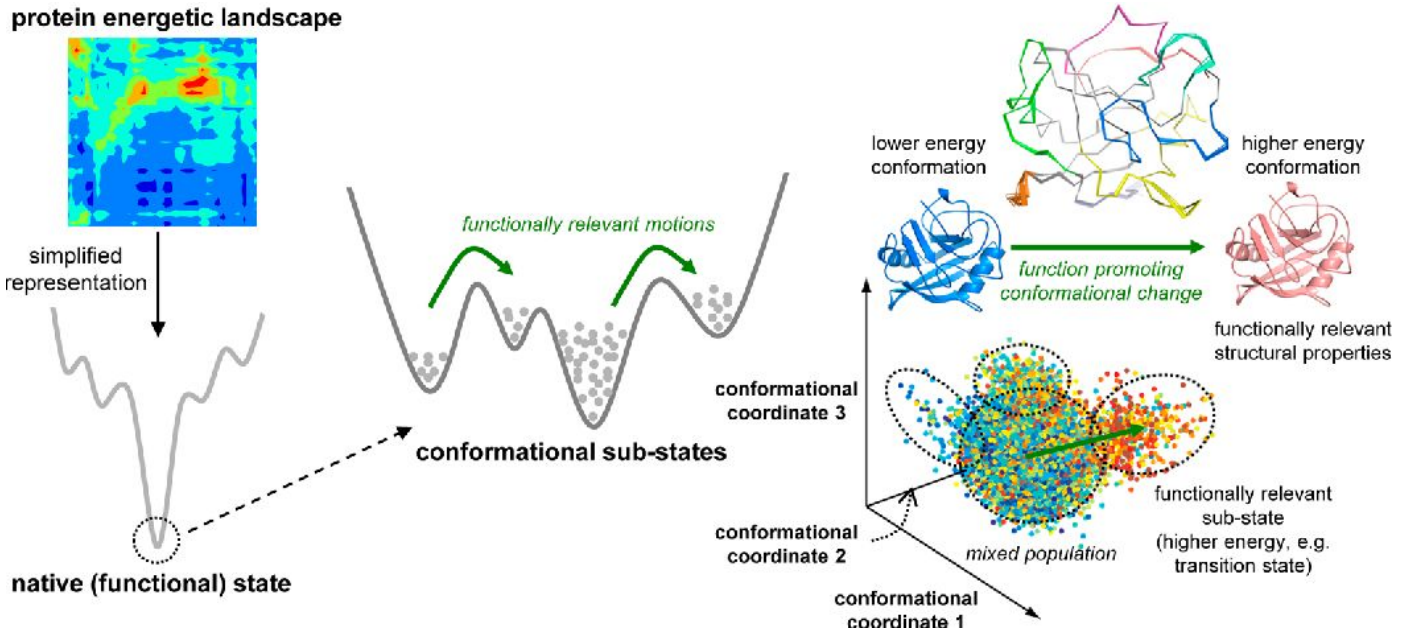
# Predicting multiple conformations via sequence clustering and AlphaFold2

Hannah K. Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M. Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell & Dorothee Kern

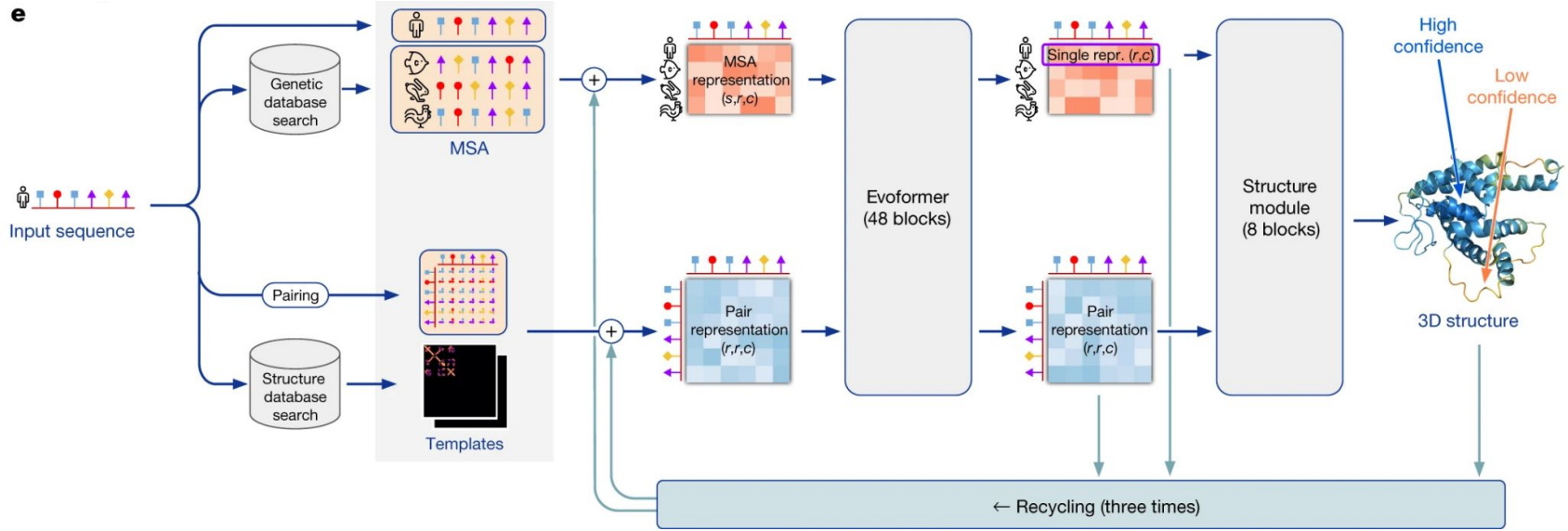
**Gökçe Uludoğan**
PhD Candidate
**LifeLU Reading Group | 12 September 2024**

