

Neural Network Architecture Search in Genomics by AMBER

Frank Zijun Zhang

Center for Computational Biology, Genomics
Flatiron Institute, Simons Foundation

Lewis-Sigler Institute for Integrative Genomics
Princeton University

June 16, 2022

Outline



Basics of Deep learning in Genomics and Neural Architecture Search (NAS)



Deep residual convolutional neural network for CRISPR/Cas9 outcomes and variant effects



Biophysics-interpretable modeling of CRISPR/Cas9 off-target effect

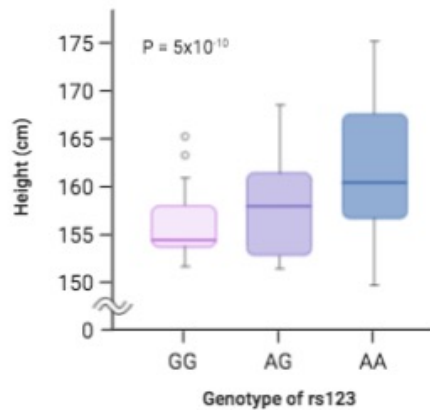
Genomics is Data-Driven

① Height and genetic data for individuals in study



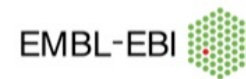
② Single-variant association test with a candidate variant, rs123

A alleles increase height on average*



* Liberties were taken with the size of the allelic effect shown in this example

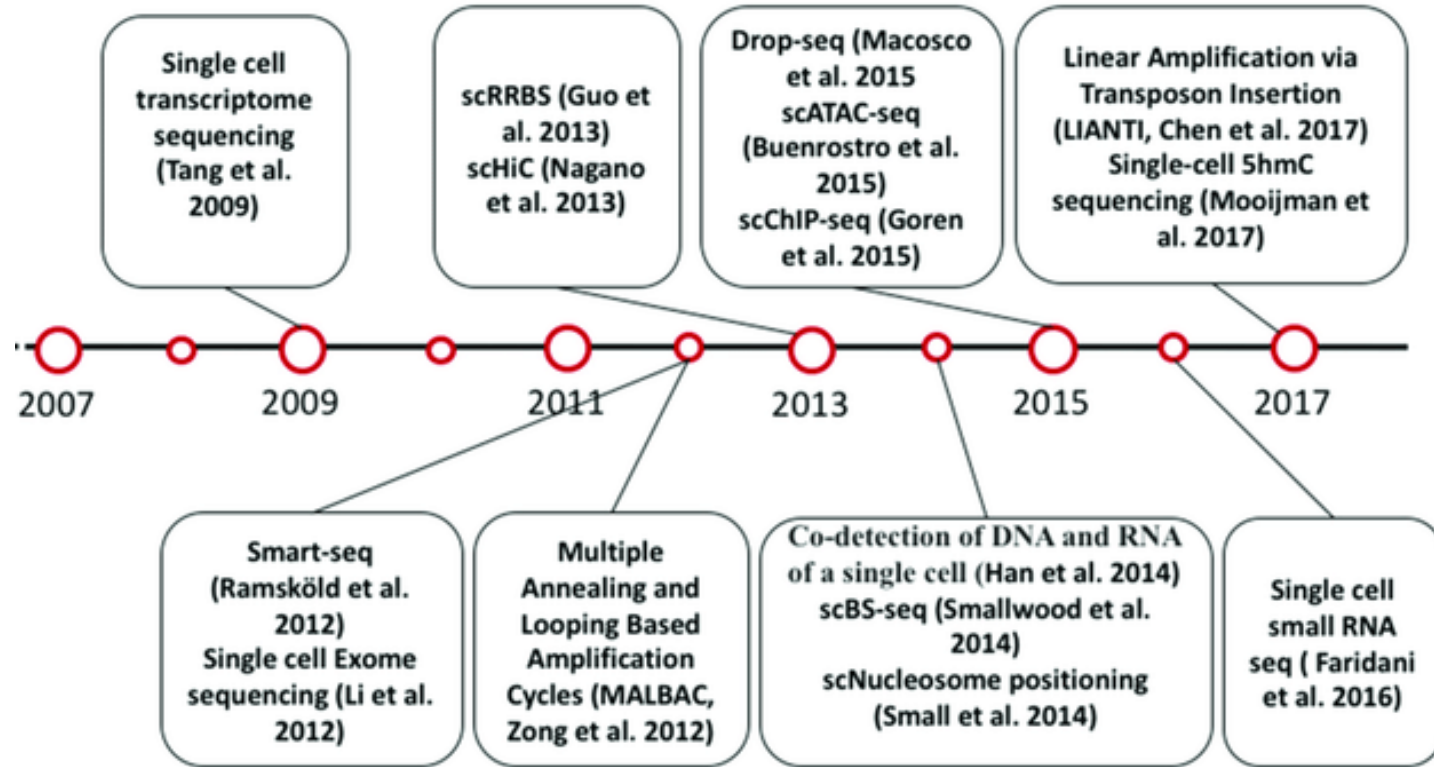
Published Genome-Wide Associations as of July 2019
 $p \leq 5 \times 10^{-8}$ for 17 trait categories



NHGRI-EBI GWAS Catalog
www.ebi.ac.uk/gwas

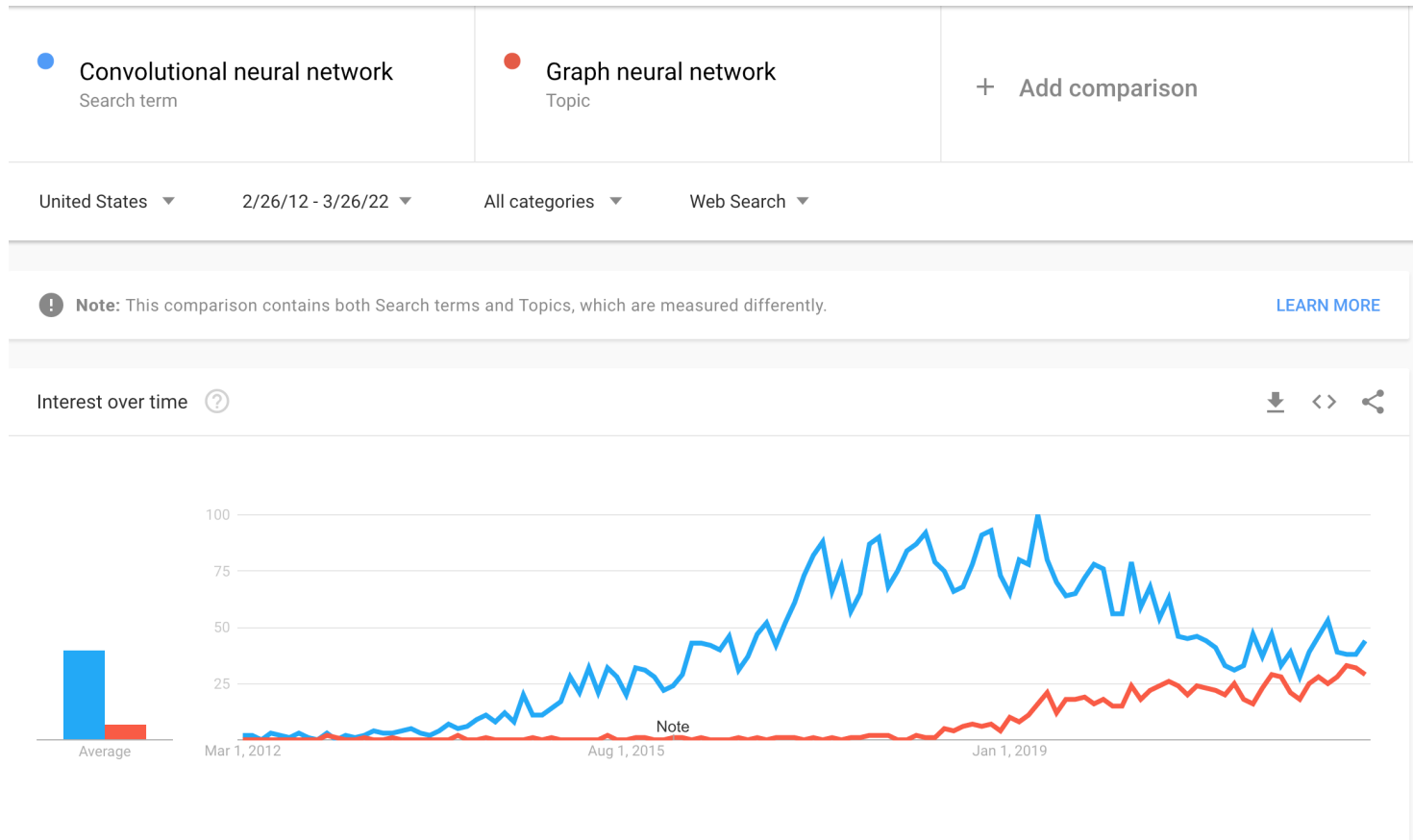
Genomics is Data-Driven

- Timeline of Single-cell sequencing milestones



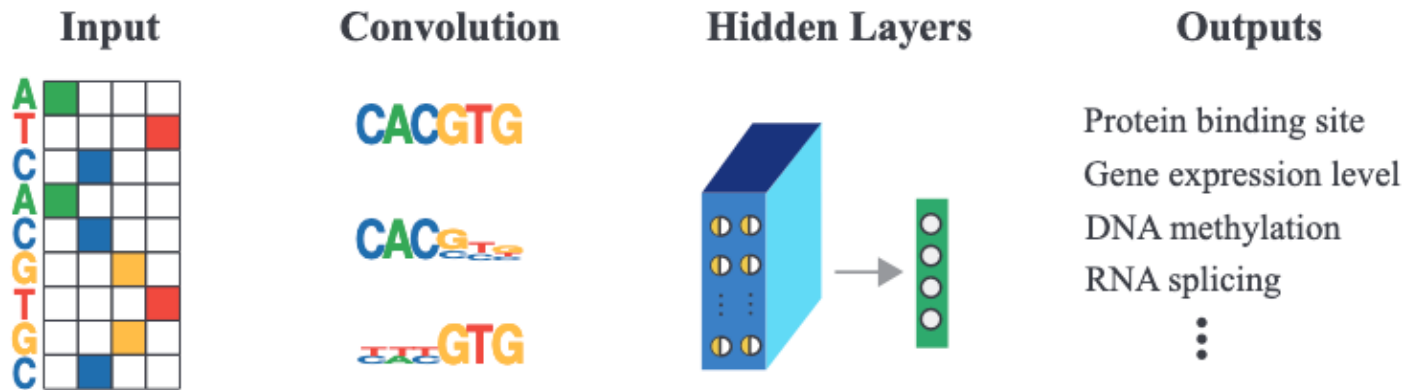
Evolution of modern deep learning methods

- CNN – popular in the last decade but plateaued
- GNN – starting to rise!



Type 1: sequence-to-molecule predictions

- General Framework: one-hot encoded sequence -> molecular

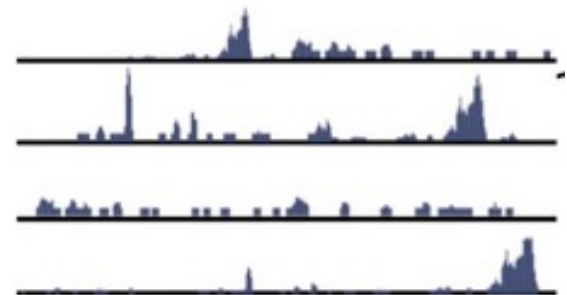


The classic Epigenetics multi-tasking model: DeepSEA (Zhou and Troyanskaya, 2015)

Input: 1000 base-pair (bp) DNA sequence

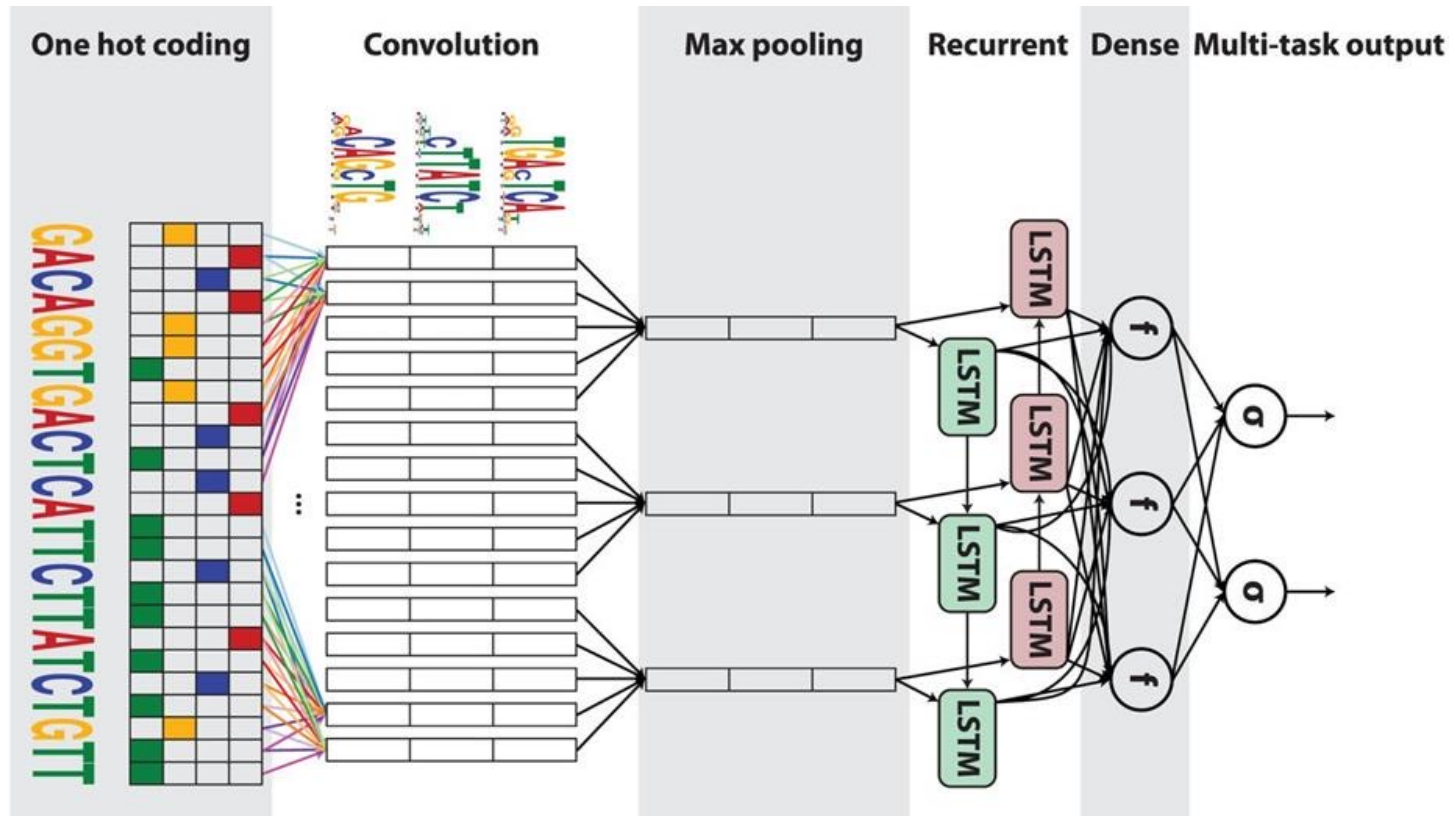
Output: multi-label classes of 919 biochemical markers

Genome-wide signals from DNA/RNA/protein sequences, for example TFs, DNase, histone



