Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

David Gifford
Lecture 14

April 4, 2019

Machine Learning Designed Therapeutics

http://mit6874.github.io

What's on tap today!

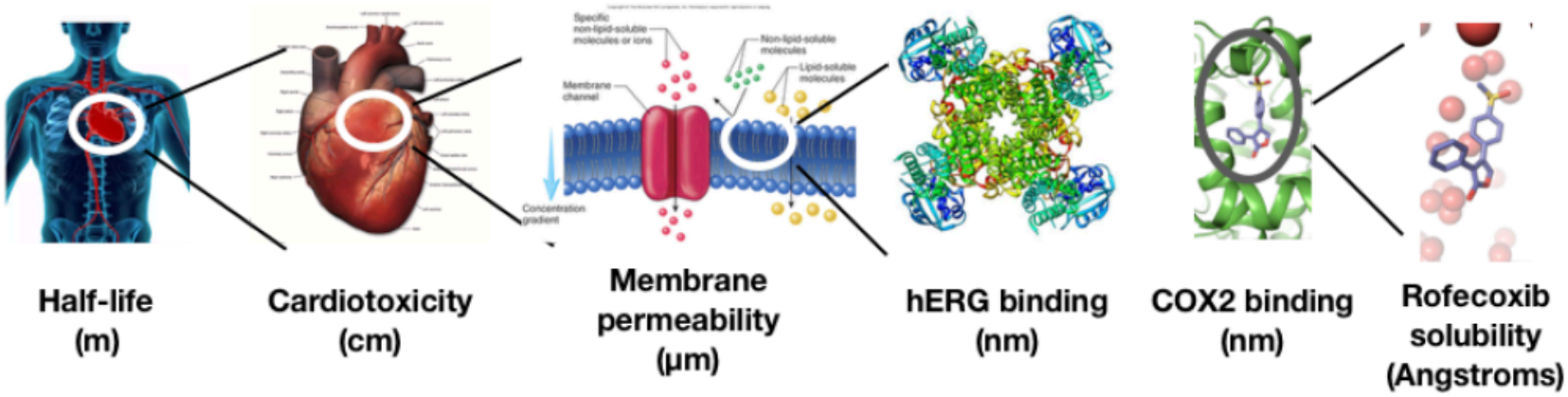
- Characterizing small molecule therapeutics
- Formulating peptide vaccines
- Designing antibodies
- Gain-of-function repairs with a DNA cut

What you should know

- Graph Convolutional Networks
- Language modes (ElMo)
- How to design novel antibody CDR sequences
- How to predict CRISPR outcomes
- Issues in the representation of small molecules

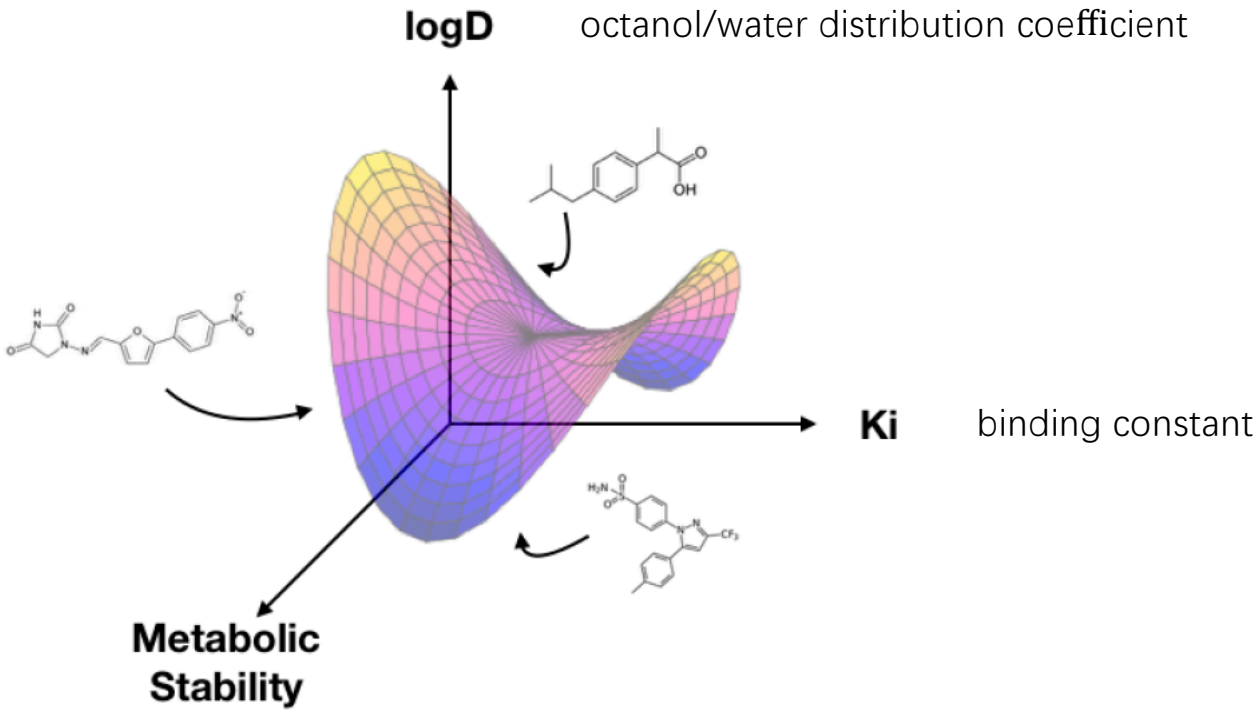
Graph Convolutional Neural Networks (GCNNs)

There are diverse prediction tasks for small molecules



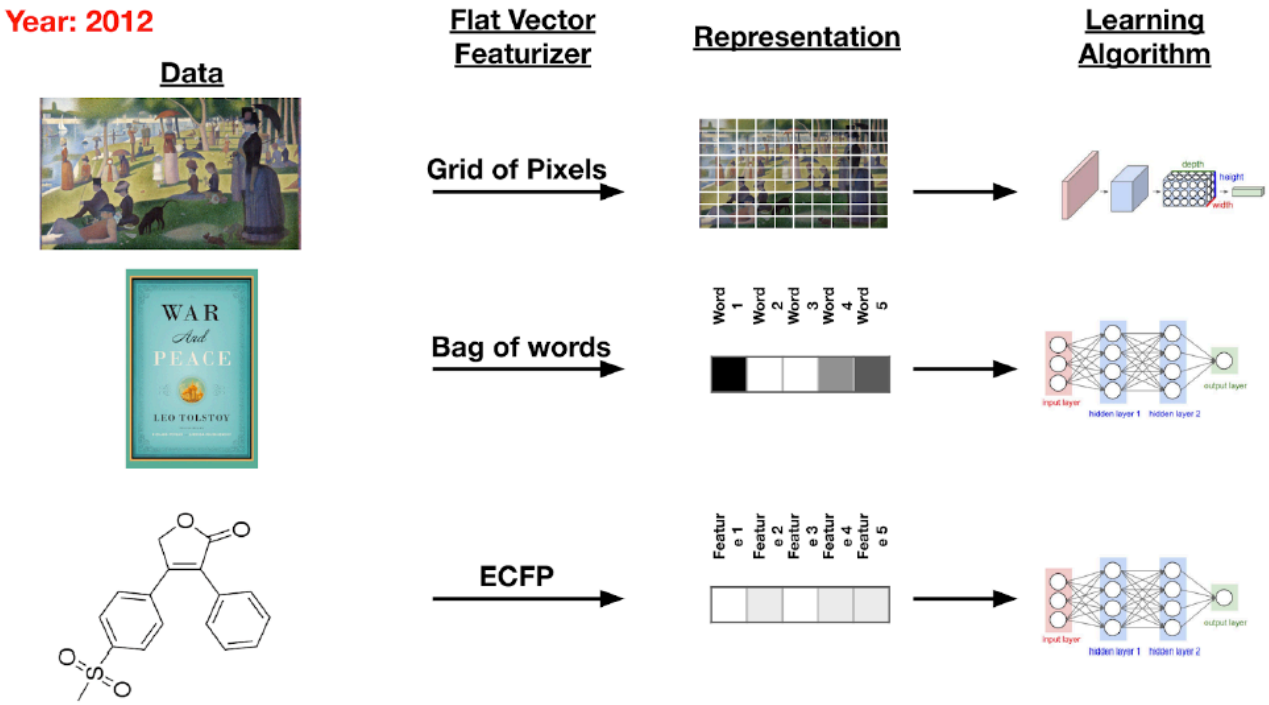
https://medium.com/@pandelab/step-change-improvement-in-molecular-property-prediction-with-potentialnet-f431ffa32a2c

How can we find molecules on the thin manifold of acceptability?



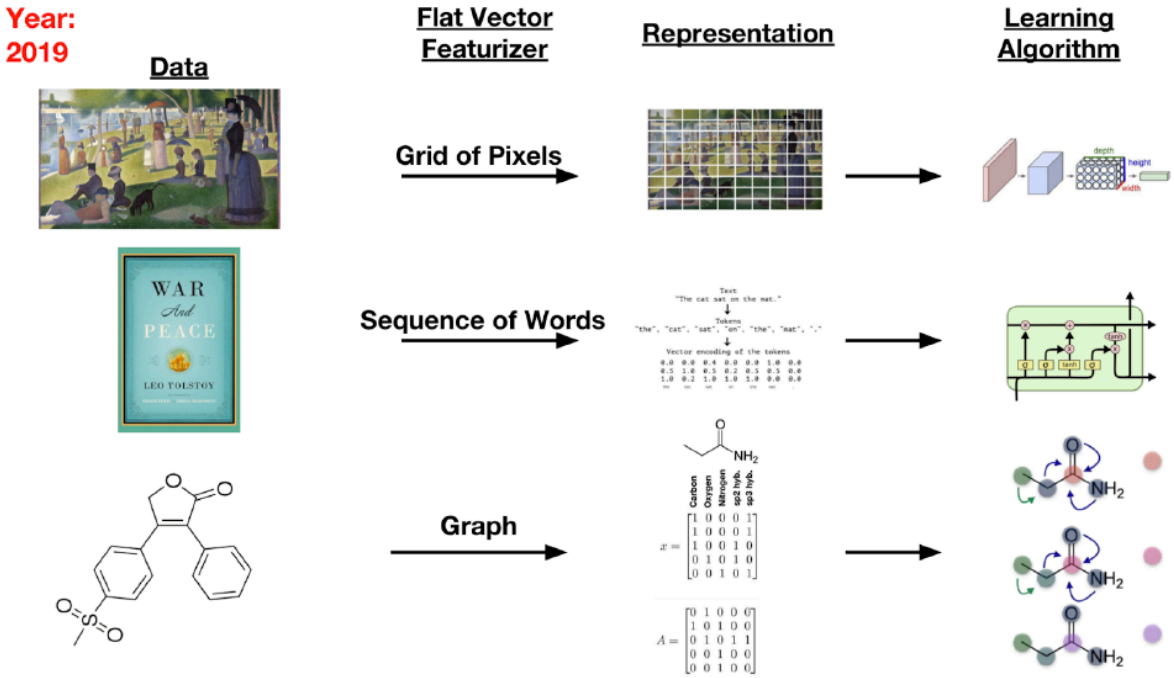
https://medium.com/@pandelab/step-change-improvement-in-molecular-property-prediction-with-potentialnet-f431ffa32a2c

How can we create features for small molecules to predict key properites?



https://medium.com/@pandelab/step-change-improvement-in-molecular-property-prediction-with-potentialnet-f431ffa32a2c

We can use graph convolutions on small molecules by representing their node features and adjacencies

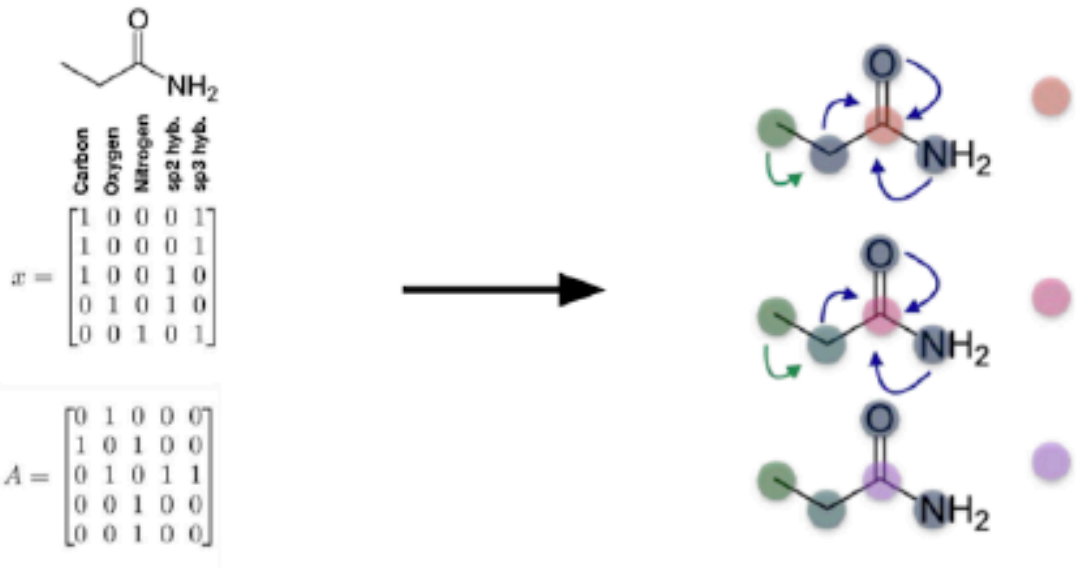


https://medium.com/@pandelab/step-change-improvement-in-molecular-property-prediction-with-potentialnet-f431ffa32a2c

We can use graph convolutions on small molecules by representing their node features and adjacencies

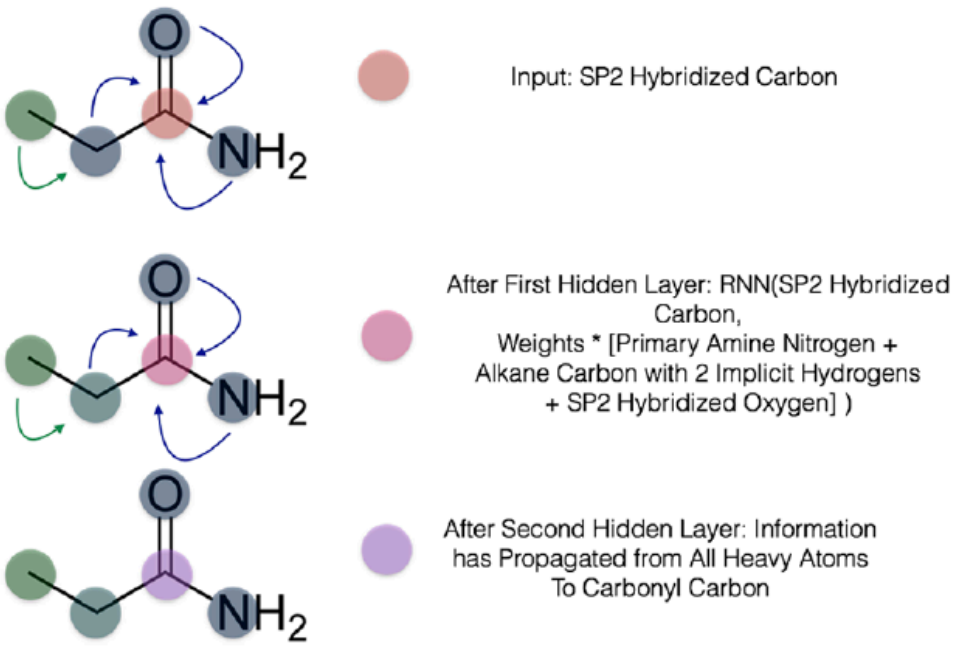
X is per-atom features of each graph node as rows

A is the per-atom adjacency matrix as rows



https://medium.com/@pandelab/step-change-improvement-in-molecular-property-prediction-with-potentialnet-f431ffa32a2c

Information propagates along adjacencies one step at a time



https://medium.com/@pandelab/step-change-improvement-in-molecular-property-prediction-with-potentialnet-f431ffa32a2c

X is per-atom features of each graph node as rows

A is the per-atom adjacency matrix as rows

K is the number of graph convolutional layers

x is flat features of molecule

Multi-Layer Perceptron / Fully Connected Neural Network	Graph Convolutional Neural Network (GCNN)
	$H^{(1)} = ReLU\left(W^{(1)} \cdot A \cdot X\right)$
	$H^{(2)} = ReLU\left(W^{(2)} \cdot A \cdot H^{(1)}\right)$
	$\vdots$
	$H^{(K)} = ReLU\left(W^{(K)} \cdot A \cdot H^{(K-1)}\right)$
	$x^{(NN)} = \sum_{atoms} H^{(K)}$
$h^{(1)} = ReLU\left(W^{(1)} \cdot x\right)$	$h^{(1)} = ReLU\left(W^{(1)} \cdot x^{(NN)}\right)$
$h^{(2)} = ReLU\left(W^{(2)} \cdot h^{(1)}\right)$	$h^{(2)} = ReLU\left(W^{(2)} \cdot h^{(1)}\right)$
$\vdots$	$\vdots$
$h^{(K)} = ReLU\left(W^{(K)} \cdot h^{(K-1)}\right)$	$h^{(K)} = ReLU\left(W^{(K)} \cdot h^{(K-1)}\right)$

X is per-atom features of each graph node

A is the atom adjacency matrix

H is the feature map for all atoms

1 bond length away

2 bond lengths away

K bond lengths away

Sum over per atom features
$x^{(NN)}$ is convolutional "fingerprint" of entire molecule

K fully connected layers

Graph Convolutional Neural Network (GCNN)

$$H^{(1)} = ReLU\left(W^{(1)} \cdot A \cdot X\right)$$

$$H^{(2)} = ReLU\left(W^{(2)} \cdot A \cdot H^{(1)}\right)$$

$$\vdots$$

$$H^{(K)} = ReLU\left(W^{(K)} \cdot A \cdot H^{(K-1)}\right)$$

$$x^{(NN)} = \sum_{atoms} H^{(K)}$$

$$h^{(1)} = ReLU\left(W^{(1)} \cdot x^{(NN)}\right)$$

$$h^{(2)} = ReLU\left(W^{(2)} \cdot h^{(1)}\right)$$

$$\vdots$$

$$h^{(K)} = ReLU\left(W^{(K)} \cdot h^{(K-1)}\right)$$

PotentialNet

h_i is the feature map for atom i

1 bond length away; edge type specific processing;
Gated Recurrent Unit enables selecting adding

$$h_i^{(1)} = GRU\left(x_i, \sum_e^{N_{\text{et}}} \sum_{j \in N^{(e)}(v_i)} NN^{(e)}(x_j)\right)$$

$$\vdots$$

K bond length away; edge type specific processing

$$h_i^{(K)} = GRU\left(h_i^{(b_{K-1})}, \sum_e^{N_{\text{et}}} \sum_{j \in N^{(e)}(v_i)} NN^{(e)}(h_j^{(b_{K-1})})\right)$$

