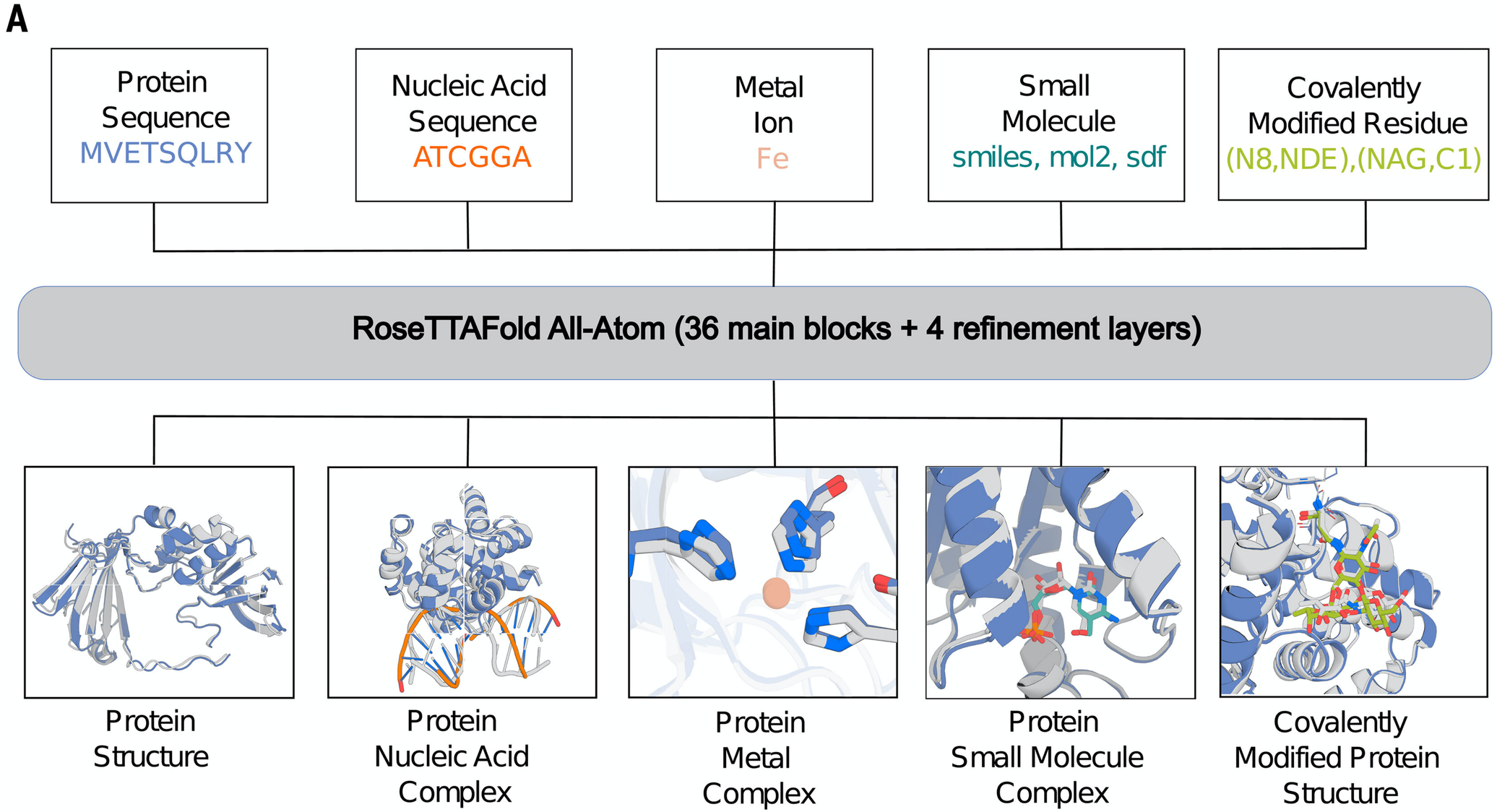
Diffusion Models III

RoseTTAFold All-Atom for bimolecular modeling and RFDiffusion AA for protein design

