XXX Symposium on Bioinformatics
& Computer-Aided Drug Discovery
September 16-18, 2024

Protein 3D Structure Identification by AlphaFold: a Physics-Based *Prediction* or *Recognition* Using Huge Databases?

Alexei V. Finkelstein

Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia
Biology Department, Lomonosov Moscow State University, Moscow, Russia

Dmitry N. Ivankov

Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia

E-mail: afinkel@vega.protres.ru

Two main problems of protein folding

«Protein folding problem» №1:

HOW a protein can fold spontaneously so fast?

⬇

Solved: "Folding funnel" with phase separation

(Finkelstein, Badretdinov, 1997-98; Garbuzynskiy et al., 2013)

«Protein folding problem» №2:

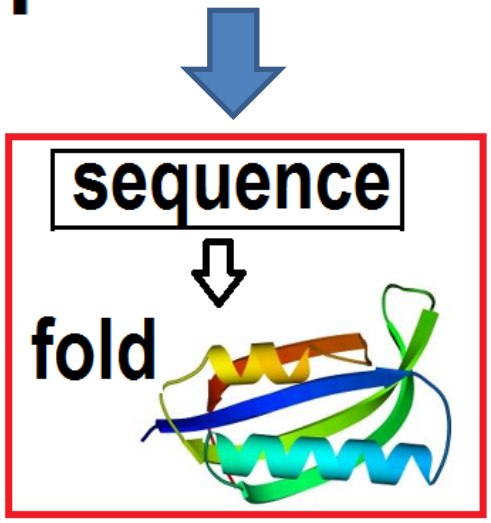
Predict 3-dimensional structure of a protein from its amino acid sequence

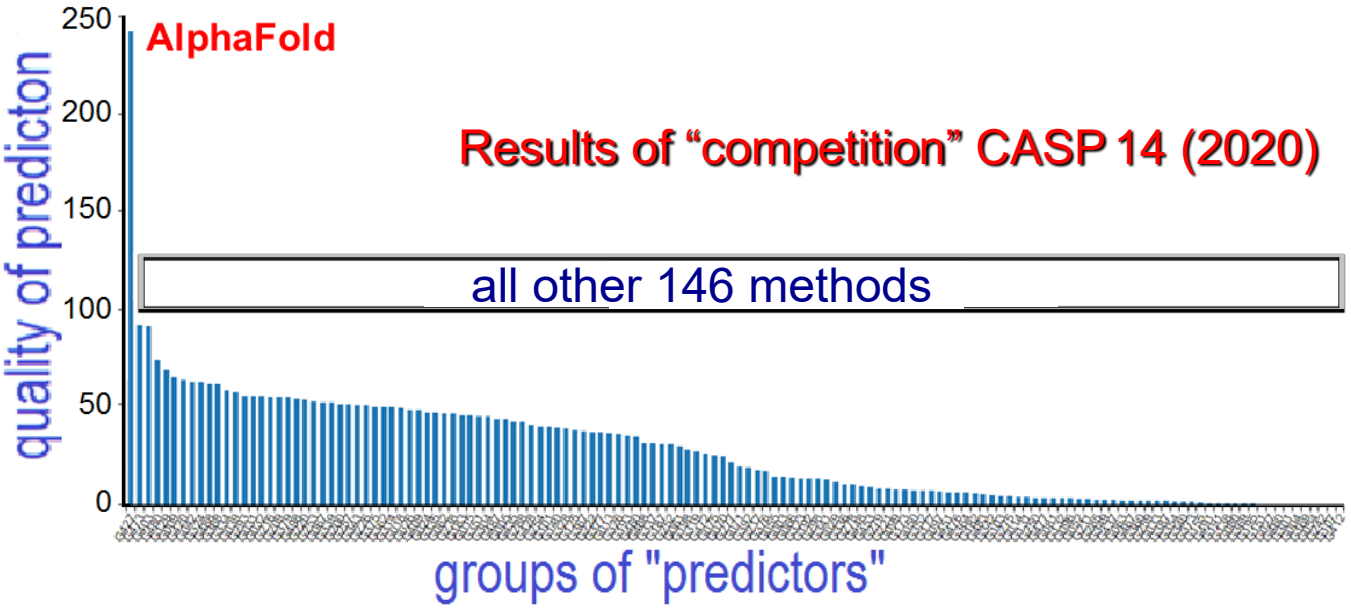
⬇

AlphaFold

(Senior *et al.*; Jumper *et al.*)

prediction





AlphaFold – great success

A. Senior, J. Jumper, et al., 2018-21

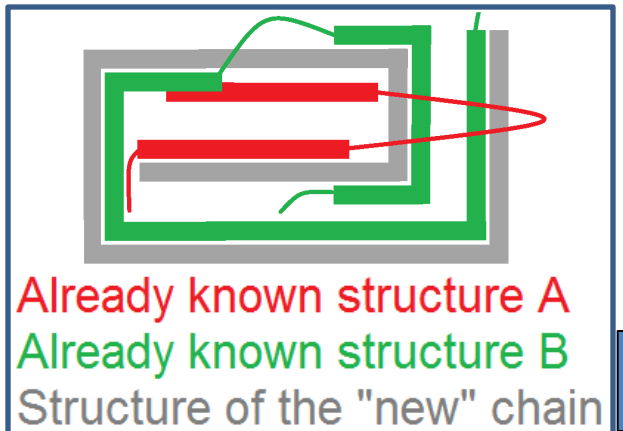
1) What is the main reason for this success?

2) What does **AlphaFold** do:

– does it *predict* protein structure from its a.a. sequence & *physics of protein chain folding?*

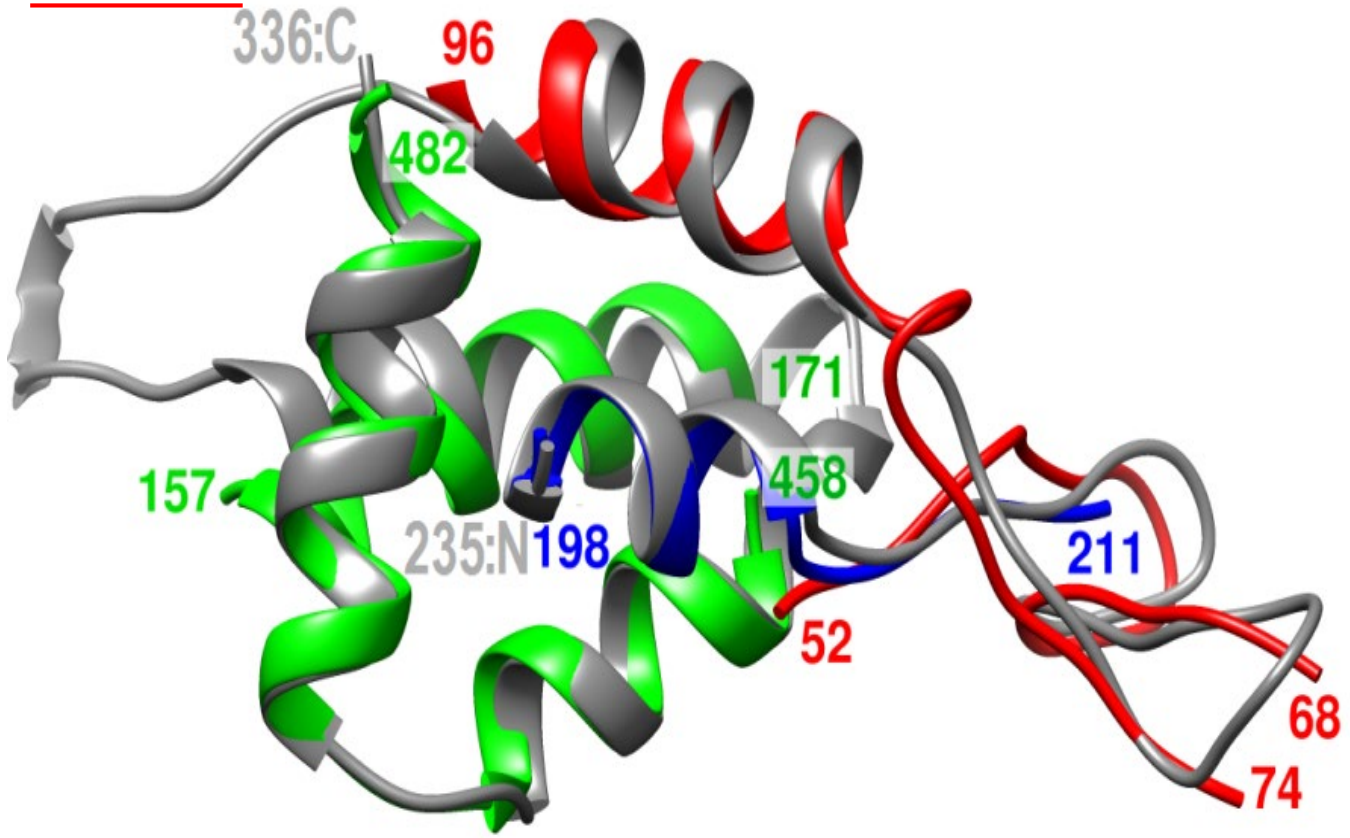
or

– *recognize* this structure by the *similarity* of large pieces of its a.a. sequence with **"joined" large pieces of sequences** that *already* are in PDB?