$h^{(NN)}$ is convolutional "fingerprint" of entire molecule

$$h^{(NN)} = \sigma\left(i(h^{(K)}, x)\right) \odot \left(j(h^{(K)})\right)$$

$$h^{(FC_0)} = \sum_{j=1}^{N_{Lig}} h_j^{(NN)}$$

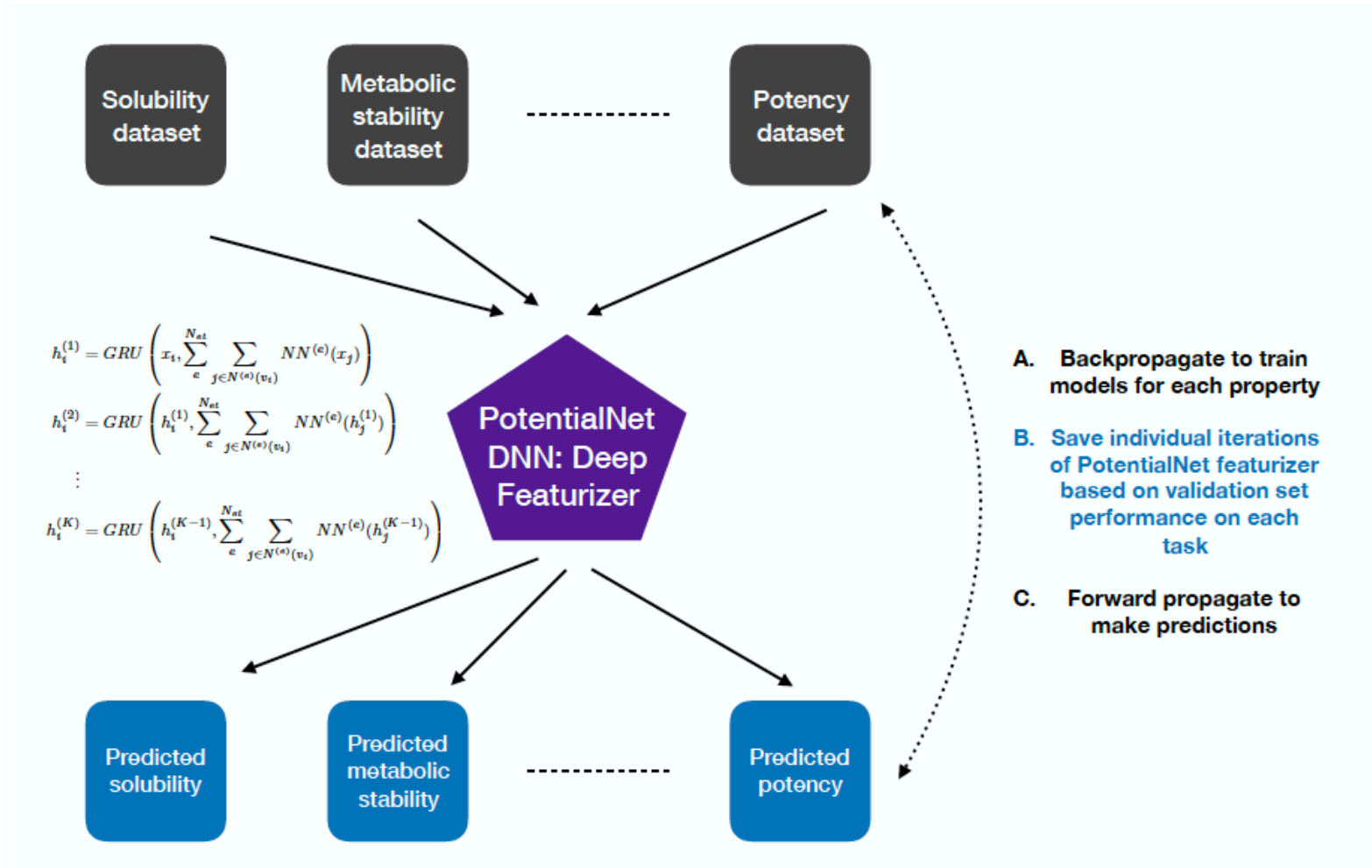
Fully connected layers

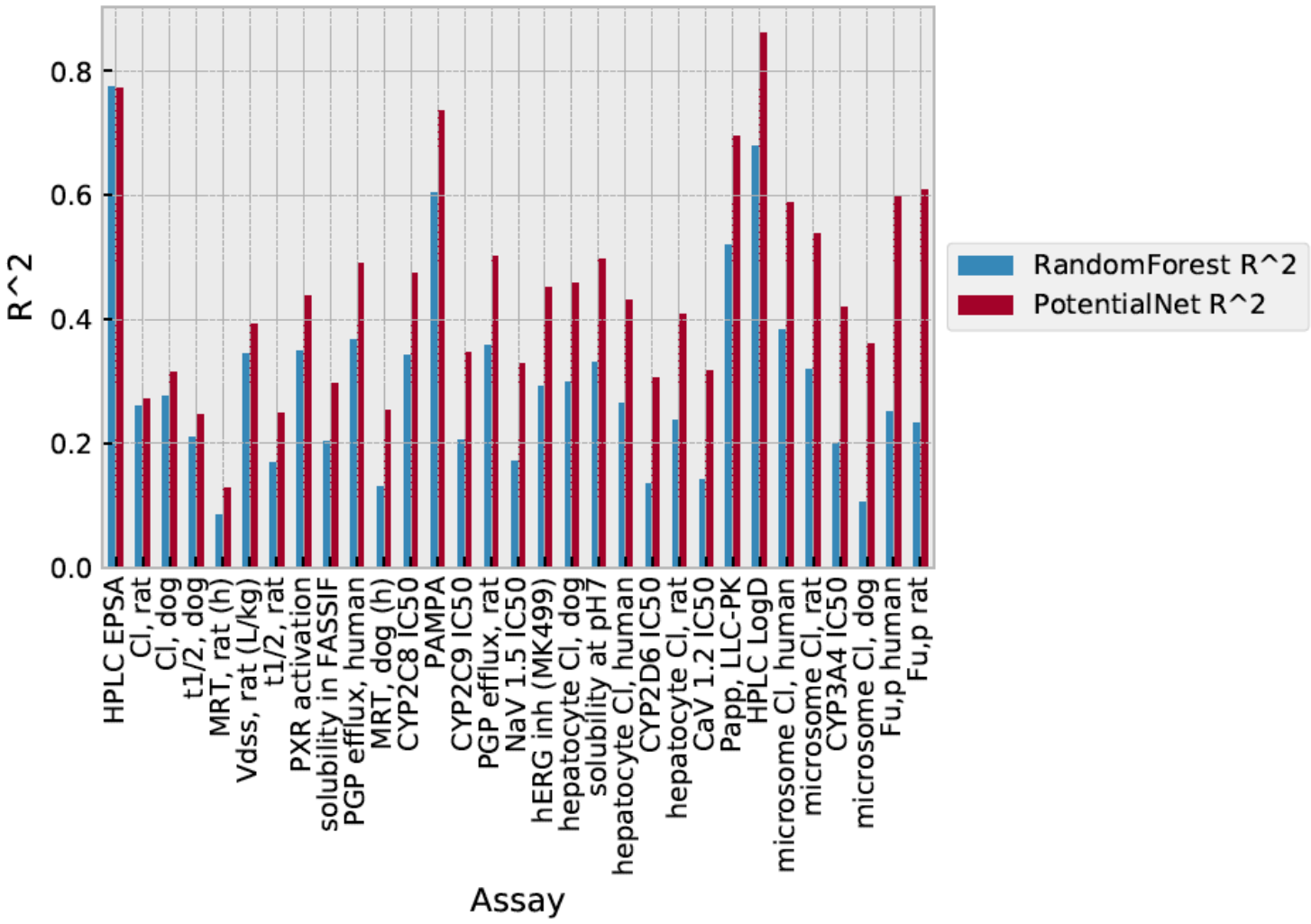
$$h^{(FC_1)} = ReLU\left(W^{(FC_1)} h^{(FC_0)}\right)$$

$$\vdots$$

$$h^{(FC_K)} = W^{(FC_K)} h^{(FC_{K-1})}$$

Training and Prediction

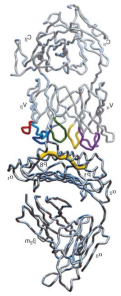



R^2
0.0
0.2
0.4
0.6
0.8
RandomForest R^2
PotentialNet R^2
HPLC EPSA
Cl, rat
Cl, dog
t1/2, dog
MRT, rat (h)
Vdss, rat (L/kg)
t1/2, rat
PXR activation
solubility in FASSIF
PGP efflux, human
MRT, dog (h)
CYP2C8 IC50
PAMPA
CYP2C9 IC50
PGP efflux, rat
NaV 1.5 IC50
hERG inh (MK499)
hepatocyte Cl, dog
solubility at pH7
hepatocyte Cl, human
CYP2D6 IC50
hepatocyte Cl, rat
CaV 1.2 IC50
Papp, LLC-PK
HPLC LogD
microsome Cl, human
microsome Cl, rat
CYP3A4 IC50
microsome Cl, dog
Fu,p human
Fu,p rat
Assay

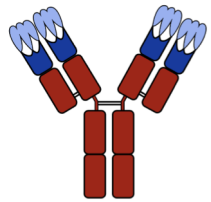
Computational immunotherapeutics

We wish to both identify therapeutic targets and therapeutic molecules with the help of machine learning

- Identifying target peptide-MHC molecules

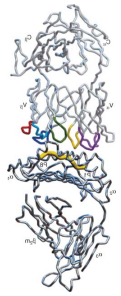


- Designing antibody Complementarity Determining Regions (CDRs)

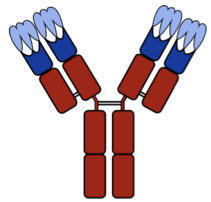


We wish to both identify therapeutic targets and therapeutic molecules with the help of machine learning

- Identifying target peptide-MHC molecules

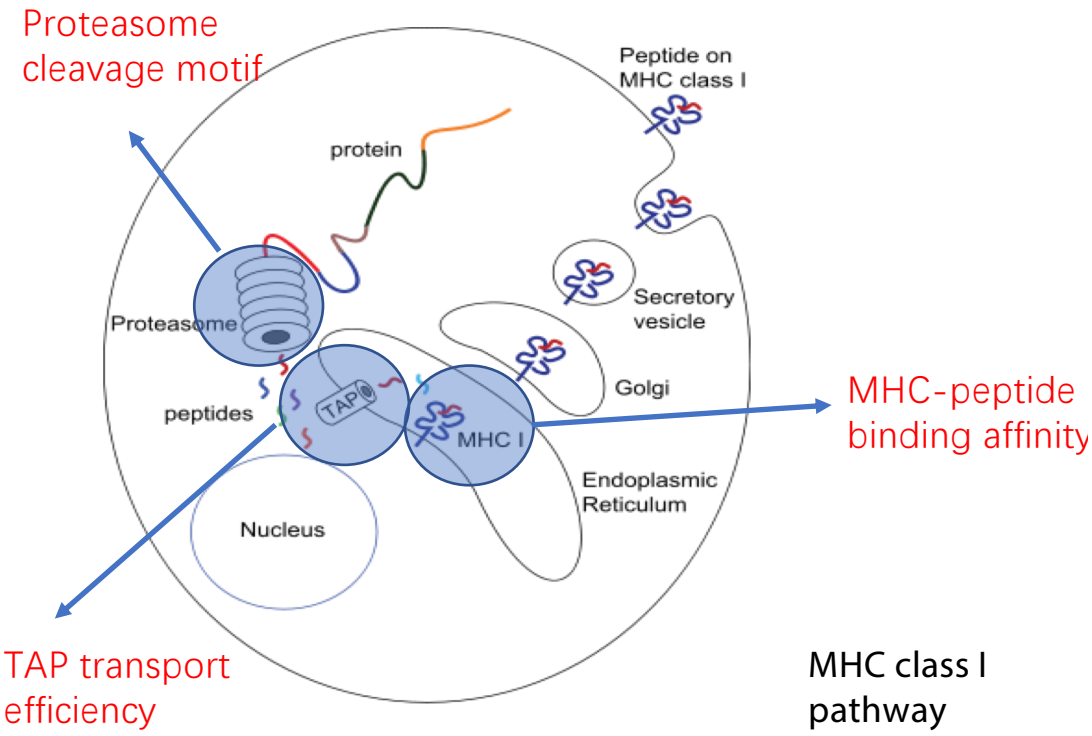


- Designing antibody Complementarity Determining Regions (CDRs)

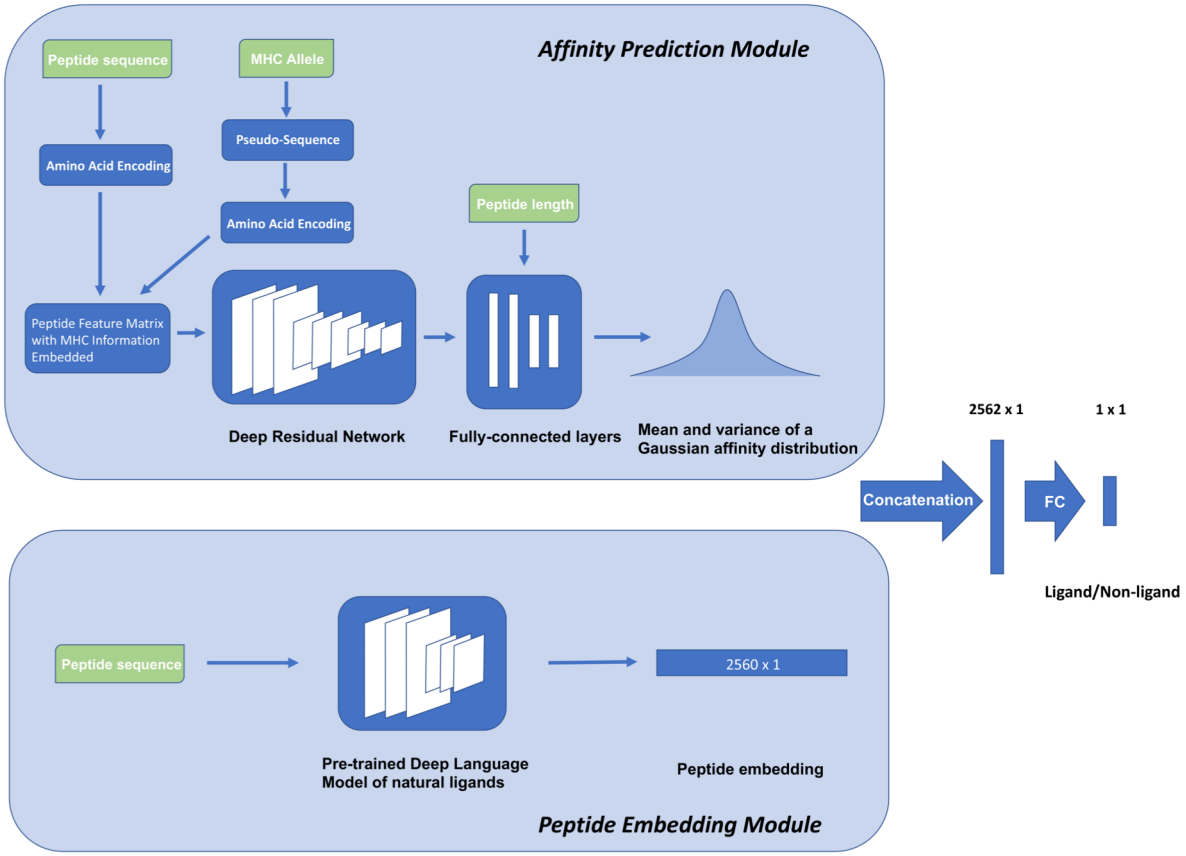


Most existing methods for peptide display focus on modeling MHC binding affinity

- Immune Epitope Database (IEDB) contains a large collection of binding affinity datasets curated from literature
- However, models trained on affinity data are not able to consider other factors in MHC ligand selection



DeepLigand predicts MHC class I peptide presentation (552,252 positive examples, 2.5M negative, 192 MHC alleles)



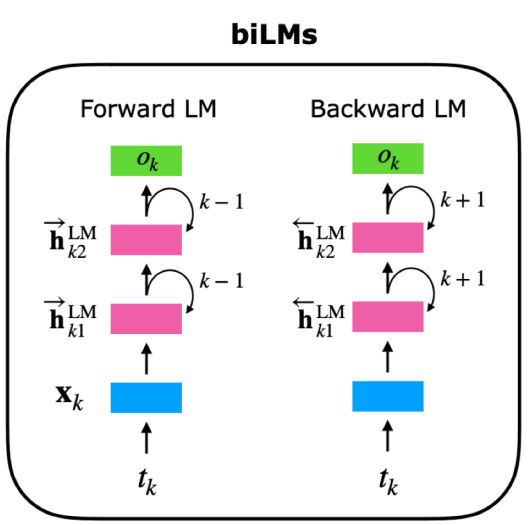
ELMo for learning contextualized word embedding

ELMo represents a word t_k as a linear combination of corresponding hidden layers (inc. its embedding)

$$\mathbf{ELMo}_k^{\text{task}} = \gamma^{\text{task}} \times \sum \left\{ s_2^{\text{task}} \times \mathbf{h}_{k2}^{\text{LM}},\ s_1^{\text{task}} \times \mathbf{h}_{k1}^{\text{LM}},\ s_0^{\text{task}} \times \mathbf{h}_{k0}^{\text{LM}} \right.$$

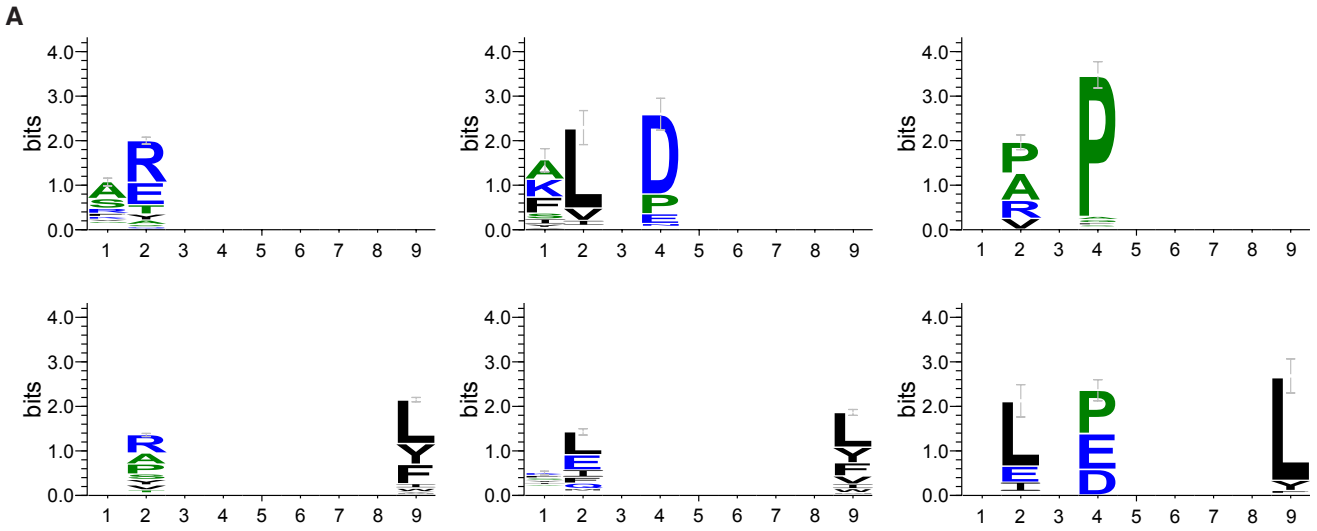
$$\left. ([\mathbf{x}_k ; \mathbf{x}_k]),\ [\overrightarrow{\mathbf{h}}_{kj}^{\text{LM}} ; \overleftarrow{\mathbf{h}}_{kj}^{\text{LM}}] \right\}$$

Concatenate hidden layers

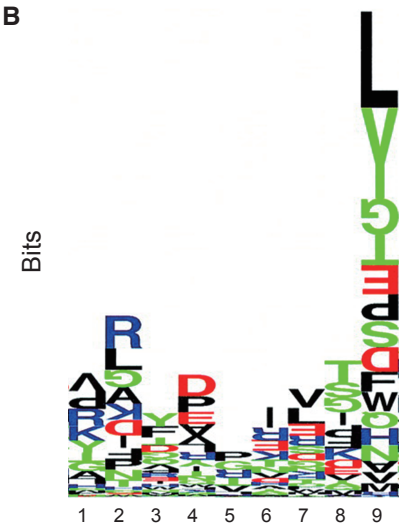


https://arxiv.org/abs/1802.05365

Class I learned language model is consistent with the known proteasome cleavage motif

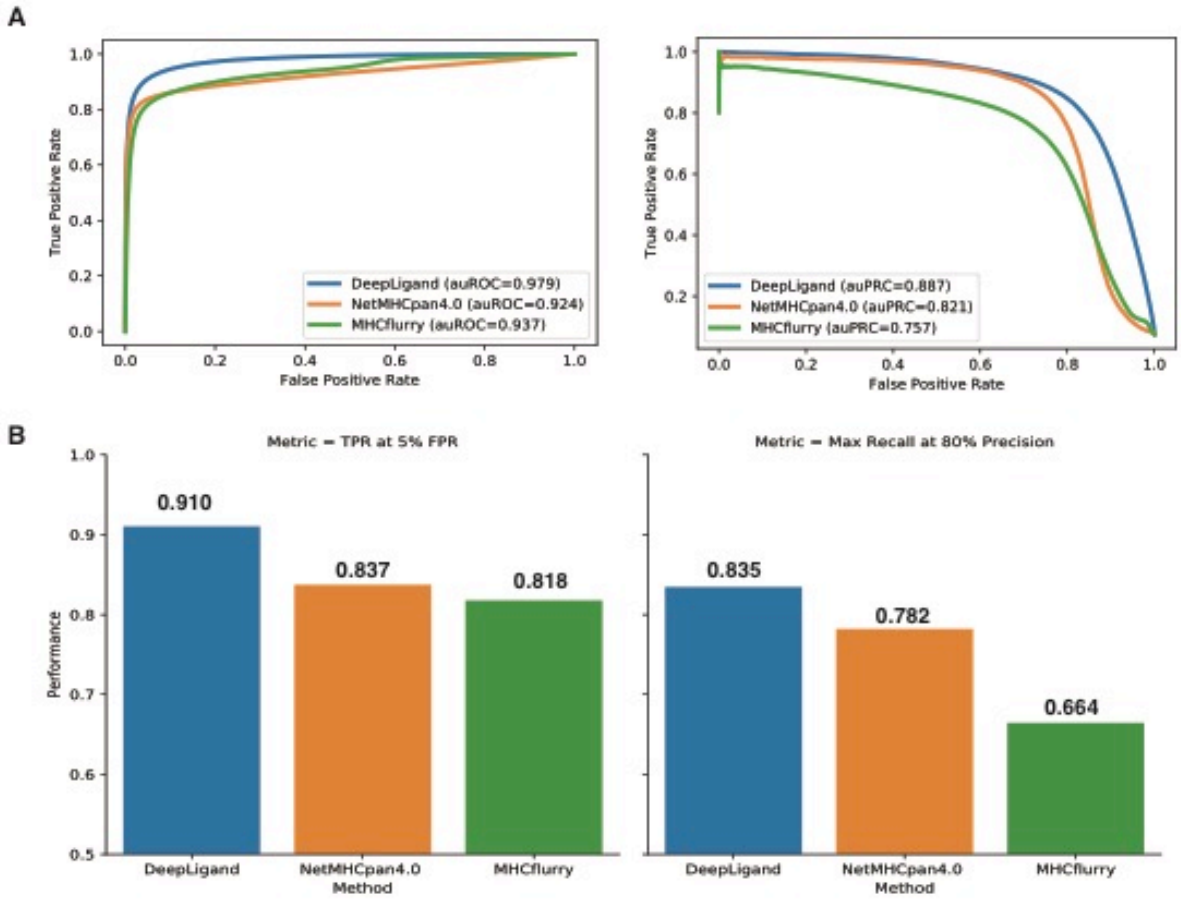


Learned Language Model

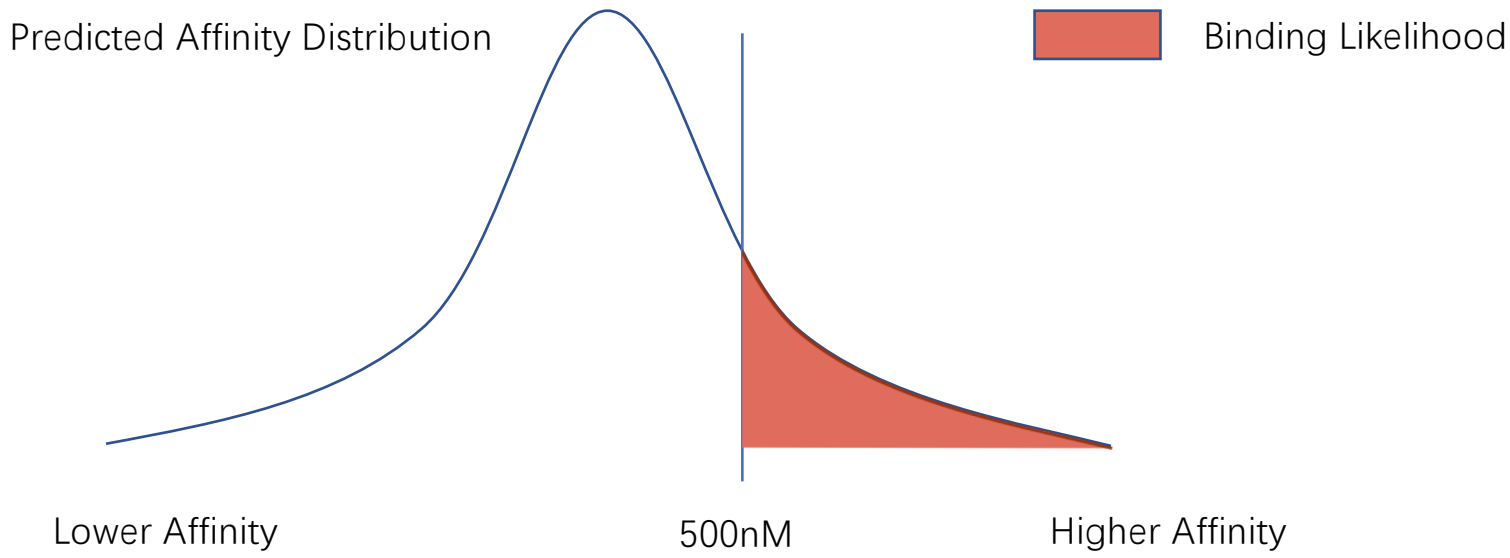


Proteasome motif

DeepLigand outperforms existing methods (Class I)