October 10th, 2024

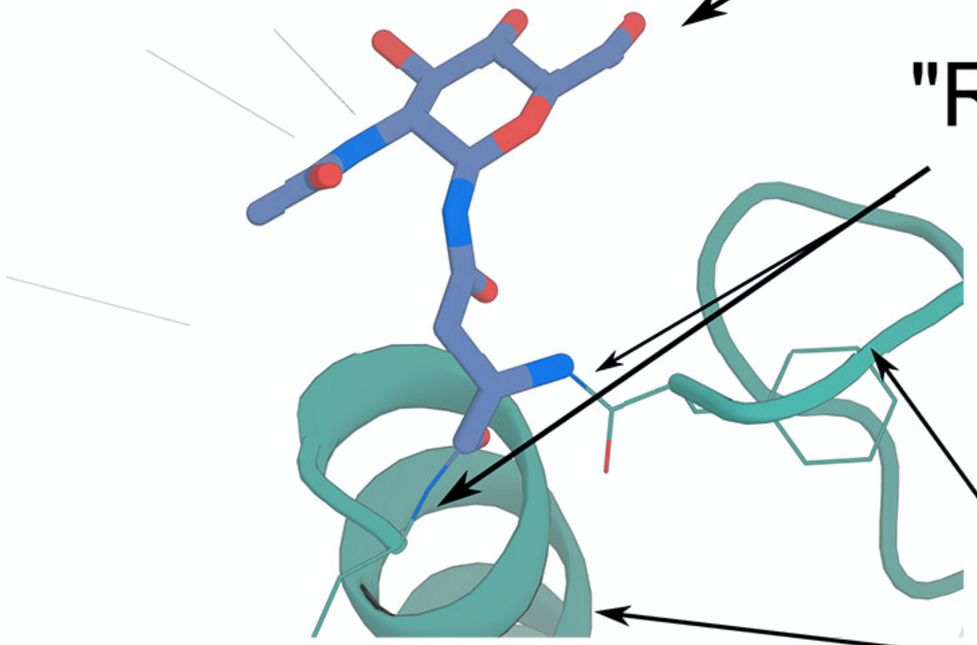


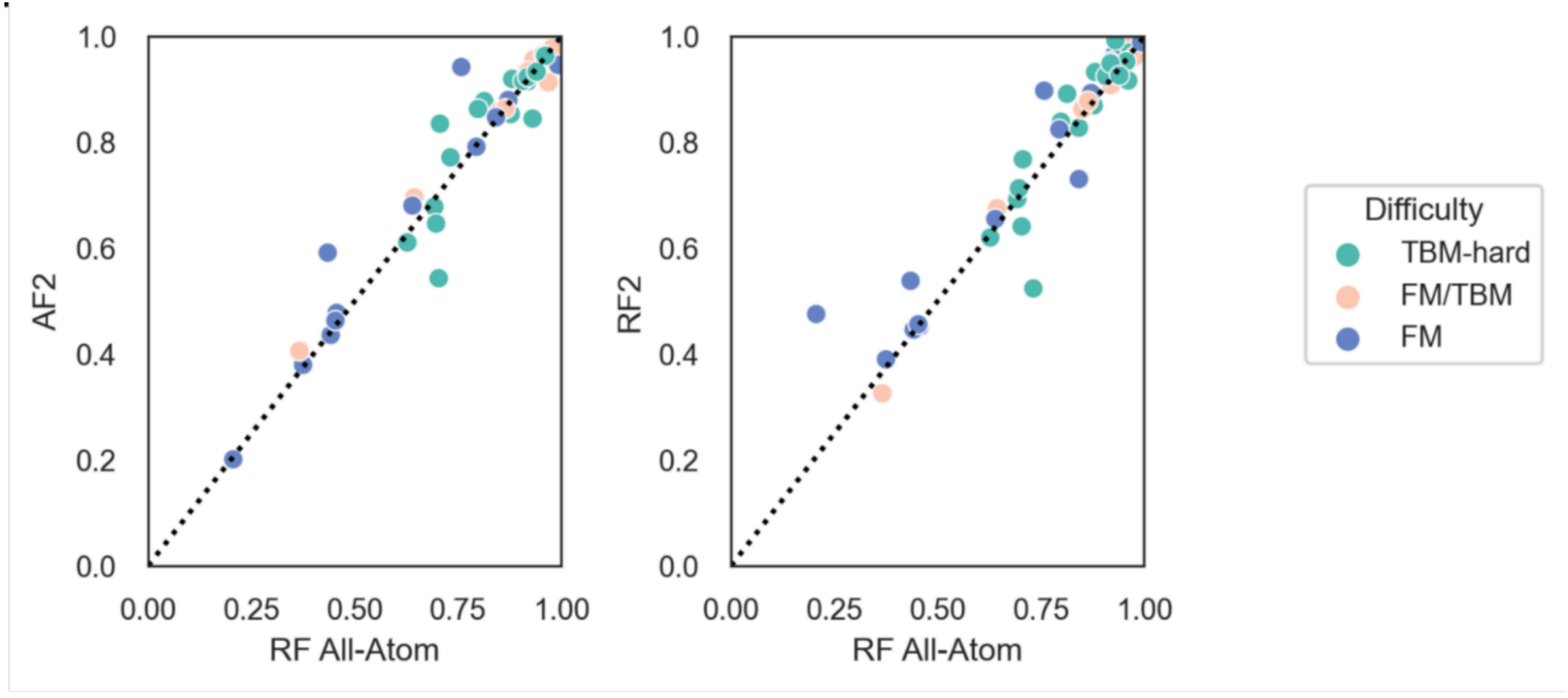
RF All-Atom can predict structures involving proteins, nucleic acids, and small molecules.

