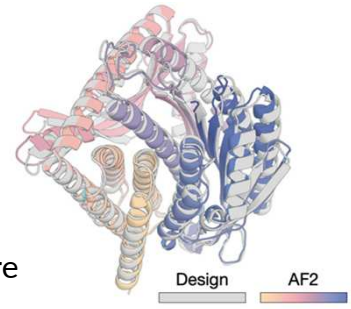
Computational Protein Design

Today's slides were adapted & edited from sets by:
Clay Kosonocky (UT Austin, "Machine Learning for Biochemical Applications", https://www.biomlsociety.org/seminar)
&
Joe Watson/David Juergens (Uwashington, "RFDiffusion: Accurate protein design using structure prediction and diffusion generative models", https://www.youtube.com/watch?v=wIHwHDt2NoI)

BCH394P/364C Systems Biology / Bioinformatics

Edward Marcotte, Univ of Texas at Austin

Why design new proteins?



- Function
  - Create enzymes & binders not found in nature
- Structure
  - Creating symmetric assemblies
  - Remove / modify antigenic structures
- Optimize properties (stability, expression, etc.)
  - Redesign natural enzymes to work at higher temp, survive organic solvents, bind new substrates, etc

How do we design new proteins?

Modified from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al.*, *Nature* 620:1089-1100 (2023)

This is a rich field with decades of effort. We're not going to review it. Instead, we'll focus only on recent efforts using ML (=AI) for protein design by leveraging AlphaFold/RosettaFold/ESMFold. For example, language models like ESM2 can predict single amino substitutions:

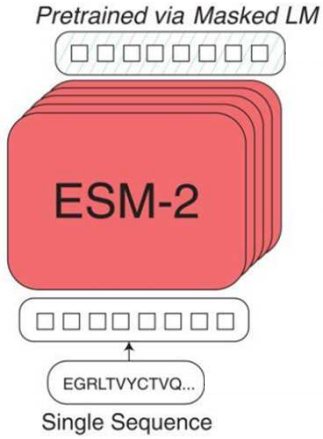
M N K F K G N K V V
…

Predict w/ model

M N K F K [?] N K V V
…

Mask residue

M N K F K T N K V V
…



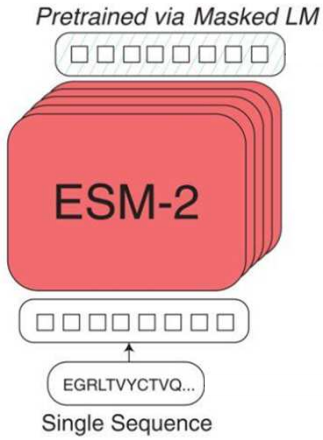
Modified from Clay Kosonocky

Lin *et al, Science* 379:1123-1130 (2023)

This is a rich field with decades of effort. We're not going to review it. Instead, we'll focus only on recent efforts using ML (=AI) for protein design by leveraging AlphaFold/RosettaFold/ESMFold. For example, language models like ESM2 can predict single amino substitutions:

Why should this do anything?

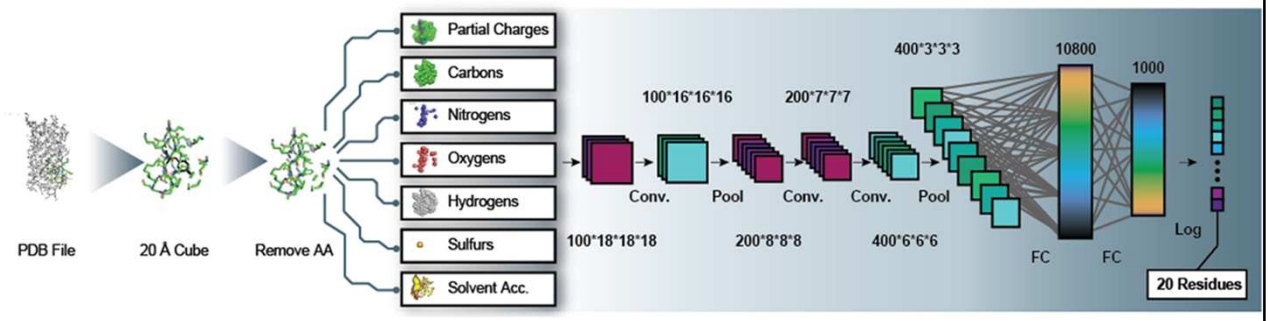
Learned evolutionary information used to predict when nature "messed up"



Modified from Clay Kosonocky

Lin *et al, Science* 379:1123-1130 (2023)

Similarly, substitutions can be predicted using a self-supervised structural approach



MutCompute Architecture

Predicts amino acid changes that would be expected to be preferred at a given position based on the structural environment

Modified from Clay Kosonocky

Shroff *et al. ACS Synth. Biol.* 9:2927–2935 (2020)

Amino acid segments can also be redesigned using LLMs

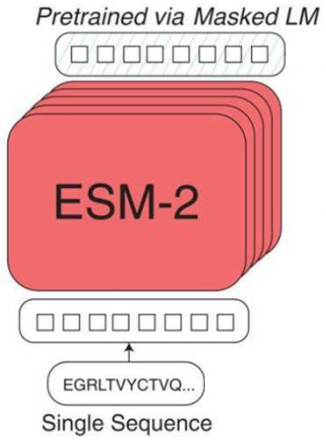
M N K H I G M K V V
…

Predict w/ model

M N [?] V V …

Mask span

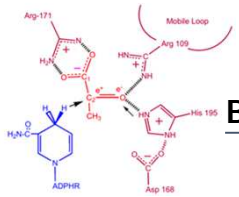
M N K F K T N K V V
…



Slide from Clay Kosonocky

Figures from Lin et al. 2022 & Dauparas et al. 2022

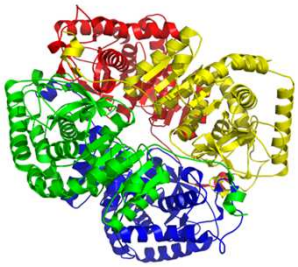
For complete redesigns, we can instead consider the following structure-based workflow for ML protein design:



