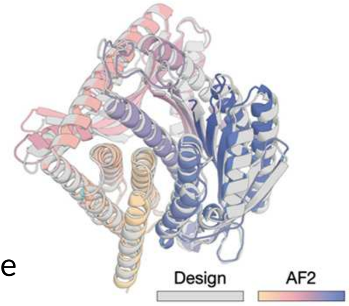
Computational Protein Design

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?



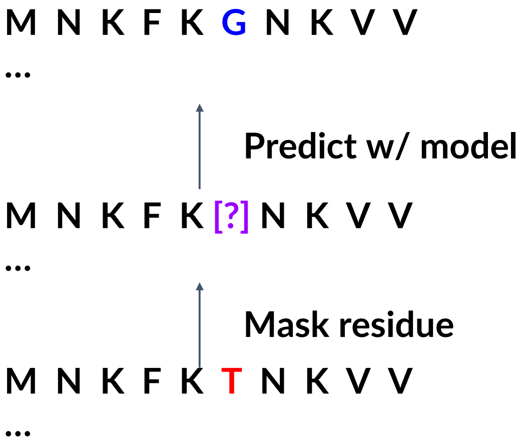
- Function
  - Create enzymes & binders not found in nature
- Structure
  - Creating symmetric assemblies
  - Remove / modify antigenic structures
- Property Optimization (stability, expression, etc.)
  - Redesign natural enzymes to work at higher temp, survive organic solvents, bind new substrates,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



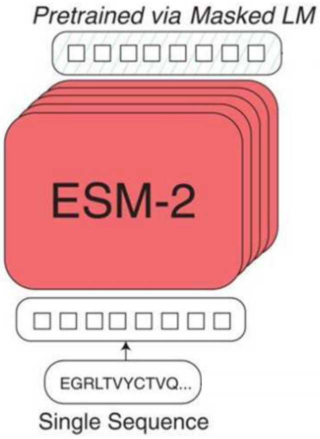
M N K F K G N K V V …

Predict w/ model

M N K F K [?] N K V V …

Mask residue

M N K F K T N K V V …



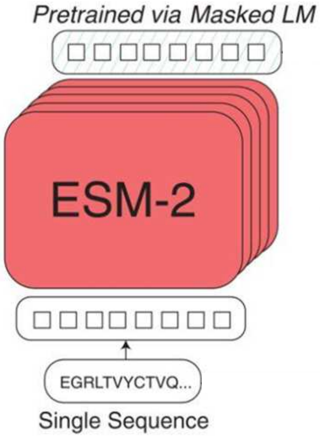
Modified from Clay Kosonocky

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?

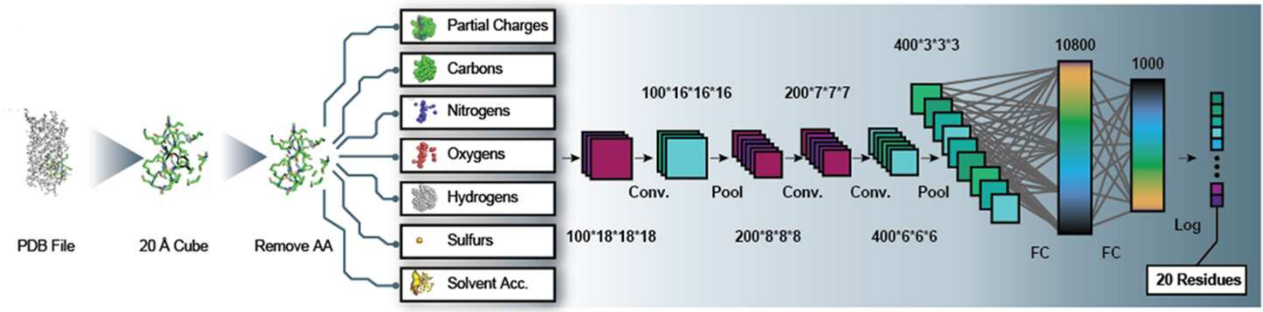
Learned evolutionary information used to predict when nature "messed up"



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

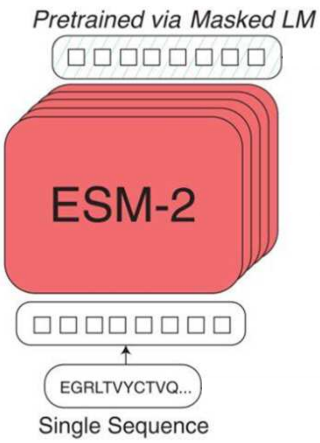
M N K H I G M K V V ...

Predict w/ model

M N [?] V V ...

Mask span

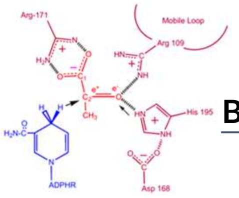
M N K F K T N K V V ...



Slide from Clay Kosonocky

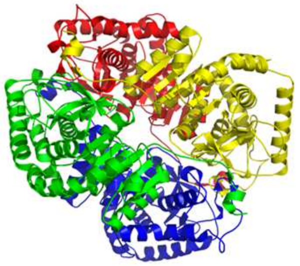
Figures from Lin et al. 2022 & Dauparas et al. 2022

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



Lactate dehydrogenase

Backbone Design →



Inverse Folding →

MNKFKGNKVVLIG
NGAVGSSYAFSLV
NQSIVD…

Protein function

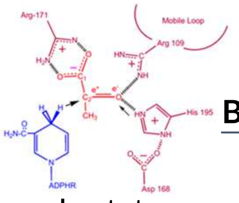
Protein backbone

Amino acid sequence

Modified from Clay Kosonocky

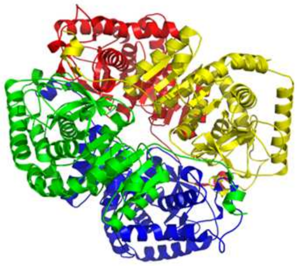
Backbone design

Which backbone will give us the desired function?