RoseTTAFold All-Atom

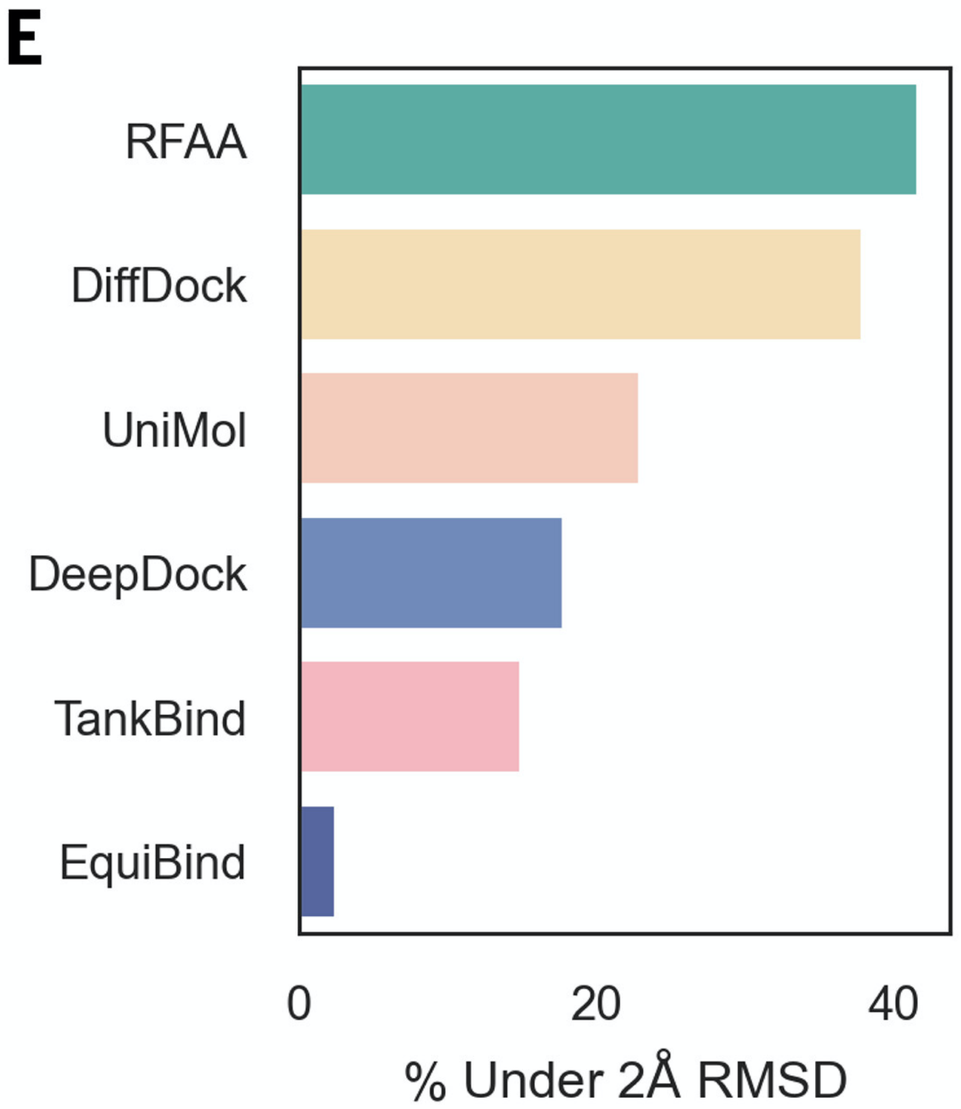
- RF2 architecture was updated to also accommodate small molecules
- Changes to the tracks
  - 1D track: chemical element of non-polymer atoms and supplemented residues and nucleic acid representation with 46 new element-type tokens
  - 2D track: atom bond information
  - 3D: chirality information
- Assigns frames for atoms in arbitrary molecules to use all atoms frame aligned point error
- Structure refinement now acts on cloud of frames and non-polymer atoms
- Data augmentation by “atomizing” protein residues (is also used for covalently modified residues)

A


Covalent modification and residue treated as atoms
"Residue to Atom" Bond Features
Remainder of protein treated as residues



RF All-Atom replicates results on CASP14 target (GDT_TS).