“Novel fold”: When it is impossible to superimpose *any* of already known 3D structures onto this “**novel** fold”

greatest success

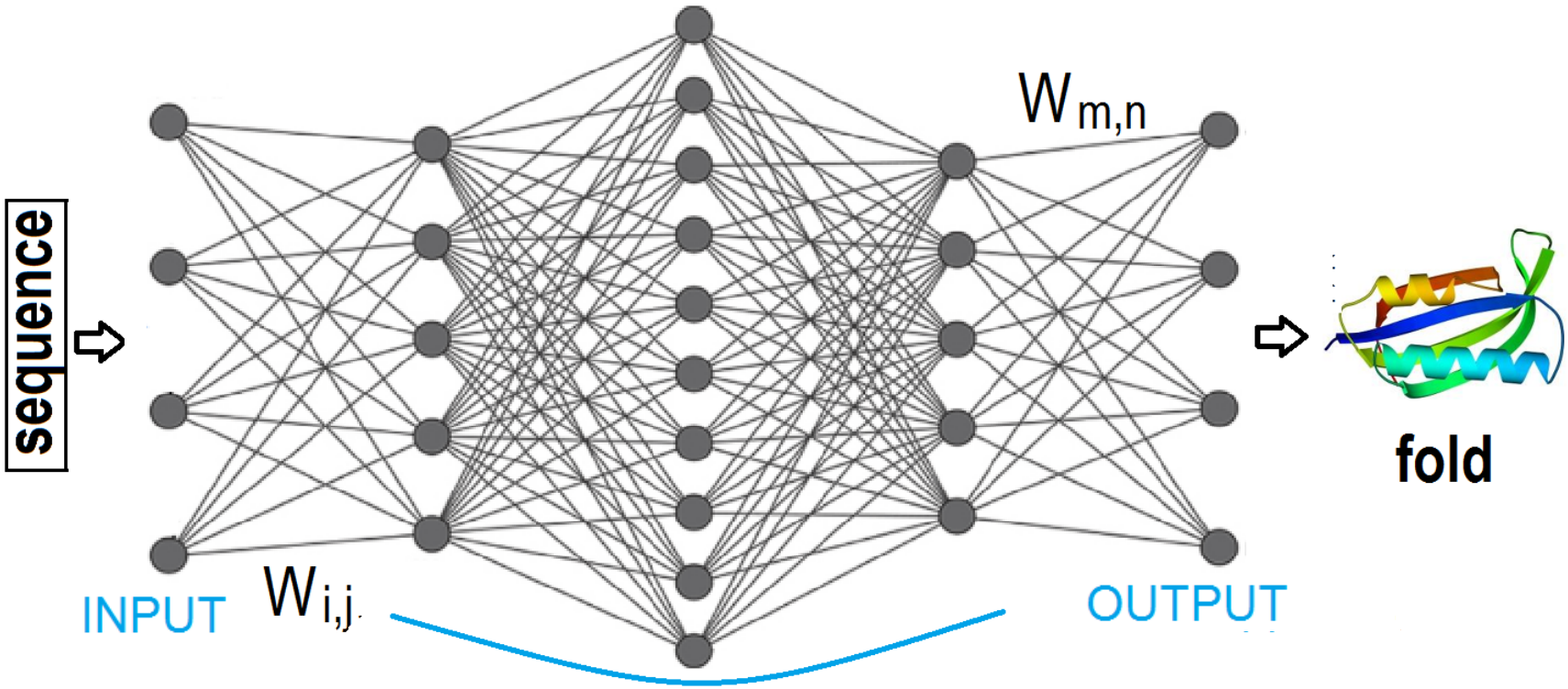


One of many dozens of examples of superposition of pieces of already known 3D structures onto a “**novel** fold”

“Novel fold” (6VR4, chain A - target T1035 from CASP 14 (2020)) as a combination of fragments of 3 already known structures available to AlphaFold during the training:

1GB3, chain A; 5A29, chain A; 5W40, chain B.

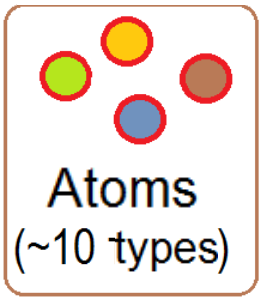
AlphaFold: neural network

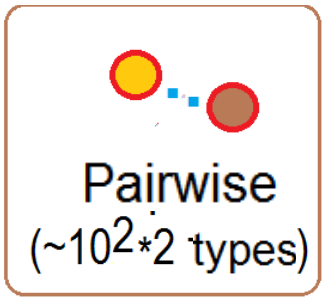


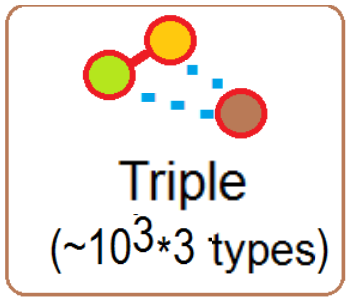
Hidden layers (many dozens)

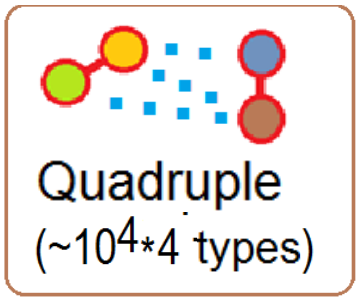
$W_{i,j}$, $W_{m,n}$ - "weights" (adjustable parameters): In Alphafold: **~21,000,000**

But physics of atomic interactions in proteins only needs **~43,200**







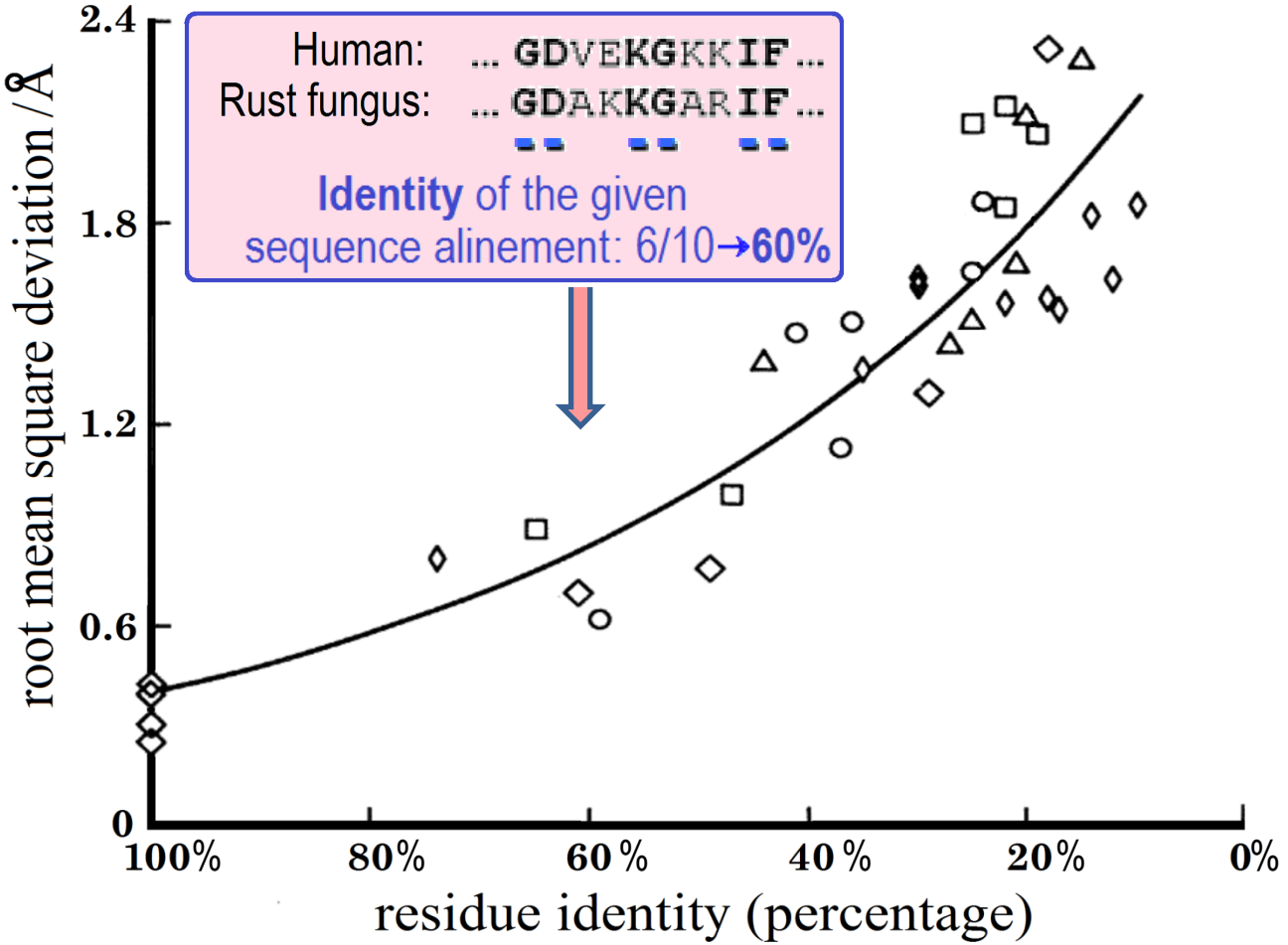


So, ~20,955,000 parameters are "trained" not in physics, but ?

