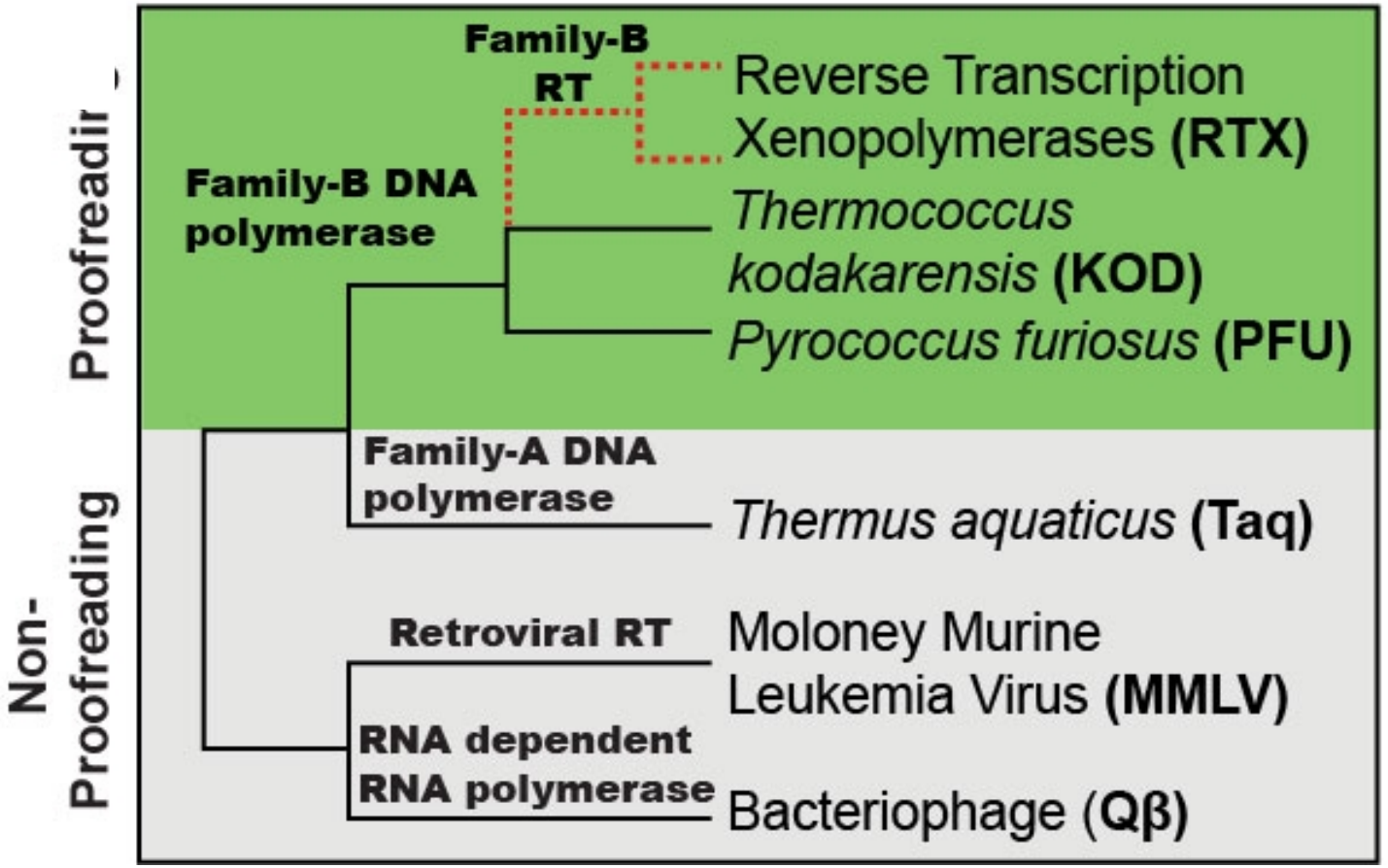
Is a blind watchmaker the same as a blind neural net?
Adventures in protein engineering

Andrew Ellington
Center for Systems and Synthetic Biology
University of Texas at Austin

NSF Nanoscience
December 7, 2023

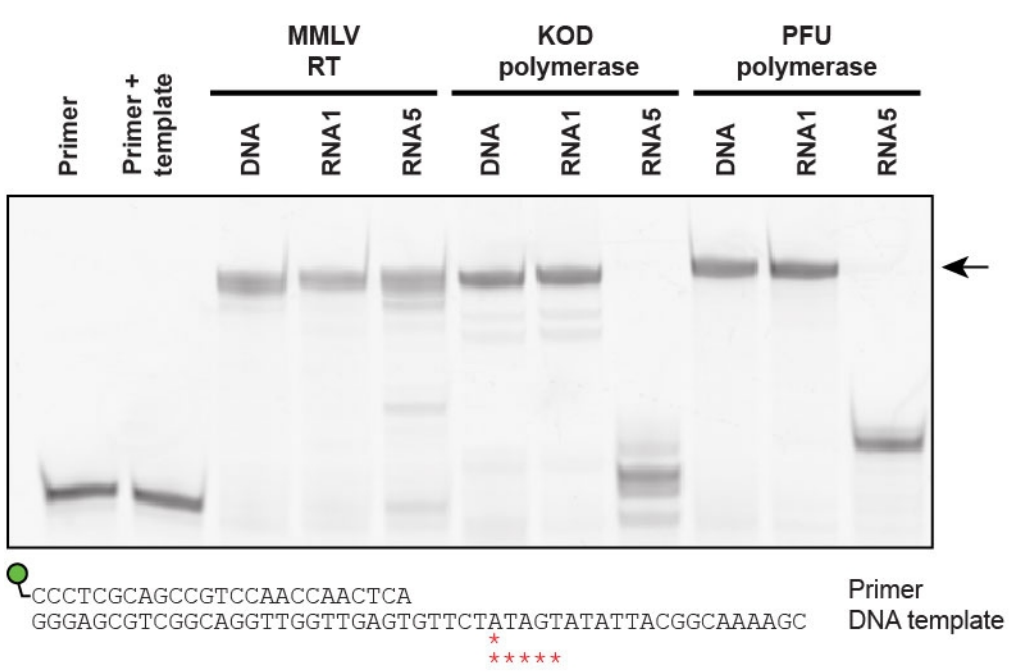
- Surfing sequence space
- Letting computers take the lead

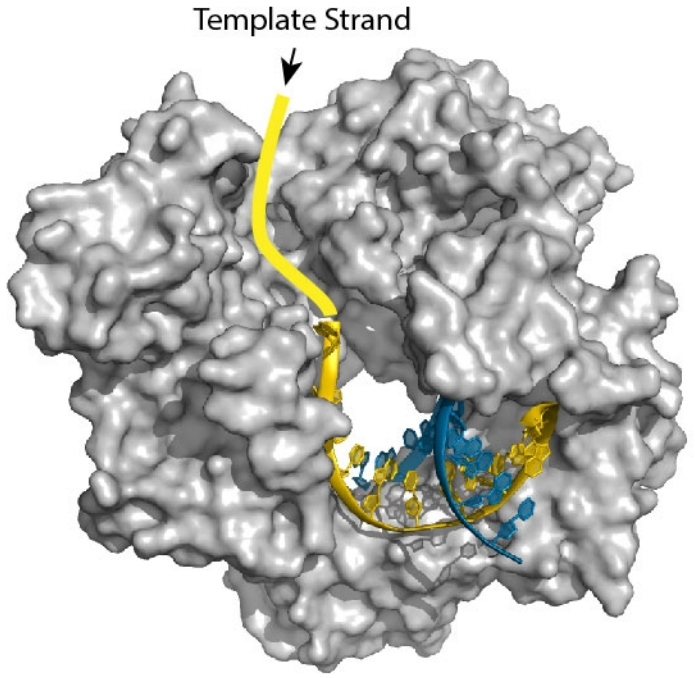
A new lineage for reverse transcriptases



