Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA

Baek et al.

Luke Elder

10/17/24

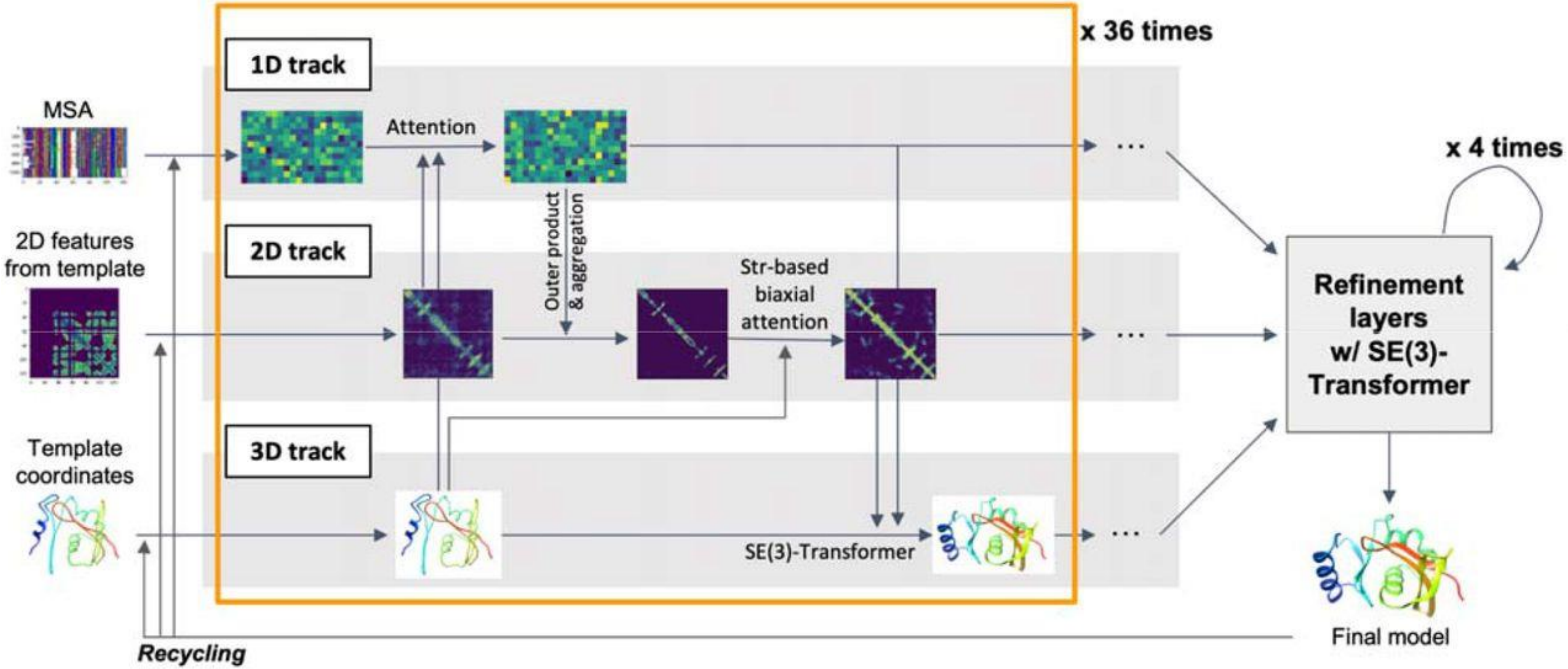
Outline

- Introduction
- RoseTTAFold2 overview
- Methods
- Results
- Conclusions

Introduction

- Previous approaches for protein-nucleic acid complex prediction involve building models of protein and nucleic acid (NA) components separately and then use docking to combine
    - Prediction of structure of complexes has lagged behind individual structures
- Goal: extend ideas of AF2 and RoseTTAFold to predict structure of nucleic acids and protein-nucleic acid complexes from sequence
    - Difficulty: Lack of data
- Train model with same data as RoseTTAFold2 augmented with RNA, protein-RNA, and protein-DNA complexes
    - Evaluate on more recently published complexes without homologs

RoseTTAFold2 Architecture


x 36 times
1D track
2D track
3D track
MSA
2D features from template
Template coordinates
Attention
Outer product & aggregation
Str-based biaxial attention
SE(3)-Transformer
x 4 times
Refinement layers w/ SE(3)-Transformer
Final model
Recycling

Efficient and accurate prediction of protein structure using RoseTTAFold2

Baek et al. bioRxiv 2023.05.24.542179; doi: https://doi.org/10.1101/2023.05.24.542179

RoseTTAFold2 architecture details

- Each of the 3 tracks is initialized through a series of embedding layers from initial MSA and template features
- 36 rounds of the main iteration where 1D, 2D and 3D tracks talk to each other
- 4 rounds of 3D update with frozen 1D and 2D data
- 0-3 Recycling passes before backprop pass (like AF2):
  - Directly for 3D
  - Used to modify 1D and 2D embeddings
- 3D track is SE(3) Transformer with fully connected graphs