Lactate dehydrogenase

Protein function

Backbone Design



Protein backbone

Inverse Folding

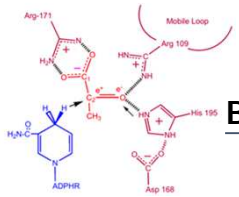
MNKFKGNKVVLIG
NGAVGSSYAFSLV
NQSIVD…

Amino acid sequence

Modified from Clay Kosonocky

Backbone design

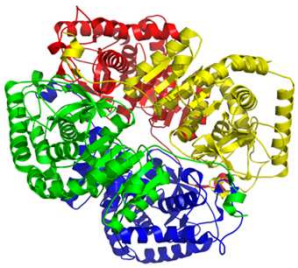
Which backbone will give us the desired function?



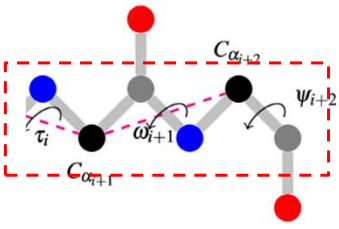
Lactate dehydrogenase

Protein function

Backbone Design



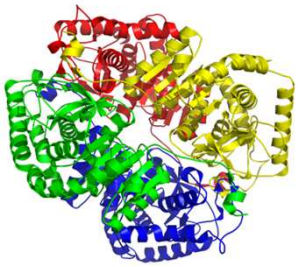
Protein backbone



Slide from Clay Kosonocky

Inverse folding

Which amino acid sequence will give us the desired backbone?



Protein backbone

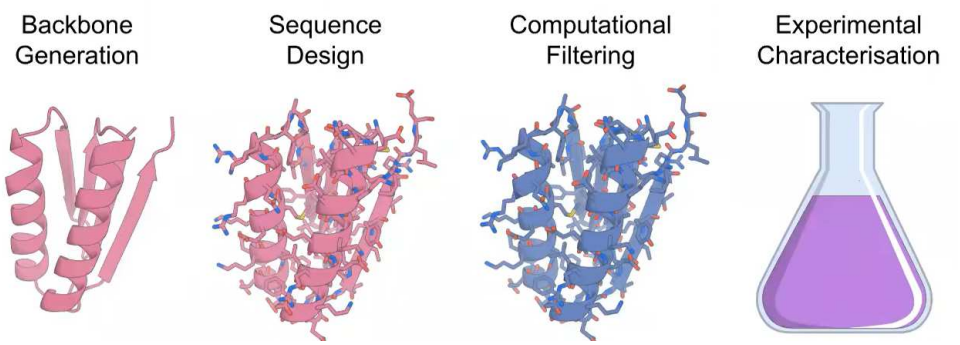
Inverse Folding →

MNKFKGNKVVLIG
NGAVGSSYAFSLV
NQSIVD…

Amino acid sequence

Slide from Clay Kosonocky

In practice, these steps can be carried out in the following way:



ProteinMPNN

AlphaFold
RoseTTAFold
ESMFold
Boltz
Chai
etc.

Clone, express, purify 100+ clones

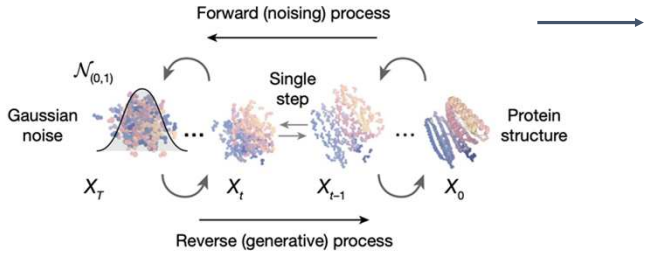
Also BoltzGen/Design

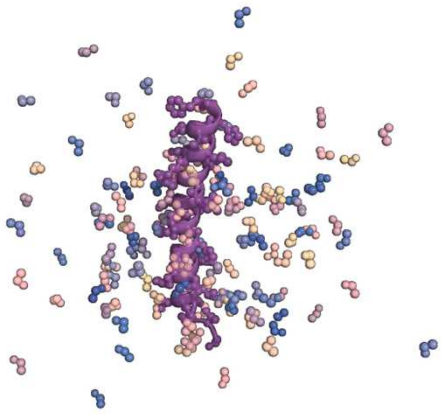
Joe Watson & David Juergens
https://www.youtube.com/watch?v=wIHwHDt2NoI

