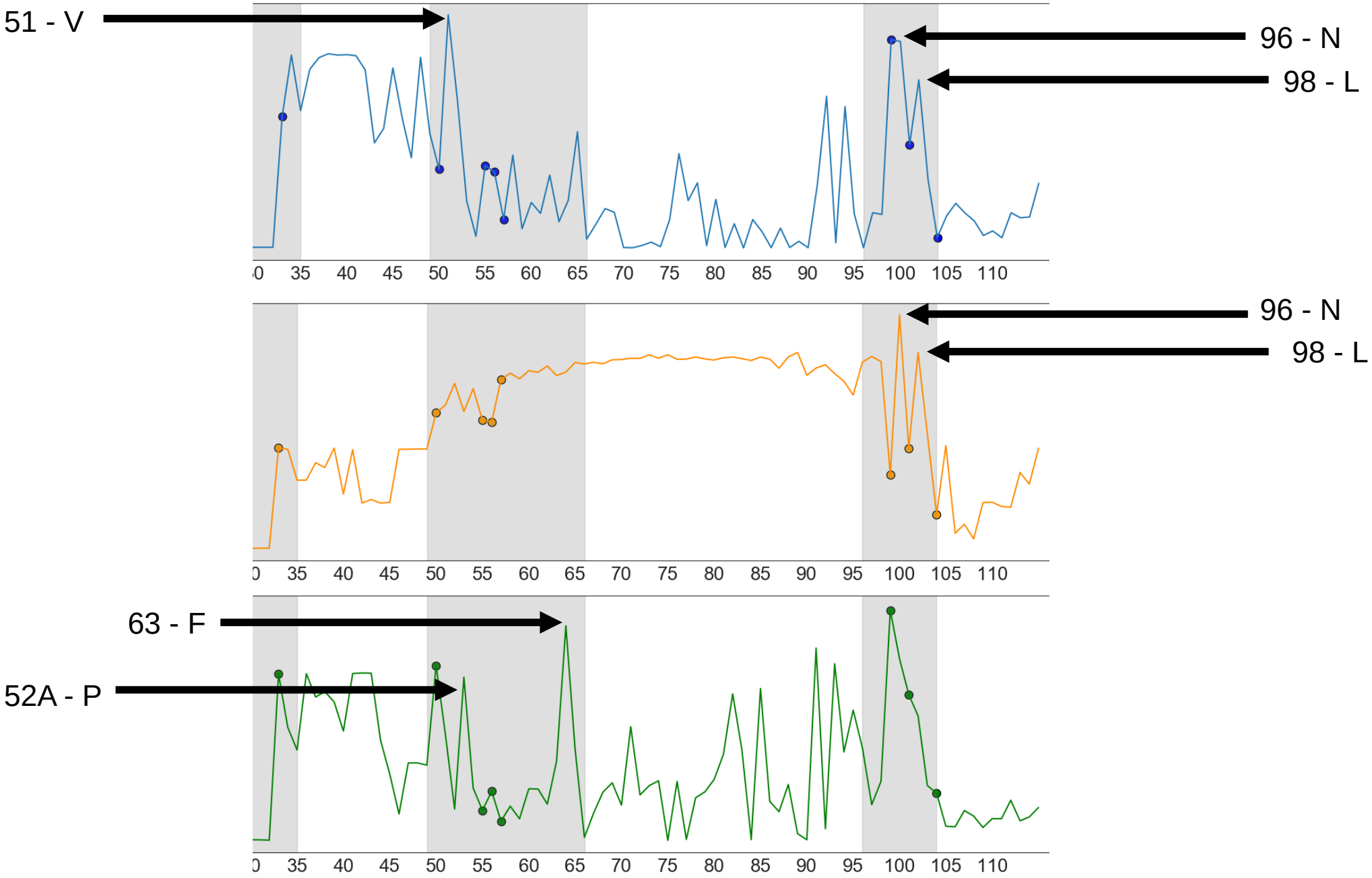
1830



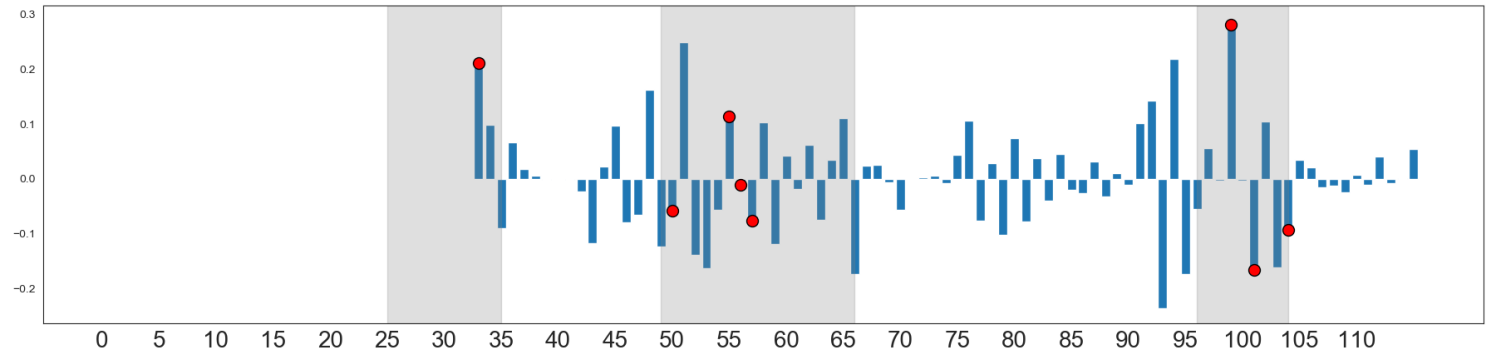
849



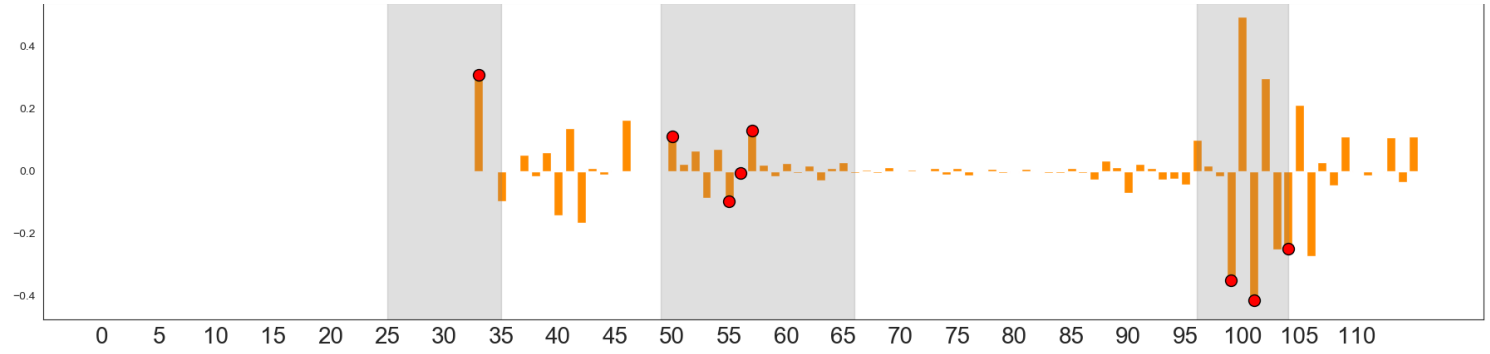
1146


51 - V
96 - N
98 - L
0 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
96 - N
98 - L