RFAA performs better on protein-ligand complexes compared to other machine learning methods. (PoseBusters dataset)

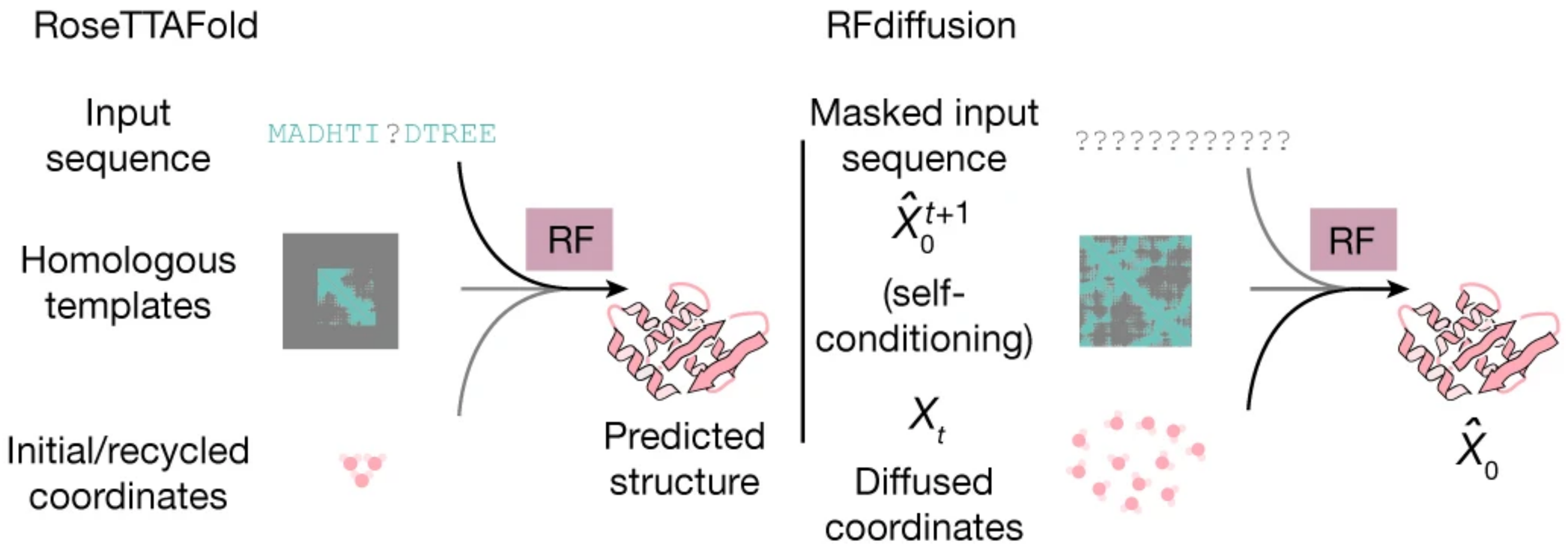
RFDiffusion and RFDiffusion AA

- RFDiffusion is a generative model for protein backbones built off of RoseTTAFold

- Previous backbone generation models (e.g. ProtDiff) suffered from poor designability: that is, very few of the generated backbones could be computationally validated by methods such as AlphaFold2

- RFDiffusion can generate highly designable backbones, and it also supports conditional generation for designing proteins.

- RFDiffusion AA expands on RFDiffusion by using RFAA to condition on other molecules.



RFdiffusion uses the much of the architecture of RF and its trained weights.