KNOWN:

SIMILAR SEQUENCES ➔ *VERY* SIMILAR STRUCTURES

but only – with identity ≳15-20%



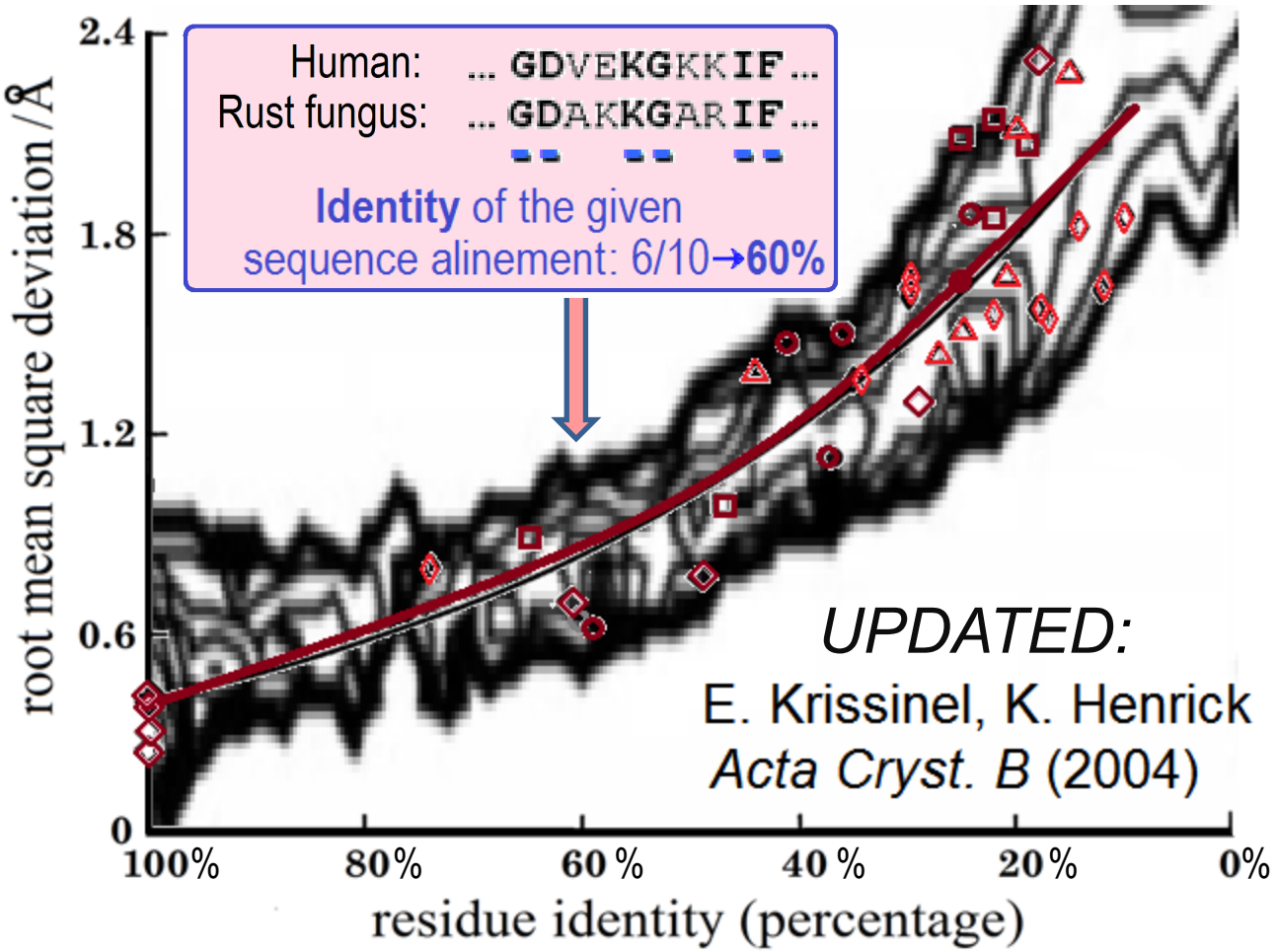
Lesk, A.M., Chothia, C. *Phil. Trans. R. Soc. Lond. A* **317**, 345–356 (1986)

WHAT IDENTITY WITH A "NEW" SEQUENCE IS EXPECTED - HAVING MODERN HUGE DATABASES?

KNOWN:

SIMILAR SEQUENCES ➔
***VERY* SIMILAR STRUCTURES**

but only – with identity ≳15-20%



Lesk, A.M., Chothia, C. *Phil. Trans. R. Soc. Lond. A* **317**, 345–356 (1986)

WHAT IDENTITY WITH A "NEW" SEQUENCE IS EXPECTED - HAVING MODERN HUGE DATABASES?

WHAT IDENTITY WITH A "NEW" SEQUENCE IS EXPECTED - HAVING MODERN HUGE DATABASES?

The probability that two "random" sequences of n a.a. residues, each of which occurres with a probability p (in proteins, ~1/20), coincide in m positions, follows from

5%

$$P_{m,pn} = \frac{(pn)^m}{m!} e^{-pn}$$

Poisson distribution

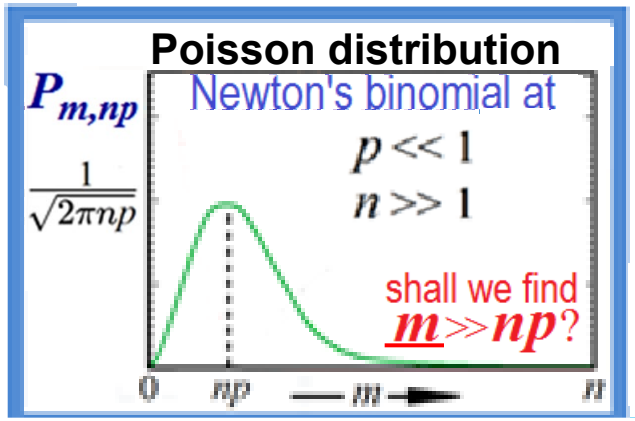
WHAT IDENTITY WITH A "NEW" SEQUENCE IS EXPECTED - HAVING MODERN HUGE DATABASES?

The probability that two "random" sequences of n a.a. residues, each of which occurres with a probability p (in proteins, ~1/20), coincide in m positions, follows from

5%

$$P_{m,pn} = \frac{(pn)^m}{m!} e^{-pn}$$

pn – average (*pairwise* comparison)



ACDEFGHI**K**LMNPQRSTVWY

KPYDSTFQ**K**HILAMNPQRST ➡ Expected for **1** pairwise comparison: $n \times 5\%$

Could we expect $n \times 20\%$ for **1** out of **1000000** pairwise comparisons?

ACDEFGHI**K**LMNPQRSTVWY	AC**D**EFGHIKLMNPQRSTVWY	ACDE**F**GHIK**L**MNPQRSTVWY
KPYDSTFQ**K**HILAMNPQRST	KP**D**DSTFQSHILAMNPQRST	KPYD**F**TFQH**L**ILAMNPQRST
1	2	1000000

and having shifts, insertions, deletions – when comparing sequences?

ACD**E**---FGHI**K**LMNPQRST**VW**Y

MNPD**A**TF**E**PYDSTFQ**K**HILA--MN**VW**RSTDSTF

WHAT IDENTITY WITH A "NEW" SEQUENCE IS EXPECTED - HAVING MODERN HUGE DATABASES?

The probability that two "random" sequences of n a.a. residues, each of which occurres with a probability p (in proteins, ~1/20), coincide in m positions, follows from

5%

$$P_{m,pn} = \frac{(pn)^m}{m!} e^{-pn}$$

pn – average (*pairwise* comparison)

Poisson distribution

because $m! \approx (m/e)^m$ (Stirling's eq.), then $P_{m,pn} \approx \left(\frac{ep}{m/n}\right)^m e^{-pn}$

When **1** sequence is compared not with **1**, but with N others, then $P_{m,pn} \cdot N = 1$ gives the maximally expected number of matches (M) with the "most similar" of them. Thus, the expected residue identity M/n follows from the equation

$$\left(\frac{M/n}{pe}\right)\ln\left(\frac{M/n}{pe}\right) + \frac{1}{e} = \left(\frac{1}{npe}\right)\ln(N)$$

expected M/n

MODERN DATABASES

no chain shifts, insertions, deletions

n - any, N=1: $M/n = p = 5\%$

n=100 (domain), N~150000 (PDB): M/n = 20%

n=100 (domain), N~190000000 (UniProt): M/n = 24%

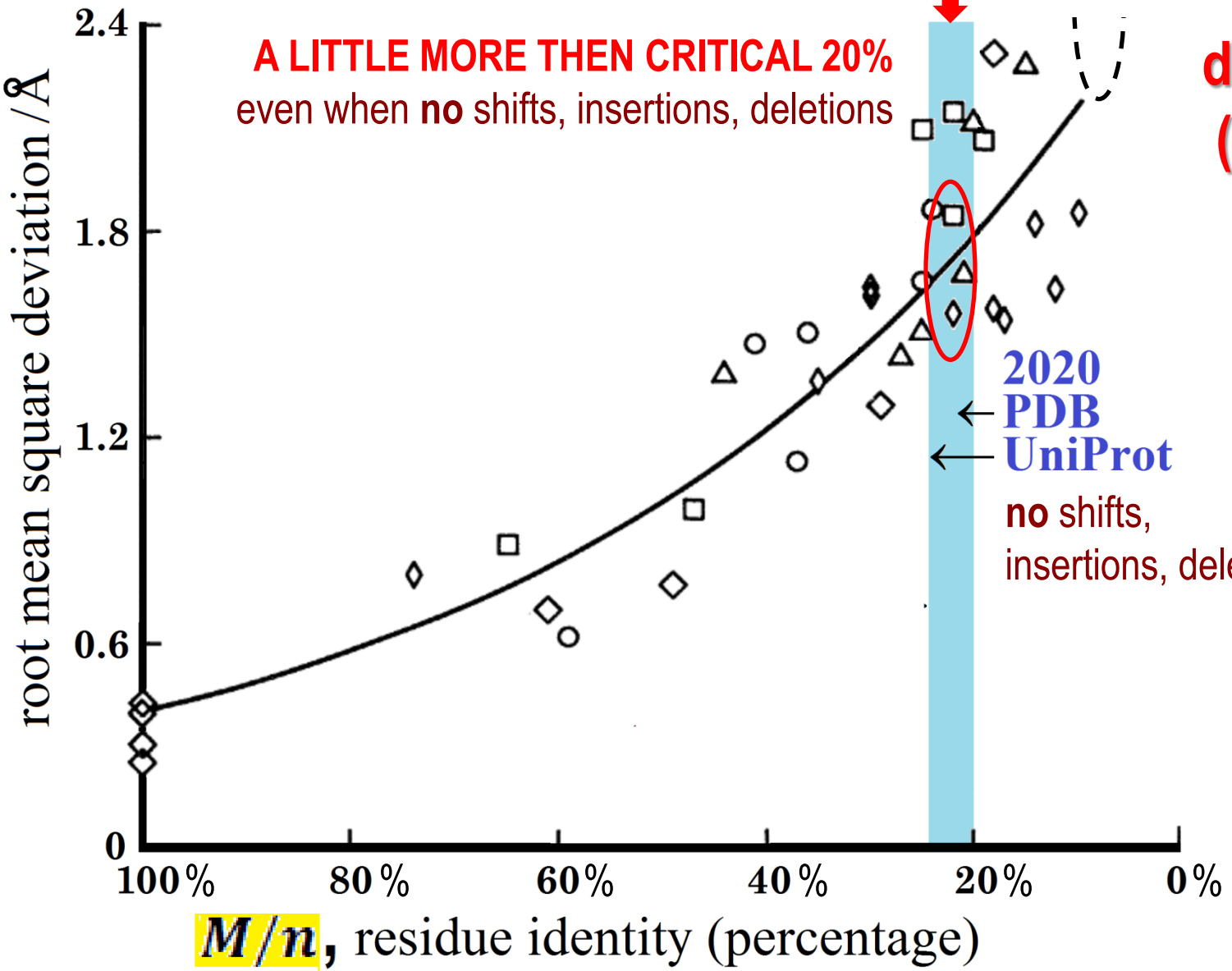
with chain shifts, insertions, deletions