0 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
63 - F
52A - P
0 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110

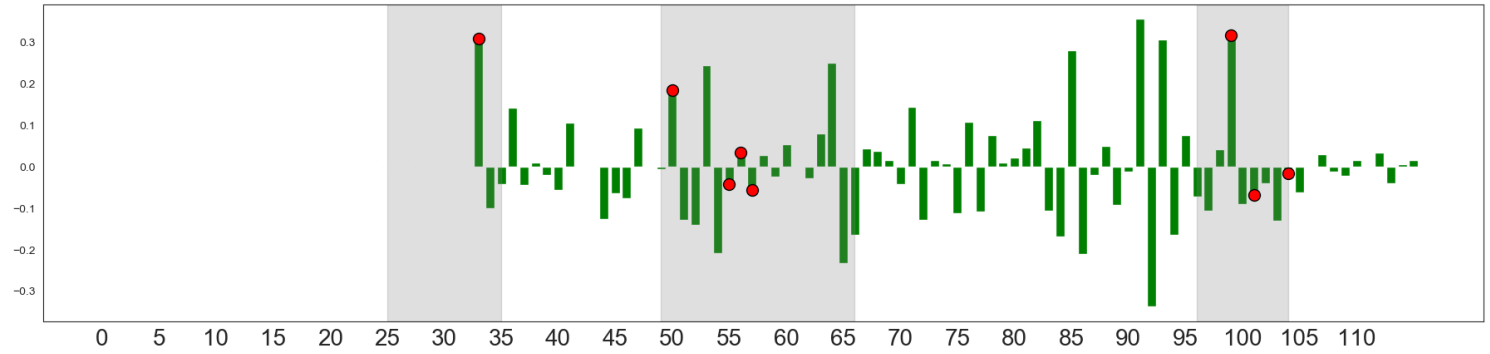
Change in the difference between antigen binding and non-antigen binding clones