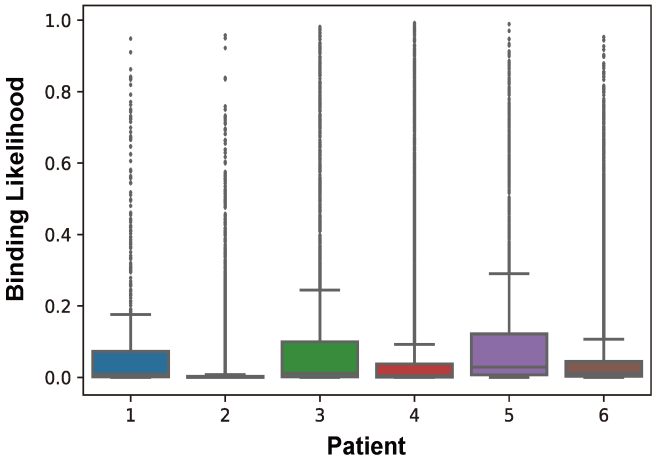


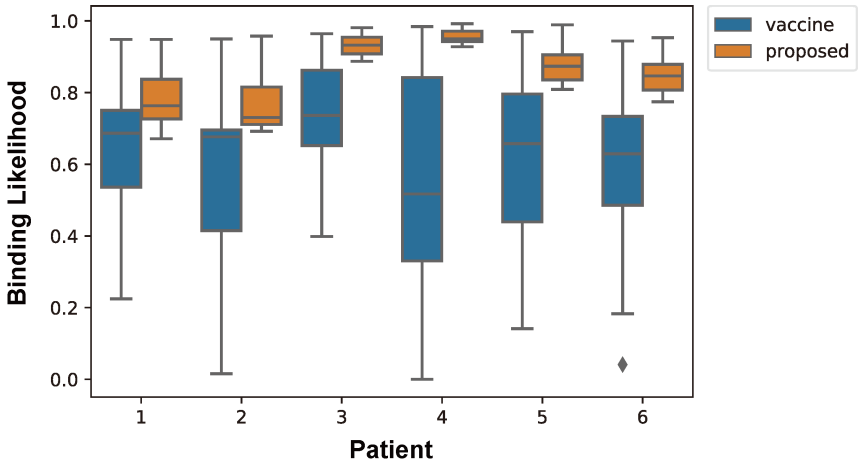
MHC class I uncertainty metrics enable binding likelihood for peptide-MHC molecules



Published MHC class I peptide vaccine formulations can be improved by uncertainty metrics

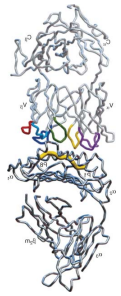
- We examined the binding likelihood of neo-antigen peptides designed by Otta et al (2017) as well as all the mutation-spanning peptides in each patient



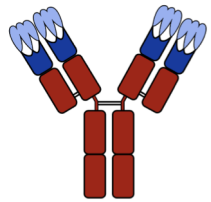


We wish to both identify therapeutic targets and therapeutic molecules with the help of machine learning

- Identifying target peptide-MHC molecules

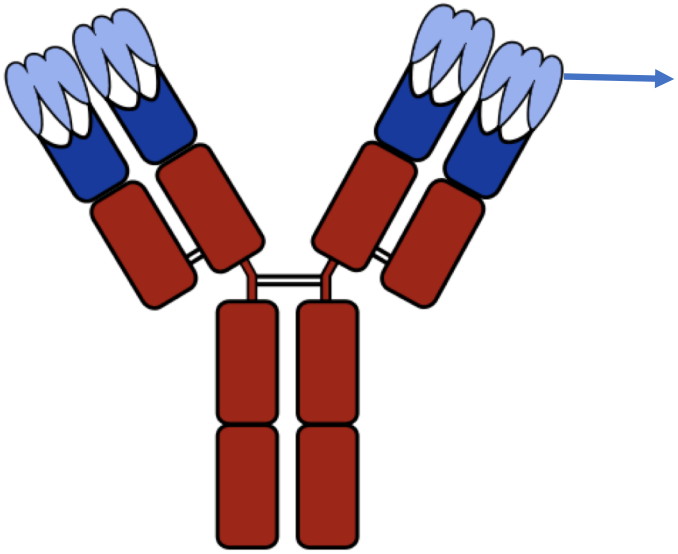


- Designing antibody Complementarity Determining Regions (CDRs)

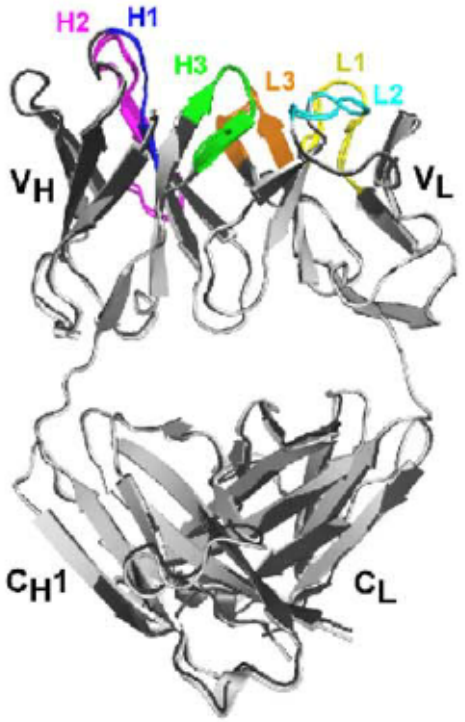


How can we design CDRs that meet multiple objectives?

Complementarity-determining regions (CDRs) largely determine target affinity



Six CDRs in total for each Fab



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101210/

Design flow

Naïve library 10^{10} CDRH3 sequences

↓

Multiplexed Affinity Assay

↓ Training Data

ML Design Methods

↓

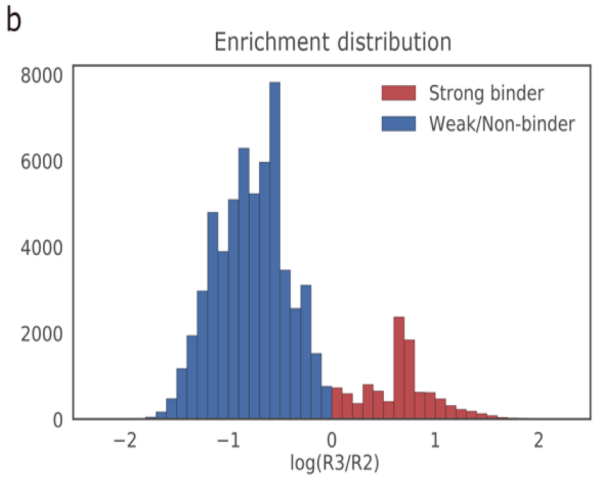
Naïve library → ML library fabrication

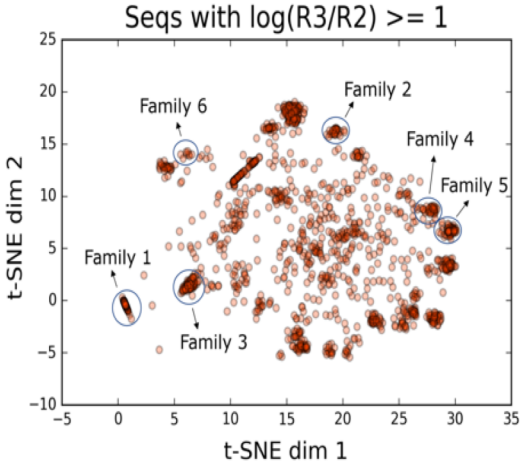
↓

Multiplexed Affinity Assay

↓

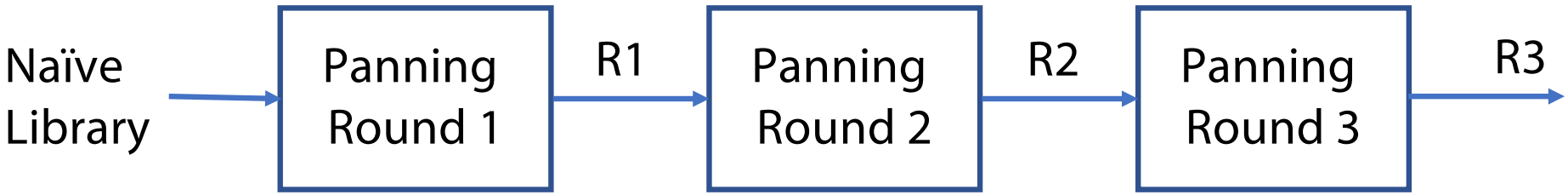
Full length antibody characterization





Training Data

Enrichment is defined by the output of three panning rounds



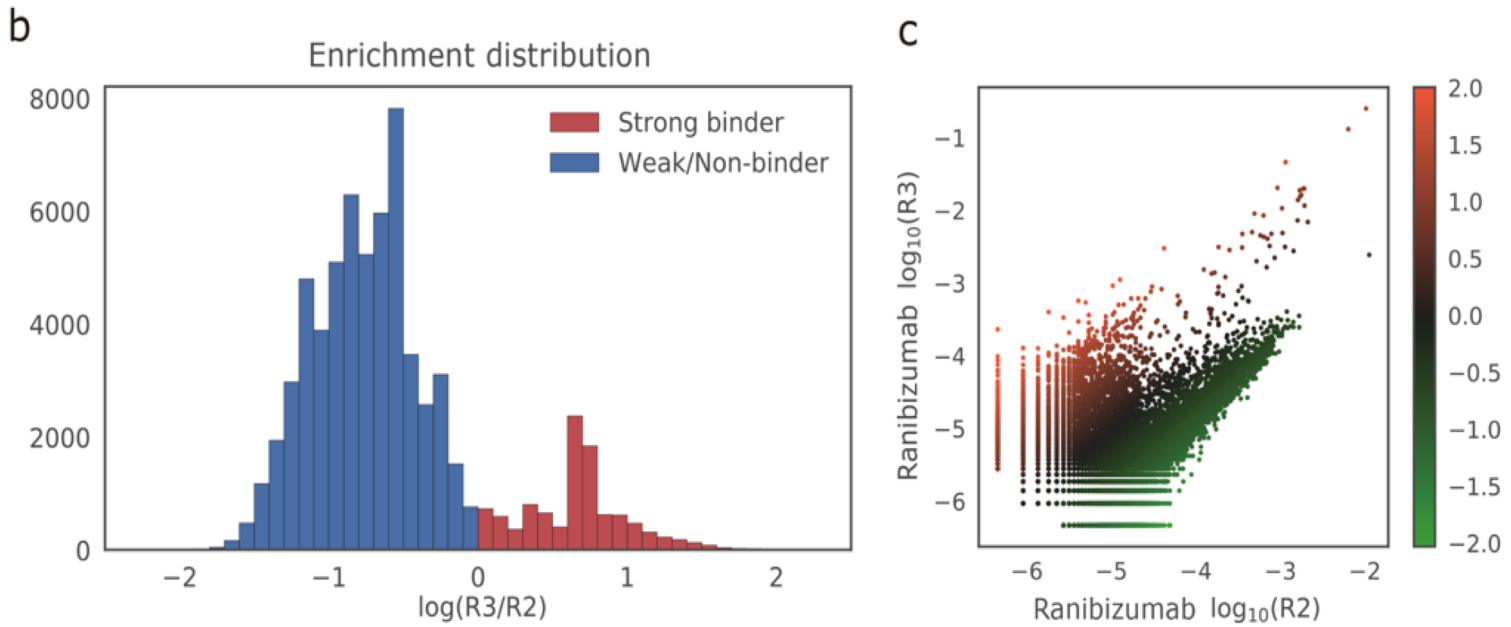
$$\text{Enrichment}_{R3/R1} = \text{Log}_{10}\,(\text{Frequency R3} / \text{Frequency R1})$$

$$\text{Enrichment}_{R3/R2} = \text{Log}_{10}\,(\text{Frequency R3} / \text{Frequency R2})$$

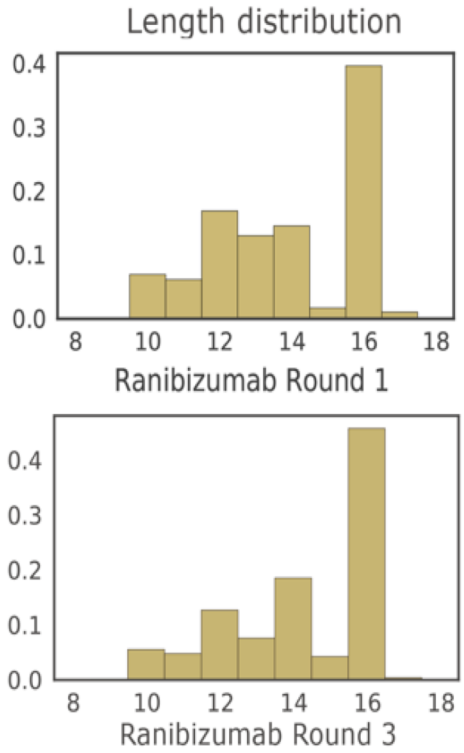
Train on CDR-H3 sequence and enrichment

Sequence	Log(R3/R2)	Enbrel_a	Avastin_a	Herceptin_a	
ADGAFDAYMDY	-0.9561	-0.5989	-0.9730	-1.2414	
ADGYRVYYYAMDY	1.2253	1.4830	0.9872	1.1175	←
ADRRPPLIFFDY	0.8519	0.8458	1.9072	1.9057	
ADWLSLLYRFDY	-0.4779	-0.9202	-0.7767	-0.8339	
AEHVAYHPRYSFDY	-0.9474	-0.7291	-0.9730	-0.8649	
AGRYWWLLDY	0.3242	0.3843	1.7872	0.6588	
AGYHQTWPYGLDY	1.0482	0.8792	0.9135	-0.2221	
AKRRRQYVYHPIYFDY	1.6727	1.4852	1.9769	2.0698	
AKYADTYGLDY	0.4839	0.2024	-0.2996	0.9655	
AKYGSYYGFDY	0.5650	0.3526	0.3929	0.5801	
DAYPGWDLWPDYPFDY	0.2757	-0.0151	-0.0842	0.4879	←
DDIHHLLYYFDY	0.9610	1.1010	0.9135	1.5183	
DDQYVGYFYGEGGLDY	-0.2620	0.0897	0.3372	0.1532	
DDVKGHSKQDLRVFDY	0.7702	-0.0341	1.7246	0.1893	
DDVYWIAAFDY	-0.5247	0.8792	-0.8859	-0.4439	←
DDWYGGLERGLIQFDY	0.2621	-0.0544	1.3027	0.3120	

High enrichment suggests high affinity sequences (Ranibizumab, 67769 sequences)



Ranibizumab binders have preferred CDR-H3 lengths

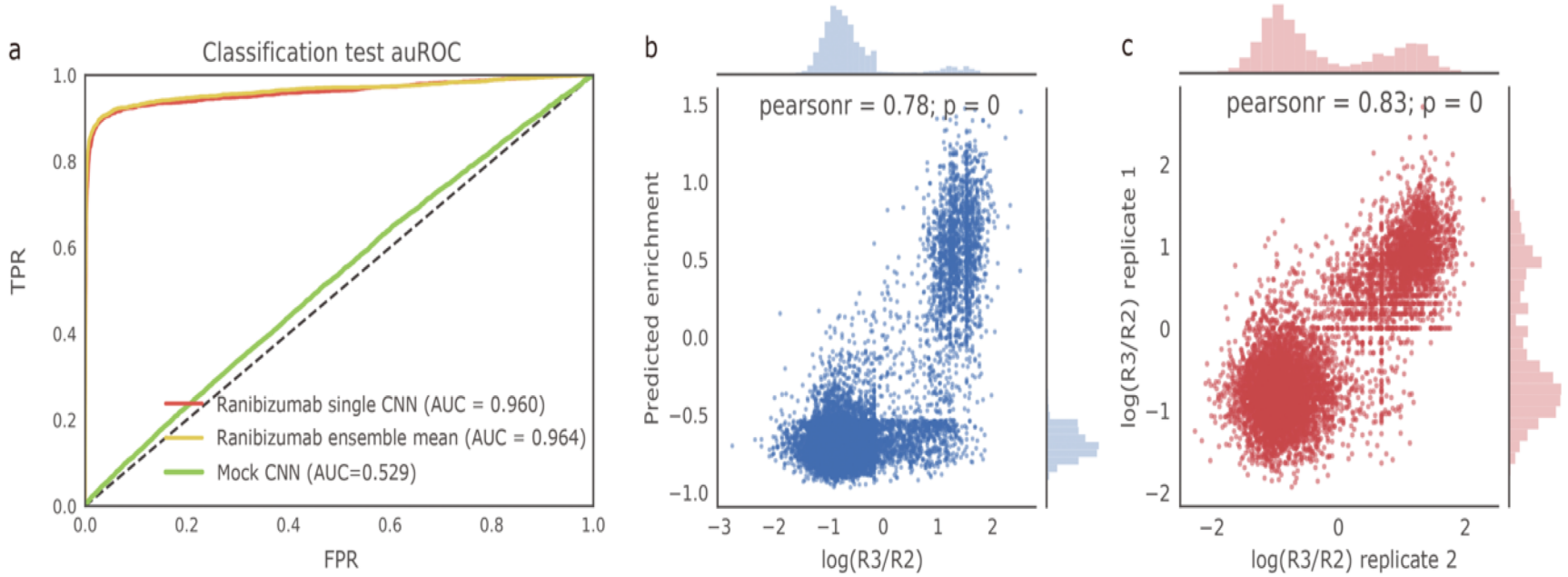


We used six different model architectures

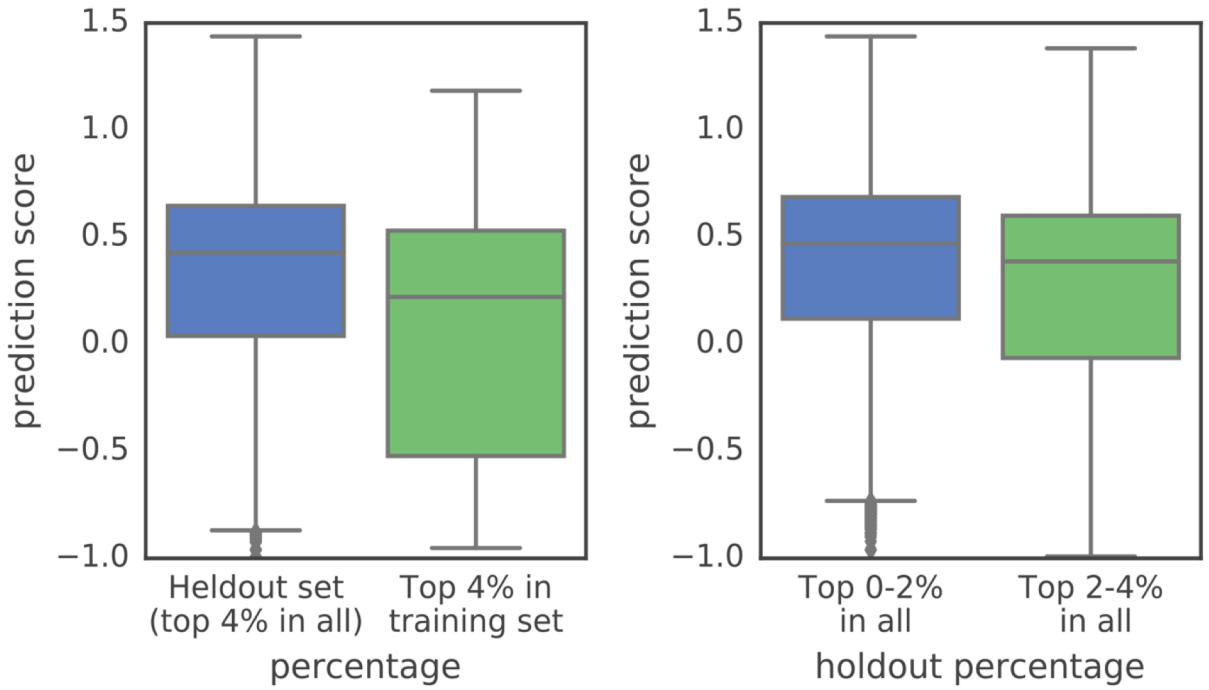
	# of Convolutional layers	# of Convolutional filters	Convolutional filter size	# of Fully connected layer	# of Fully connected neurons	# of parameters in total
2fc	0	0	0	2	32	13954
1conv(32*5)+1fc	1	32	5	1	16	8402
2conv(32*5_64*5)+1fc	2	32,64	5	1	16	18706
1conv(64*5)+1fc	1	64	5	1	16	16754
1conv(32*3)+1fc	1	32	3	1	16	7122
2conv(8*1_64*5)+1fc	2	8, 32	1, 5	1	16	13082

