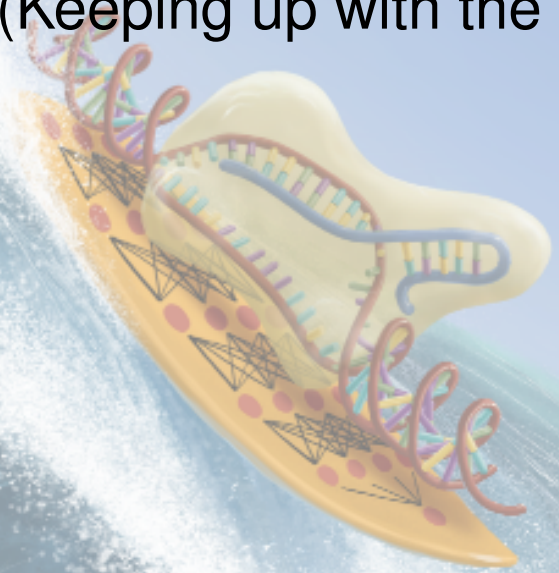
CRISPR/Cas9 gRNA design for base editing

(Keeping up with the CRISPR Tsunami)



Jan Gorodkin

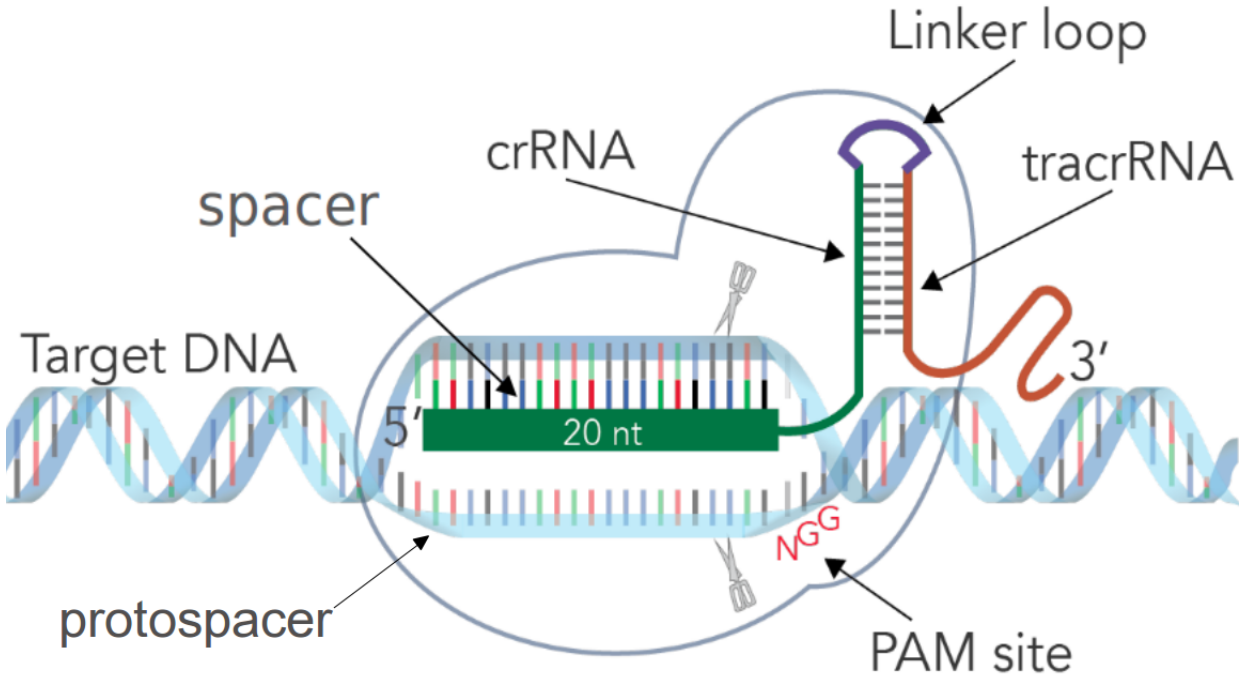
Center for non-coding RNA in Technology and Health
Department Veterinary and Animal Science
University of Copenhagen

39th TBI Winterseminar in Bled 2024

(Artist: David Deen)

CRISPR - Cas9

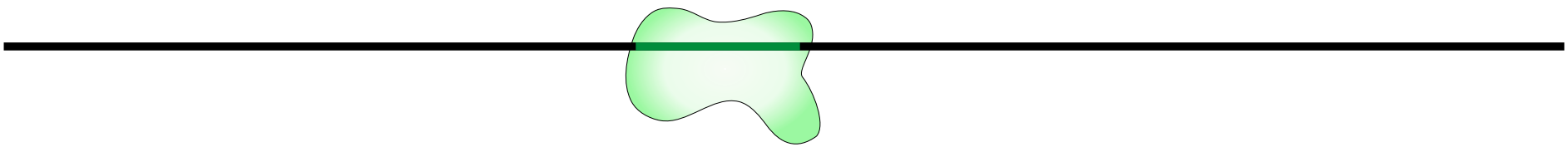
CRISPR: Clustered, Regularly Interspaced, Short Palindromic Repeats



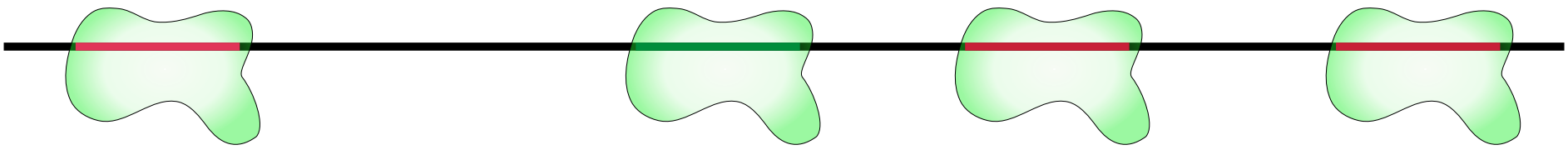
- Produces double strand breaks (DSBs) on DNA
- A single guide RNA (sgRNAs) drives the spCas9 endonuclease enzyme
  - 20nt complementary sequence
  - Adjacent to the protospacer adjacent motif (PAM) site

(Belhke, Genetic Engineering & Biotechnology News, 2016)

gRNA design



gRNA design



CRISPRon data for training and testing

Creating training, validation and independent test set of the in total 23,902 gRNAs[‡].

- 6 fold; one held out as independent test set
- 5-fold cross-validation
- gRNAs with up to 4nt differences were grouped together
- gRNAs > 4nt to other gRNAs were distributed randomly over the folds

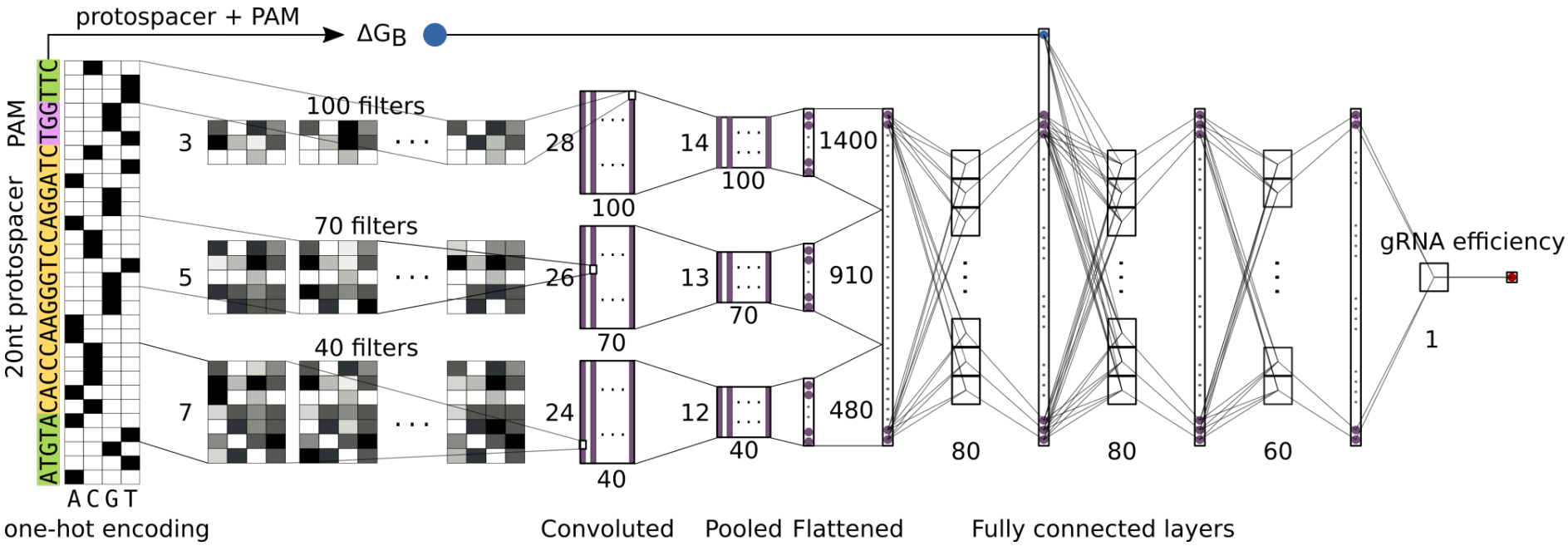
Yonglun Luo
DREAM team

Giulia Corsi
RTH

[‡] Xiang[¶], Corsi[¶], Anthon[¶], *et al.*, Nat Comm, 2021 ; [†] Pan[¶], *et al.*,. Nat Comm, 2022

CRISPRon network

Deep network for gRNA efficiency prediction‡



ΔG_B developed in the *CRISPRoff* program, is the resulting gRNA:DNA binding energy taking gRNA self-folding and DNA opening energy into account†.