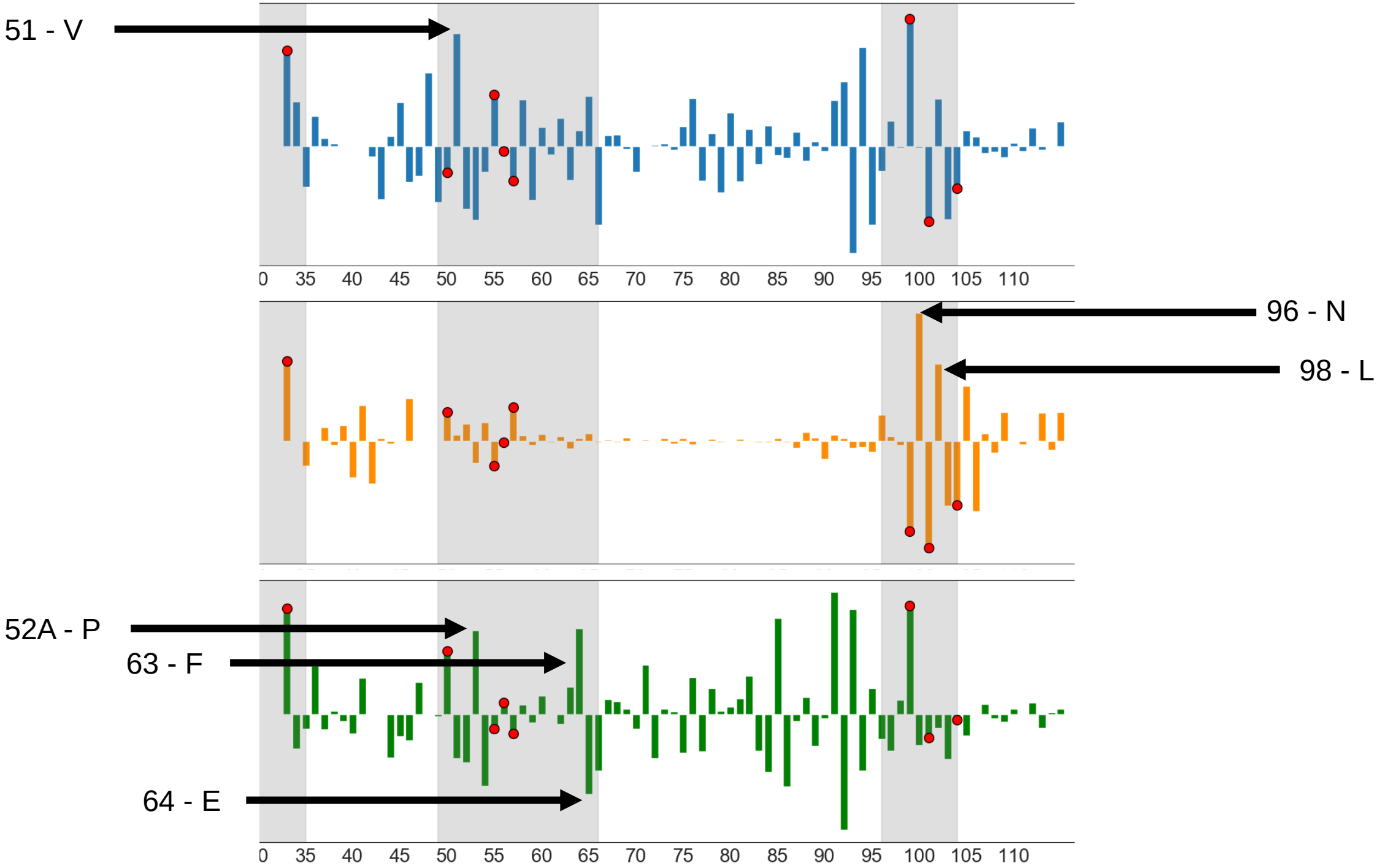
1830



849



1146


51 - V
96 - N
98 - L
52A - P
63 - F
64 - E
0 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
0 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110