Output layer: Classification – binary cross entropy loss Regression – mean squared error

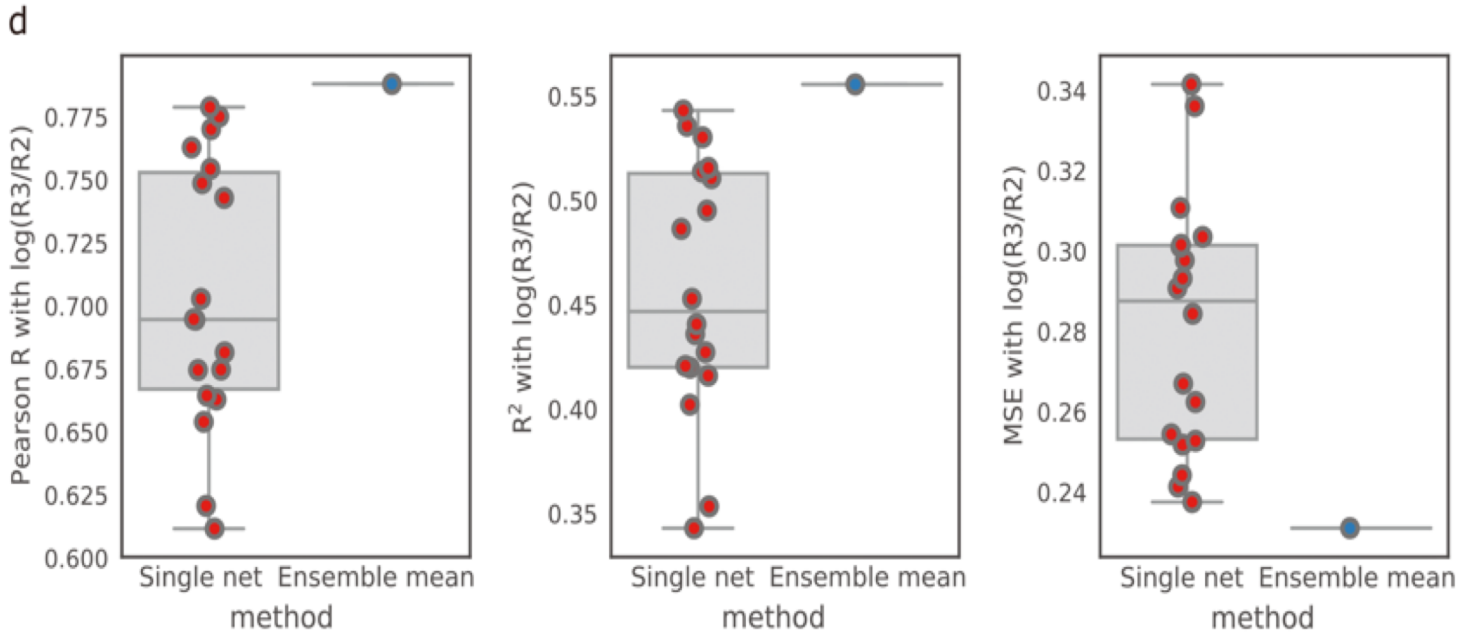
Regression performance is comparable to replicate experiment performance



CNNs produce better scores than they have seen in training for top sequences



An ensemble of 24 networks is more robust than the individual networks



How can we optimize CDRs?

Design flow

Naïve library 10^{10} CDRH3 sequences

↓

Multiplexed Affinity Assay

↓ Training Data

ML Design Methods

↓

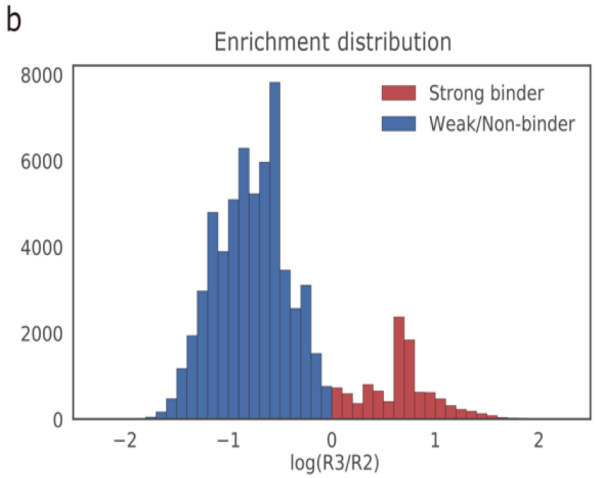
Naïve library → ML library fabrication

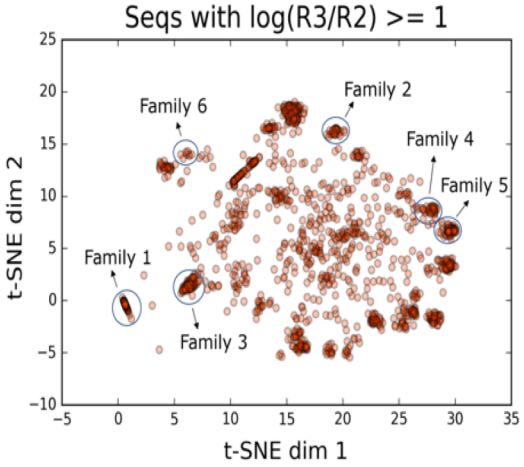
↓

Multiplexed Affinity Assay

↓

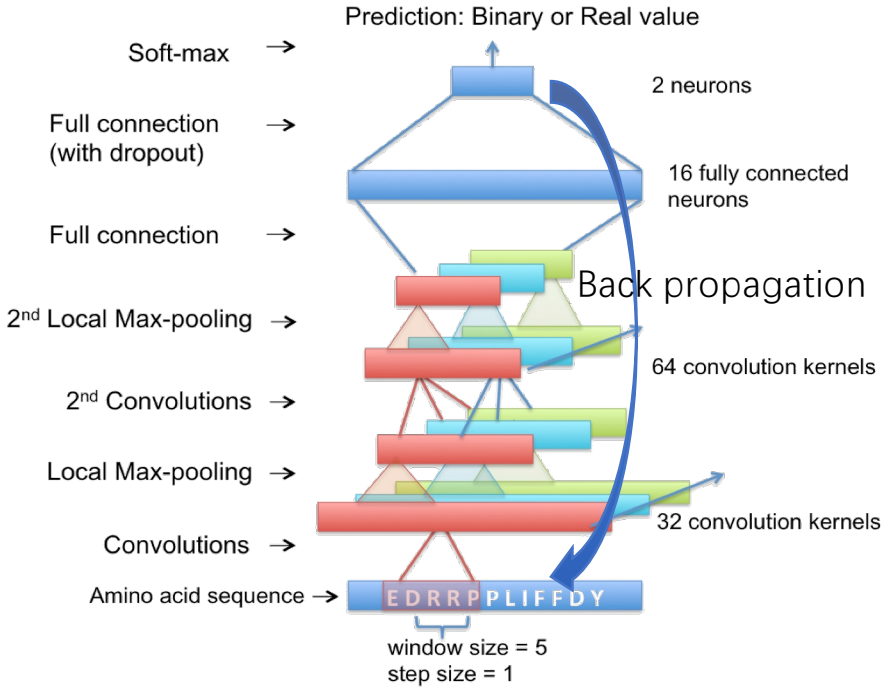
Full length antibody characterization



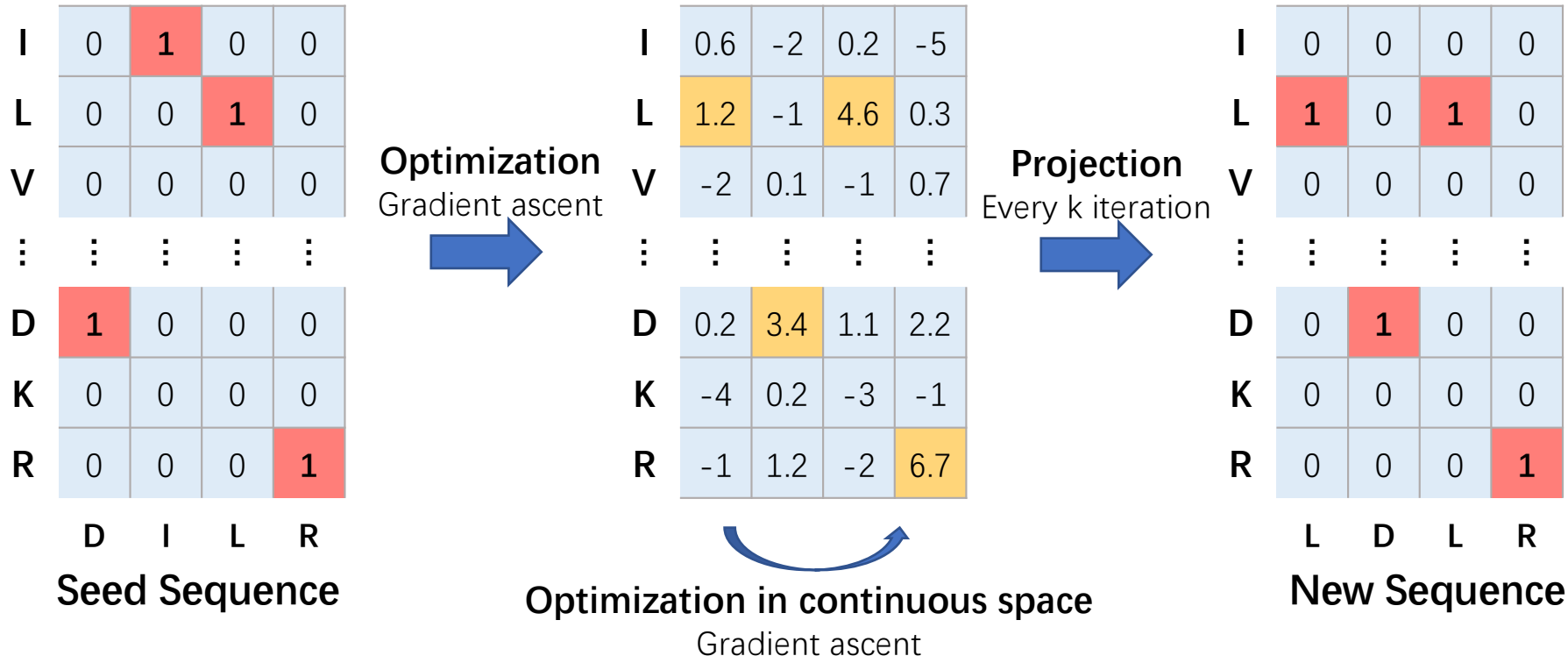


Training Data

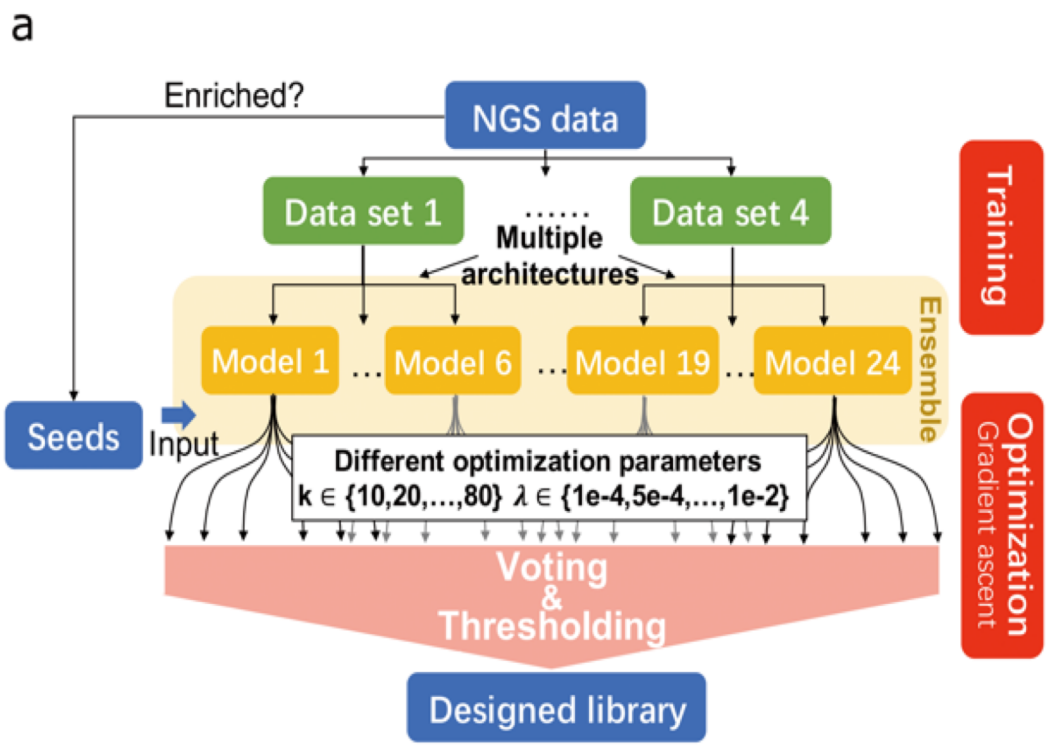
Our model from sequence to enrichment is differentiable
Method 1 - Optimization with gradients



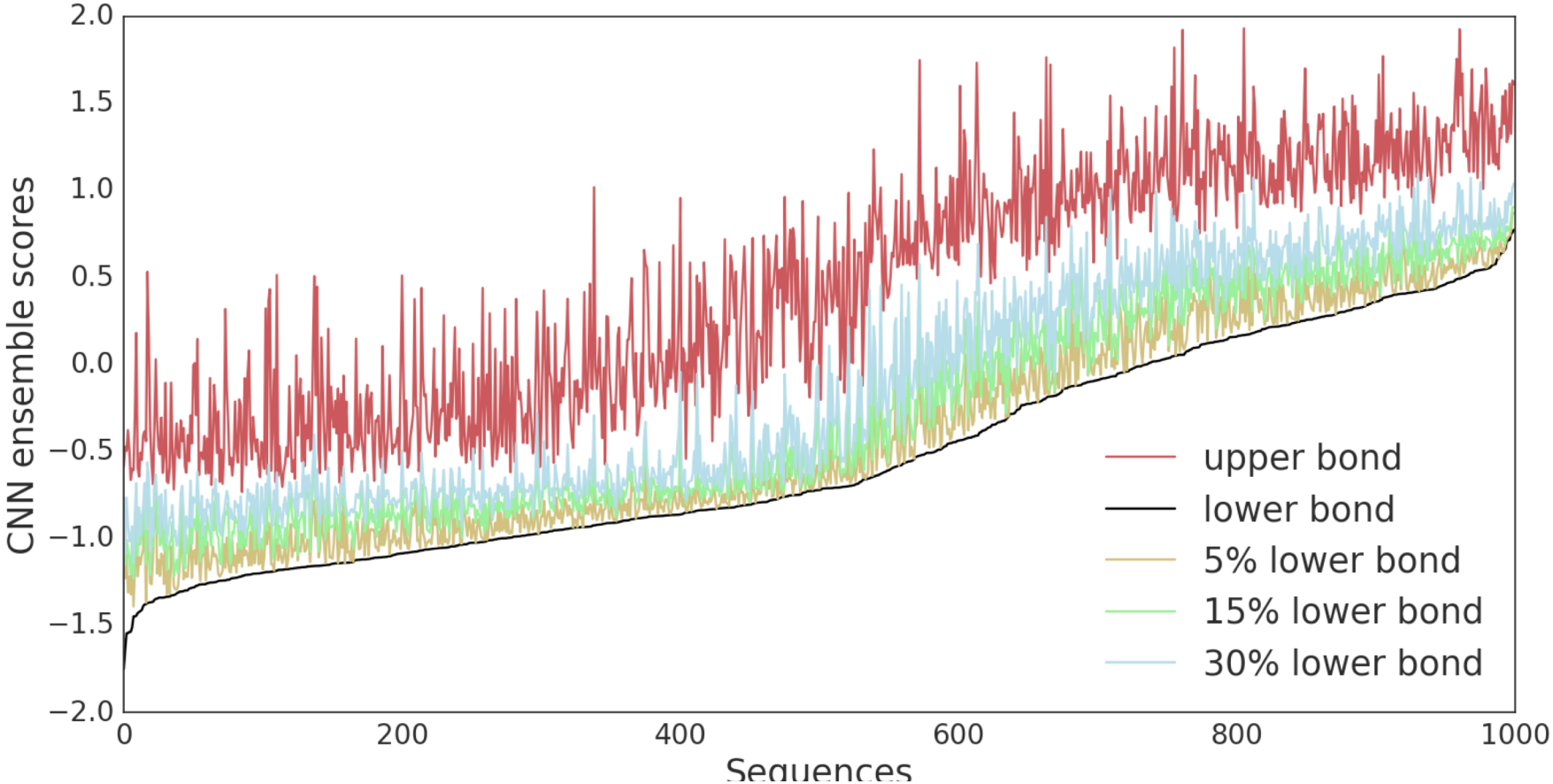
Projecting continuous representation into one-hot representation



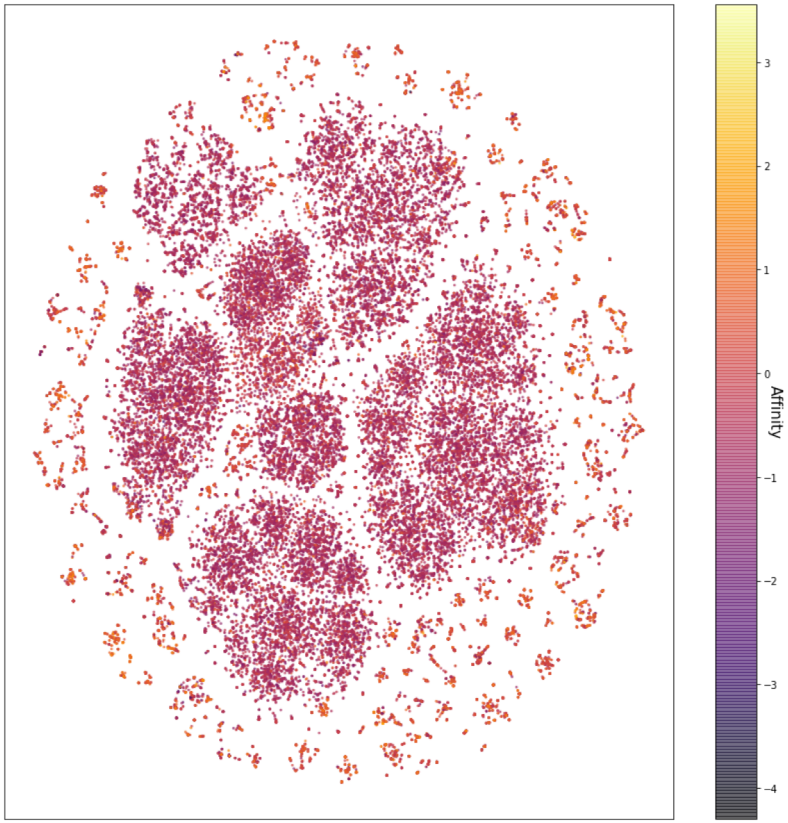
Ens-Grad uses voting across ensembles and hyper-parameters to choose sequences