Architecture details

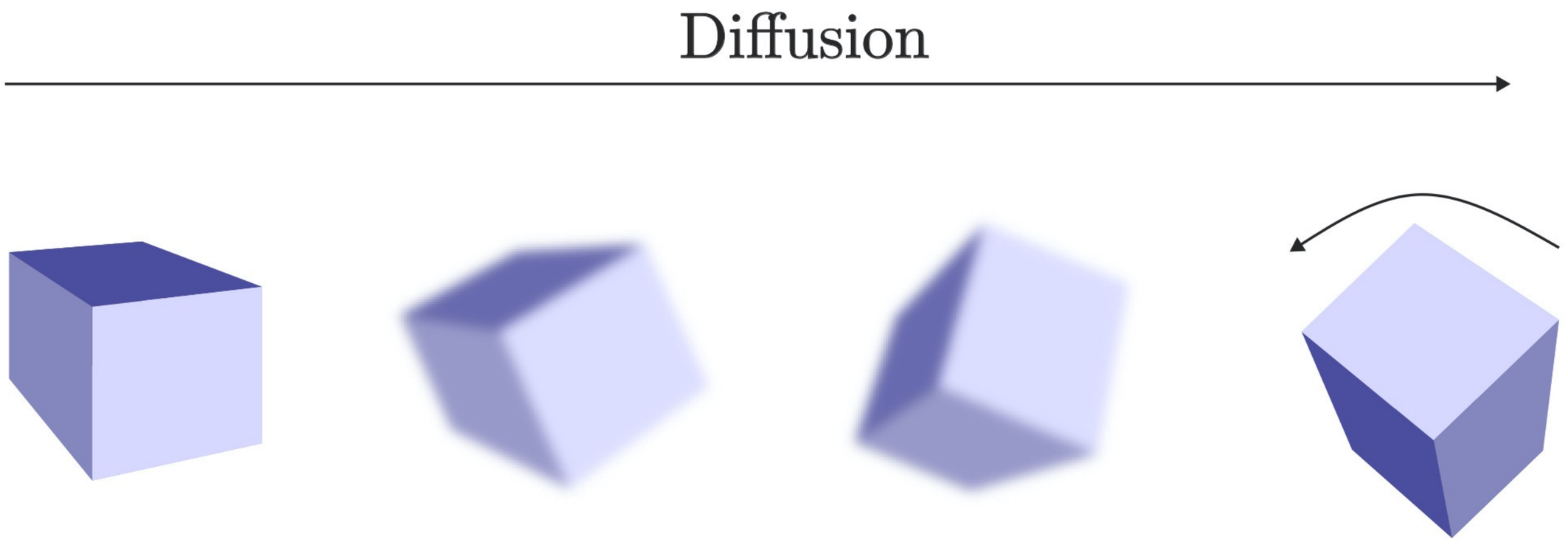
- The score network is parameterized so that the predicted denoised structure is predicted at each step.
  - $\mathrm{RF}(\mathbf{X}_t) = \hat{\mathbf{X}}_0$ and the next step moves toward the predicted denoised structure
- Recall that RF represents proteins as frames (Ca atom + frame constructed from C, Ca, and N). Thus RFdiffusion must diffuse on frames, not just points
  - Adding noise to translations -> VP SDE
  - Adding noise to orientation -> ?
  - Orientations can be described by SO(3), so we would like to know how to implement a score model on this manifold

Riemannian Score-Based Generative Modelling

Valentin De Bortoli[*] [†], **Émile Mathieu**[*] [‡], **Michael Hutchinson**[*] [‡]

James Thornton[‡], **Yee Whye Teh**[‡], **Arnaud Doucet**[‡]

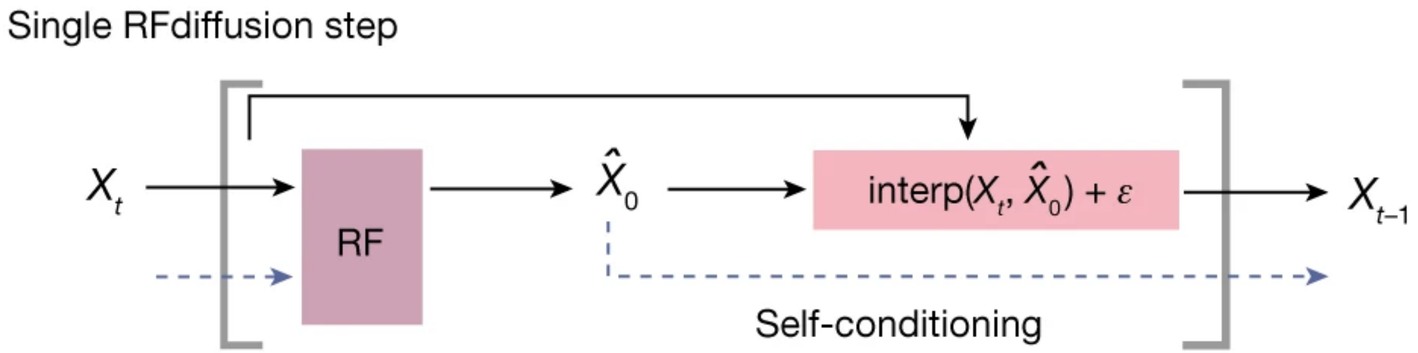
Describes how to implement SDE score-based models on Riemannian manifolds (e.g. SO(n)).