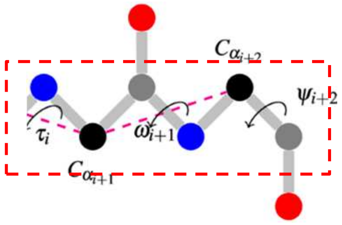
Lactate dehydrogenase

Backbone Design →



Protein function

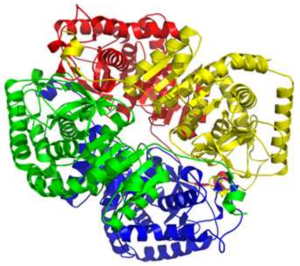
Protein backbone →



Slide from Clay Kosonocky

Inverse folding

Which amino acid sequence will give us the desired backbone?



Inverse Folding →

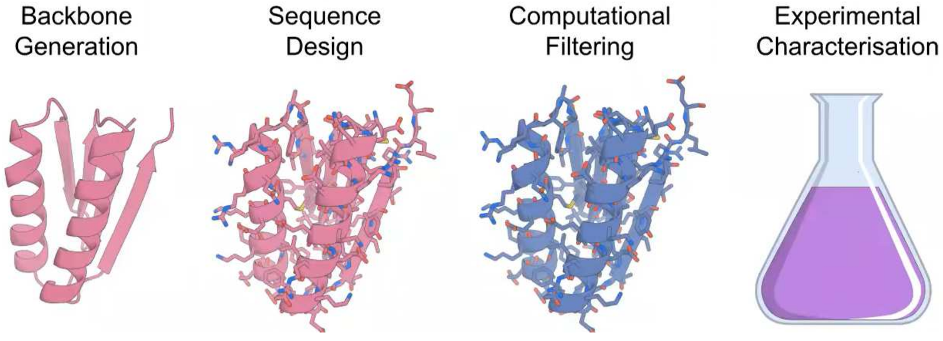
MNKFKGNKVVLIG
NGAVGSSYAFSLV
NQSIVD…

Protein backbone

Amino acid sequence

Slide from Clay Kosonocky

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



RFDiffusion

ProteinMPNN

AlphaFold2
RoseTTAFold
ESMFold
OmegaFold

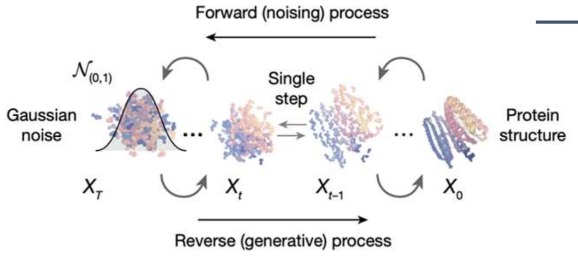
Clone, express, purify ~100 clones

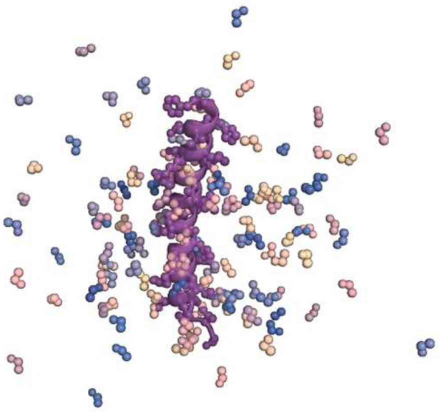
Joe Watson & David Juergens
https://www.youtube.com/watch?v=wIHwHDt2NoI

Protein backbone design: RFdiffusion

RoseTTAFold

+

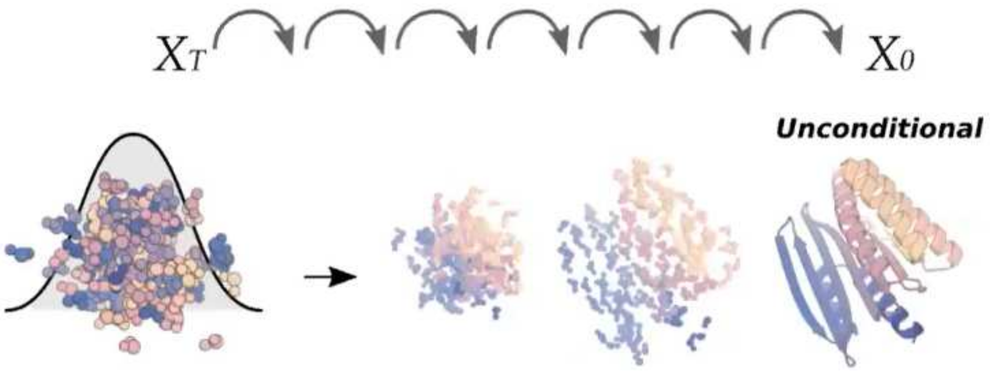




Slide from Clay Kosonocky

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

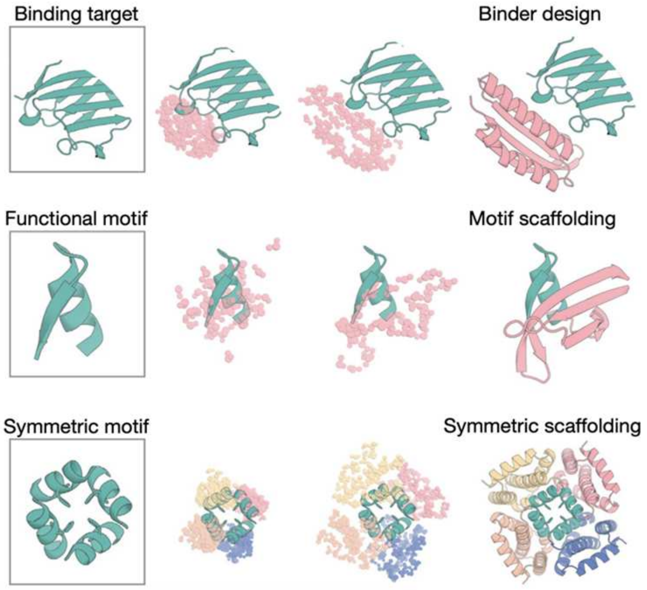
RFdiffusion can generate entirely new folds starting from noise