Jared
Ellefson

KOD is sensitive to RNA Templates

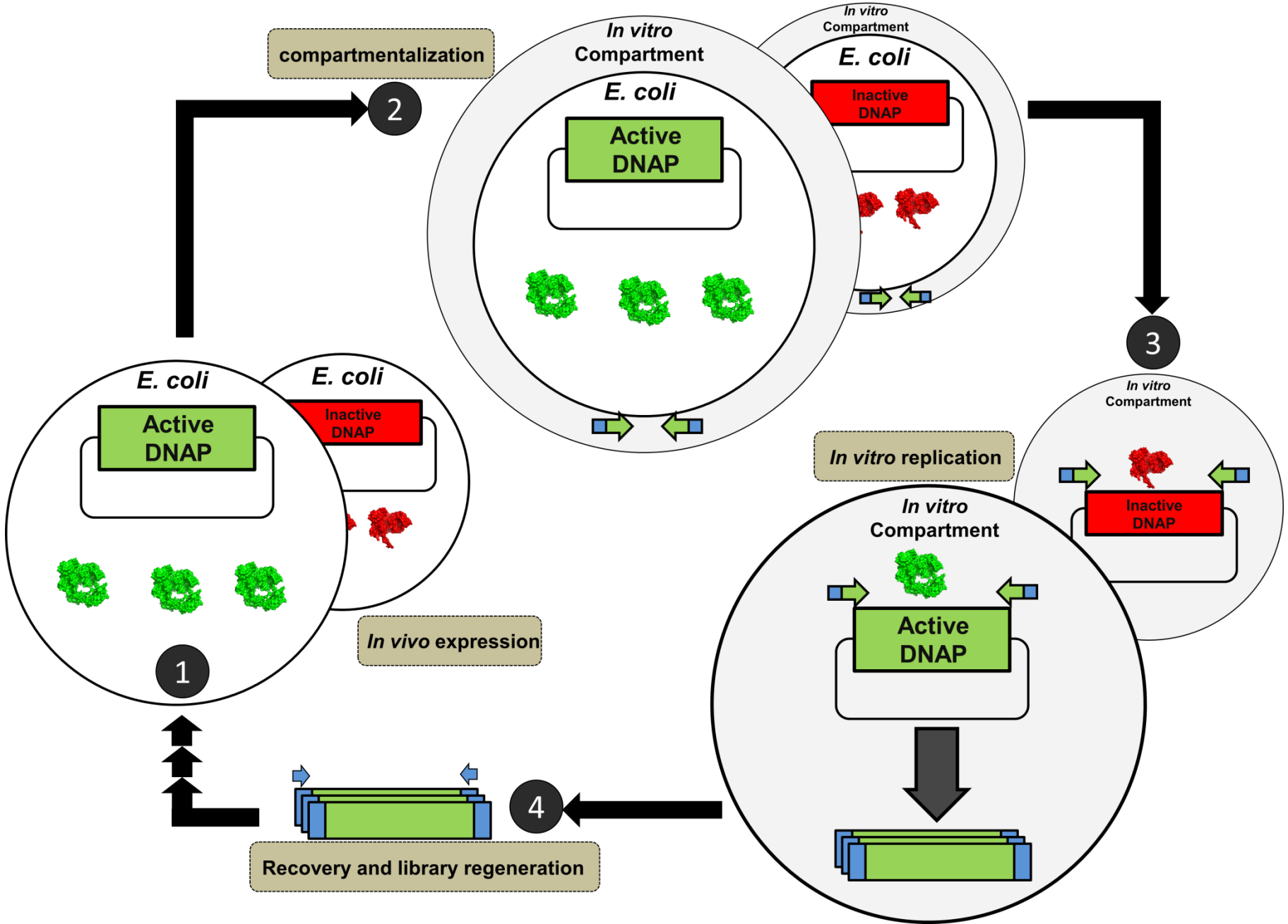




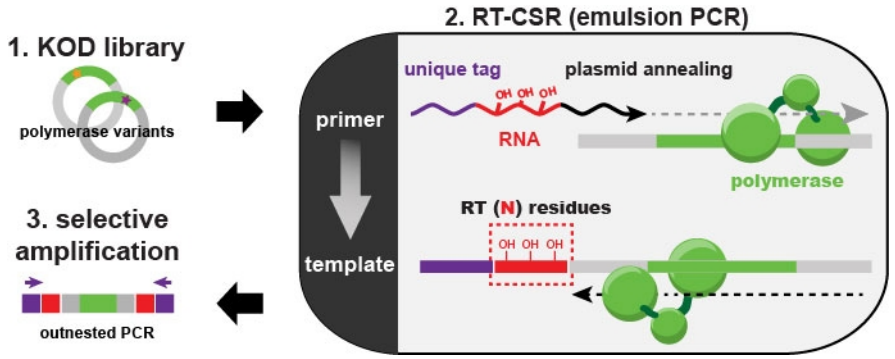
Essentially no initial activity

Complex template-polymerase interface = no rational library design

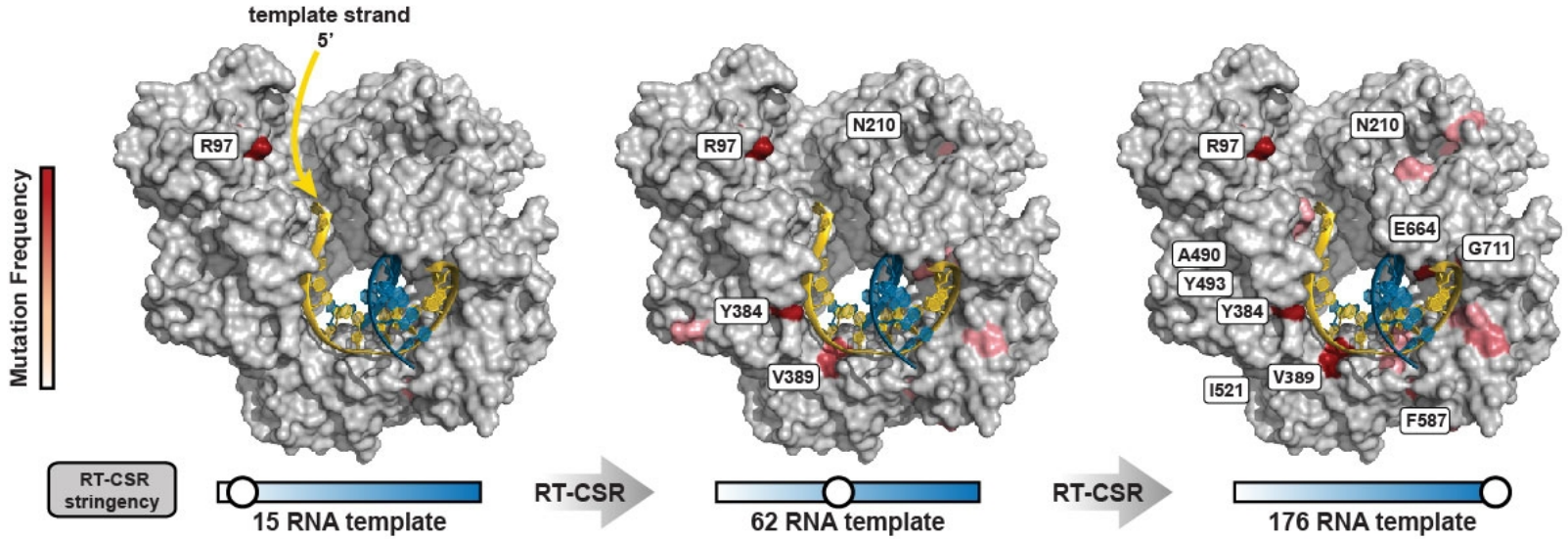
Compartmentalized self-replication (CSR; Phil Holliger)



Transforming KOD Polymerase into an Efficient Reverse Transcriptase



Increased Stringency by Increasing # of RNA bases in Primers

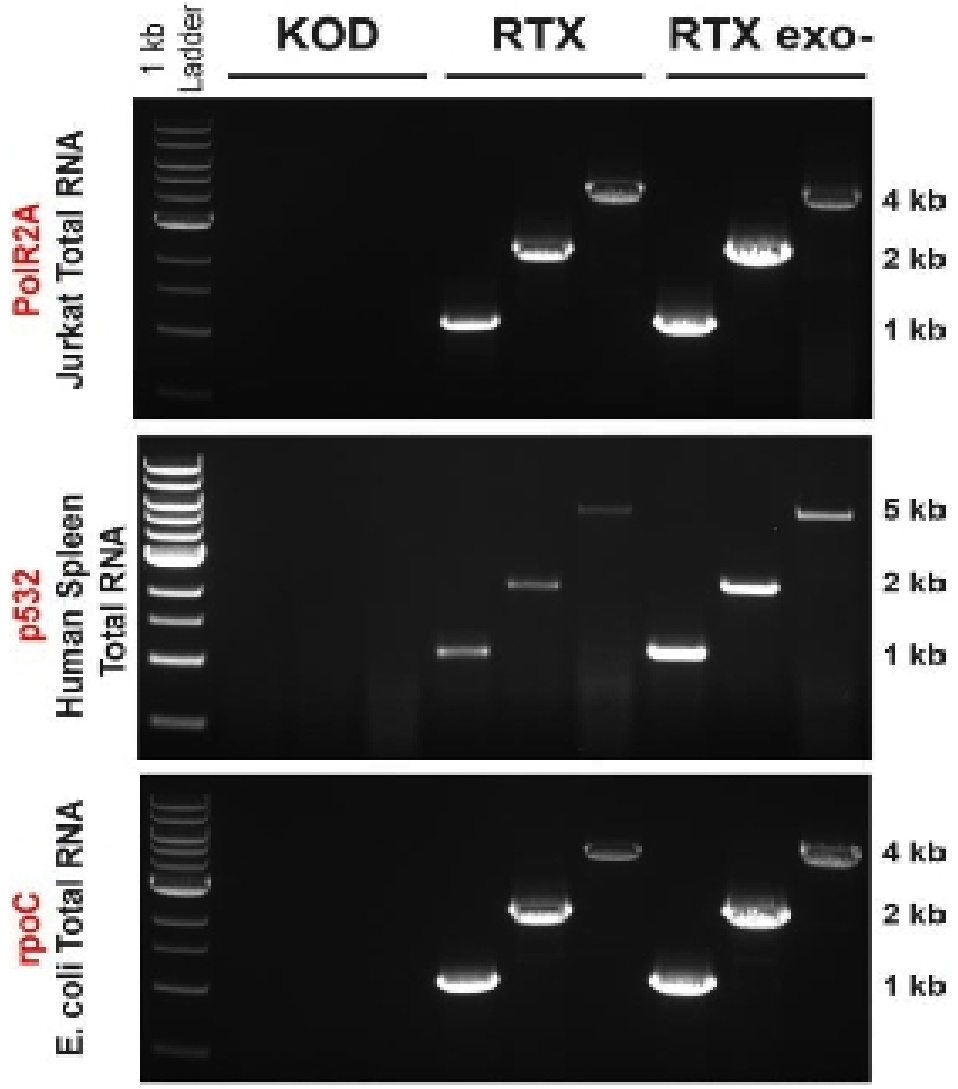


Next-Gen Sequencing of Libraries Recapitulates Evolutionary History

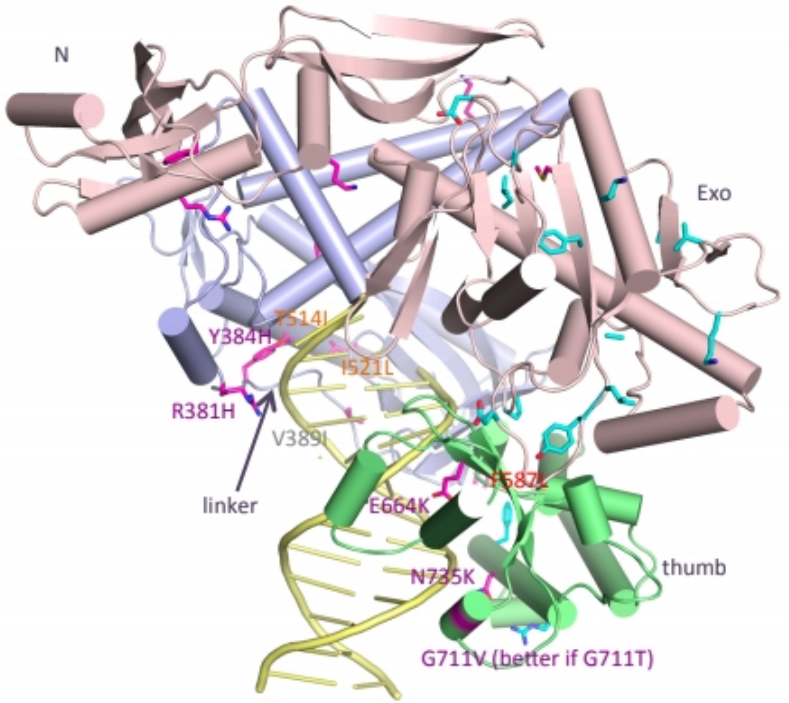
in vitro (RT-CSR) Selection Process

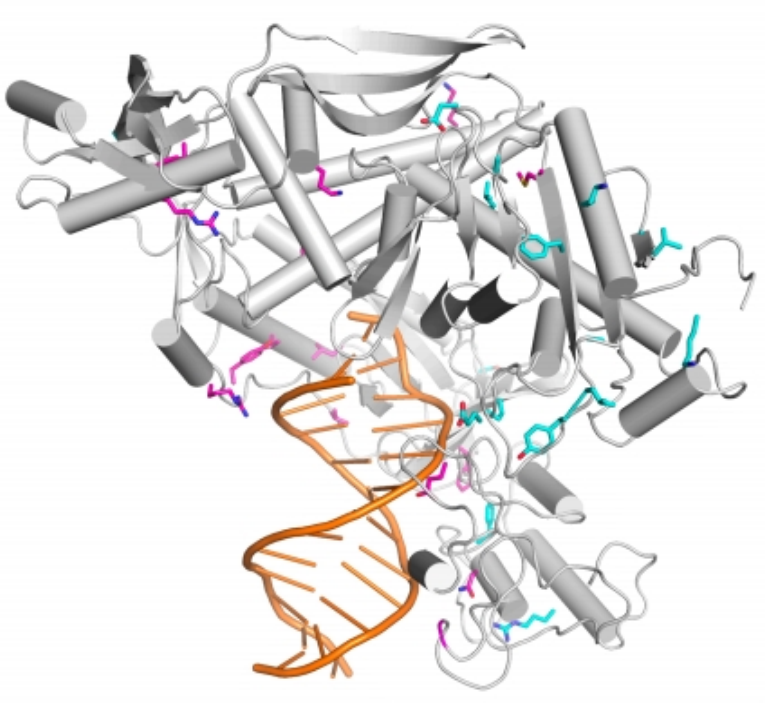
RT-CSR Round	Challenge RNA
1	10
2	10
3	15
4	15
5	20
6	20
7	25
8	25
9	62
10	62
11	100
12	100
13	100
14	100
15	100
16	100
17	176
18	176

From RNA directly to dsDNA, via PCR



What has evolution actually done?





Wei Yang, NIH

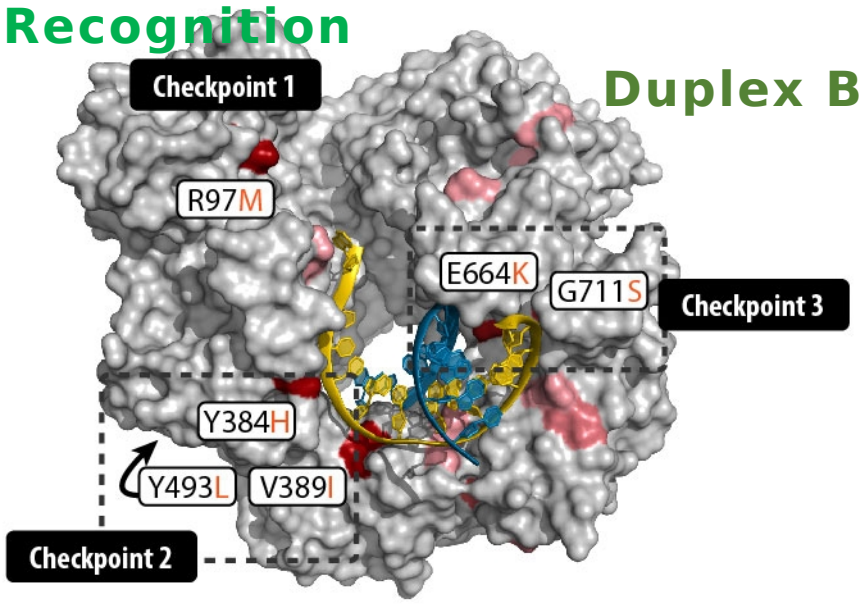
Comparison of RTX (left) and KOD (right) structures co-crystallized with RNA:DNA (RTX) or DNA (KOD) templates. Relevant residues and regions leading to accommodation of the RNA template are listed.