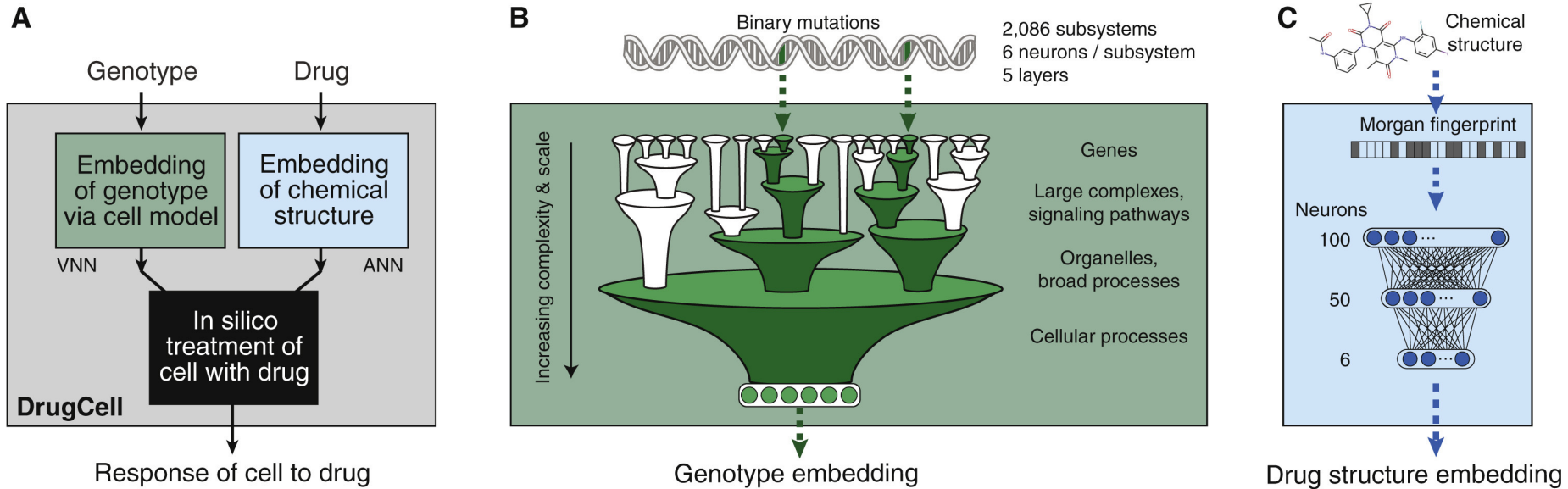
Type 1: sequence-to-molecule predictions

- A follow-up hybrid CNN-RNN for the same task; Quang and Xie, 2016



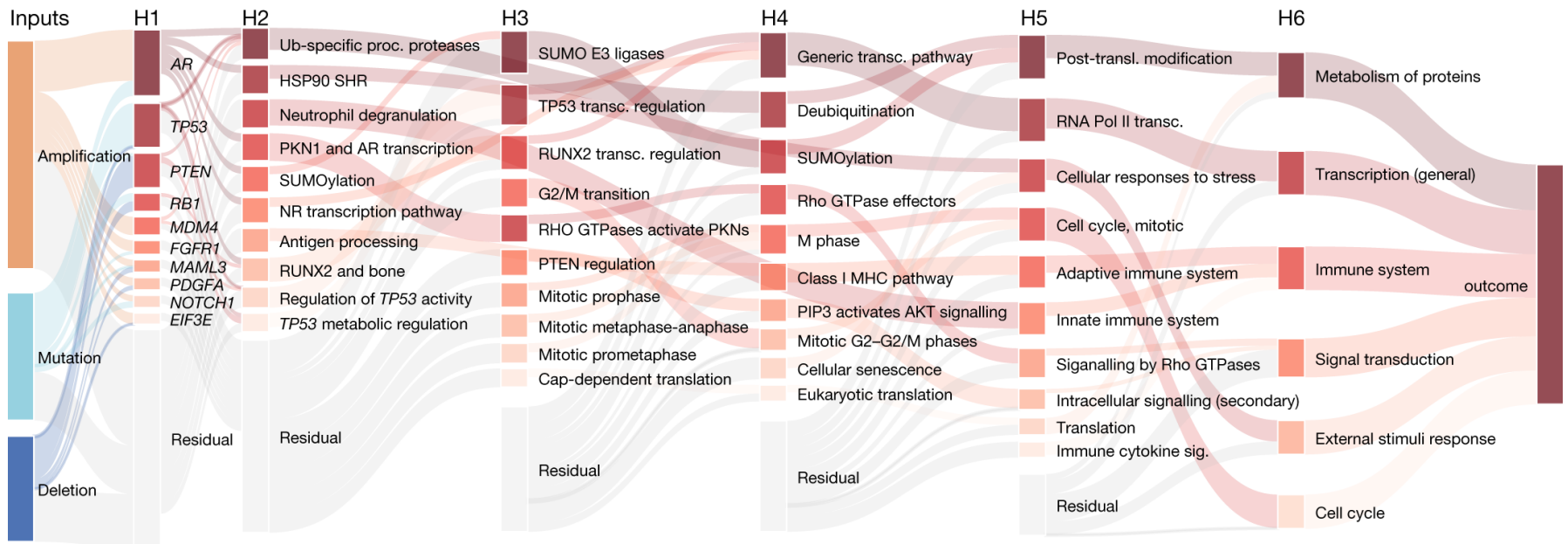
Type 2: molecule-to-phenotype predictions

- DrugCell: interpretable deep learning model of human cancer cells and drug interactions (Kuenzi and Park et al., 2020)

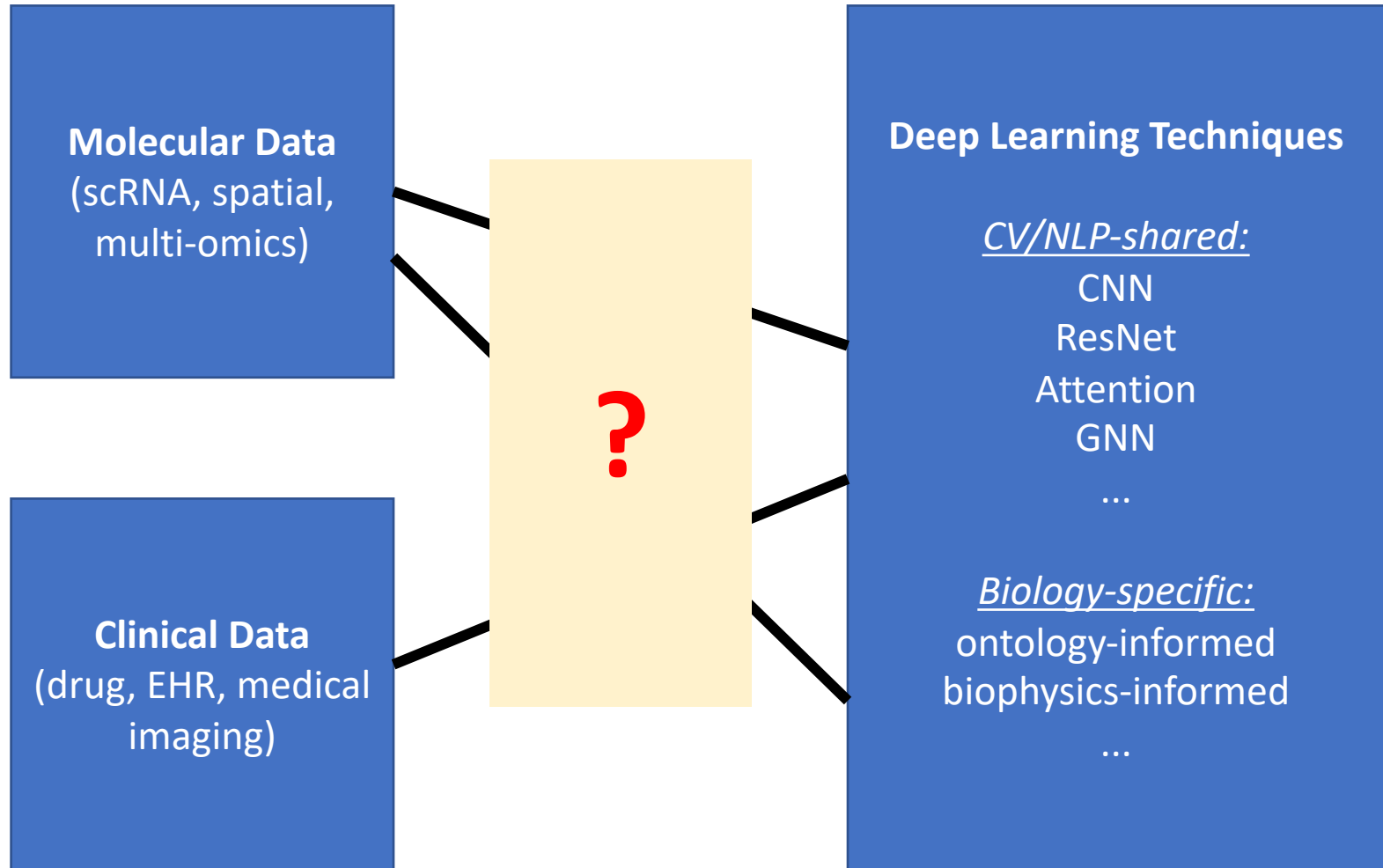


Type 2: molecule-to-phenotype predictions

- P-net: primary vs metastatic prostate cancer predictions from tumor mutations (Elmarakeby et al., 2021)



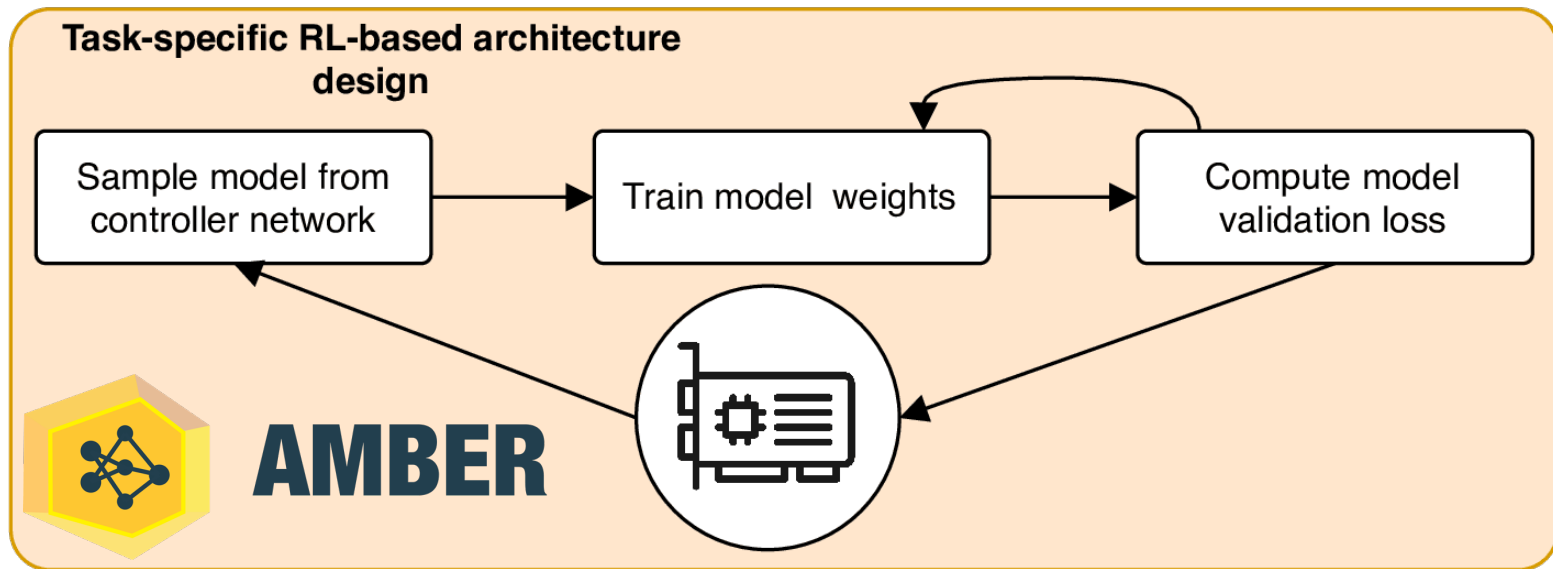
Challenges and Opportunities for AI in Medicine



AMBER Automates Deep Learning Deployment

Automated Modeling for Biological Evidence-based Research

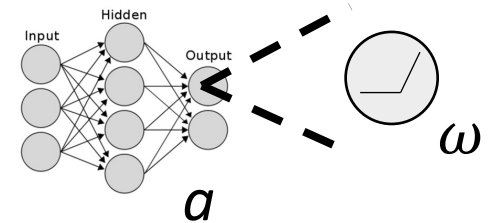
- The process of architecture tuning is automated by Reinforcement learning (RL).
- AMBER is efficient and data-driven, searching $>10^{30}$ models in 72 GPU hours.



Formulations of NAS Basics

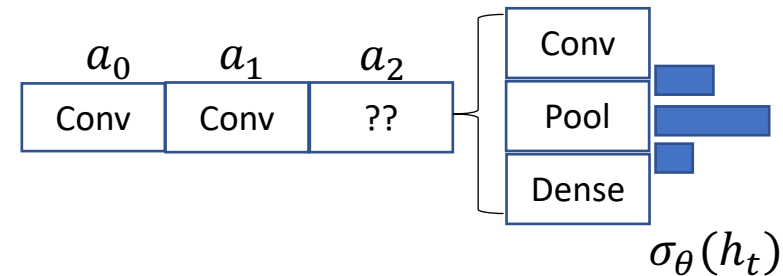
1. To learn a function that maps x to y , optimize its architectures a :

$$y_i = f_{\omega; a}(x_i)$$



2. Sample a_t from the conditional probability $P(a_t|a_{t-1}, \dots, a_0)$ by a Recurrent Neural Network $\sigma_\theta(\cdot)$ with parameters θ :

$$a_t \sim P(a_t|a_{t-1}, \dots, a_0) = \sigma_\theta(h_t)$$



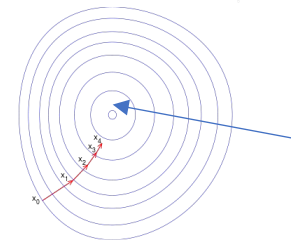
$\pi(a_k; \theta)$: log-likelihood of a_k

R_k : reward for a_k

b : moving average of R

3. Optimize θ w.r.t. to a reward R (usually validation accuracy):

$$\frac{1}{m} \sum_{k=1}^m \nabla_{\theta} \pi(a_k; \theta) (R_k - b)$$



Objective: high reward with large likelihood

AMBER-searched Model is Accurate and Parameter-efficient

- Applied AMBER to 919 epigenetics markers (i.e. DeepSEA task)
- AMBER searched architectures matched or exceed expert model

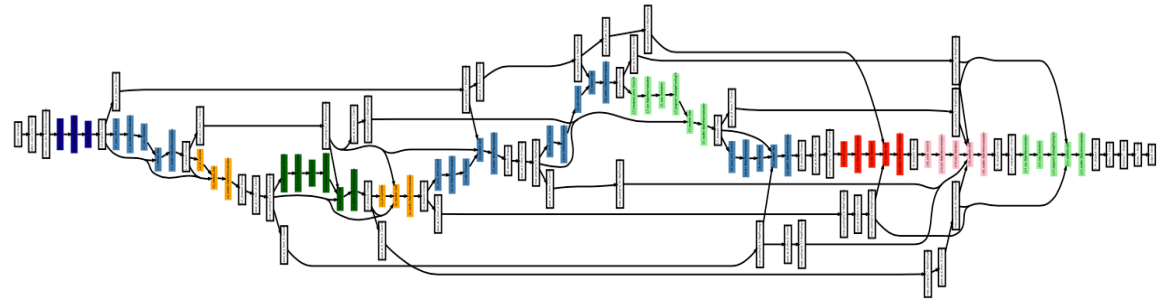
Color code for Operations

- conv8
- conv4
- dconv8
- dconv4
- maxpool
- avgpool
- identity

AMBER-Base

Total Parameters: 13.8M

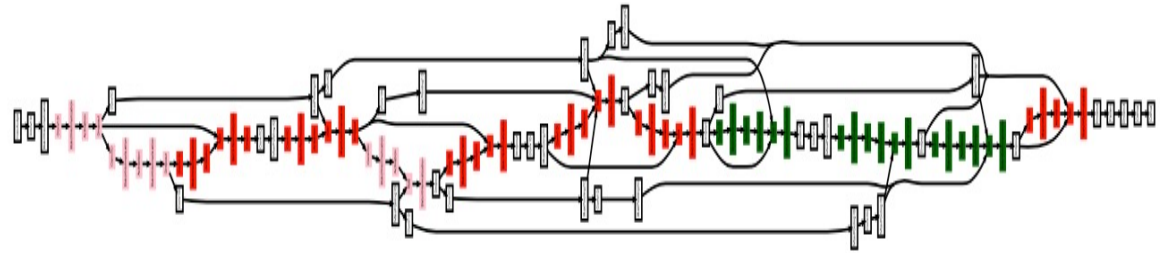
Avg. AUPRC = 0.337



AMBER-Seq

Total Parameters: 13.5M

Avg. AUPRC = 0.367

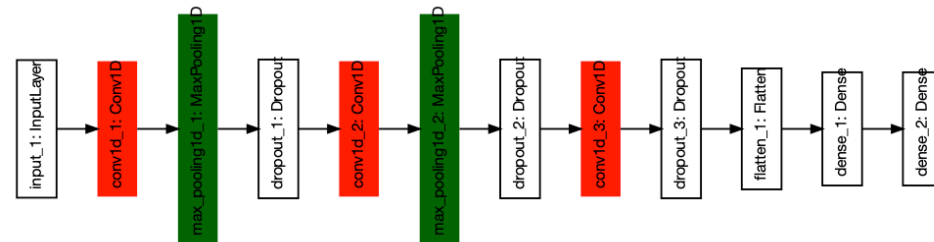


DeepSEA

(expert model)

Total Parameters: 52.8M

Avg. AUPRC = 0.338



AMBER: Publicly Available and Reusable across Biological Domains

<https://github.com/zj-zhang/AMBER>

Automated Modeling for Biological Evidence-based Research

AMBER is a toolkit for designing high-performance neural network models automatically in Genomics and Bioinformatics.