Protein backbone design: RFdiffusion

RoseTTAFold

+

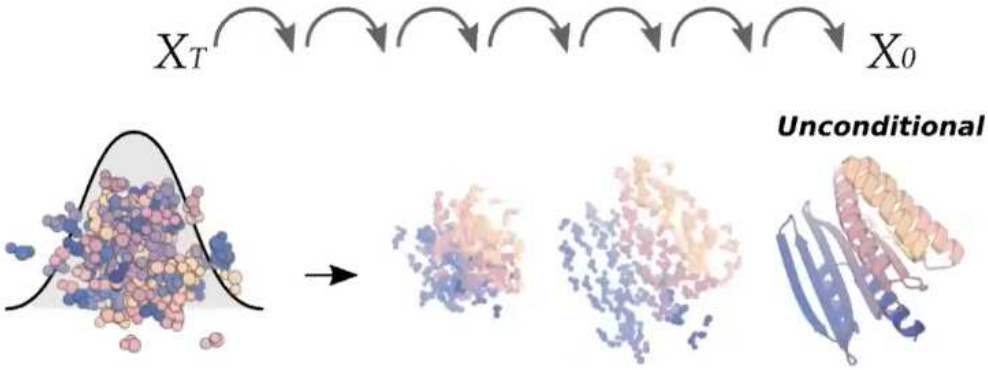




Slide from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

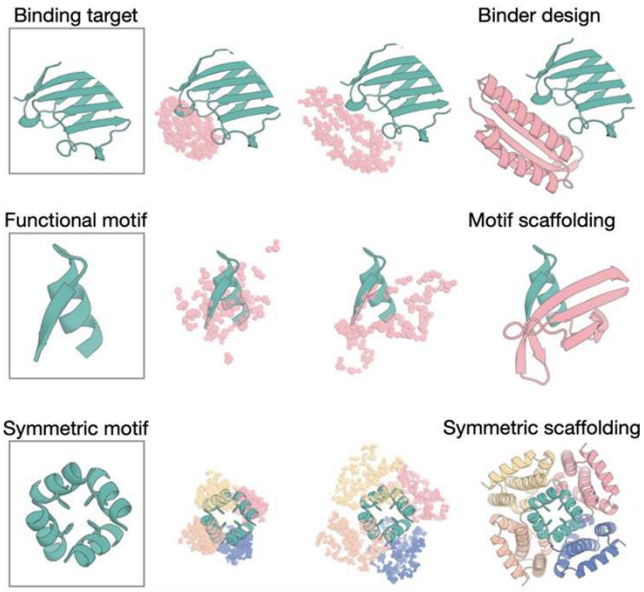
RFdiffusion can generate entirely new folds starting from noise



Slide from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

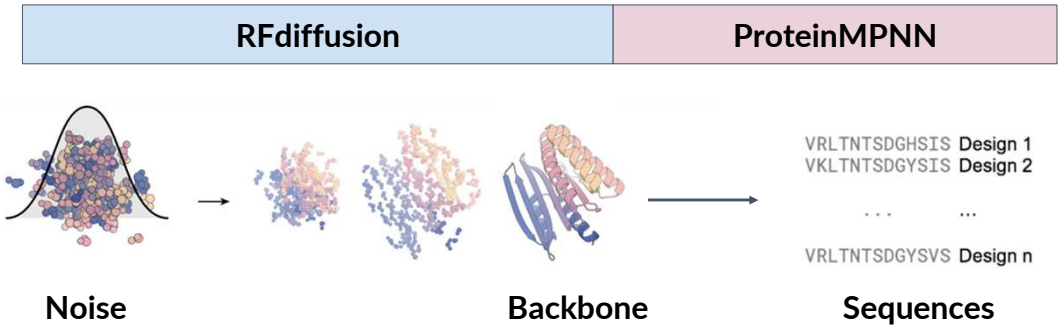
RFdiffusion can also conditionally generate new backbones, e.g. :



Slide from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

Combining RFdiffusion for backbone design + ProteinMPNN for amino acid sequence selection

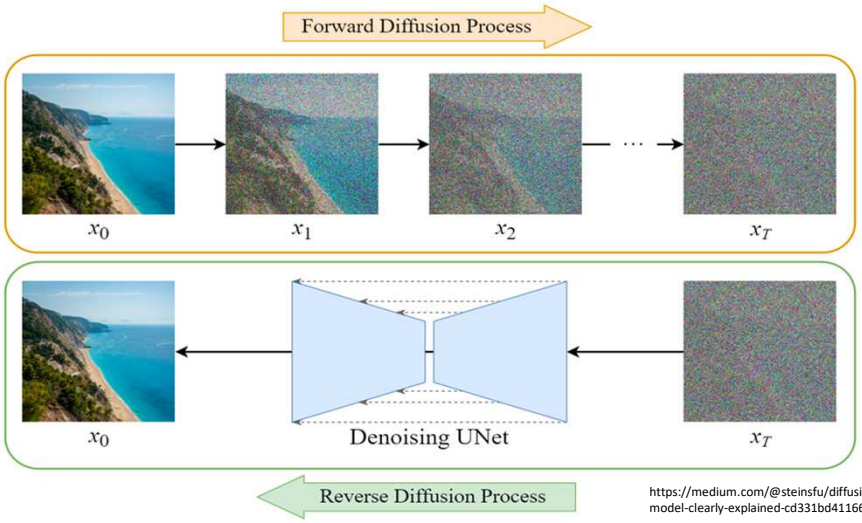


Slide from Clay Kosonocky
Figures from Watson et al. 2023 & Dauparas et al. 2022

Before we look at how these models work, let's try a live demo of RFDiffusion + MPNN to design 2 proteins designed to bind each other

https://colab.research.google.com/github/sokrypton/ColabDesign/blob/main/rf/examples/diffusion.ipynb

RFDiffusion is an example of a diffusion model, e.g.:



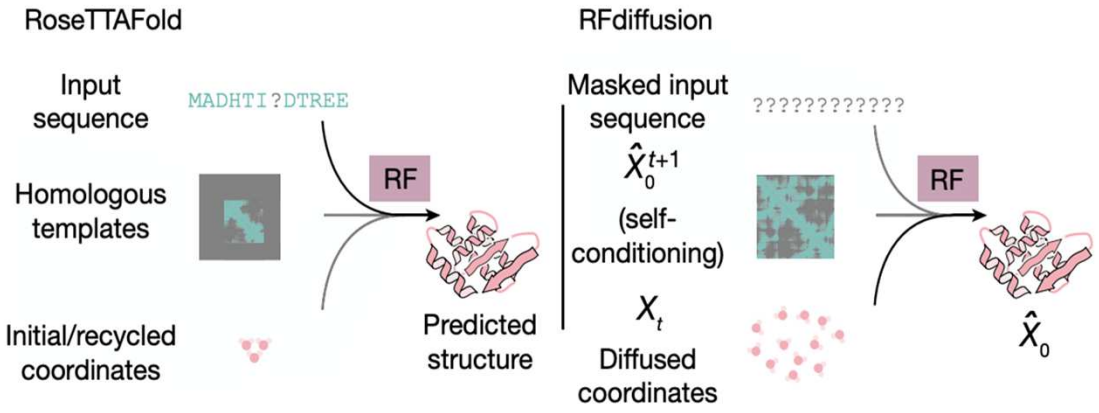
https://medium.com/@steinsfu/diffusion-model-clearly-explained-cd331bd41166#b614