Molecular Checkpoints in KOD Polymerase for Alternate Template Recognition (RNA)

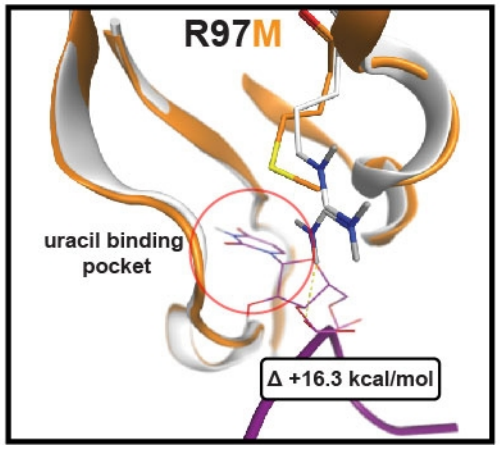
Checkpoint	Round 18			
	Amino acid Position	Mutation Frequency	Amino Acid Change	Variant Frequency
2	384		Y -> H	96.00%
1	97	93.3%	R -> A	20.80%
			R -> F	18.00%
			Other	54.50%
2	389		V -> I	91.90%
	210		N -> D	84.90%
2	493	83.3%	Y -> C	59.00%
			Y -> L	13.20%
			Y -> F	11.10%
3	664	82.7%	E -> K	60.40%
			E -> Q	22.30%
3	711	75.0%	G -> S	46.80%
			G -> V	28.20%
2	521		I -> L	59.40%
2	490		A -> T	58.50%
	587	55.1%	F -> L	36.80%
			F -> I	18.30%

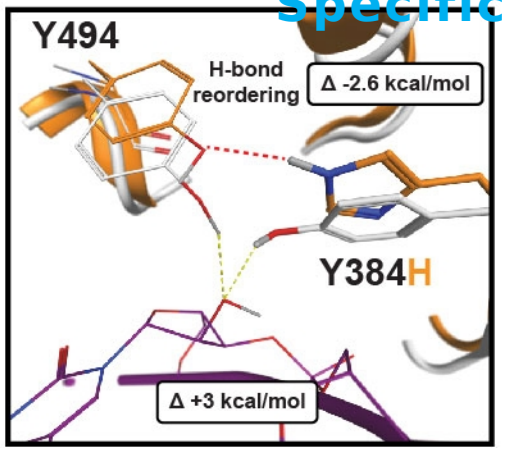
Uracil Recognition

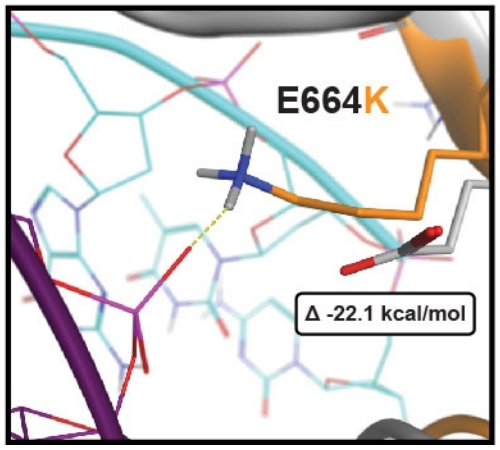
Duplex Binding



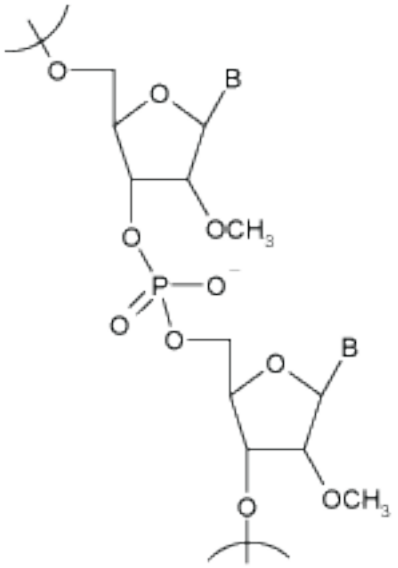
Active Site Specificity

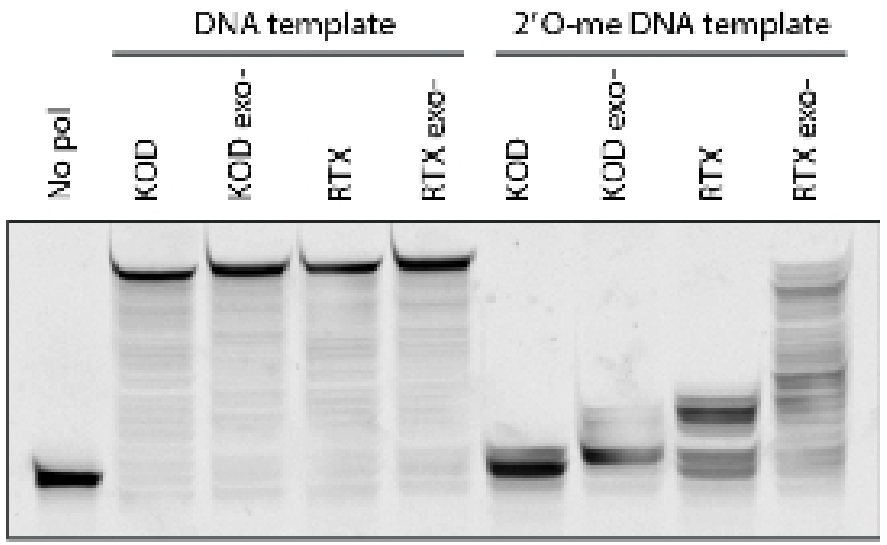






Bred for RNA … or just away from DNA?

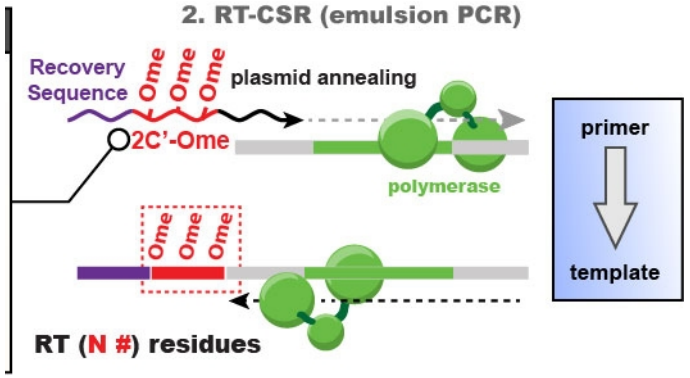


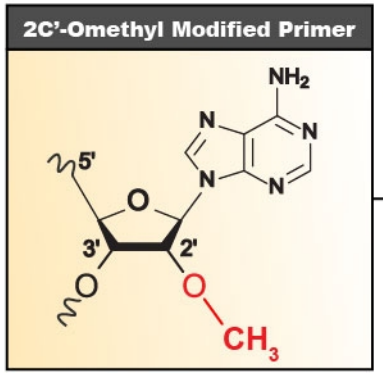


Supplementary Figure 13. Primer extension reactions on DNA and 2' O-methyl DNA substrates using KOD, KOD exo-, RTX, and RTX exo-. KOD polymerases were not capable of primer extension indicating 2' O-methyl DNA is not a substrate. RTX enzymes could polymerize across 2' O-methyl substrates, but stimulated proofreading preventing fully extended products.

We now present the possibility of a future with a RTX lineage for many XNAs.

Evolving the 2 Ome RTX Reverse Transcriptase





"Challenge" Ome RNA

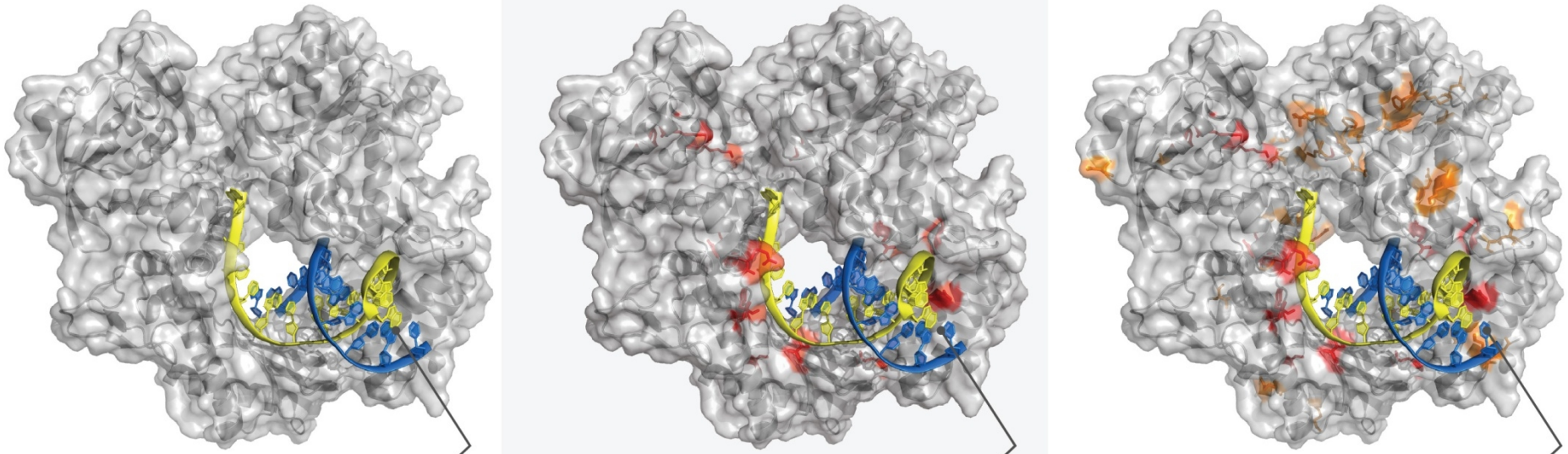
Round #	For. Primer (# Mods.)	Rev. Primer (# Mods.)	Total
1	5	5	10
2	5	5	10
3	10	5	15
4	10	10	20
5	10	10	20
6	20	10	30
7	20	10	30
8	20	20	40
9	20	20	40
10	20	52	72
11	20	52	72
12	20	52	72
13	20	52	72
14	20	52	72
15	20	52	72
16	30	52	82
17	30	52	82
18	30	52	82

Incremental Restructuring of the Template Specificity

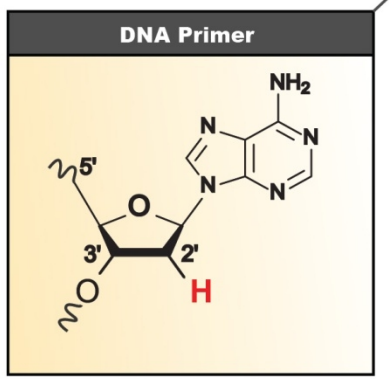
Primer
Template
RTX Mutations
Ome-RTX Mutations

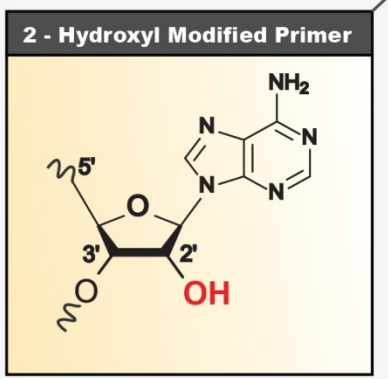
KOD (DNA) → **RTX (RNA / DNA)** → **Ome-RTX (Ome / RNA / DNA)**

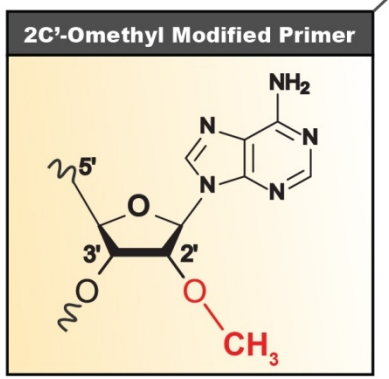
(18 Rounds) (36 Rounds)



Challenge Primer (CSR)

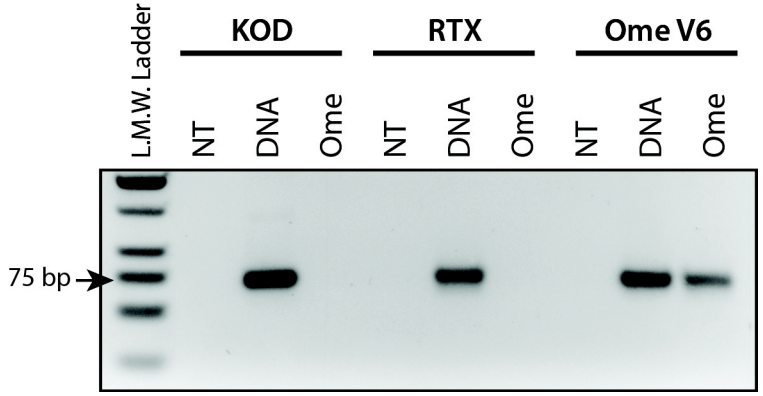






Test for RT-PCR decoding of Ome Templates

RT-PCR Test for Oligonucleotide Replication



- Ome V6 can effectively RT-PCR fully modified templates.
- Ome V6 can decode Omethyl RNA messages

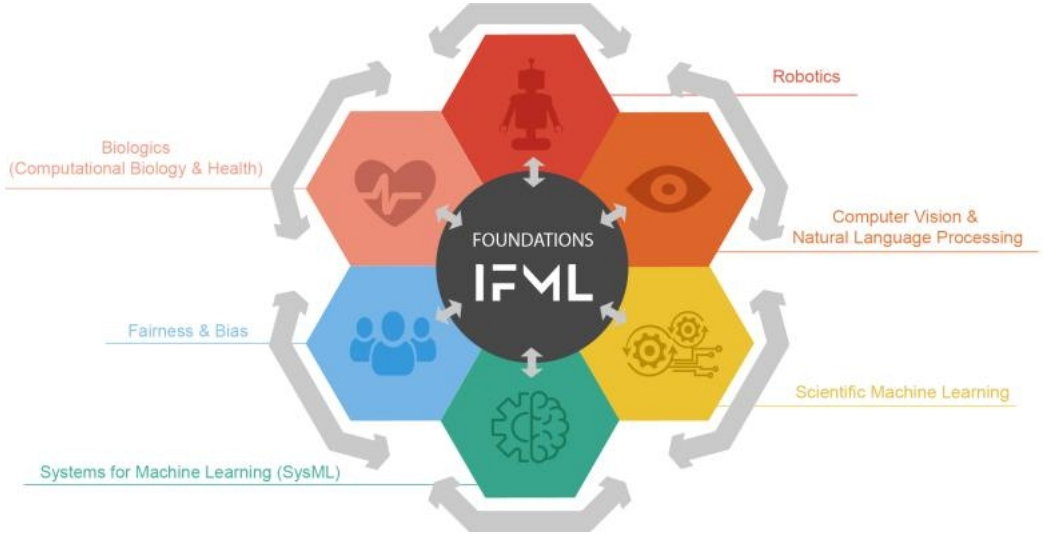
- Surfing sequence space
- Letting computers take the lead

While directed evolution is a powerful tool, it can also be a slow and cumbersome one. The ‘hunt-and-peck’ nature of mutation is fundamentally different than how a human engineer would approach the problem of making a new molecule. Enter machine learning.