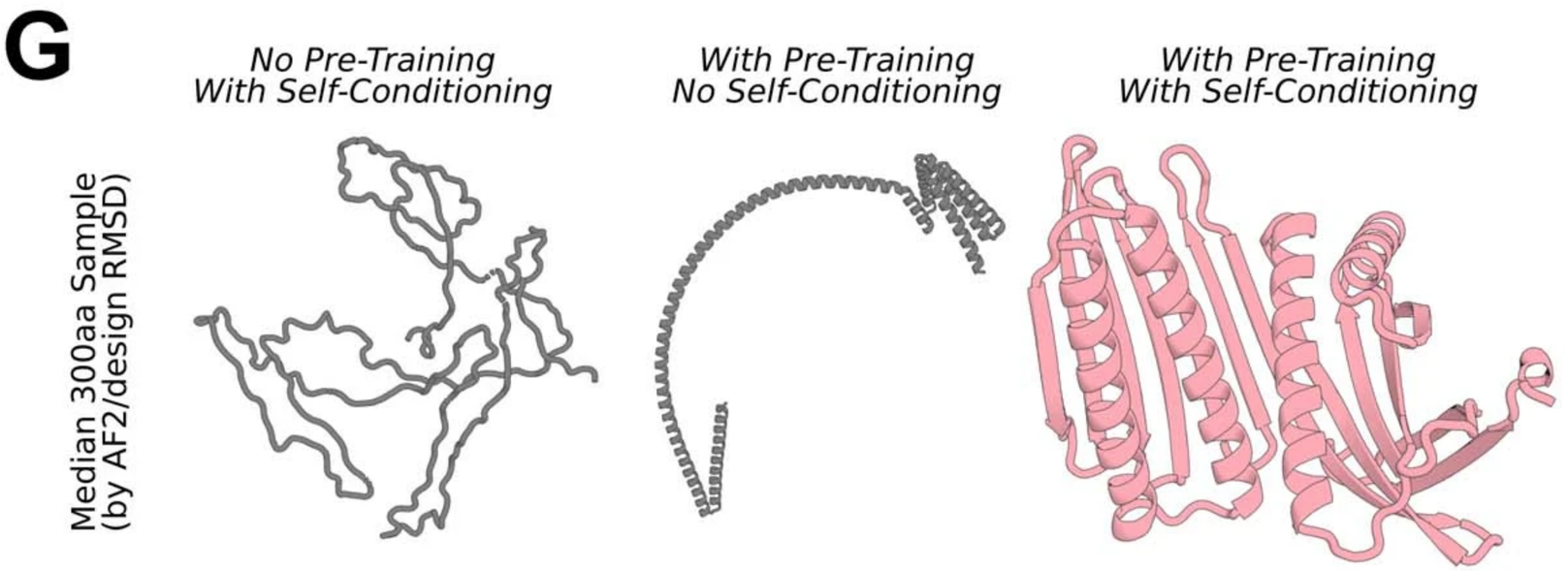


Example of diffusion in SO(3). Starting orientation is perturbed with IGSO(3). Limiting distribution is uniform