The overview, tutorials, API documentation can be found at: <https://amber-automl.readthedocs.io/en/latest/>

To get quick started, use this Google Colab notebook.  [Open in Colab](#)

- Predicting 919 epigenetics regulatory markers
 - 1000 bp sequence → 919 binary epigenetic markers

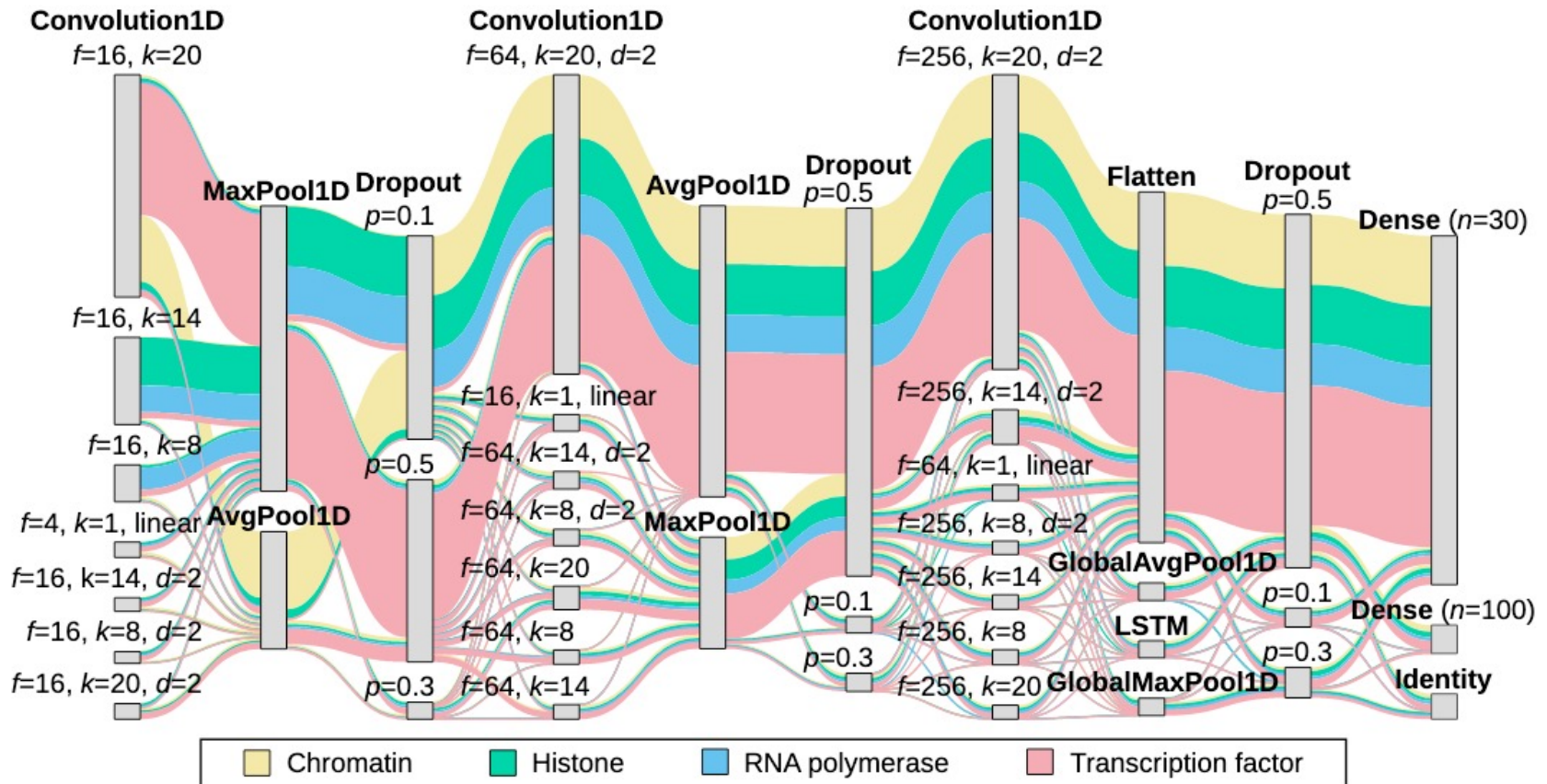


- Predicting 6 genome editing outcomes induced by CRISPR/Cas9
 - 60 bp sequence → probabilities of 6 editing outcomes



AMBIENT: towards data-specific, training-free NAS

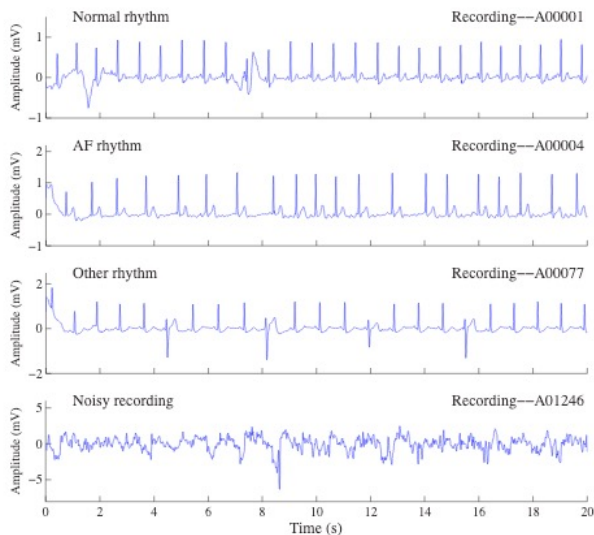
- Datasets from different biology factors use different neural network architectures!



AMBER Benchmarked on Electrocardiograms (ECG)

- NAS-bench-360: Tu *et al.*, 2021
- Input: 9 to 60-second ECG recordings sampled at 300 Hz
- Output: four classes, normal, disease, other, or noisy rhythms

Figure 1. Examples of the ECG waveforms.



Model Space	Algorithm	ECG	DeepSEA
WRN	default	0.57±0.01	0.60±0.001
DenseNAS	random	0.58±0.01	0.60±0.001
DenseNAS	original	0.60±0.01	0.60±0.001
WRN	ASHA	0.57±0.01	0.59±0.002
DARTS	GAEA	0.66±0.01	0.64±0.02
AMBER	ENAS	0.67±0.015	0.68±0.01

AMBER is Easy to Use

https://github.com/rtu715/NAS-Bench-360/blob/main/AMBER/examples/amber_ecg.py

[30 lines for Model Space Setup – from Example Script]

```
# Next, define the specifics
wd = "./outputs/AmberECG/"
X_train, Y_train, X_val, Y_val = read_data_physionet_4_with_val('.')
Y_train = to_categorical(Y_train, num_classes=4)
Y_val = to_categorical(Y_val, num_classes=4)
train_data = (X_train, Y_train)
val_data = (X_val, Y_val)
input_node = Operation('input', shape=(1000, 1), name="input")
output_node = Operation('dense', units=4, activation='sigmoid')
```

Replaced by
pickled
configs
since v0.1.2

[70 lines for Run Configuration – from Example Script]

```
# finally, run program
amb = Amber(types=type_dict, specs=specs)
amb.run()
```

Outline



Basics of Deep learning in Genomics and Neural Architecture Search (NAS)



Deep residual convolutional neural network for CRISPR/Cas9 outcomes and variant effects



Biophysics-interpretable modeling of CRISPR/Cas9 off-target effect

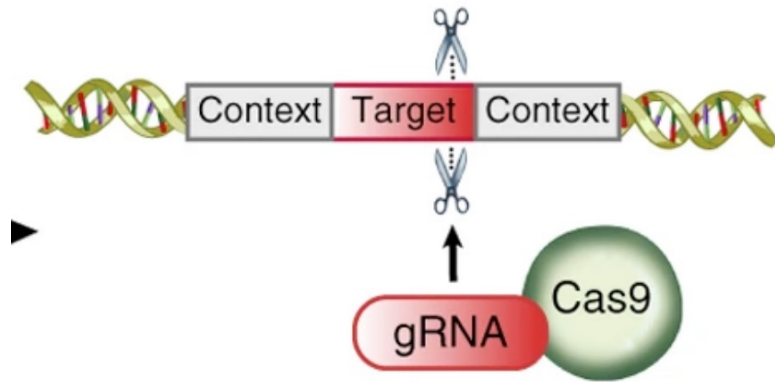
Victoria Li



Hunter College High School

Predictable CRISPR/Cas9 Editing Outcomes




Cas9 cuts target and generates mutations

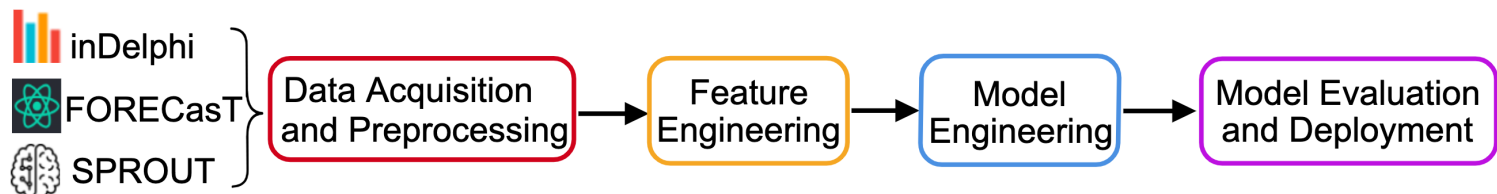


Cas9 Editing Outcomes

	CAGGCTTGGCTGCAAGAGCATCGGCCTGAAAGC	AGTGAGGAGGCAGCGGCCCTGGTGGTAGACTTG ACC	
I1	CAGGCTTGGCTGCAAGAGCATCGGCCTGAAAGC	AAGTGAGGAGGCAGCGGCCCTGGTGGTAGACTT GAC	57.1%
D3	CAGGCTTGGCTGCAAGAGCATCGGCCTGAAAG	TGAGGAGGCAGCGGCCCTGGTGGTAGACTTGAC C	4.0%
D25	CAGGCTTGGCTGCAAGAGCATCGGCC	CTGGTGGTAGACTTGACC	1.6%
D9	CAGGCTTGGCTGCAAGAGCATCGGCCTGA	GGAGGCAGCGGCCCTGGTGGTAGACTTGACC	1.6%
D25	CAGGCTTGGCTGCAAGAGCA	GCGGCCCTGGTGGTAGACTTGACC	1.6%

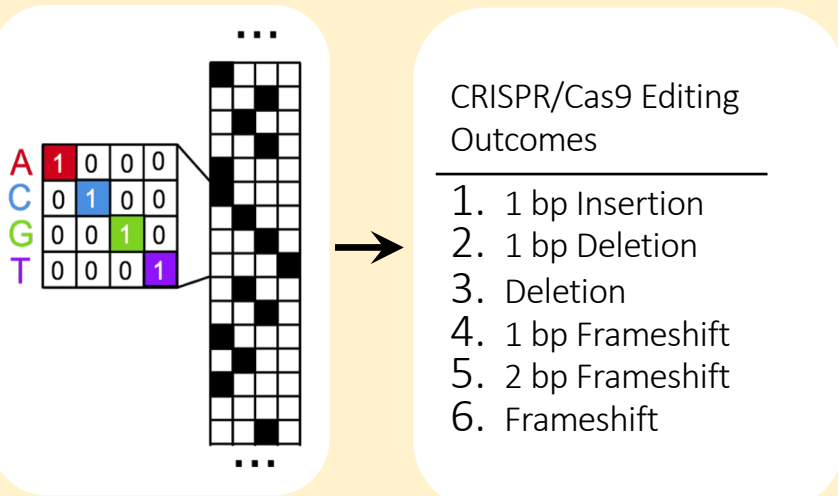
Existing CRISPR/Cas9 editing outcome predictors are reliant on feature and model engineering

	 (1) inDelphi (Shen et al. 2018)	 (2) FORECasT (Allen et al. 2019)	 (3) SPROUT (Leenay et al. 2019)
Number of gRNAs	2,000	~40,000	1,656
Cell Line	mESC, HCT116, HEK293, K562, U2OS	Cas9-expressing K562 (Artificial)	Primary T cells
Method	Neural Networks and k-nearest neighbors	Multinomial Logistic Regression	Gradient-boosting Decision Trees

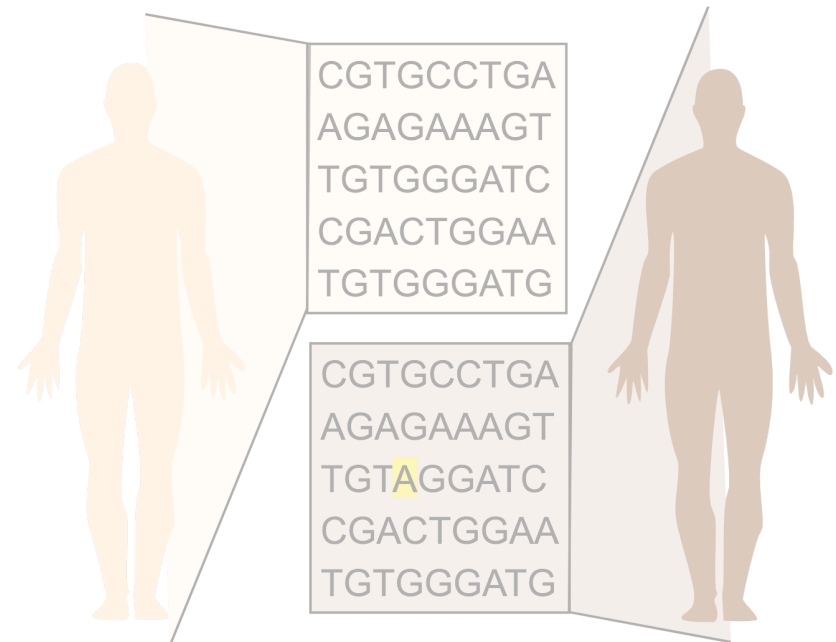


Objective: generating an automated and variant-aware CRISPR/Cas9 outcome predictor