The rise of AlphaZero

“In certain kinds of positions, it sees so deeply that it plays like God.” – Gary Kasparov

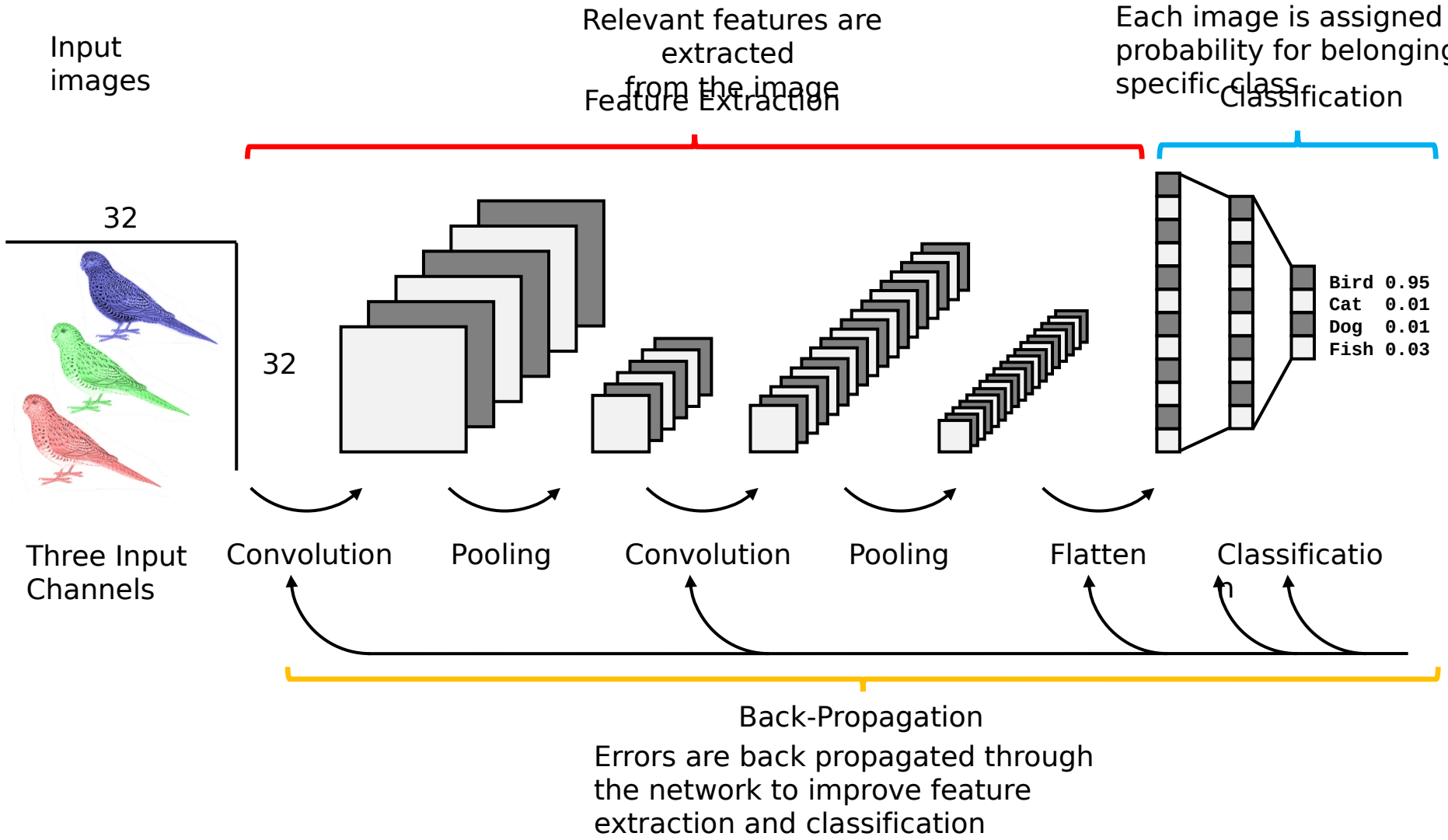
Advances in other sciences are possible in part because of the Institute for Foundations in Machine Learning (IFML)



Adam Klivans

Alex Dimakis

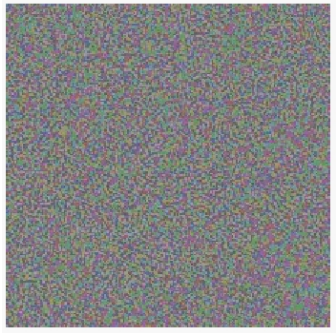
Supervised Deep Learning Framework



- essentially one big ***non-linear*** math equation with millions/billions of parameters that are optimized by minimizing the error between the correct answer and the predicted answer

Data driven feature extraction makes deep learning very powerful

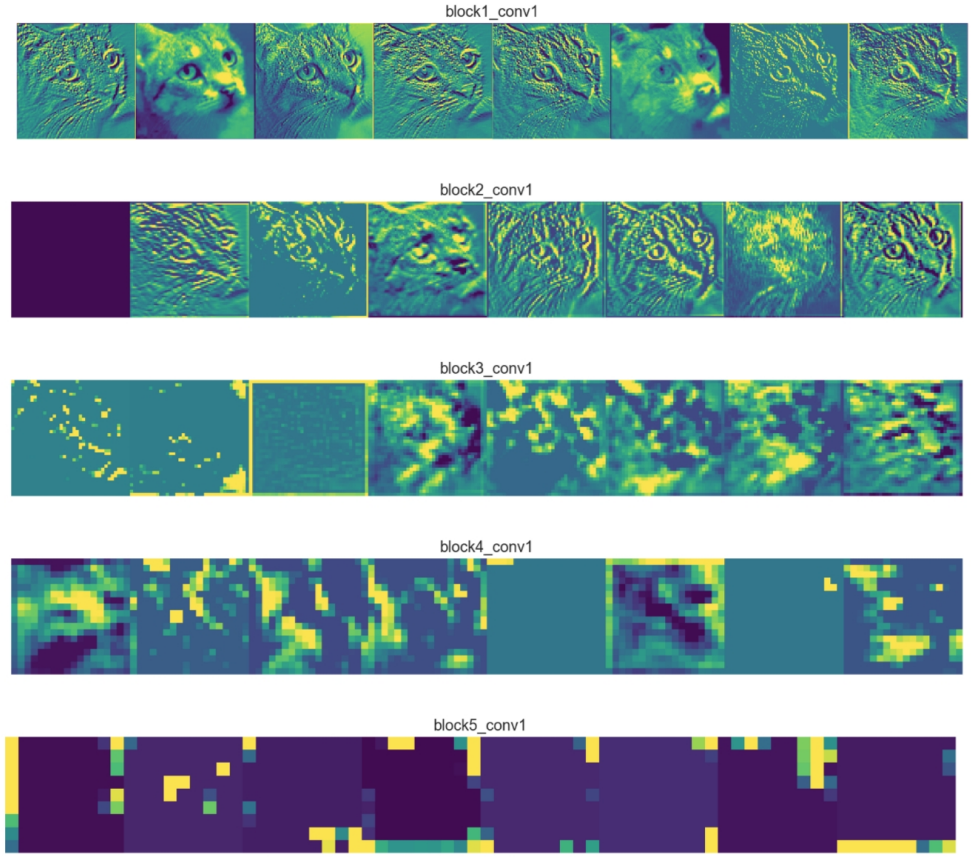
Apply Convolutional layer



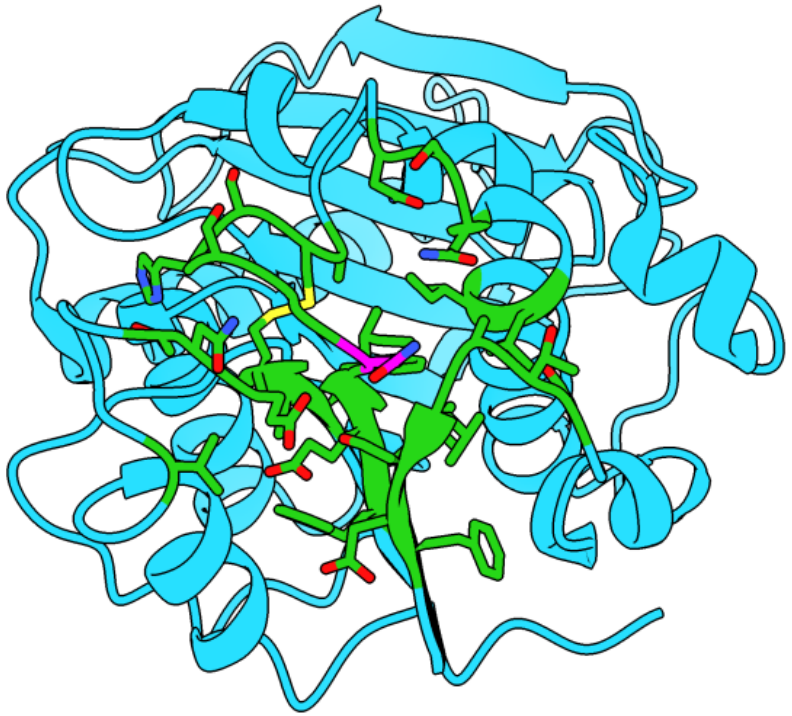
Feature Maps

Train on lots of data

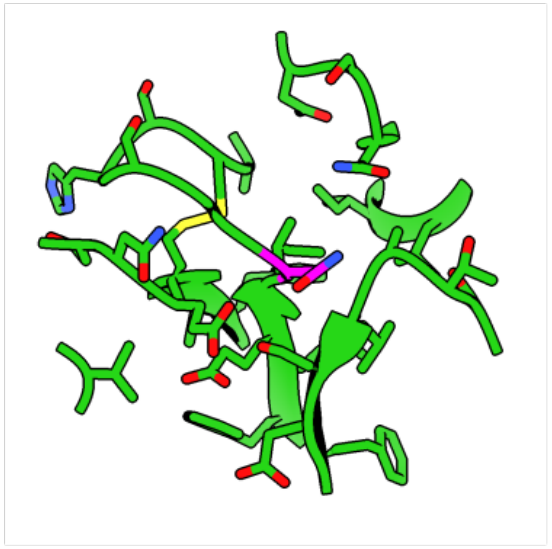
Convolutional filters learn salient features through training



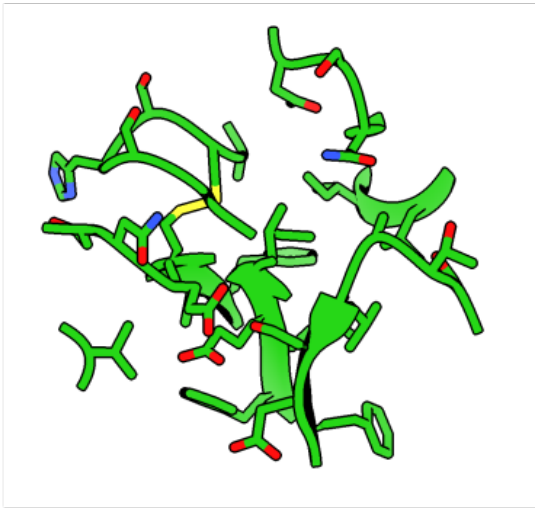
A self-supervised learning task enables evolution to teach us what proteins '*should*' look like