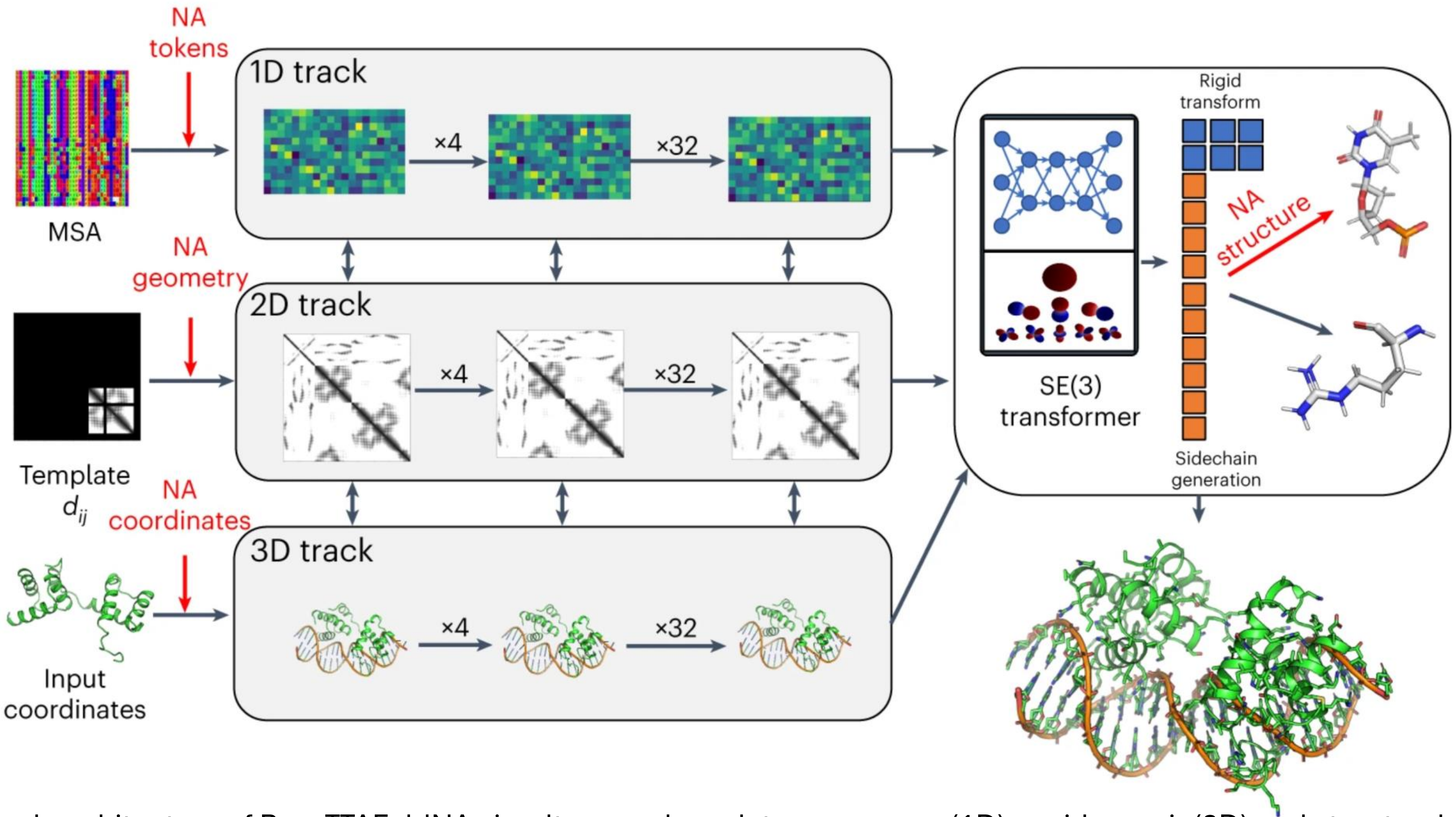
RoseTTAFold2 dataset and training

- All protein structures in the PDB published before April 30, 2020
    - 280k structures in 20k clusters
    - More permissive than RF
- ‘distilled’ protein structures
    - High confidence AF2 predictions
    - 3.6M sequence/structure pairs in 1.0M clusters
- Multimers dataset
    - Heteromeric interactions between different chains
    - Homomeric interactions between 2 copies of same chain
- Training is run using a 50%/25%/25% split between distillation data, PDB monomers/homoligomers, and PDB heterooligomers
- MSAs are randomly masked

RoseTTAFold2 predictions

- Backbone and sidechain predictions made every iteration
    - Made up of coordinates of backbone atoms and 10 total predicted angles
    - Only fed forward indirectly through state features (3D track)
- Auxiliary Heads:
    - Distogram and orientation from 2D track
        - Binned prediction of distance and 3 angle terms (5 total angles) – like trRosetta
    - Amino-acid logits from 1D track – used for masked AA prediction
    - pLDDT – binned lddt per residue prediction – from state features
    - pAE – error per residue – computed from pair features

RoseTTAFoldNA Architecture



The three-track architecture of RoseTTAFoldNA simultaneously updates sequence (1D), residue-pair (2D) and structural (3D) representations of protein–nucleic acid complexes. The areas in red highlight key changes necessary for the incorporation of nucleic acids: inputs to the 1D track include additional NA tokens, inputs to the 2D track represent template protein–NA and NA–NA distances (and orientations) and inputs to the 3D track represent template or recycled NA coordinates. Finally, the 3D track as well as the structure refinement module (upper right) can build all-atom nucleic acid models from a coordinate frame (representing the phosphate group) and a set of 10 torsion angles (six backbone, three ribose ring and one nucleoside). In this figure, d_{ij} are the template inter-residue distances, and SE(3) refers to the Special Euclidean Group in three dimensions.

Modifications from RoseTTAFold2 (RF2)