The key idea is to add "structured" noise to data (like images) in a series of consecutive time steps = Gaussian noise of known variance. Then, train a NN to undo this noise. If starting from a full noise starting point, this process will then converge the image to something that resembles starting training data

Modified from Clay Kosonocky

RFdiffusion architecture

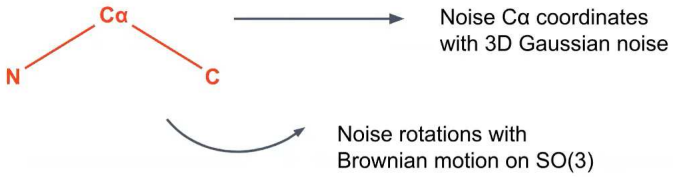
Essentially, Watson et al. took the RoseTTAFold architecture and model parameters and fine tuned on the denoising task:

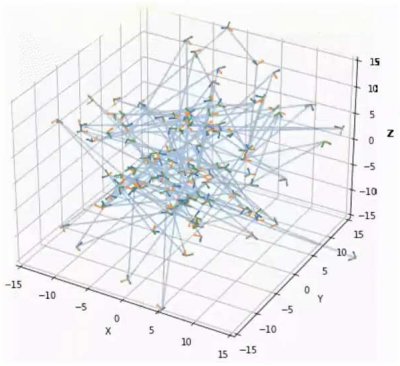


Slide from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

RFDiffusion treats each peptide bond N-Cα-C as a triangle, a simple representation that retains full information about the backbone, then "diffuses" (translates and rotates) these triangles



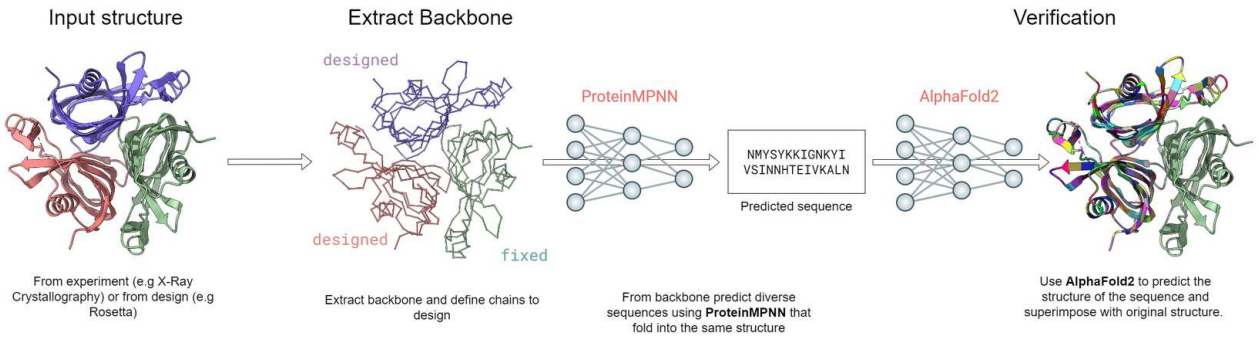


→ *Train a neural network to remove noise ('reverse') this process*

Slide from Jason Lim *via* the Watson/Juergens lecture

ProteinMPNN generates amino acid sequences for backbones

Input = a protein structure (just needs the backbone)
Output = predicts new sequences that should fold into that backbone

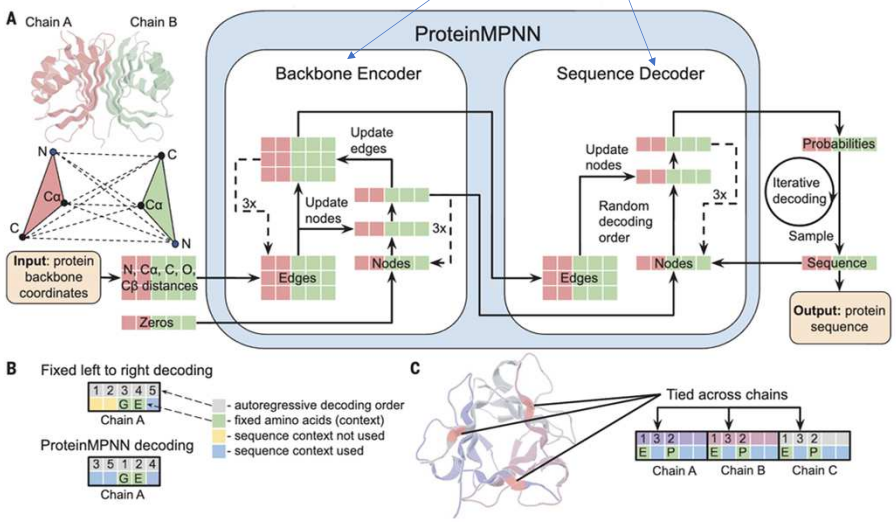


Live Demo:
https://huggingface.co/spaces/simonduerr/ProteinMPNN
https://doi.org/10.5281/zenodo.7630417

ProteinMPNN paper: Dauparas *et al. Science* 378, 49–56 (2022)

