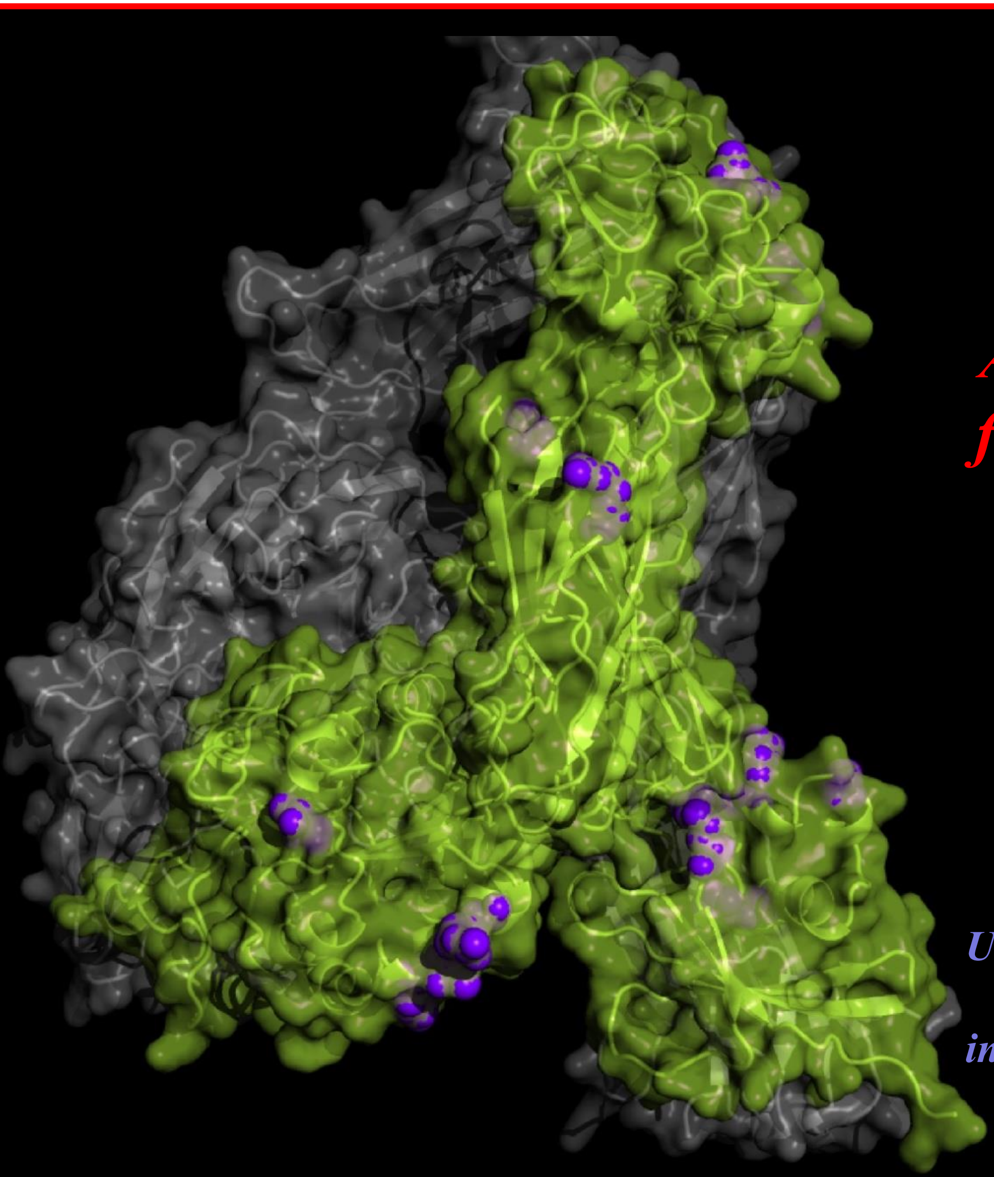


Artificial Intelligence Strategies for Protein Structure, Dynamics and Design

Alexandre G. de Brevern
w/ Dr. Yasser Mohseni Behbahani
& Pr. Jean-Christophe Gelly

DSIMB Bioinformatics team,
Université Paris Cité & Université de la Réunion,
INSERM, EFS, BIGR U1134,
inIdEx GREx, Necker Hospital, Paris, FRANCE.

short outline of the presentation

1. First AI approaches in Structural Bioinformatics
2. AlphaFold
3. Protein flexibility prediction
4. Pathogenicity prediction
5. Recent developments
6. Conclusion(s)

1. FIRST AI APPROACHES IN STRUCTURAL BIOINFORMATICS

First approaches in Bioinformatics

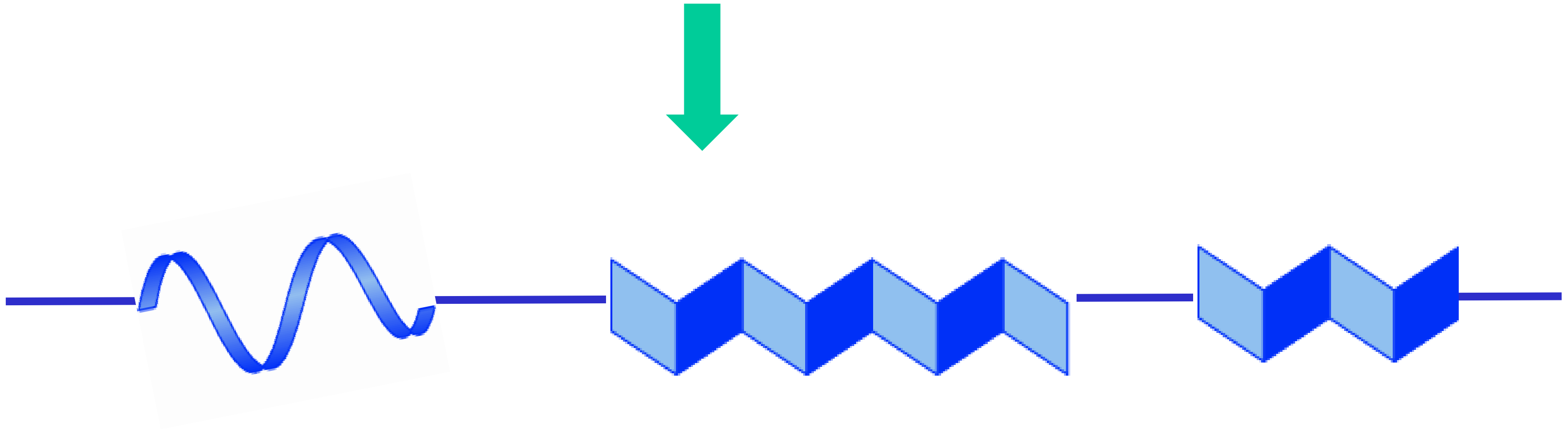
- Prediction of secondary structures from the sequences

…SVAWCLPKPLPEGTEDKDQTATIPSLSAMLGALFLWMFWPSFNSALLRSPIERKNAVFN…

First approaches in Bioinformatics

- Prediction of secondary structures from the sequences

…SVAWCLPKPLPEGTEDKDQTATIPSLSAMLGALFLWMFWPSFNSALLRSPIERKNAVFN…



First approaches in Bioinformatics

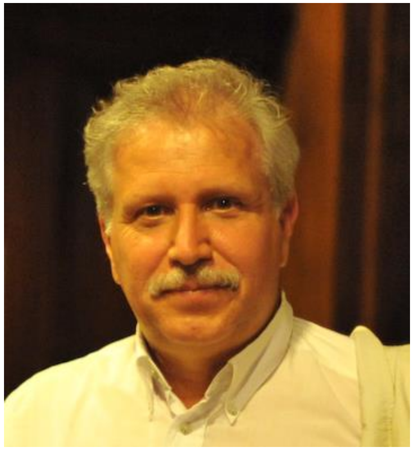
- **Stats -** Garnier–Osguthorpe–Robson (and later with Gibrat)

GOR I (1978) information theory, single-residue statistics Q_3= 58.0%

GOR II (1985) Improved dataset, expanded statistics Q_3 =**61.5%**

GOR III(1996) Conditional probability including neighbors Q_3=64.0%

GOR IV(1997) Longer windows, improved parameters, uses large curated datasets Q_3=**64.4%**