# Understanding the mechanics of any protein's functions requires understanding its conformational substates




Ramanathan, A., Savol, A.J., Burger, V.M., Chennubhotla, C.S., & Agarwal, P.K. (2014). Protein conformational populations and functionally relevant substates. *Accounts of chemical research, 47 1,* 149-56 .

# AlphaFold2 advanced protein single-structure prediction by inferring interaction patterns between related sequences in a multiple-sequence alignment (MSA)




Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *nature* 596.7873 (2021): 583-589.

# Evolutionary couplings for multiple states are present in MSAs

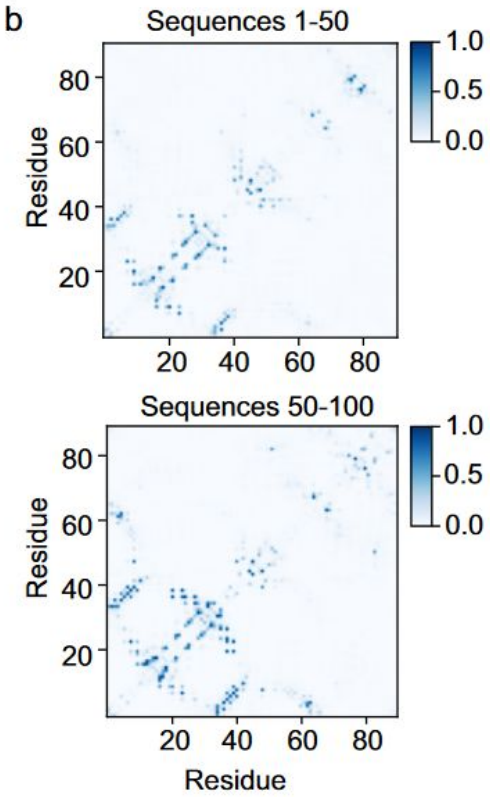
- Amino acids **co-evolve**, reflecting underlying protein structure.
- Proteins evolve considering **multiple conformational states**.
- AF2 excels at **single-structure prediction** but struggles with **multiple conformations.**
- **MSA subsampling** allows AF2 to predict conformational changes in transporters.
- Successful MSA subsampling suggests **multiple state contacts are present in complete MSAs.**
- Methods are needed to **separate signals from multiple states in MSAs.**

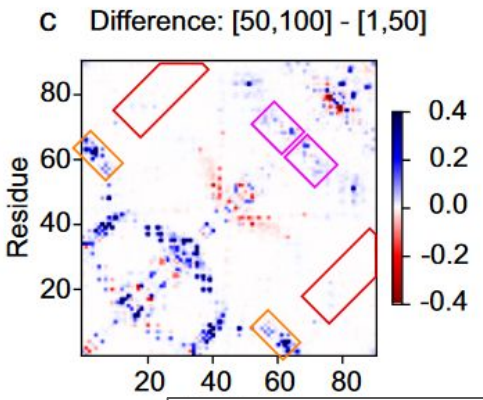
# Metamorphic proteins

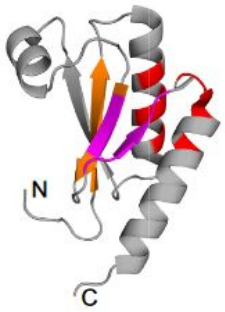
- Proteins occupying **multiple distinct secondary structures** for biological function
- **Model proteins** for conformational ensemble prediction methods, due to **significant structural changes**
- **Example:**

  KaiB (108 residues) switches between thioredoxin-like and alternative conformations, affecting ~40 C-terminal residues

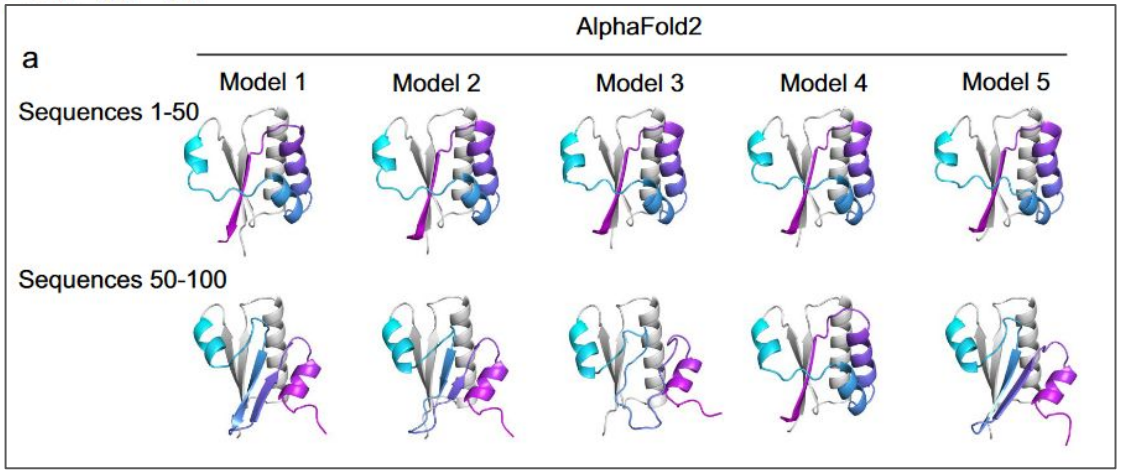
# Investigating two highly-similar sets of sequences in the KaiB$^{TE}$ MSA: Contacts predicted by MSA-Transformer and AF2 models