Center **microenvironment** around an **amino acid**



Delete **remaining** protein atoms



Masked Microenvironment

Delete centered **amino acid** and use as the **label**

Evolution provides the learning signal during model training

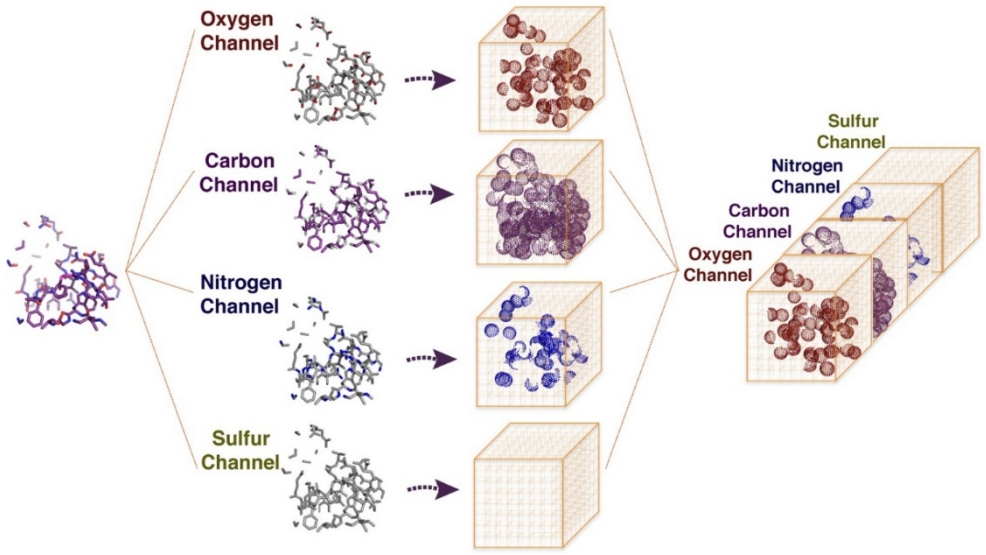
Austin Cole, Aperiam

A 3D CNN can predict amino acid identity

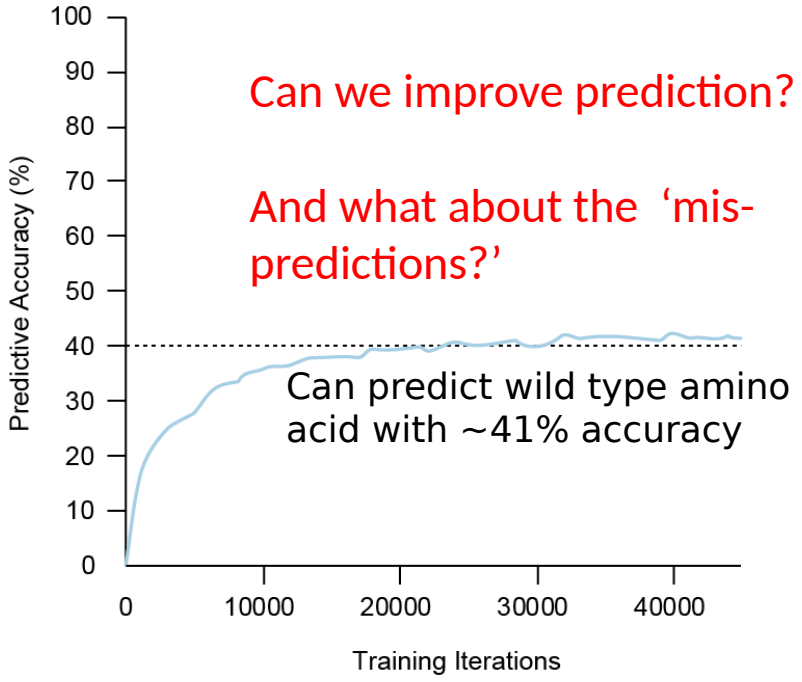
Trains a neural network to learn what residues fits given a chemical environment

- 32,760 structures used for training – 1600 for testing
- 600,000 unique environments for training
- 20 amino acid environments sampled per iteration (~20,000 per epoch)

Raghav Shroff

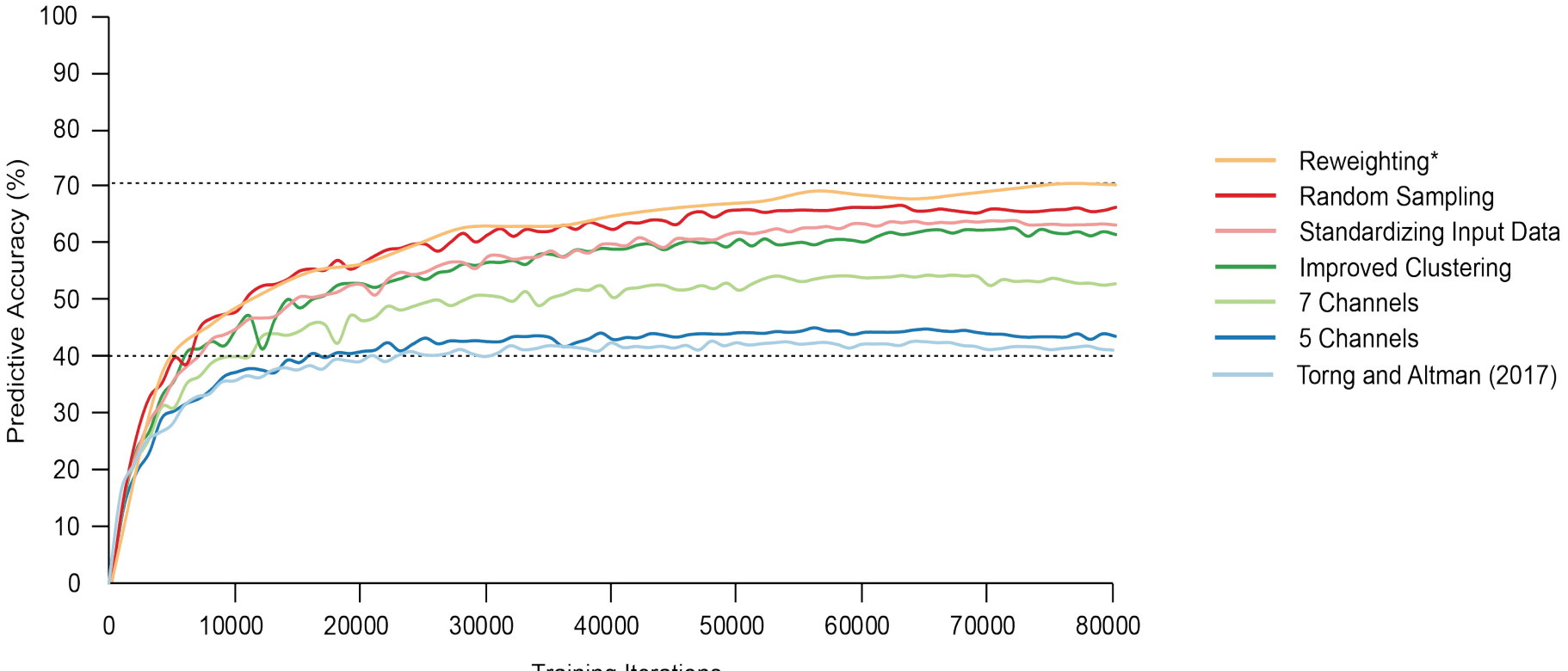


Adapted from Torng and Altman *BMC Bioinf.* (2017)

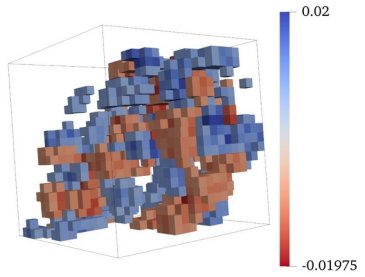


Improving predictive accuracy lowers the sequence space

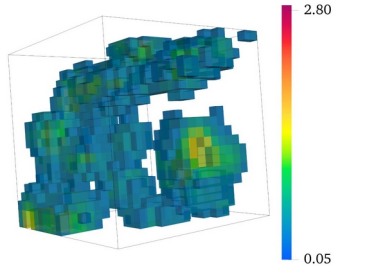
1. Use the Torng and Altman model as a starting point
2. (5 channels) Add hydrogen channel
3. (7 channels) Add partial charges and solvent accessibility channels
4. (Improved Clustering) Cluster input sequences to 50% similarity
5. (Standardize Input Data) Use refined and rebuilt protein structures from pdbRedo
6. (Random Sampling) Randomly sample residues in input proteins rather than spatially sample
7. (Reweighting) Bias residues towards natural frequencies



Physical Channels

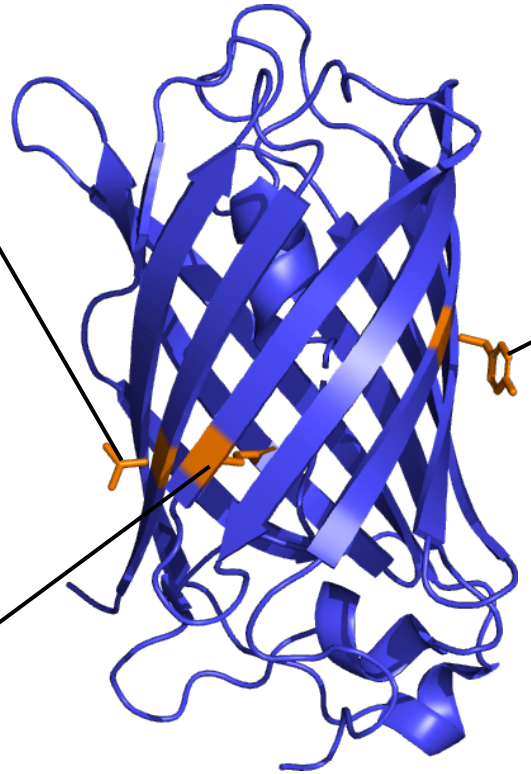


Partial Charge



Solvent Accessibility

How we leverage ML models to guide protein engineering



```
WT Residue = Val 216

P(Ser | Context) ~ 0.20
P(Val | Context) ~ 0.20
P(Ala | Context) ~ 0.14
```

ML Model Thinks Many Residues May Fit

```
WT Residue = Tyr 14

P(Phe | Context) ~ 0.77
P(Tyr | Context) ~ 0.14
P(His | Context) ~ 0.05
```

An Aromatic Probably Belongs Here

```
WT Residue = Gln 39

P(Arg | Context) ~ 0.68
P(Lys | Context) ~ 0.30
P(His | Context) ~ 0.01
```

Gln Probably Does not Belong Here

Mutations at residues like Gln 39 could improve a diverse set of protein functions.

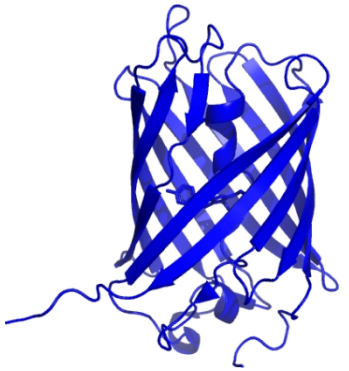
MutCompute can flag positions in the protein string that are likely contributing to instability

BFP: Sites predicted by the NN yield stabilizing mutants

Blue fluorescent proteins have a long history optimization for brightness, solubility and folding.

Can we use our neural network to improve secBFP2?

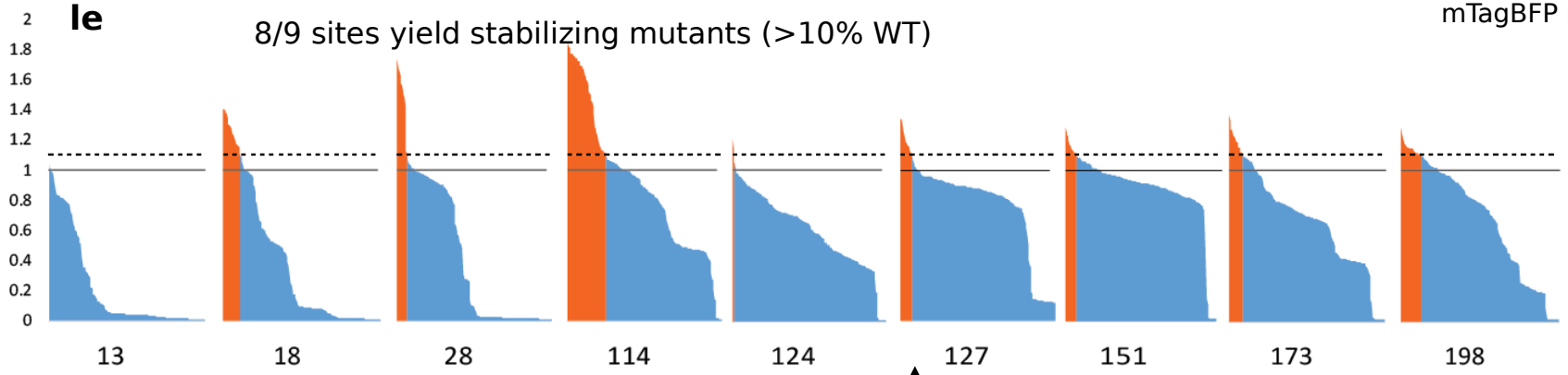
Selected residues with the lowest wild-type probability, built NNS libraries, and assayed approximately 200 variants. Sequenced the highest variants.