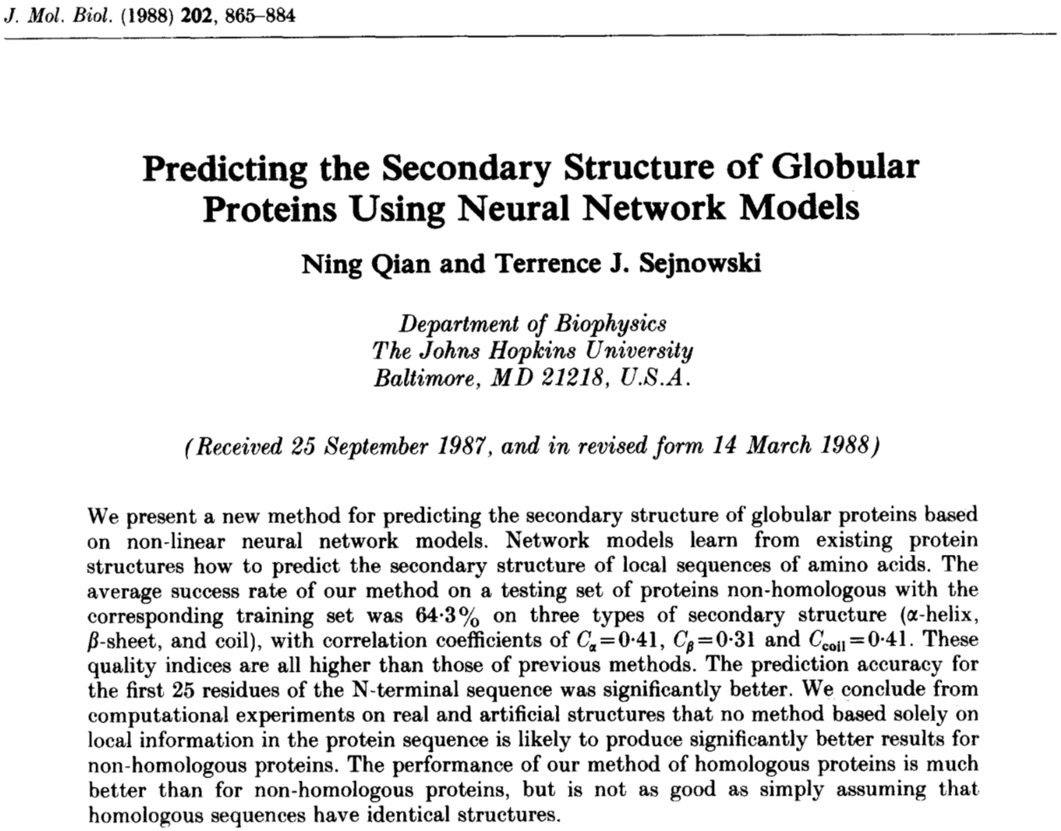


$$Q_3 = \frac{\text{right state predicted}}{\text{total number res.}}$$

First approaches in Bioinformatics

1988

- Qian and Sejnowski – Artificial Neural Networks



J. Mol. Biol. (1988) **202**, 865–884

Predicting the Secondary Structure of Globular Proteins Using Neural Network Models

Ning Qian and Terrence J. Sejnowski

Department of Biophysics
The Johns Hopkins University
Baltimore, MD 21218, U.S.A.

(Received 25 September 1987, and in revised form 14 March 1988)

We present a new method for predicting the secondary structure of globular proteins based on non-linear neural network models. Network models learn from existing protein structures how to predict the secondary structure of local sequences of amino acids. The average success rate of our method on a testing set of proteins non-homologous with the corresponding training set was 64·3% on three types of secondary structure (α-helix, β-sheet, and coil), with correlation coefficients of $C_\alpha = 0{\cdot}41$, $C_\beta = 0{\cdot}31$ and $C_{\text{coil}} = 0{\cdot}41$. These quality indices are all higher than those of previous methods. The prediction accuracy for the first 25 residues of the N-terminal sequence was significantly better. We conclude from computational experiments on real and artificial structures that no method based solely on local information in the protein sequence is likely to produce significantly better results for non-homologous proteins. The performance of our method of homologous proteins is much better than for non-homologous proteins, but is not as good as simply assuming that homologous sequences have identical structures.

First approaches in Bioinformatics

1986

- Qian and Sejnowski

Single network $Q_3 = 62.3\%$

GOR II (1985) $Q_3 = 61.5\%$

GOR III (1996) $Q_3 = 64.0\%$

GOR IV(1997) $Q_3 = 64.4\%$

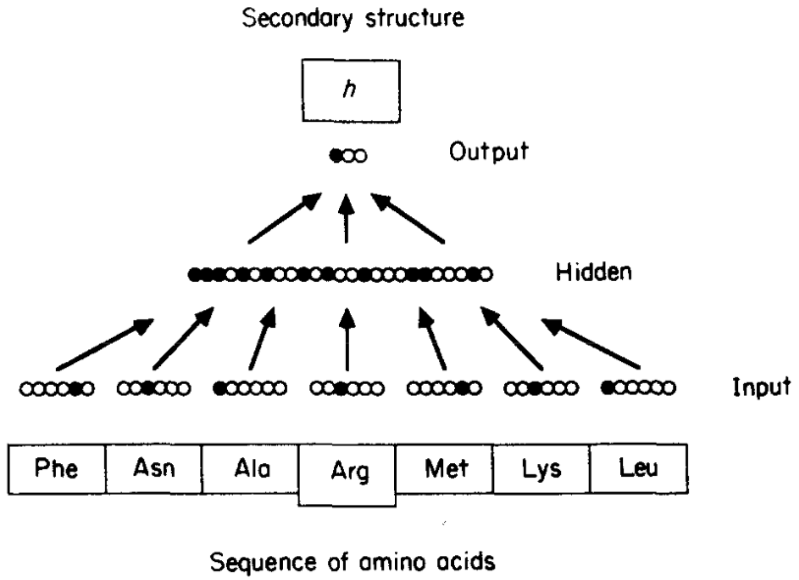


Figure 1. A diagram of network architecture. The standard network had 13 input groups, with 21 units/group, representing a stretch of 13 contiguous amino acids (only 7 input groups and 7 units/group are illustrated). Information from the input layer is transformed by an intermediate layer of "hidden" units to produce a pattern of activity in 3 output units, which represent the secondary structure prediction for the central amino acid.

First approaches in Bioinformatics

1986

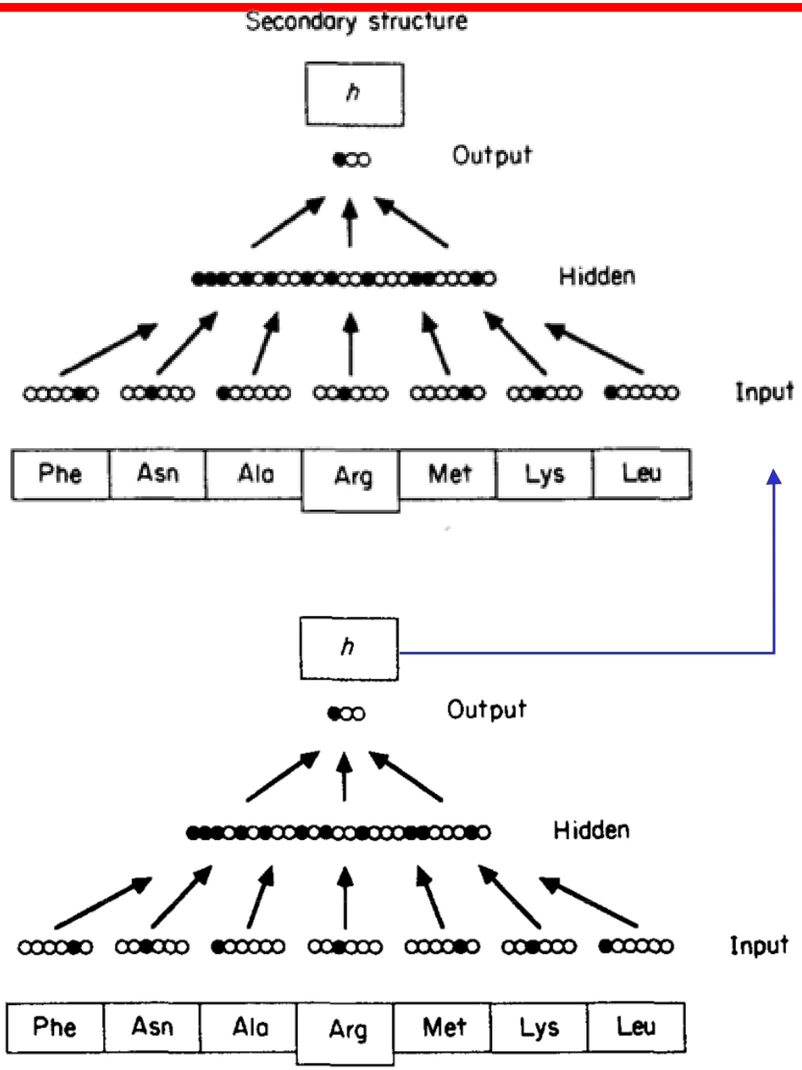
- Qian and Sejnowski

Single network $Q_3 = 62.3\%$

2-nets* $Q_3 = \mathbf{64.3\%}$

GOR IV(1997) $Q_3 = \mathbf{64.4\%}$

Not bad but GOR is statistics

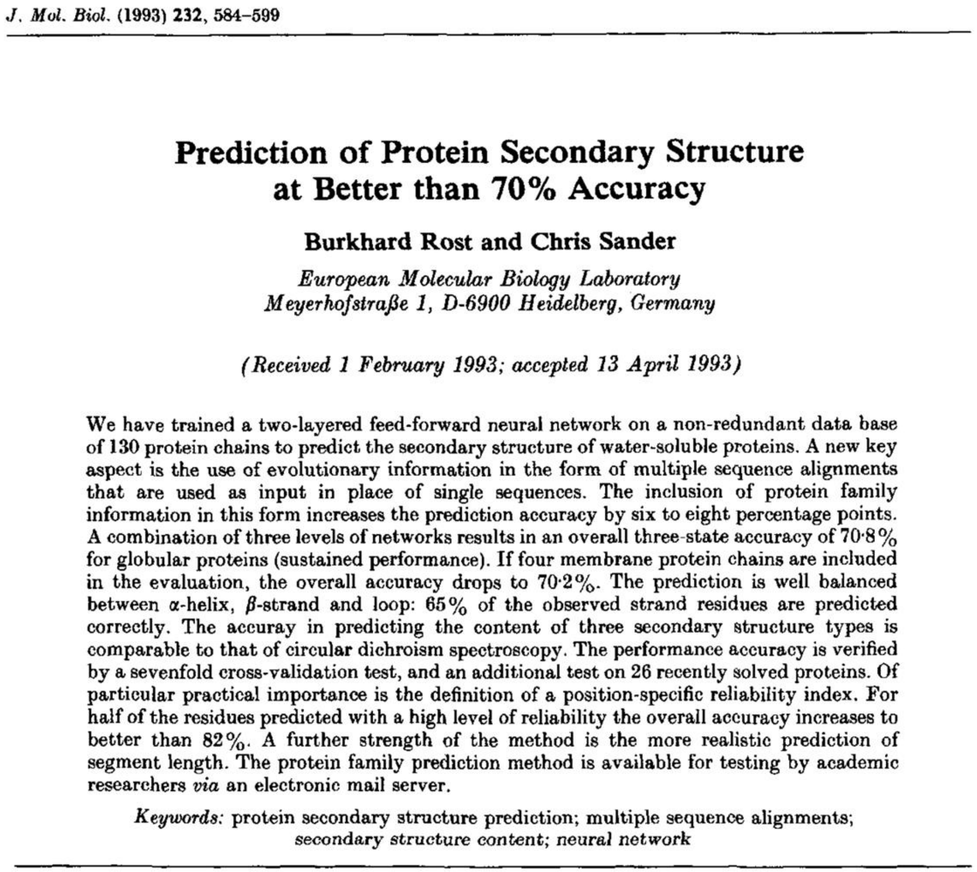


**Network cascade*

First approaches in Bioinformatics

1993

- The Breaking point: PHD by Rost and Sander



J. Mol. Biol. (1993) **232**, 584–599

Prediction of Protein Secondary Structure at Better than 70% Accuracy

Burkhard Rost and Chris Sander

European Molecular Biology Laboratory
Meyerhofstraße 1, D-6900 Heidelberg, Germany