Residues suggested to be interesting for antigen binding

- Mutations suggested by the mutational scan (200 seqs on pareto frontier – optimized for both psy and ant)
  - 51, 52, 52A, 53, 54, 55, 56, 57, 58, 60, 62, 98

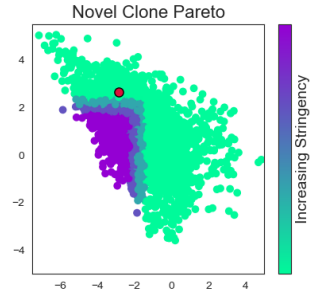
- Mutations suggested by the mutational scan (either high divergence between classes or high change in divergence between classes for specificity only)
  - 51, 52A, 63, 64, 96, 98

Most frequently suggested
Commonly suggested

How do they work together?

Analog data from isolated clones

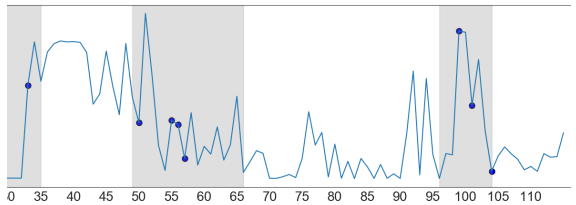
NGS data from screen/sort



Pareto Optimization

New suggested mutations/sites

Feature analysis



New suggested mutations/sites

Do the suggested mutations and sites match for each set of analysis?

Overlap of suggested mutations

- Overlapping suggested mutations
  - 51, 52, 53, 52A, 98
    - 51-V, 52-N, 53-N, 52A-P, 98-L

- Other strong candidates
  - 57, 63, 96
    - 57-T, 63-F, 96-N

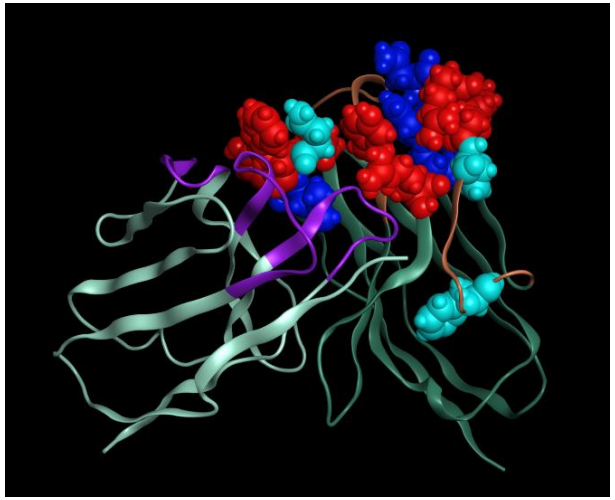
Suggested mutations

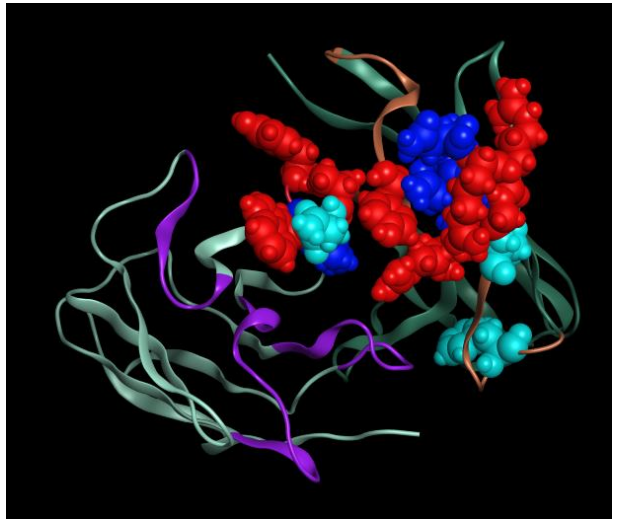
Red – mutated in previous library
Royal blue – overlapping mutations suggested by both analyses
Turquoise – mutation strongly suggested by one analysis, not both