Giulia Corsi
RTH

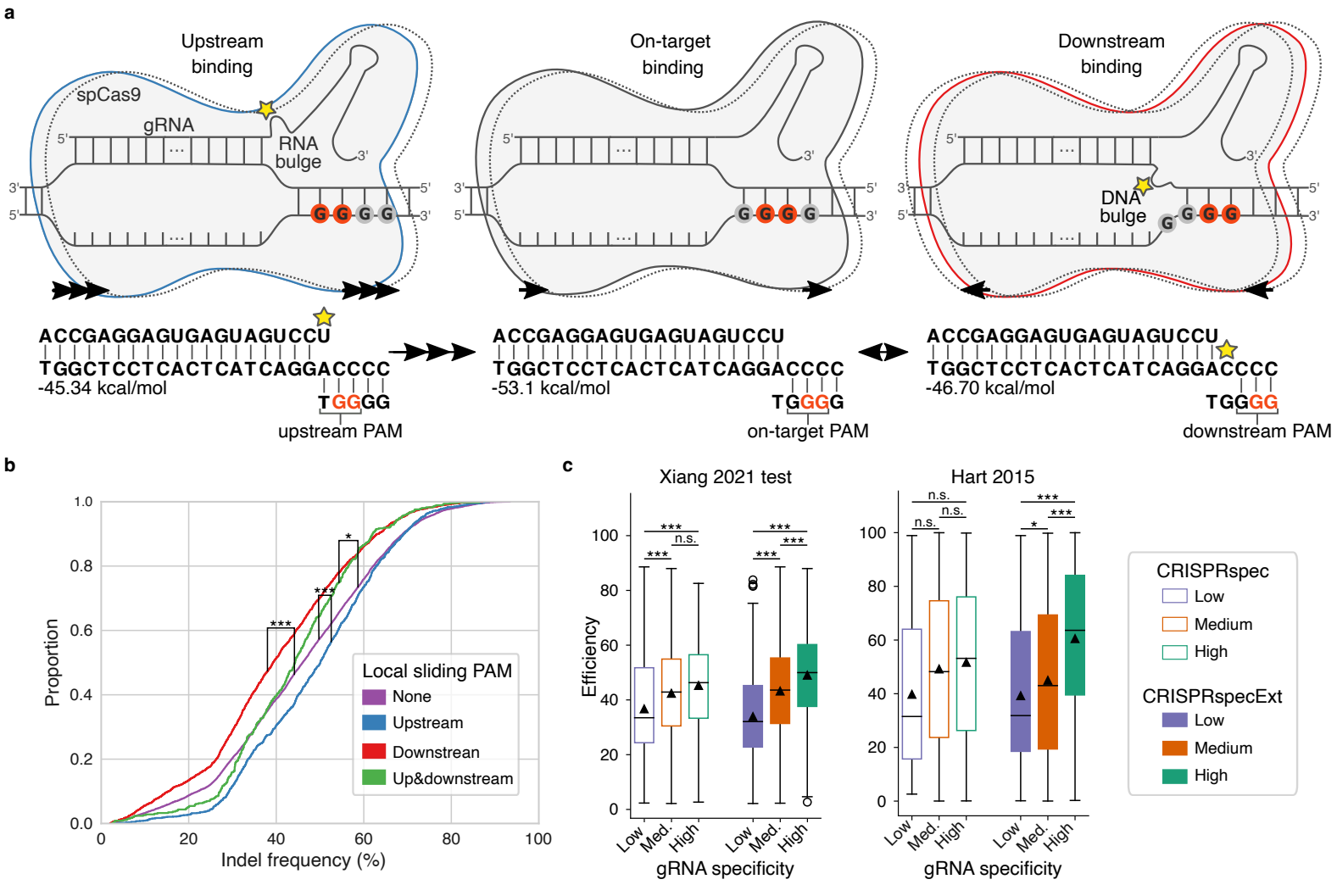
Christian Anthon
RTH

‡ Xiang¶, Corsi¶, Anthon¶, *et al.*, Nat Comm, 2021

† Alkan, *et al.*, Genome Biol, 2018.

gRNA context matters

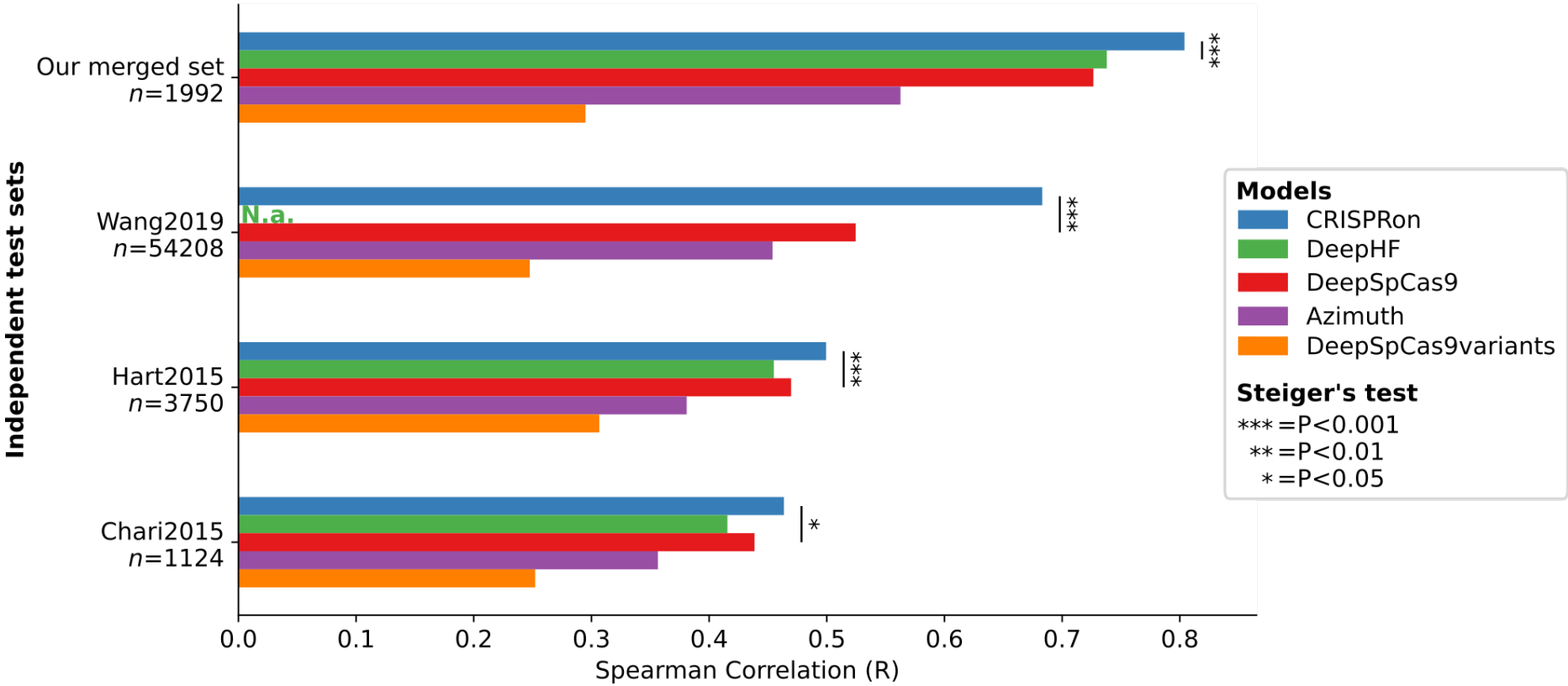
Cas9 gRNAs work in a constrained binding energy interval while being PAM context dependent†