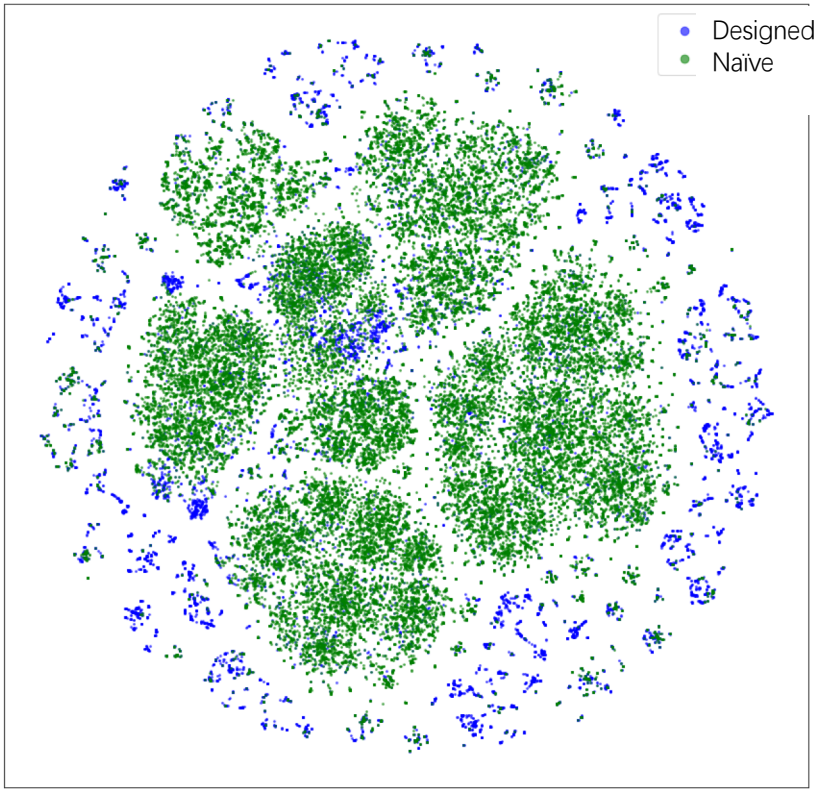


Ranked sequences are filtered by the lower bound from all 18 networks

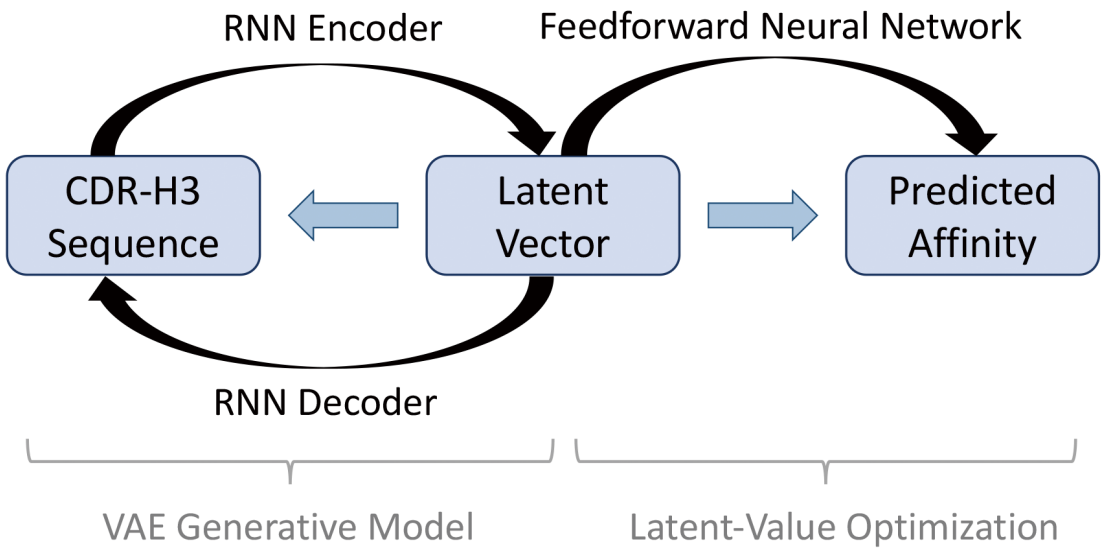


Designed sequences appear in islands of enrichment





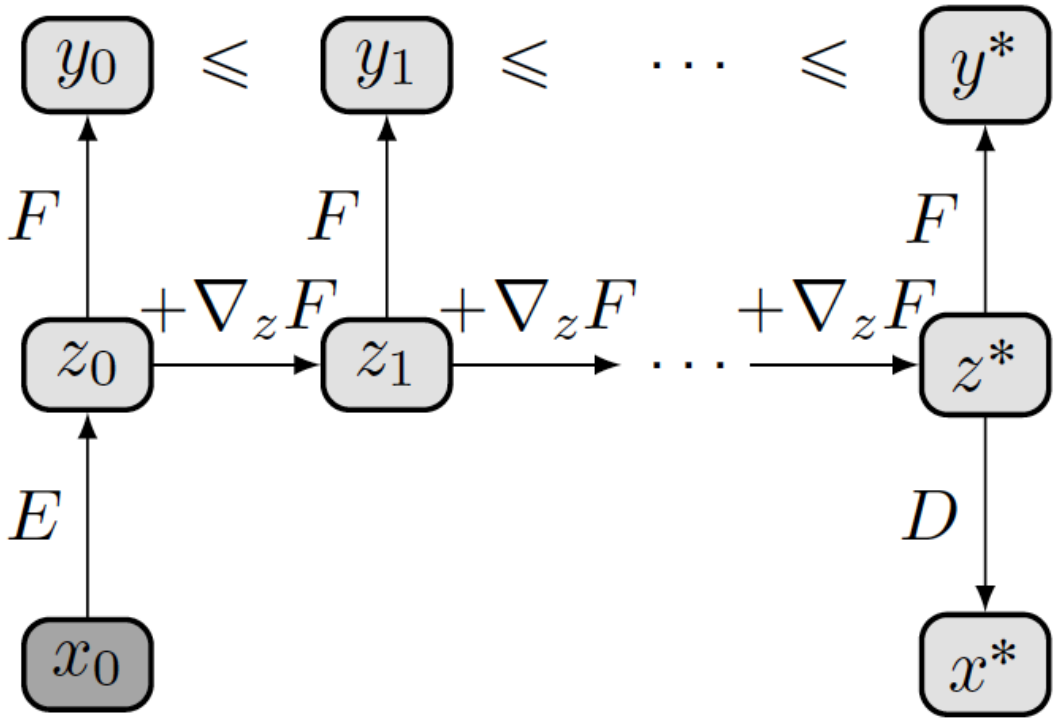
Method 2- Optimize in latent space with a variational autoencoder



VAE diagram

Jonas Mueller, David Gifford, Tommi Jaakkola ;
Proceedings of the 34th International Conference on Machine Learning, PMLR 70:2536-2544, 2017.

Method 2- Optimize in latent space with a variational autoencoder



Design flow

Naïve library 10^{10} CDRH3 sequences

↓

Multiplexed Affinity Assay

↓ Training Data

ML Design Methods

↓

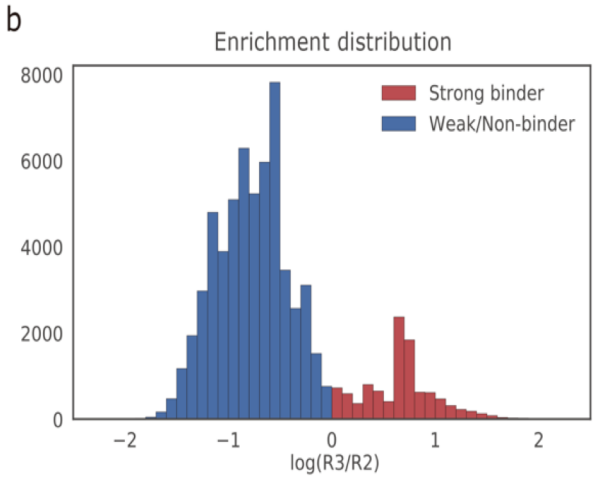
Naïve library → ML library fabrication

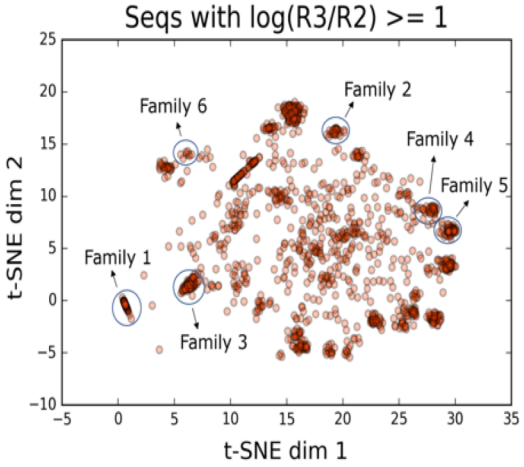
↓

Multiplexed Affinity Assay

↓

Full length antibody characterization



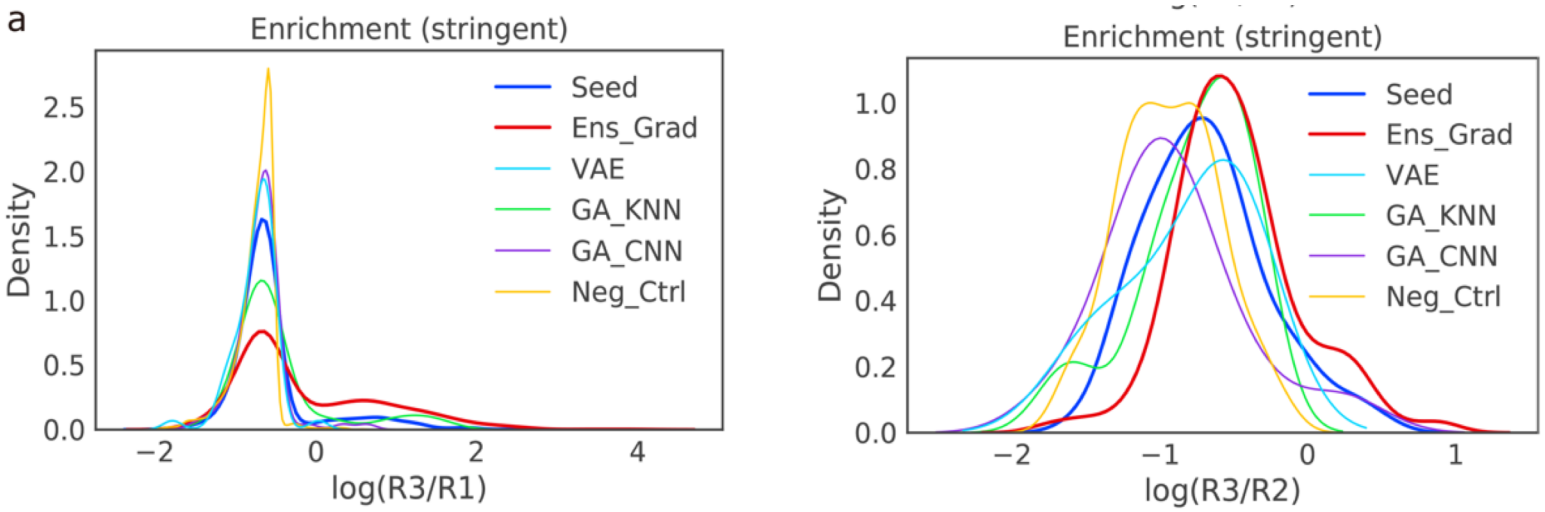


Training Data

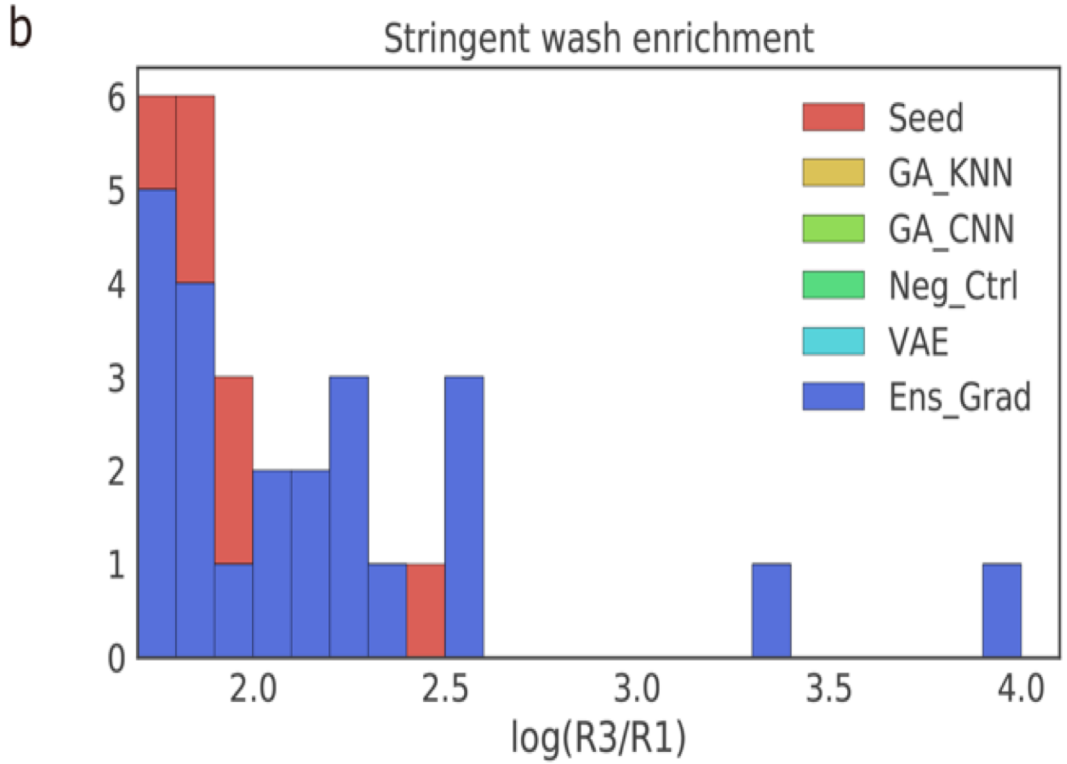
Testing of Fab sequences by direct synthesis

- We computed 77,596 novel machine learning proposed CDR-H3 sequences (Ens-Grad 5,467 sequences)
- We added 26,939 controls and synthesized a total of 104,525 oligonucleotides encoding CDR-H3 sequences
- The oligonuclotides were cloned into a Fab framework and expressed on phage
- Our library with complexity 10^5 was mixed 1:100 into a native library of complexity ~10^{10}
- The combined library was subject to rounds of panning

Ens-Grad sequences are on average more enriched than seeds and the synthetic results of other ML methods

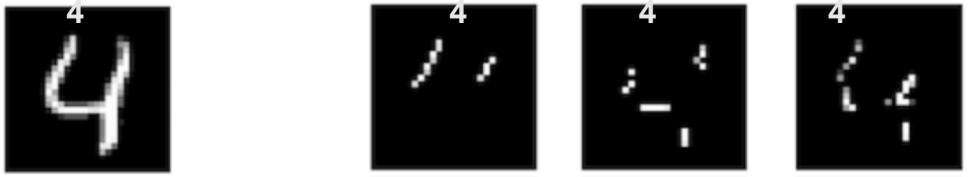


Top scoring synthetic CDR_H3 sequences were designed by Ens Grad (Stringent Wash)



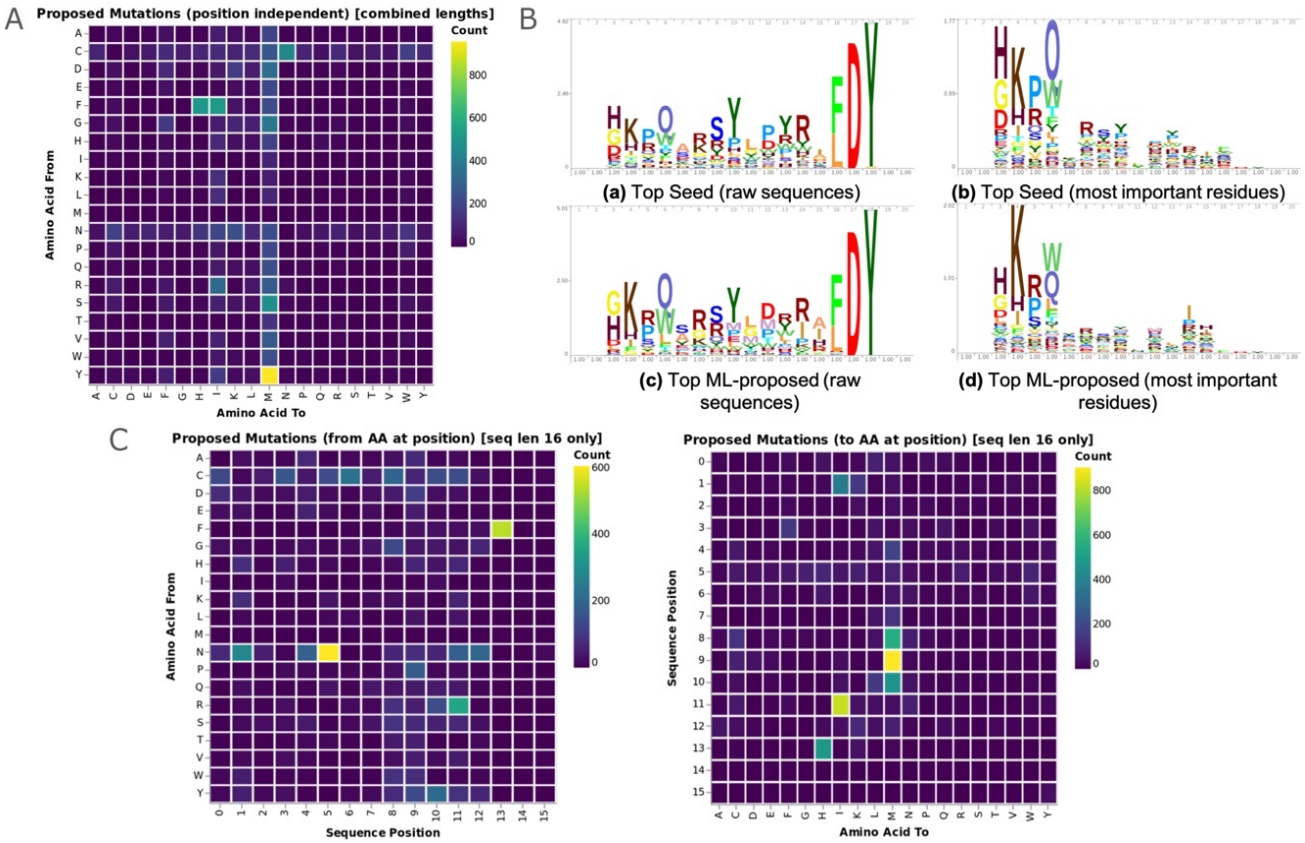
Sufficient Input Subsets provide model interpretation

- One simple rationale for ***why*** a black-box decision is reached is a sparse subset of the input features whose values form the basis for the decision
- A ***sufficient input subset*** (SIS) is a minimal feature subset whose values alone suffice for the model to reach the same decision (even without information about the rest of the features' values)



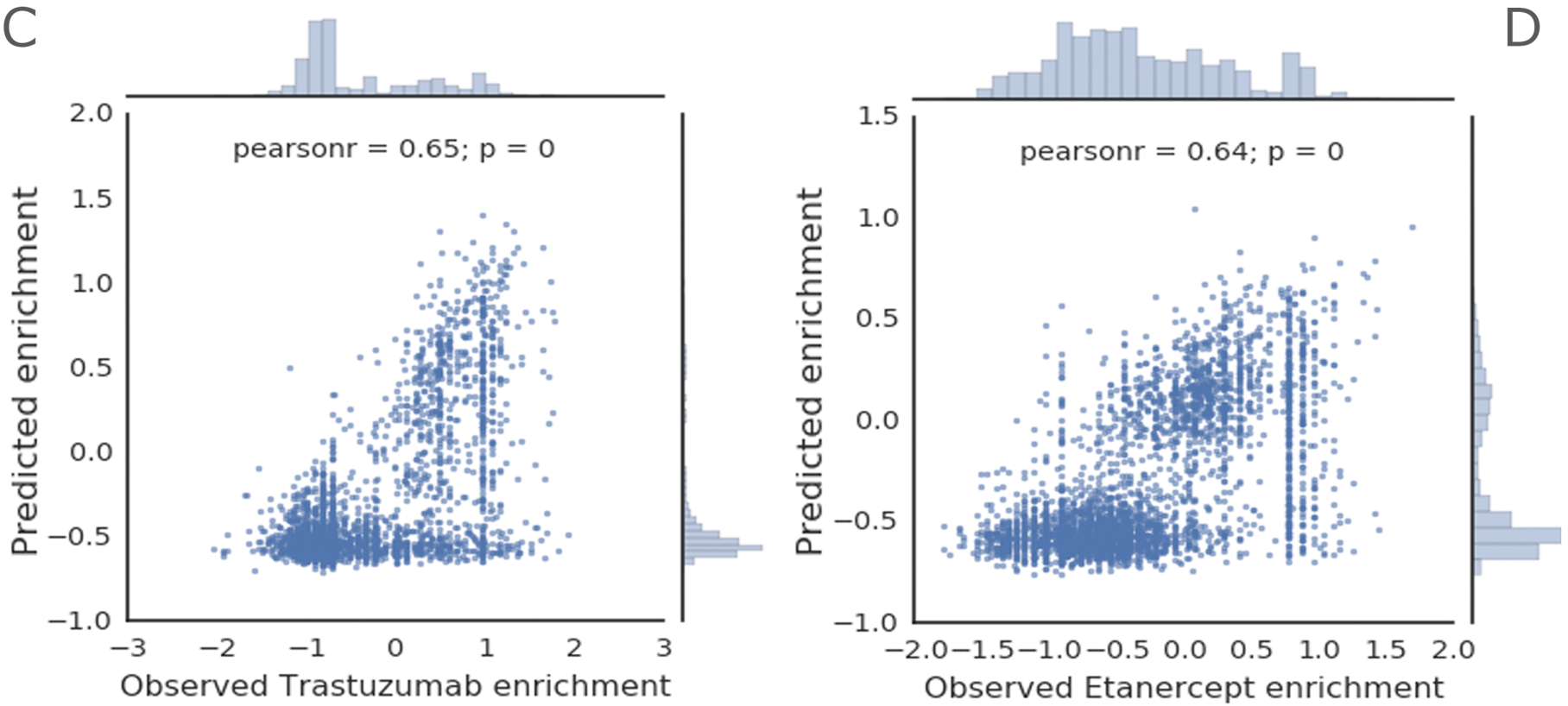
	CDR-H3 Sequence	Group	R^2	EC50(nM)	Standard log(R3/R1)	Stringent log(R3/R1)
Family 1	HKPQAKSYLPLRLLDY	Ens_Grad	0.99	0.47	3.369	2.399
	HKPQAISYLPYRLLDY	Ens_Grad	0.998	0.5	2.61	2.577
	HKPQAISYLPYRILDY	Seed	0.993	0.62	2.418	2.467
	HKPQAKSYLPMRLLDY	Ens_Grad	0.98	0.93	2.409	0.836
	HKPQAVSYLPYRILDY	Ens_Grad	0.994	0.98	2.915	2.561
	HKPQAKSYLPYRLLDY	Seed	0.996	1.48	2.693	1.128
	HKPQAKSYLPYRTLDY	Seed	0.993	2.49	2.371	1.986
	HKPQSKSYLPYRLLDY	Seed	0.995	4.78	2.634	0.445
	HKPQAKSYLPYRILDY	Seed	0.992	6.55	1.41	1.112
Family 2	YRSPHHRGGATWQFDY	Seed	0.992	5.79	-0.037	0.036
Family 3	DLFRYYYFMWPLDY	Ens_Grad	0.986	34.05	2.638	0.523
	DLFRYYYFFWPLDY	Seed	0.99	109.5	2.988	1.283
Family 4	MHYYDIGVFPWDTFDY	Ens-Grad	0.971	**0.29**	2.089	3.381
	GHYYDIGVFPWDTFDY	Seed	0.99	0.49	0.703	1.593
Family 5	WQQWAGYPRQKYSFDY	Seed	0.986	3.31	2.657	1.888
	WQQWSGYPRQKYSFDY	Seed	0.975	66.81	0.264	-0.219
Family 6	GKSLYGQETTWPHFDY	Seed	0.99	0.67	2.002	0.946

What ML changes facilitate enrichment improvement?