Slide from Clay Kosonocky

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

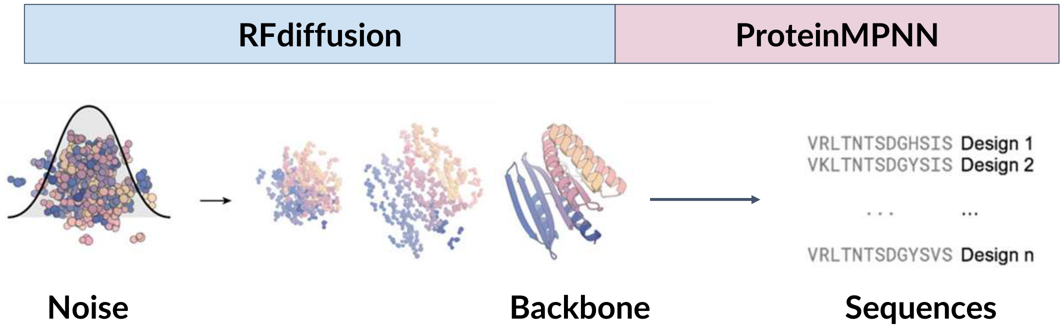
RFdiffusion can also conditionally generate new backbones, e.g. :



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


RFdiffusion | ProteinMPNN

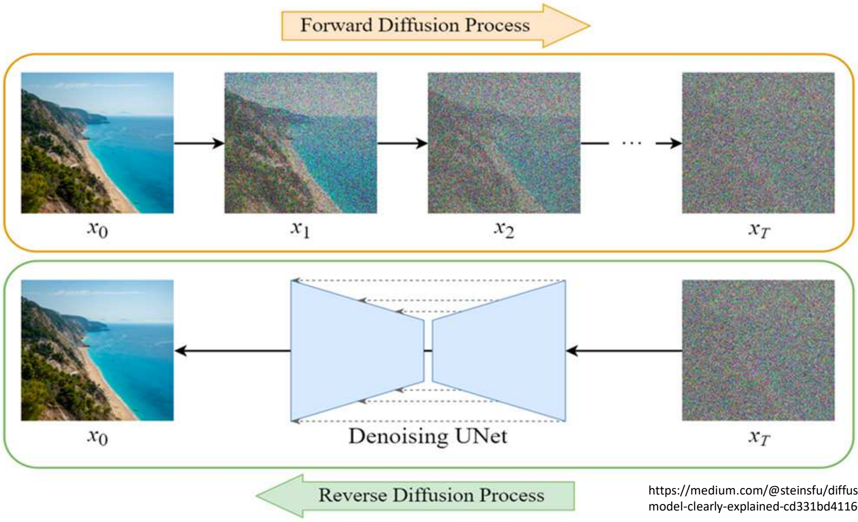
```
VRLTNTSDGHSIS Design 1
VKLTNTSDGYSIS Design 2
   . . .        ...
VRLTNTSDGYSVS Design n
```

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

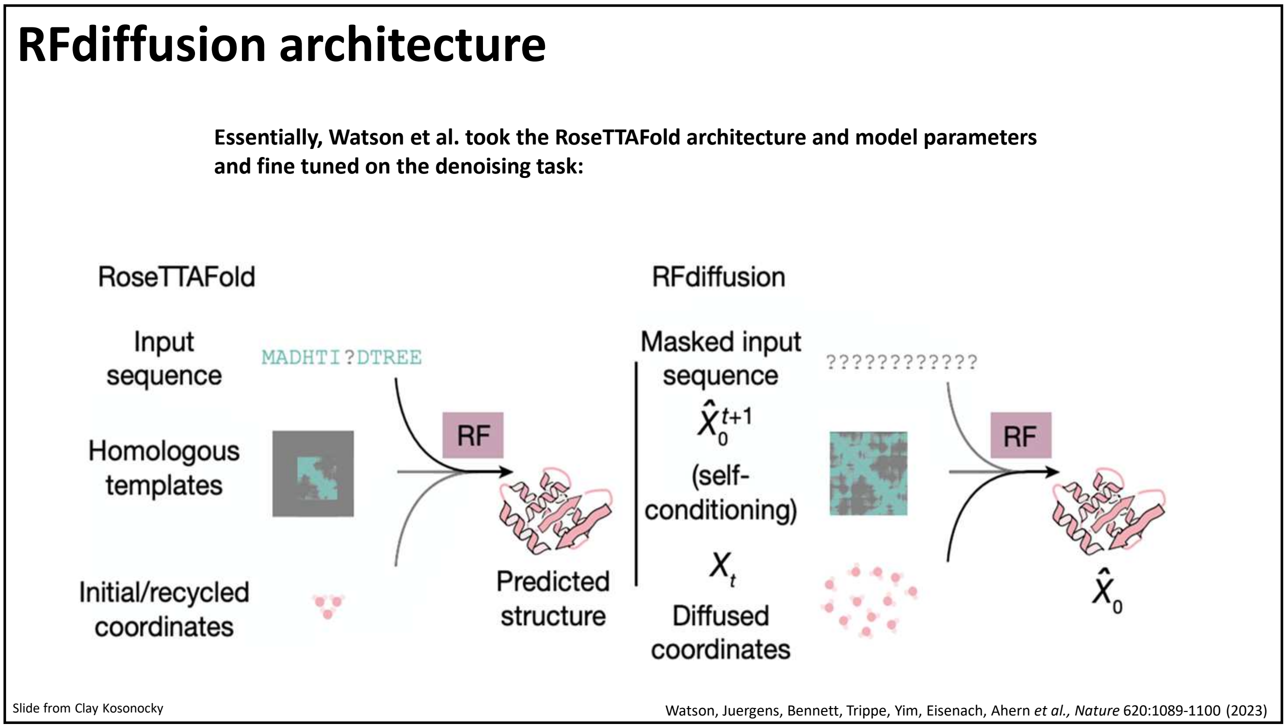
RFDiffusion is an example of a diffusion model, e.g.:



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:

RoseTTAFold

Input sequence

MADHTI?DTREE

Homologous templates

Initial/recycled coordinates

RF

Predicted structure

RFdiffusion

Masked input sequence

????????????