† Corsi¶, Qu¶ *et al.*, Nat Comm, 2022

CRISPRon performance

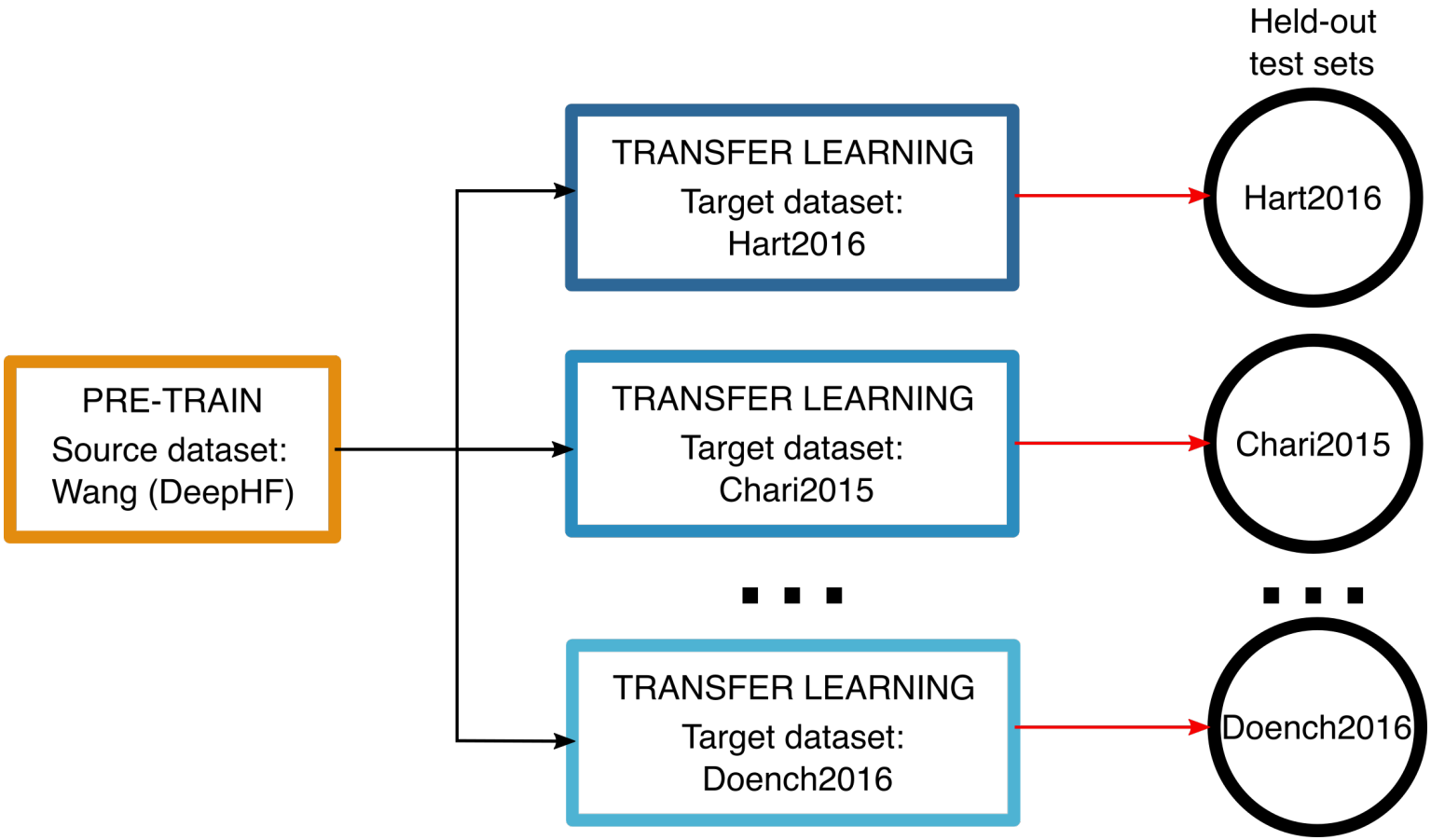
Evaluation on external data sets is critically important‡



‡ Corsi *et al.*, Letter to the editor, Bioinformatics, 2023

Benchmarking on external data is needed

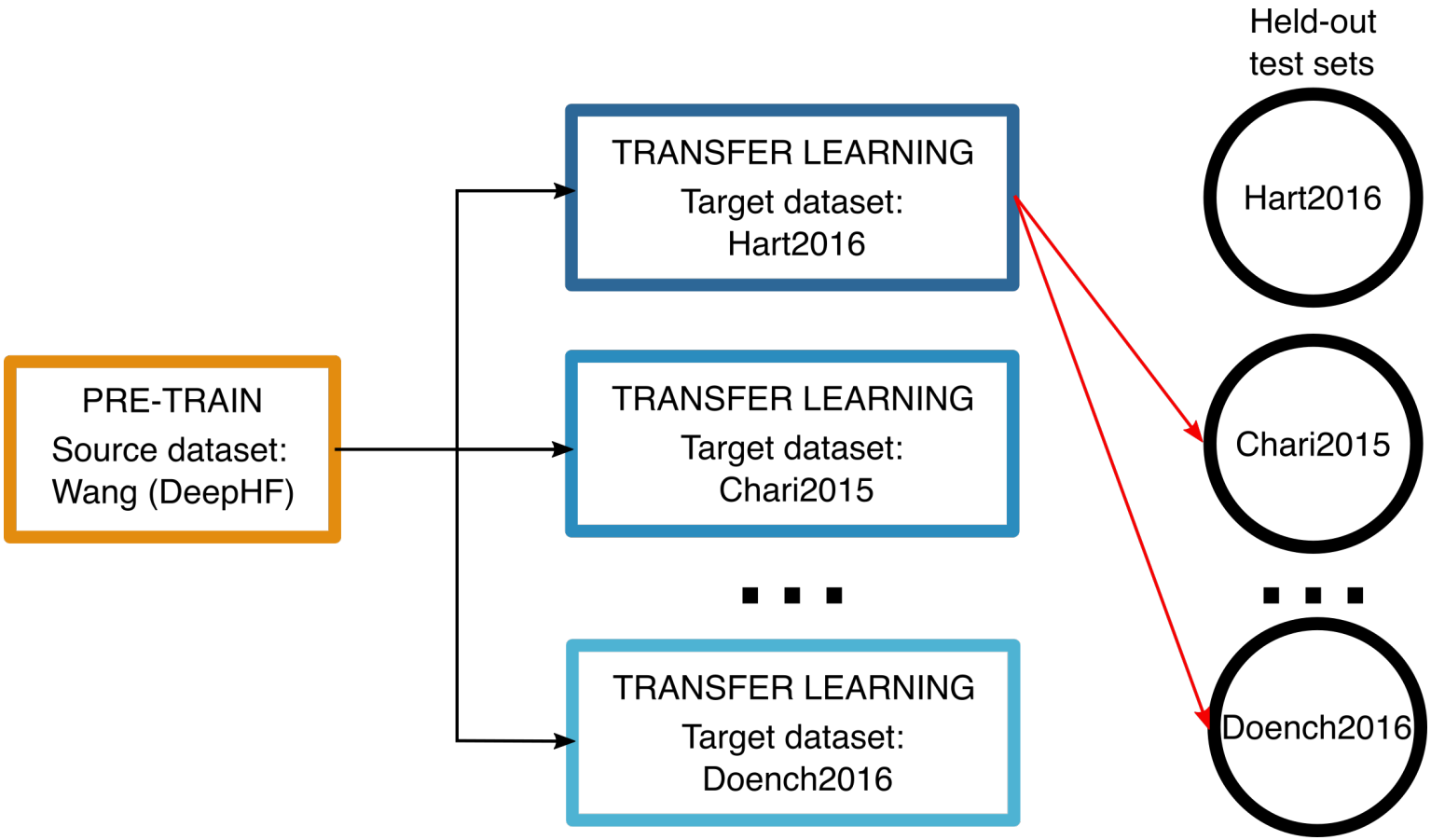
- DeepCRISTL[†]: novel set of models
  - pre-trained on large-scale datasets (surrogate gRNAs)
  - refined by transfer learning on smaller datasets (non-surrogates)



[†] Elkayam and Orenstein, Bioinformatics, 2022.