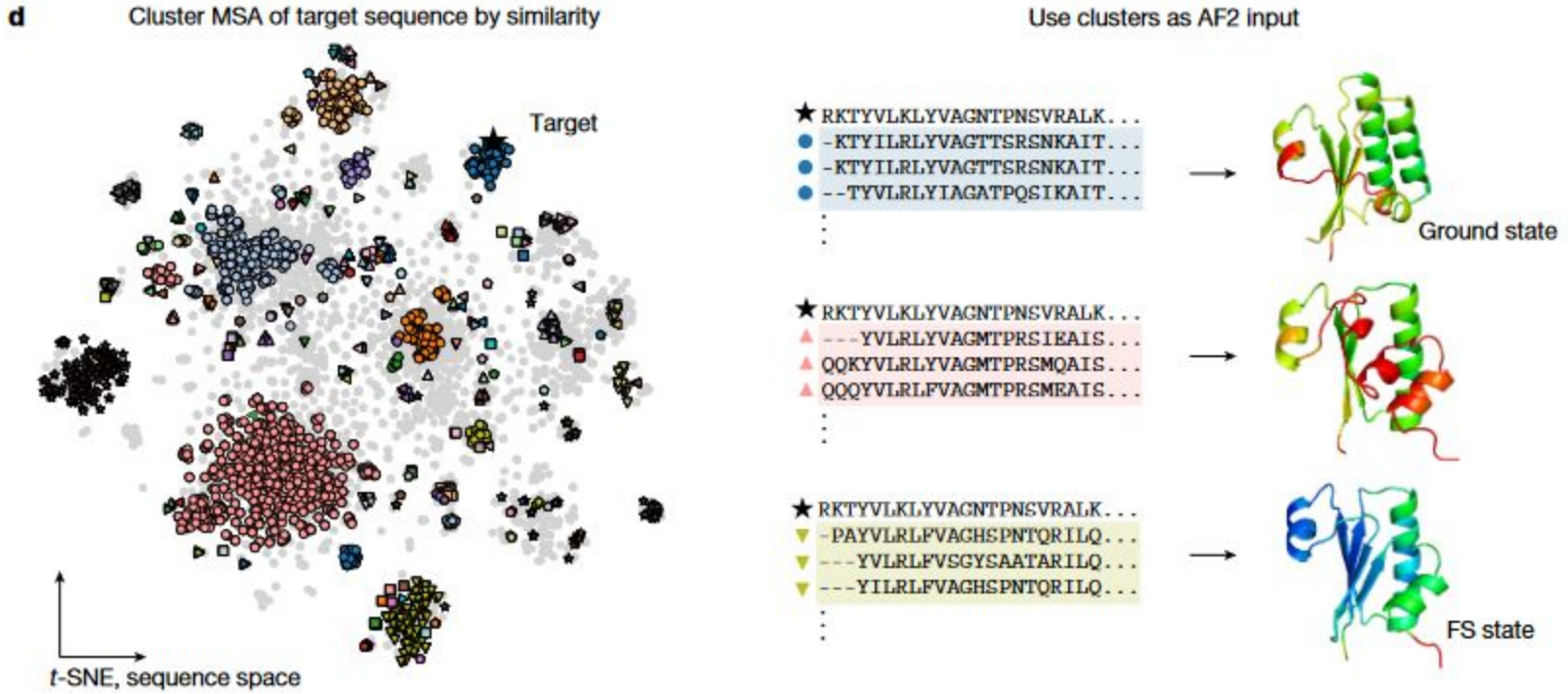




**Beta-strands**
**Beta-strands**
**Helix-helix interaction**



# AF2-Cluster


d
Cluster MSA of target sequence by similarity
Target
t-SNE, sequence space
Use clusters as AF2 input
RKTYVLKLYVAGNTPNSVRALK...
-KTYILRLYVAGTTSRSNKAIT...
-KTYILRLYVAGTTSRSNKAIT...
--TYVLRLYIAGATPQSIKAIT...
Ground state
RKTYVLKLYVAGNTPNSVRALK...
---YVLRLYVAGMTPRSIEAIS...
QQKYVLRLYVAGMTPRSMQAIS...
QQQYVLRLFVAGMTPRSMEAIS...
RKTYVLKLYVAGNTPNSVRALK...
-PAYVLRLFVAGHSPNTQRILQ...
---YVLRLFVSGYSAATARILQ...
---YILRLFVAGHSPNTQRILQ...
FS state