(Received 1 February 1993; accepted 13 April 1993)

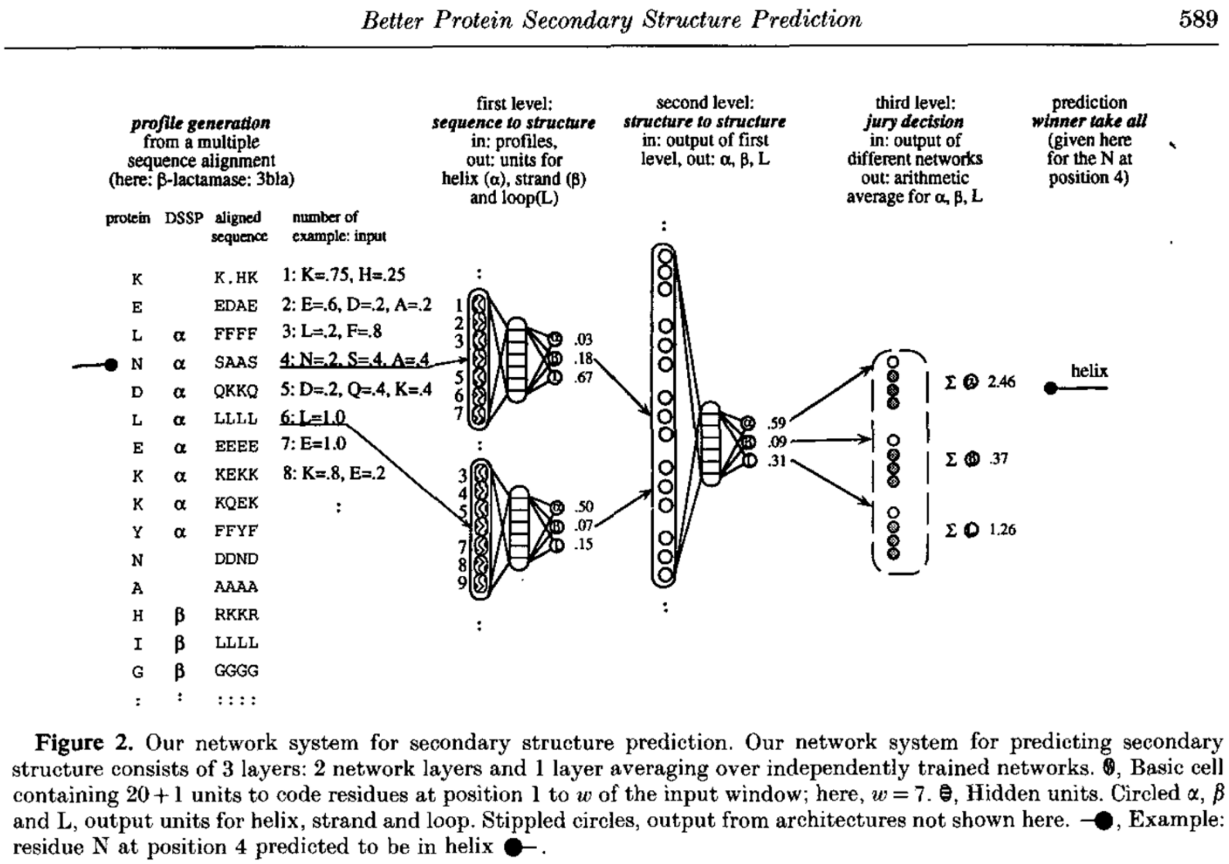
We have trained a two-layered feed-forward neural network on a non-redundant data base of 130 protein chains to predict the secondary structure of water-soluble proteins. A new key aspect is the use of evolutionary information in the form of multiple sequence alignments that are used as input in place of single sequences. The inclusion of protein family information in this form increases the prediction accuracy by six to eight percentage points. A combination of three levels of networks results in an overall three-state accuracy of 70·8% for globular proteins (sustained performance). If four membrane protein chains are included in the evaluation, the overall accuracy drops to 70·2%. The prediction is well balanced between α-helix, β-strand and loop: 65% of the observed strand residues are predicted correctly. The accuray in predicting the content of three secondary structure types is comparable to that of circular dichroism spectroscopy. The performance accuracy is verified by a sevenfold cross-validation test, and an additional test on 26 recently solved proteins. Of particular practical importance is the definition of a position-specific reliability index. For half of the residues predicted with a high level of reliability the overall accuracy increases to better than 82%. A further strength of the method is the more realistic prediction of segment length. The protein family prediction method is available for testing by academic researchers *via* an electronic mail server.

Keywords: protein secondary structure prediction; multiple sequence alignments; secondary structure content; neural network

First approaches in Bioinformatics

1993

- The Breaking point: PHD by Rost and Sander



Better Protein Secondary Structure Prediction 589

protein	DSSP	aligned sequence	number of example: input
K		K.HK	1: K=.75, H=.25
E		EDAE	2: E=.6, D=.2, A=.2
L	α	FFFF	3: L=.2, F=.8
N	α	SAAS	4: N=.2, S=.4, A=.4
D	α	QKKQ	5: D=.2, Q=.4, K=.4
L	α	LLLL	6: L=1.0
E	α	EEEE	7: E=1.0
K	α	KEKK	8: K=.8, E=.2
K	α	KQEK	:
Y	α	FFYF	
N		DDND	
A		AAAA	
H	β	RKKR	
I	β	LLLL	
G	β	GGGG	
:	:	::::	

Figure 2. Our network system for secondary structure prediction. Our network system for predicting secondary structure consists of 3 layers: 2 network layers and 1 layer averaging over independently trained networks. ⊛, Basic cell containing 20 + 1 units to code residues at position 1 to w of the input window; here, $w = 7$. ⊖, Hidden units. Circled α, β and L, output units for helix, strand and loop. Stippled circles, output from architectures not shown here. —●, Example: residue N at position 4 predicted to be in helix ●—.

First approaches in Bioinformatics

1993

➢ The Breaking point: PHD by Rost and Sander

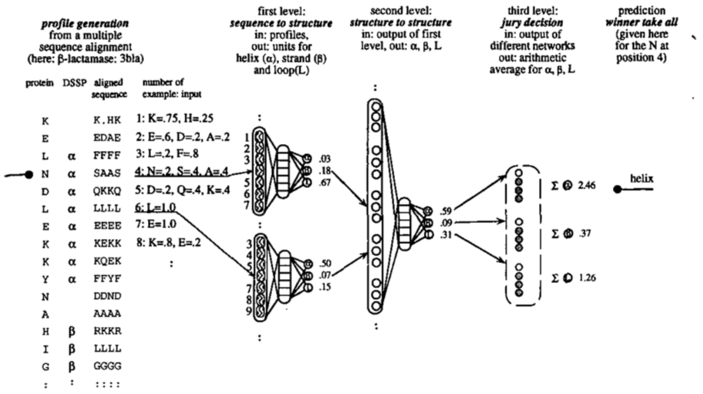
PHD (Profile network from HeiDelberg)

Similar networks !

But not the same data :

Evolution (PSI-BLAST-type profiles)

Better Protein Secondary Structure Prediction 589



protein	DSSP	aligned sequence	number of example: input
K		K.HK	1: K=.75, H=.25
E		EDAE	2: E=.6, D=.2, A=.2
L	α	FFFF	3: L=.2, F=.8
N	α	SAAS	4: N=.2, S=.4, A=.4
D	α	QKKQ	5: D=.2, Q=.4, K=.4
L	α	LLLL	6: L=1.0
E	α	EEEE	7: E=1.0
K	α	KEKK	8: K=.8, E=.2
K	α	KQEK	:
Y	α	FFYF	
N		DDND	
A		AAAA	
H	β	RKKR	
I	β	LLLL	
G	β	GGGG	
:	:	::::	

Figure 2. Our network system for secondary structure prediction. Our network system for predicting secondary structure consists of 3 layers: 2 network layers and 1 layer averaging over independently trained networks. ⓢ, Basic cell containing 20 + 1 units to code residues at position 1 to w of the input window; here, $w = 7$. ⊖, Hidden units. Circled α, β and L, output units for helix, strand and loop. Stippled circles, output from architectures not shown here. —●, Example: residue N at position 4 predicted to be in helix ●—.

First approaches in Bioinformatics

1993-97

- Evolution of PHD by Rost and Sander

(1993) Q_3= 71%

(1995) Q_3 = 72.2% on RS126 and a webserver !

(1997) Q_3 = 76.0% on RS126 and 513

Improvements included refine datasets, change in the window, and real PSSM from PSI-BLAST.