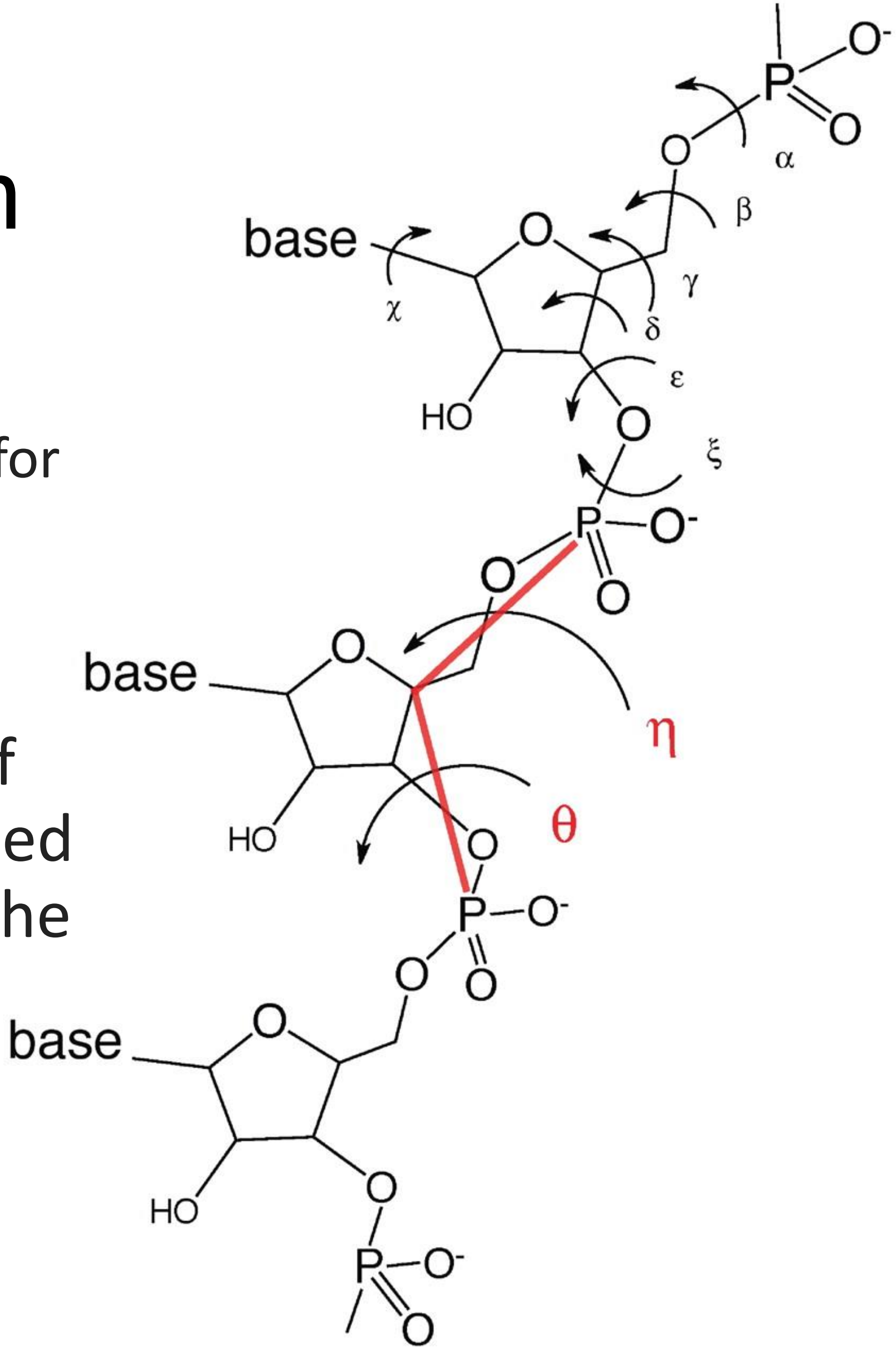
- Based on 3-track architecture of RoseTTAFold
  - Each track extended to support nucleic acids
- 1D track:
  - RF – 22 tokens (20 amino acids, 1 unknown or gap, 1 mask)
  - RFNA - 10 tokens added – 4 DNA nucleotides, 4 RNA nucleotides, 1 unknown for each
- 2D track:
  - RF – builds representation of interaction of all AA pairs
  - RFNA – generalized to model interactions between nucleic acid bases and AAs
- 3D track:
  - RF – represents position and orientation of each AA frame
  - RFNA – Define frame for nucleotides (3 atoms and 10 torsion angles)

Data processing

- Protein and protein complex data used is the same as for RF2
- Added RNA and protein-nucleic acid complexes
  - Include PDBs better than 4.5 Å resolution published before April 30, 2020
  - All RNA single chains, all RNA duplexes, all interacting protein-nucleic acid pairs
  - 7,396 (1632 clusters) RNA chains, 23,583 (1556 clusters) complexes – clustered and split into train and validation
- MSAs created for all protein and RNA sequences
- Added random nucleotide padding to DNA to improve generalizability – 580 protein-DNA complexes
- Test set: PDBs published May 1, 2020 or later
  - 91 complexes with one protein molecule plus a single RNA chain or DNA duplex
  - 43 cases with a single RNA chain
  - 106 cases with more than one protein chain or more than a single RNA chain or DNA duplex.

All atom nucleotide generation

- Represent each nucleotide as a rigid frame
  - Orientation of phosphate group O-P-O (N–Cα–C for AA)
  - Ten torsion angles: 6 backbone, 1 sidechain, three controlling ribose 'pucker' (v_0, v_1 and v_2)
- When all atom models are generated as part of the loss calculation, they are kinematically folded outward from the phosphate group following the chain of torsions connecting them.



Loss function

$$\text{loss} = w_{\text{seq}} \times \text{seq} + w_{\text{6D}} \times \text{6D} + w_{\text{str}} \times \text{str} + w_{\text{tors}} \times \text{tors} + w_{\text{err}} \times \text{err}$$

- seq is masked amino acid recovery loss
- 6D is 6 dimensional distogram loss
- str is structure loss
  - average backbone FAPE loss over all 40 structure layers of the network plus the all-atom FAPE loss for the final model
- tors is the torsion prediction loss averaged over the 40 structure layers
- err is the loss in pLDDT prediction
- w_{seq} = 3.0, w_{6d} = 1.0, w_{str} = 10.0, w_{tors} = 10.0 and w_{err} = 0.1

