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
KOD is sensitive to RNA Templates

Essentially no initial activity

Complex template-polymerase interface = no rational library design
Increased Stringency by Increasing # of RNA bases in Primers

Next-Gen Sequencing of Libraries Recapitulates Evolutionary History
From RNA directly to dsDNA, via PCR
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)

<table>
<thead>
<tr>
<th>Amino acid Position</th>
<th>Mutation Frequency</th>
<th>Amino Acid Change</th>
<th>Variant Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>384</td>
<td>93.3%</td>
<td>Y -&gt; H</td>
<td>96.00%</td>
</tr>
<tr>
<td>97</td>
<td>93.9%</td>
<td>R -&gt; A</td>
<td>20.80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R -&gt; F</td>
<td>18.00%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other</td>
<td>54.50%</td>
</tr>
<tr>
<td>389</td>
<td>83.3%</td>
<td>V -&gt; I</td>
<td>91.90%</td>
</tr>
<tr>
<td>210</td>
<td>84.3%</td>
<td>N -&gt; D</td>
<td>84.90%</td>
</tr>
<tr>
<td>493</td>
<td>80.0%</td>
<td>Y -&gt; C</td>
<td>59.00%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y -&gt; L</td>
<td>13.20%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y -&gt; F</td>
<td>11.10%</td>
</tr>
<tr>
<td>664</td>
<td>82.7%</td>
<td>E -&gt; K</td>
<td>60.40%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E -&gt; Q</td>
<td>22.30%</td>
</tr>
<tr>
<td>711</td>
<td>75.0%</td>
<td>G -&gt; S</td>
<td>46.80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G -&gt; V</td>
<td>28.20%</td>
</tr>
<tr>
<td>521</td>
<td>64.0%</td>
<td>I -&gt; L</td>
<td>59.40%</td>
</tr>
<tr>
<td>490</td>
<td>58.50%</td>
<td>A -&gt; T</td>
<td>58.50%</td>
</tr>
<tr>
<td>587</td>
<td>55.1%</td>
<td>F -&gt; L</td>
<td>36.80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F -&gt; I</td>
<td>18.30%</td>
</tr>
</tbody>
</table>

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 prevents 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

<table>
<thead>
<tr>
<th>Round #</th>
<th>For. Primer (# Mods.)</th>
<th>Rev. Primer (# Mods.)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>52</td>
<td>72</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>52</td>
<td>72</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>52</td>
<td>72</td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>52</td>
<td>72</td>
</tr>
<tr>
<td>14</td>
<td>20</td>
<td>52</td>
<td>72</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>52</td>
<td>72</td>
</tr>
<tr>
<td>16</td>
<td>30</td>
<td>52</td>
<td>82</td>
</tr>
<tr>
<td>17</td>
<td>30</td>
<td>52</td>
<td>82</td>
</tr>
<tr>
<td>18</td>
<td>30</td>
<td>52</td>
<td>82</td>
</tr>
</tbody>
</table>
Incremental Restructuring of the Template Specificity

KOD (DNA) \[\rightarrow\] RTX (RNA / DNA) \[\rightarrow\] Ome-RTX (Ome / RNA / DNA)

(18 Rounds) \[\rightarrow\] (36 Rounds)
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.

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

The rise of AlphaZero
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

Input images

Three Input Channels

Convolution

Pooling

Convolution

Pooling

Flatten

Classification

- Relevant features are extracted from the image
- Each image is assigned a probability for belonging to a specific class
- Errors are back propagated through the network to improve feature extraction and classification
- 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

Bird 0.95
Cat 0.01
Dog 0.01
Fish 0.03
Data driven feature extraction makes deep learning very powerful.

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

Delete centered amino acid and use as the label

Evolution provides the learning signal during model training.
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)

Can predict wild type amino acid with ~41% accuracy

Can we improve prediction?
And what about the ‘mis-predictions?’

Adapted from Tormg 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
How we leverage ML models to guide protein engineering

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

**WT Residue = Val 216**

- $P(\text{Ser} \mid \text{Context}) \sim 0.20$
- $P(\text{Val} \mid \text{Context}) \sim 0.20$
- $P(\text{Ala} \mid \text{Context}) \sim 0.14$

ML Model Thinks Many Residues May Fit

**WT Residue = Tyr 14**

- $P(\text{Phe} \mid \text{Context}) \sim 0.77$
- $P(\text{Tyr} \mid \text{Context}) \sim 0.14$
- $P(\text{His} \mid \text{Context}) \sim 0.05$

An Aromatic Probably Belongs Here

**WT Residue = Gln 39**

- $P(\text{Arg} \mid \text{Context}) \sim 0.68$
- $P(\text{Lys} \mid \text{Context}) \sim 0.30$
- $P(\text{His} \mid \text{Context}) \sim 0.01$

Gln Probably Does not Belong Here

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

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
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

**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)

Paik et al. ACS Biochemistry 2021

MutCompute predictions available at [https://mutcompute.com/view/3tan](https://mutcompute.com/view/3tan)
Combined Variants are Inhibitor Resistant

Urine Sample contains Urea ~300mM

Urea 50mM blocks PCR
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?

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

Hal Alper, ChemE
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
MutCompute designed variant outperforms the literature on a PET depolymerization.

*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)


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

Visualize FAST-PETase at https://mutcompute.com/view/7sh6
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](https://mutcompute.com/view/6ij6).
PET degradation time-lapse

Sourced from Walmart
48 hour time lapse at 50C (122F)
Synergize MutComputeX with AlphaFold and Docking for Substrate Specificity Engineering

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

Blue is experimental
Green is AlphaFold

10 mutations in sequence space computationally modeled

TM-Score: 92

Experimental ligand
Al-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

And not make: 3-OMe Norbelladine

Galantamine: FDA Approved drug to treat Alzheimer

**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

**Directed evolution:**
- RTX: Jared Ellefson, Raghav Shroff
- T7 RNAP: Adam Meyer

**Funding:**
- DTRA
- Exxon
- NASA
- NIH
- Welch Foundation