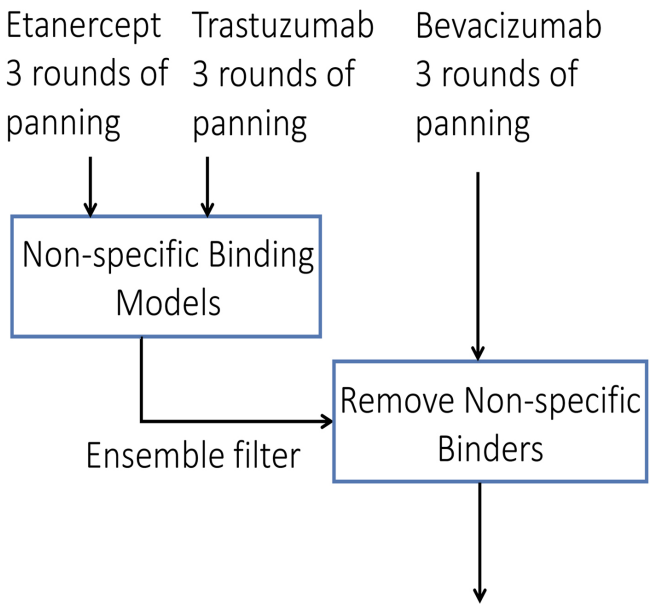


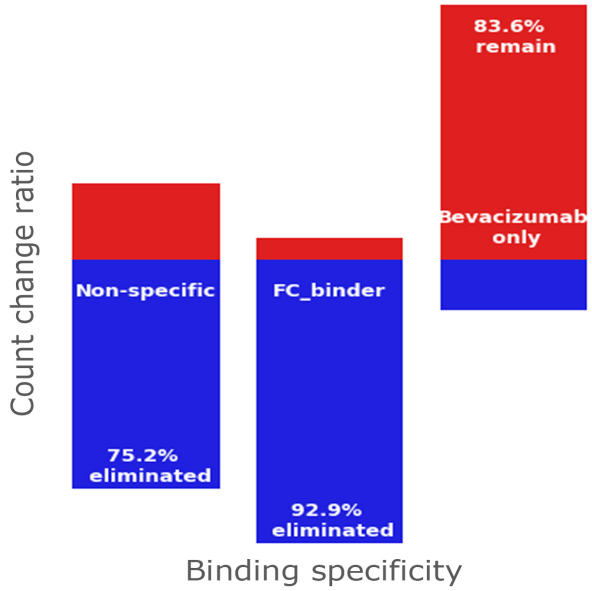
Multi-objective optimization for specificity

We can build a joint model of Trastuzumab and Entanercpet binding



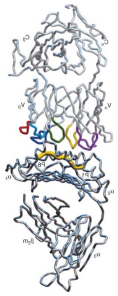
We can remove non-specific binders

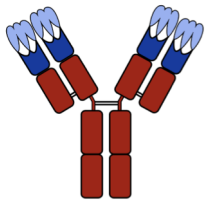




We wish to both identify therapeutic targets and therapeutic molecules with the help of machine learning

- Identifying target peptide-MHC molecules
  - Best prediction model
  - Likelihood metrics can optimize vaccine formulation
- Designing antibody Complementarity Determining Regions (CDRs)
  - Differentiable models from sequence to objective
  - Enables multi-objective optimization





Acknowledgements

MIT

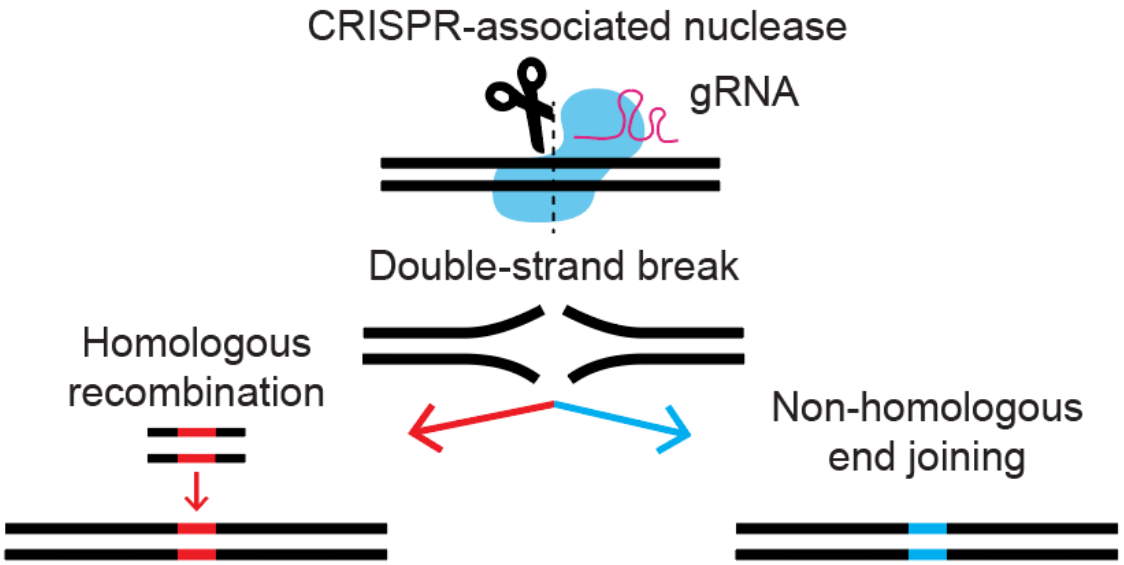
- Saber (Ge Liu)
- Haoyang Zeng
- Jonas Mueller
- Brandon Carter
- Ziheng Wang (Tony)
- Nathan Hunt
- Michael Birnbaum

Novartis

- Stefan Ewert
- Jonas Schilz
- Geraldine Horny

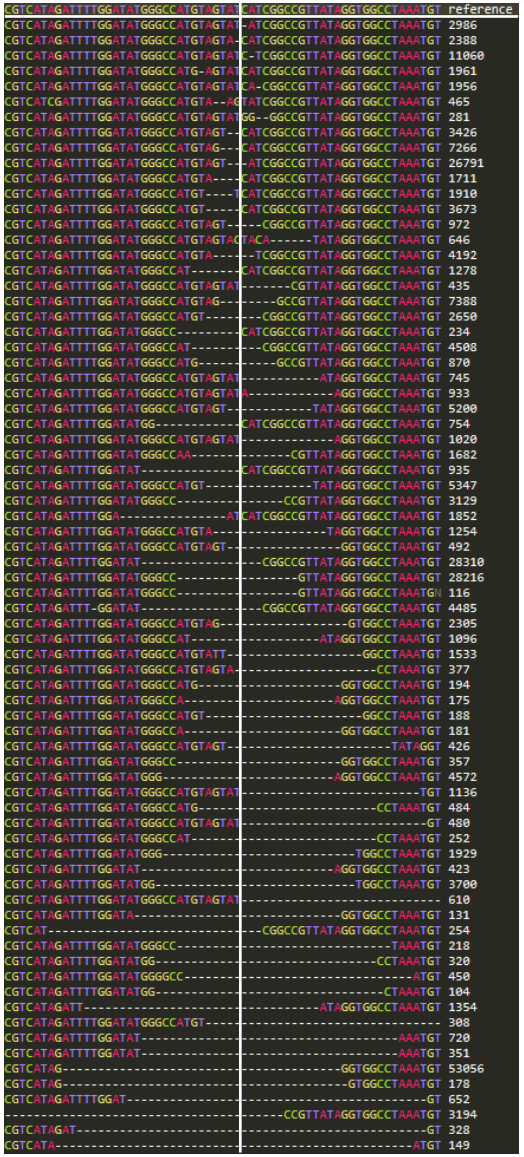
Predictable and precise template-free editing of pathogenic mutations by CRISPR-Cas9 nuclease

The state of CRISPR genome editing

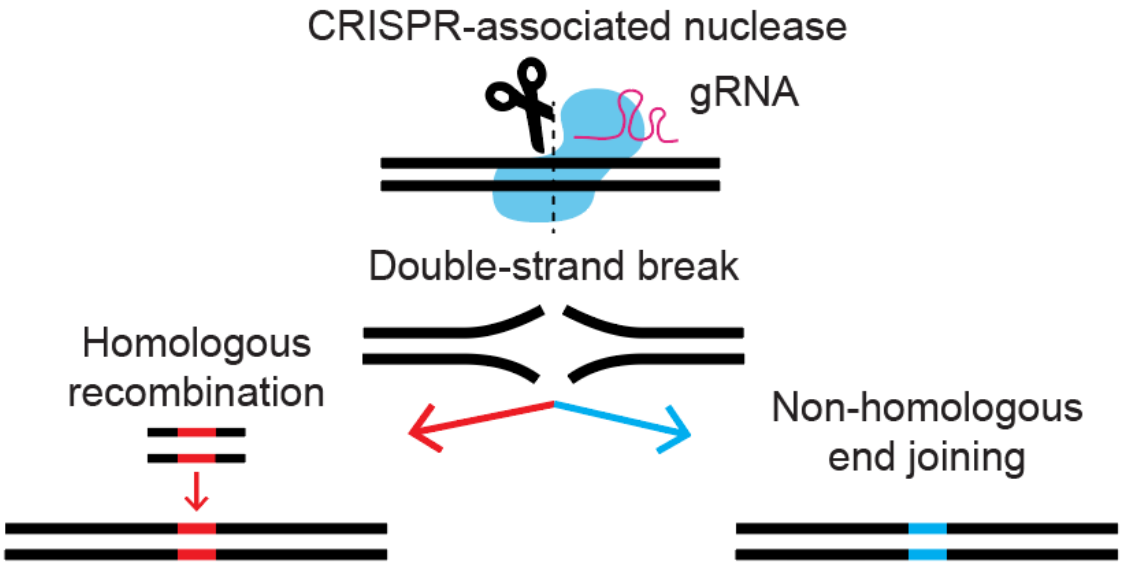


- often inefficient
- designable
- undesirable byproducts

- efficient
- random indels
- highly heterogeneous
- impractical beyond gene disruption

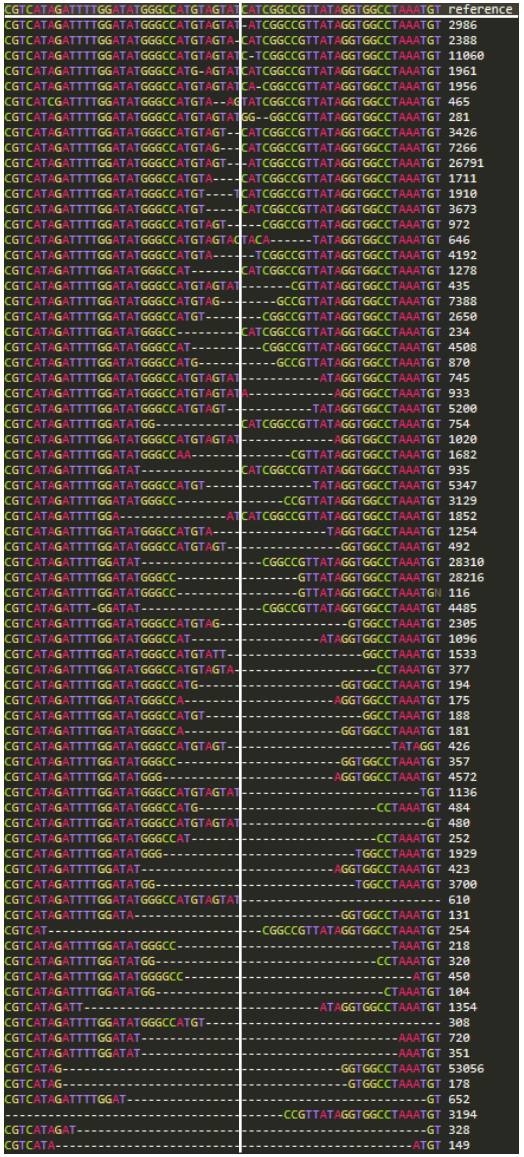


The state of CRISPR genome editing

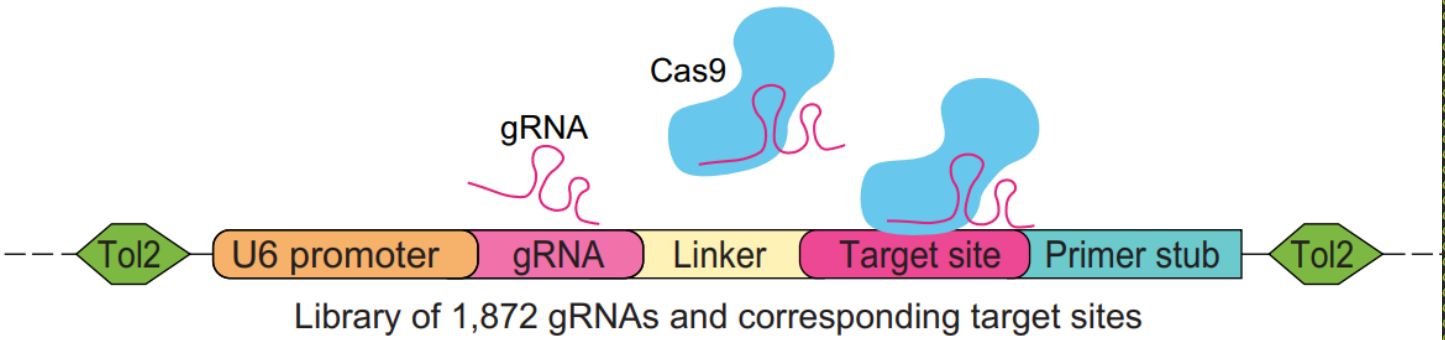


- often inefficient
- designable
- predictable byproducts

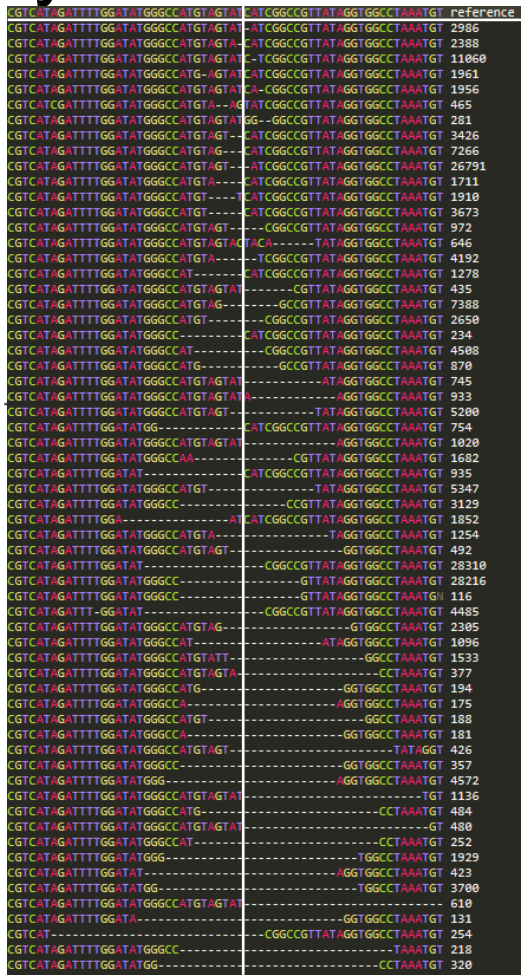
- efficient
- predictable indels
- can be homogeneous
- practical: repair of pathogenic alleles to wild-type



High-throughput genome-integrated assay of Cas9-mediated DNA repair



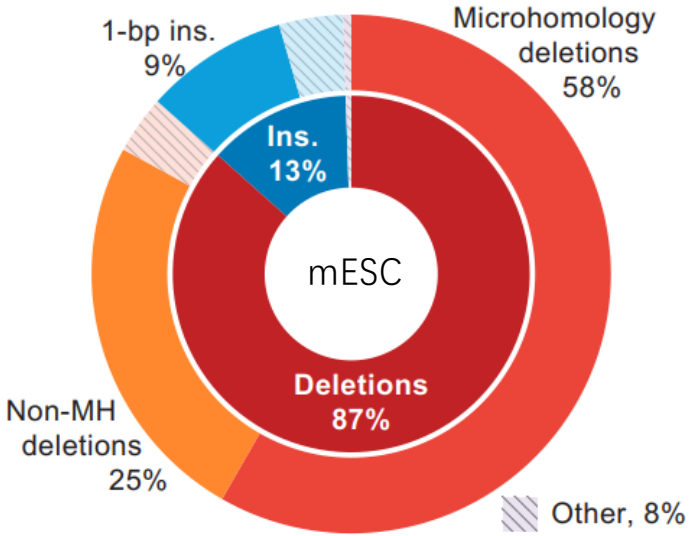
Library of 1,872 gRNAs and corresponding target sites



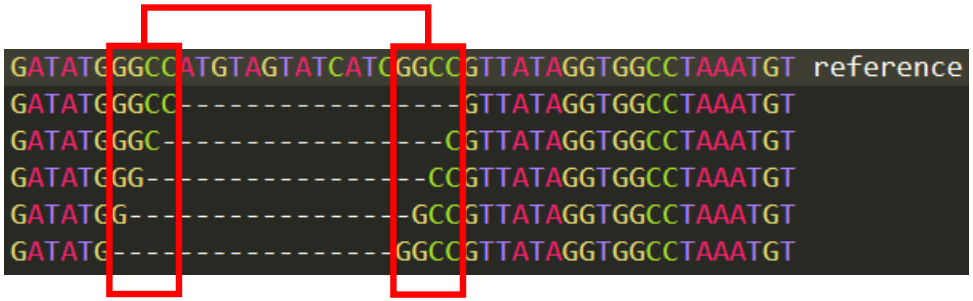
- 96 target sites in largest previous study
- Designed 1,872 target sites (55-bp) based on the human genome

Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

Cas9 primarily causes microhomology deletions in genome-integrated and endogenous settings