$\hat{X}_0^{t+1}$

(self-conditioning)

X_t

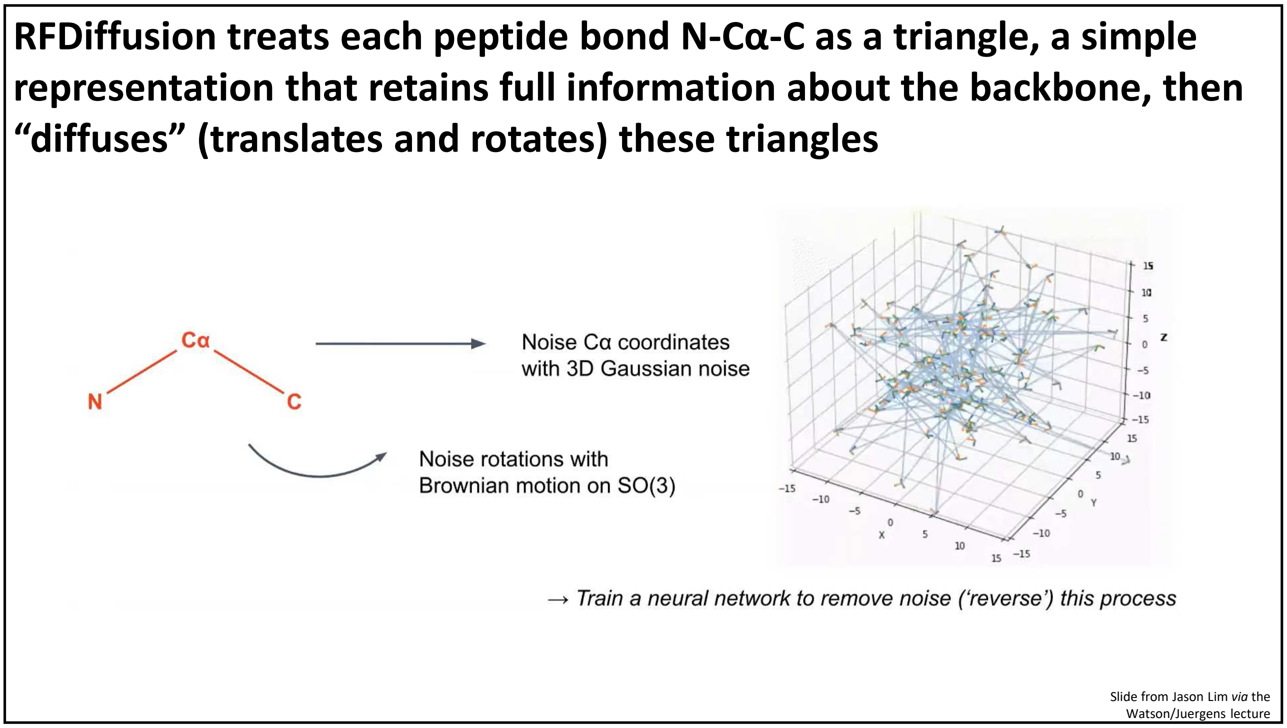
Diffused coordinates

RF

$\hat{X}_0$

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

Cα

N

C

Noise Cα coordinates with 3D Gaussian noise

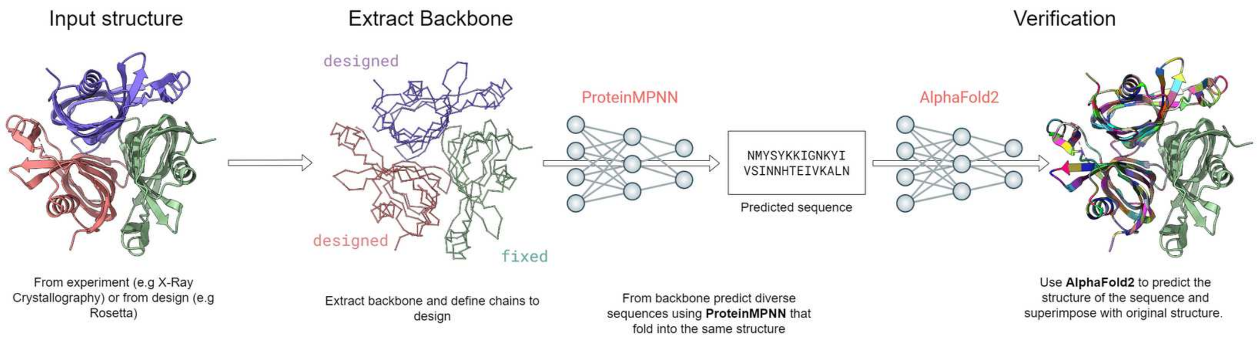
Noise rotations with Brownian motion on SO(3)

→ 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



From experiment (e.g X-Ray Crystallography) or from design (e.g Rosetta)

Extract backbone and define chains to design

From backbone predict diverse sequences using **ProteinMPNN** that fold into the same structure

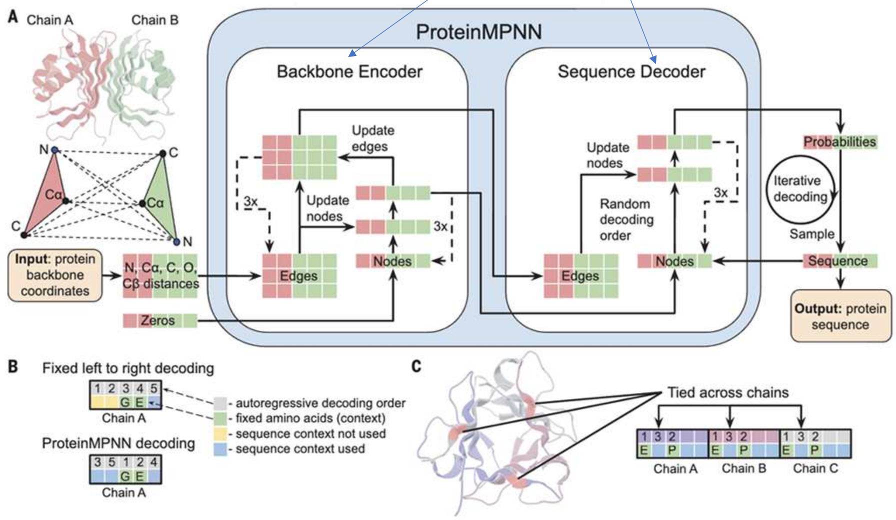
Use **AlphaFold2** to predict the structure of the sequence and superimpose with original structure.