First approaches in Bioinformatics

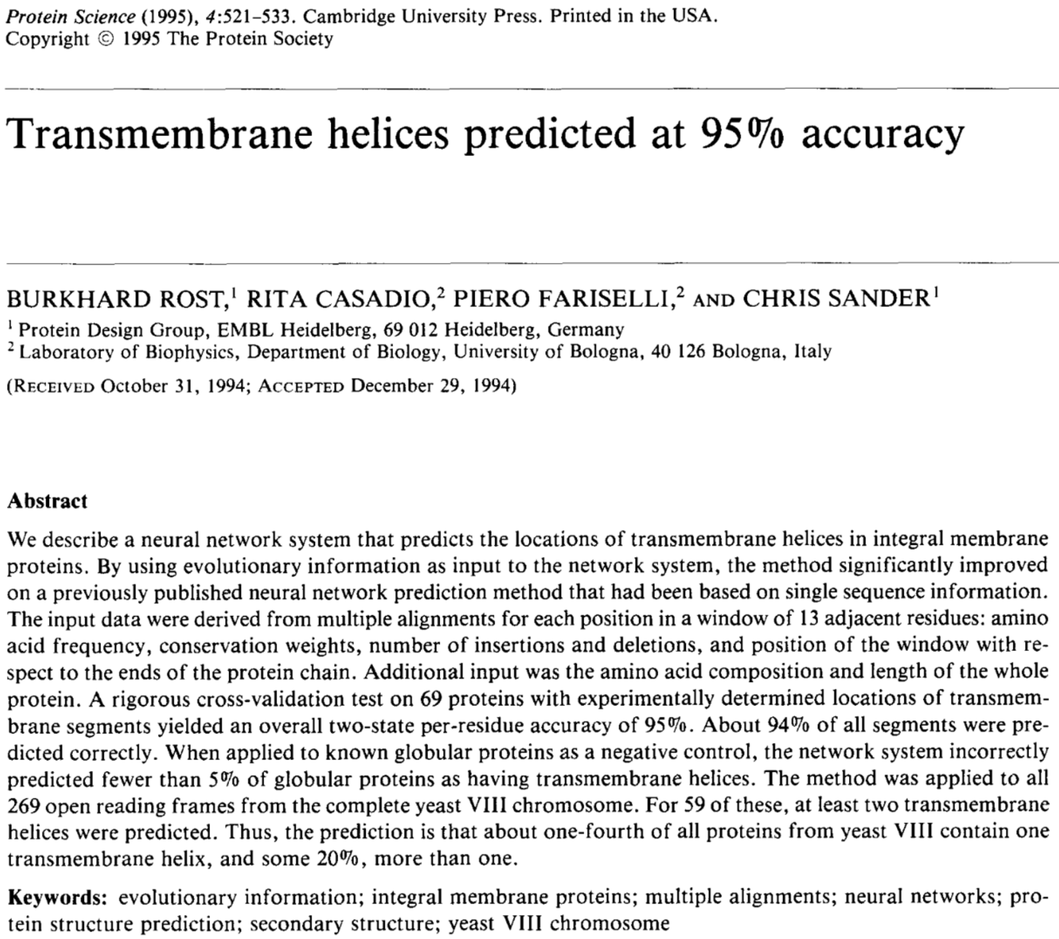
1993-97

- Evolution of PHD by Rost and Sander

(1993)	Q_3= 71%
(1995)	Q_3 = 72.2% on RS126
(1997)	Q_3 = **76.0%** on RS126
Qian and Sejnowski (1988)	Q_3 = 64.3%
GOR IV(1997)	Q_3= **64.4%**

First approaches in Bioinformatics

- More impressive: PHDtm for transmembrane helices



Protein Science (1995), *4*:521–533. Cambridge University Press. Printed in the USA.

Transmembrane helices predicted at 95% accuracy

BURKHARD ROST,[1] RITA CASADIO,[2] PIERO FARISELLI,[2] AND CHRIS SANDER[1]
[1] Protein Design Group, EMBL Heidelberg, 69 012 Heidelberg, Germany
[2] Laboratory of Biophysics, Department of Biology, University of Bologna, 40 126 Bologna, Italy

(RECEIVED October 31, 1994; ACCEPTED December 29, 1994)

Abstract

We describe a neural network system that predicts the locations of transmembrane helices in integral membrane proteins. By using evolutionary information as input to the network system, the method significantly improved on a previously published neural network prediction method that had been based on single sequence information. The input data were derived from multiple alignments for each position in a window of 13 adjacent residues: amino acid frequency, conservation weights, number of insertions and deletions, and position of the window with respect to the ends of the protein chain. Additional input was the amino acid composition and length of the whole protein. A rigorous cross-validation test on 69 proteins with experimentally determined locations of transmembrane segments yielded an overall two-state per-residue accuracy of 95%. About 94% of all segments were predicted correctly. When applied to known globular proteins as a negative control, the network system incorrectly predicted fewer than 5% of globular proteins as having transmembrane helices. The method was applied to all 269 open reading frames from the complete yeast VIII chromosome. For 59 of these, at least two transmembrane helices were predicted. Thus, the prediction is that about one-fourth of all proteins from yeast VIII contain one transmembrane helix, and some 20%, more than one.

Keywords: evolutionary information; integral membrane proteins; multiple alignments; neural networks; protein structure prediction; secondary structure; yeast VIII chromosome

First approaches in Bioinformatics

- PHDtm for transmembrane helices (TMb or not)

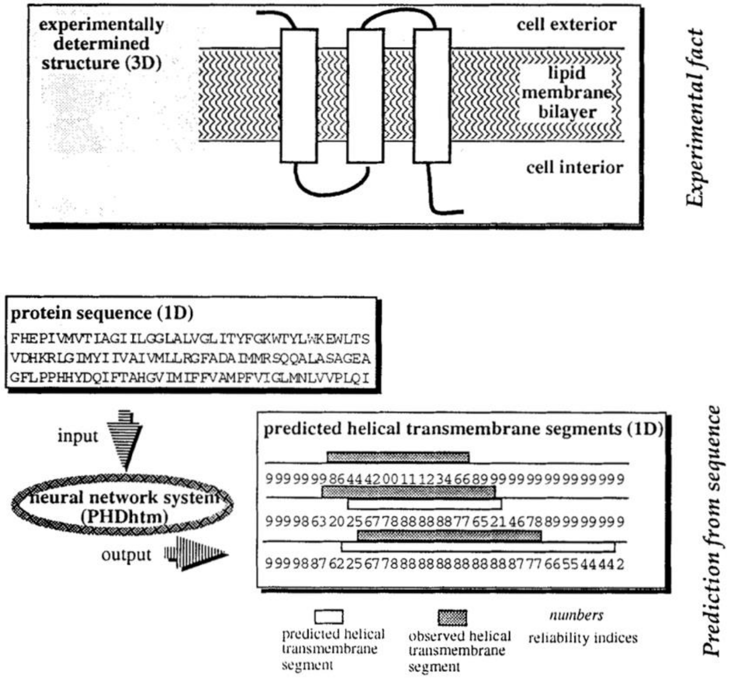
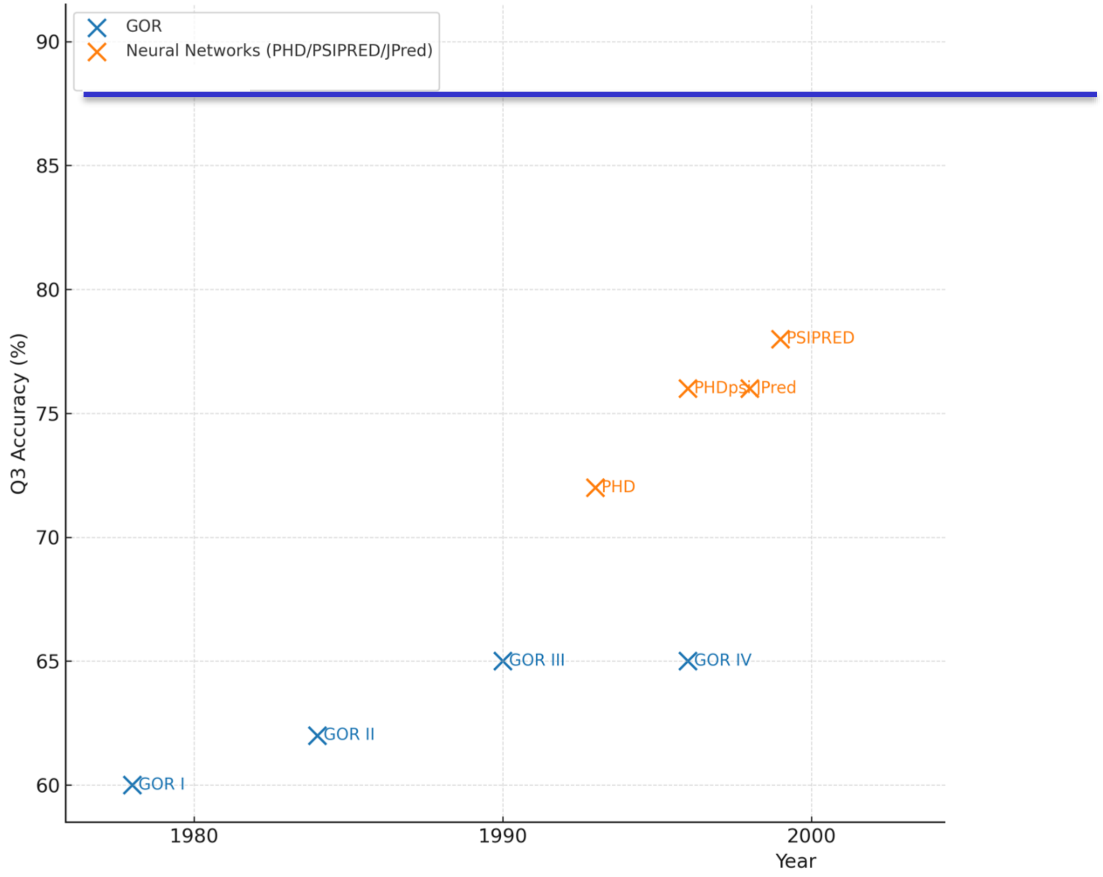

experimentally determined structure (3D)
cell exterior
lipid membrane bilayer
cell interior
Experimental fact
protein sequence (1D)
FHEPIVMVTIAGIILGGLALVGLITYFGKWTYLWKEWLTS
VDHKRLGIMYIIVAIVMLLRGFADAIMMRSQQALASAGEA
GFLPPHHYDQIFTAHGVIMIFFVAMPFVIGLMNLVVPLQI
input
neural network system (PHDhtm)
output
predicted helical transmembrane segments (1D)
9999999864442001112346689999999999999999
9999863202567788888887765214678899999999
9999887622567788888888888888777665544442
predicted helical transmembrane segment
observed helical transmembrane segment
numbers
reliability indices
Prediction from sequence