Benchmarking on external data is needed

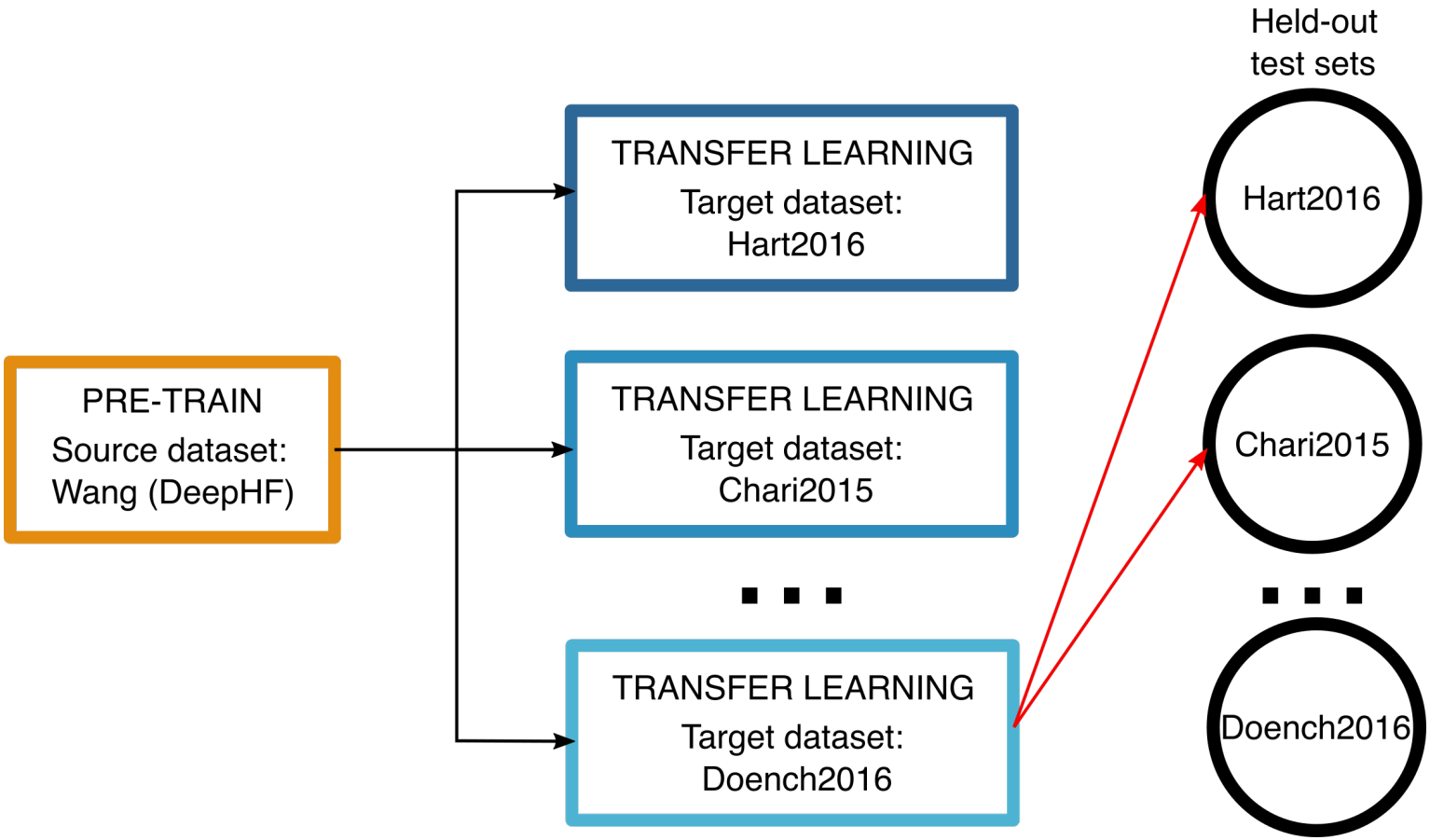
- DeepCRISTL[†]: novel set of models
  - pre-trained on large-scale datasets (surrogate gRNAs)
  - refined by transfer learning on smaller datasets (non-surrogates)



[†] Elkayam and Orenstein, Bioinformatics, 2022.

Benchmarking on external data is needed

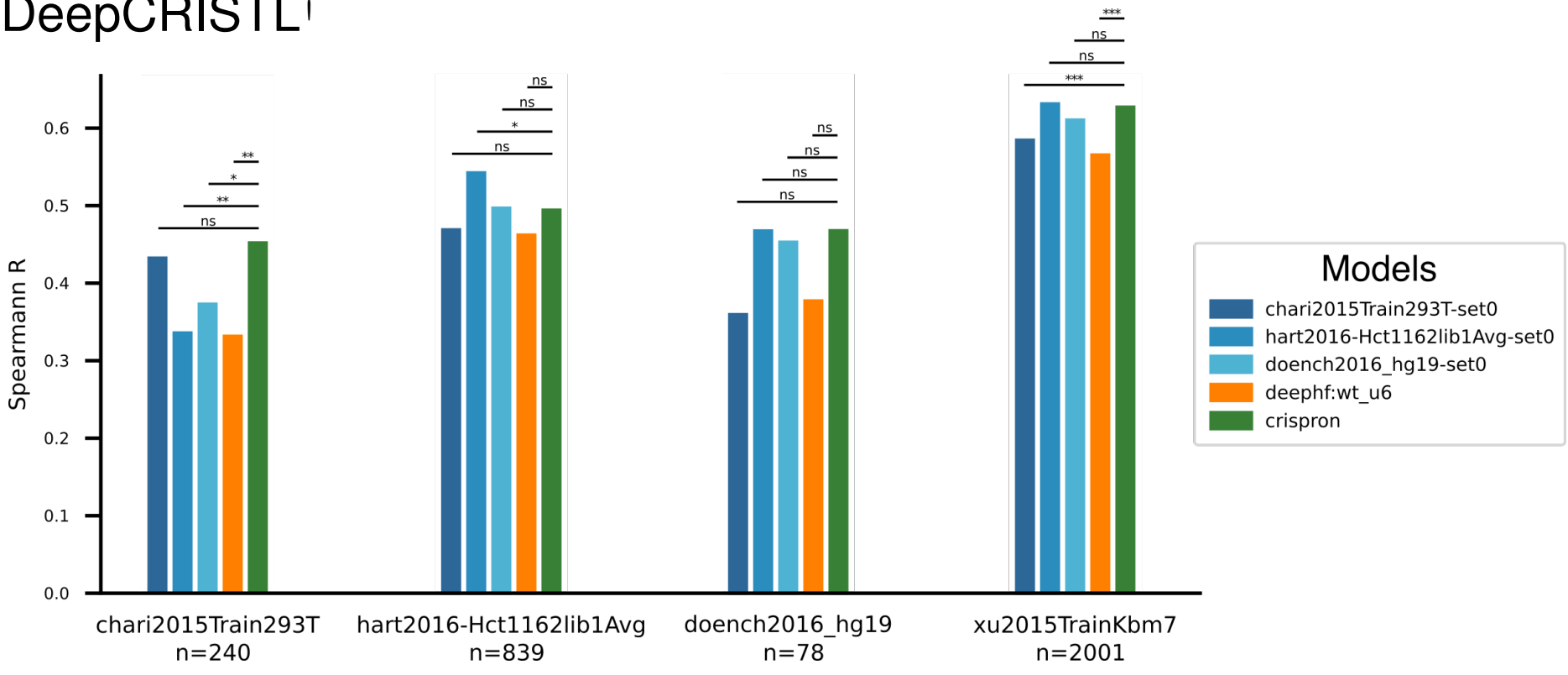
- DeepCRISTL[†]: novel set of models
  - pre-trained on large-scale datasets (surrogate gRNAs)
  - refined by transfer learning on smaller datasets (non-surrogates)



[†] Elkayam and Orenstein, Bioinformatics, 2022.