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

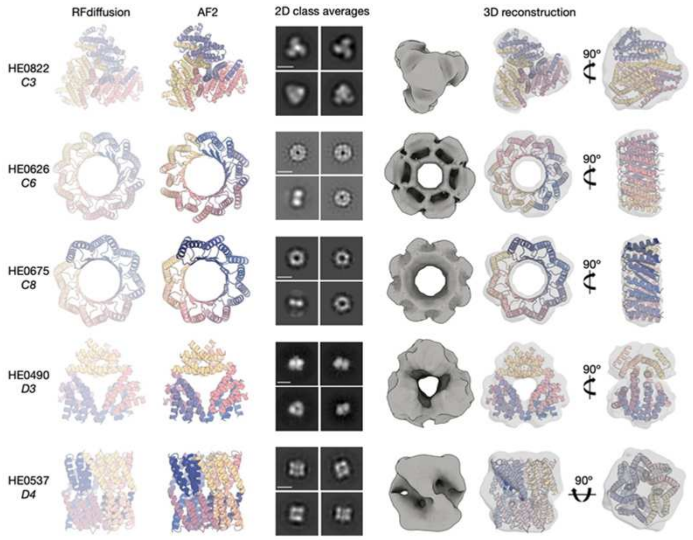
Encoder encodes the backbone features,
the decoder runs iteratively to predict each aa