mTagBFP PDB: 3M24

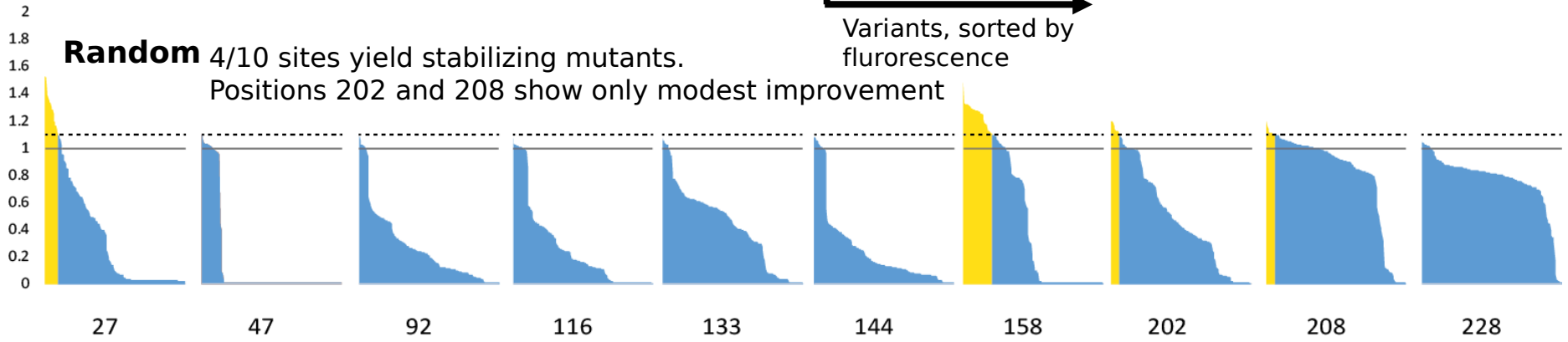
Unfavorable

8/9 sites yield stabilizing mutants (>10% WT)



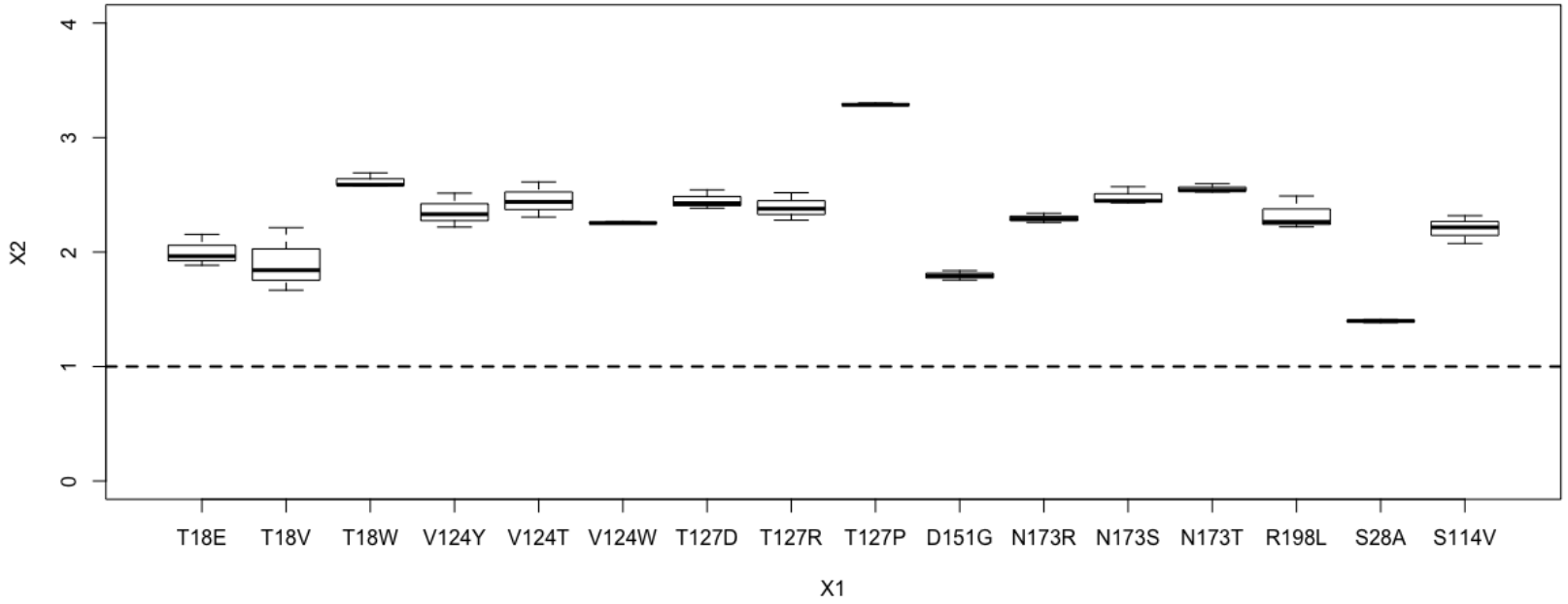
Does this outperform randomly choosing sites?

Random 4/10 sites yield stabilizing mutants.
Positions 202 and 208 show only modest improvement

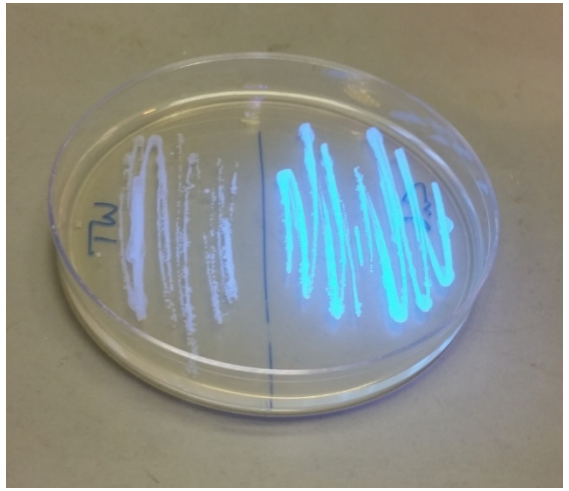


3D CNN stabilizing mutants can be combined for greater effect

- While the effects of stabilizing mutations are typically modest, they are usually additive



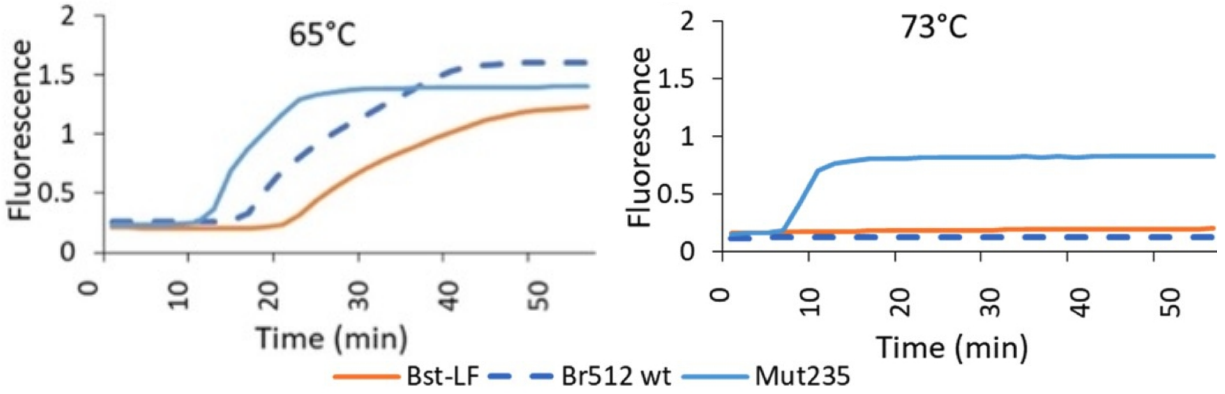
Wild-type



Bluebonnet:
Combining 8 mutations

MutCompute guided the thermal stabilization of a polymerase for single temperature COVID19 diagnostic applications

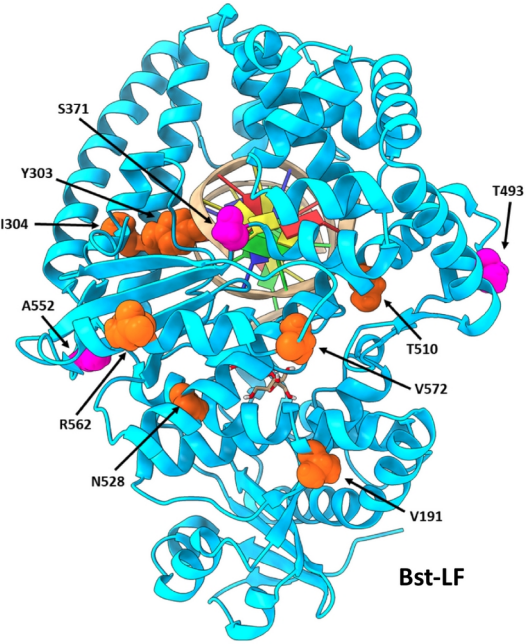
LAMP-OSD Assays



Left shift: protein more thermostable/active
Right shift: protein is unfolding/inactive

MutCompute Top 10 WT Mispredictions

Name	Predicted Mutcompute Mutations (WT-Pred)	Wild Type Amino Acid Probability	Predicted Amino Acid Probability
Mut1	V 191 L	0.001	0.738
Mut2	T 493 N	0.001	0.85
Mut3	A 552 G	0.004	0.996
Mut4	R 562 V	0.004	0.58
Mut5	S 371 D	0.01	0.872
Mut6	N 528 E	0.014	0.468
Mut7	T 510 F	0.017	0.946
Mut8	I 304 V	0.018	0.981
Mut9	Y 303 H	0.019	0.522
Mut10	V 572 A	0.019	0.876

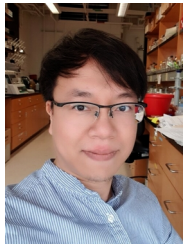


LAMP-OSD:

- *Isothermal* nucleotide amplification technique
- rivals speed and sensitivity of PCR
- **Does not require thermal cycling and associated instrumentation**
- More convenient for clinical and field use

ML-designed polymerase (Mut235) enabled single temperature COVID19 diagnostic in under ~20 minutes (and as little as ~10 minutes)