Benchmarking on external data is needed

CRISPRon perform overall better on independent data than DeepCRISTL†

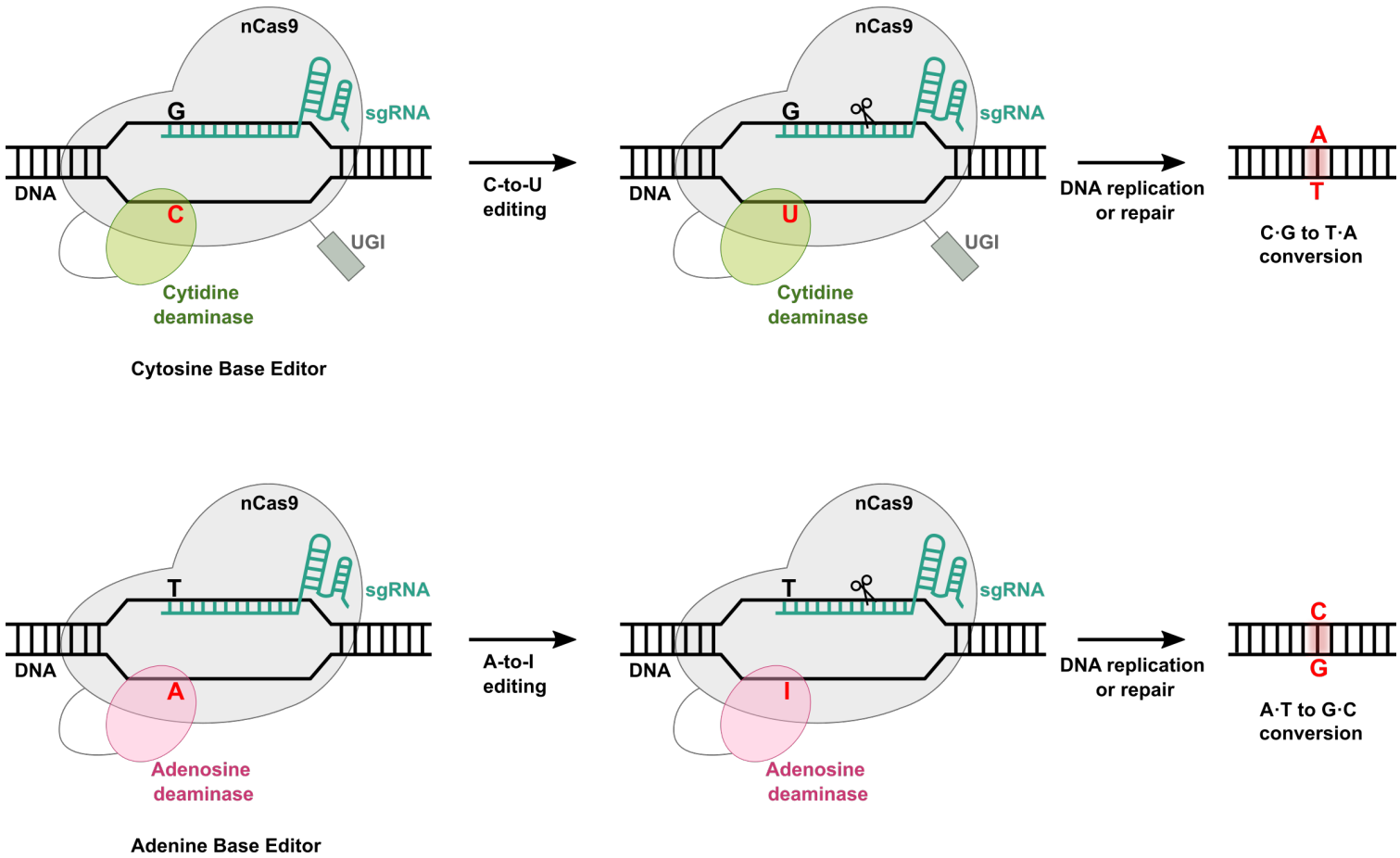


10 models on 10 data sets:
63 of 100 no difference; CRISPRon best 32 of 100; DeepCRISTL best 5 of 100

† Corsi, Bioinformatics, 2023.

Base editing

Precise genome editing by directly changing a targeted base



(Illustration by SeHee Park: https://biotech.ucdavis.edu/news/dna-base-editors-genome-editing)

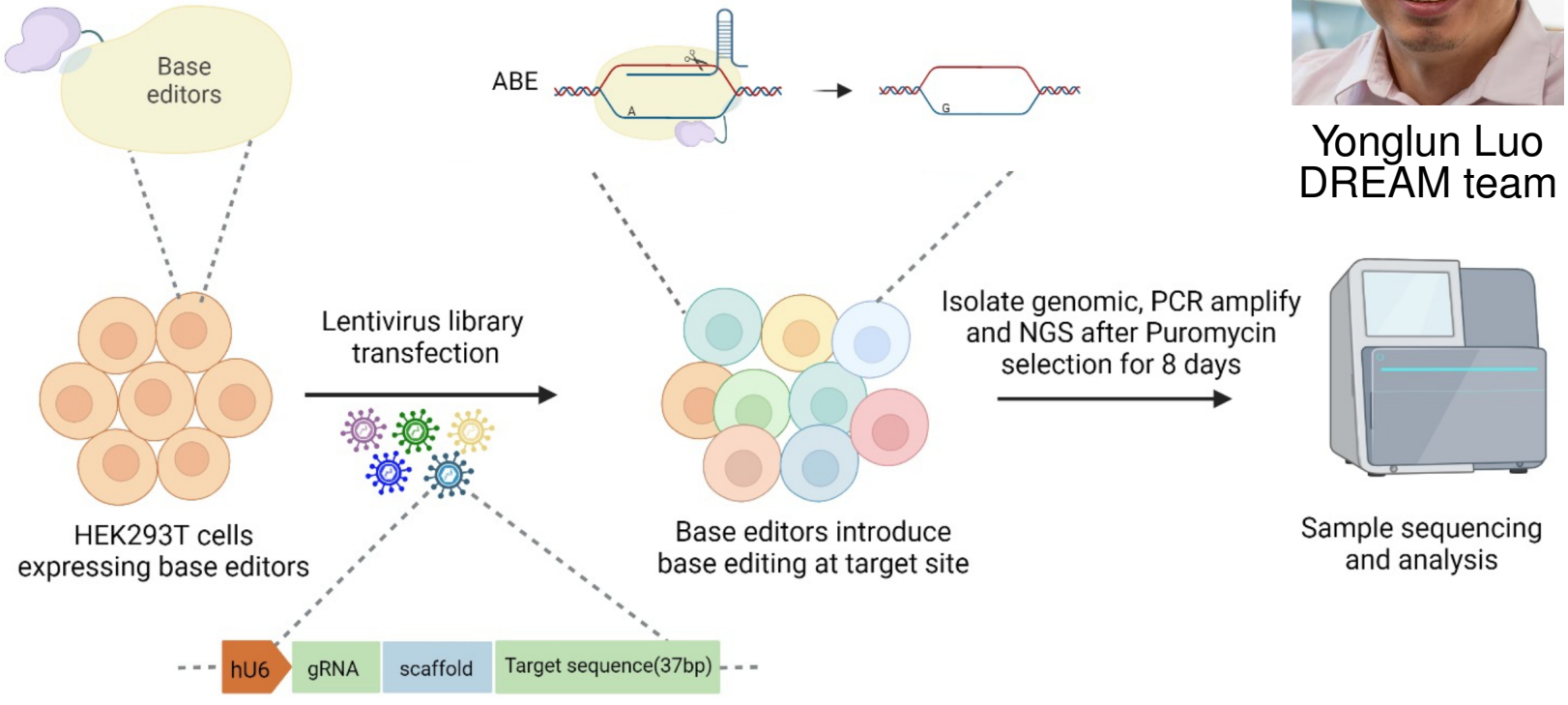
Pros: no double-strand breaks / no donor DNA template required