n=100 (domain), N~150000 (PDB)$*10^6$: M/n = 25%

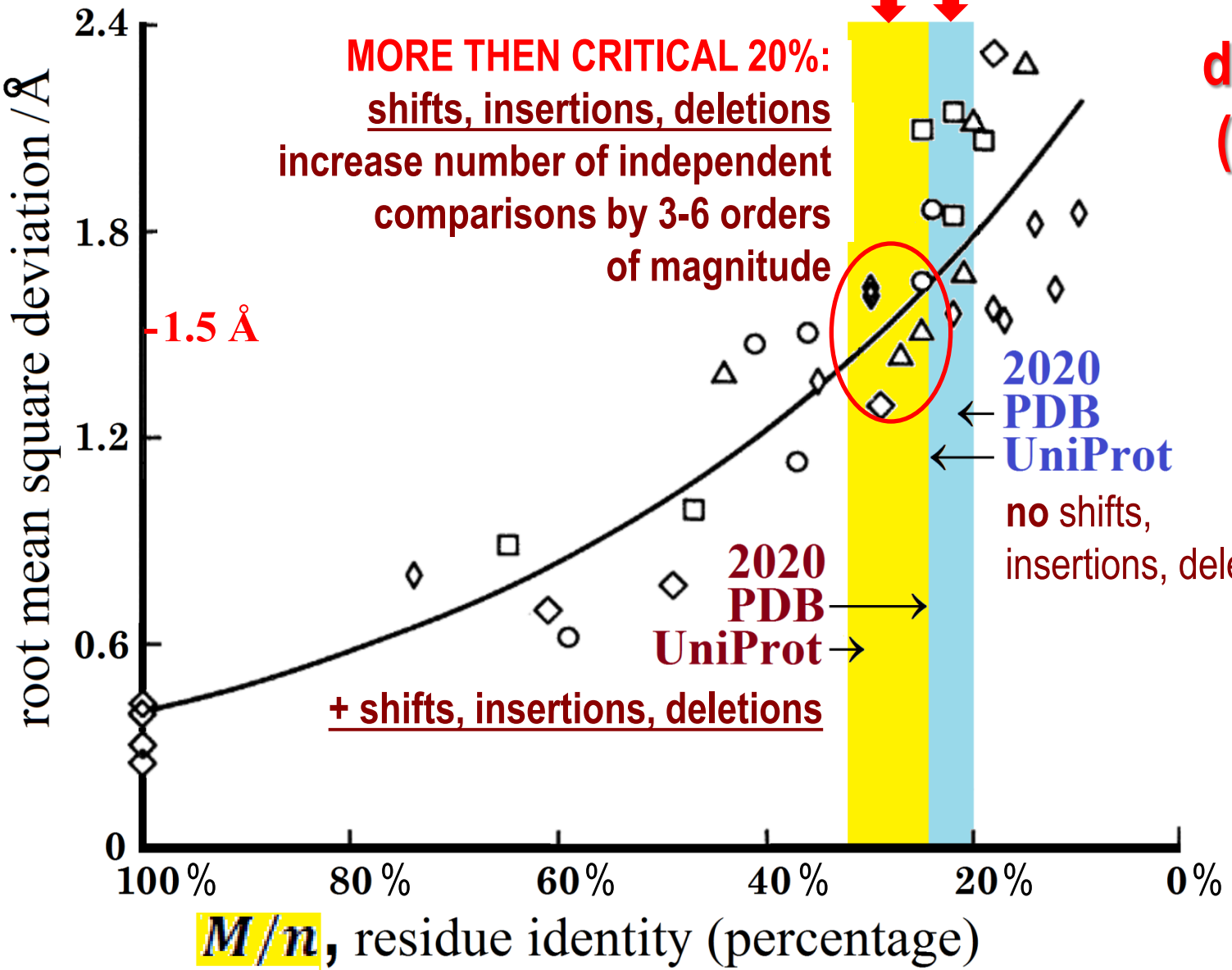
n=100 (domain), N~190000000 (UniProt)$*10^6$: M/n = 32%

8”

EXPECT TO FIND – HAVING MODERN DATABASES



EXPECT TO FIND – HAVING MODERN DATABASES

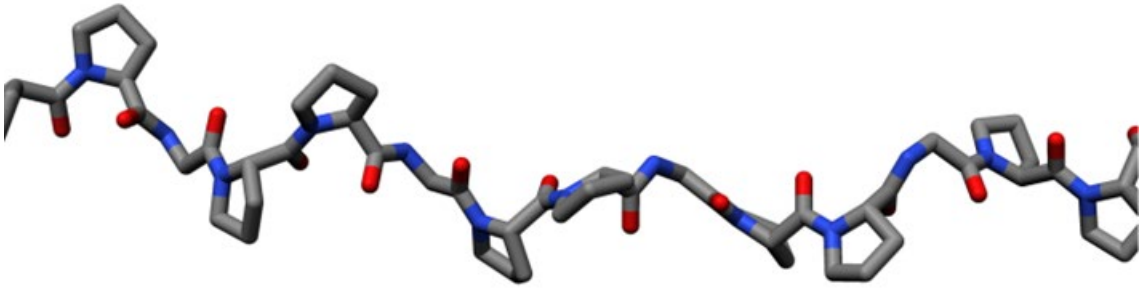


WITH MODERN DATABASES, ALPHAFOLD CAN RECOGNIZE PROTEIN STRUCTURE

WITH MODERN DATABASES, ALPHAFOLD CAN RECOGNIZE PROTEIN STRUCTURE

NOTE:

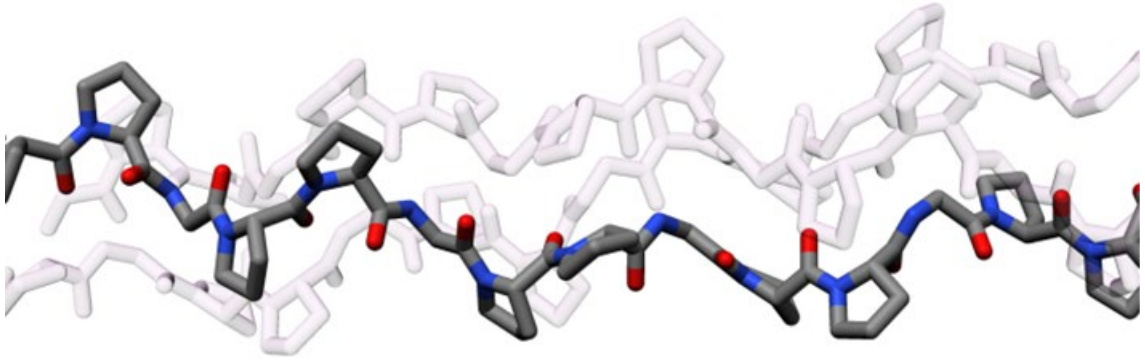
Bioinformatics is much more important than *physics* for AlphaFold predictions:



AlphaFold -predicted piece of $(Gly\text{-}Pro\text{-}Pro)_{13}$

- is a *contradicting to physics* prediction of a *non-compact* structure of separate collagen-like $(Gly\text{-}Pro\text{-}Pro)_{13}$ chain, which **lacks interactions** that can **support** it.

In collagen, such a chain is fixed by ***surrounding*** chains:



Collagen **PDB**: 4CTD

- but these have been *not* introduced to AlphaFold, asked to predict a structure of the **separate** $(Gly\text{-}Pro\text{-}Pro)_{13}$ chain!

Knowing similar complexes, AlphaFold makes correct *bioinformatic recognition, though contradicting to physics* of this separate chain.

A LITTLE PHILOSOPHY

1) "Predict fold" $\neq$ **"Predict folding"** (*folding rate*)

result — process ↓

AlphaFold

*(Garbuzynskiy et al,. PNAS, **110**:147–150, 2013;*
*Ivankov, Finkelstein, Biomolecules **10**:E250, 2020)*

2) Does *AlphaFold* know protein physics?

- *it knows only the **frequency of occurrence** in **PDB** of elements of protein structures, which is **related to their stability***
  (Finkelstein et al., Proteins, 23: 142-150, 1995)
- ***AlphaFold** relies on bioinformatics, and (yet) knows **nothing** about the process of protein **folding***

A LITTLE PHILOSOPHY

1) "Predict fold" $\neq$ **"Predict folding"** (*folding rate*)

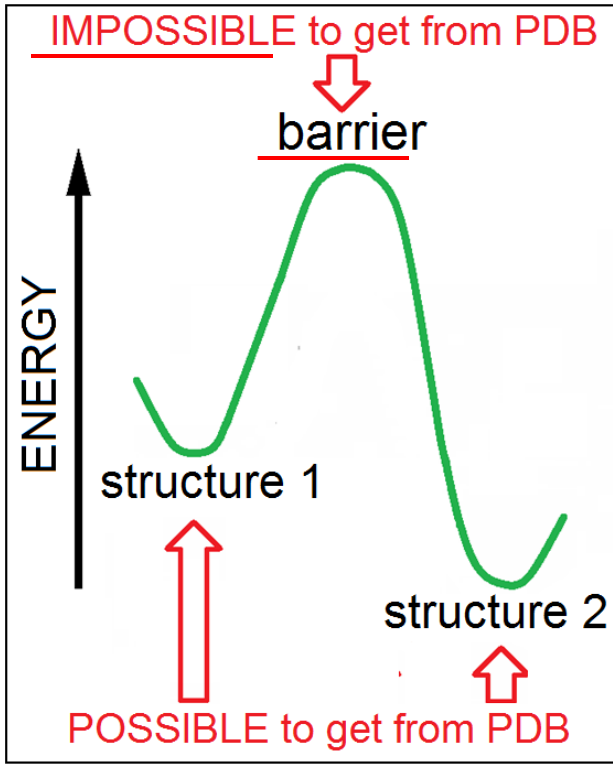
result — process

AlphaFold ↓

*(Garbuzynskiy et al,. PNAS, **110**:147–150, 2013;*
*Ivankov, Finkelstein, Biomolecules **10**:E250, 2020)*

2) Does *AlphaFold* know protein physics?