Andre Maranhao, PhD



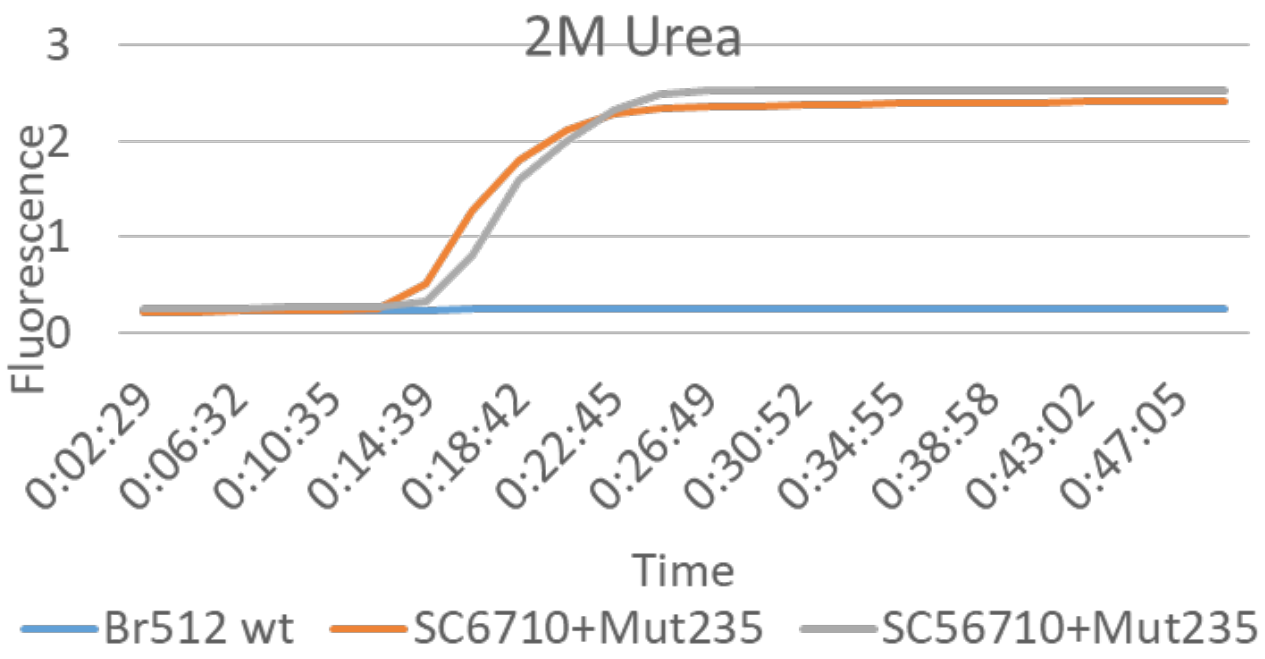
Inyup Paik, PhD

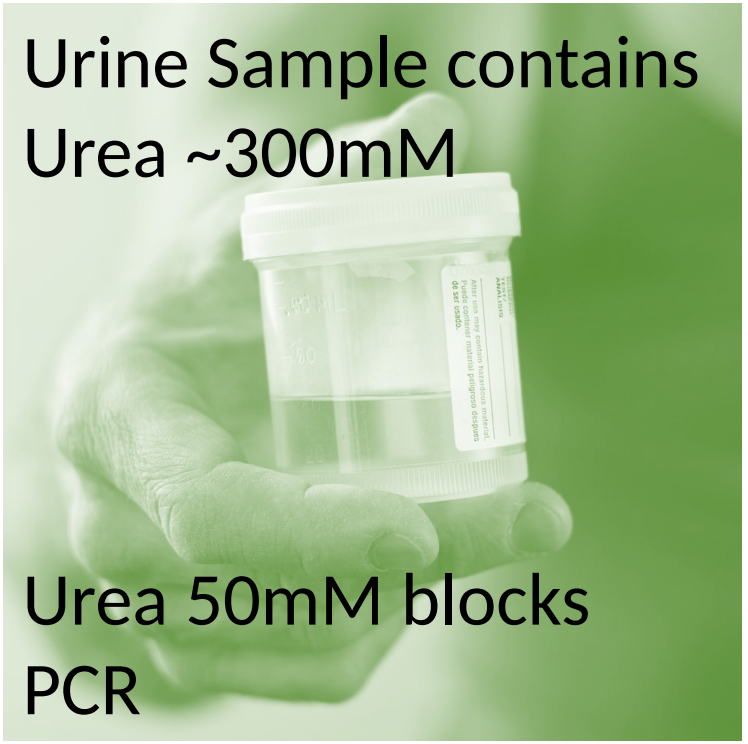
Sanchita Bhadra, PhD

Paik et al. ACS Biochemistry **2021**

MutCompute predictions available at https://mutcompute.com/view/3tan

Combined Variants are Inhibitor Resistant



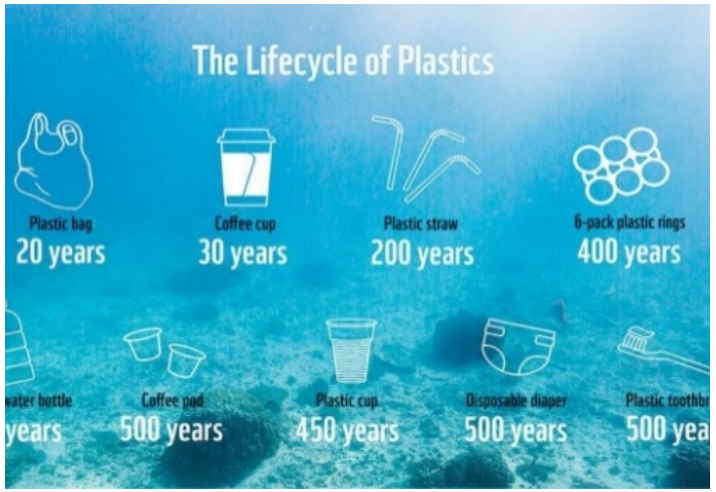


Ellington Lab's Distribution Efforts (06.01.2020 ~ Current)

Plastic pollution is a global problem

"Every minute, the equivalent of one garbage truck of plastic is dumped into our ocean."

- United Nations Environment Programme



Plastic was invented because it's durability

~12 million tonnes/year entered the ocean

Now we eat microplastics, yay

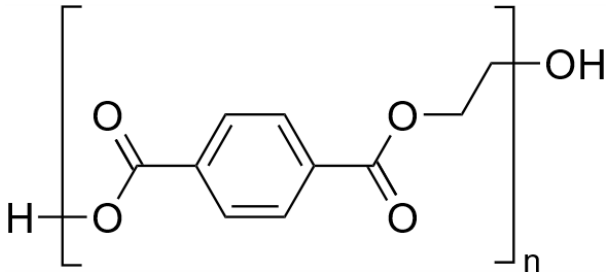
It took nature ~60M years to learn how to efficiently breakdown wood and end the Carboniferous period. With machine learning, can we accelerate this process for plastic into a few years?

Hal Alper, ChemE

https://earth.org/plastic-pollution-statistics/

Turning to Nature:
Enzymatic PET depolymerization

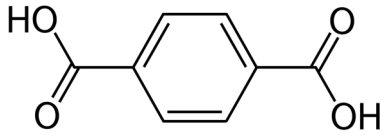
- PETase: a PET hydrolase enzyme first discovered in *Ideonella sakaiensis* in 2016
- Cutinase: Cutin hydrolase enzyme also capable of depolymerization of PET
- 48% sequence similarity between the two scaffolds



PET

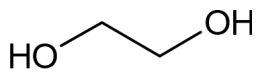
plastic

PETase or **Cutinase**



TPA

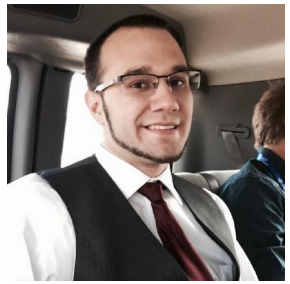
+



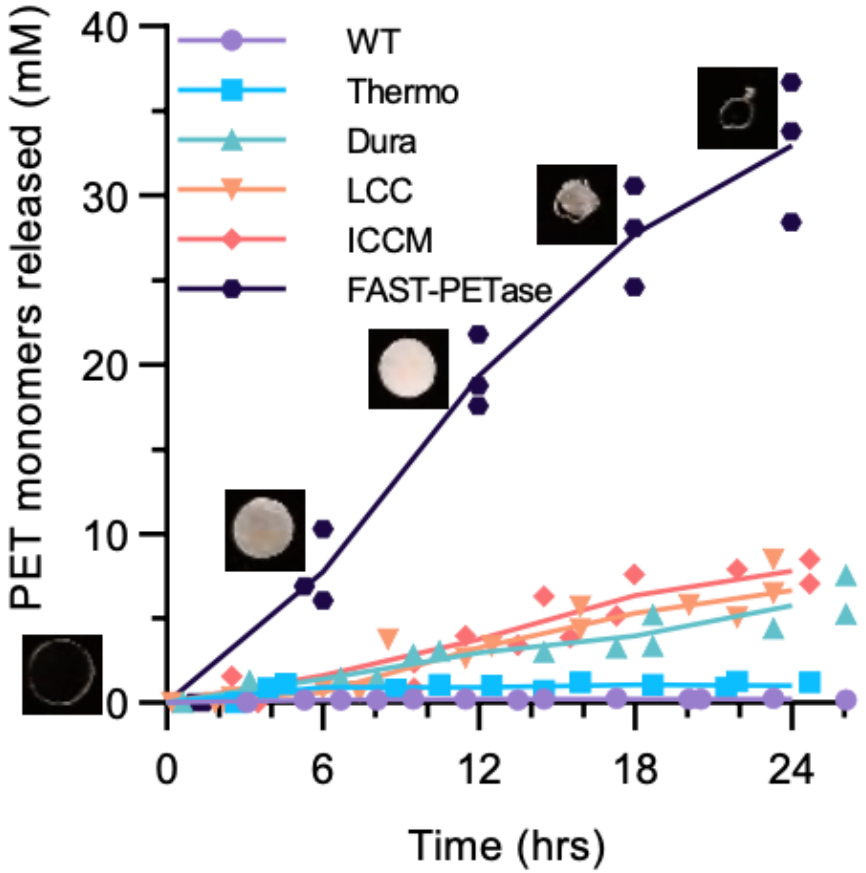
EG

Raw materials

MutCompute designed variant outperforms the literature on a PET depolymerization



Danny Diaz,
Ellington Lab



Hongyuan Lu, PhD

MutCompute predictions available at https://mutcompute.com/view/6ij6

Visualize FAST-PETase at https://mutcompute.com/view/7sh6

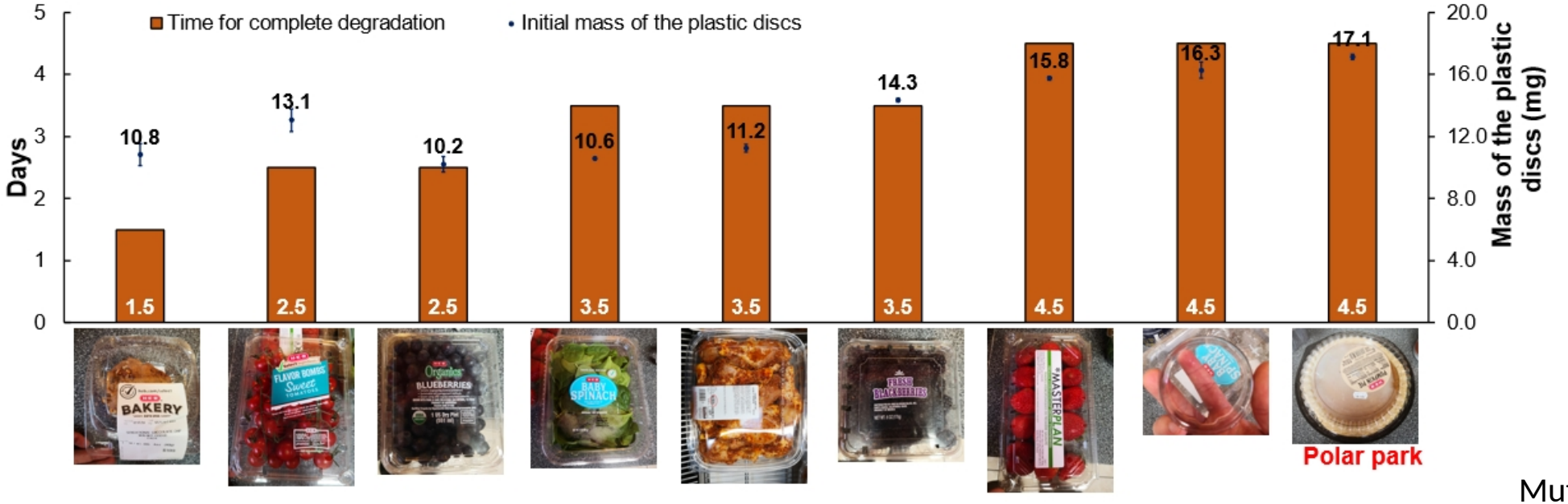
***FAST-PETase: S121E /R224Q/N233K (All 3 predicted by MutCompute)**

***MutCompute designed variants displayed significantly improved protein expression yield (data in supplementary slide)**

ThermoPETase: Son et al. ACS Catalysis (2019)
DuraPETase: Cui et al. ACS Catalysis (2021)
Cutinase Engineering (LCC and ICCM): Tournier et al. Nature (2020)

H. Lu, **D. J. Diaz**, N. J. Czarnecki, C. Zhu, W. Kim, R. Shroff, D. J. Acosta, B. Alexander, H. Cole, Y. J. Zhang, N. Lynd, A. D. Ellington, H. S. Alper
Machine learning-aided engineering of hydrolases for PET depolymerization. (2022) Nature, in press.

With MutCompute, we engineered *FAST-PETase* that can achieve 100% degradation of retail PET in days



Pink is the anomalous chemistry MutCompute identified

MutCompute predictions available at https://mutcompute.com/view/6ij6

H. Lu, **D. J. Diaz**, N. J. Czarnecki, C. Zhu, W. Kim, R. Shroff, D. J. Acosta, B. Alexander, H. Cole, Y. J. Zhang, N. Lynd, A. D. Ellington, H. S. Alper
Deep Learning redesign of PETase for practical PET degrading applications. (2021) Nature, in review.