(1) Create CROTON

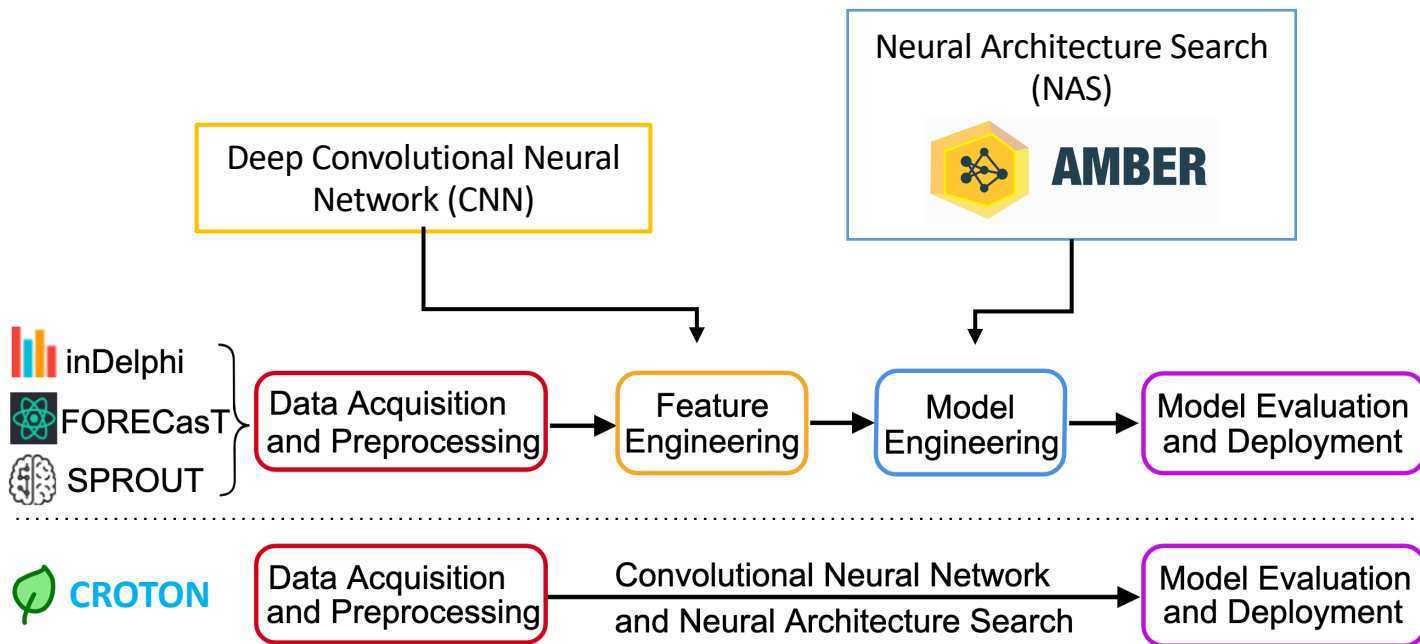


(2) Genomic Variants in Patients



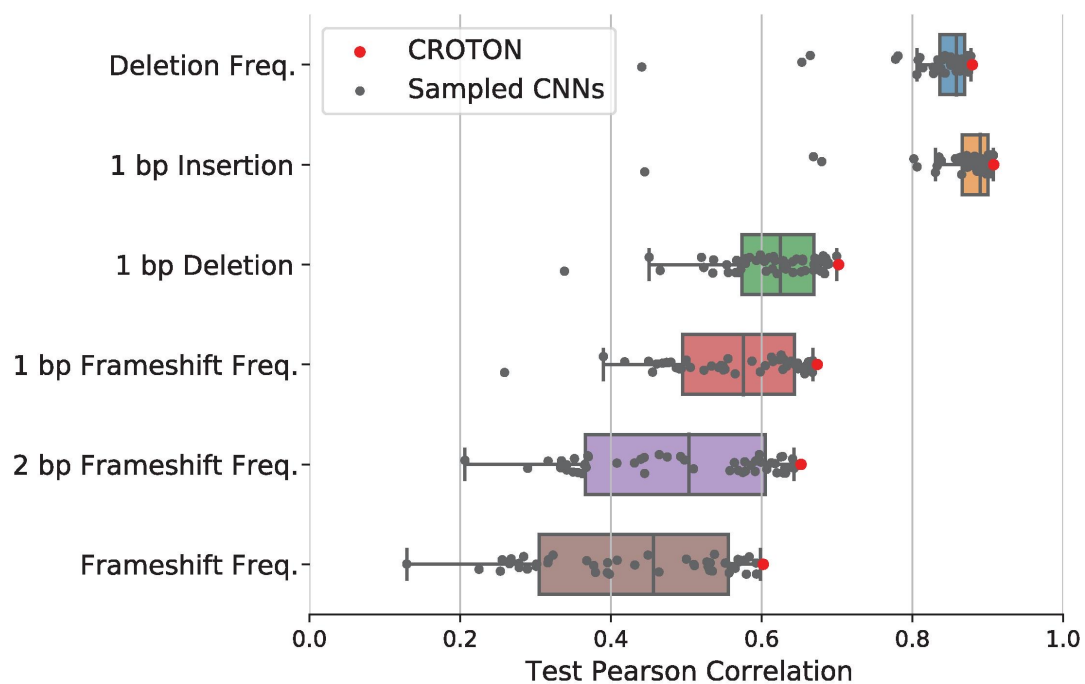
The CROTON ML pipeline is highly automated

- **CROTON**: **CR**ISPR **O**utcomes **T**hrough **cON**volutional neural networks

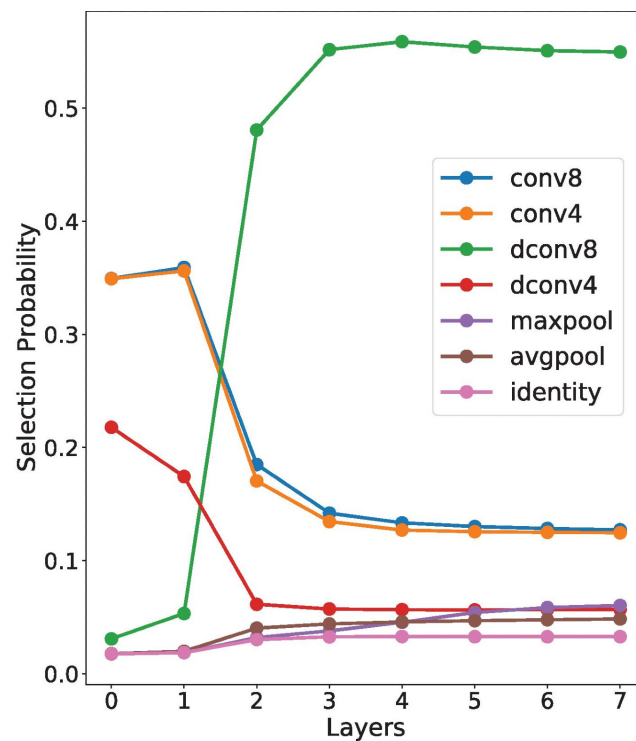


NAS designs effective multi-task deep CNN architectures

- Sample architectures from the model search space

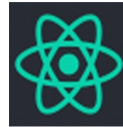


- Layer selection probabilities



CROTON Outperforms Existing Models

- Trained on synthetic sequences in K562, tested on endogenous genomic sequences in primary human T cells.



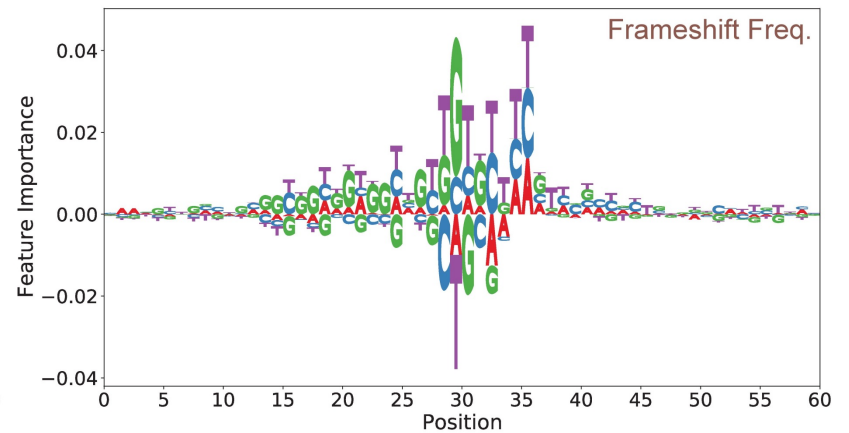
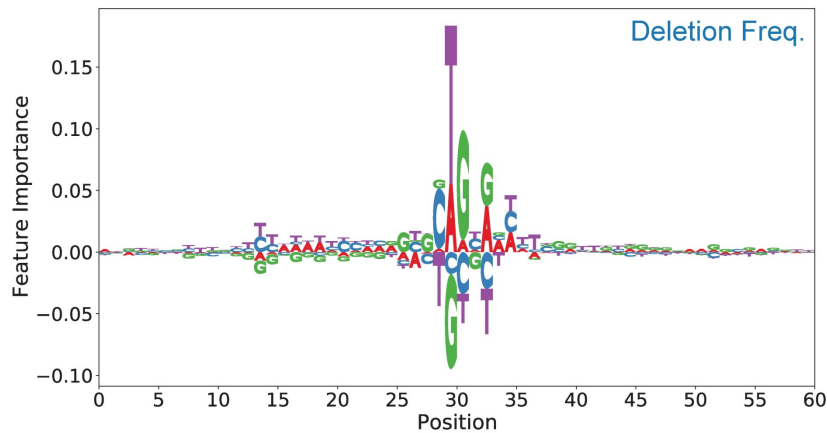
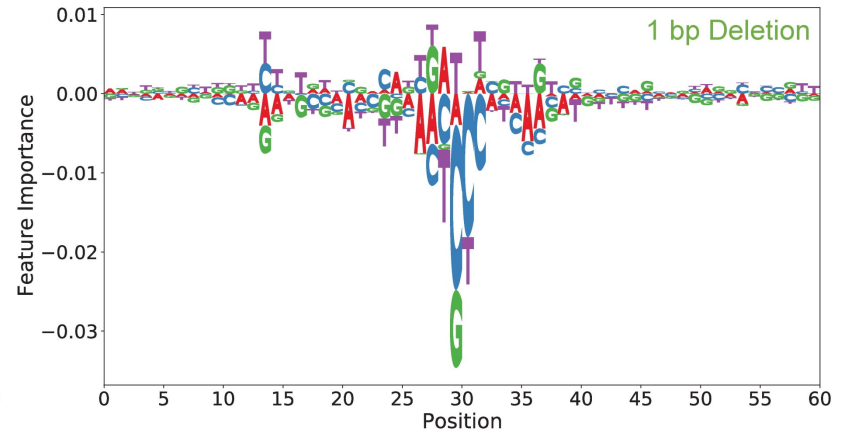
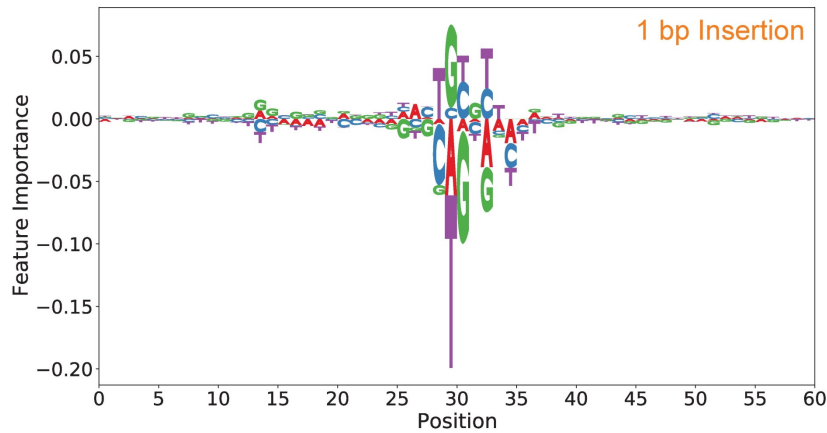
	CROTON	inDelphi	FORECasT
Deletion*	81.12	51.00	73.17
1 bp Insertion*	82.42	52.40	75.10
1 bp Deletion*	57.51	21.45	30.36
1 bp Frameshift*	73.84	54.69	66.71
2 bp Frameshift*	64.30	42.40	50.04
Frameshift*	55.56	51.54	57.94

	CROTON	SPROUT
Deletion*	81.12	77
1 bp Insertion**	65.22	62
1 bp Deletion**	43.81	40

*Pearson's Correlation, **Kendall's Tau

(Since testing was conducted on SPROUT (T cell) data, CROTON was compared to SPROUT's published metrics)

Nucleotides upstream of the PAM sequence are important to CRISPR/Cas9 editing outcomes



CROTON is publicly-available

- github.com/vli31/CROTON

CROTON

Please input a 60-nucleotide target sequence:
(ex. TCCAGGGCCTAATCTGACCGTCCTAGATACCTCAGGGTGG
GCAATACGAGGTAATGGCAG)

Your sequence...

I

Predict

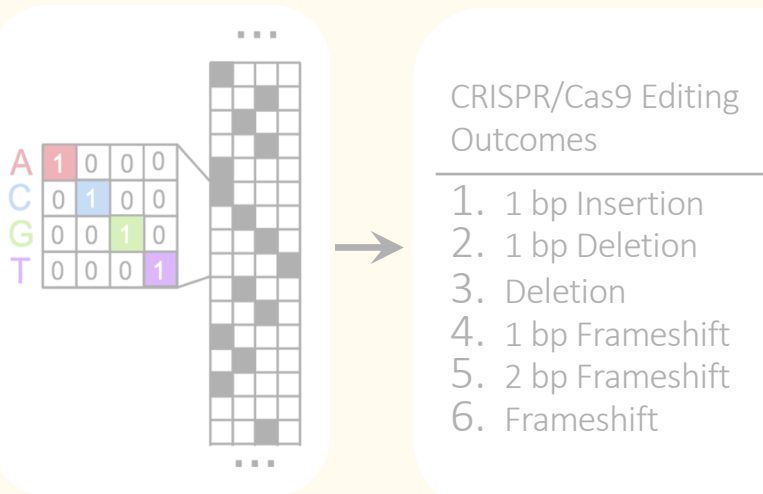
Input:

Output:

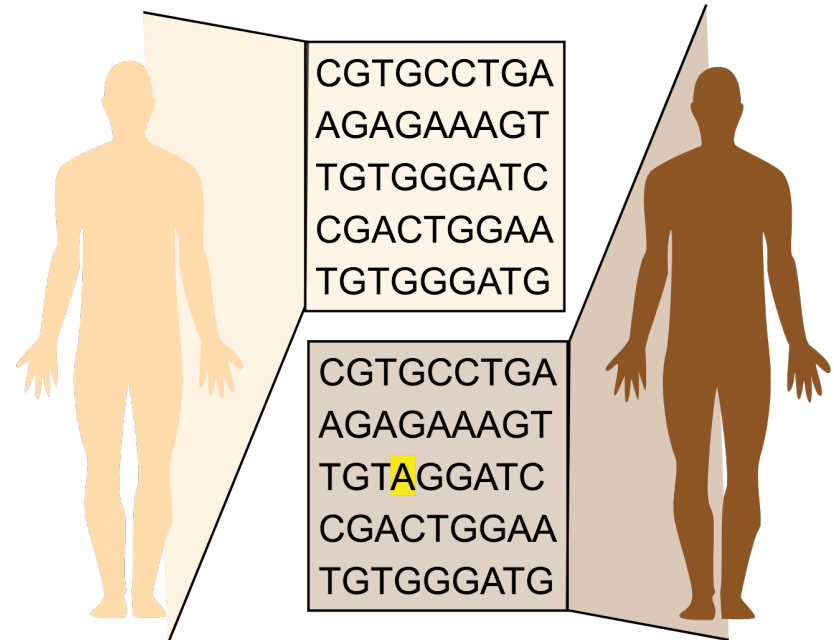
- 1 bp Insertion Probability:
- 1 bp Deletion Probability:
- Deletion Frequency:
- 1 bp Frameshift Frequency:
- 2 bp Frameshift Frequency:
- Frameshift Frequency:

Objective: generating an automated and variant-aware CRISPR/Cas9 outcome predictor