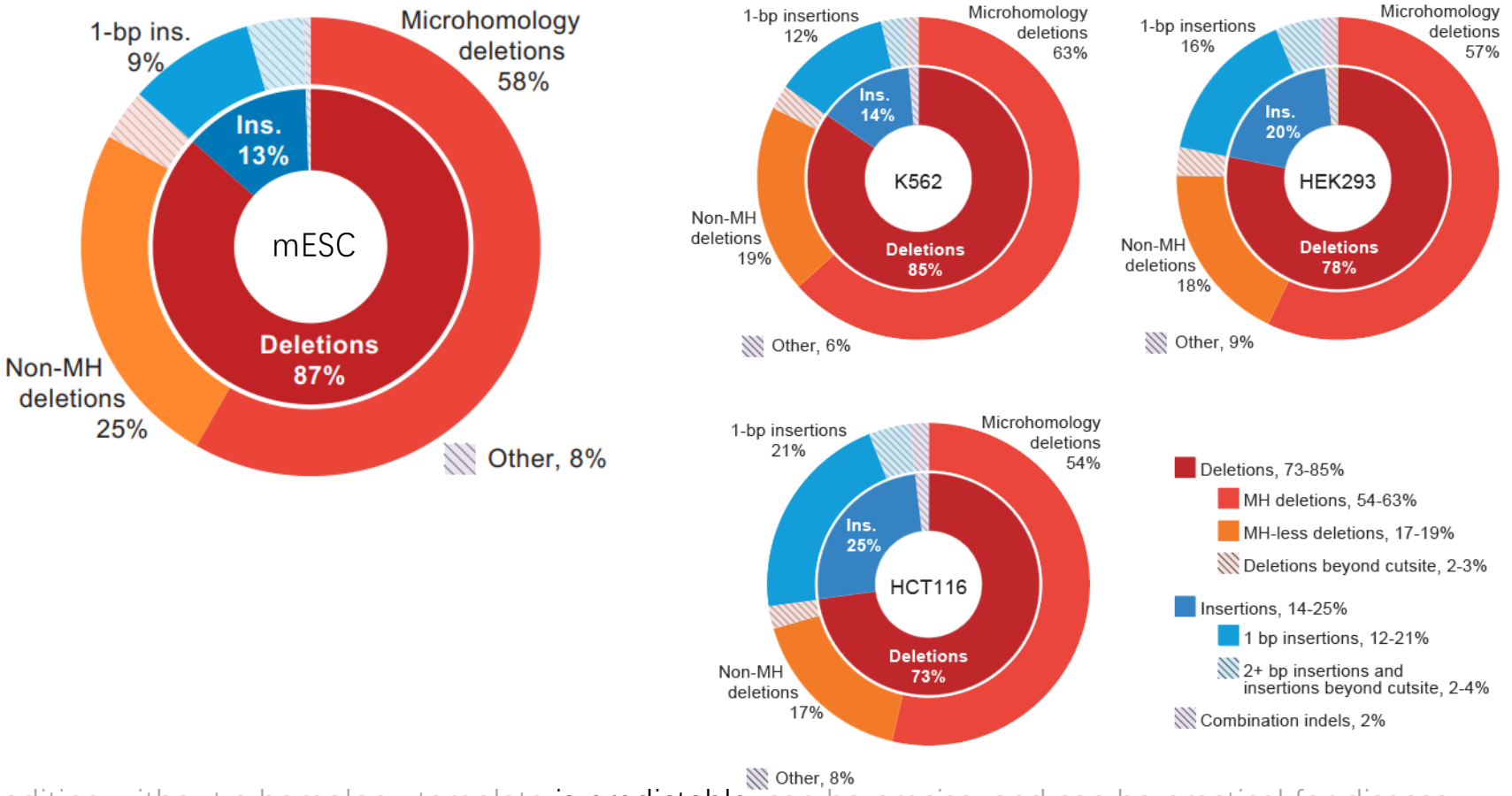


A microhomology deletion is a deletion with multiple equal-scoring alignments



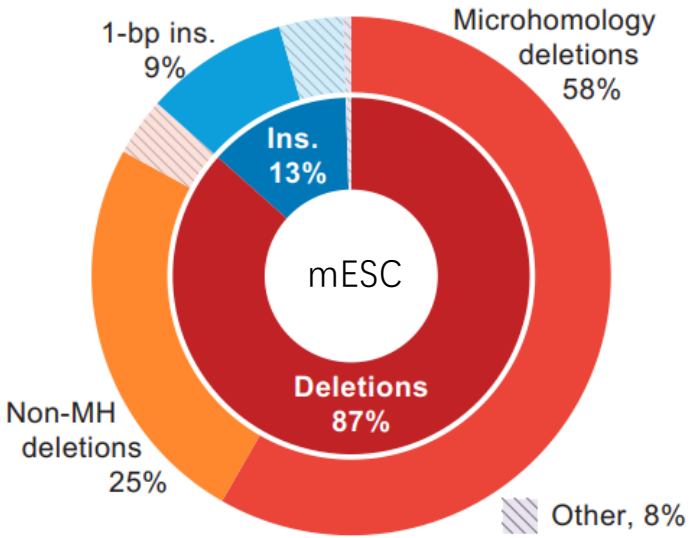
Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

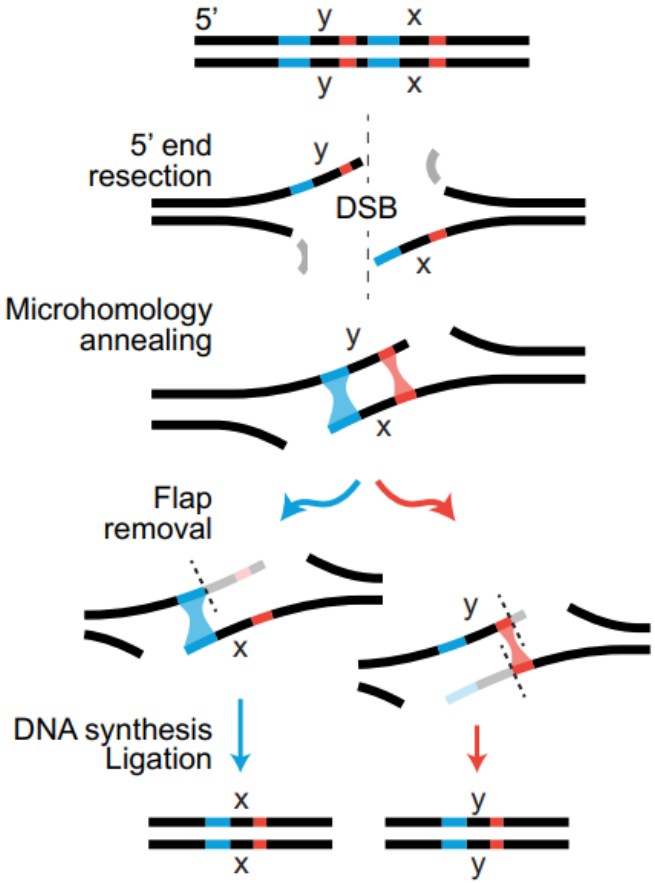
Cas9 primarily causes microhomology deletions in genome-integrated and endogenous settings



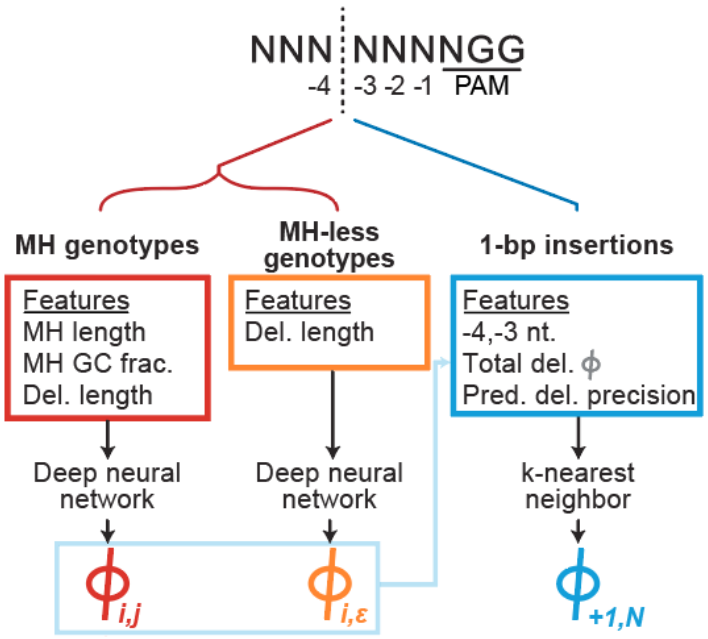
Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

Majority of repair products arise from microhomology-mediated end-joining (MMEJ)

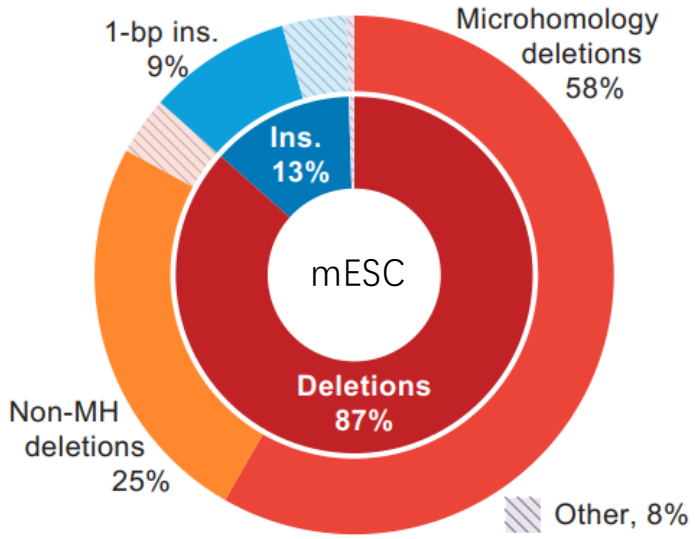




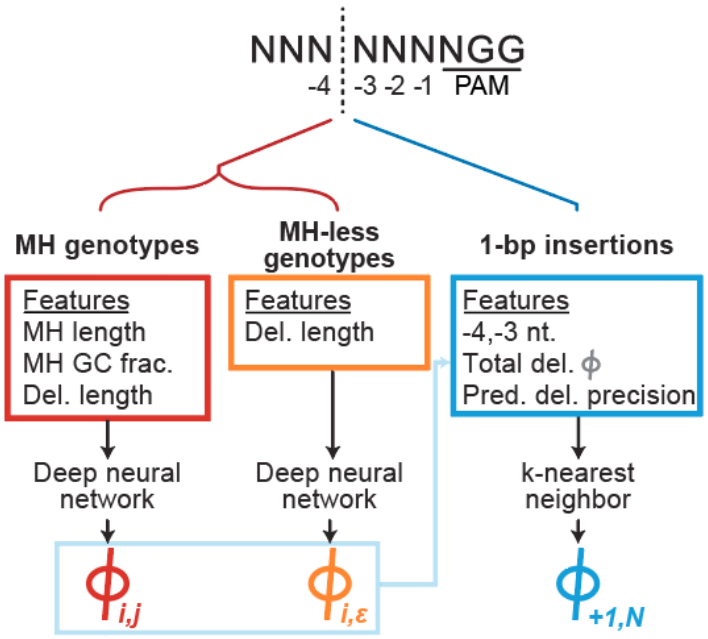
Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease



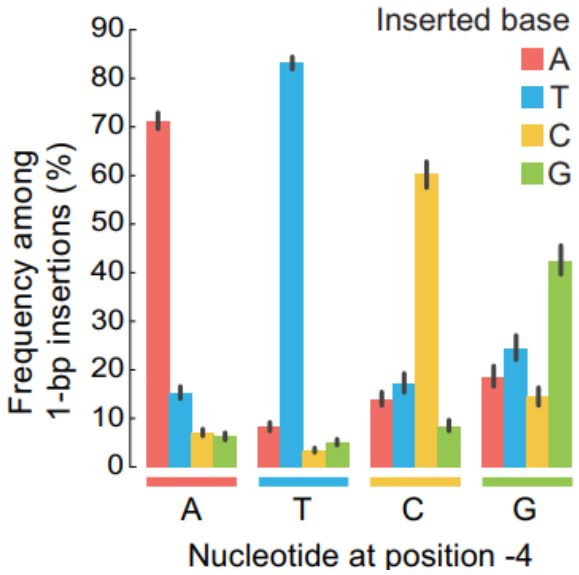
inDelphi predicts 90% of repair products from 3 major repair classes



Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

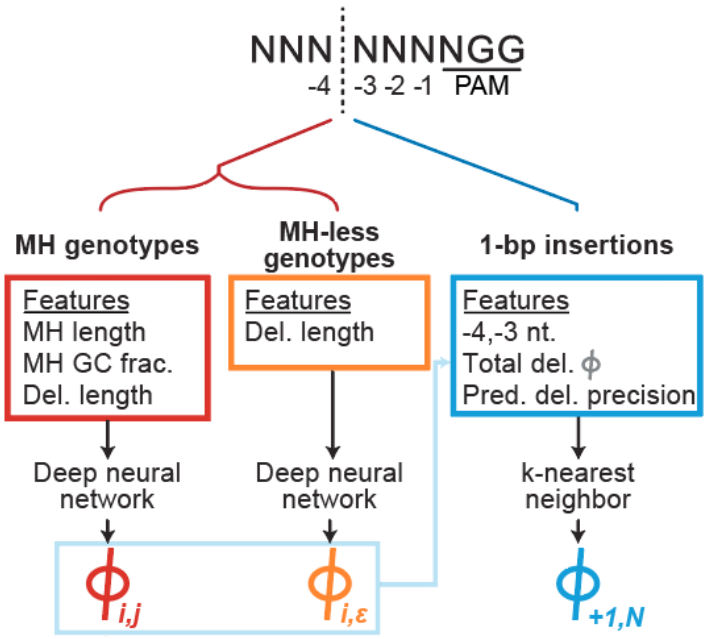


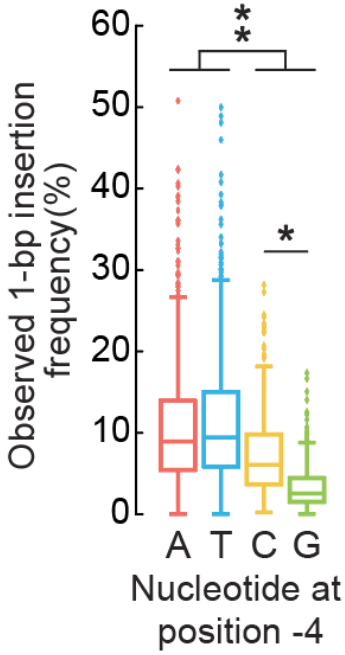
1-bp insertions copy the adjacent nucleotide

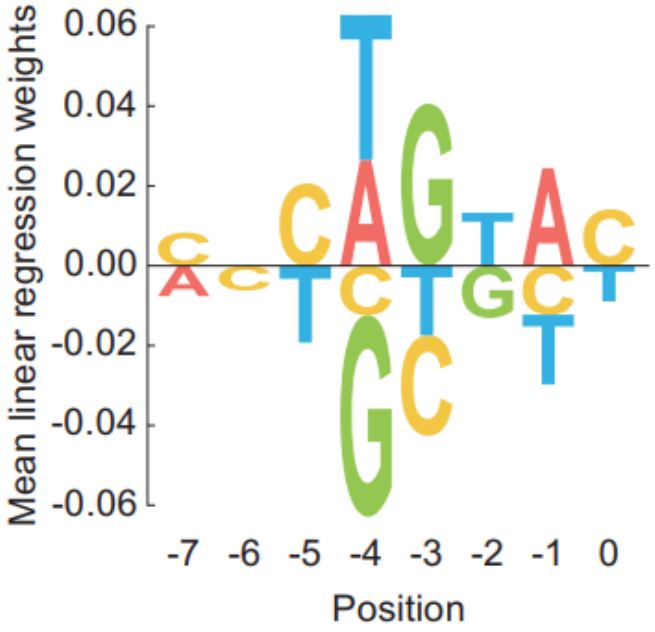


Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

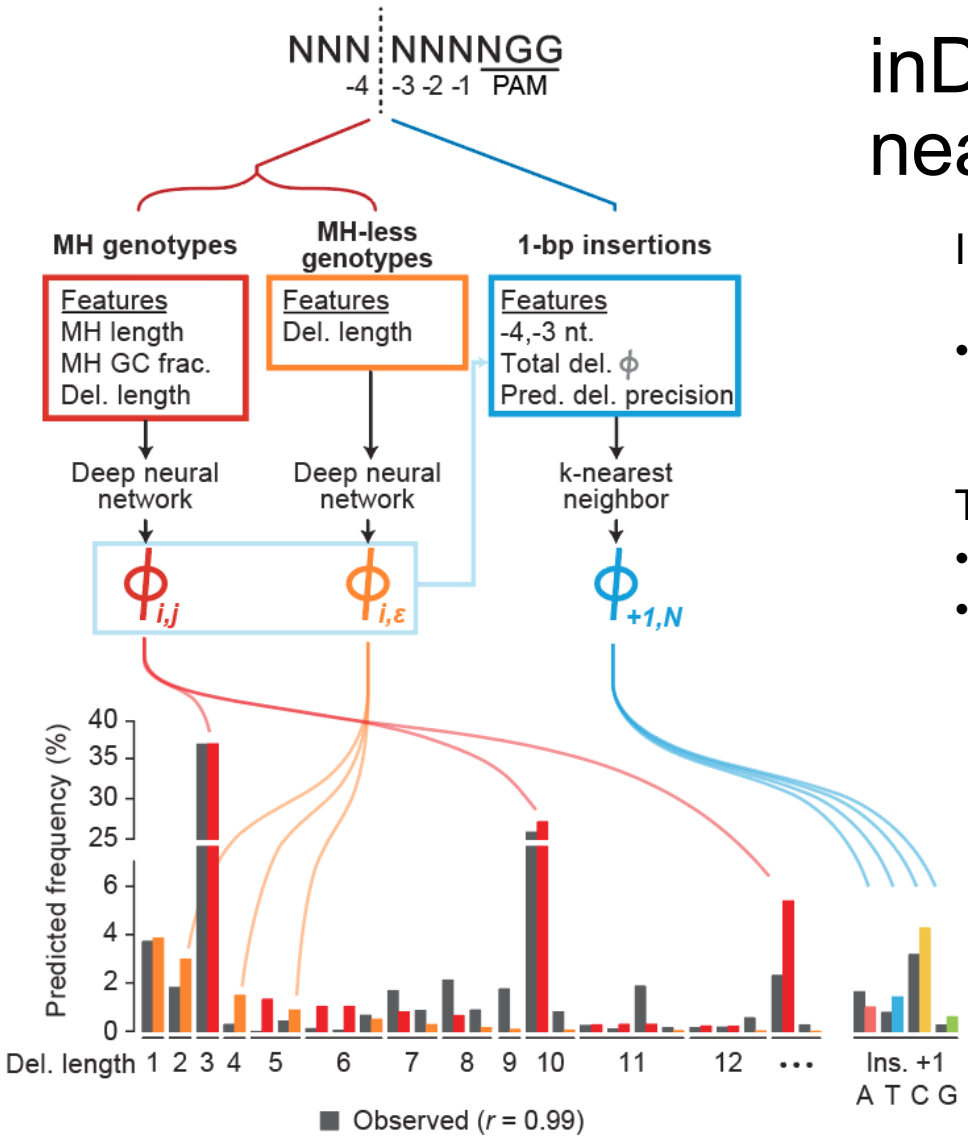
1-bp insertion frequency depends on local sequence context







Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

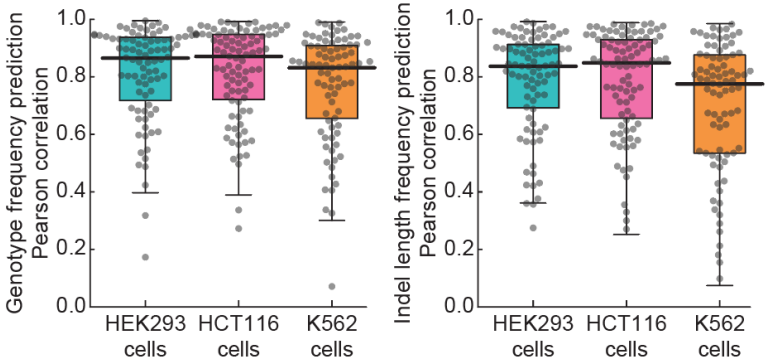


inDelphi accurately predicts nearly all repair outcomes

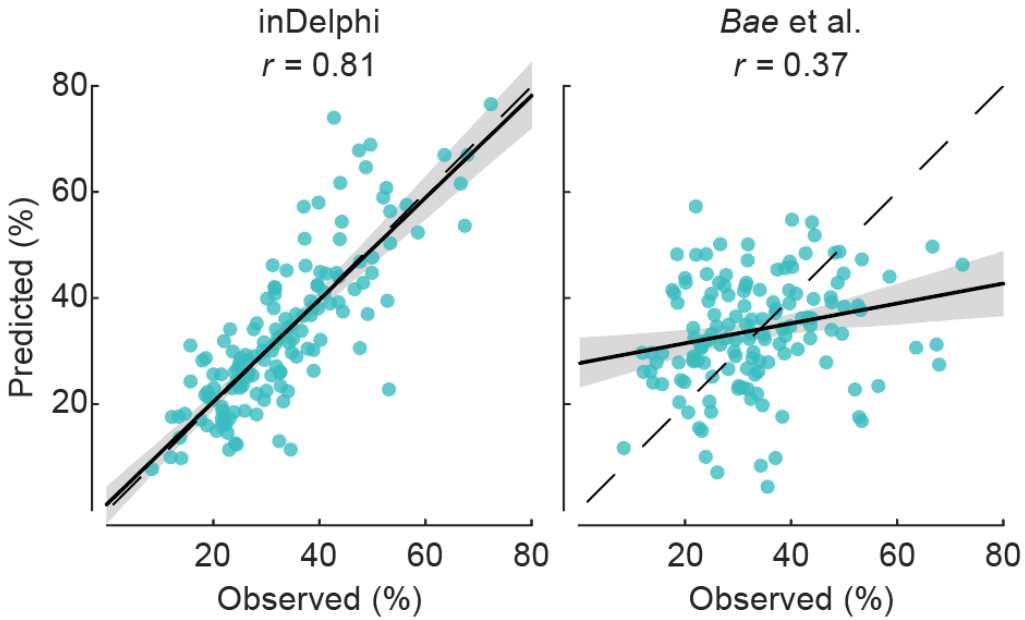
Input: Sequence, cutsite

- Predicts 90% of observed repair outcomes
  - 70% at single-base resolution

Training & testing on held-out cell-types
- Median r = 0.87 on genotype prediction
- Median r = 0.84 on indel length prediction