# AF2-Cluster

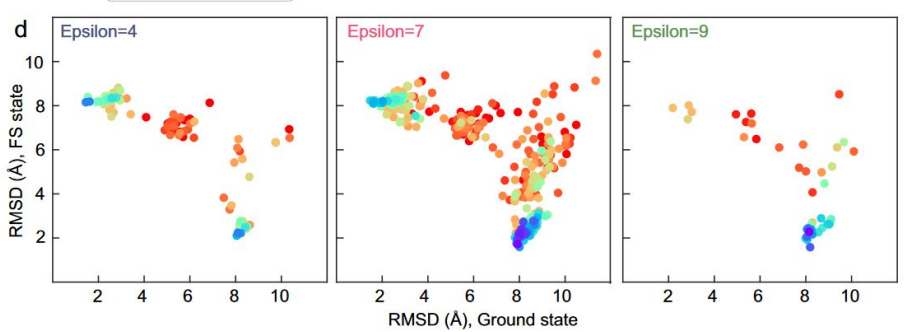
## MSA Generation

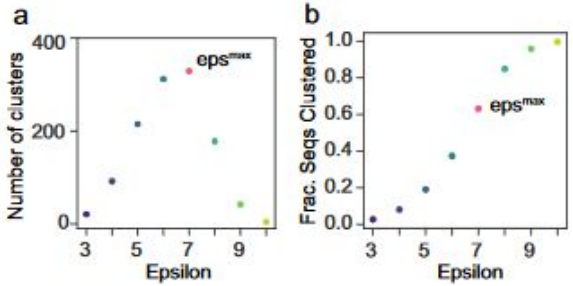
- MSAs generated using **MMseqs2-based routine** in **ColabFold**.
- **Search Process**:
  - Query sequence searched against **UniRef30** database (3 iterations).
- **Re-alignment**:
  - Hits' corresponding **UniRef100 cluster members** realigned to the profile from the previous iteration.
- **Filtering Criteria**:
  - Max sequence identity per cluster: **≤ 95%**.
  - **Top 3,000** most-diverse sequences filtered across identity buckets:
    - [0.0–0.2], (0.2–0.4], (0.4–0.6], (0.6–0.8], (0.8–1.0).
- **Gap Removal**: Sequences with more than **25% gaps** removed before clustering.

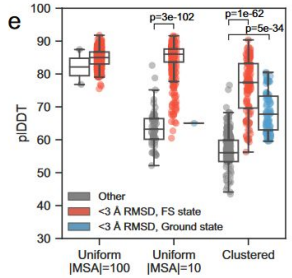
## Clustering

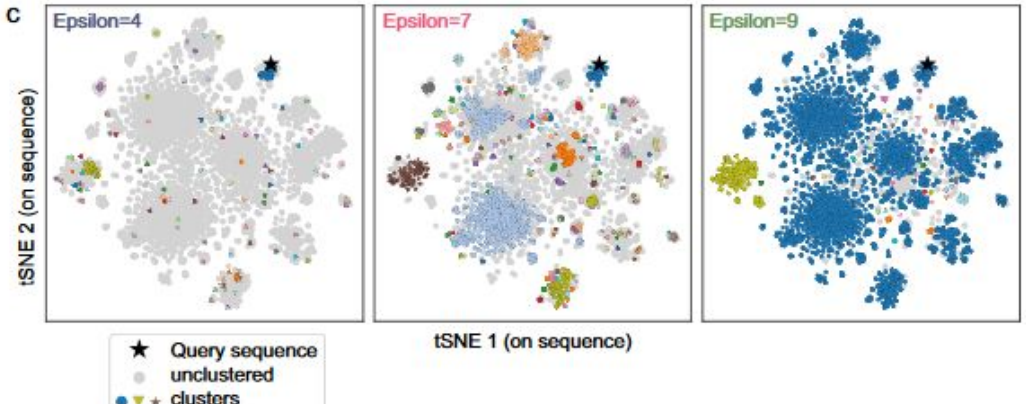
- **Cluster MSA using edit distance**
- **DBSCAN Clustering** involving **core density regions**:
  - At least **k points** within distance **epsilon** form clusters
  - Points outside core regions are considered **noise**.
- **Cluster Size Balance**
  - **Too small**: Not enough signal to capture states.
  - **Too large**: Dilutes signal from some states.
  - Example: KaiB predicted only **FS state** using entire MSA.
  - Epsilon varied from **3 to 20** in steps of **0.5**.