Fig. 1. Prediction of the location of transmembrane helices. In one class of membrane proteins, typically apolar helical segments are embedded in the lipid bilayer oriented perpendicular to the surface of the membrane. Helical segments can be regarded as more or less rigid cylinders. Thus, the 3D structure of the membrane spanning protein region can be determined by: the location of segments with respect to sequence; the orientation of helical axes; the inclination of helical axes with respect to lipid bilayer; and the phase of helices with respect to each other (orientation of helical wheel). Here, we simplify extremely by projecting 3D structure onto a 1D string describing which residues of the protein are part of a transmembrane helices. Input to the prediction tool (neural network system) is a protein sequence (in general a sequence alignment), output is a prediction of the location of transmembrane segments. The example shown (sequence of cytochrome O ubiquinol oxidase subunit I, cyob_eco in SWISS-PROT; Bairoch & Boeckmann, 1994) contained one of the few segments that were underpredicted (missed). The numbers give the reliability of the prediction for each residue on a scale of 0–9 (Fig. 2). Nontransmembrane regions, when predicted correctly, usually reached the highest reliability (9). Thus, the unusually low reliability values for the underpredicted segment might have enabled the expert user to improve the automatic prediction by interpreting this region as nonloop.

First approaches in Bioinformatics

- Gain in prediction accuracy



Theoretical limit (88%)

Made with ChatGPT

First approaches in Bioinformatics

- But is everything really so beautiful …

First approaches in Bioinformatics

- But is everything really so beautiful …
- dataset RS 126 … is not really non – redundant ... so it bias the results …

First approaches in Bioinformatics

- But is everything really so beautiful …
- dataset RS 126 … is not really non – redundant ... so it bias the results …
- And the dataset 513 is in fact also not really cleaned…

First approaches in Bioinformatics

- But is everything really so beautiful …

- dataset RS 126 … is not really non – redundant ... so it bias the results …

- And the dataset 513 is in fact also not really cleaned…

- The issue is that some non-specialists still use them !!!

First approaches in Bioinformatics

- But is everything really so beautiful …
- Dataset RS 126 … is not really non – redundant ... so it bias the results …
- And the dataset 513 is in fact also not really cleaned…
- The issue is that some non-specialists still use them !!!

Important point: Data, data, data …

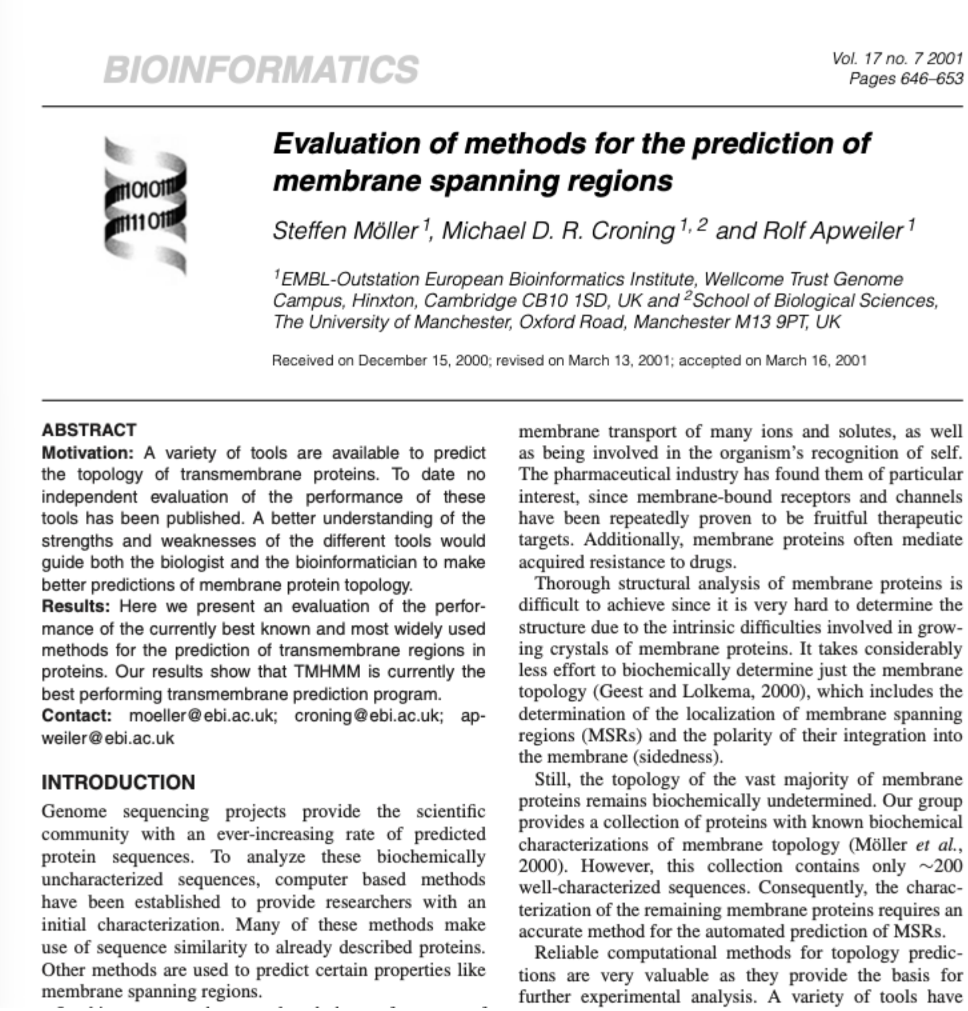
Inserm

Institut national
de la santé et de la recherche médicale

First approaches in Bioinformatics

Bioinformatique & Interactions

- But is everything really so beautiful …



BIOINFORMATICS

Vol. 17 no. 7 2001
Pages 646–653

Evaluation of methods for the prediction of membrane spanning regions

Steffen Möller[1], *Michael D. R. Croning*[1, 2] *and Rolf Apweiler*[1]

[1]*EMBL-Outstation European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and* [2]*School of Biological Sciences, The University of Manchester, Oxford Road, Manchester M13 9PT, UK*

Received on December 15, 2000; revised on March 13, 2001; accepted on March 16, 2001

ABSTRACT

Motivation: A variety of tools are available to predict the topology of transmembrane proteins. To date no independent evaluation of the performance of these tools has been published. A better understanding of the strengths and weaknesses of the different tools would guide both the biologist and the bioinformatician to make better predictions of membrane protein topology.

Results: Here we present an evaluation of the performance of the currently best known and most widely used methods for the prediction of transmembrane regions in proteins. Our results show that TMHMM is currently the best performing transmembrane prediction program.

Contact: moeller@ebi.ac.uk; croning@ebi.ac.uk; apweiler@ebi.ac.uk

INTRODUCTION

Genome sequencing projects provide the scientific community with an ever-increasing rate of predicted protein sequences. To analyze these biochemically uncharacterized sequences, computer based methods have been established to provide researchers with an initial characterization. Many of these methods make use of sequence similarity to already described proteins. Other methods are used to predict certain properties like membrane spanning regions.

membrane transport of many ions and solutes, as well as being involved in the organism's recognition of self. The pharmaceutical industry has found them of particular interest, since membrane-bound receptors and channels have been repeatedly proven to be fruitful therapeutic targets. Additionally, membrane proteins often mediate acquired resistance to drugs.

Thorough structural analysis of membrane proteins is difficult to achieve since it is very hard to determine the structure due to the intrinsic difficulties involved in growing crystals of membrane proteins. It takes considerably less effort to biochemically determine just the membrane topology (Geest and Lolkema, 2000), which includes the determination of the localization of membrane spanning regions (MSRs) and the polarity of their integration into the membrane (sidedness).

Still, the topology of the vast majority of membrane proteins remains biochemically undetermined. Our group provides a collection of proteins with known biochemical characterizations of membrane topology (Möller *et al.*, 2000). However, this collection contains only ~200 well-characterized sequences. Consequently, the characterization of the remaining membrane proteins requires an accurate method for the automated prediction of MSRs.