ProteinMPNN architecture

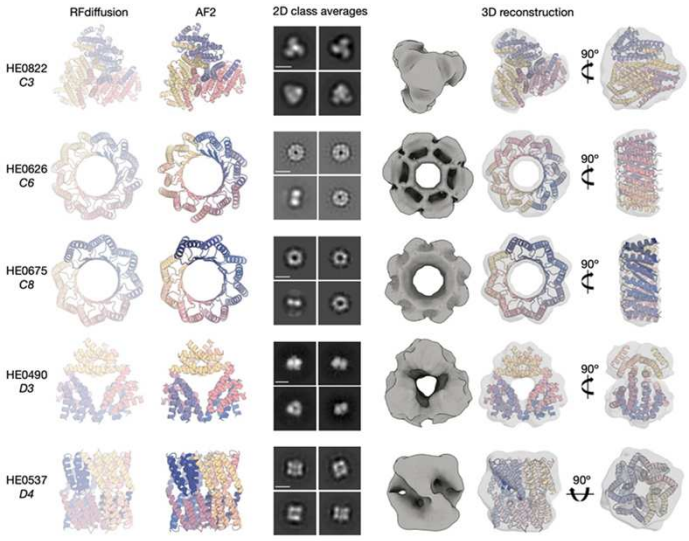
Encoder encodes the backbone features, the decoder runs iteratively to predict each aa