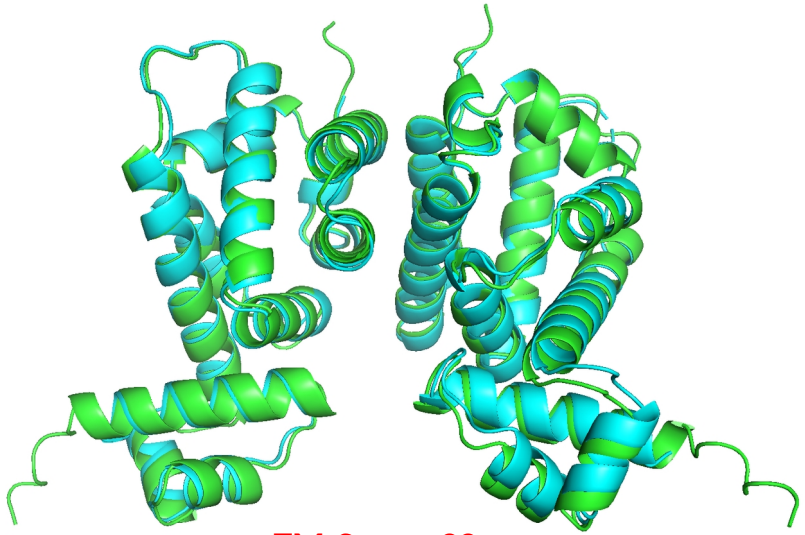
PET degradation time-lapse

Sourced from Walmart
48 hour time lapse at 50C (122F)

Synergize MutComputeX with AlphaFold and Docking for Substrate Specificity Engineering

10 mutations in sequence space computationally modeled

Blue is experimental **Green is AlphaFold**



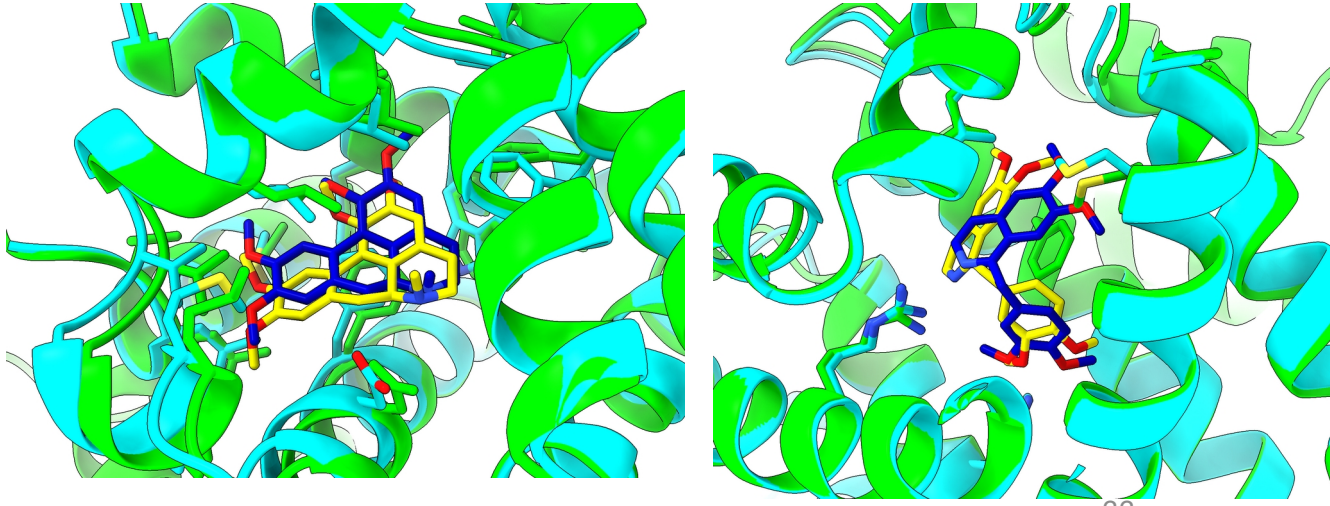
TM-Score: 92

Workflow:

- Alphafold a protein variant
- Sample ligand conformer space
- dock a library of ligand conformers with AI
- Design ligand specific libraries with MutComputeX
- Directed Evolution/Site Directed Mutagenesis Experiments
- Repeat

Apply to Transcription Factors and Enzymes

Simon d'Oelsnitz, PhD



Experimental ligand **AI-docked ligand**

Active Site Enzyme Engineering Without a Structure I

Enzyme: Methyl Transferase

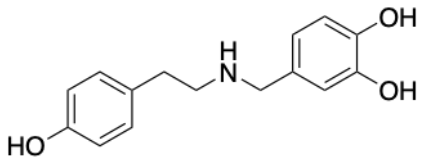
Previous Attempts:

- Error Prone PCR failed to provide any improved variants

AI Pipeline:

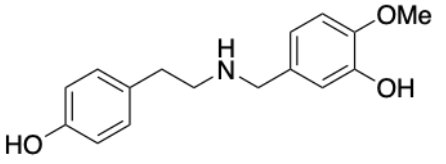
- AlphaFold protein
- AI dock SAM cofactor
- AI dock substrate
- Generate mutational designs with MutComputeX
- Screen Variants
- Stack gain of function variants

Trying to make 4-OMe Norbelladine

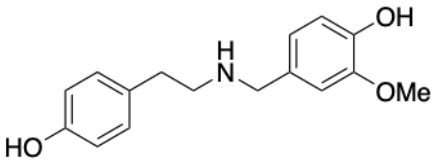


Substrate

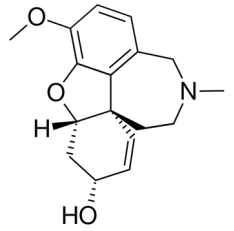
Enzyme
SAM cofactor



Major Product



Minor Product



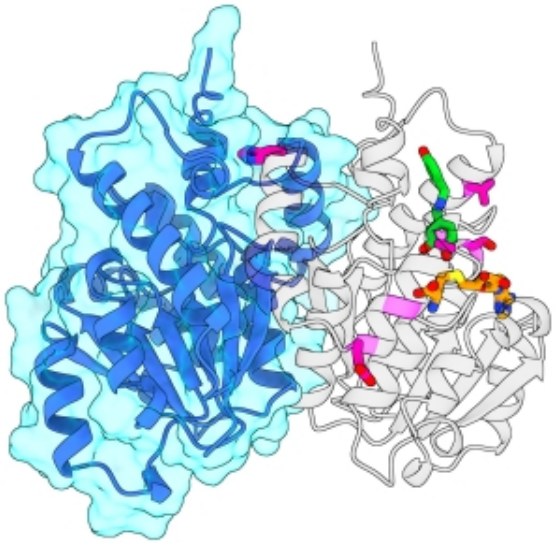
Galantamine:
FDA Approved drug to treat Alzheimer

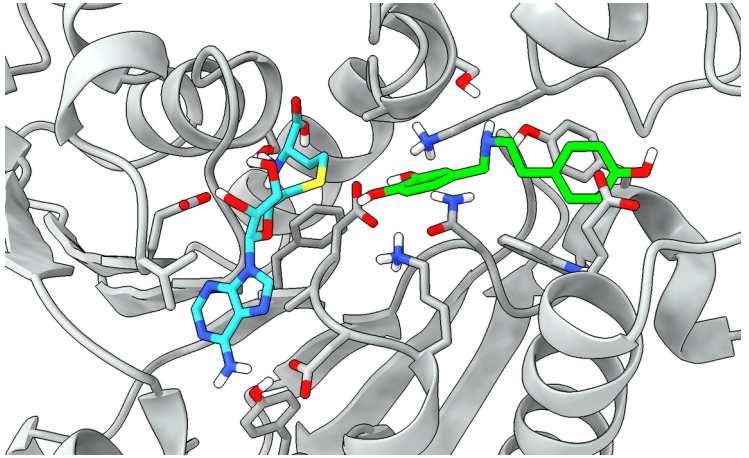
And not make: 3-OMe Norbelladine

Matt Minus, PhD

Simon d'Oelsnitz, PhD

James Howard



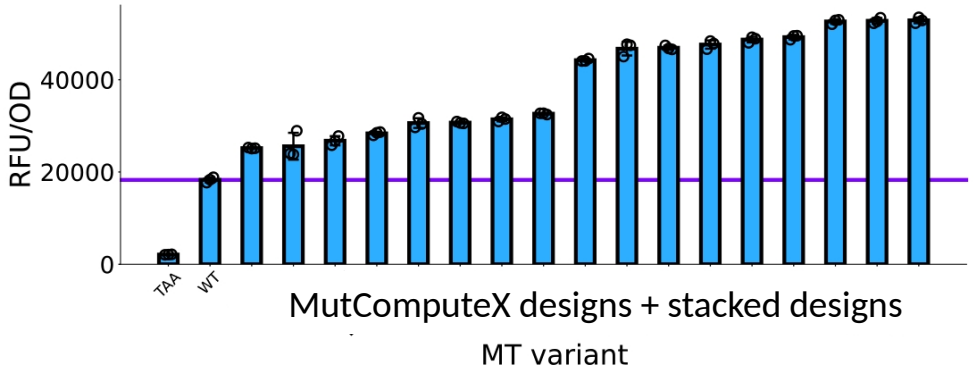


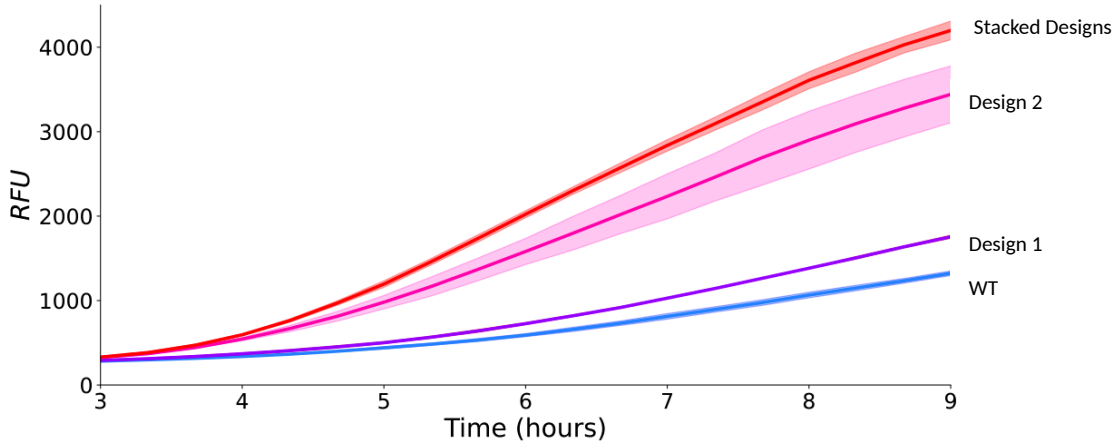
Active Site Enzyme Engineering Without a Structure II

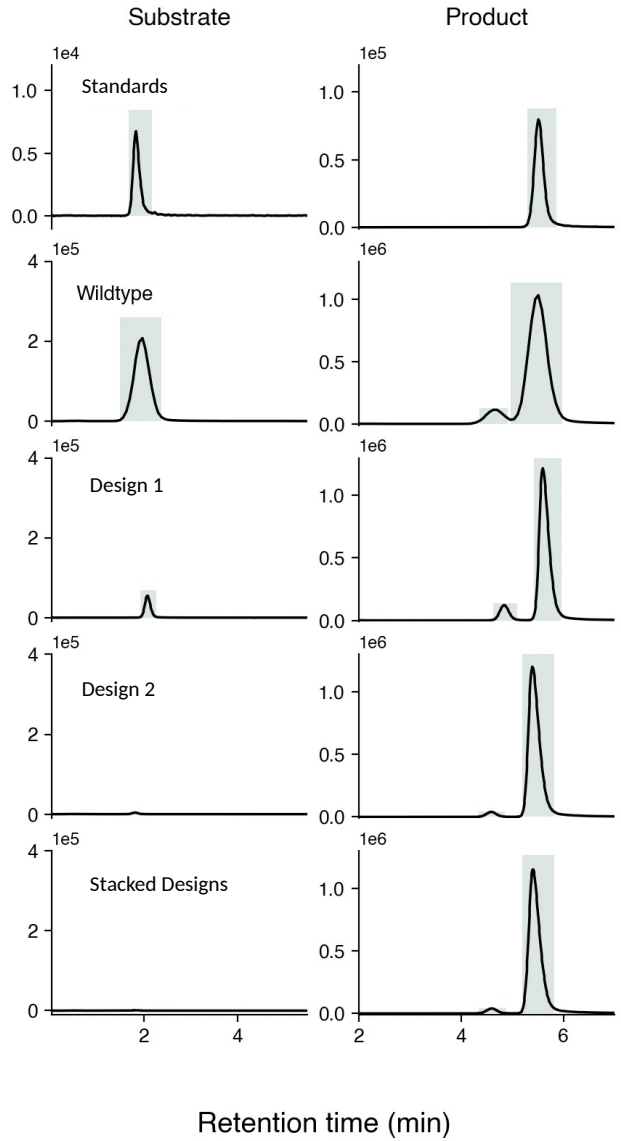
Provided 22 mutagenesis designs, 7 of improved enzyme activity

Conclusion:

- Improved activity of Methyl Transferase by 3X with active site mutations without an experimentally solved structure
- Currently writing manuscript







Conclusions

- Directed evolution is still excellent at evaluating entire structures / functions, especially where many mutations may be required to attain a given phenotype
- Even so, directed evolution will be largely displaced by machine learning coupled to synthetic biology (DBTL) approaches
- Increasingly, there will be no requirement for solved protein structures in order to carry out engineering campaigns
- Increasingly, there will be no requirement for deep chemical or biological understanding in order to carry out engineering campaigns

Acknowledgements

Computational:

- James Loy, PhD
- Raghav Shroff, PhD
- Chengyue Gong

Experimentalists:

- Ebru Cayir, PhD
- Simon d'Oelsnitz, PhD
- Matt Minus, PhD
- James Howard
- Alper lab, Hong Lu

Funding:

- DTRA
- Exxon
- NASA
- NIH
- Welch Foundation

Directed evolution:

RTX: Jared Ellefson, Raghav Shroff

T7 RNAP: Adam Meyer

IFML Institute for Foundations of MACHINE LEARNING

AMD

Adam Klivans