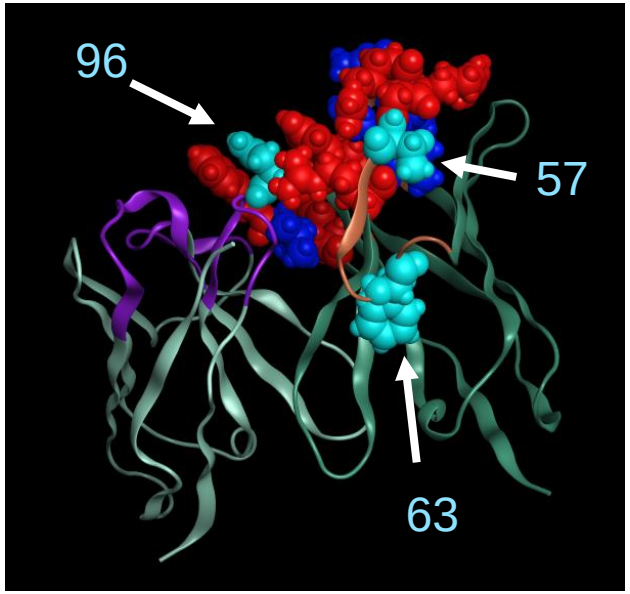
	26	27	28	29	30	31	32	33	34	35
CDR1	G	Y	T	F	T	D	Y	Y	M	H

	50	51	52	52A	53	54	55	56	57	58	59	60	61	62	63	64	65
CDR2	R	V	N	P	N	R	R	G	T	T	Y	N	Q	K	F	E	G

	93	94	95	96	97	98	99	100	101	102
CDR3	A	R	A	N	W	L	-	-	D	Y







Yulei's paper – residues important for specificity

	H1	H2				H3			Spearman's correlation coeff (ρ)		
	33	50	54	55	56	93	97	102			
	Y	R	R	R	G	A	W	Y	PSR	OVA	Avg
1	F		T		D			A	0.76	0.83	0.80
2	F		T		D			D	0.75	0.83	0.79
3	V		T	G	D				0.65	0.81	0.73
4					D	S	L	V	0.75	0.69	0.72
5	V				D		L	D	0.77	0.68	0.72
6		K			D		L	D	0.75	0.68	0.72
7	V		T		D			D	0.69	0.74	0.71
8	F			G	D		G		0.66	0.73	0.69
9				G	D		L	V	0.68	0.70	0.69
10	F			G	D		L		0.65	0.72	0.68

New Library Design

Base Sequences

	33	50	54	55	56	93	97	102
WT	Y	R	R	R	G	A	W	Y
	Y	R	R	G	D	A	G	D
	Y	R	R	G	D	A	A	V
	Y	R	T	A	G	A	L	Y
	Y	R	A	G	G	A	L	D
	Y	R	T	G	A	A	L	A
	Y	R	A	G	G	A	A	F

Never mutated - Important for antigen binding

Mutated in Yulei's findings for specificity by not here when considering antigen binding

Never mutated – important for specificity and antigen binding

Not mutated in Yulei's findings or here

Both could be removed from new library design

New Library Design

From old library

	54	55	56	97	102
WT	R	R	G	W	Y

New suggestions

	51	52	52A	53	57	63	96	98
WT	V	N	P	N	T	F	N	L
Freq	82% I		(27% P)		56% T	50% V		

Additional considerations

- Natural diversity
- Too many mutations too close
- Yulei's past analysis to make the old library
- Properties of residues being mutated
- Properties of residues sampled
- …etc