~1.7M parameters (vs ~100M for alphafold), trained on 1 A100 for 2 days

Figure from Dauparas et al. 2022

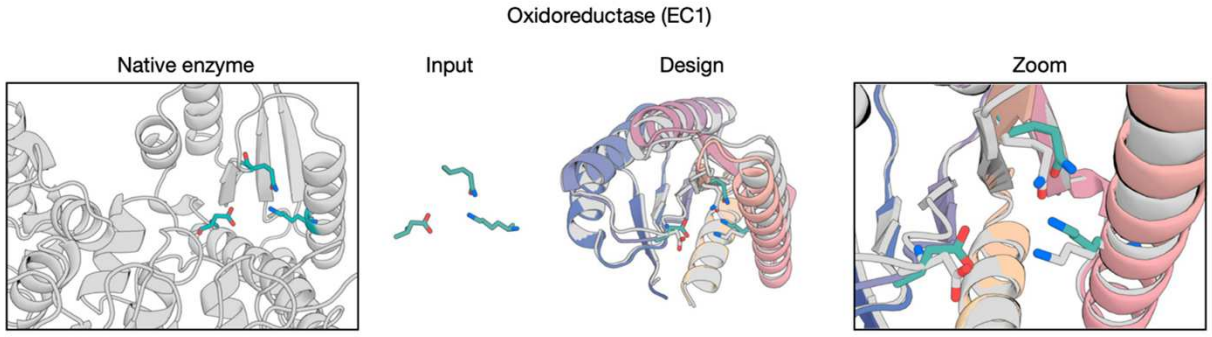
Designs: Symmetric assemblies



Slide from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

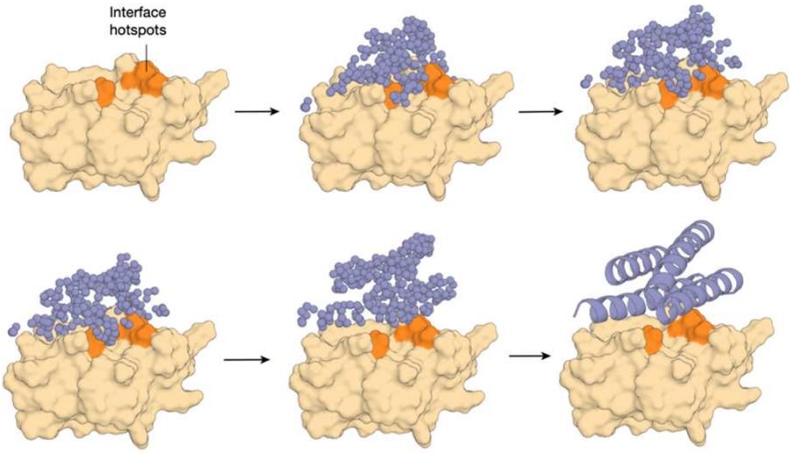
Designs: Constrained active site



Slide from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

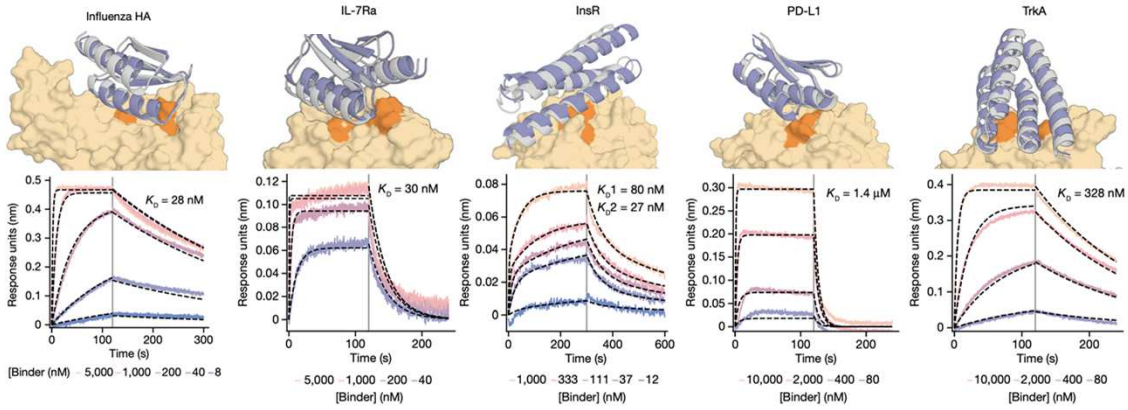
Designs: Protein binders



Slide from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

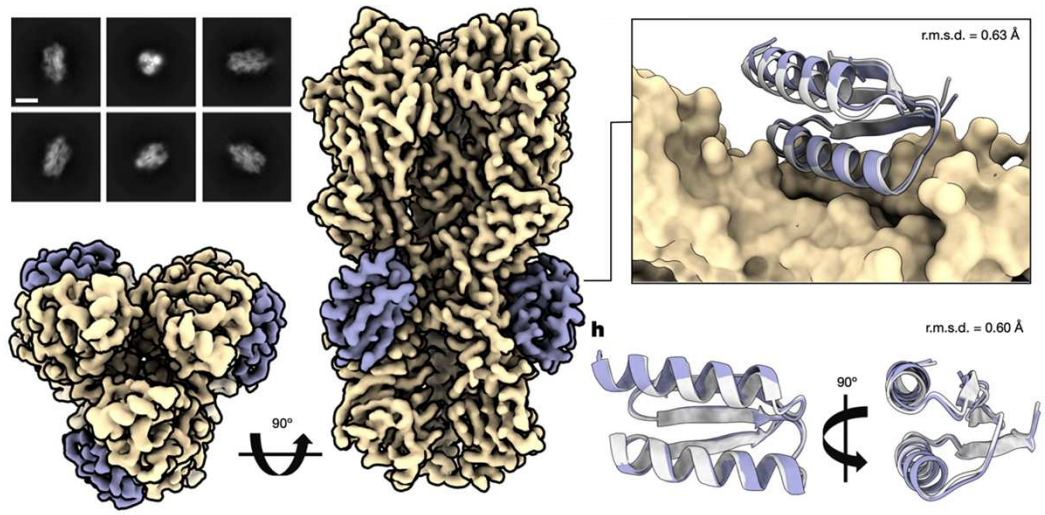
Protein binder experimental validation



Slide from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

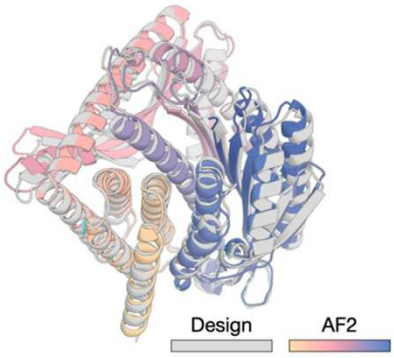
Influenza Haemagglutinin (HA) binder structure

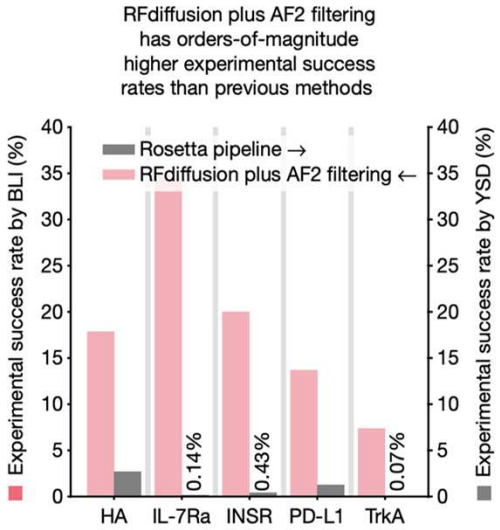


Slide from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

Rfdiffusion+MPNN designs can be independently computationally verified with AlphaFold

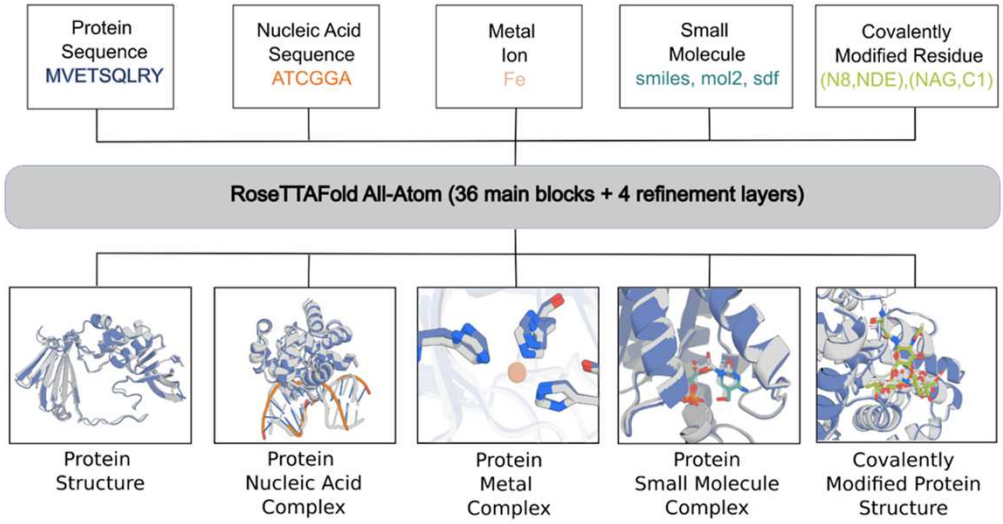




Slide from Clay Kosonocky

Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

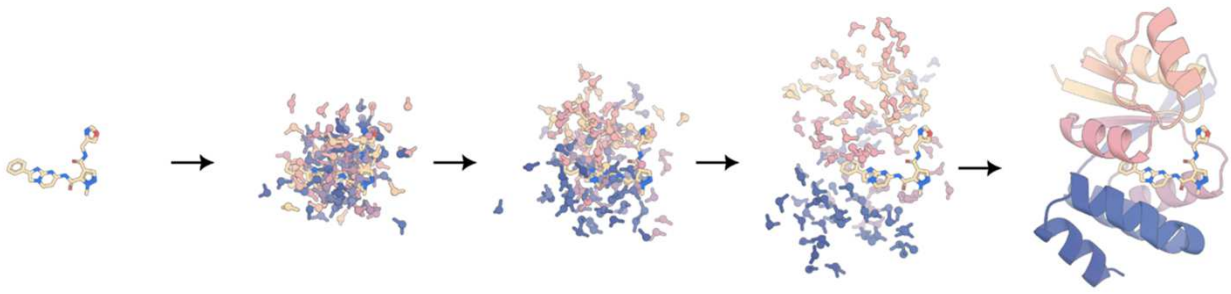
Ligands are now modelable with RoseTTAFold All-Atom