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

Figure from Dauparas et al. 2022

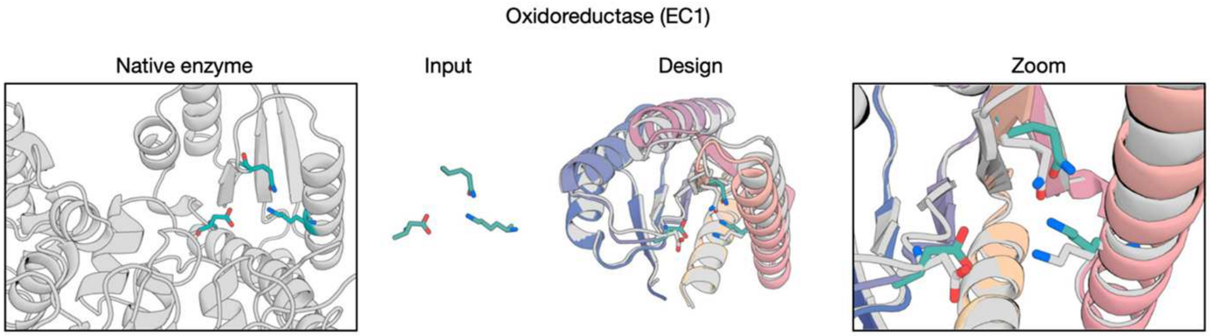
Designs: Symmetric assemblies



Slide from Clay Kosonocky

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

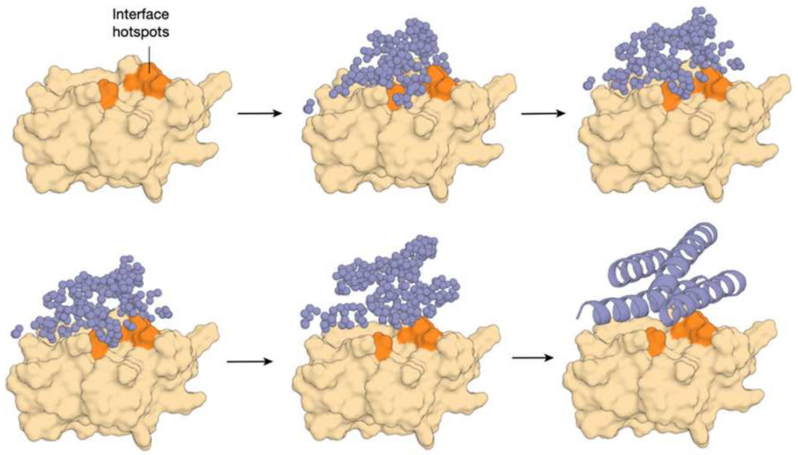
Designs: Constrained active site



Slide from Clay Kosonocky

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

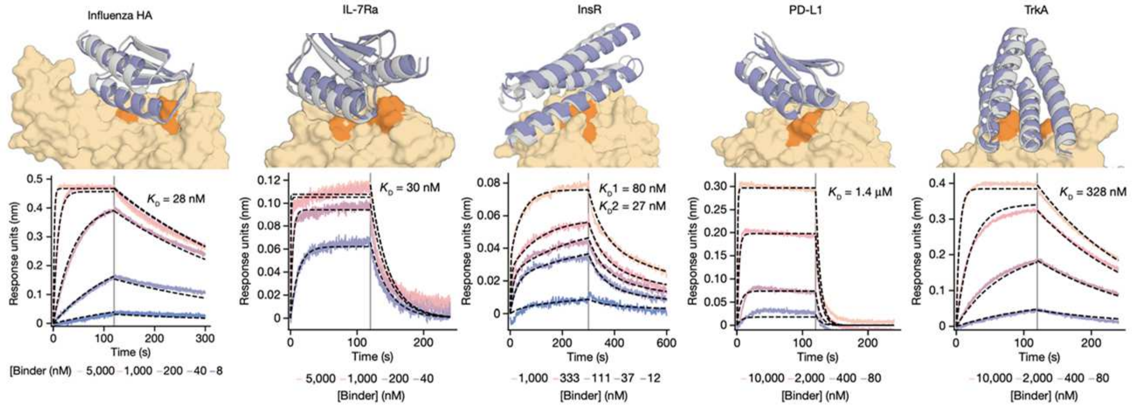
Designs: Protein binders



Slide from Clay Kosonocky

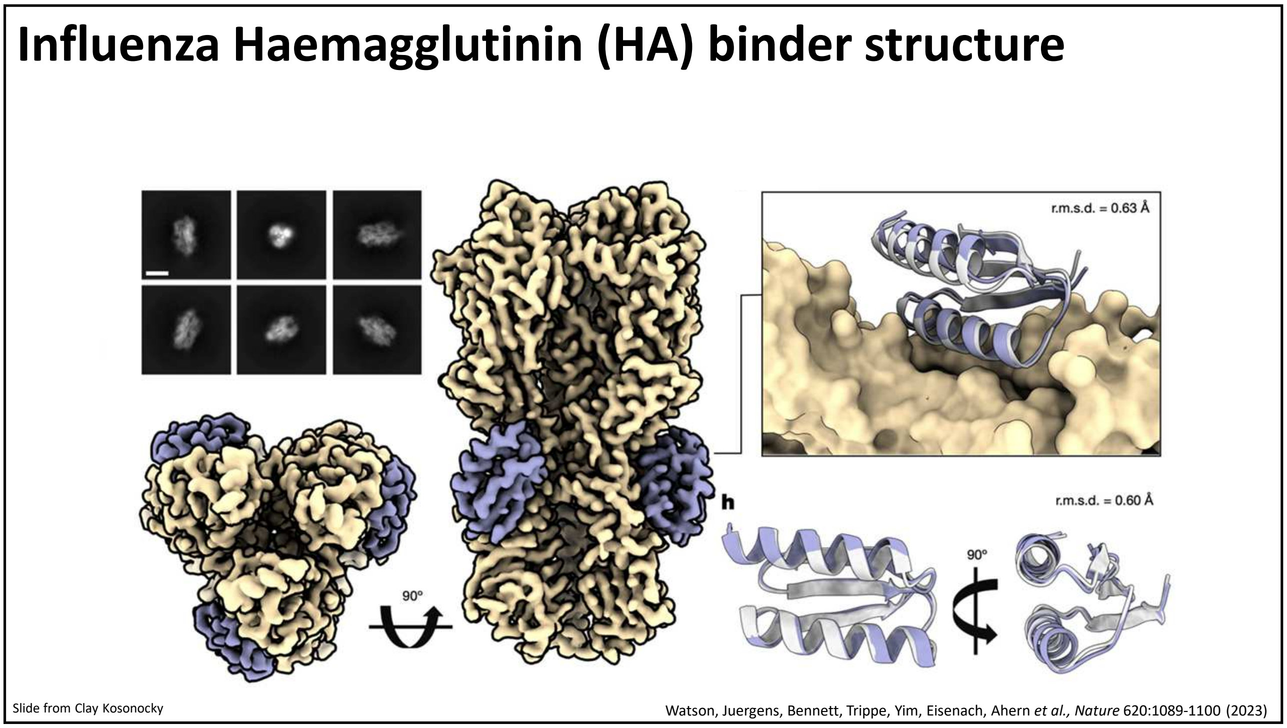
Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

Protein binder experimental validation



Slide from Clay Kosonocky

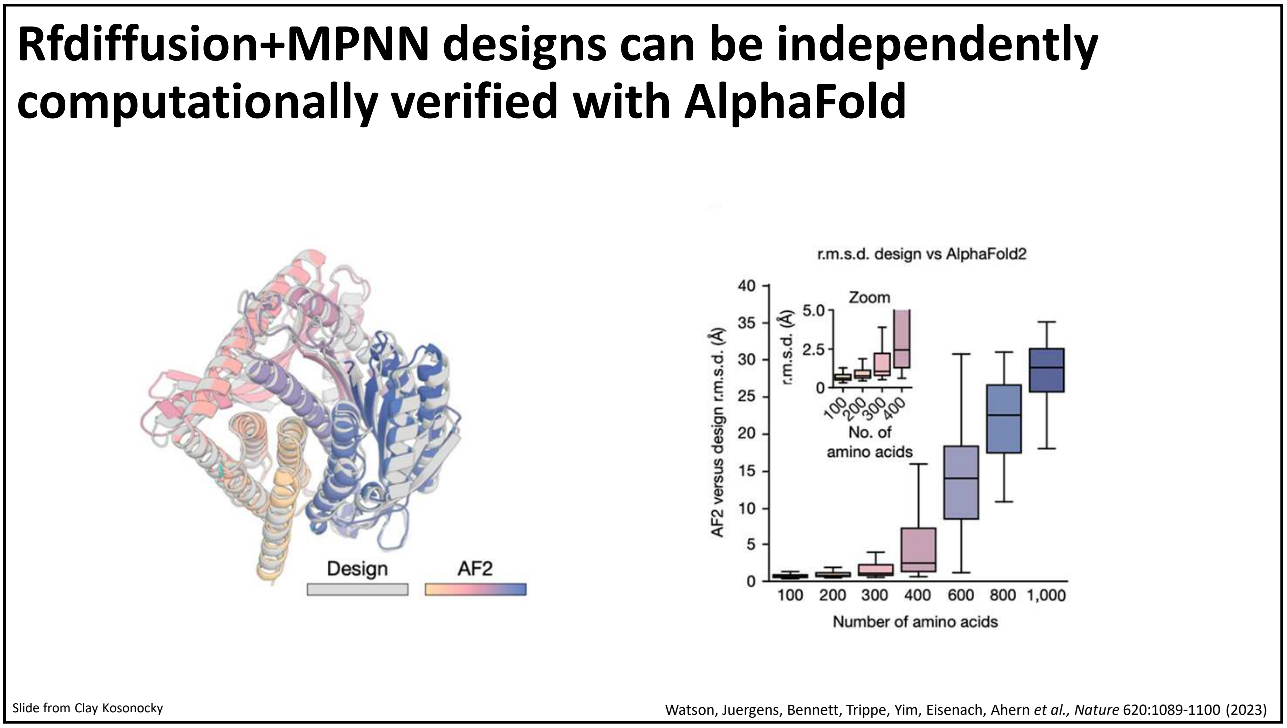
Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)



Influenza Haemagglutinin (HA) binder structure

Slide from Clay Kosonocky

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



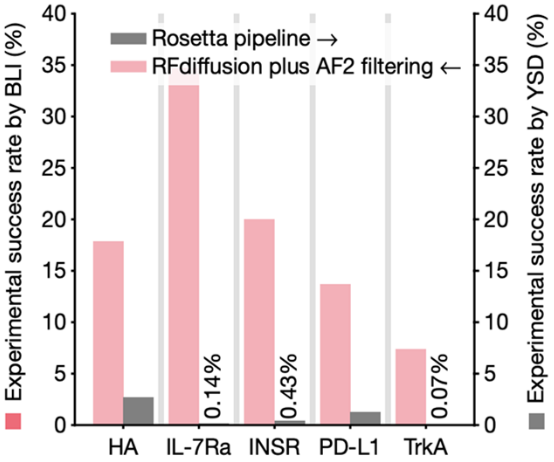
Rfdiffusion+MPNN designs can be independently computationally verified with AlphaFold