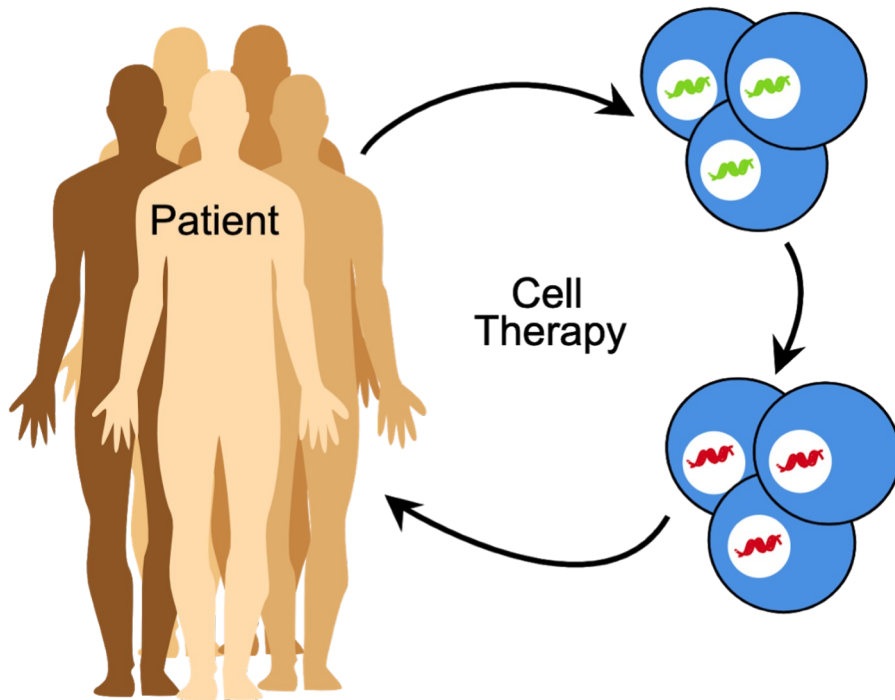
(1) Create CROTON



(2) Genomic Variants in Patients

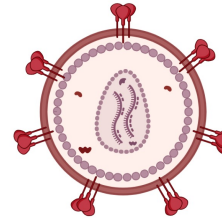


CRISPR/Cas9 is used to inactivate genes in clinical trials

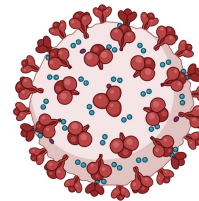


Antiviral Therapy

(1) *CCR5* and HIV

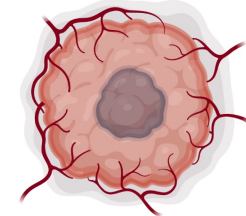


(2) *ACE2* and SARS-Cov-2



Cancer Immunotherapy

(3) *PDCD1*, *CTLA4* and Cancer

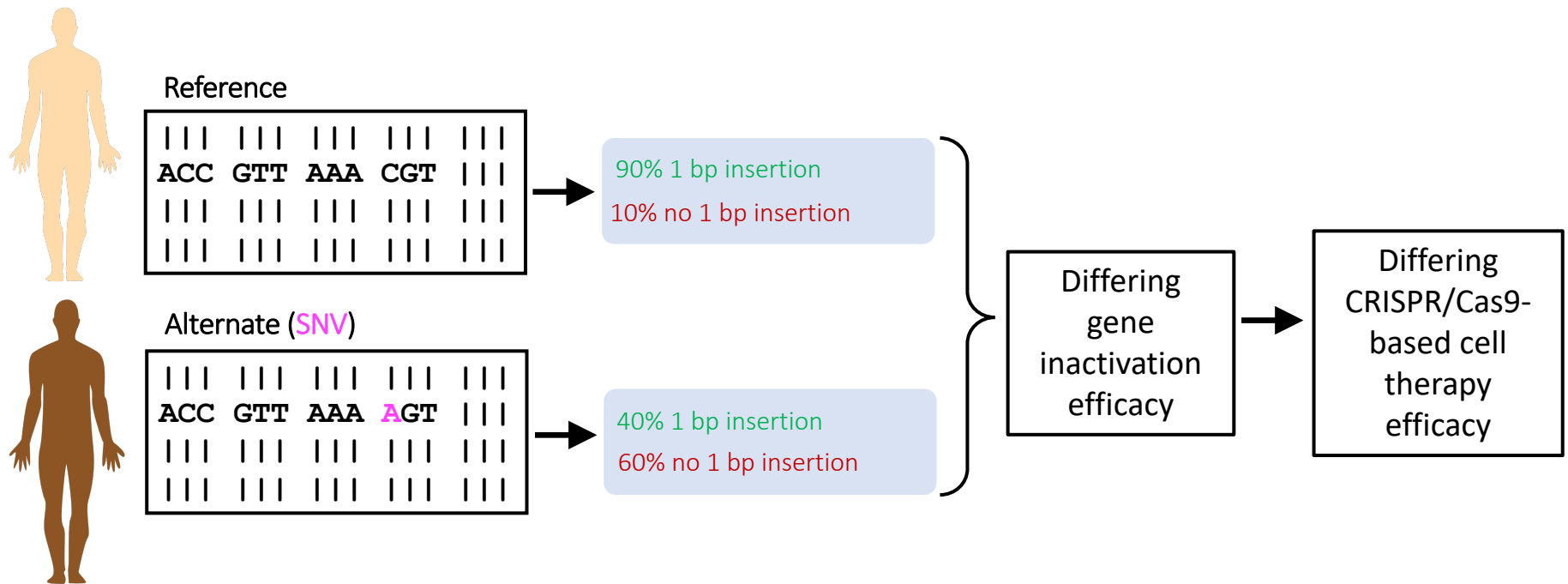


- *PDCD1*: 1st CRISPR/Cas9-based clinical trial to clear safety concerns

○ Xue et al. 2020

Single nucleotide variants can substantially impact CRISPR/Cas9 editing outcomes

- There are ~10-15 million common human SNVs, which can impact CRISPR/Cas9 editing outcomes (Eichler et al., 2007)



Single nucleotide variants substantially impact CRISPR/Cas9 editing outcomes

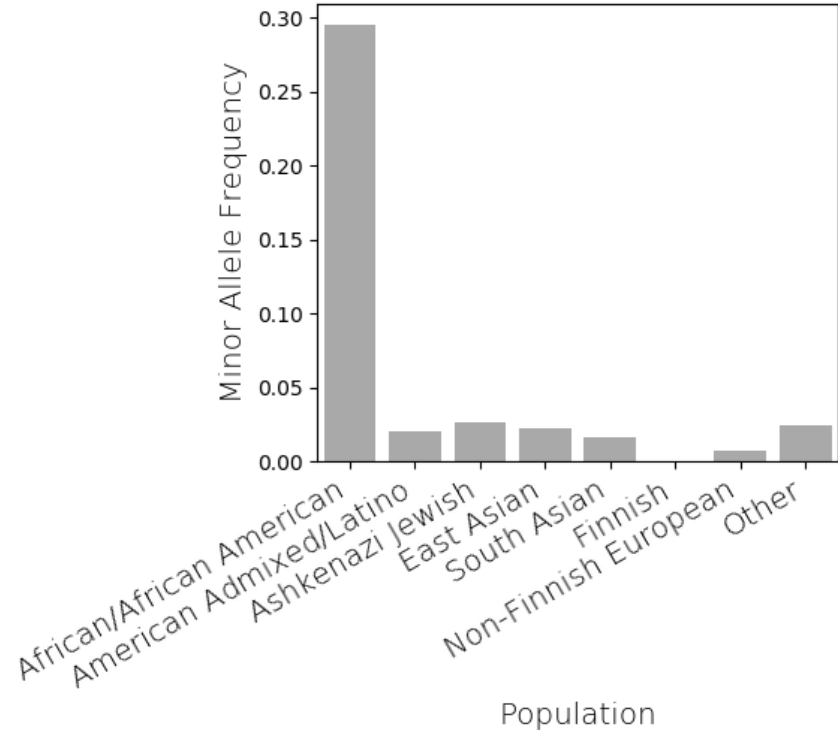
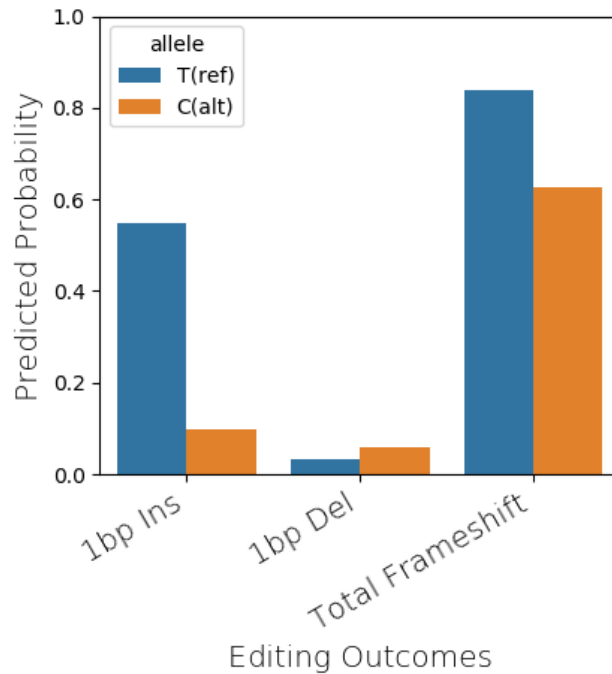
- SNVs with a high impact on 1 bp insertion prediction

Gene	Variant	Reference Pred.	Alternate Pred.	Absolute Difference
<i>PDCD1</i>	rs1284638279	0.576	0.110	0.466
<i>ACE2</i>	rs1482922566	0.656	0.222	0.434
<i>ACE2</i>	rs370610075	0.056	0.489	0.432
<i>PDCD1</i>	rs535799968	0.029	0.429	0.399
<i>PDCD1</i>	rs141119263	0.202	0.601	0.398
<i>PDCD1</i>	rs769685838	0.130	0.524	0.394
<i>PDCD1</i>	rs371902970	0.132	0.515	0.382
<i>PDCD1</i>	rs370660750	0.116	0.497	0.381
<i>PDCD1</i>	rs1021665035	0.110	0.475	0.365
<i>PDCD1</i>	rs1185044781	0.399	0.036	0.363
<i>CCR5</i>	rs1032906612	0.060	0.422	0.362
<i>CCR5</i>	rs139737901	0.190	0.552	0.362
<i>CCR5</i>	rs767205045	0.546	0.186	0.360

CROTON Identifies Cas9-altering Genetic Variants

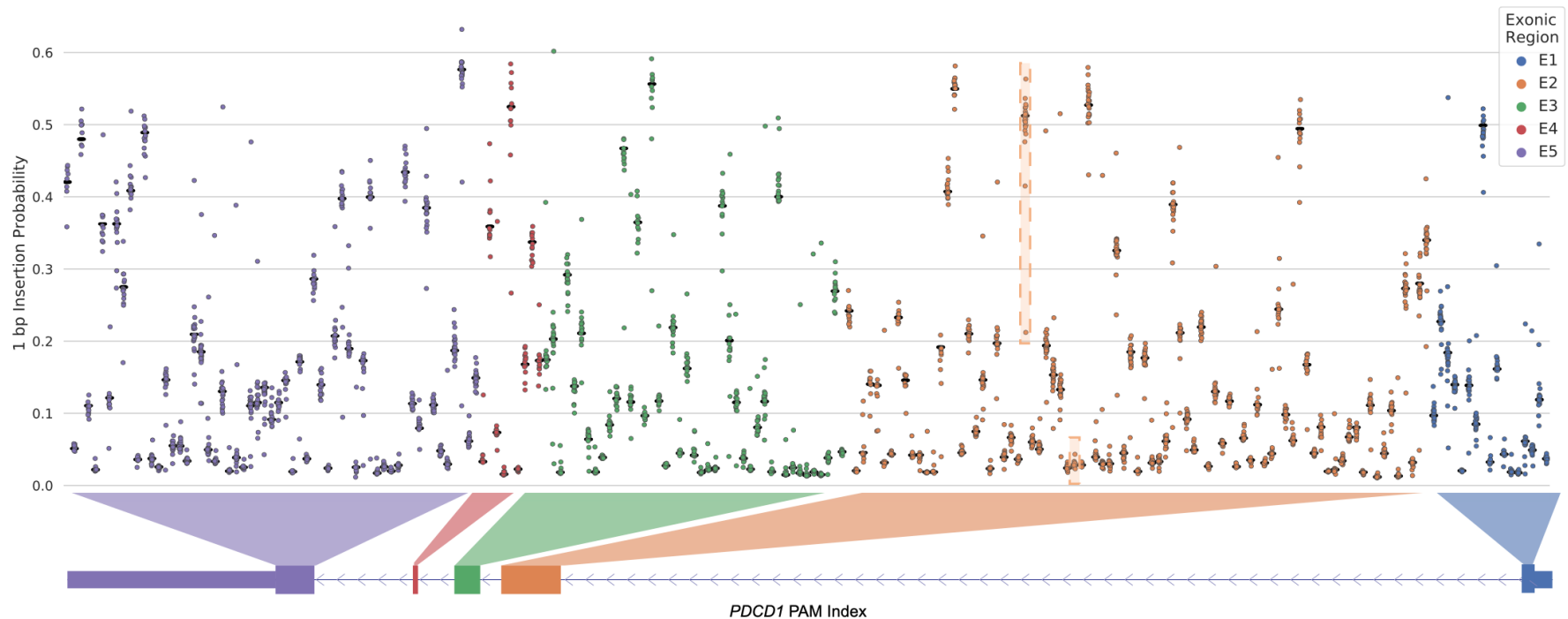
- Inheritable, population-stratified genetic variants can substantially influence Cas9 editing outcomes.

FGFR3
rs2305181, PAM id 147



Variant Effect Analysis for gRNAs in Clinical Trials

- *PDCD1* is knocked-out in non-small cell lung carcinoma (ClinicalTrials.gov NCT02793856).
- Each column is a PAM; each dot is a variant.



CROTONdb: variant effect prediction database for CRISPR/Cas9 editing outcomes

<https://croton.princeton.edu>

Enter gene entrez or gene symbol to see the predictions

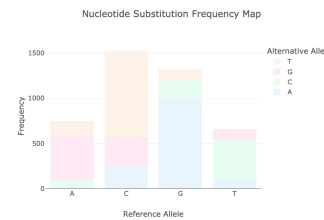
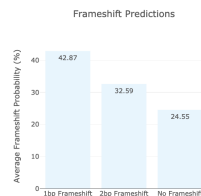
Submit Clear

[Example Gene](#)

The query interval has **Typical Frameshift**.
 Predicted: **74.98%**
 Percentile over genome: **53.71%**

Variant load for this query is **Substantially Higher**.
 Observed variants: **4242**
 Expected variants: **3939**

The query interval has **Typical Variant Impact to Frameshift**.
 Predicted: **1.27%**
 Percentile over genome: **65.48%**



gRNA published previously:

Cell

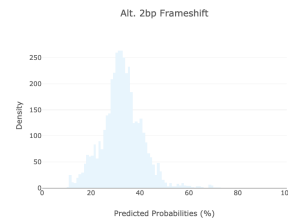
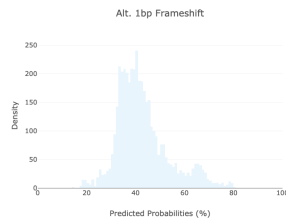
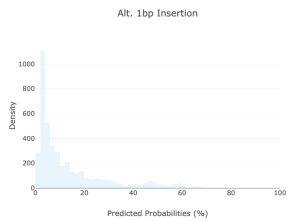
Article

Genetic Inactivation of CD33 in Hematopoietic Stem Cells to Enable CAR T Cell Immunotherapy for Acute Myeloid Leukemia

Kim et al., 2018, *Cell* 173, 1439–1453

May 31, 2018 © 2018 Elsevier Inc.

<https://doi.org/10.1016/j.cell.2018.05.013>



Variant ID	Variant Position	Ref. Allele	Alt. Allele	PAM ID	PAM Range	Max Variant Effect (%)	Ref. 1bp Insertion (%)	Ref. 1bp Frameshift (%)	Ref. 2bp Frameshift (%)	Alt. 1bp Insertion (%)	Alt. 1bp Frameshift (%)	Alt. 2bp Frameshift (%)
rs1339188502	51225275	A	G	CD33 17	+ : 51225245 - 51225305	45.8	51.5	64.3	20.6	5.7	34.1	43.1
rs770795199	51225276	G	A	CD33 17	+ : 51225245 - 51225305	12.8	51.5	64.3	20.6	38.7	54.4	29.0
rs201510739	51225270	G	A	CD33 17	+ : 51225245 - 51225305	8.7	51.5	64.3	20.6	60.1	72.3	16.0

CROTON-db:

5.38 million gRNA targets

90.82 million estimated variant effects

CROTON-db, *In preparation*

Summary of CROTON

- CROTON is a fully automated, publicly-available deep learning predictor for CRISPR/Cas9 editing outcomes.
- CROTON achieves SOTA performance and outperforms existing models manually tuned by experts.
- We use CROTON to identify that SNVs can substantially affect genome editing outcomes.
- These effects are systematically documented and analyzed in CROTONdb, facilitating safer and more effective CRISPR/Cas0-based cell therapies.