Cons: unwanted concurrent mutations

Base editing data

In complement to published data, we generated *in house* data

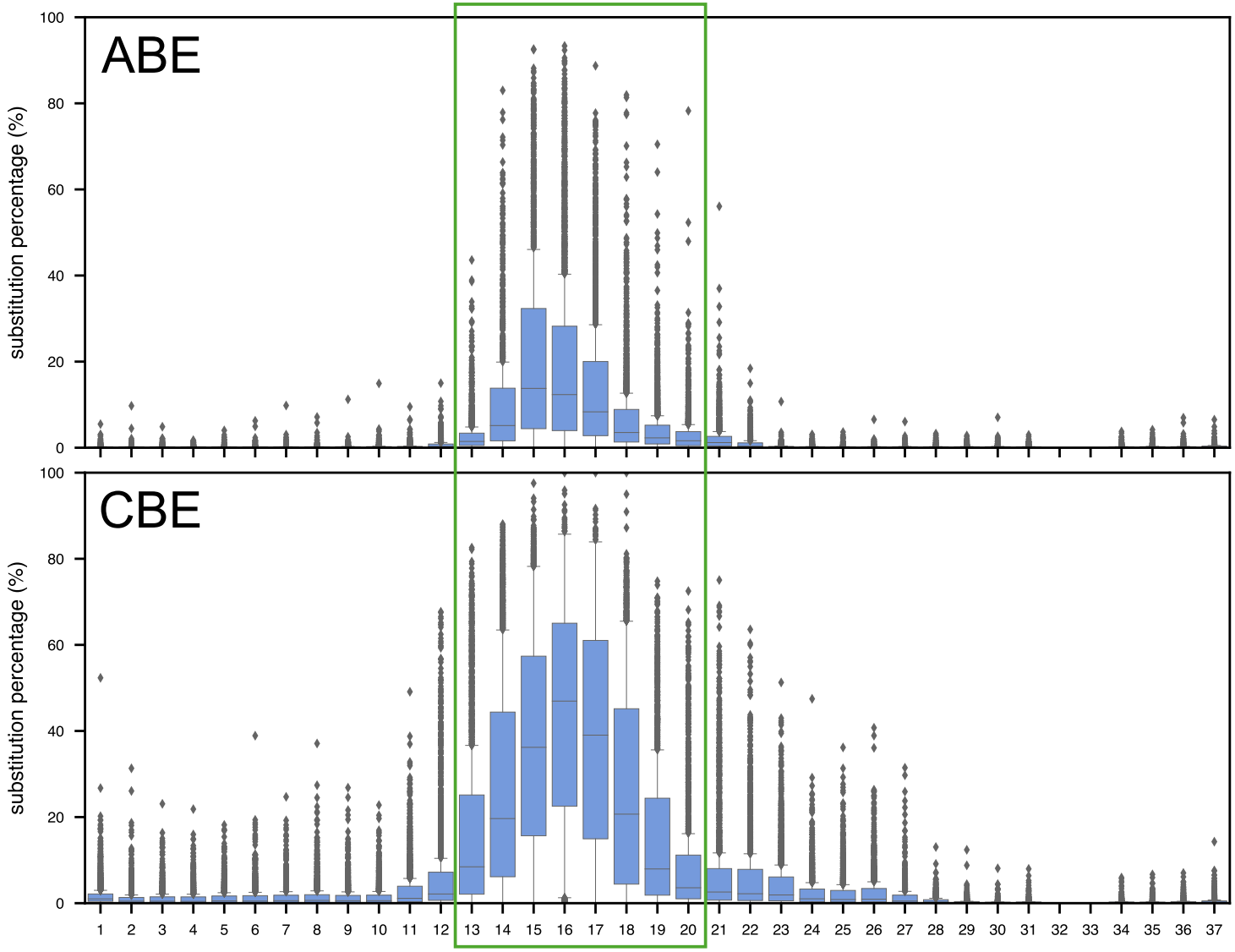
Yonglun Luo
DREAM team



† Pan, *et al.*, Nature Comm, 2022.

Base editing window

Bystander bases are edited as well



Base editing outcome

Two numbers: gRNA editing efficiency and outcome frequency

Example (ABE):

	upstream	gRNA	PAM	downstream	
Target sequence	TATCTCCAGG	GGAGGTGGTACGGCTGTAGC	GGG	GGAC	# reads measured by sequencing
Outcome1 (WT):	TATCTCCAGG	GGAGGTGGTACGGCTGTAGC	GGG	GGAC	r1
Outcome2:	TATCTCCAGG	GGGGGTGGTACGGCTGTAGC	GGG	GGAC	**r2**
Outcome3:	TATCTCCAGG	GGAGGTGGTGCGGCTGTAGC	GGG	GGAC	**r3**
Outcome4:	TATCTCCAGG	GGGGGTGGTGCGGCTGTAGC	GGG	GGAC	**r4**

$$total = r1 + r2 + r3 + r4$$

$$\textbf{gRNA editing efficiency} = \frac{\text{\# total reads of all sequences with intended target nucleotide transitions}}{\text{\# total reads}} = \frac{\text{r2+r3+r4}}{\text{total}}$$

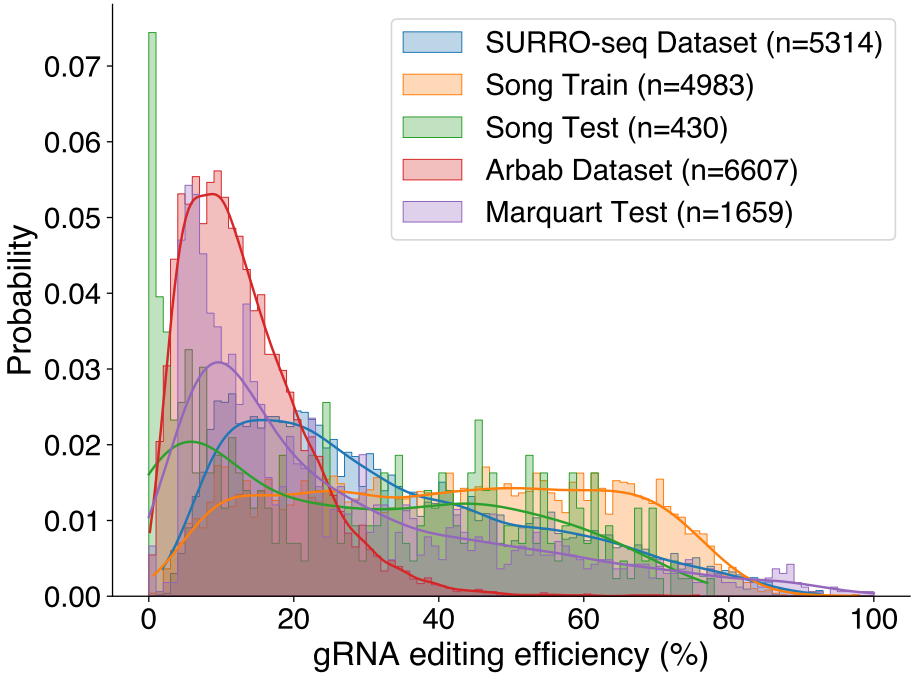
$$\textbf{outcome frequency} = \frac{\text{\# reads of specific base–edited outcome sequence}}{\text{\# total reads}} = \frac{\text{r2}}{\text{total}}$$

$$\text{gRNA editing efficiency} = \sum \text{edited outcome frequency}$$

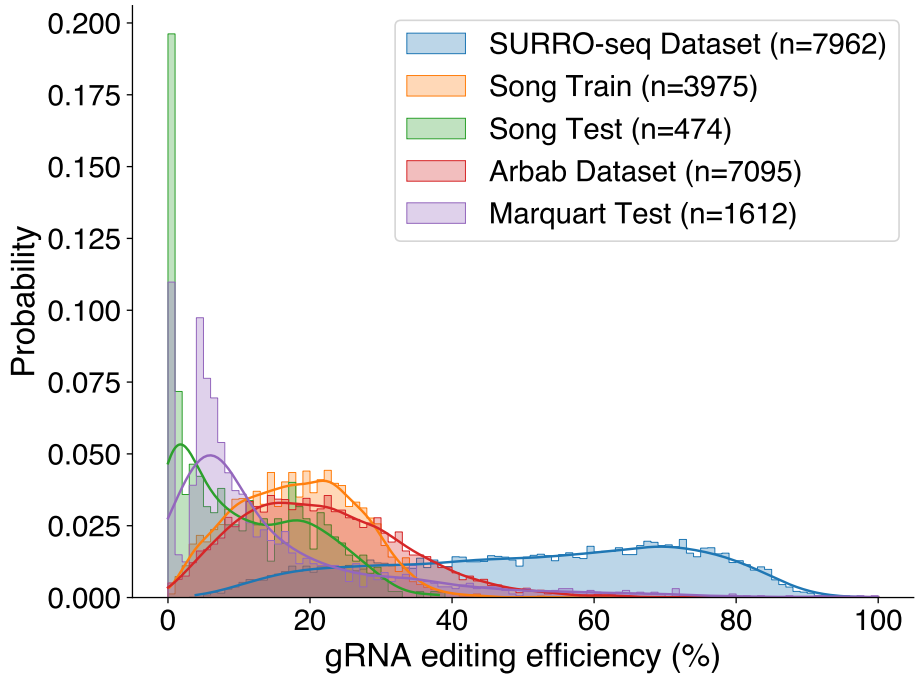
Base editing data

gRNA edtiting efficiencies

ABE:



CBE:



Data sources: our dataset: Sun *et al.*, (in prep); Song training and test: Song, et al., Nat. Biotechnol., 2020; Arbab dataset: Arbab, *et al.*, Cell, 2020; Marquart test set: Marquart, et al., Nat. Comm., 2021

Base editing data

Current prediction methods evaluate the performance individually of gRNA efficiency and outcome frequency.

Here: evaluate the numbers jointly with a fused correlation coefficient[†]

[†] Gorodkin, Comput Chem, 2004.

Extending Pearson's correlation coefficient

Consider two $N \times K$ tables: $\underline{\underline{X}}$ and $\underline{\underline{Y}}$. Define[†]

$$COV(\underline{\underline{X}}, \underline{\underline{Y}}) = \sum_{k=1}^{K} w_k COV(\underline{\underline{X}}_k, \underline{\underline{Y}}_k) = \frac{1}{K} \sum_{n=1}^{N} \sum_{k=1}^{K} (X_{nk} - \overline{X}_k)(Y_{nk} - \overline{Y}_k)$$

where $\overline{X}_k = \frac{1}{N} \sum_{n=1}^{N} X_{nk}$ and $\overline{Y}_k$ are the respective means of column k. Use ("prior") $w_k = 1/K$.

$$R_K = \frac{COV(\underline{\underline{X}}, \underline{\underline{Y}})}{\sqrt{COV(\underline{\underline{X}}, \underline{\underline{X}}) COV(\underline{\underline{Y}}, \underline{\underline{Y}})}}$$

[†] Gorodkin, Comput Chem, 2004.