Reliable computational methods for topology predictions are very valuable as they provide the basis for further experimental analysis. A variety of tools have

First approaches in Bioinformatics

- But is everything really so beautiful …

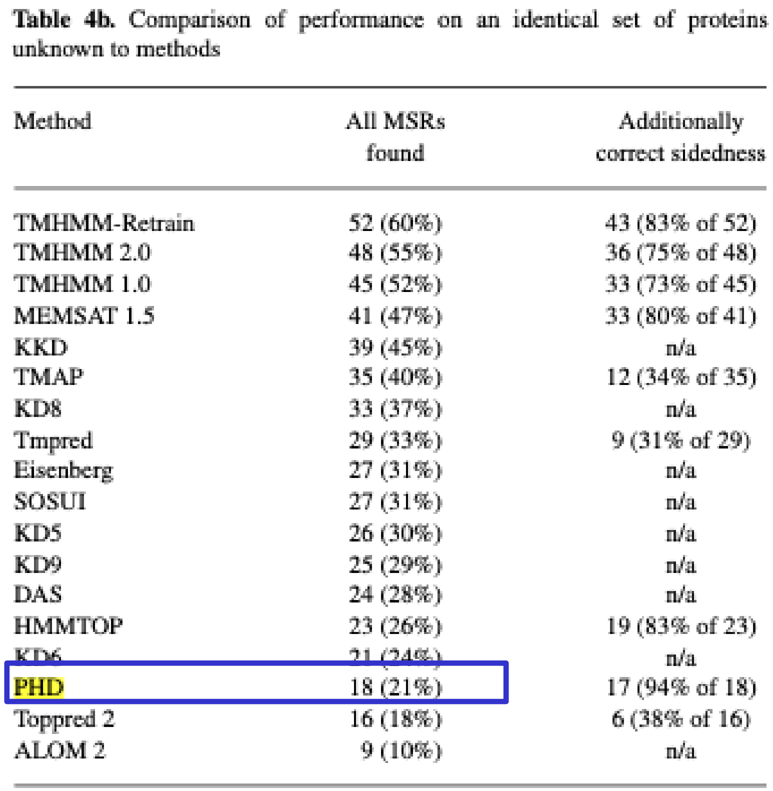


Table 4b. Comparison of performance on an identical set of proteins unknown to methods

Method	All MSRs found	Additionally correct sidedness
TMHMM-Retrain	52 (60%)	43 (83% of 52)
TMHMM 2.0	48 (55%)	36 (75% of 48)
TMHMM 1.0	45 (52%)	33 (73% of 45)
MEMSAT 1.5	41 (47%)	33 (80% of 41)
KKD	39 (45%)	n/a
TMAP	35 (40%)	12 (34% of 35)
KD8	33 (37%)	n/a
Tmpred	29 (33%)	9 (31% of 29)
Eisenberg	27 (31%)	n/a
SOSUI	27 (31%)	n/a
KD5	26 (30%)	n/a
KD9	25 (29%)	n/a
DAS	24 (28%)	n/a
HMMTOP	23 (26%)	19 (83% of 23)
KD6	21 (24%)	n/a
PHD	18 (21%)	17 (94% of 18)
Toppred 2	16 (18%)	6 (38% of 16)
ALOM 2	9 (10%)	n/a

PHDtm

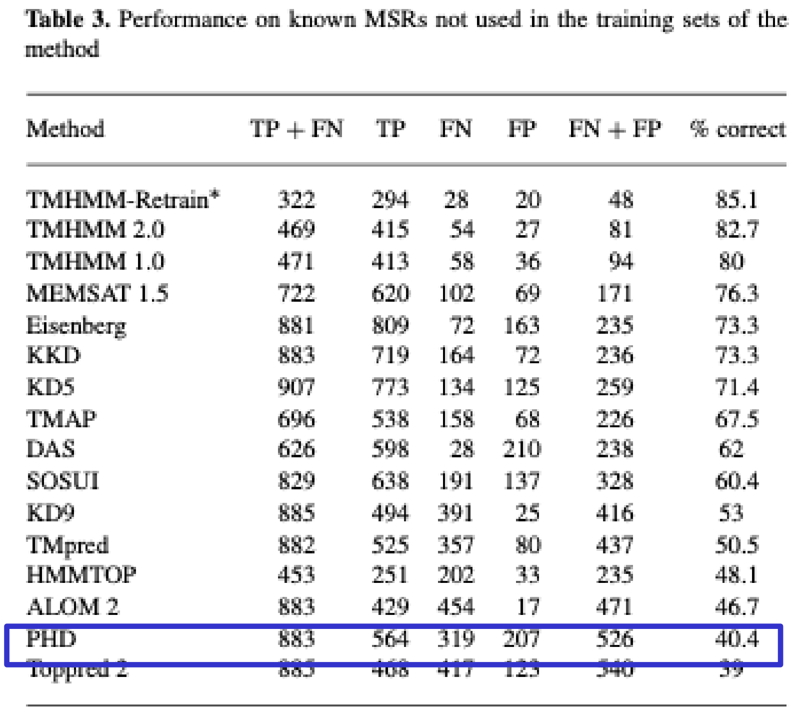


Table 3. Performance on known MSRs not used in the training sets of the method

Method	TP + FN	TP	FN	FP	FN + FP	% correct
TMHMM-Retrain*	322	294	28	20	48	85.1
TMHMM 2.0	469	415	54	27	81	82.7
TMHMM 1.0	471	413	58	36	94	80
MEMSAT 1.5	722	620	102	69	171	76.3
Eisenberg	881	809	72	163	235	73.3
KKD	883	719	164	72	236	73.3
KD5	907	773	134	125	259	71.4
TMAP	696	538	158	68	226	67.5
DAS	626	598	28	210	238	62
SOSUI	829	638	191	137	328	60.4
KD9	885	494	391	25	416	53
TMpred	882	525	357	80	437	50.5
HMMTOP	453	251	202	33	235	48.1
ALOM 2	883	429	454	17	471	46.7
PHD	883	564	319	207	526	40.4
Toppred 2	885	468	417	123	540	39

First approaches in Bioinformatics

- But is everything really so beautiful …

How PHDtm can drops from 95% to > 50% when using new data ?

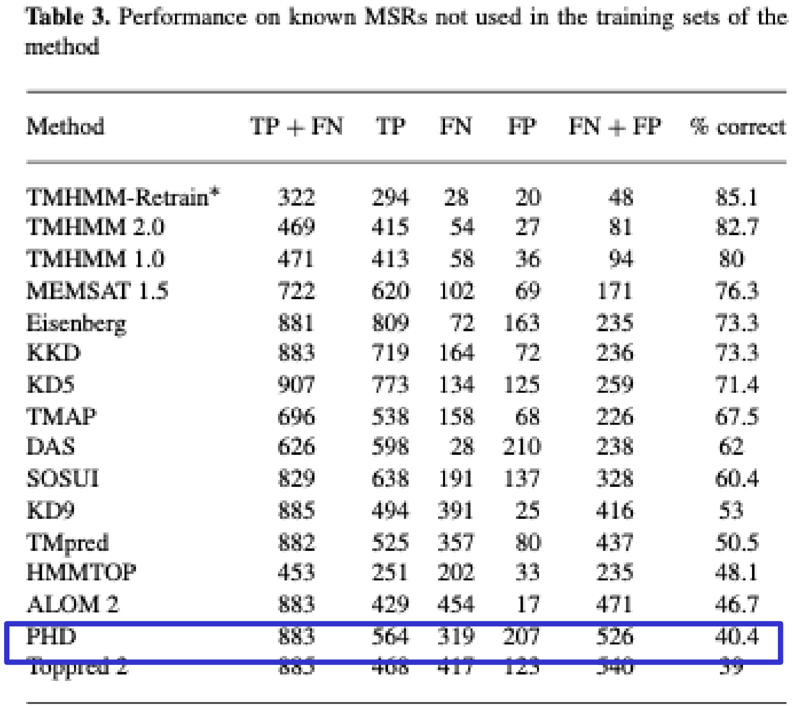


Table 3. Performance on known MSRs not used in the training sets of the method

Method	TP + FN	TP	FN	FP	FN + FP	% correct
TMHMM-Retrain*	322	294	28	20	48	85.1
TMHMM 2.0	469	415	54	27	81	82.7
TMHMM 1.0	471	413	58	36	94	80
MEMSAT 1.5	722	620	102	69	171	76.3
Eisenberg	881	809	72	163	235	73.3
KKD	883	719	164	72	236	73.3
KD5	907	773	134	125	259	71.4
TMAP	696	538	158	68	226	67.5
DAS	626	598	28	210	238	62
SOSUI	829	638	191	137	328	60.4
KD9	885	494	391	25	416	53
TMpred	882	525	357	80	437	50.5
HMMTOP	453	251	202	33	235	48.1
ALOM 2	883	429	454	17	471	46.7
PHD	883	564	319	207	526	40.4
Toppred 2	885	468	417	123	540	39

First approaches in Bioinformatics

- But is everything really so beautiful …

Original training dataset:

0 structure … (logical)

So delineation of TMb were taken from UniProt, i.e. prediction, and it is Not good to do prediction on prediction …

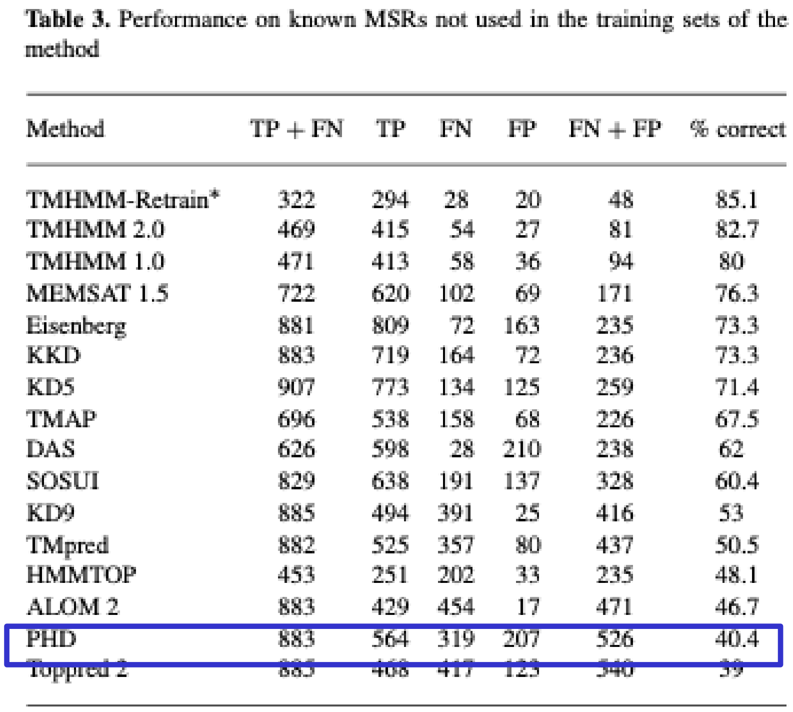


Table 3. Performance on known MSRs not used in the training sets of the method

Method	TP + FN	TP	FN	FP	FN + FP	% correct
TMHMM-Retrain*	322	294	28	20	48	85.1
TMHMM 2.0	469	415	54	27	81	82.7
TMHMM 1.0	471	413	58	36	94	80
MEMSAT 1.5	722	620	102	69	171	76.3
Eisenberg	881	809	72	163	235	73.3
KKD	883	719	164	72	236	73.3
KD5	907	773	134	125	259	71.4
TMAP	696	538	158	68	226	67.5
DAS	626	598	28	210	238	62
SOSUI	829	638	191	137	328	60.4
KD9	885	494	391	25	416	53
TMpred	882	525	357	80	437	50.5
HMMTOP	453	251	202	33	235	48.1
ALOM 2	883	429	454	17	471	46.7
PHD	883	564	319	207	526	40.4
Toppred 2	883	468	417	123	340	39