Fine tuning loss

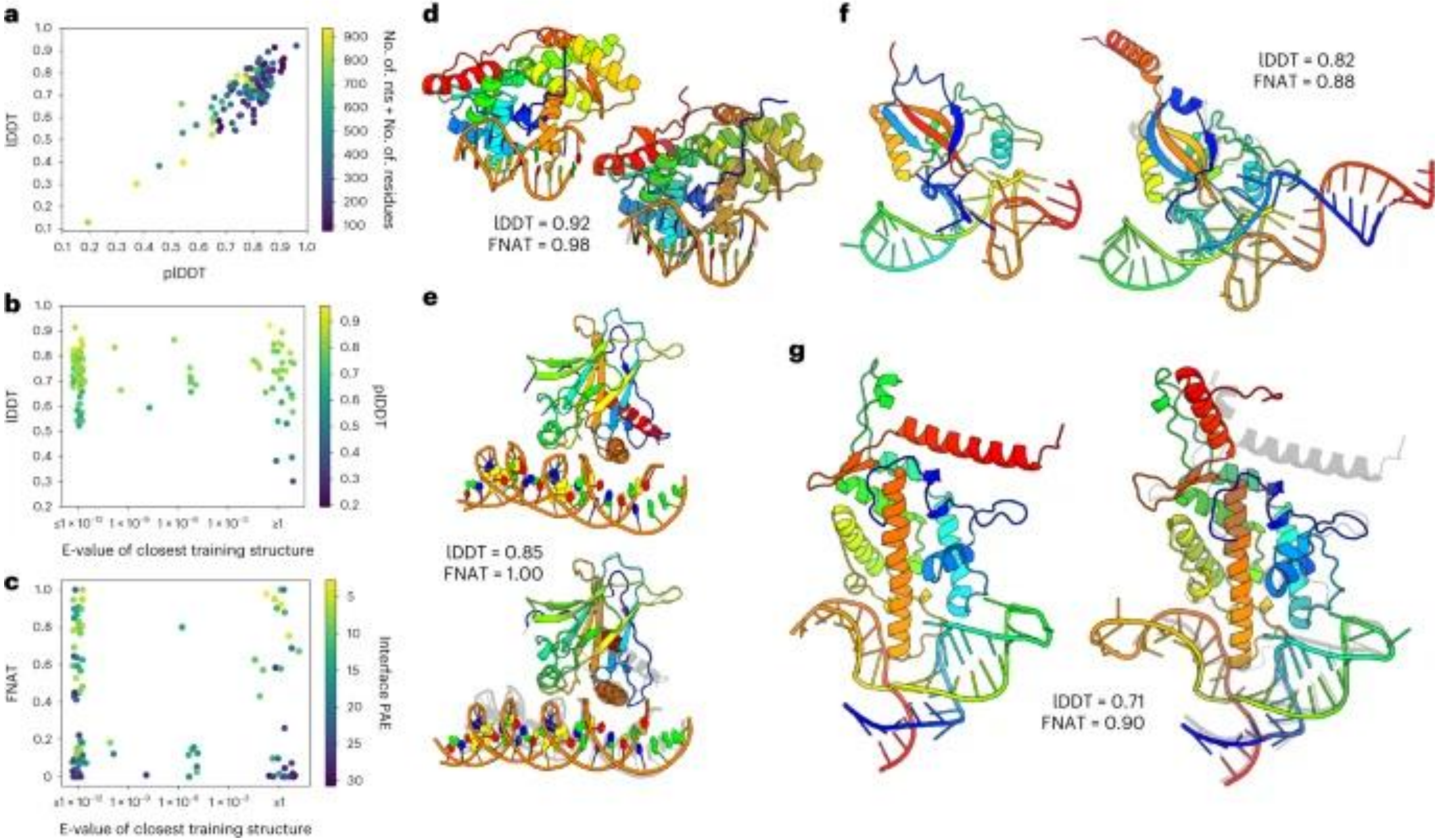
$$\begin{aligned}\text{loss}_{\text{finetune}} = \text{loss} + w_{\text{LJ}} \times \text{LJ} + w_{\text{hbond}} \times \text{hbond} \\ + w_{\text{geom}} \times \text{geom} + w_{\text{pairerr}} \times \text{pairerr}\end{aligned}$$

- LJ and hbond are Lennard-Jones and hydrogen bond energies of final structures
- geom enforces ideal bond lengths and bond angles
- pairerr is predicted residue-pair error
- $w_{\text{geom}} = 0.1$, $w_{\text{LJ}} = 0.02$, $w_{\text{hbond}} = 0.05$ and $w_{\text{pairerr}} = 0.1$