The Discrete version of R_K

The $K \times K$ *confusion* matrix $\underline{\underline{C}}$ [†]

$$R_K = \frac{N\,Tr(\underline{\underline{C}}) - \sum_{kl} \tilde{\underline{\underline{C}}}_k \hat{\underline{\underline{C}}}_l}{\sqrt{N^2 - \sum_{kl} \tilde{\underline{\underline{C}}}_k (\hat{\underline{\underline{C}}^{\top}})_l}\sqrt{N^2 - \sum_{kl} (\tilde{\underline{\underline{C}}^{\top}})_k \hat{\underline{\underline{C}}}_l}}$$

- $\tilde{\underline{\underline{C}}}_k$ the kth row of $\underline{\underline{C}}$.
- $\hat{\underline{\underline{C}}}_l$ the lth column of $\underline{\underline{C}}$.
- $\underline{\underline{C}}^{\top}$ is $\underline{\underline{C}}$ transposed.

[†] Gorodkin, Comput Chem, 2004.

The Rank version of R_K

Using ranks for k vectors each with n numbers.
Equivantly for the distance $d_{nk} = (x_{nk} - y_{nk})$ one can obtain[†]

$$\rho_K = 1 - \frac{1}{K} \sum_{k=1}^{K} \frac{6 \sum_{n=1}^{N} d_{nk}^2}{N(N^2 - 1)}$$

With ties (two or more variables with the same rank) we use the full version.

[†] Sun & Gorodkin (in prep)

CRISPRon-ABE data for training and testing

Our set is matched into the splits as for CRISPRon[‡].

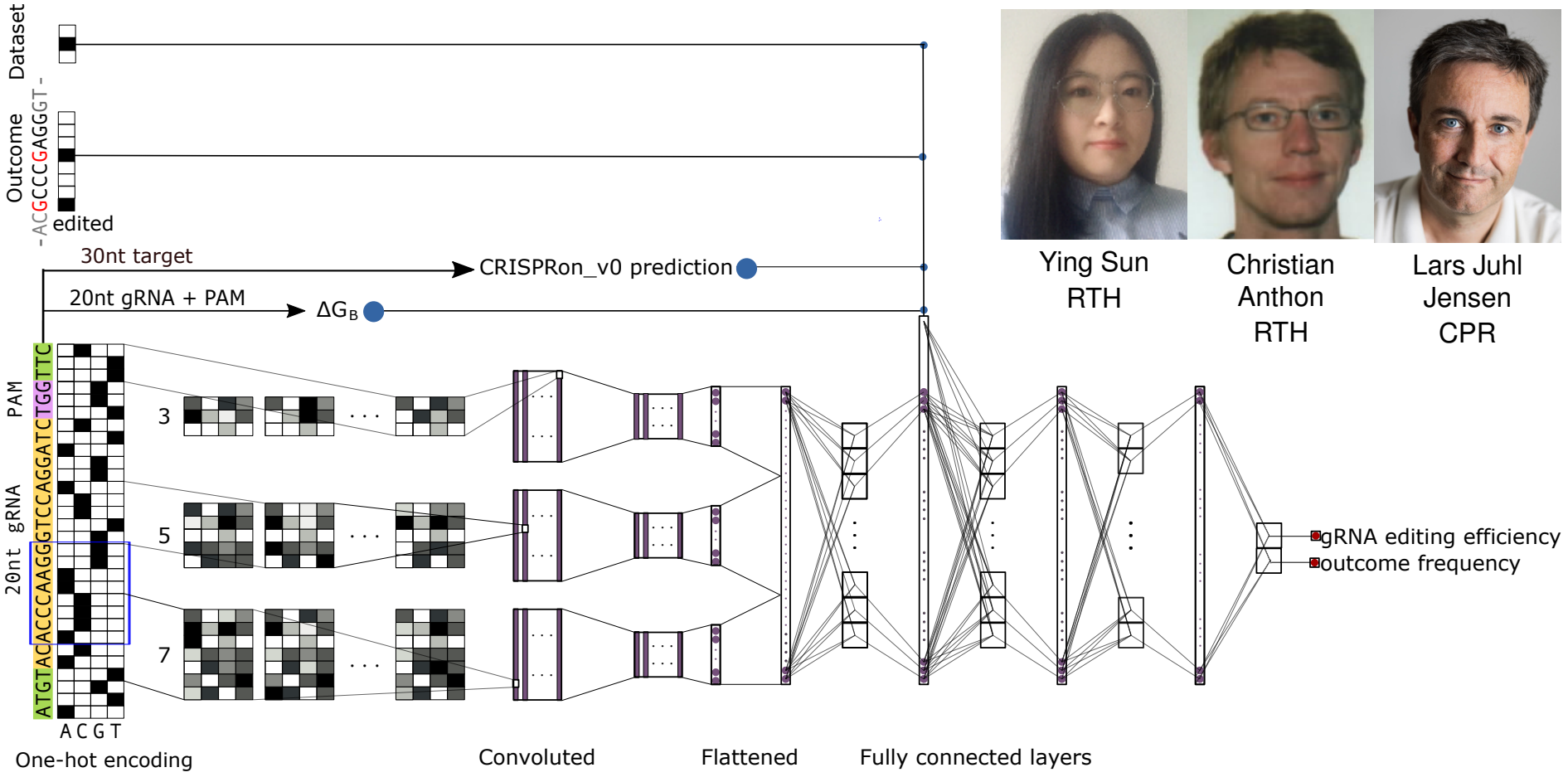
- 6 fold; same fold as independent test set
- Same 5-fold cross-validation for training
- gRNAs with up to 4nt differences were grouped together when adding new datasets
- gRNAs > 4nt to other gRNAs were distributed randomly over the folds