# maximizing information content of clustering using DBSCAN

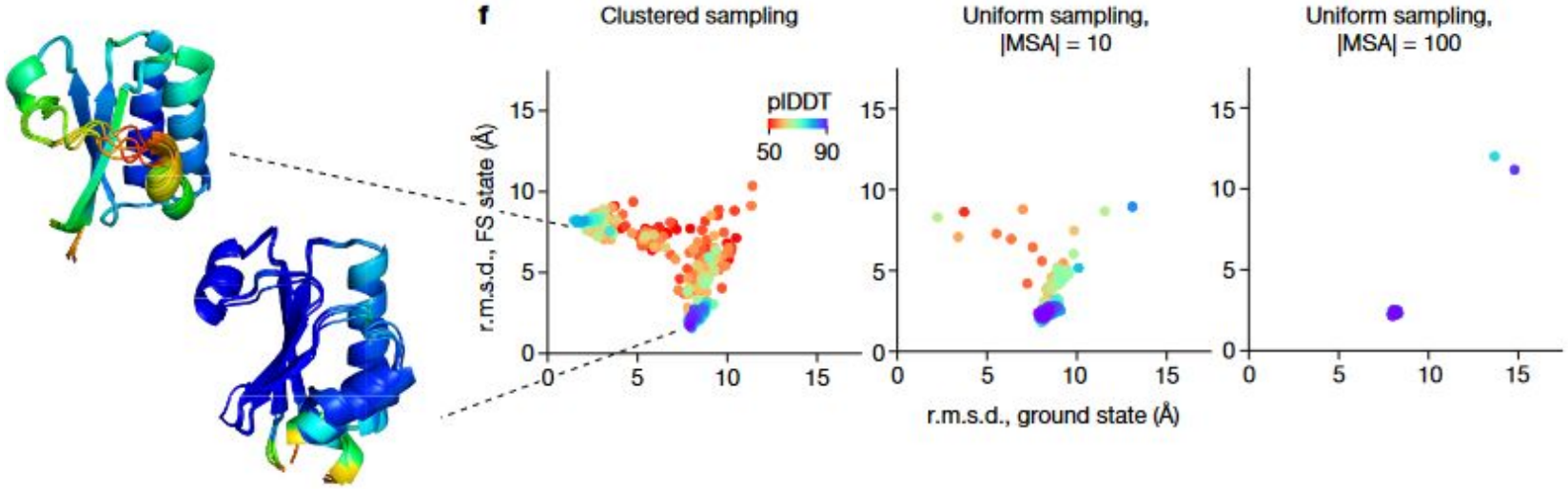


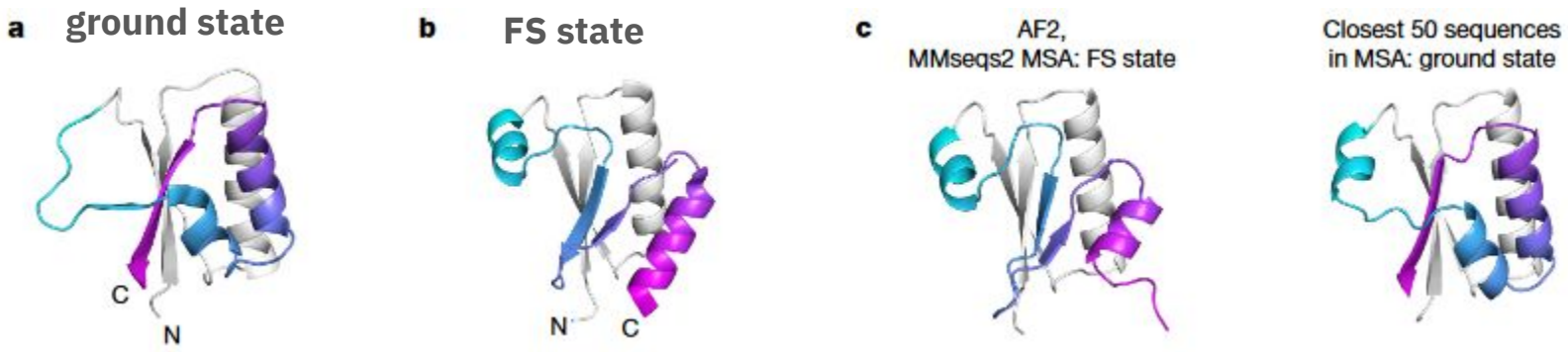



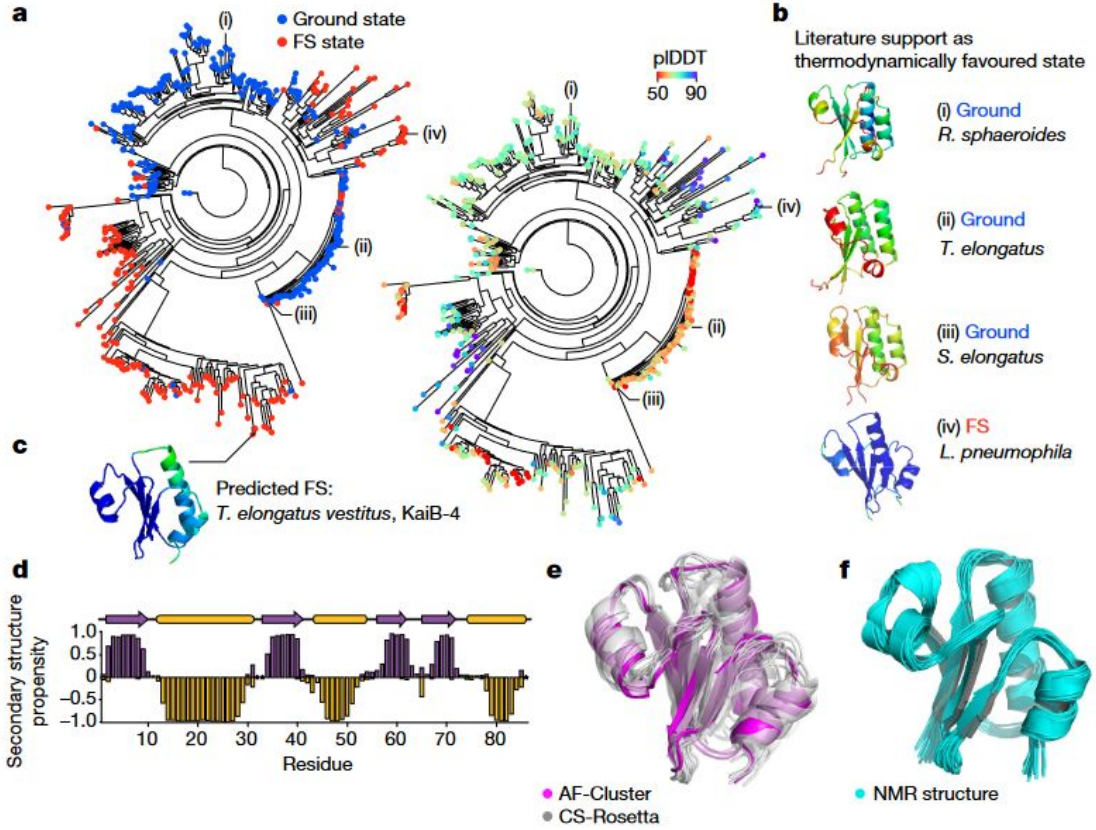
# AF-Cluster for KaiB$^{TE}$ aligns with both ground