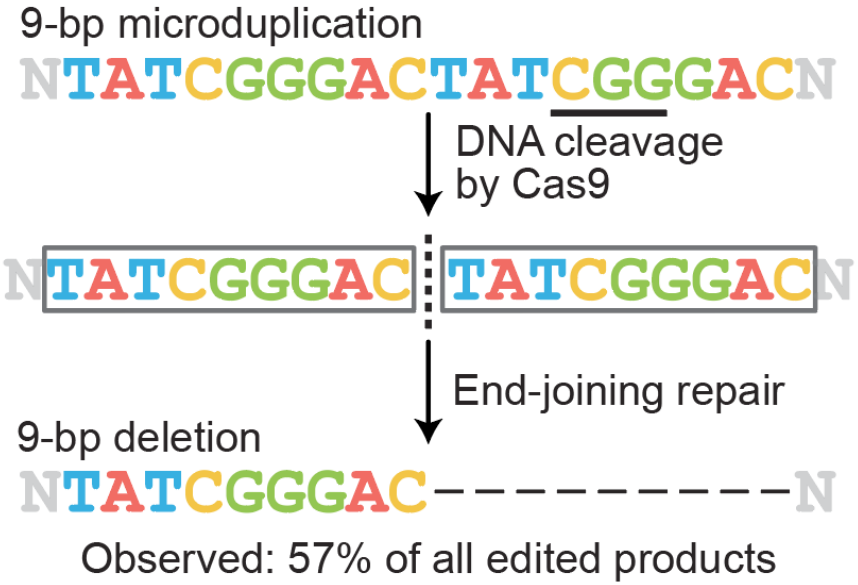


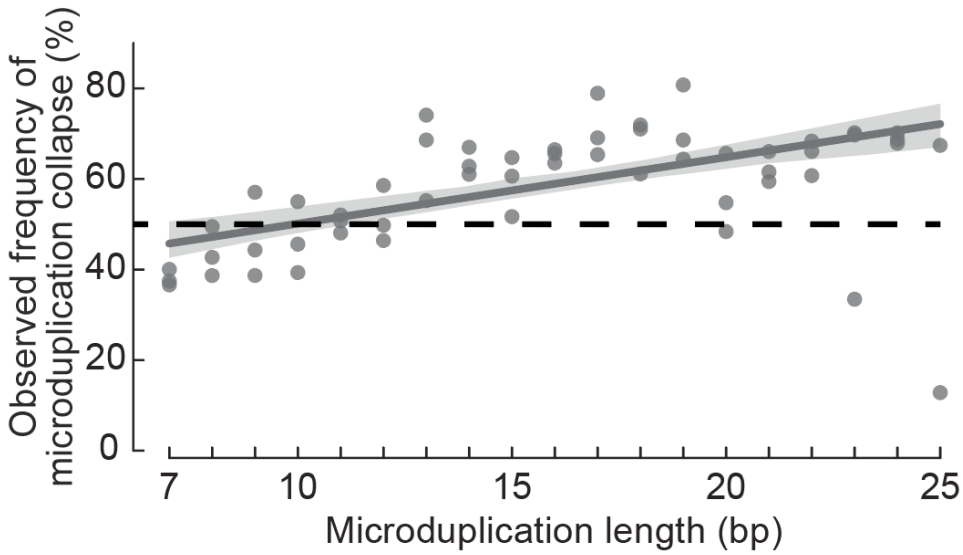
inDelphi accurately predicts frameshifts



Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

Target sites yielding a single deletion repair genotype >50% of the time



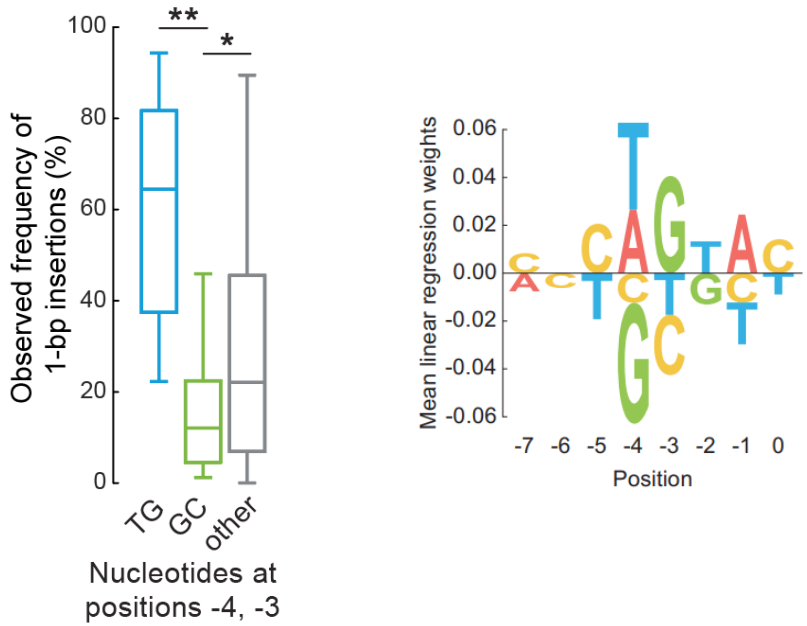


Target sites yielding a single insertion repair genotype >50% of the time

Weak microhomology

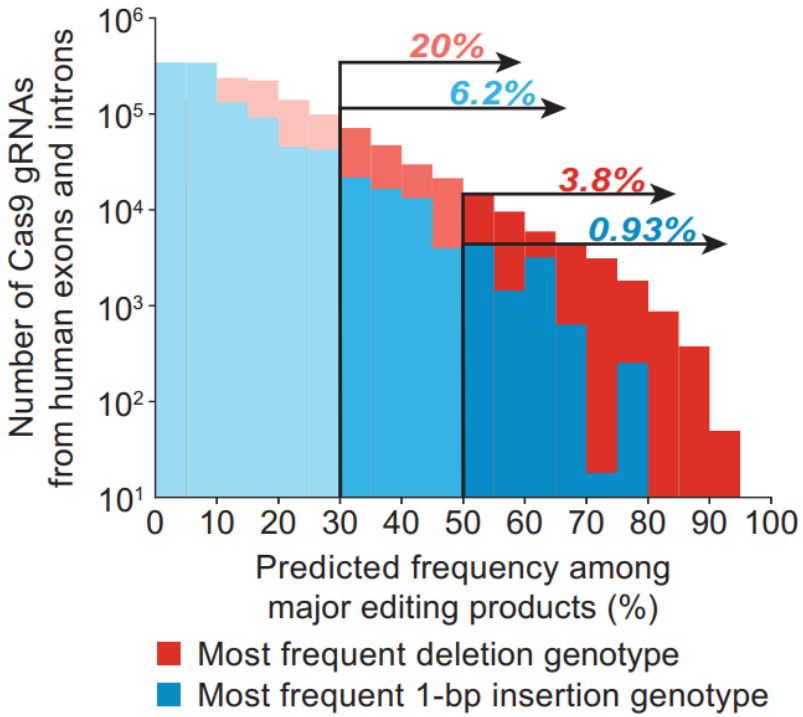
```
GCCCCCACAAAGCGGACCAGCCAGCNN NNTAGGGAATAAGGCCAACTTGCACCCC
AGTGCACGAGTGACGGGATAATGTGNN NNCAGGTAATTGATAGGCTTCCAACTTA
CCGAGACCTACGCGAACCGAAATCGNN NNCCGGTGCGCAAGCAGGGCTGGCGTCC
```

Local sequence context

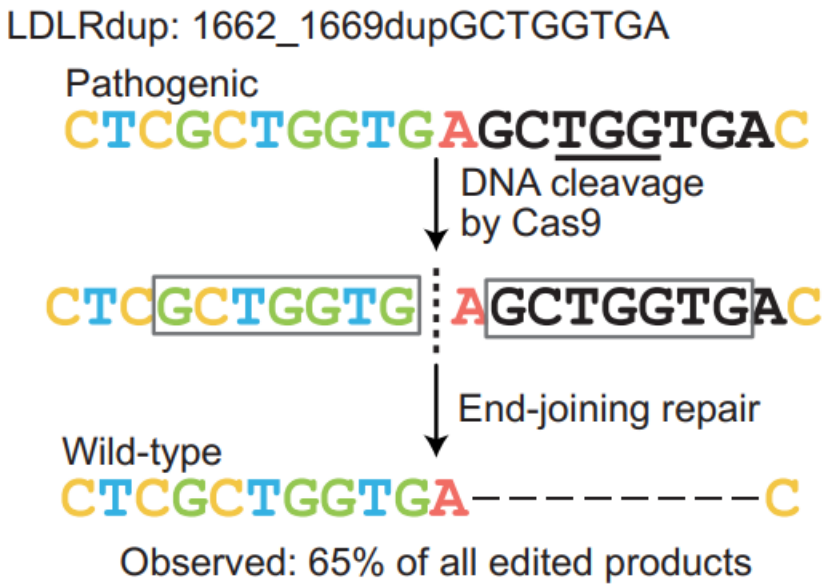


Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

inDelphi predicts that 5% of gRNAs yield a single repair genotype the majority of the time



Pathogenic microduplications are efficiently repairable to wild-type with simple Cas9 cutting



Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

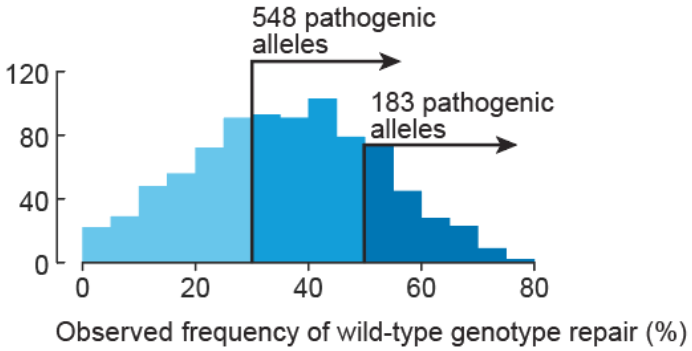
inDelphi identified 183 pathogenic alleles corrected to wild-type at >50% frequency (*r* = 0.64)

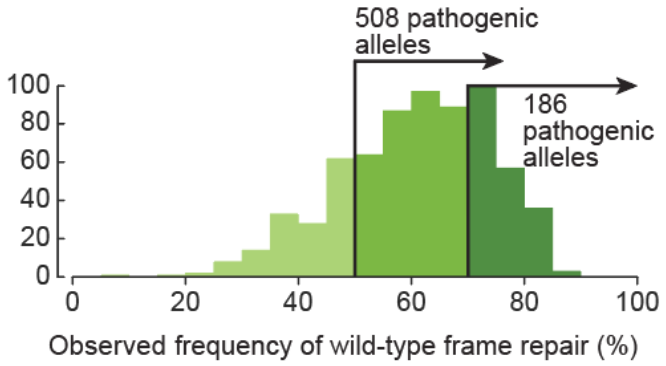
23,018
- Clinvar and HGMD

1,592
- Identified and designed with highest repair %

865
- Candidates for wild-type repair after quality filtering

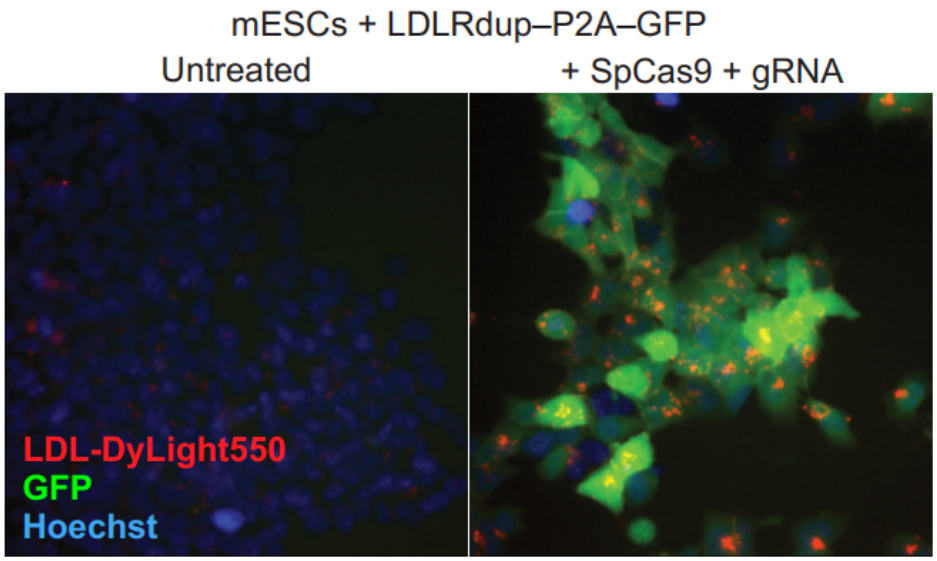
682
- Candidates for wild-type frame repair after quality filtering





Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

Efficient repair of pathogenic alleles to wild-type with template-free Cas9-nuclease treatment



- Primary patient-derived fibroblasts
  Human and mouse cell lines
- SpCas9 and SaCas9
- HPS1 71%
  LDLR 77%
  PORCN 48%
  GAA 68%
  GLB1 42%

Cas9 editing without a homology template is predictable, can be precise, and can be practical for disease

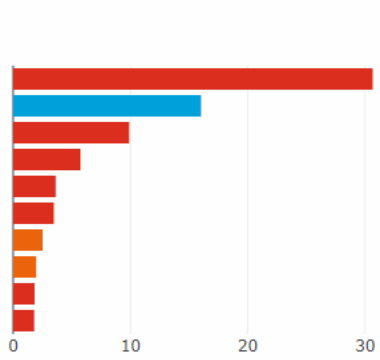
inDelphi

AATAGTTTCTAGCACAGCGCTGGTGTGC | GTCGTGTGGCTGAAGGCATAGTAATTCTGA

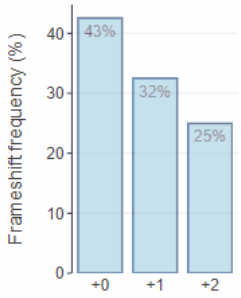
◄ DSB ►

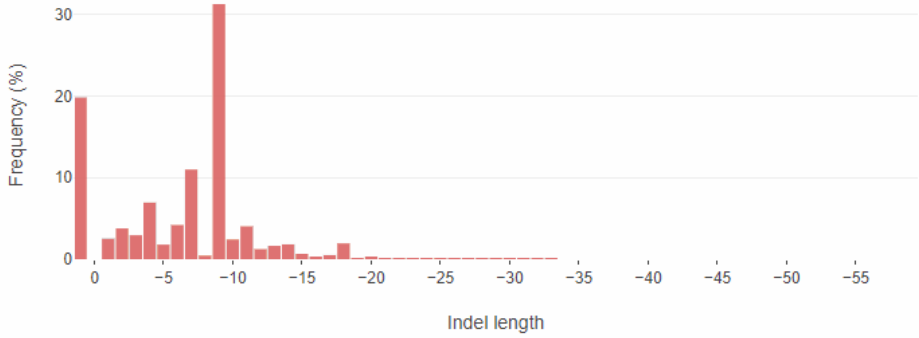
Summary of predictions: Top 10 frequent events

Alignment	Category	%
AATAGTTTCTAGCACAGCGCTGGTGTGCGTCGTGTGGCTGAAGGCATAGTAATTCTGA	Reference	-
AATAGTTTCTAGCACAGCGCTGGTGTG---------GCTGAAGGCATAGTAATTCTGA	9-bp deletion	30.6
AATAGTTTCTAGCACAGCGCTGGTGTGCCGTCGTGTGGCTGAAGGCATAGTAATTCTGA	1-bp insertion	16.0
AATAGTTTCTAGCACAGCGCTGGTGTG-------TGGCTGAAGGCATAGTAATTCTGA	7-bp deletion	9.9
AATAGTTTCTAGCACAGCGCTGGTGT----CGTGTGGCTGAAGGCATAGTAATTCTGA	4-bp deletion	5.7
AATAGTTTCTAGCACAGCGCTGGTG-----------GCTGAAGGCATAGTAATTCTGA	11-bp deletion	3.6
AATAGTTTCTAGCACAGCGCTGGT------CGTGTGGCTGAAGGCATAGTAATTCTGA	6-bp deletion	3.5
multiple deletion genotypes	1-bp deletion	2.5
multiple deletion genotypes	2-bp deletion	2.0
AATAGTTTCTAGCACAGCGCTG-----------------AAGGCATAGTAATTCTGA	18-bp deletion	1.8
AATAGTTTCTAGCACAGCGCTGGTGTG--TCGTGTGGCTGAAGGCATAGTAATTCTGA	2-bp deletion	1.8



Indel length predictions





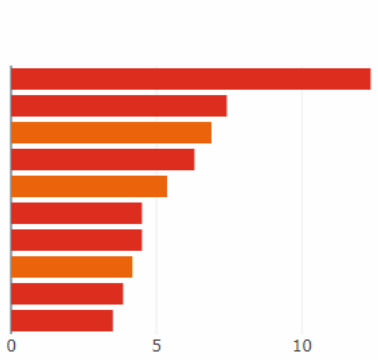
inDelphi

AATAGTTTCTAGCACAGCGCTG | ACGAGTGTGCGTCGTGTGGCTGAAGGCATA

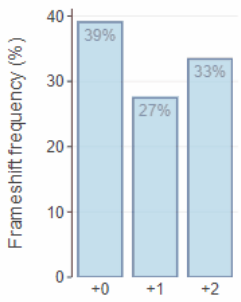
◄ DSB ►

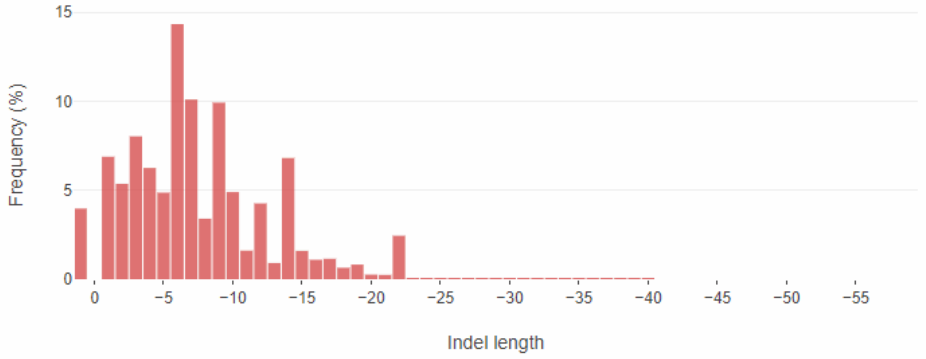
Summary of predictions: Top 10 frequent events

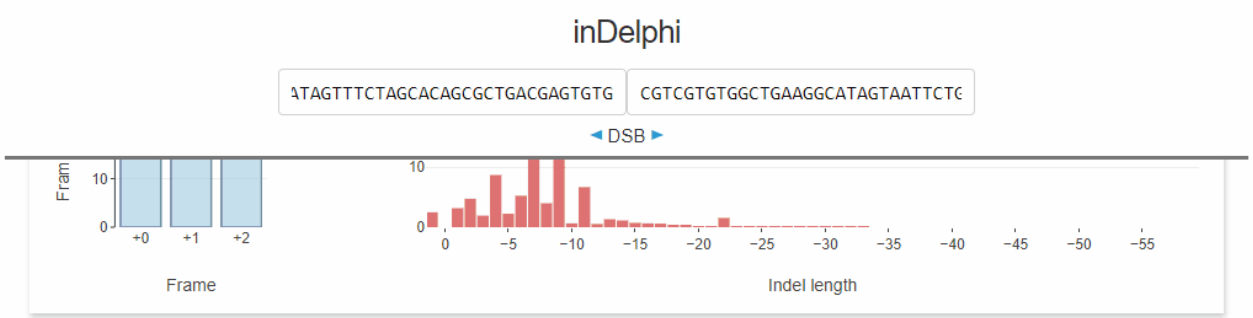
Alignment	Category	%
\TAGTTTCTAGCACAGCGCTGACGAGTGTGCGTCGTGTGGCTGAAGGCATAGTAATTCT(	Reference	-
\TAGTTTCTAGCACAGCG------AGTGTGCGTCGTGTGGCTGAAGGCATAGTAATTCT(	6-bp deletion	12.4
\TAGTTTCTAGCACAGCGCTG-------TGCGTCGTGTGGCTGAAGGCATAGTAATTCT(	7-bp deletion	7.4
multiple deletion genotypes	1-bp deletion	6.9
\TAGTTTCTAGCACAGCG--------------TCGTGTGGCTGAAGGCATAGTAATTCT(	14-bp deletion	6.3
multiple deletion genotypes	2-bp deletion	5.4
\TAGTTTCTAGCACAGCGCTG---------CGTCGTGTGGCTGAAGGCATAGTAATTCT(	9-bp deletion	4.5
\TAGTTTCTAGCAC---------GAGTGTGCGTCGTGTGGCTGAAGGCATAGTAATTCT(	9-bp deletion	4.5
multiple deletion genotypes	3-bp deletion	4.2
\TAGTTTCTAGCACAGCGCTG---AGTGTGCGTCGTGTGGCTGAAGGCATAGTAATTCT(	3-bp deletion	3.9
\TAGTTTCTAGCACAG----------TGTGCGTCGTGTGGCTGAAGGCATAGTAATTCT(	10-bp deletion	3.5



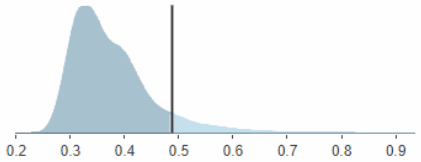
Indel length predictions



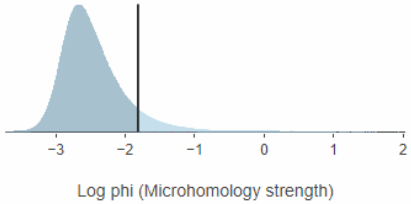




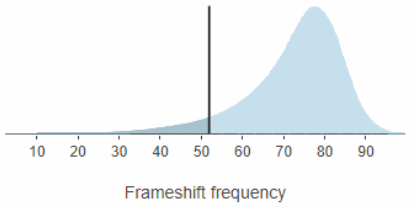
Comparison to predictions at 12,691,956 SpCas9 target sites in human exons and introns



This target site has high precision.
Precision score: 0.49
Percentile: 91.9



This target site has high microhomology strength.
Log phi: -1.81
Percentile: 91.8



This target site has low frameshift frequency.
Frameshift frequency: 51.9%
Percentile: 6.1

Acknowledgements

Max Shen

Mandana Arbab

David R. Liu

Richard I. Sherwood

Daniel Worstell

Olga Krabbe

Christopher A. Cassa

Jonathan Hsu