Model training

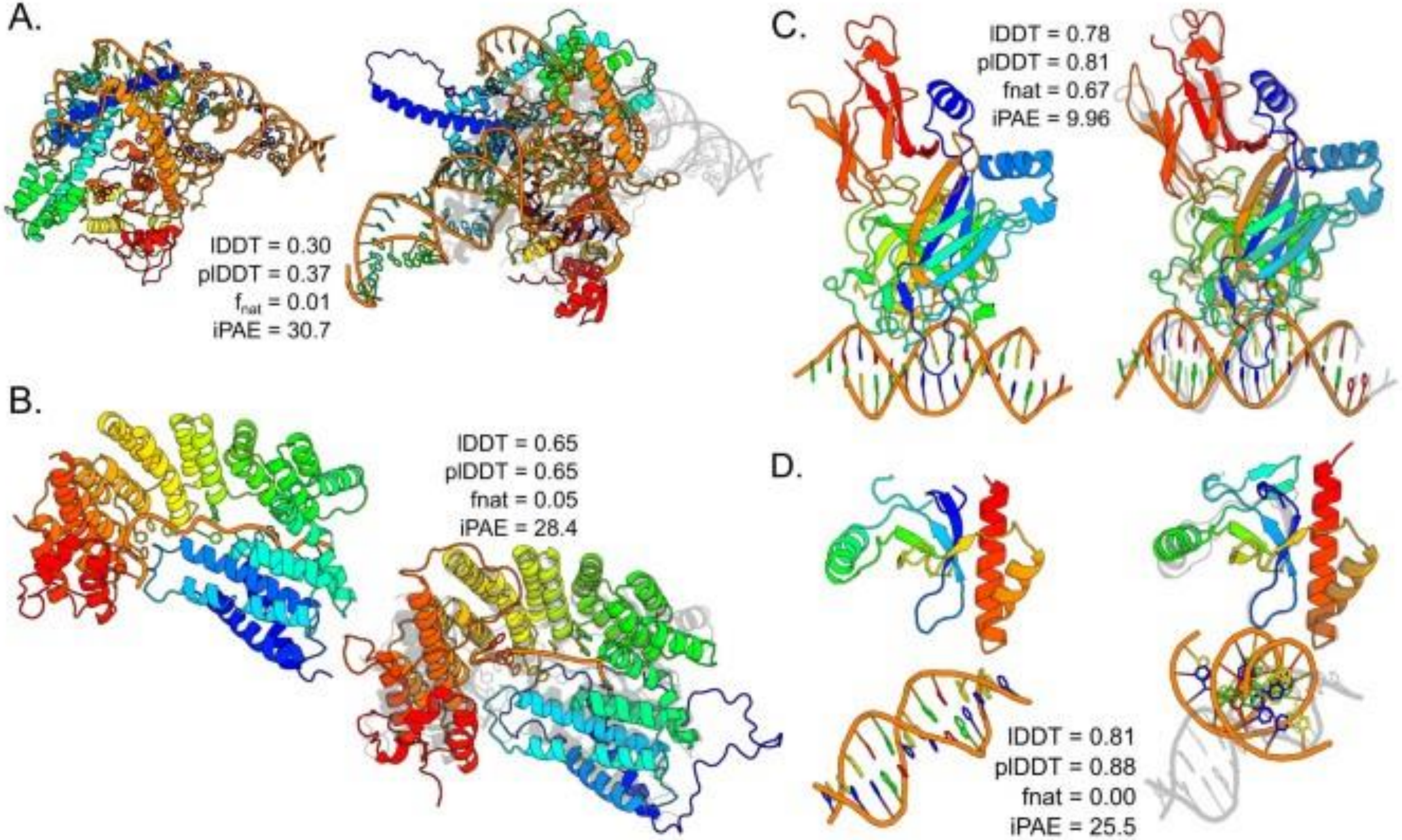
- 5 input pools sampled with equal probability:
  - Protein structures, ‘distilled’ protein structures (from AF2), protein complexes, protein-NA complexes, and RNA structures
- For both pools containing ‘complexes,’ an equal number of positive and negative examples were used in training
- Sequences cropped to 256 residues
- Batch size of 64 with learning rate of 0.001, decaying every 5,000 steps
  - The Adam optimizer was used, with L2 regularization (coeff = 0.01)
- After $\sim 1 \times 10^5$ optimization steps, fine-tuning training was carried out
  - Crop size = 384, batch size = 128, lr = 5×10^{-4}, and 30,000 steps
- Took 4 weeks on 64 GPUs

Predicting protein-NA complexes



a–**c**, Summary of results on 32 protein–NA cluster representatives from the validation set and 84 protein–NA structures released since May 2020. **d**–**g**, Four examples of protein–NA complexes without homologs in the training set