and FS crystal structures, while uniform MSA sampling only captures FS state confidently.



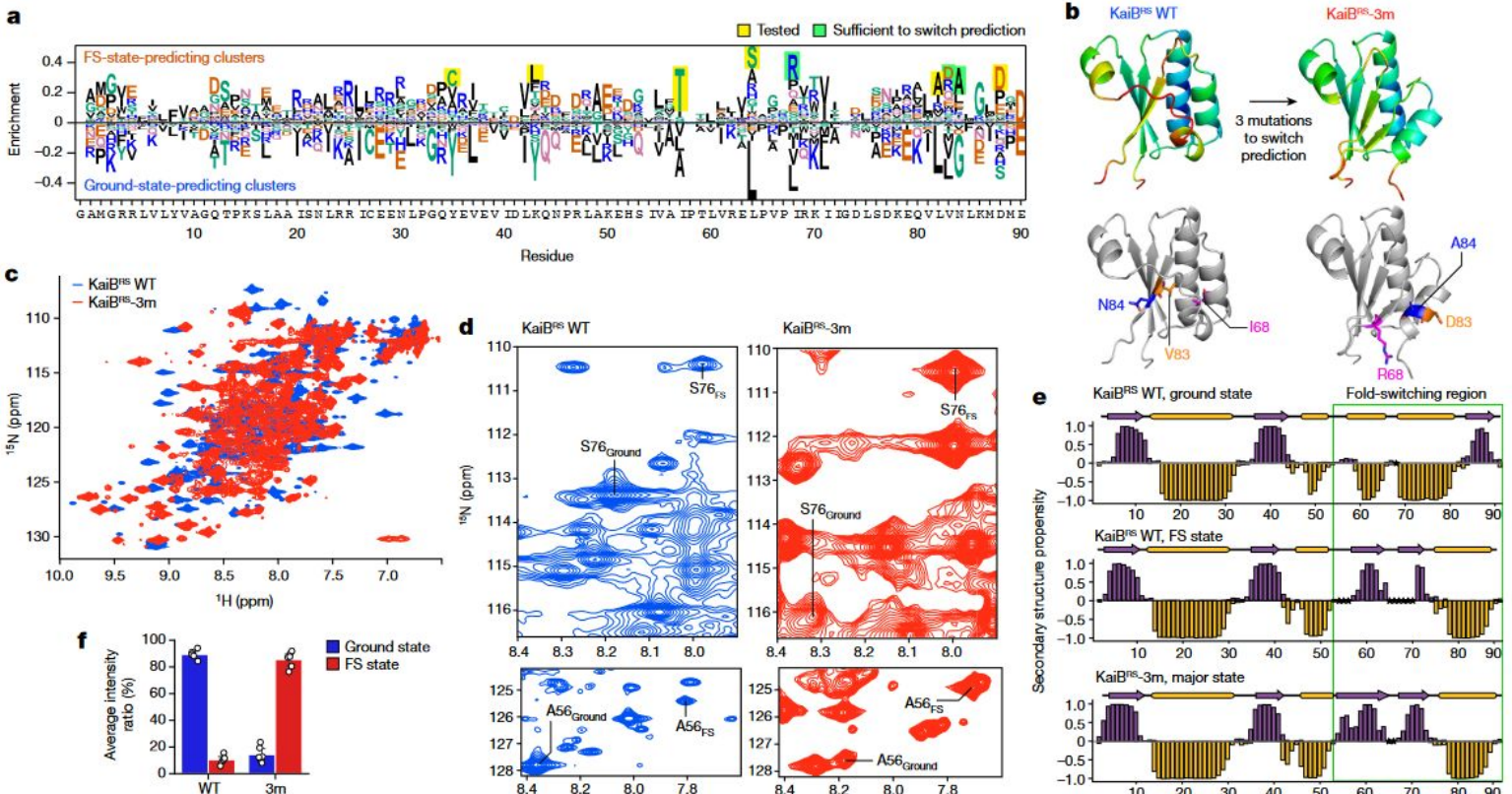
# AF2 predictions from MSA clusters for the fold-switching protein KaiB return both known structures