- *it knows only the **frequency of occurrence** in **PDB** of elements of protein structures, which is **related to their stability***
  (Finkelstein et al., Proteins, 23: 142-150, 1995)
- ***AlphaFold** relies on bioinformatics, and (yet) knows **nothing** about the process of protein **folding***



A LITTLE PHILOSOPHY

1) "Predict fold" $\neq$ **"Predict folding"** (*folding rate*)

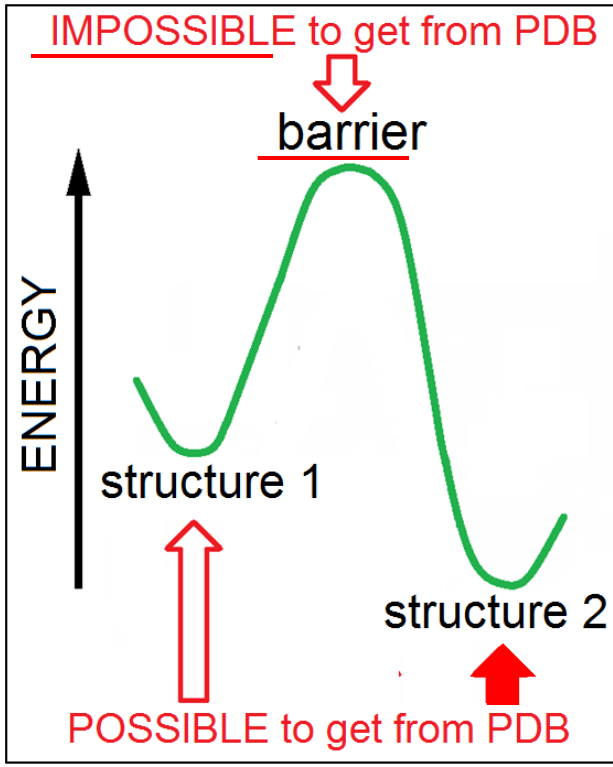
result — process

AlphaFold ↓

*(Garbuzynskiy et al,. PNAS, **110**:147–150, 2013;*
*Ivankov, Finkelstein, Biomolecules **10**:E250, 2020)*

2) Does *AlphaFold* know protein physics?

- *it knows only the **frequency of occurrence** in **PDB** of elements of protein structures, which is **related to their stability***
  (Finkelstein et al., Proteins, 23: 142-150, 1995)
- ***AlphaFold** relies on bioinformatics, and (yet) knows **nothing** about the process of protein **folding***



3) Does a **good prediction** mean a **correct understanding**?

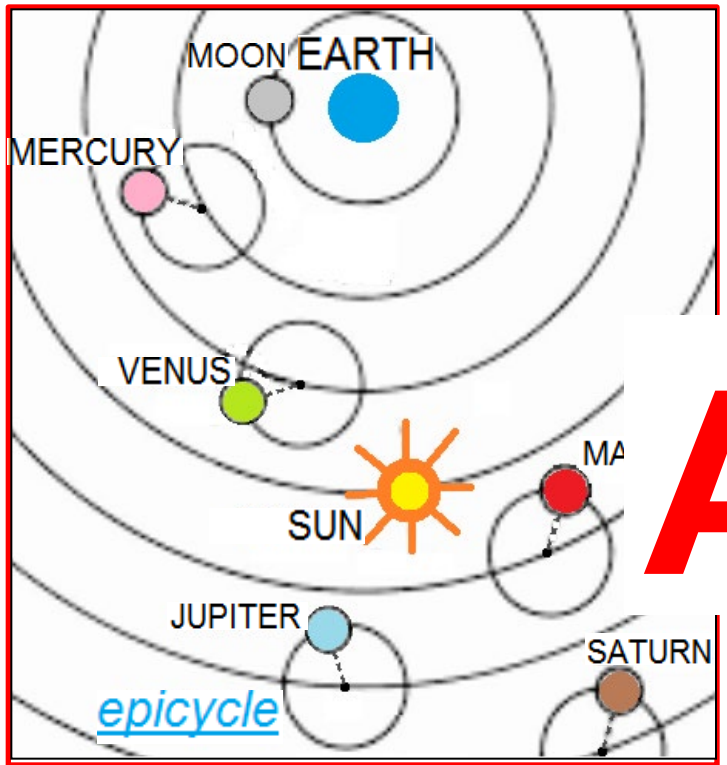
- ***NO***. ↓ *An example from the history of astronomy*

GOOD PREDICTION ⇏ CORRECT UNDERSTANDING

Priests of Egypt and Babylon:

GOOD PREDICTIONS
of eclipses of the Sun and Moon
(based on huge archives spanning 2500 years!),

BUT: *fundamentally* WRONG UNDERSTANDING
(The Earth is flat!)



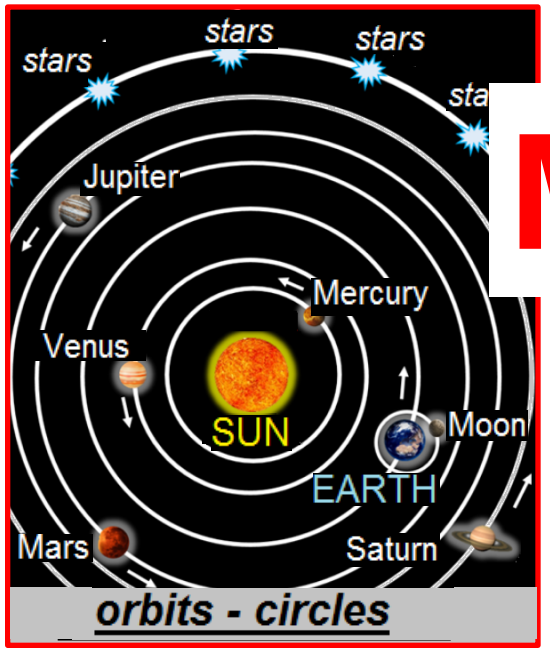
PTOLEMAEUS
(*using huge archives*):

GOOD PREDICTION

AlphaFold

***WRONG* UNDERSTANDING of the *PROCESS*!**

GOOD PREDICTION ⇎ CORRECT UNDERSTANDING

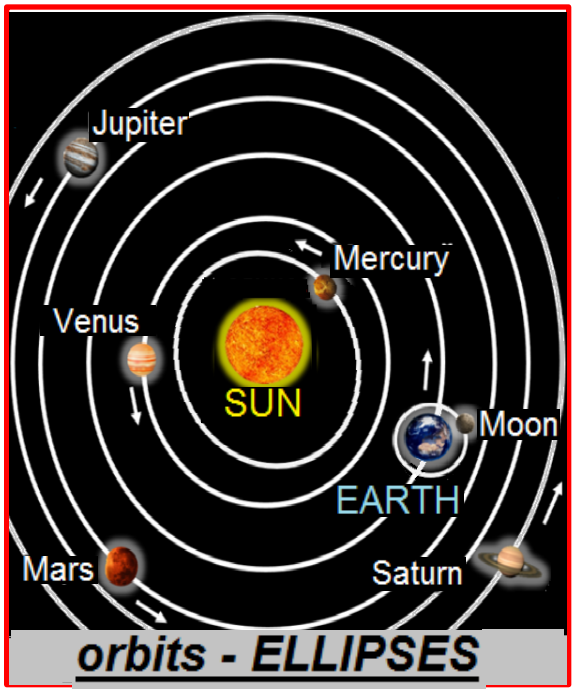


Copernicus:

Mol. dynamics

***(BUT – SMALL ERROR:** in parameters*),

IMPERFECT (worse than by Ptolemaeus)
PREDICTION OF PLANETARY MOVEMENTS

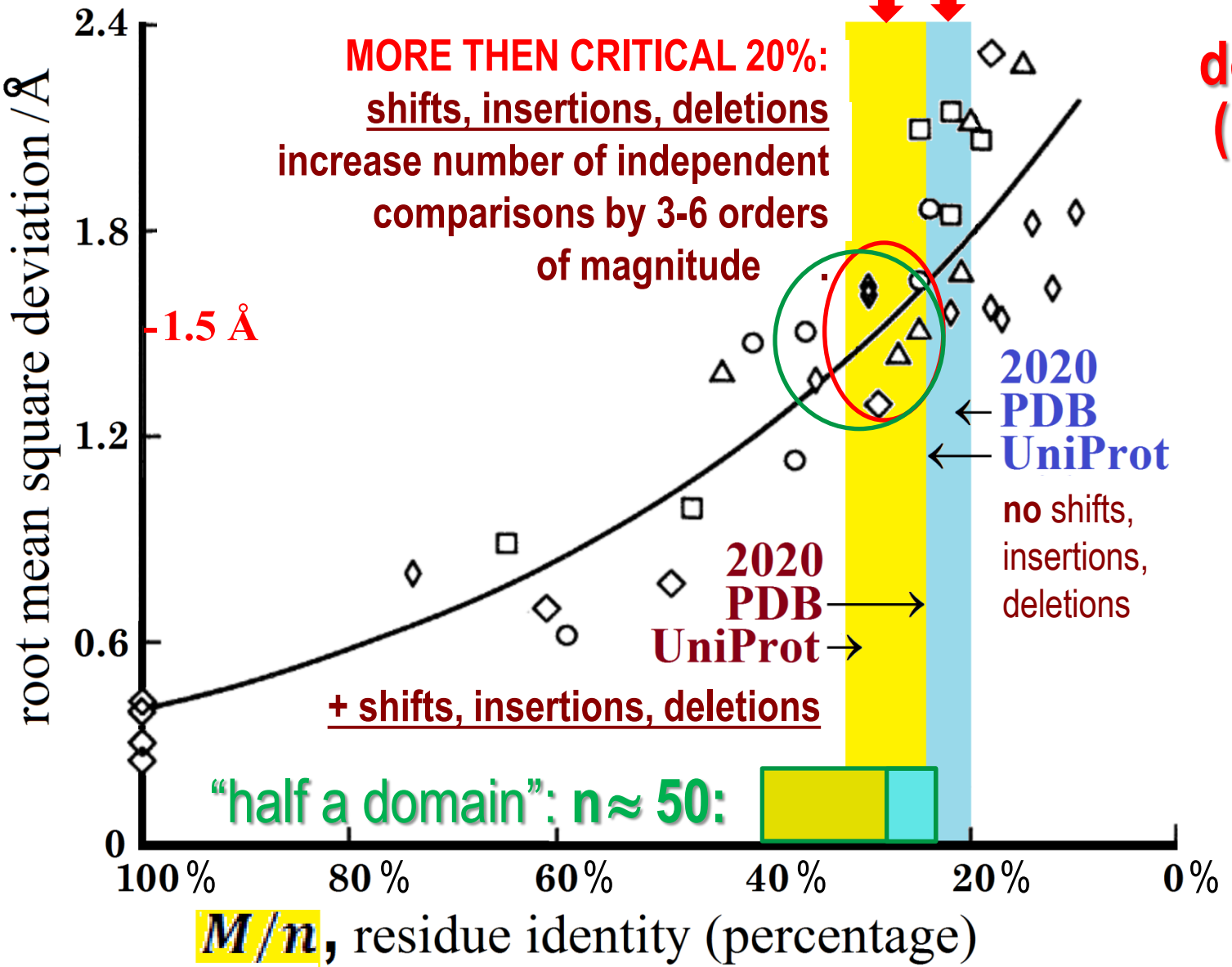


Kepler, Newton:

***CORRECT* UNDERSTANDING**
(exact equations of celestial mechanics!),

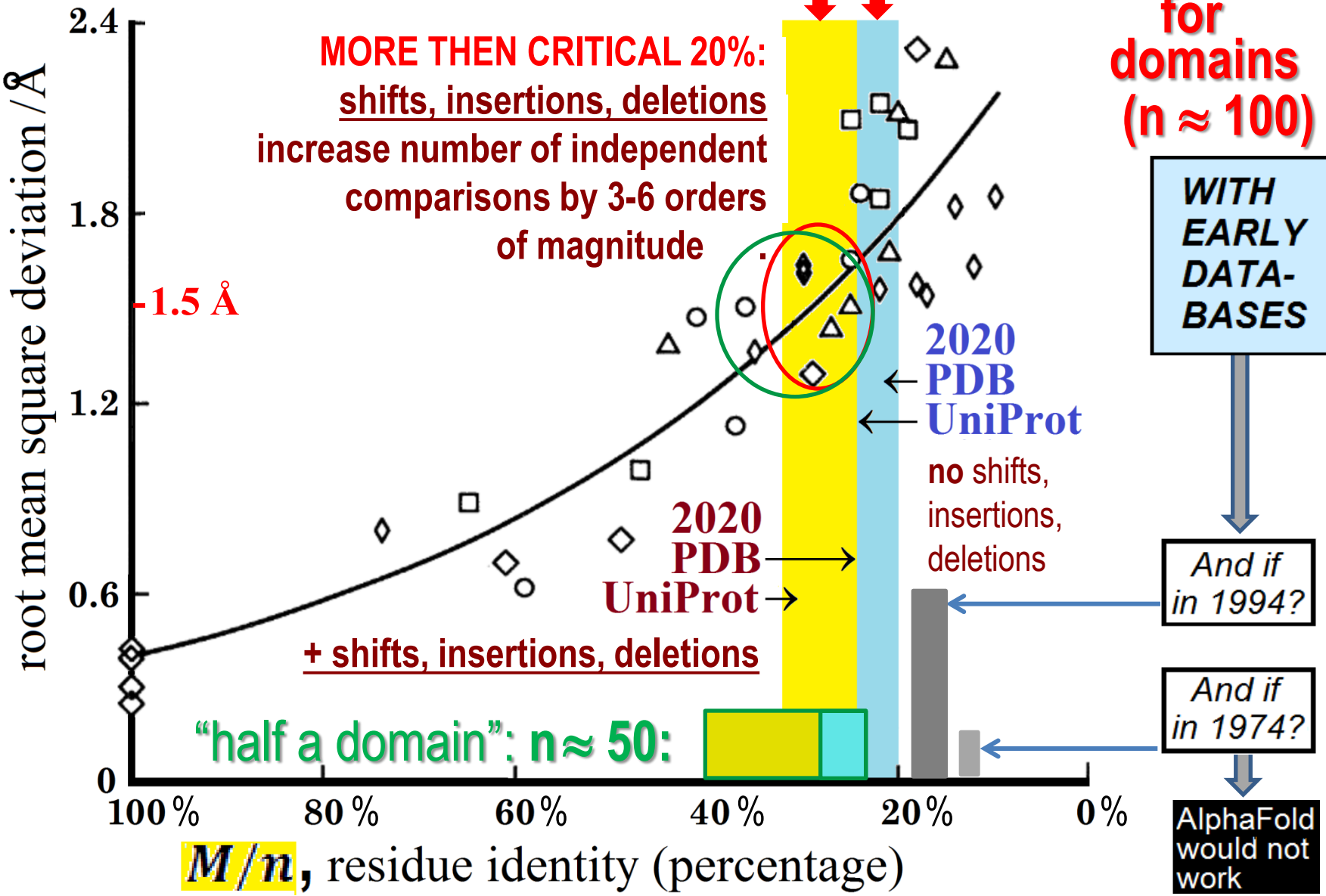
PERFECT PREDICTION OF MOVEMENTS OF PLANETS ,
COMETS, ROCKETS AND EVERYTHING ELSE

EXPECT TO FIND – HAVING MODERN DATABASES



WITH MODERN DATABASES, ALPHAFOLD CAN RECOGNIZE PROTEIN STRUCTURE

EXPECT TO FIND – HAVING MODERN DATABASES



WITH MODERN DATABASES, ALPHAFOLD CAN RECOGNIZE PROTEIN STRUCTURE

The basis of AlphaFold's great success is a skillful usage of huge protein databases collected during 60 years and clearly presenting evolutionary conservation of stable features of 3D protein structures.
Now AlphaFold gives a possibility to predict, or rather recognize stable protein structures from their a.a. sequences without considering the process of protein folding that creates these structures.

We emphasize that the this study *does not* diminish the merit and utility of AlphaFolds; it only explains the basis of their success.

On the basis of AlphaFold:

RoseTTAFold:
Anishchenko I., … , Baker D. - De novo protein **design** by deep network hallucination. *Nature* **600**, 547–552 (2021). https://doi.org/10.1038/s41586-021-04184-w.

AF-multimer:
Gao, M., …, Skolnick J. - AF2Complex predicts direct physical **interactions in multimeric proteins** with deep learning. *Nat Commun* **13**, 1744 (2022). https://doi.org/10.1038/s41467-022-29394-2

OpenFold:
Ahdritz G., …, AlQuraishi M. - OpenFold: retraining AlphaFold2 yields new **insights into its learning** mechanisms and capacity for generalization. *Nat Methods* **21**, 1514–1524 (2024). https://doi.org/10.1038/s41592-024-02272-z

AlphaFold 3:
Abramson J., …, Jumper J.M. - Accurate structure prediction of **biomolecular interactions** with AlphaFold 3. *Nature* **630**, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w

etc.

Thanks for your attention!

Protein 3D Structure Identification by AlphaFold: a Physics-Based *Prediction* or *Recognition* Based on Huge Databases?

Alexei V. Finkelstein[1,2], Dmitry N. Ivankov[3]

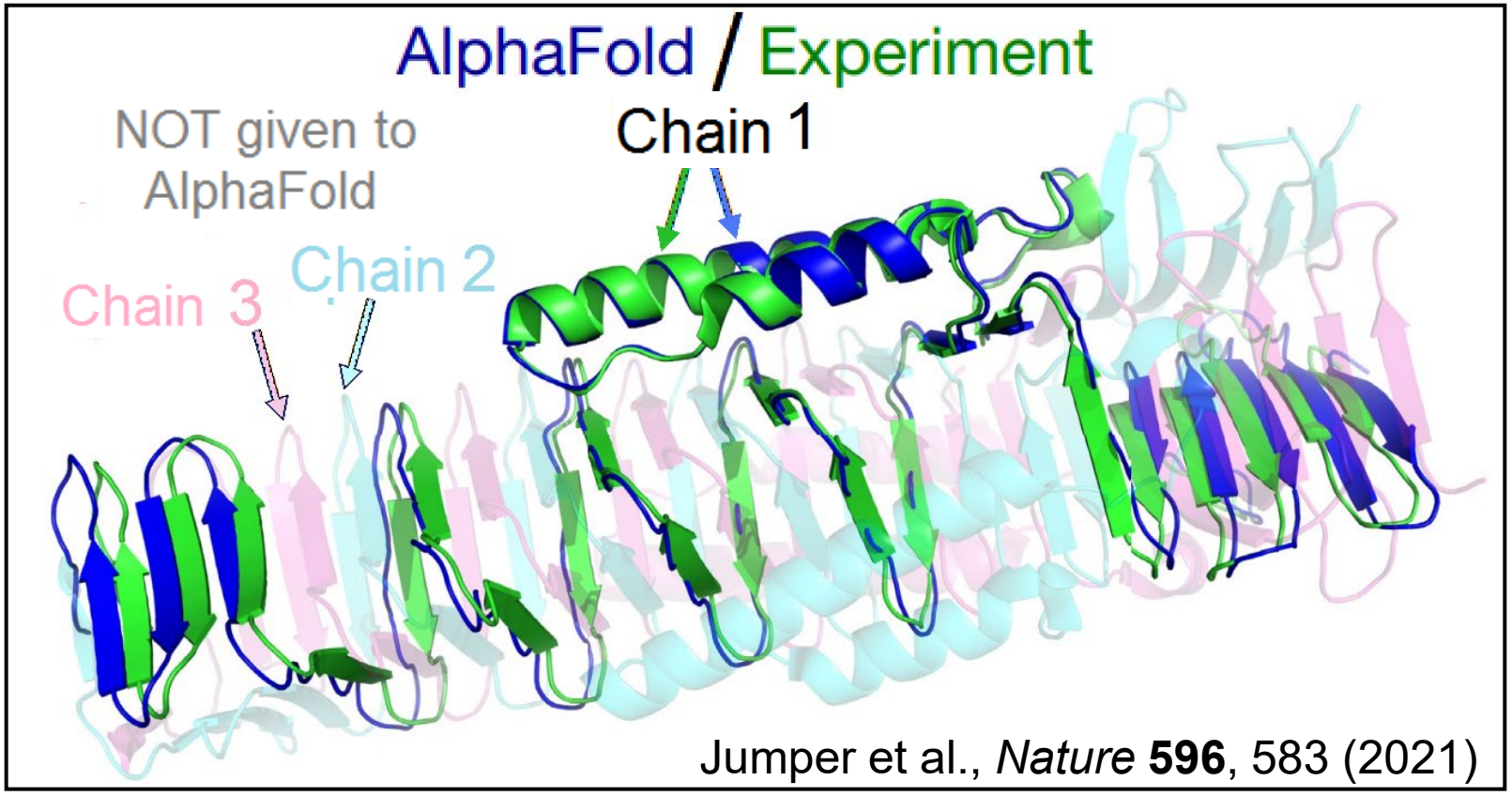
[1]Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia
[2]Biology Department, Lomonosov Moscow State University, Moscow, Russia
[3]Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia

E-mail: afinkel@vega.protres.ru

We are grateful to N.V. Dovidchenko, S.O. Garbuzynskiy, and especially J. Jumper for discussions, and the RSF (grant № 21-14-00268) for support

AlphaFold is NOT driven by physics:

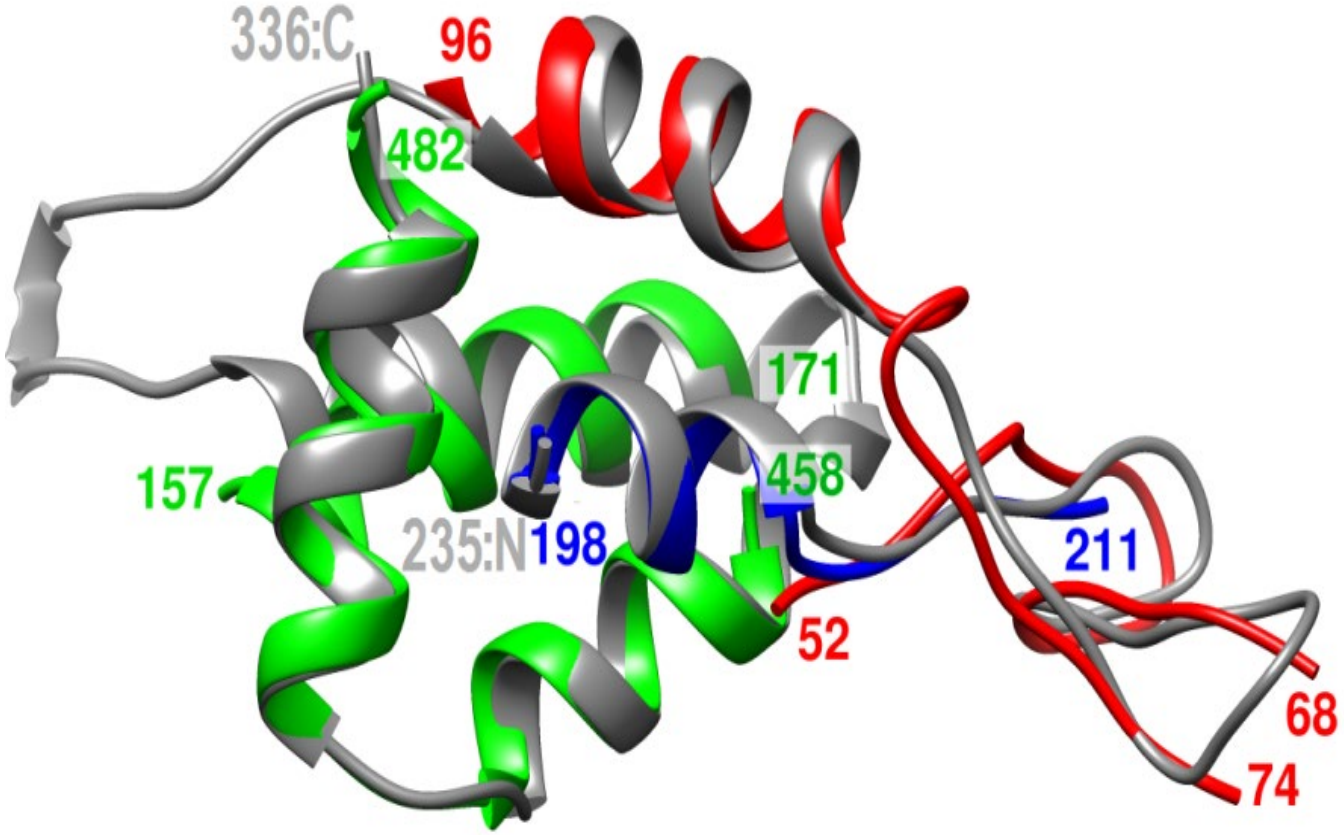
AlphaFold, which is given a.a. sequence of only one of the three intertwined protein chains, recognizes its spatial structure



— which, due to its complete non-compactness,

— **cannot** be stable on its own!

One of many dozens of examples of superposition of pieces of already known 3D structures on a novel fold



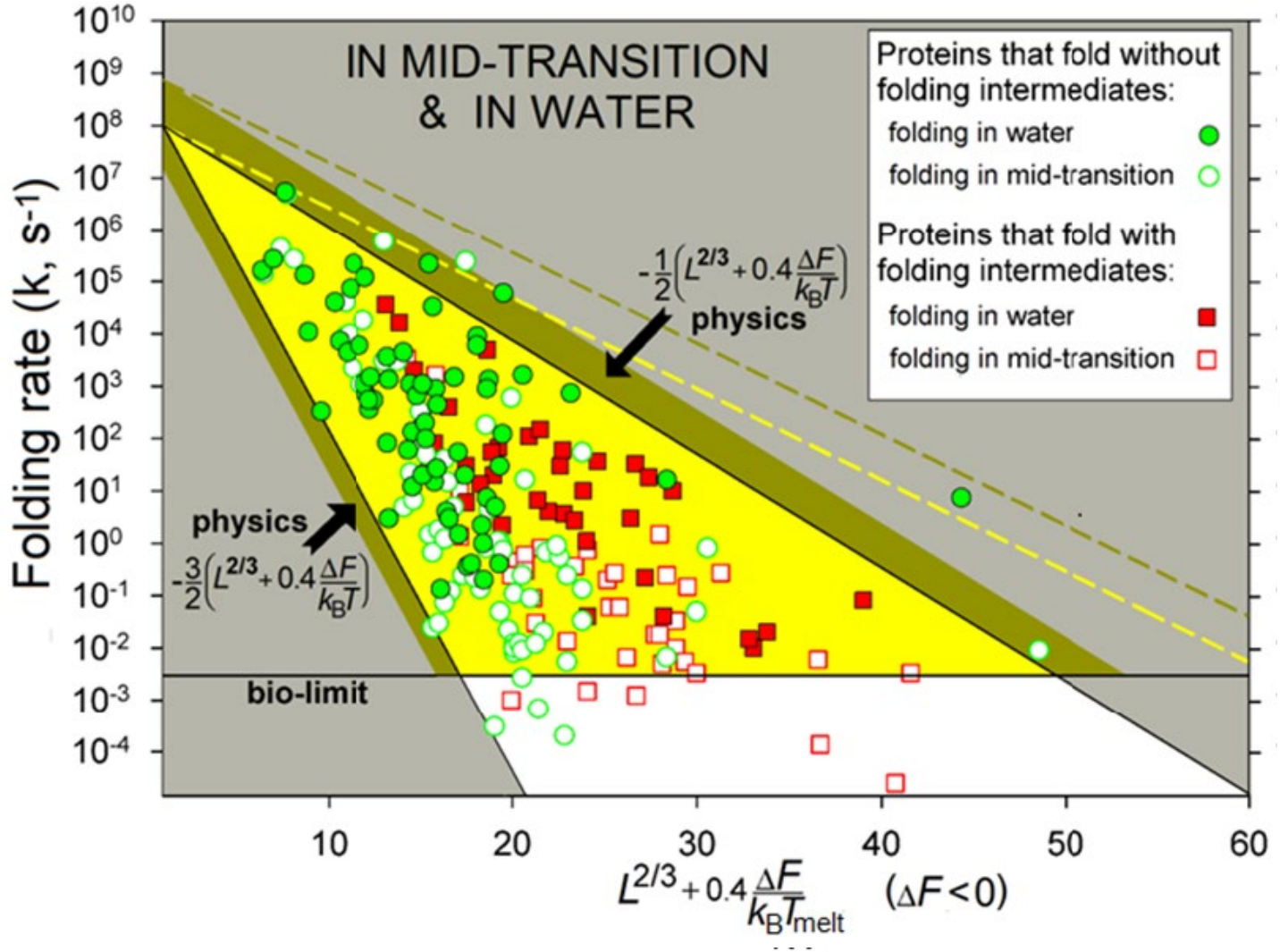
"Novel fold" (6VR4, chain A - target T1035 from CASP 14) as a combination of fragments of 3 already known structures available to AlphaFold during the training:

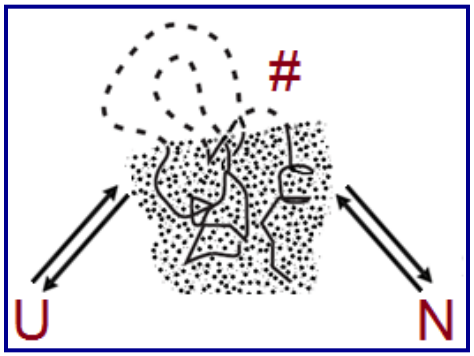
1GB3, chain A; 5A29, chain A; 5W40, chain B.