First approaches in Bioinformatics

- But is everything really so beautiful …

Original training dataset:

0 structure … (logical)
so delineation of TMb
were taken from UniProt,
i.e. prediction, and it is
Not good to do prediction on
prediction …

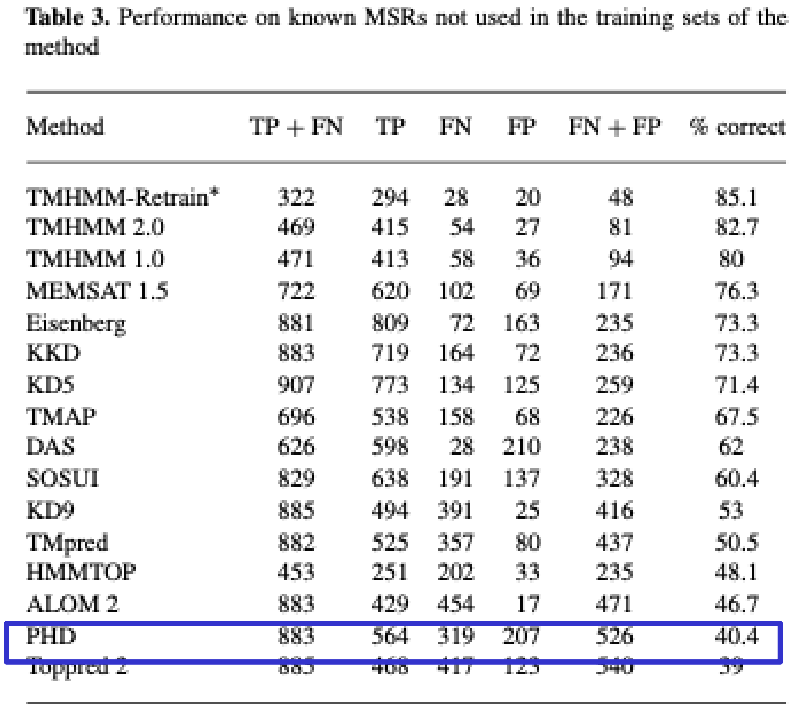


Table 3. Performance on known MSRs not used in the training sets of the method

Method	TP + FN	TP	FN	FP	FN + FP	% correct
TMHMM-Retrain*	322	294	28	20	48	85.1
TMHMM 2.0	469	415	54	27	81	82.7
TMHMM 1.0	471	413	58	36	94	80
MEMSAT 1.5	722	620	102	69	171	76.3
Eisenberg	881	809	72	163	235	73.3
KKD	883	719	164	72	236	73.3
KD5	907	773	134	125	259	71.4
TMAP	696	538	158	68	226	67.5
DAS	626	598	28	210	238	62
SOSUI	829	638	191	137	328	60.4
KD9	885	494	391	25	416	53
TMpred	882	525	357	80	437	50.5
HMMTOP	453	251	202	33	235	48.1
ALOM 2	883	429	454	17	471	46.7
PHD	883	564	319	207	526	40.4
Toppred 2	885	468	417	123	540	39

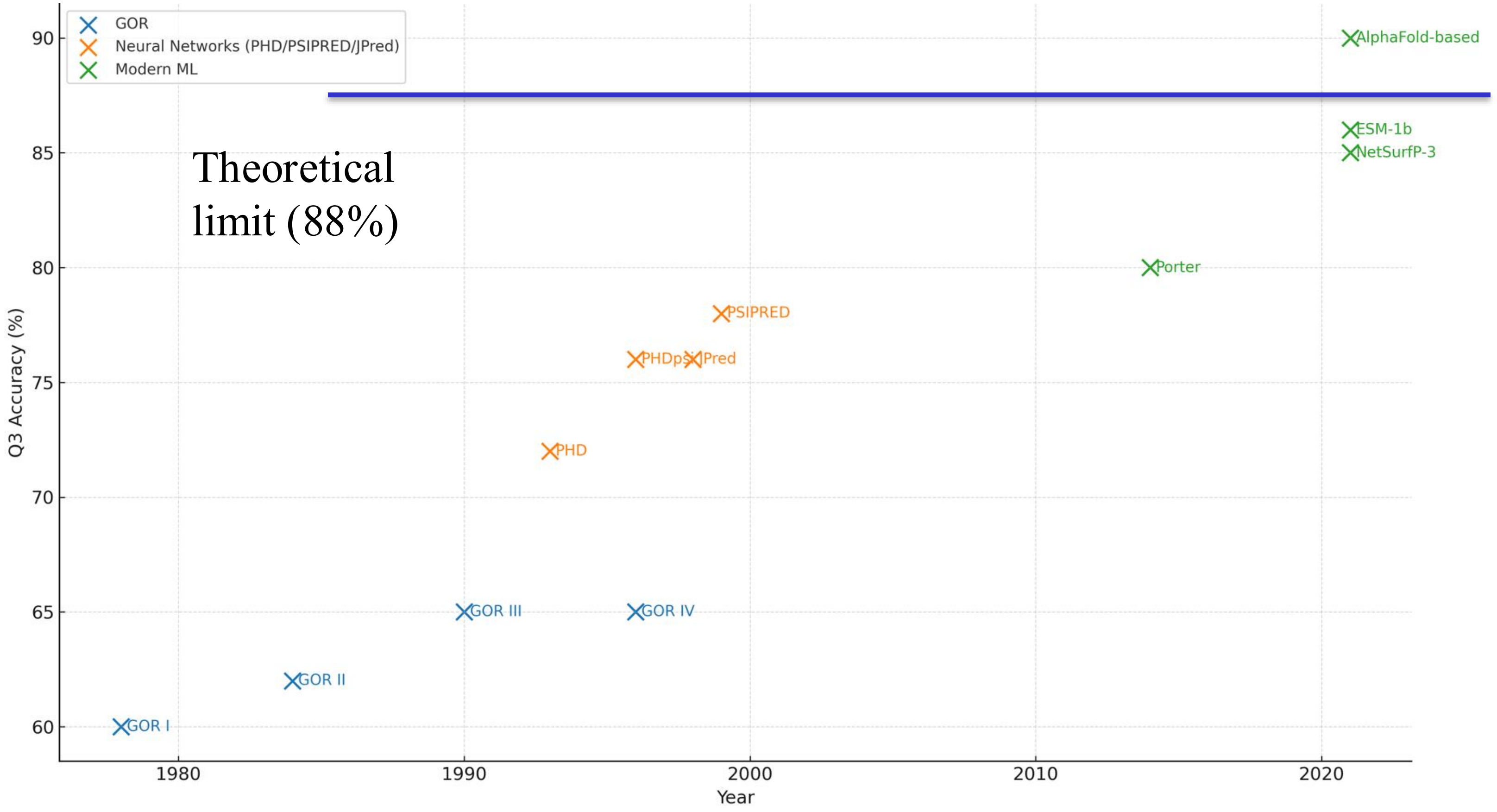
Important point: Data, data, data …

First approaches in Bioinformatics

- So we must be careful !