Modified from Clay Kosonocky

Krishna *et al, Science* March 7, 2024 DOI: 10.1126/science.adl2528

Similarly, now there's Rfdiffusion All-Atom (RFdiffusionAA)

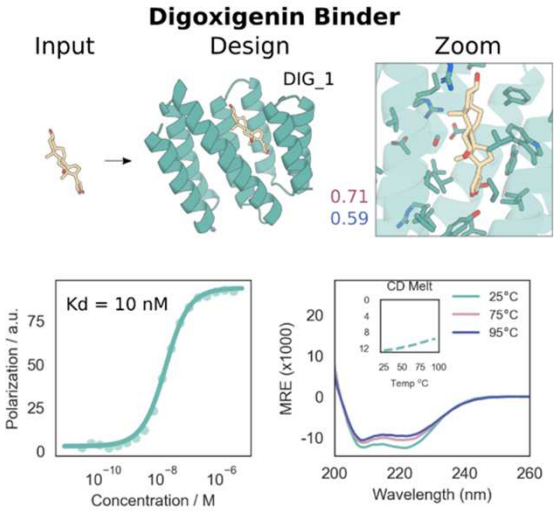


RFAA can be converted to a diffusion model just like RoseTTAFold

Slide from Clay Kosonocky

Krishna *et al, Science* March 7, 2024 DOI: 10.1126/science.adl2528

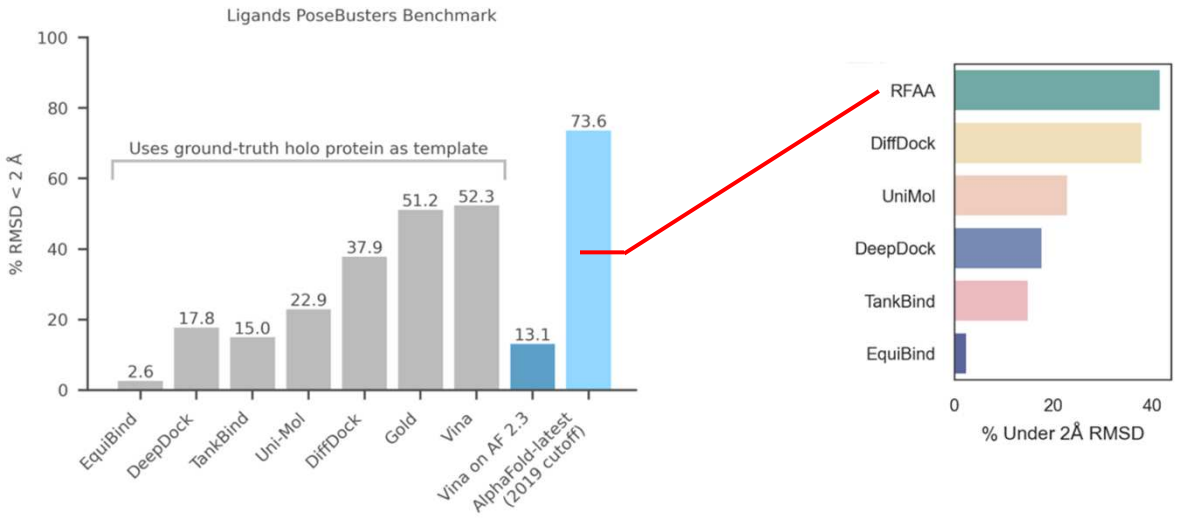
RFdiffusionAA ligand binder design: Digoxigenin



Slide from Clay Kosonocky

Krishna *et al, Science* March 7, 2024 DOI: 10.1126/science.adl2528

AlphaFold All Atom claims strong performance, although its functionality hasn't been extensively independently tested yet



Slide from Clay Kosonocky

Figure from Krishna et al. 2023 & Deepmind

…which brings us to an important question:

nature
NEWS FEATURE | 15 October 2024

AI has dreamt up a blizzard of new proteins. Do any of them actually work?

Emerging protein-design competitions aim to sift out the functional from the fantastical. But researchers hope that the real prize will be a revolution for the field.

By Ewen Callaway



"… in **Bits to Binders**, researchers are vying to create small proteins that could be used in a T-cell cancer therapy. **Run by the BioML Society, a graduate-student-led group at the University of Texas at Austin, it attracted 64 teams from 42 countries** — including Nigeria, Colombia, Iran and India. Around 18,000 designs are now being tested, with results due in early 2025. "We were quite surprised with the turnout," says co-organizer Clayton Kosonocky, a biochemistry PhD student at the university.

→ Teams designed the **80 amino acid antigen binding domain** of a Chimeric Antigen Receptor (CAR) to engage a CD20 cancer antigen target, with the goal of activating a CAR-T cell killing and proliferation response.

→ DNA was synthesized by Twist for the top 12,000 designs and are currently being tested by Leah Labs for T-cell activation.
(Sneak peak as of yesterday: some designs appear to be working!)

So many protein design tools. Which is best?