Self-conditioning

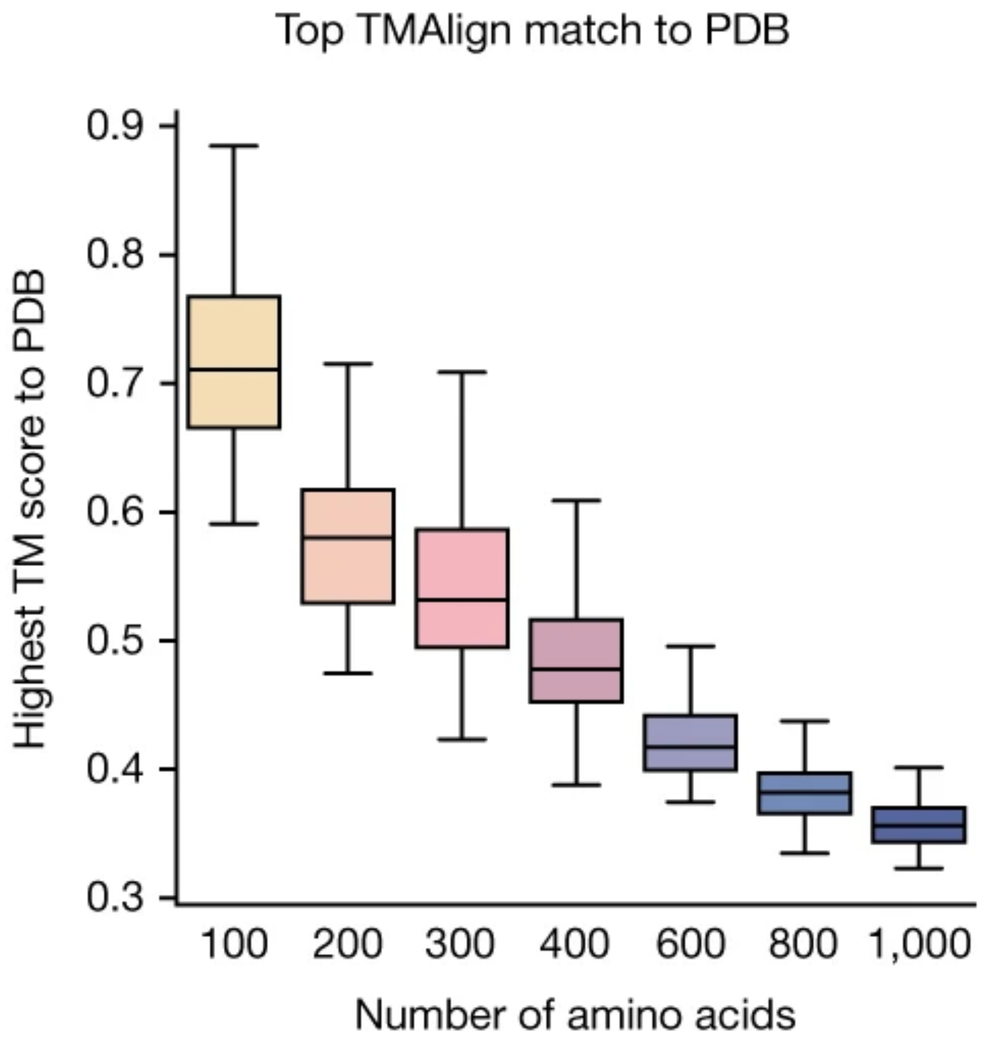
- Inspired by the recycling in AF2, RFdiffusion uses previous predictions of the denoised structure during later denoising steps.
- Self-conditioning uses previous guesses of the final structure for subsequent guesses.





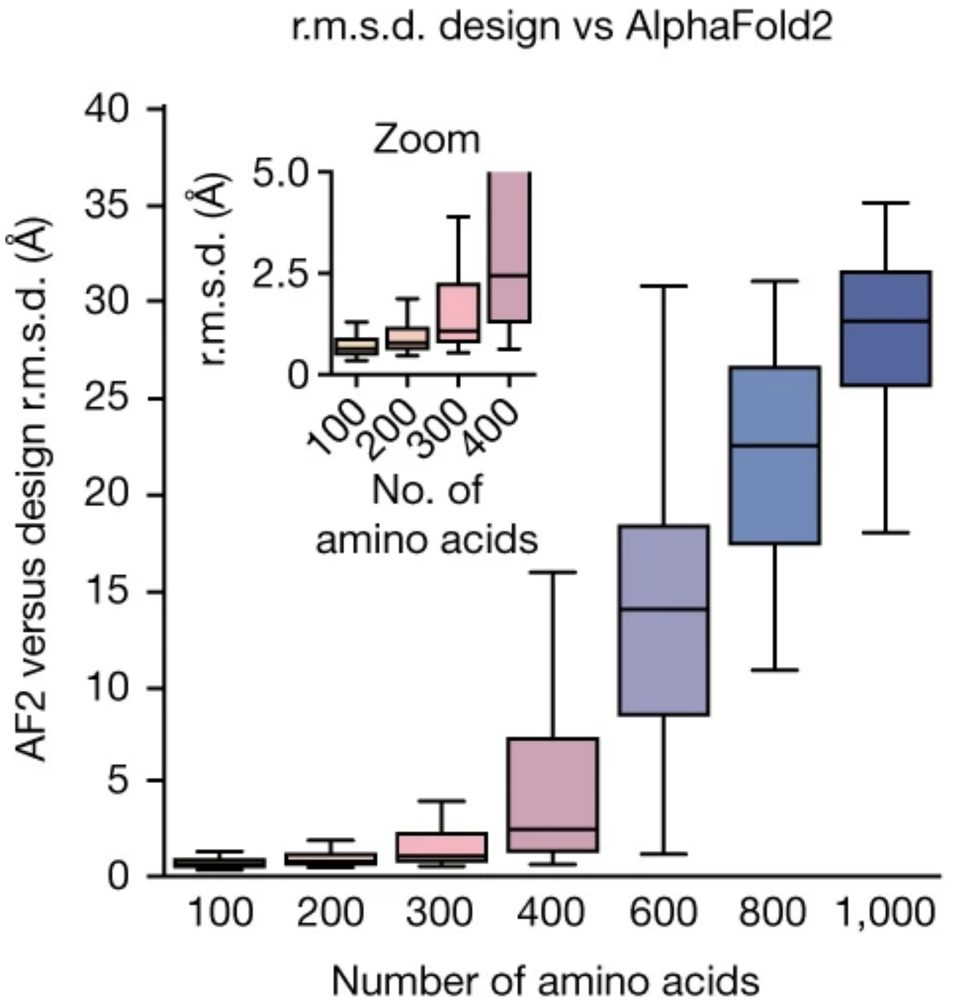
Median 300a samples by AF2 vs design RMSD. Without pre-trained RF weights, generated structures are not very protein-like. With no self-conditioning, secondary structure is represented but lack core-packing.

b



RFdiffusion produces novel backbone designs.