Ying Sun
RTH

[‡] Xiang[¶], Corsi[¶], Anthon[¶], *et al.*, Nat Comm, 2021

CRISPRon-ABE deep network

Deep network extended on the one for CRISPRon‡



Ying Sun
RTH

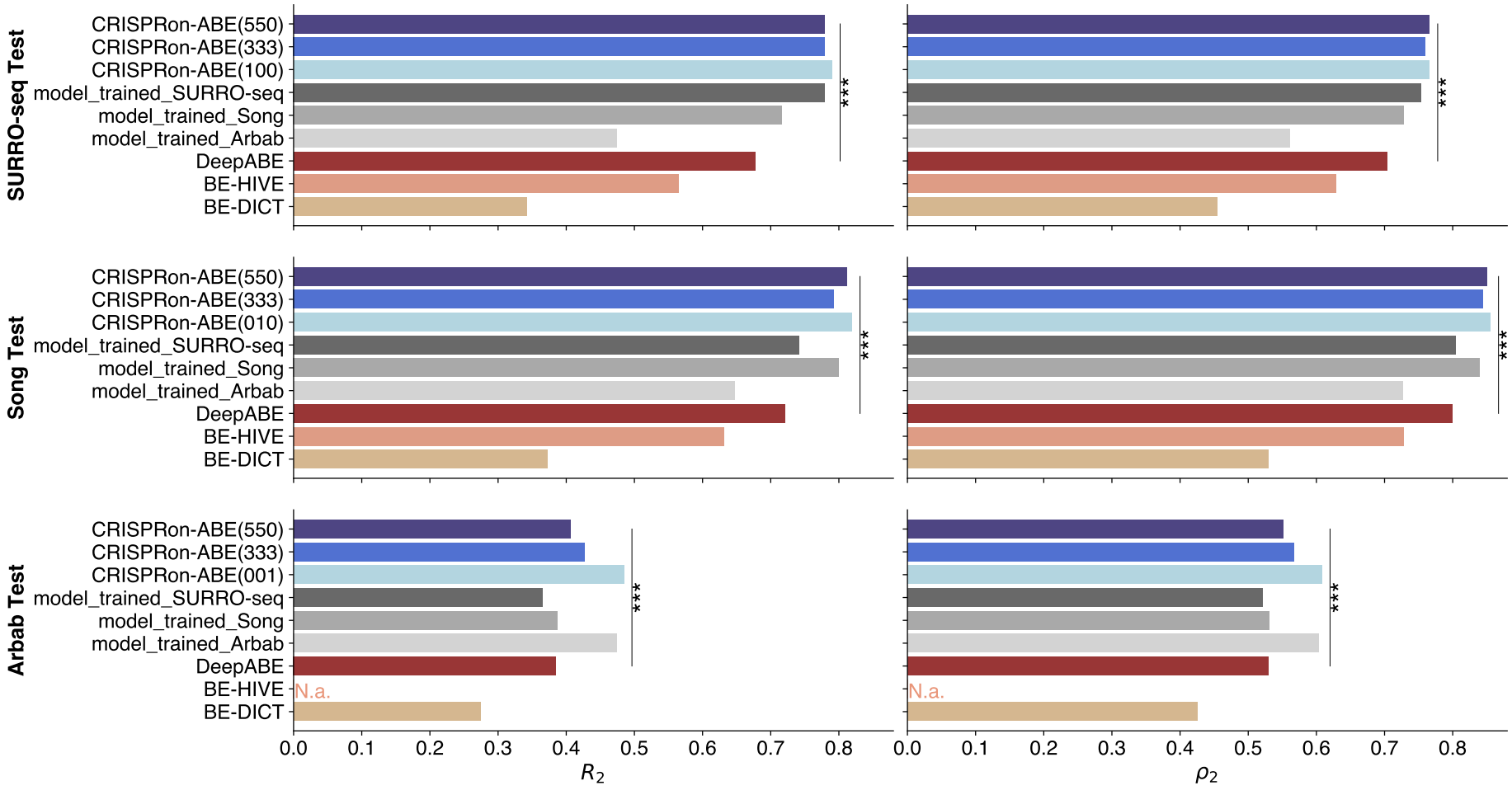
Christian
Anthon
RTH

Lars Juhl
Jensen
CPR

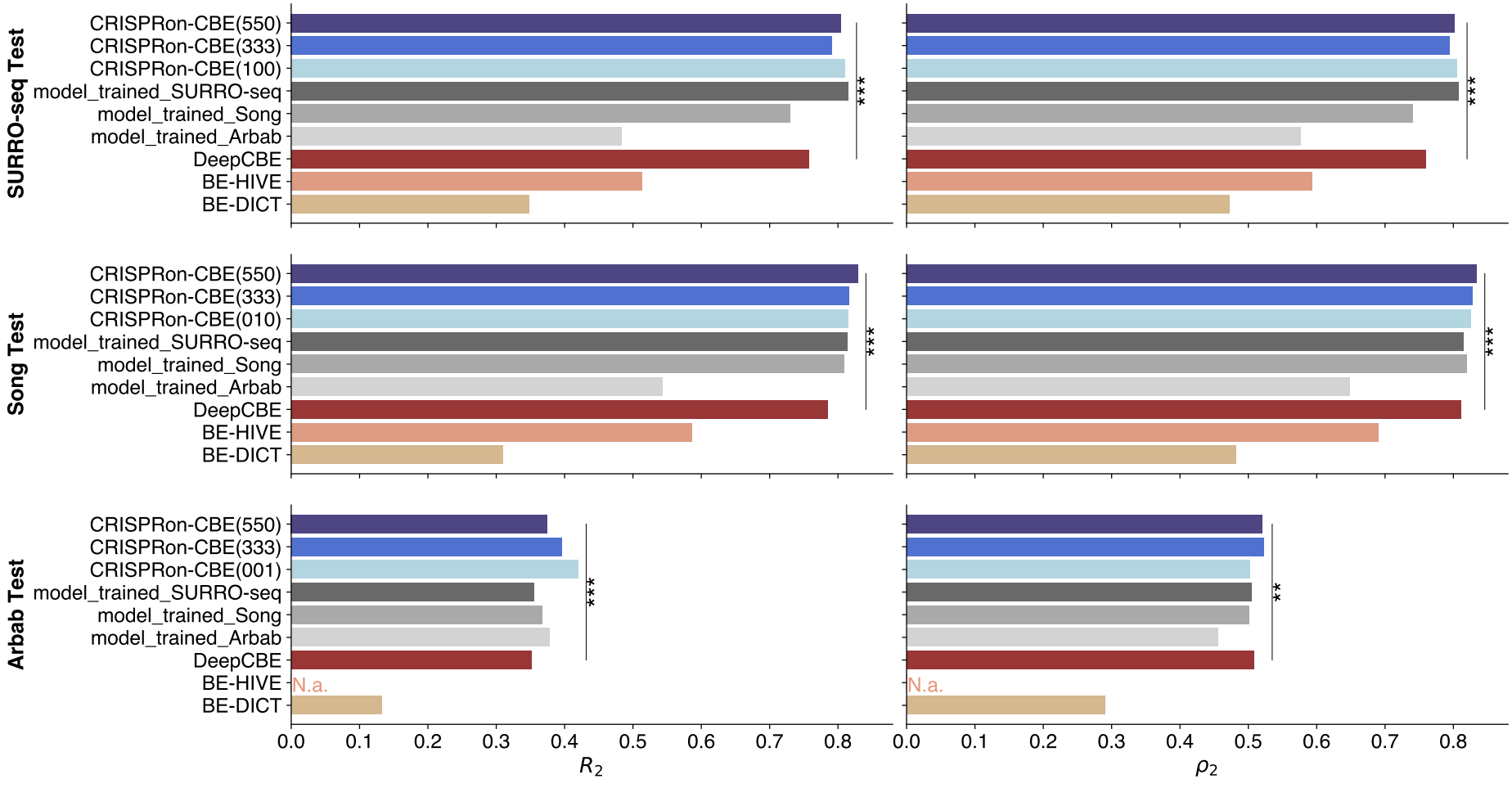
Editing with indicating outcome; CRISPRon predictions; Binding energy features
Data set indication

‡ Xiang¶, Corsi¶, Anthon¶, *et al.*, Nat Comm, 2021

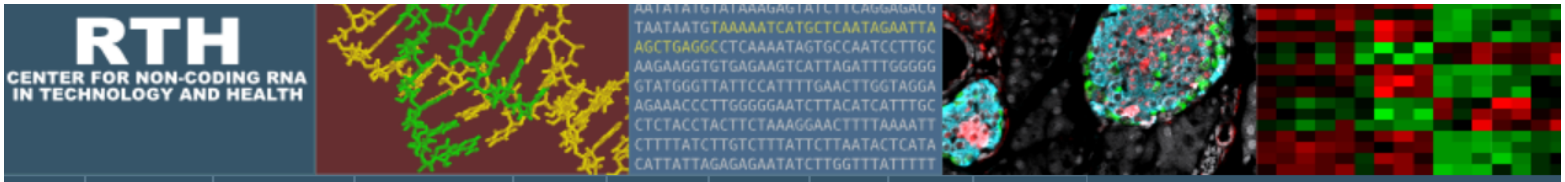
CRISPRon-ABE performance



CRISPRon-CBE performance



CRISPR tools at RTH



CRISPR

Webservers for CRISPR Cas9 on- and off-target predictions.

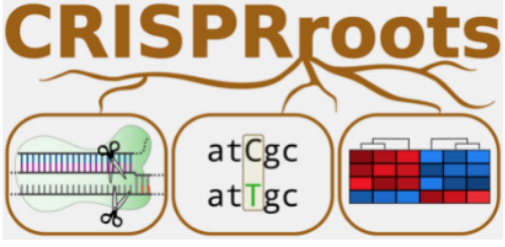
CRISPRon

State of the art on-target efficiency predictions for CRISPR-Cas9 based on deep learning utilizing the binding energy model developed for CRISPRoff.

Try the CRISPRon webserver for on-target effiency prediction.

CRISPRroots

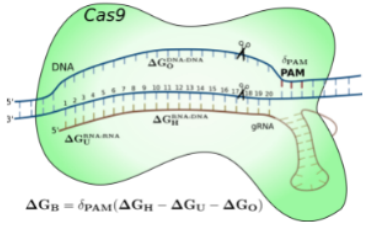
Computational pipeline for the analysis of RNA-seq data from CRISPR/Cas9 edited and control cells. The pipeline offers on-target edit verification and detection of possible off-targets affecting the transcripome.



Download the CRISPRroots pipeline here.

CRISPRoff

Off-target predictions for CRISPR-Cas9 based on an energy model for the RNA-DNA duplex binding. The model out-performs machine learning models on existing off-target data.



Try the CRISPRoff webserver to predict CRISPR-Cas9 specificity and off-targets.

https://rth.dk/resources/crispr/

Conclusions and perspectives

- CRISPR data is crucial to make good design models
- More is desirable
- Evaluation simultaneous on gRNA efficiency and outcome frequency
- Evaluation on external data sets (although data sets are diverse)
- Deep learning with flagging specific data sets
- Advancing base editing prediction

Acknowledgements

Involved team members @ UCPH:

- **Ying Sun**
- **Christian Anthon**
- **Giulia Corsi** (Alumni)
- **Ferhat Alkan** (Alumni)
- Adrian Geissler
- Dhouha Grissa
- Dhvani Vora
- Xueer Han
- Wenhao Gao
- Ziyi Sheng
- Jakob H. Havgaard
- Stefan E. Seemann

External collaborators:

- Yonglun Luo + team, Lars Bolund Institute & University of Aarhus
- Kunli Qu Lars Bolund Institute
- Xiaoguang Pan, Lars Bolund Institute
- Lars Juhl Jensen, CPR, UCPH

Funding:

- Innovation Fund Denmark
- Danish Research Councils
- Danish Center for Scientific Computing / DeiC
- Novo Nordisk Foundation

Web servers and software:
http://rth.dk/resources
http://rth.dk/resources/crispr

Open positions:
PhD position available
Postdoc (to be announced shortly)
Contact me (gorodkin@rth.dk) for further info.