Slide from Clay Kosonocky

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

RFdiffusion + AF filtering outperforms previous methods

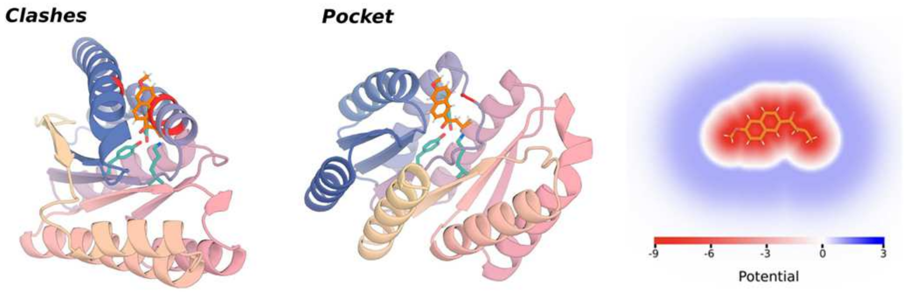
RFdiffusion plus AF2 filtering has orders-of-magnitude higher experimental success rates than previous methods



Slide from Clay Kosonocky

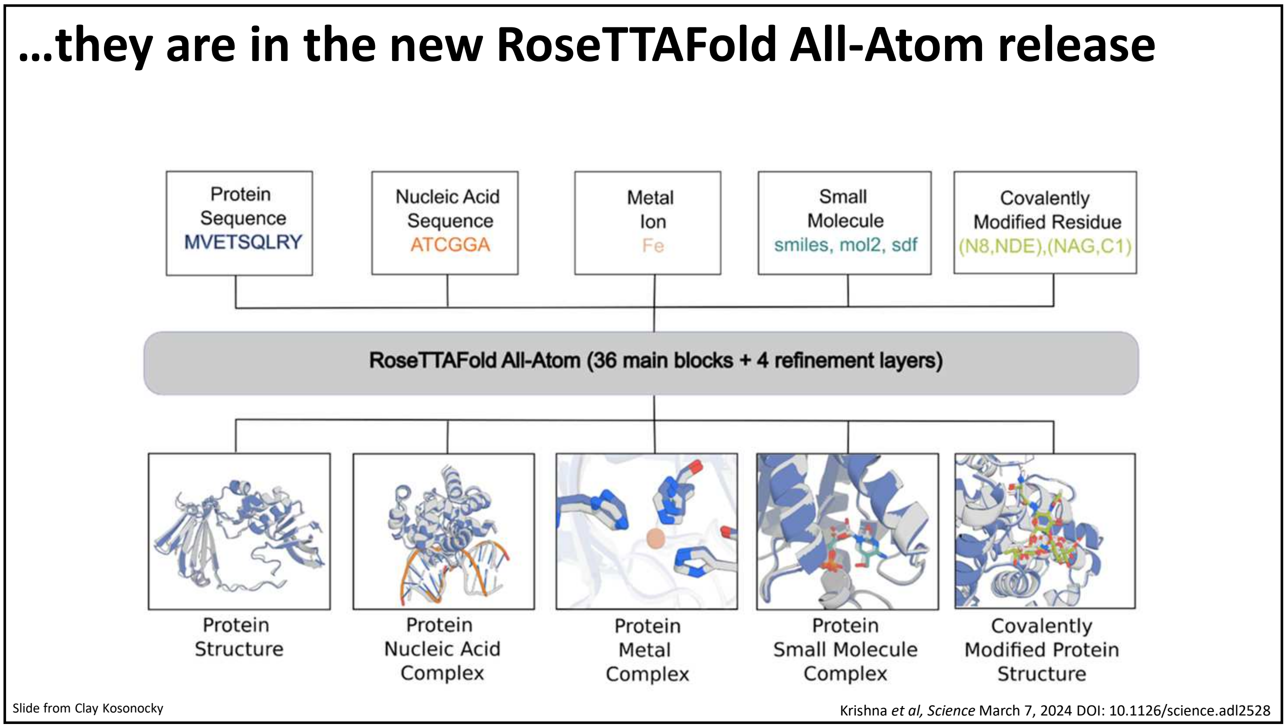
Watson, Juergens, Bennett, Trippe, Yim, Eisenach, Ahern *et al., Nature* 620:1089-1100 (2023)