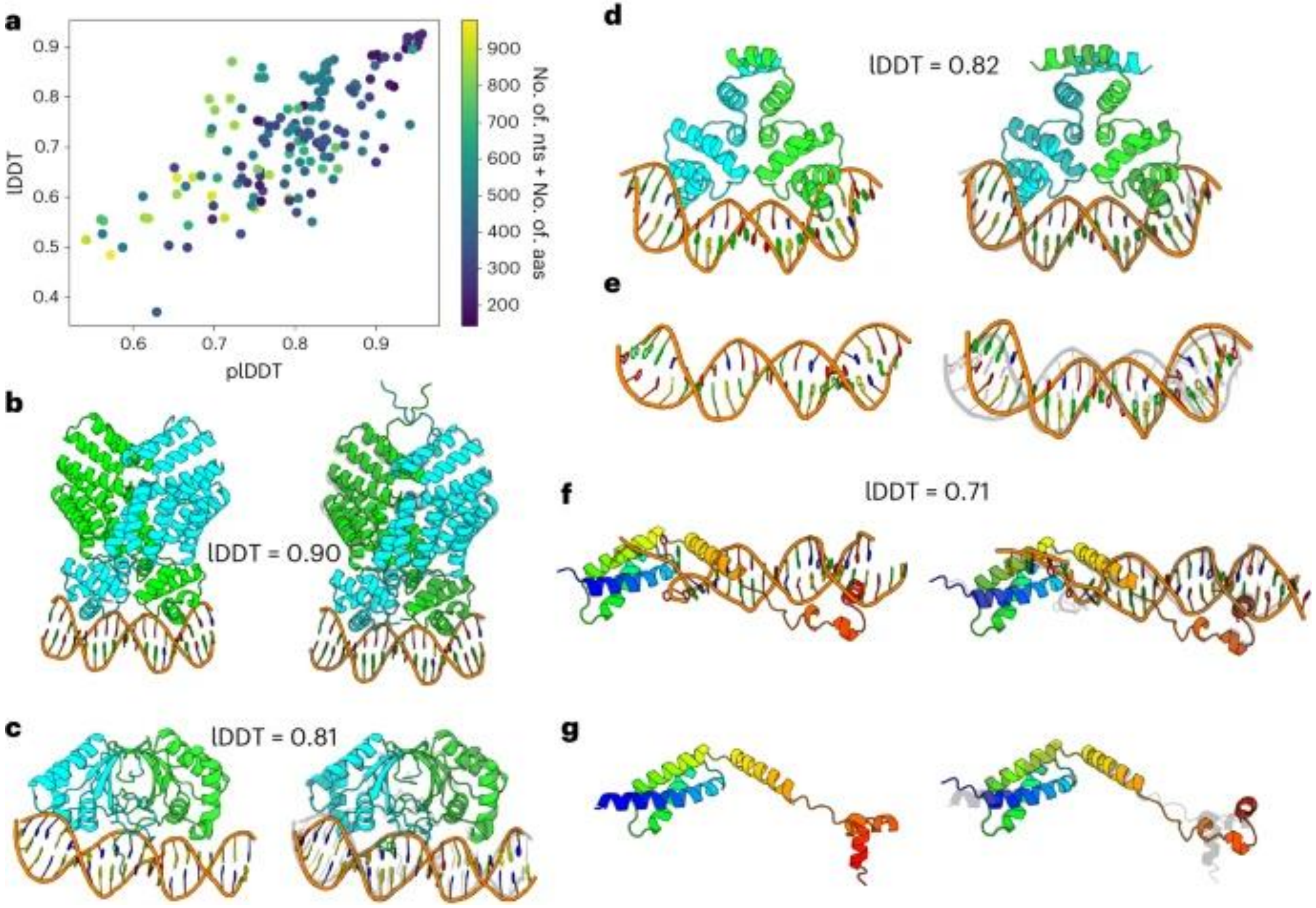
Protein-NA failure modes



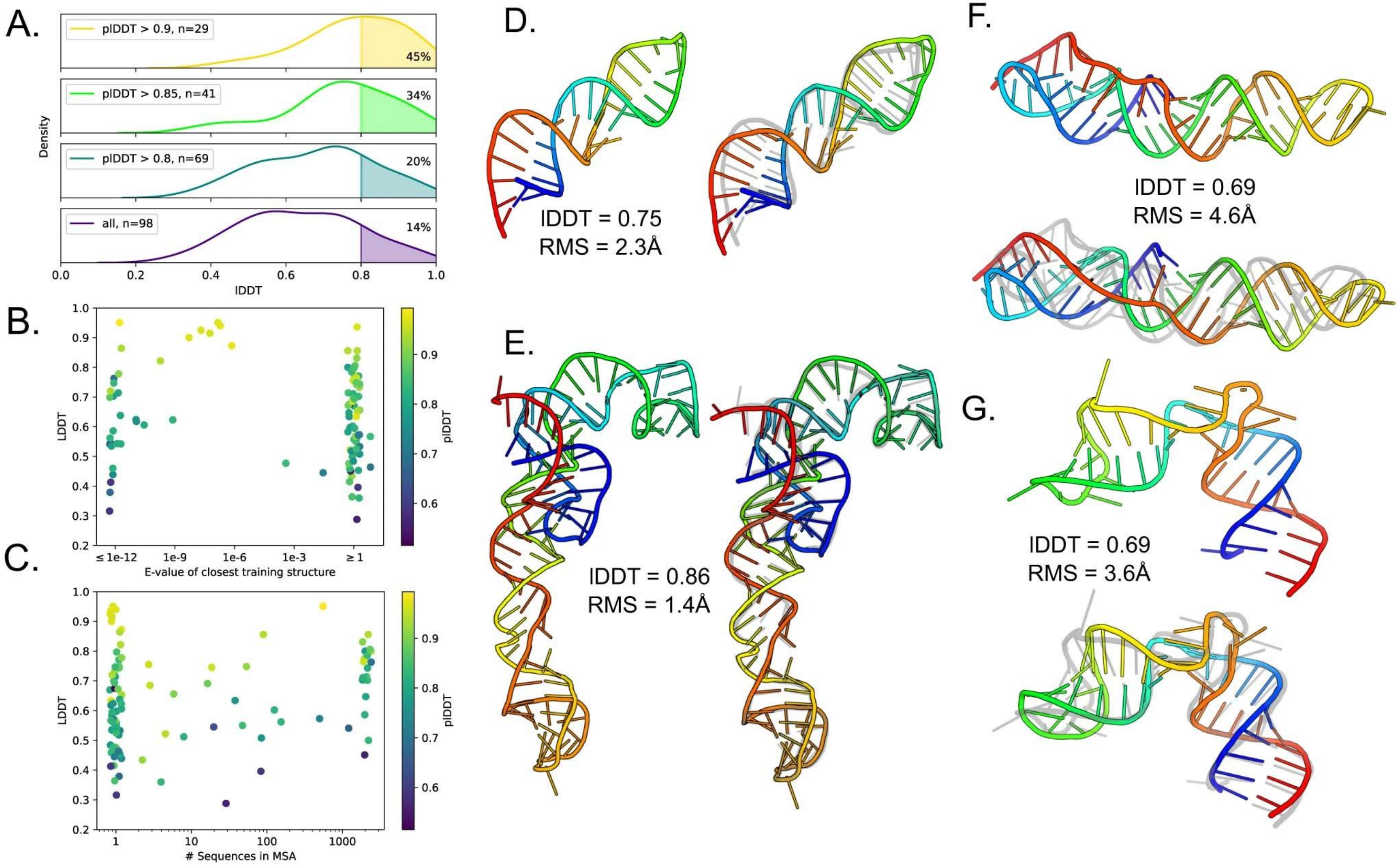
Comparisons of representative predictions showing common failure modes of predictions in cases with no training-set homologs. Left is the deposited model, and right is the prediction. (A) Example where the individual subunits predict with poor accuracy, resulting in an incorrect overall (50% of errors). (B) Example where the subunits predict with reasonable accuracy and the relative orientation is correct but the details of the interface are wrong (20%). (C) Example where the subunits predict with high accuracy and the backbone-backbone binding mode is correct, but the interface is predicted at the wrong site on the DNA (10%). (D) Example where both subunits predict correctly but the relative orientation and interface are incorrect (20%).

Multichain protein-NA complexes

a, Scatterplot of predicted model accuracy versus actual model accuracy for 161 protein–NA complexes with multiple protein chains or multiple nucleic acid chains/duplexes. **b**–**d**,**f**, Examples of successful predictions without homologs in the training set, shown as the deposited model (left) and prediction (right). **e**,**g**, Example showing different predicted conformations of the same protein or DNA duplex alone (left) and with the other component (right), from the same complexes shown in **d** (**e**) and **f** (**g**).