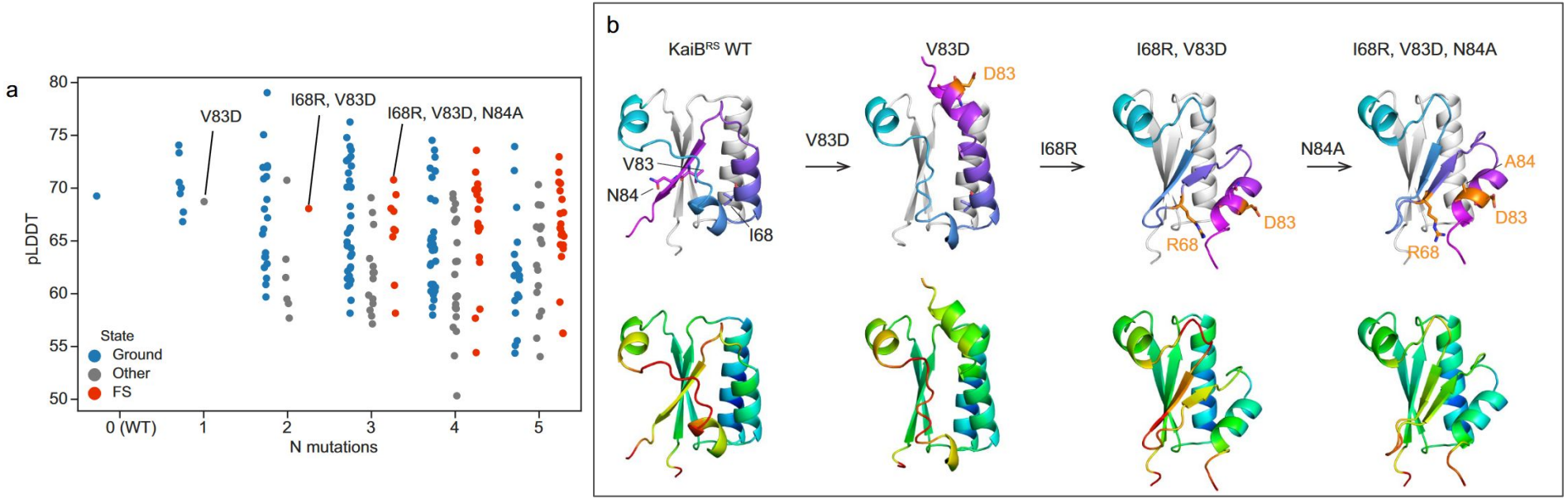
# The KaiB family contains pockets of sequences predicted to be stabilized for both states.



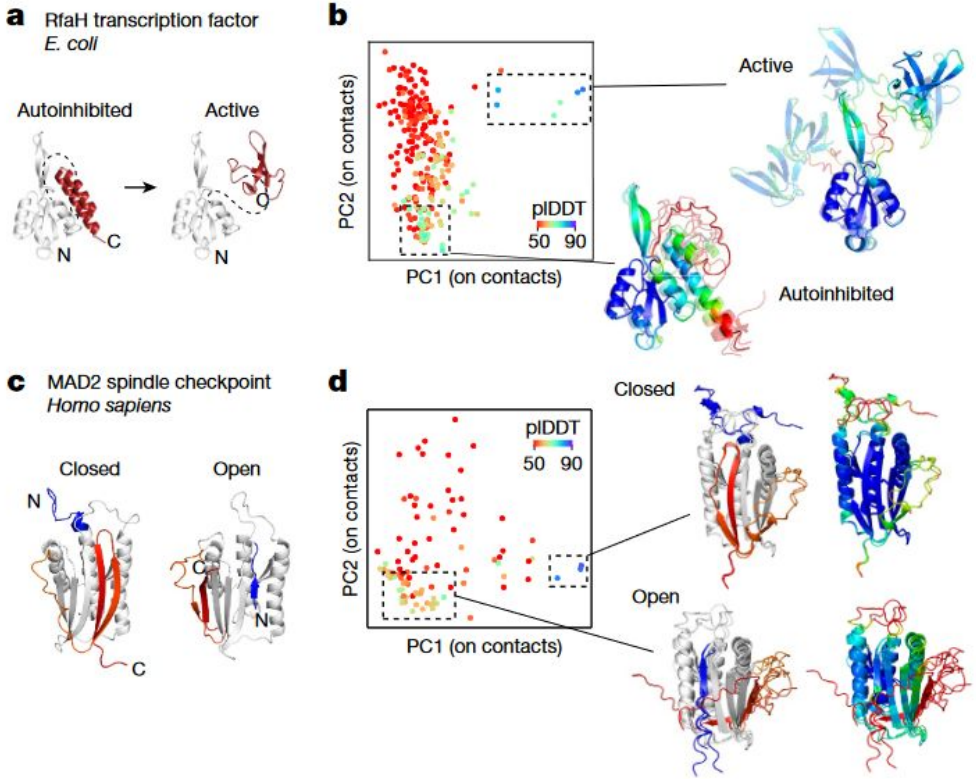
# AF-Cluster's ability to identify key mutations that could switch AF2's predictions between states was tested by analyzing clusters predicting different states.