Ligands are not explicitly modeled in RFdiffusion, but…



Slide from Clay Kosonocky

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

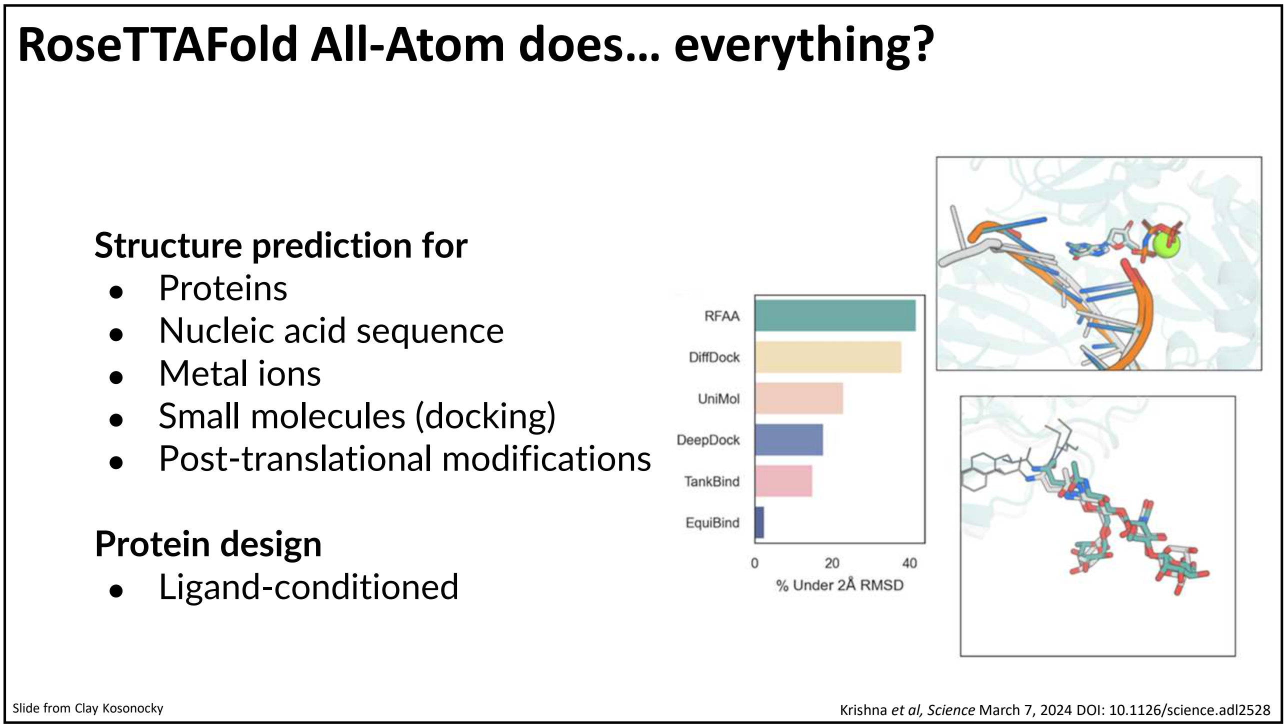


…they are in the new RoseTTAFold All-Atom release

Protein Sequence MVETSQLRY
Nucleic Acid Sequence ATCGGA
Metal Ion Fe
Small Molecule smiles, mol2, sdf
Covalently Modified Residue (N8,NDE),(NAG,C1)
RoseTTAFold All-Atom (36 main blocks + 4 refinement layers)
Protein Structure
Protein Nucleic Acid Complex
Protein Metal Complex
Protein Small Molecule Complex
Covalently Modified Protein Structure

Slide from Clay Kosonocky

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



RoseTTAFold All-Atom does… everything?

Structure prediction for

- Proteins
- Nucleic acid sequence
- Metal ions
- Small molecules (docking)
- Post-translational modifications

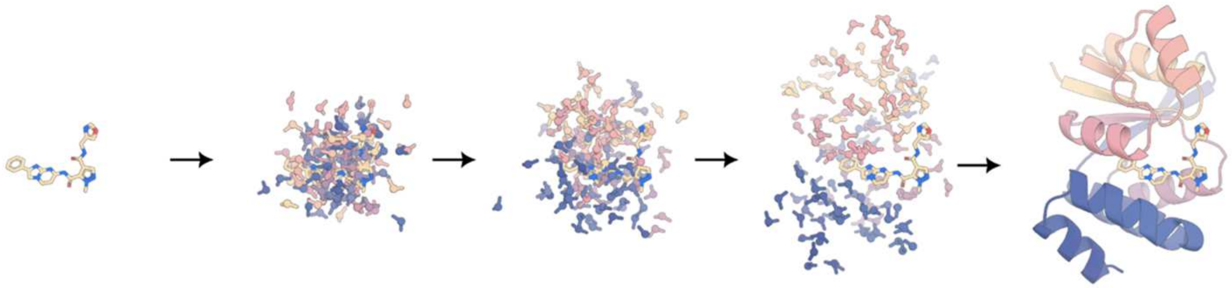
Protein design

- Ligand-conditioned

Slide from Clay Kosonocky

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

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

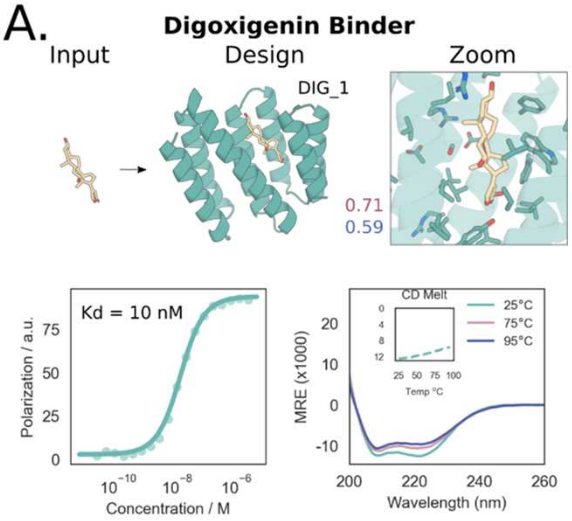


RFAA can be converted to a diffusion model just like RoseTTAFold

Slide from Clay Kosonocky

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

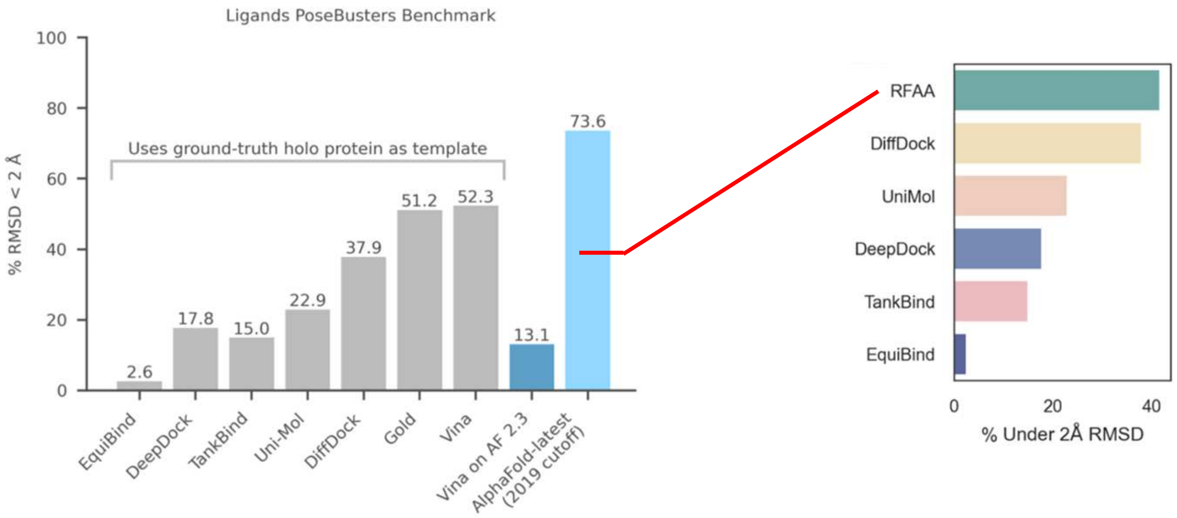
RFdiffusionAA ligand binder design: Digoxigenin



Slide from Clay Kosonocky

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

But wait… AlphaFold All Atom wins? (However, it isn't available publically and hasn't been independently tested)



Slide from Clay Kosonocky

Figure from Krishna et al. 2023 & Deepmind

Where does this bring us?

- Protein design is getting better and better
- Conditional generation options growing
  - Motifs, ligands, active sites, protein binding, etc.
- Challenges
  - **Needs broader experimental validation**
  - Designing around conformation changes
  - Antibody-antigen designs have so far not been generally solved for high affinity binders
  - Non-immunogenic designs

Modified from Clay Kosonocky