Outline



Basics of Deep learning in Genomics and Neural Architecture Search (NAS)



Deep residual convolutional neural network for CRISPR/Cas9 outcomes and variant effects



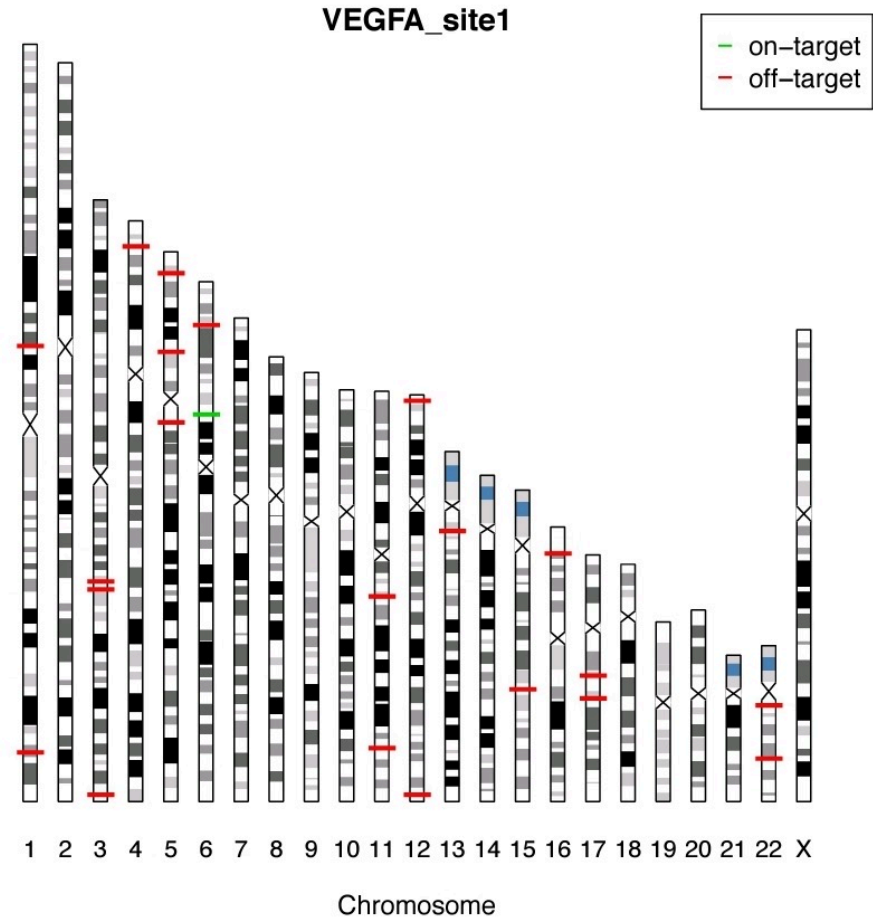
Biophysics-interpretable modeling of CRISPR/Cas9 off-target effect



Adam Lamson
Flatiron Institute, Simons Foundation

CRISPR/Cas9 Off Target effects

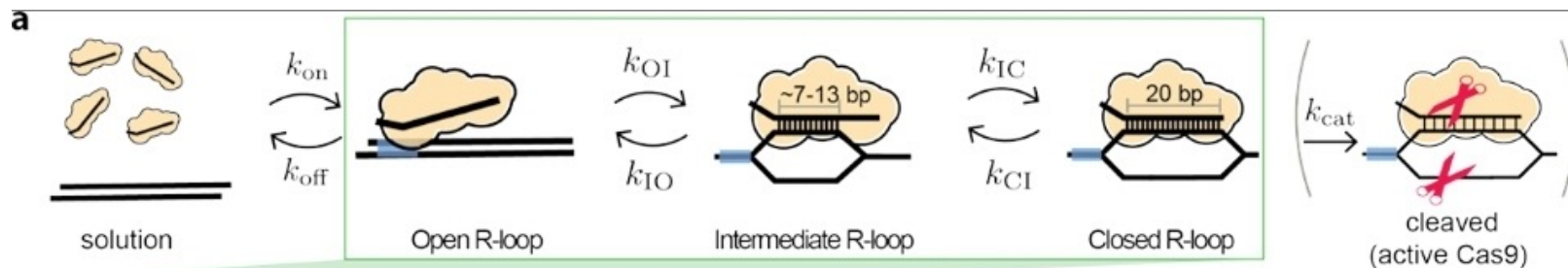
- Off Target: unintended cleavage at genomic sites w/ a similar but not an identical sequence
- Therapeutic uses need minimize the risk of deleterious outcomes
 - Even low frequency off-target can be dangerous! (clonal expansion)



Chromosome ideogram of CRISPR-Cas9 on/off-target sites for VEGFA.
Tsai et al., 2015, *Nat. Biotech*

Deciphering Cas9 Kinetics

- Existing off-target data and predictors can't profile kinetics rate directly.
 - uses hi-seq read counts as surrogates
 - can't differentiate enzyme-intrinsic kinetic parameters from-
 - exposure time
 - genetic context
 - cell cycle phase
 - DNA break repair pathway
- How many states are valid during the binding and cleavage process?
- Which of the transition is the slowest/fastest?

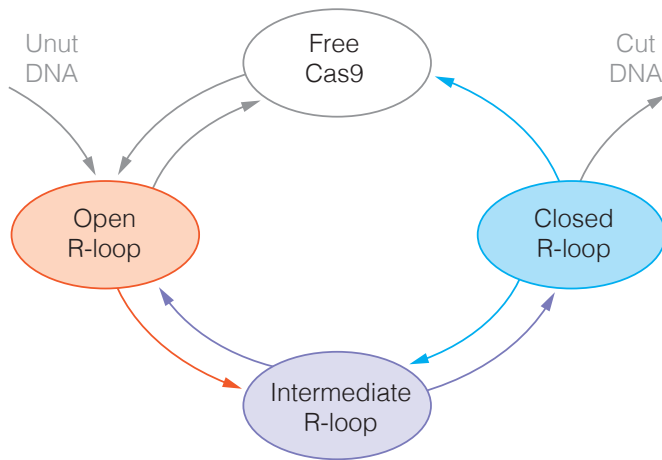


Reaction rate modeled by Kinetics Informed Neural Network (KINN)

Master equation for first-order kinetics

$$\frac{\partial S_\alpha}{\partial t} = \sum_{\beta} k_{\beta\alpha} S_\beta - S_\alpha \sum_{\beta} k_{\alpha\beta}$$

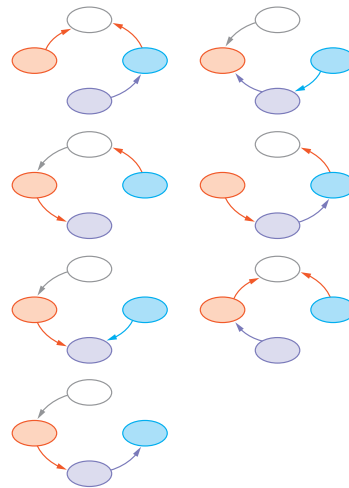
Kinetic model diagram for Cas9 cleavage



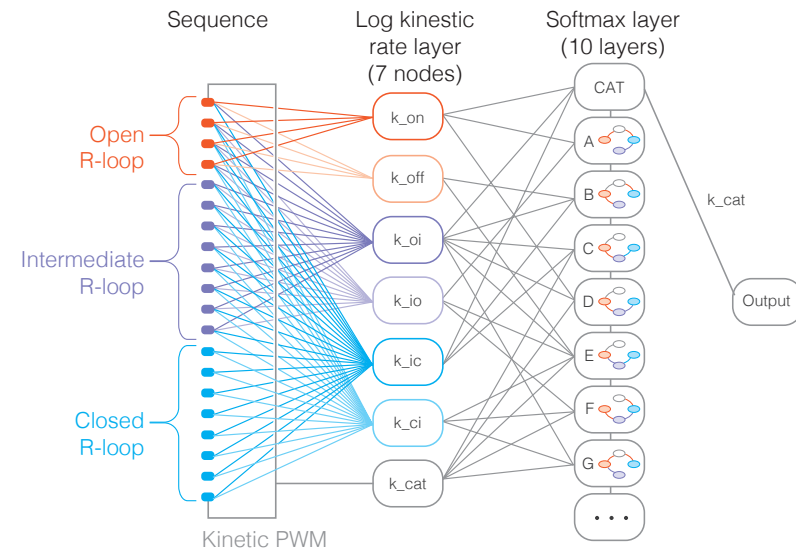
King-Altman (KA) diagrammatic method for steady state

$$(s_\infty)_\alpha = s_\alpha = \frac{\sum_{\ell \rightarrow \alpha} \kappa_\ell}{\sum_{\ell'} \kappa_{\ell'}}$$

$$\text{KA terms: } \kappa_\ell = \prod_{ij \in \ell} k_{ij}$$



$$\nu \propto \sum_{\alpha \beta \in \text{act}} \tilde{k}_{\alpha\beta} s_\alpha,$$



Build ODEs by Searching KINN

- Kinetic rates = $f(\text{seq})$

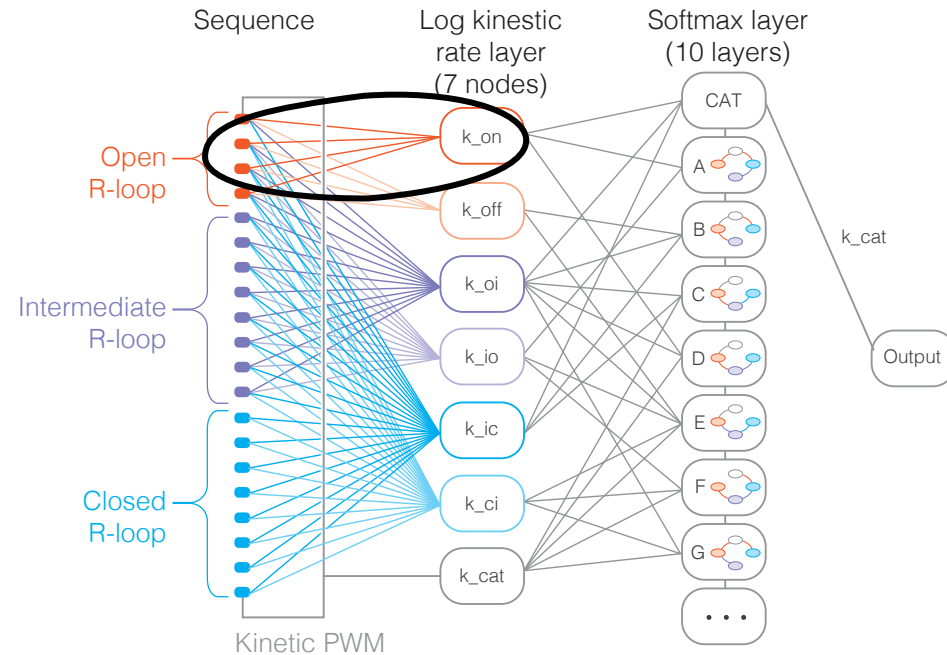
f is parameterized by convolution neural nets.

$$\log(k_{\alpha\beta}) = f(x_{i:j})$$

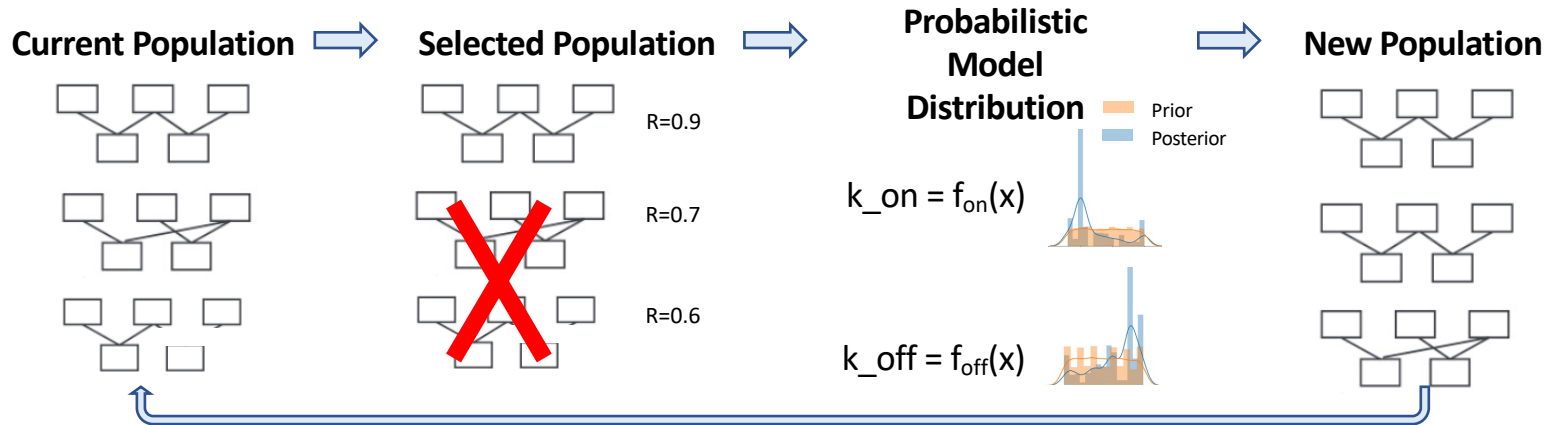
$f \in \{\text{CNNs}\}$

- range of sequence determinants for each rate

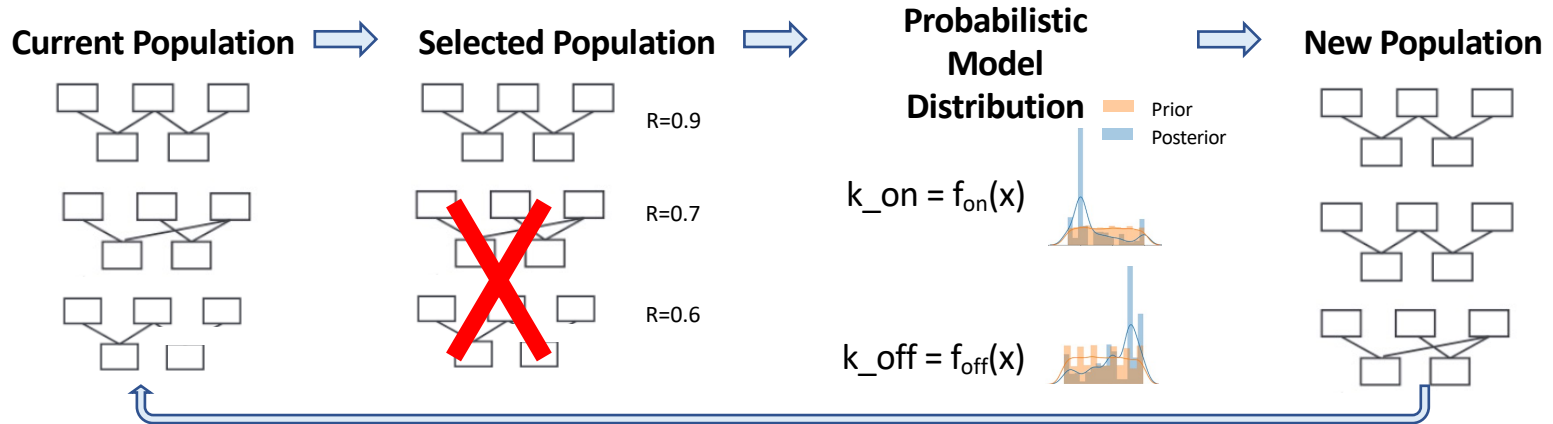
- e.g., **$k_3 = f(\text{seq}[10\text{bp}, 20\text{bp}])$**
on the right: k_3 is determined by the 10th -20th nt input seq



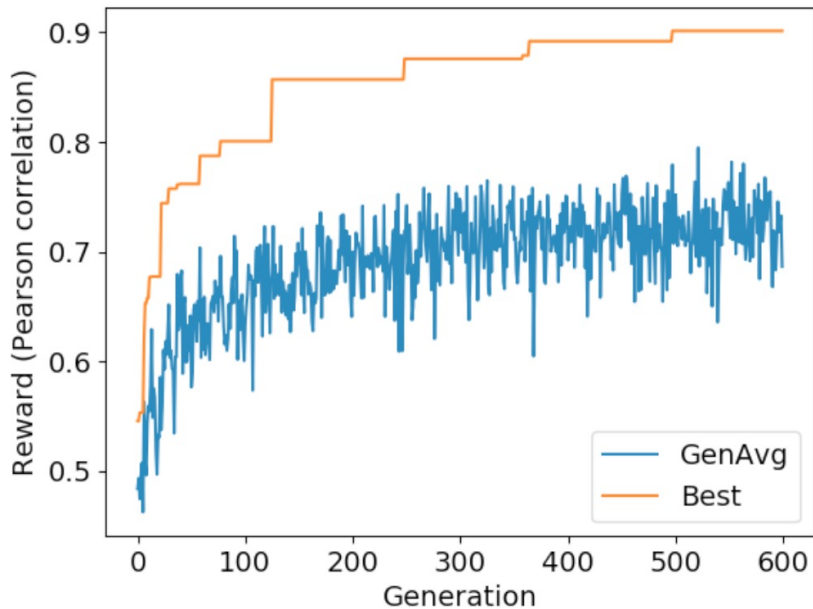
AMBER searches for KINN architectures by a probabilistic genetic algorithm



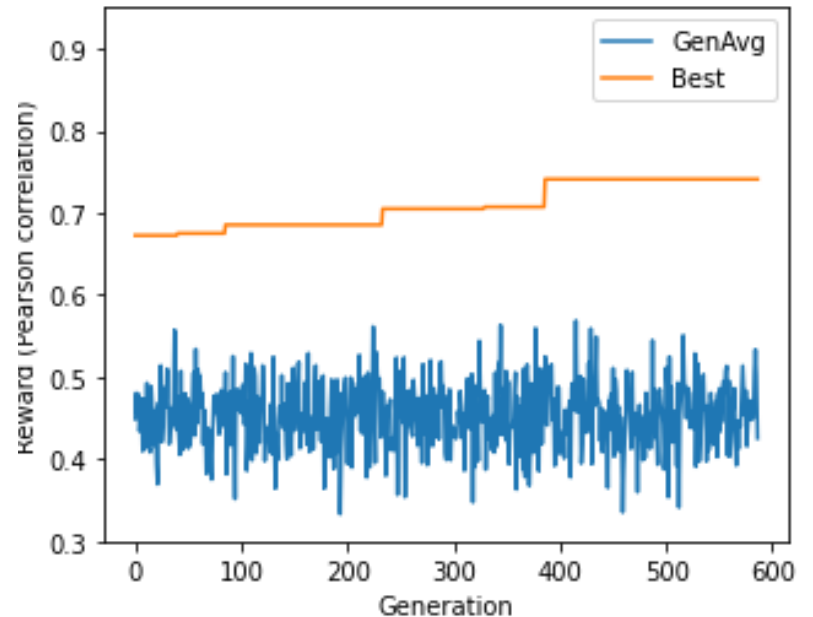
Benchmark with synthetic data



Sampling from Posterior
(update with selected models)