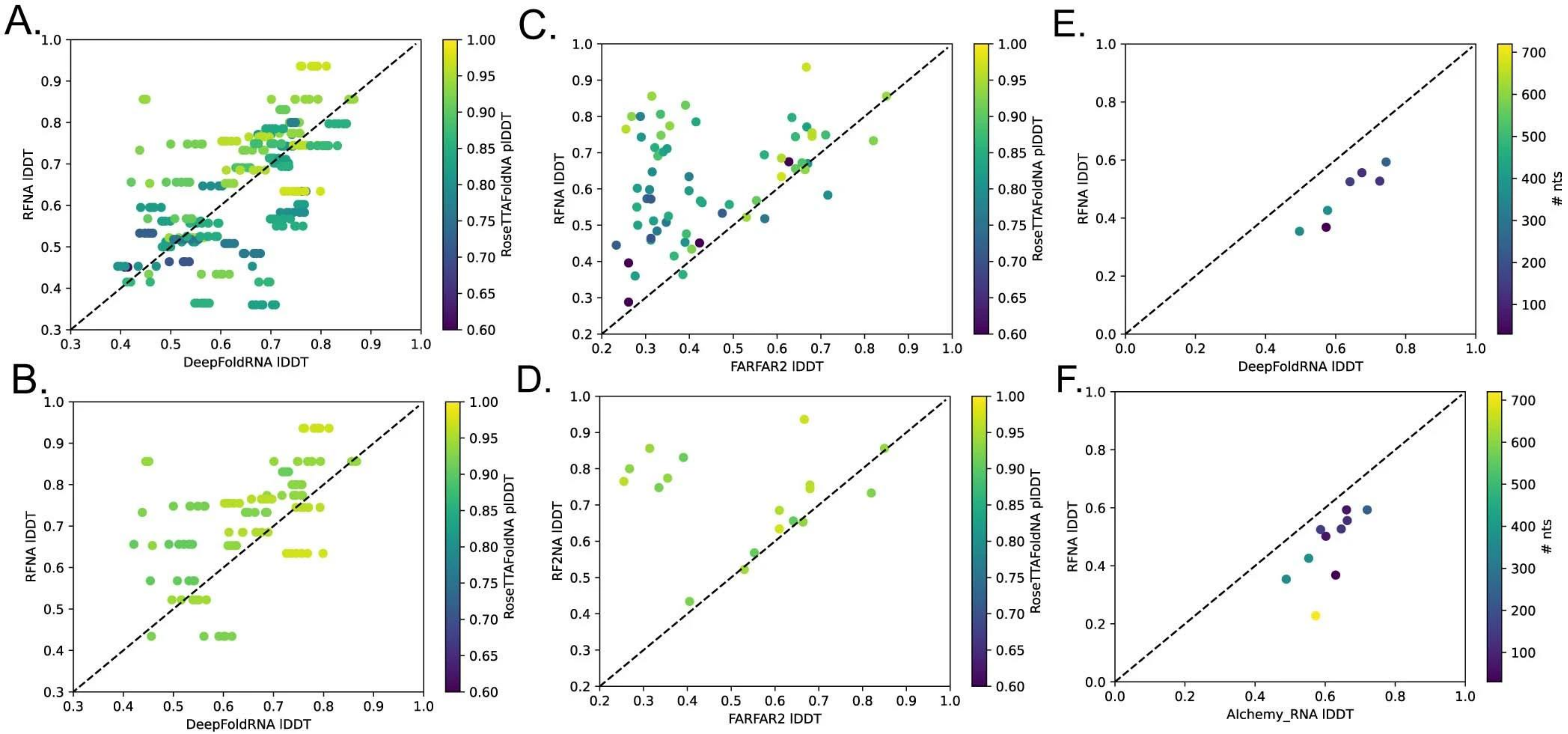


RNA structure prediction



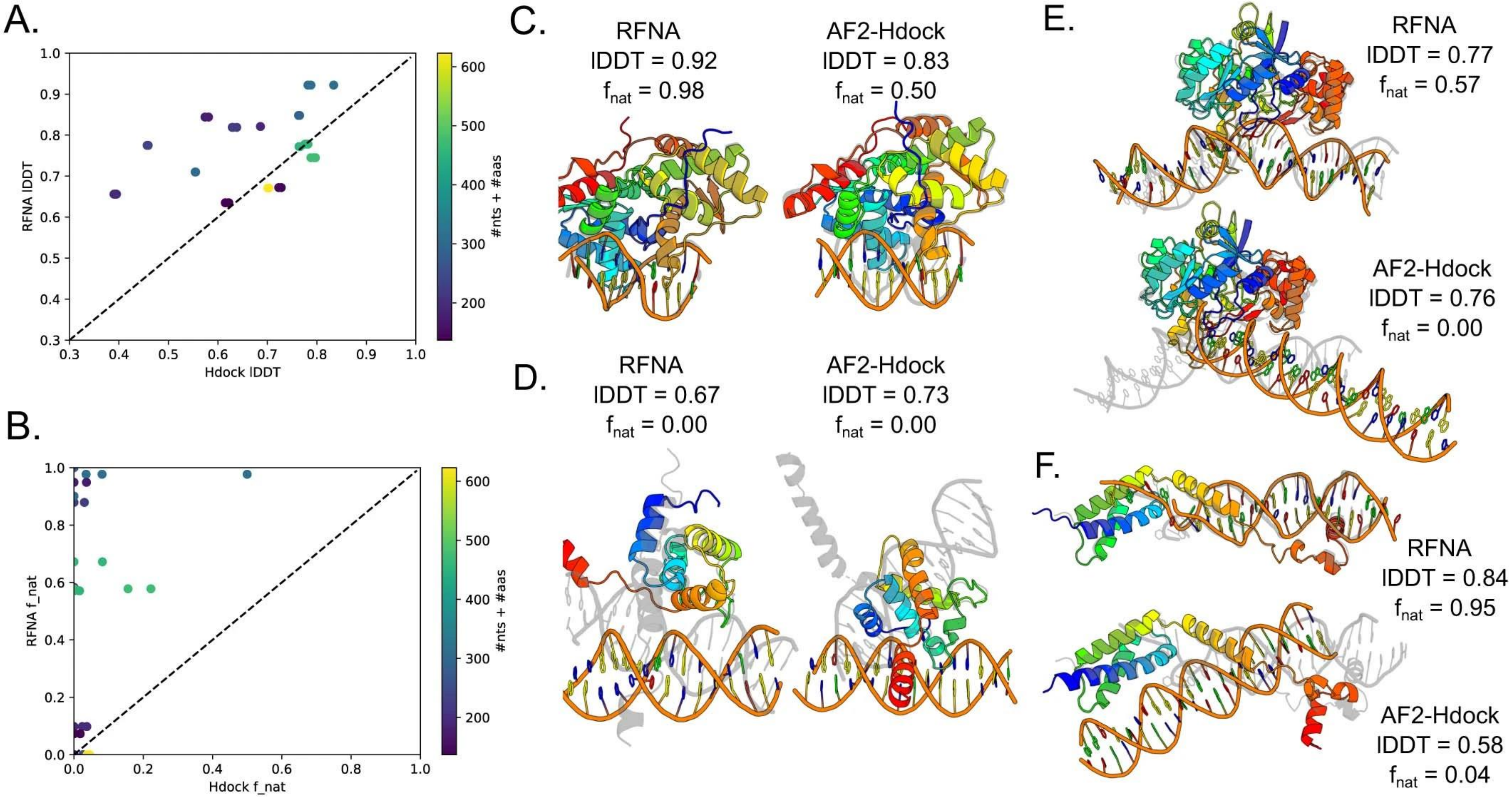
(**a**–**c**) Summary of results on 55 RNA cluster representatives from the validation set and 43 RNA structures released since May 2020. Overall average lDDT is 0.64 (**d**–**f**) Four example predictions of RNA models with no detectable sequence homologs in the training set