Structure Prediction:
AlphaFold
ESMFold
RoseTTAFold
RoseTTAFoldAllAtom
UMol
AlphaFold2-RAVE
Boltz
OpenFold

Generative Models:
RFDiffusion/AllAtom
Chroma
Protpardelle
ProGen
FrameDiff
Genie
Foldflow
Frameflow
Proteus
Multiflow
Pepflow
Bindcraft
Boltz-Gen

Language Models:
ESM-2
SaProt
Prost-T5
ProtBERT

Inverse Folding:
ProteinMPNN
LigandMPNN
CarbonDesign
ESM-IF1

Other tools:
Molecular Dynamics
Rosetta
FoldSeek
BLAST / MMseqs2
MutCompute
Stability Oracle

Can open science decide?

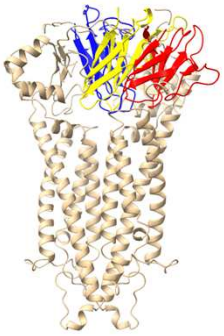
Real world scenario: CAR-T Therapy

Kosonocky et al., *bioRxiv* (2026), DOI:10.64898/2026.03.03.709355

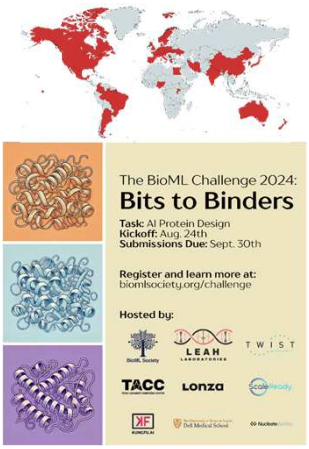
Bits to Binders AI Design Competition

62 teams entered, 85+ institutions, 40+ countries
~30 teams submitted
12,000 binders ordered

CAR-T + AI:
Generalizable cancer therapy?



Task: Bind the lymphoma cancer cell surface antigen CD20



Design: Aug '24 - Oct '24
DNA: Oct '24 - Feb '25
Testing: Feb '24 - Jan '26

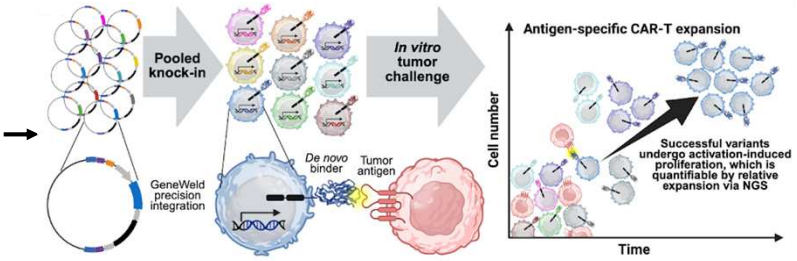
3 designed binders courtesy of Ty Gardner and Alia Clark-ElSayed
Kosonocky et al., *bioRxiv* (2026), DOI:10.64898/2026.03.03.709355

Bits to Binders AI Design Competition



DNA (courtesy of TWIST)

T cells with binders (from Leah Labs)

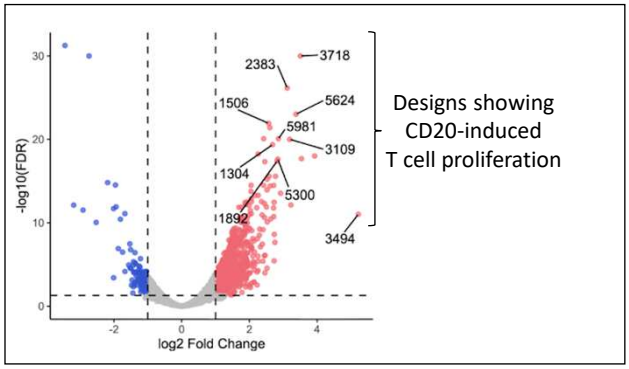


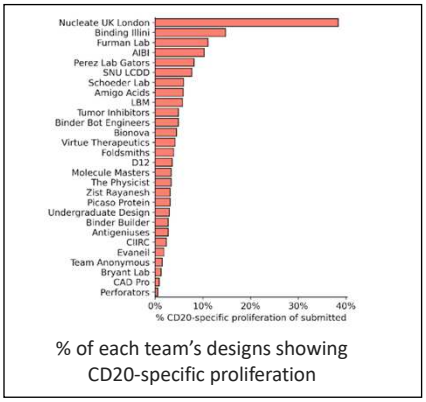
February 2025:
- 2.5 x 10^8 cells after 10-day outgrowth & selection
- 25% CAR+
  - Clinical CAR ~ 90% (w/ CAR known to work)
- Of 12,000 designs, down to 25%, or **3,000 candidates**

April 2025:
- **Approx. ~100 of the designs worked!**

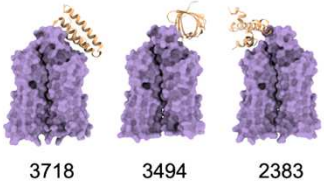
April 2025 – Jan 2026:
- Replicates
- Higher resolution assay on top hits
- Binding assays
- Write up paper and present awards

Kosonocky et al., *bioRxiv* (2026), DOI:10.64898/2026.03.03.709355





% of each team's designs showing CD20-specific proliferation



Example binder designs

A few design lessons learned:

(1) High GC content and repetitive sequences led to clone failure

(2) Nearly 40% of designs failed to express functional receptors, attributed to alpha helices excessively enriched in lysine (K) and glutamate (E), thought to induce premature translation termination

Bottom line: Don't just design for affinity; you have to consider compatibility with cellular machinery for proper expression, folding, & localization

Where's the field now?

- Protein design is getting better and better
- Conditional generation options growing
  - Motifs, ligands, active sites, protein binding, etc.
- Challenges
  - ***<u>**Needs broader experimental validation**</u>***
  - Designing around conformation changes, cell expression, solubility, not aggregating, not immunogenic
  - Antibody-antigen designs are hard but important, models are improving, but still hard and needs more validation

Modified from Clay Kosonocky