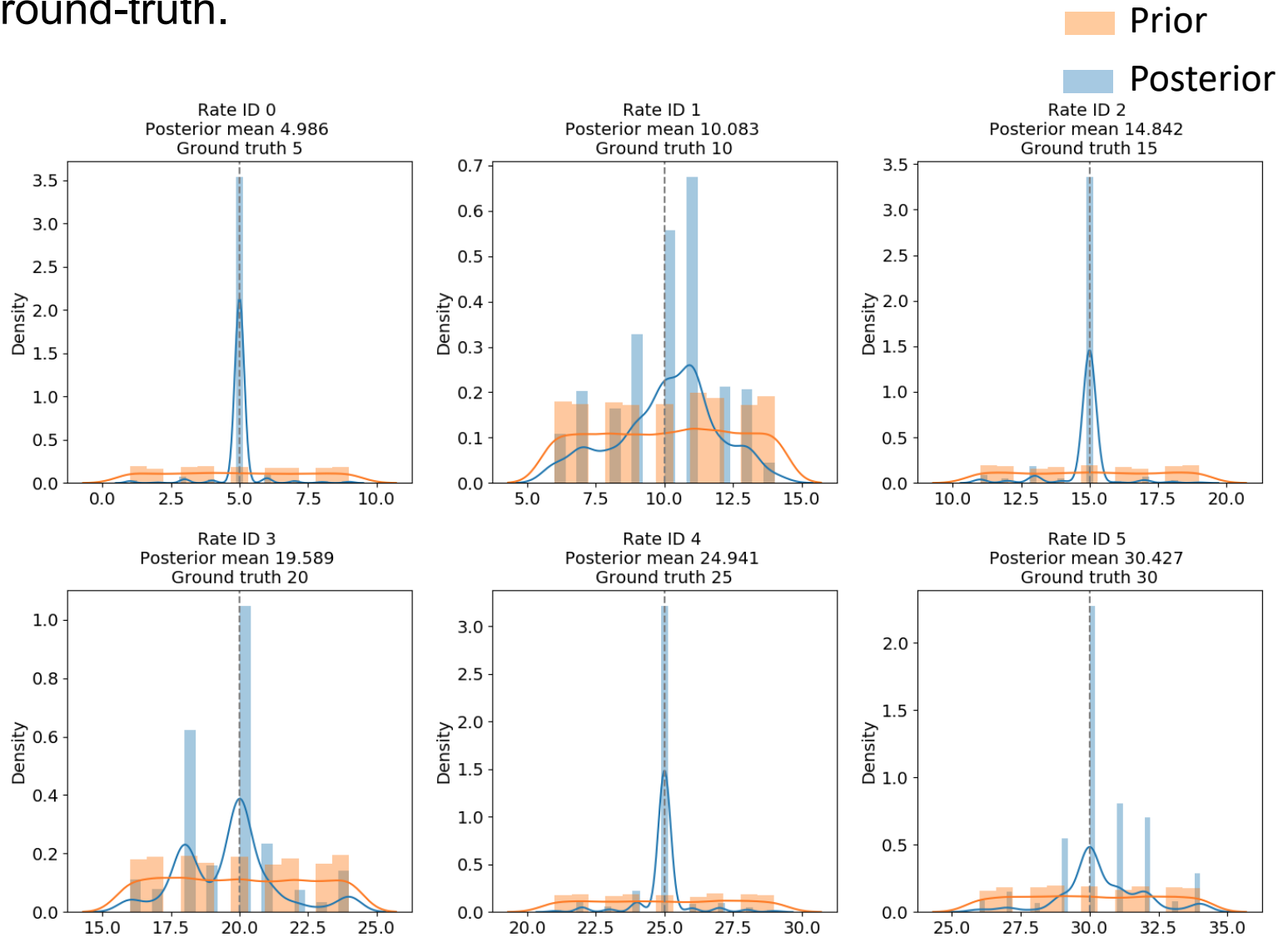


Sampling from Prior
(disable posterior update)



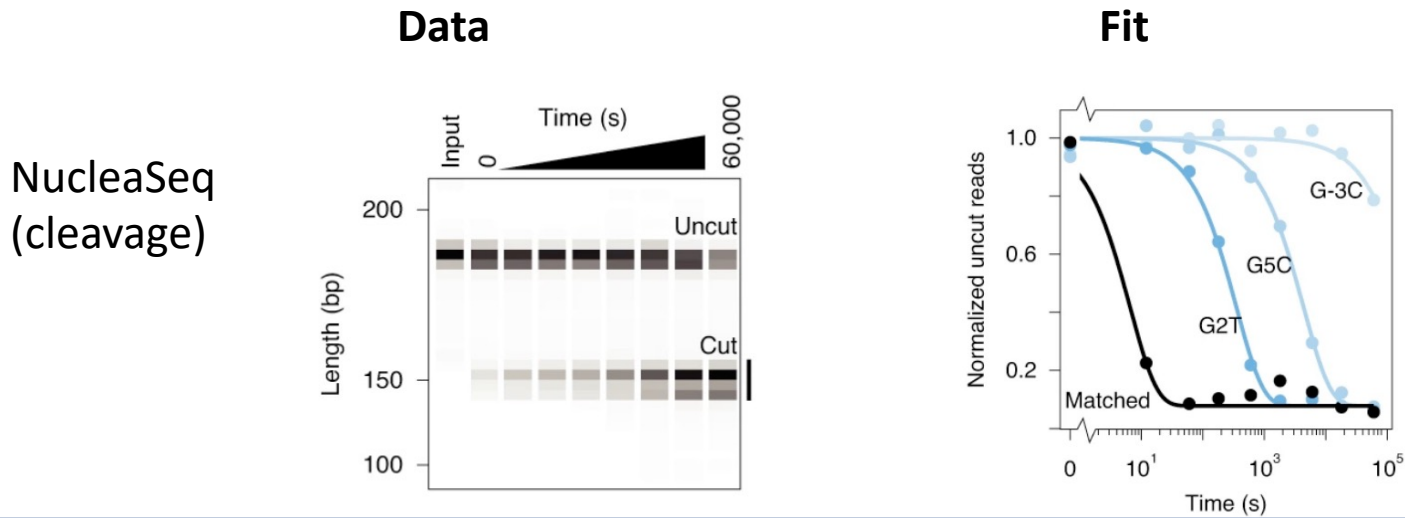
Benchmark with synthetic data

- Searched posterior for model architecture mode is aligned with ground-truth.

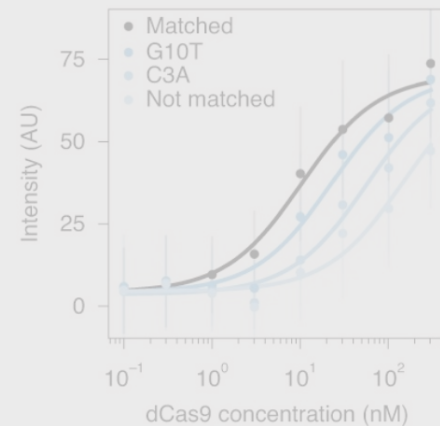
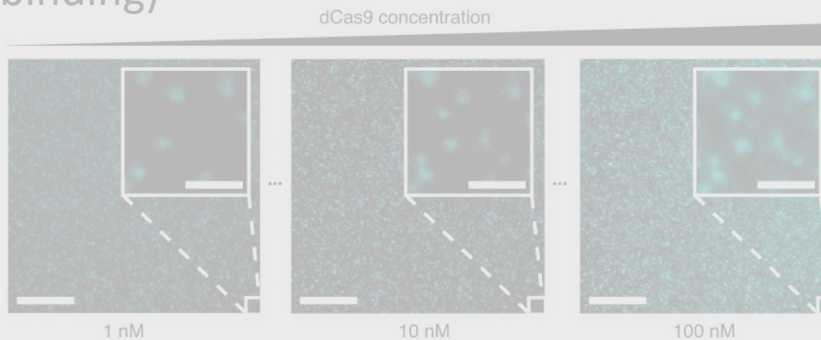


Massively Parallel Kinetic Profiling for CRISPR/Cas9

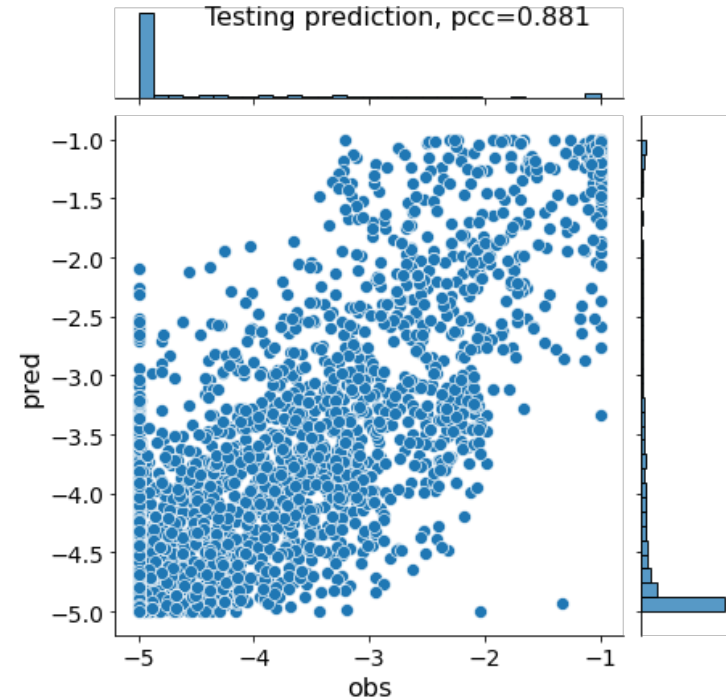
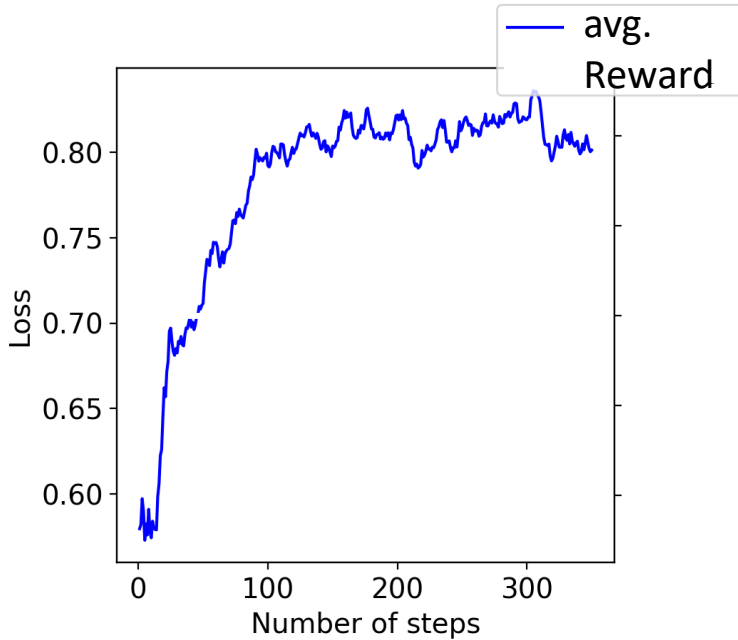
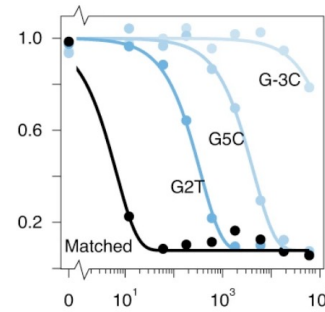
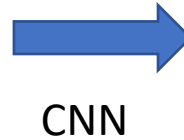
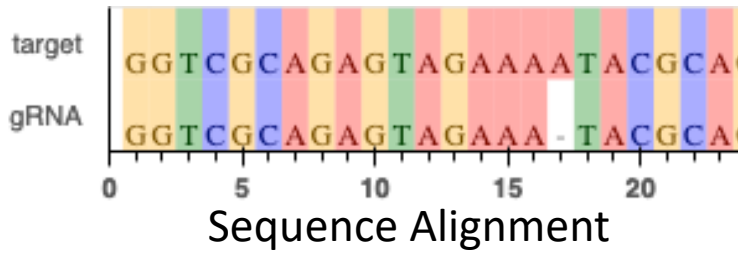
- Profiled 2 sgRNAs *in vitro*



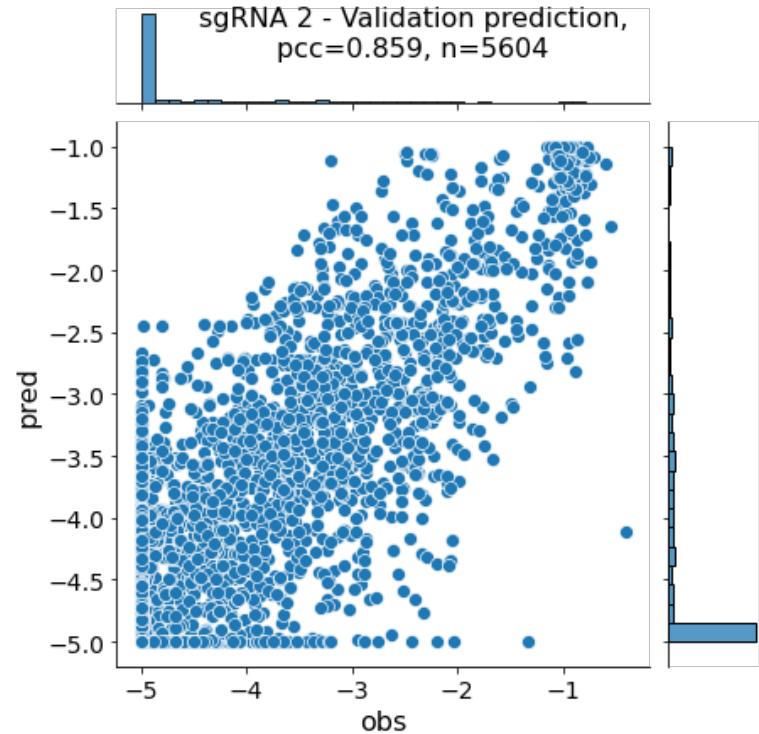
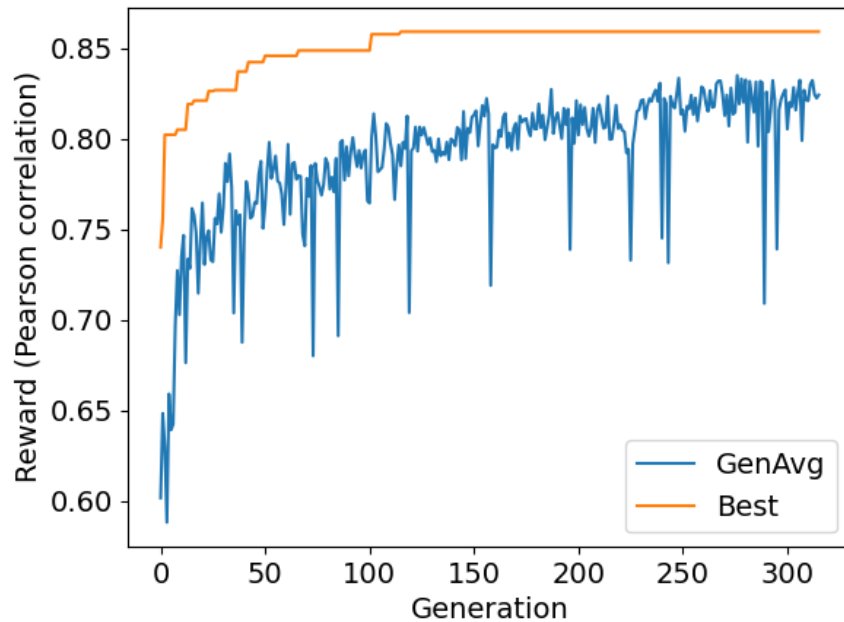
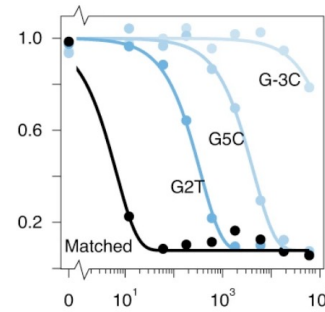
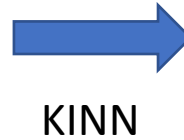
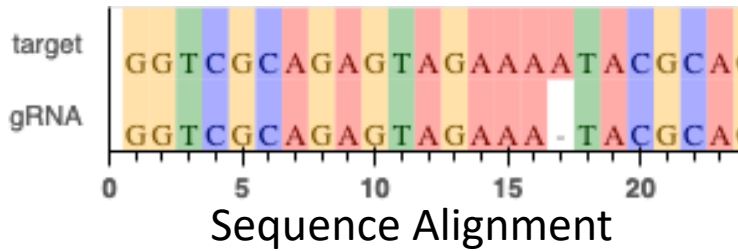
CHAMP
(binding)