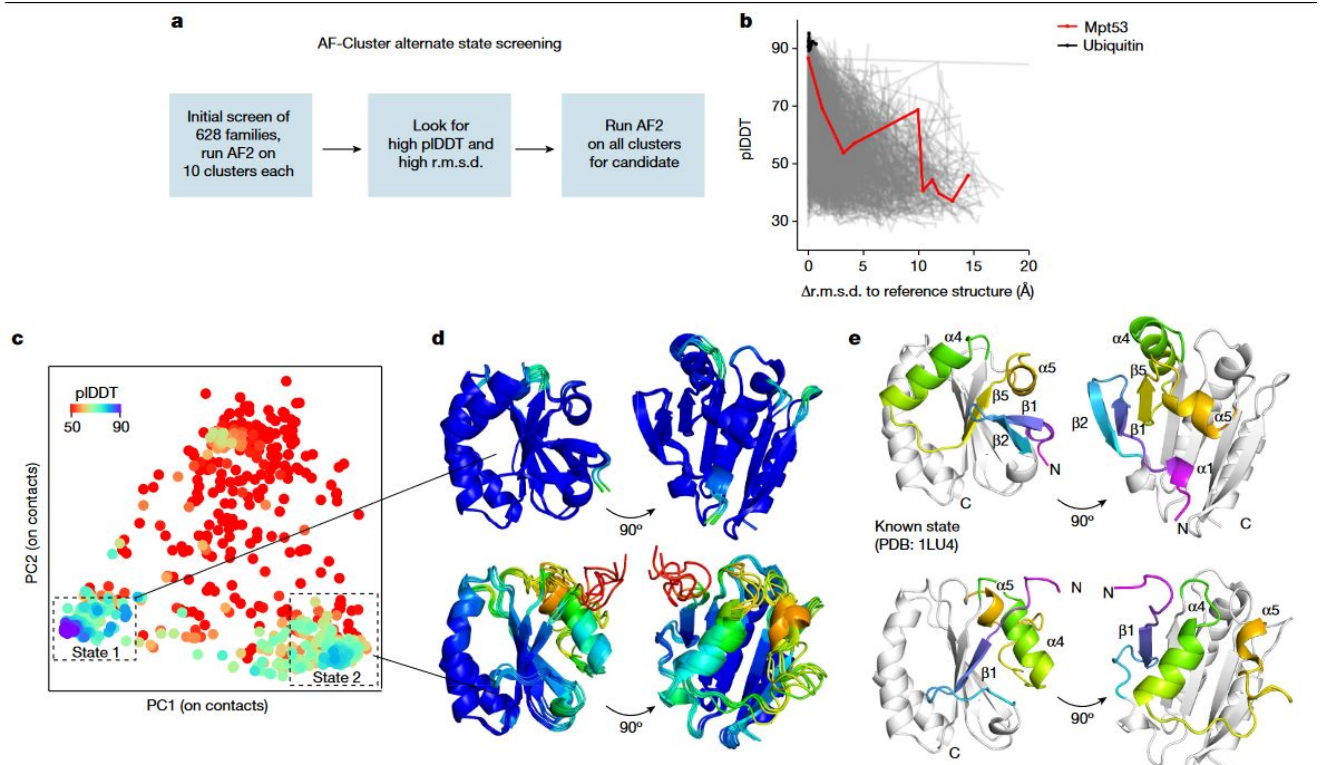
# Three mutations are sufficient to switch KaiB^RS AF2 prediction to high-confidence FS state prediction



# AF-Cluster predicts fold switching for the proteins RfaH and MAD2.



# Screening for fold switching in many protein families predicts a putative alternative fold for the M. tuberculosis secreted protein Mpt53.



# Discussion

- Clustering MSA sequences enables AF2 to sample **multiple biologically relevant conformations of metamorphic proteins.**
- **KaiB variants** in phylogenetic tree pockets predicted to **stabilize in specific states.**
- Single-sequence AF2 prediction **incorrectly predicts ground state**, highlighting the importance of clustering for isolating evolutionary couplings.
- pLDDT metric **not indicative of free energy**; AF-Cluster models show **higher pLDDT for thermodynamically disfavored states.**

# Discussion

- AF-Cluster
  - informed design of **mutations switching** KaiB^RS equilibrium from ground to FS state.
  - screening identified potential **alternative state** for M. tuberculosis oxidoreductase Mpt53.
  - may reveal numerous **uncharacterized functional protein states**.
- Alternative methods needed for **conformational substates** *absent in evolutionary signal.*