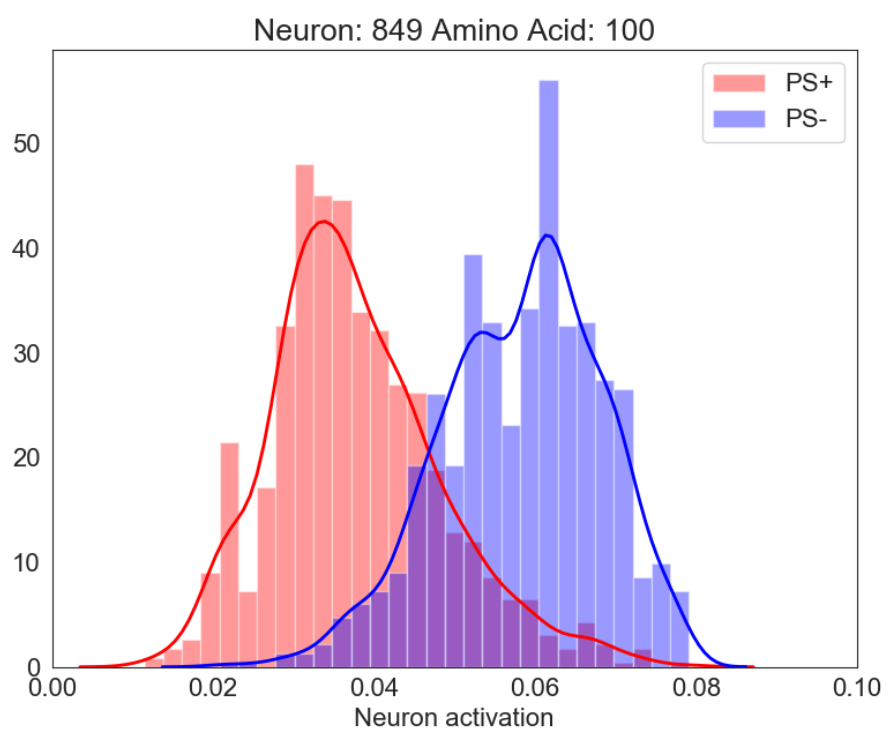
Next Steps

- Finalize decision of new clones to produce and evaluate
  - Order gene blocks and clone
- Library design
  - Work with Yulei and Matt
  - Heuristic decisions about what to include

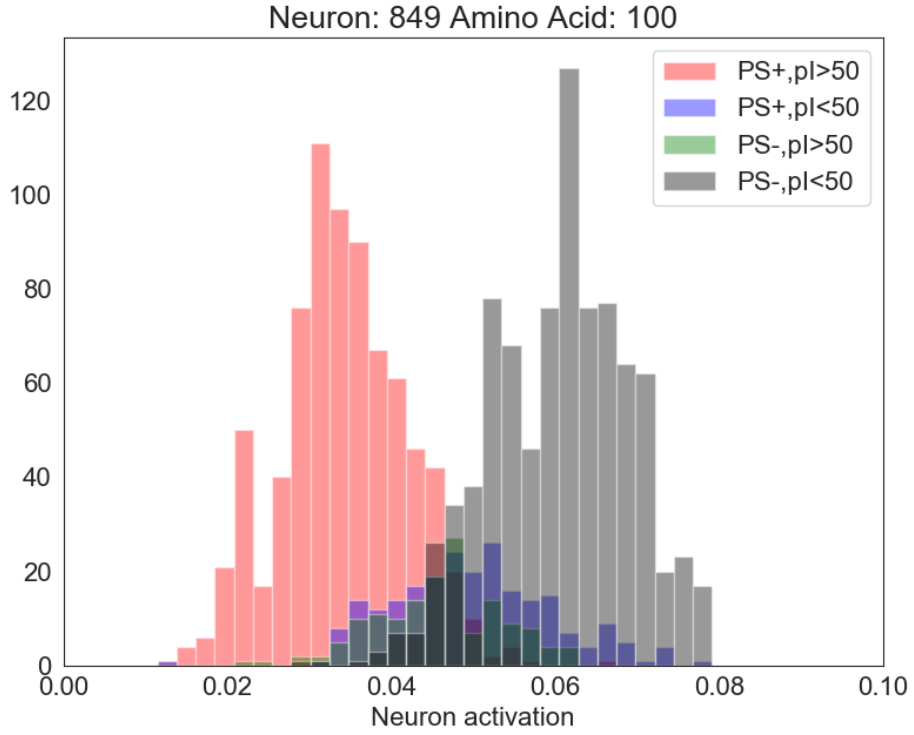
Bonus info! 

Segregating features by biophysical property

Distribution of sequences that emerges from unirep features analysis. Each sequence also has a set of biophysical descriptors



If we know both the neuron activation for a new sequence, **and the sequence pI**, we can be even more confident in our prediction of PS+ vs PS-.



Process flow chart – Feature analysis