AMBER deep CNN search for Cas9 cleavage

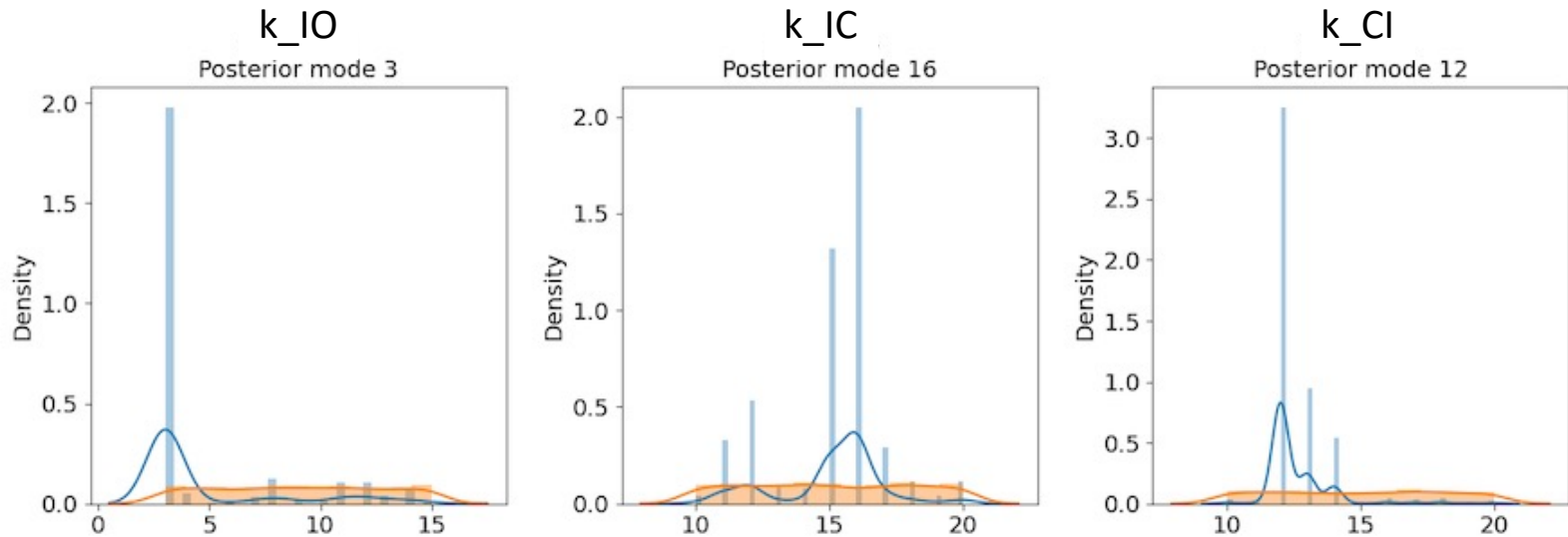


AMBER KINN search for Cas9 cleavage

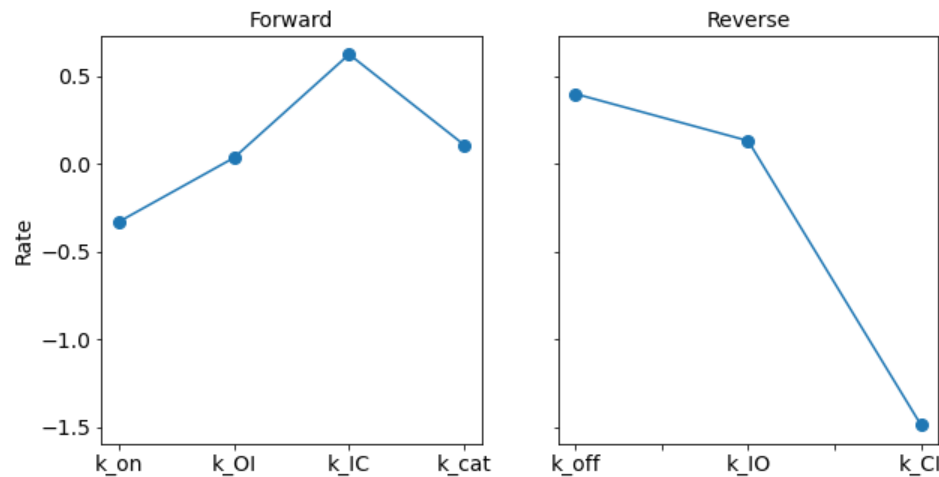


KINN is interpretable and physical

Sequence determinants of each kinetic rate

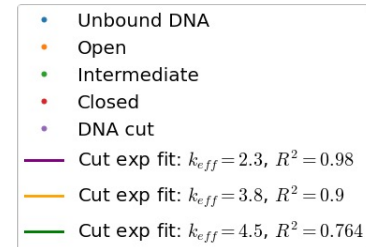
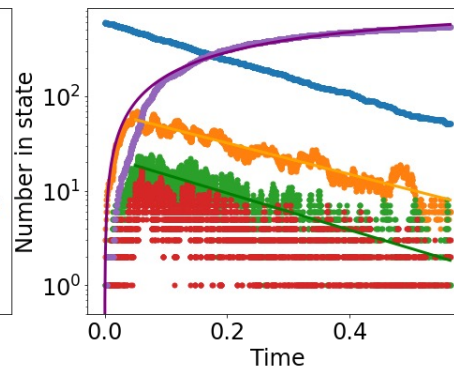
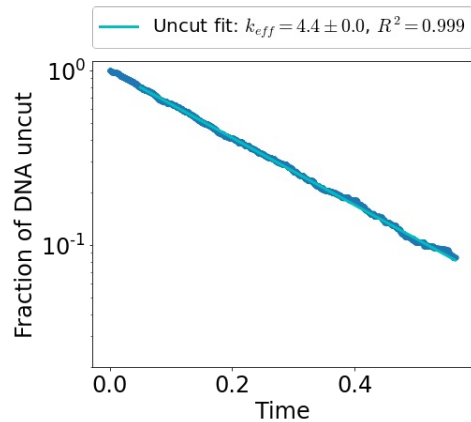
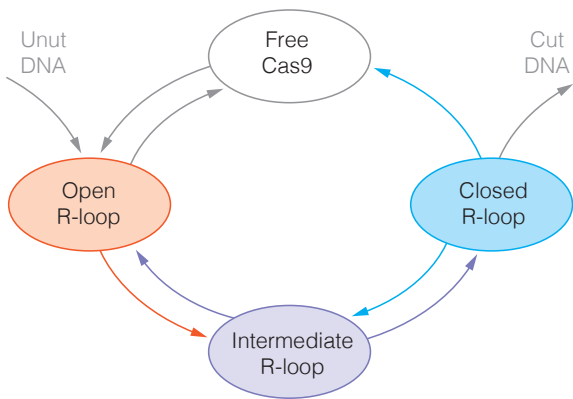


Learned kinetic rates with physical meaning (s^{-1})

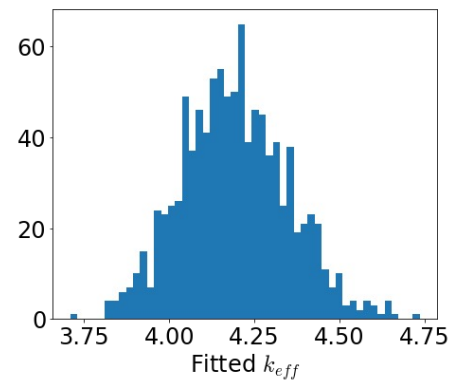
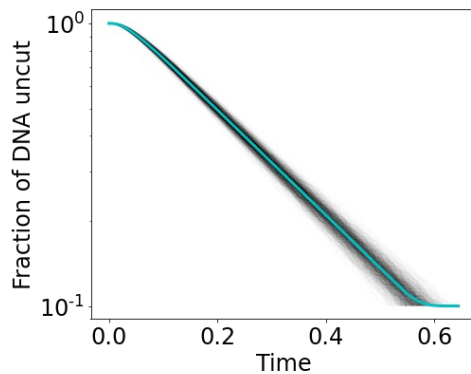


■ Prior
■ Posterior

Physics simulation of experiments from KINN learned kinetic rates



Extract variance of expect cleavage rate

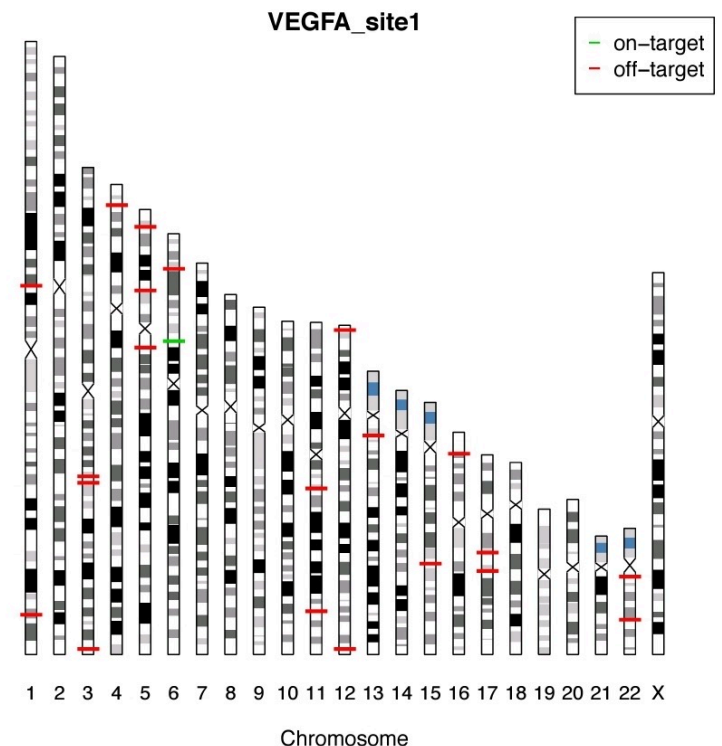


Comparison to existing Cas9 Off-targets predictors

- Test data is Guide-seq datasets *in vivo* (train data is *in vitro*)
- Task: edited off-targets vs non-edited sequences with the same Hamming distance

Performance Comparison by AUPR

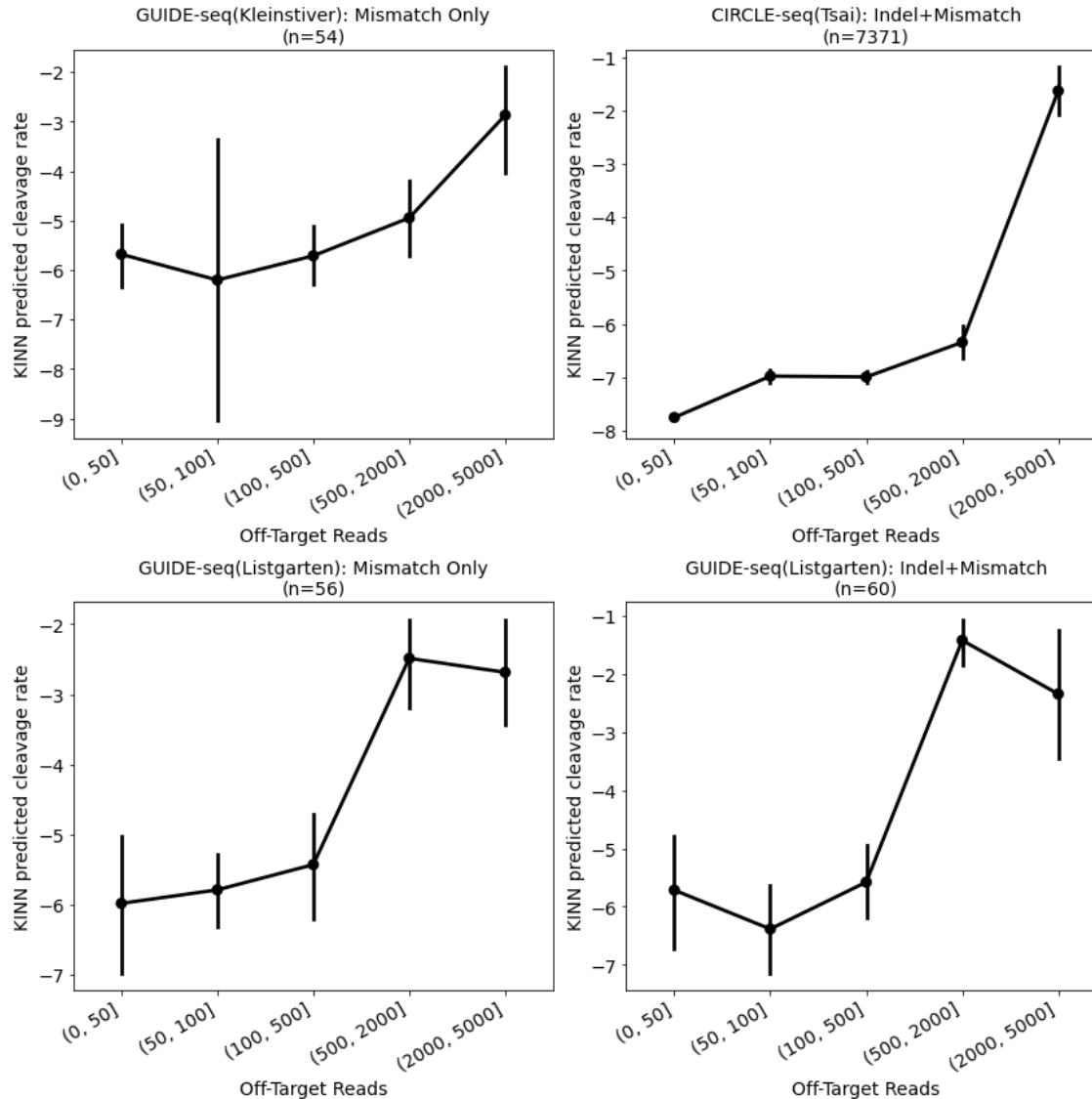
method	GUIDE (Kleinstiver)	GUIDE (Listgarten)
AMBER-KINN	0.202	0.079
AMBER-CNN	0.128	0.060
AttnToMismatch	0.071	0.025
Elevation-score	0.131	0.078
CFD	0.066	0.030
Ensemble SVM	0.113	0.048
CNN_std	0.115	0.034
CRISPROff	0.104	0.046



Chromosome ideogram of CRISPR-Cas9 on/off-target sites for VEGFA.
Tsai et al., 2015, *Nat. Biotech*

Predicted cleavage rate consistent with independent experiment measurements

- Among off-target sites, some are *more* edited than others.



Summary of AMBER/KINN

- AMBER search algorithm provides a general optimization method for building biophysics-interpretable neural networks.
- When applied on CRISPR/Cas9 kinetic data, we built a KINN that performs on par with the conventional AMBER-optimized CNN.
- KINN shed mechanistic insights on Cas9 kinetics.
- KINN outperforms existing SOTA methods for off-target predictions on external datasets, including AMBER-optimized CNN.

Outline



Basics of Deep learning in Genomics and Neural Architecture Search (NAS)



Deep residual convolutional neural network for CRISPR/Cas9 outcomes and variant effects



Biophysics-interpretable modeling of CRISPR/Cas9 off-target effect

Acknowledgements



Olga Troyanskaya
Adam Lamson
Victoria Li
Natalie Sauerwald
Christopher Park
Evan Cofer
Chandra Theesfeld
Alicja Tadych
Aaron Wong
Troyanskaya lab
members



Zhang Lab is starting in September 2022!!

Division of Artificial Intelligence in Medicine
Department of Medicine
Cedars-Sinai Medical Center
Los Angeles, CA 90048

Interested in exploring deep learning and machine learning in genomics and biomedicine? Reach out to me!

Email: zj.z@ucla.edu

Website: <https://zhanglab-aim.github.io>