“Predict folding” (*folding rate*!) ≠ “Predict fold”

$$k_f \simeq \exp\left\{-\left(\frac{1}{2}\div\frac{3}{2}\right)\left[L^{2/3}+0.4\left[\frac{\Delta F}{k_B T}\right]\right]\right\}\frac{0.1}{\text{ns}}$$

Solution of the “Levinthal’s paradox





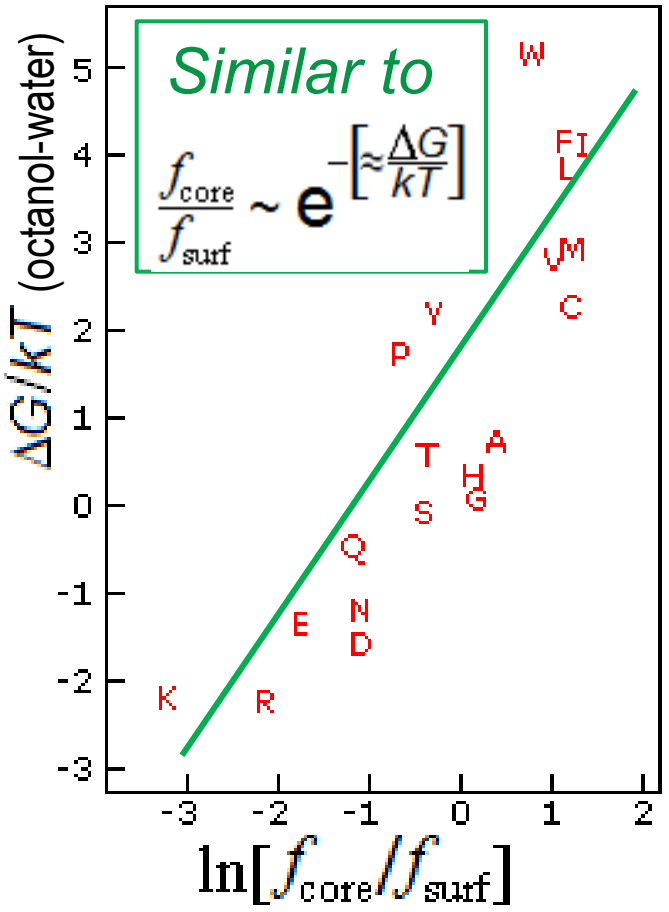
Phase transition

Finkelstein, Badretdinov, *1997,1998*; Garbuzynskiy, Ivankov, Bogatyreva, Finkelstein, *PNAS*, 2013

The occurrence of elements of protein structures is associated with their stability (Finkelstein et al., Proteins, 23: 142-150, 1995)

Small details of protein structures

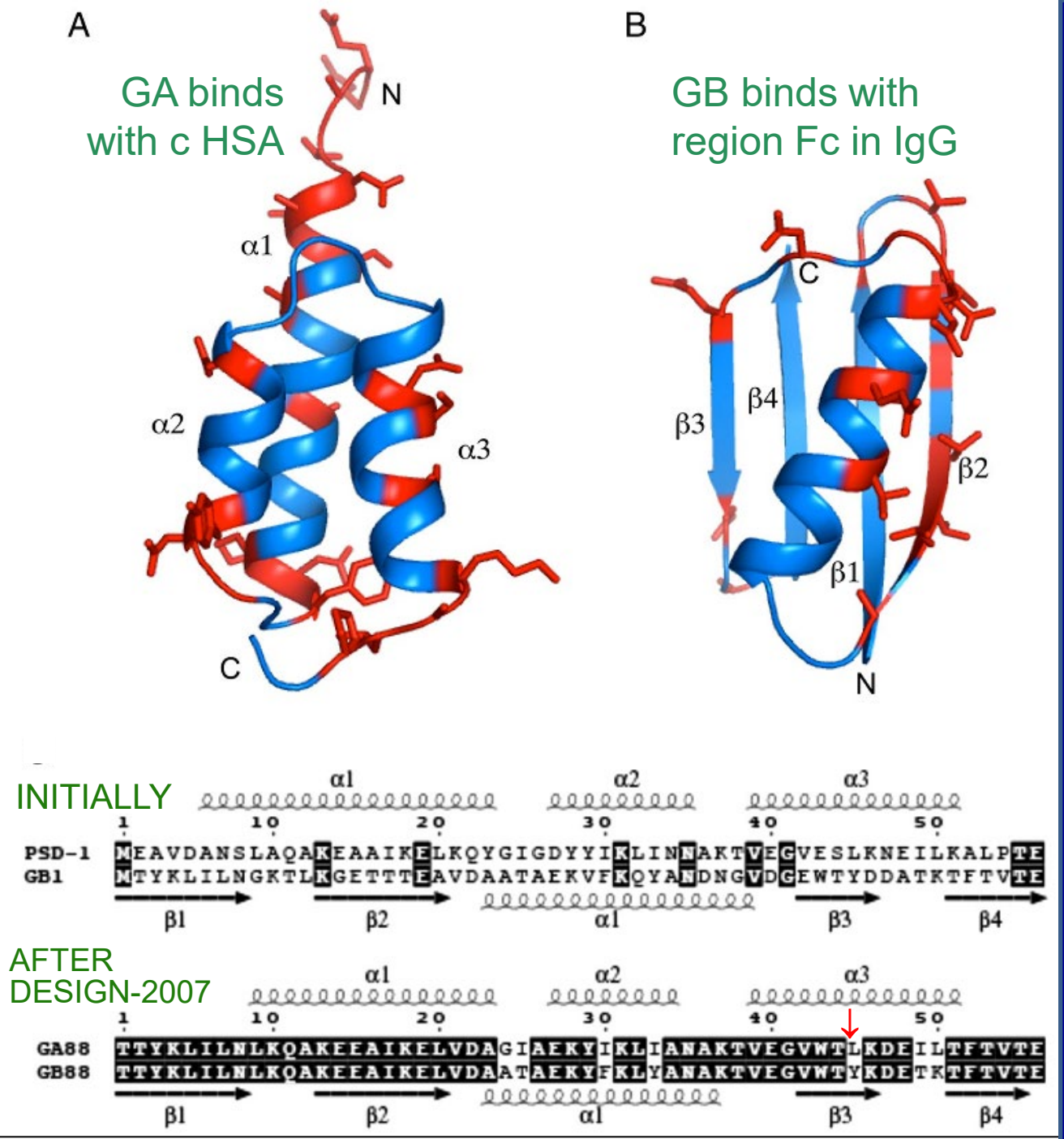
Example:



Miller, Janin, Chothia, 1984

Similar to Boltzmann-Gibbs statistics (F.M. Pohl, 1971)

The reason is a *selection* of *stable* structures



Needed:

Old chain fold and old activity – with a completely new a.a. sequence

P.A.Alexander, Y.He, Y.Chen, J.Orban, P.N.Bryan

PNAS, 2007, 104, 11963-8

The design and characterization of two proteins with 88% sequence identity but different structure and function

Y.He, Y.Chen, P.Alexander, P.N.Bryan, J.Orban

PNAS, 2008, 105, 14412-7

NMR structures of two designed proteins with high sequence identity but different fold and function

2012 (*Structure*, 20, 283-91):

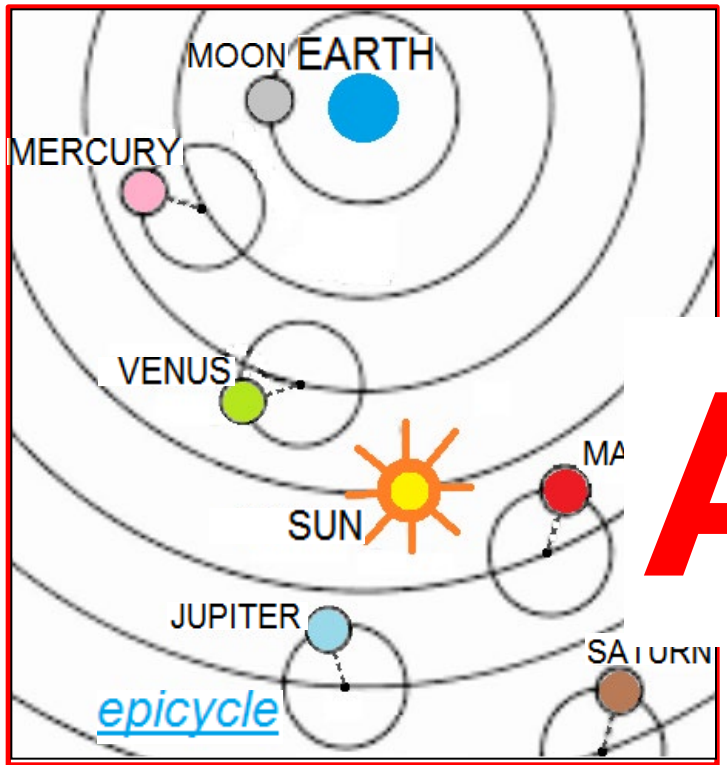
Difference: **ONE** a.a. residue!

GOOD PREDICTION ⇎ CORRECT UNDERSTANDING

Priests of Egypt and Babylon:

GOOD PREDICTIONS
of eclipses of the Sun and Moon
(based on huge archives spanning 2500 years!),

BUT: *fundamentally* WRONG UNDERSTANDING
(The Earth is flat!)



PTOLEMAEUS
(*using huge archives*):

GOOD PREDICTION

AlphaFold

***WRONG* UNDERSTANDING of the PROCESS!**