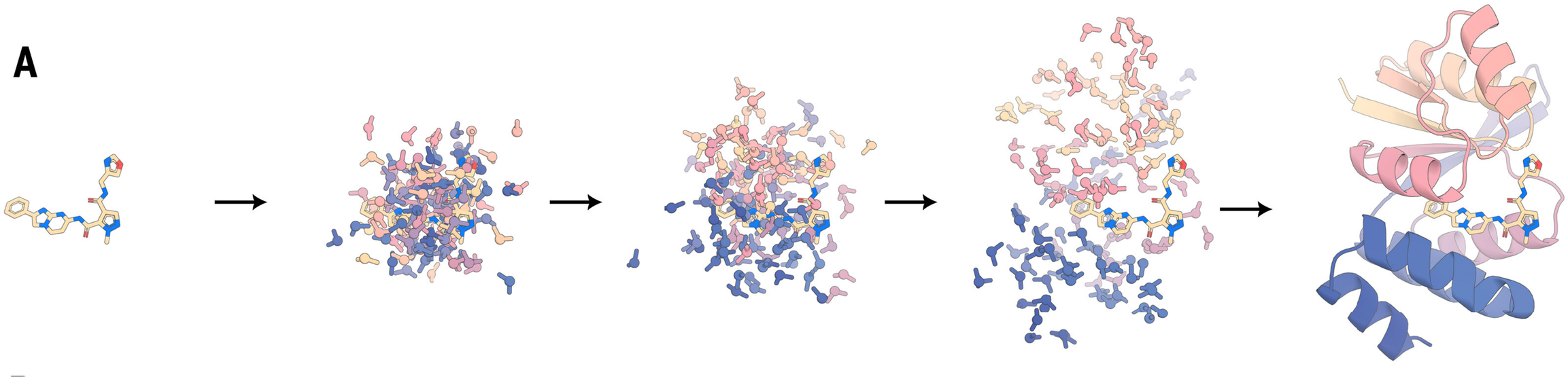
c



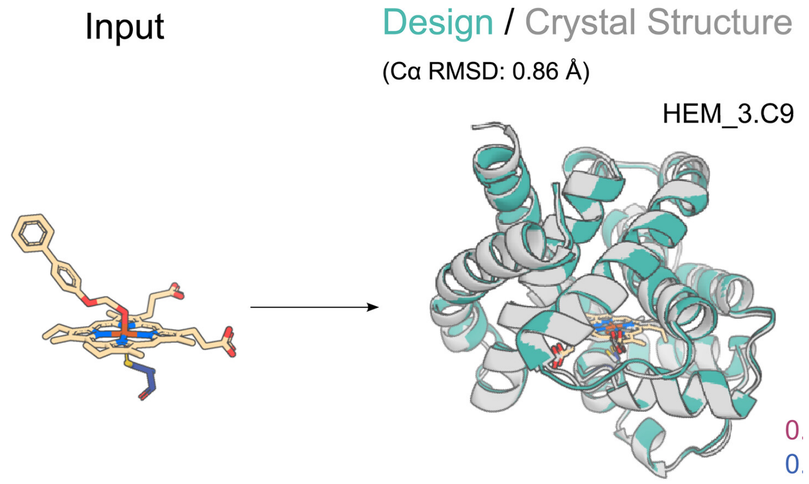
RFdiffusion results on monomeric generation. Refolding accomplished with ProteinMPNN and AF2.

RFdiffusion AA and conditional generation

- RFdiffusion AA follows the same framework as RFdiffusion but use the pertained weights from RFAA.
- RFdiffusion and RFdiffusion AA allow conditional generation of proteins by specifying the position of certain atoms/residues
  - Diffusion is done in an amortized way (motif is fixed and remaining atoms are denoised)
  - For RFdiffusion AA, this allows the design of proteins that bind to molecules.
  - The authors give a pipeline for designing a heme-binding protein. Designs were validated experimentally.



Example of generating protein structure conditioned on ligand.



Example design of heme-binding protein.