RoseTTAFoldNA vs other methods for RNA



(**e**, **f**) Comparisons between RoseTTAFoldNA and other machine learning methods on the CASP15 RNA targets (using model 1 of each method). RFNA performs somewhat worse than DeepFoldRNA and significantly worse than AIchemy_RNA, the leading machine learning method from the competition.

RoseTTAFoldNA vs Hdock



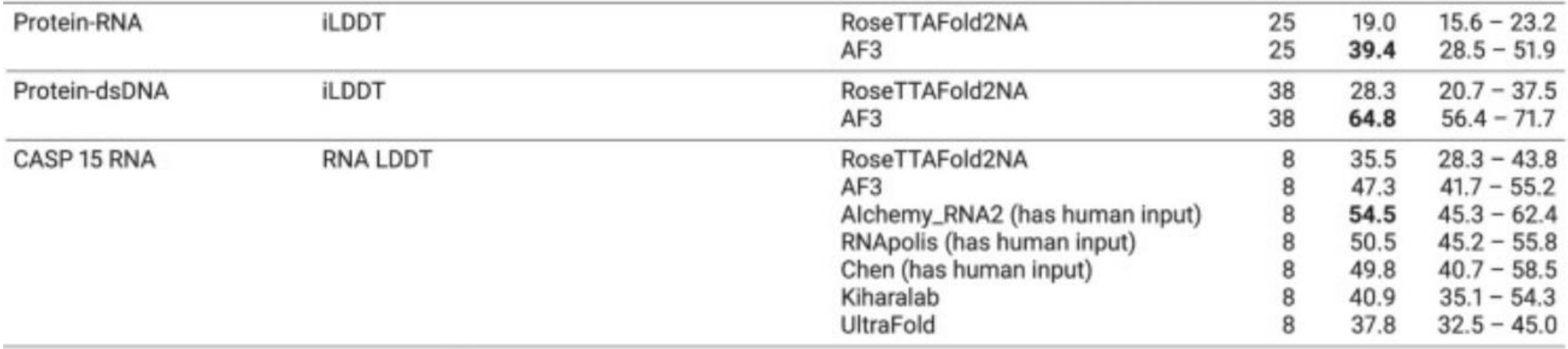
Conclusion

- At outset, unclear if there was enough data for protein-NA structures
  - Results show that it is sufficient for accurate prediction in roughly 31% of cases
- For RNA prediction, still perform worse than the state of the art on CASP15
  - Most targets are large and several are synthetic with no MSAs
- Protein structure prediction – 0.87 TM-score vs 0.88 for AF2
- Strength of model is predicting protein-NA complexes
  - Comparisons more difficult since no equivalent methods (pre AF3)
  - Performs much better than traditional docking –> improvement on state of the art

Future directions they suggest

- Larger more expressive network may improve things
  - They use 40 layer network with ~67M parameters
  - AF2 is 93M parameters
- Use of high-confidence predicted structures
  - Similar to ‘distilled’ dataset for proteins

Comparison with AF3



Protein-RNA	iLDDT	RoseTTAFold2NA	25	19.0	15.6 – 23.2
		AF3	25	**39.4**	28.5 – 51.9
Protein-dsDNA	iLDDT	RoseTTAFold2NA	38	28.3	20.7 – 37.5
		AF3	38	**64.8**	56.4 – 71.7
CASP 15 RNA	RNA LDDT	RoseTTAFold2NA	8	35.5	28.3 – 43.8
		AF3	8	47.3	41.7 – 55.2
		Alchemy_RNA2 (has human input)	8	**54.5**	45.3 – 62.4
		RNApolis (has human input)	8	50.5	45.2 – 55.8
		Chen (has human input)	8	49.8	40.7 – 58.5
		Kiharalab	8	40.9	35.1 – 54.3
		UltraFold	8	37.8	32.5 – 45.0

- RFNA represented a large improvement in protein-NA structure prediction but AF3 overshadows RFNA with more generalizable and accurate model published just 6 months later

Questions?