First approaches in Bioinformatics

➢ Gain in prediction accuracy

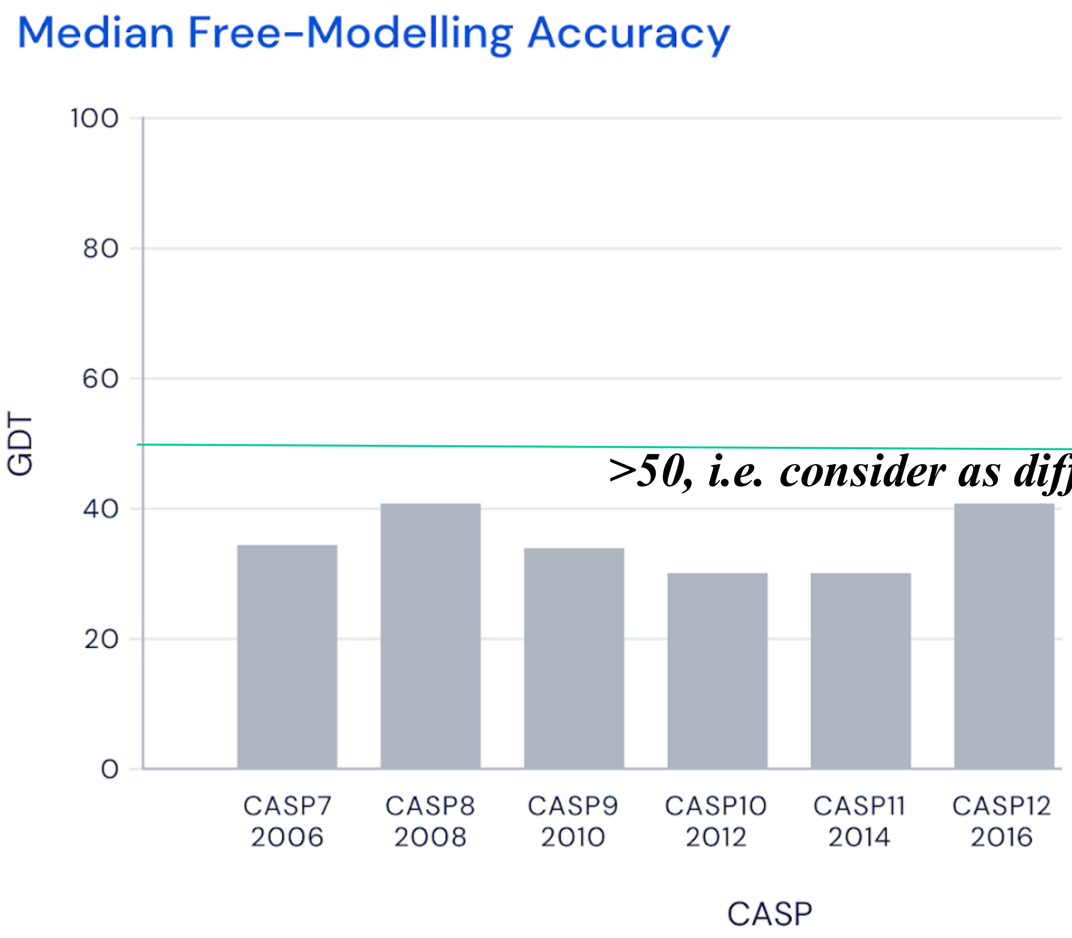


Theoretical limit (88%)

Made with ChatGPT

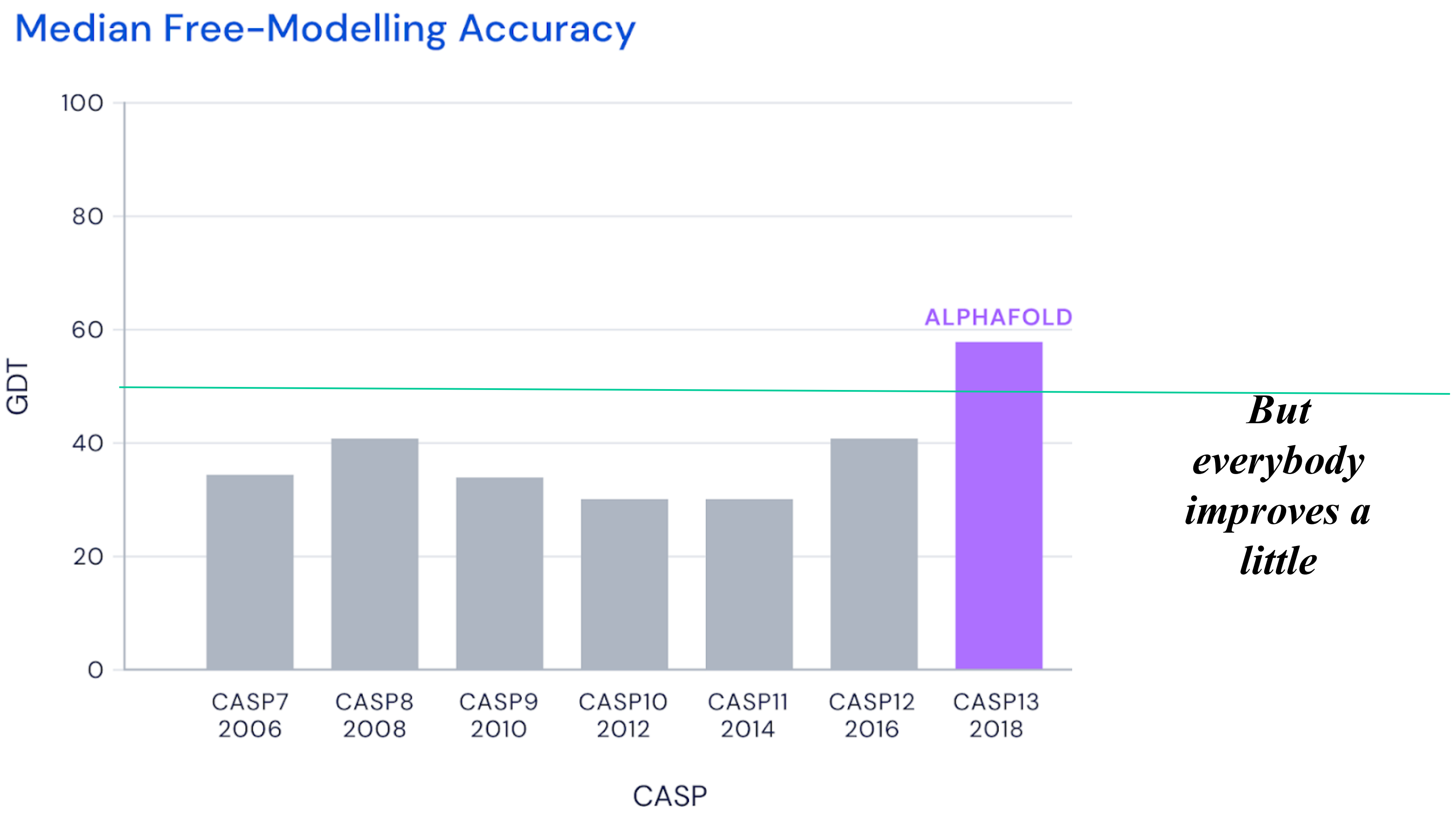
3. ALPHAFOLD

AlphaFold2



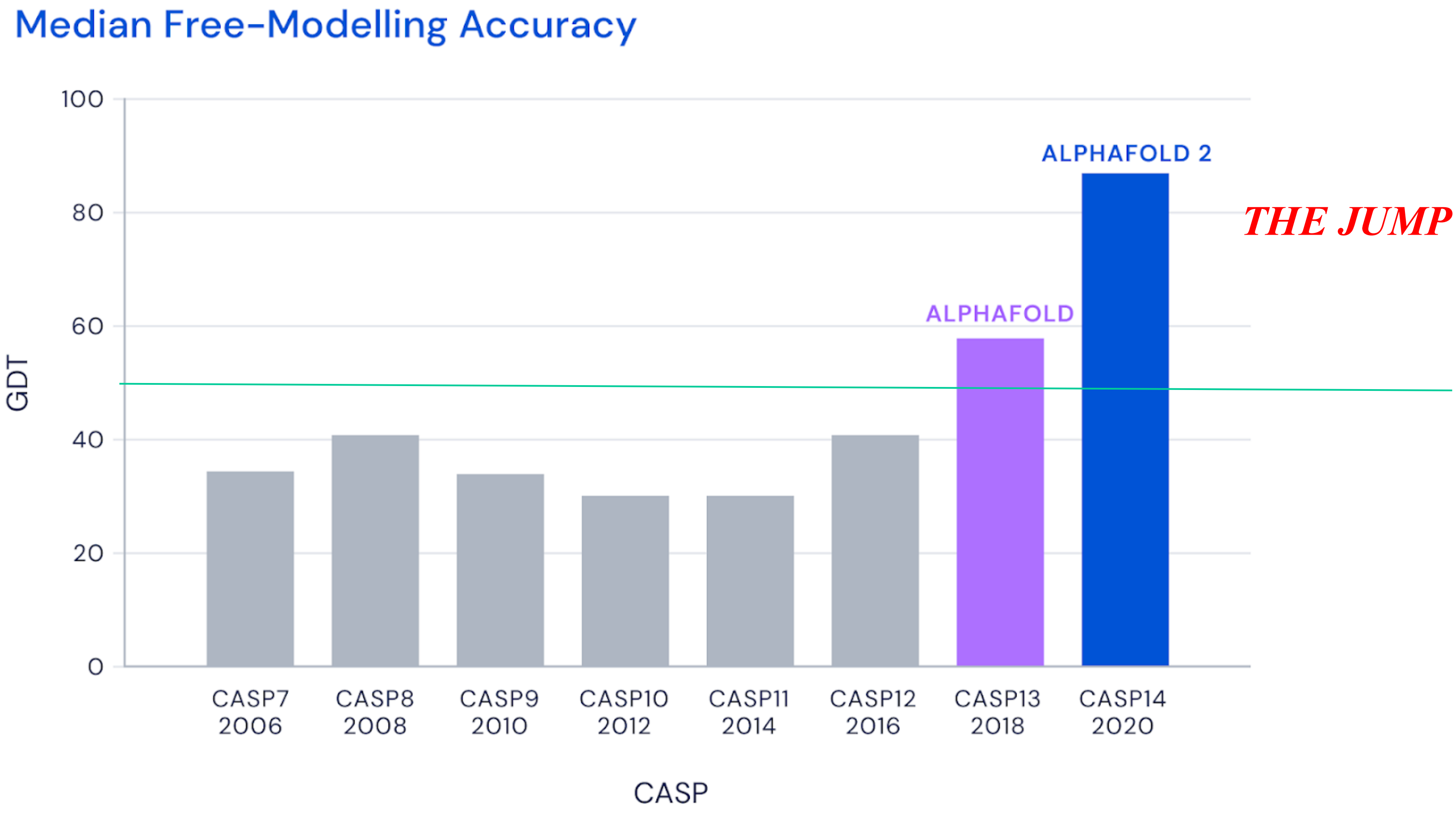
>50, i.e. consider as different fold, i.e. cannot be use

AlphaFold2



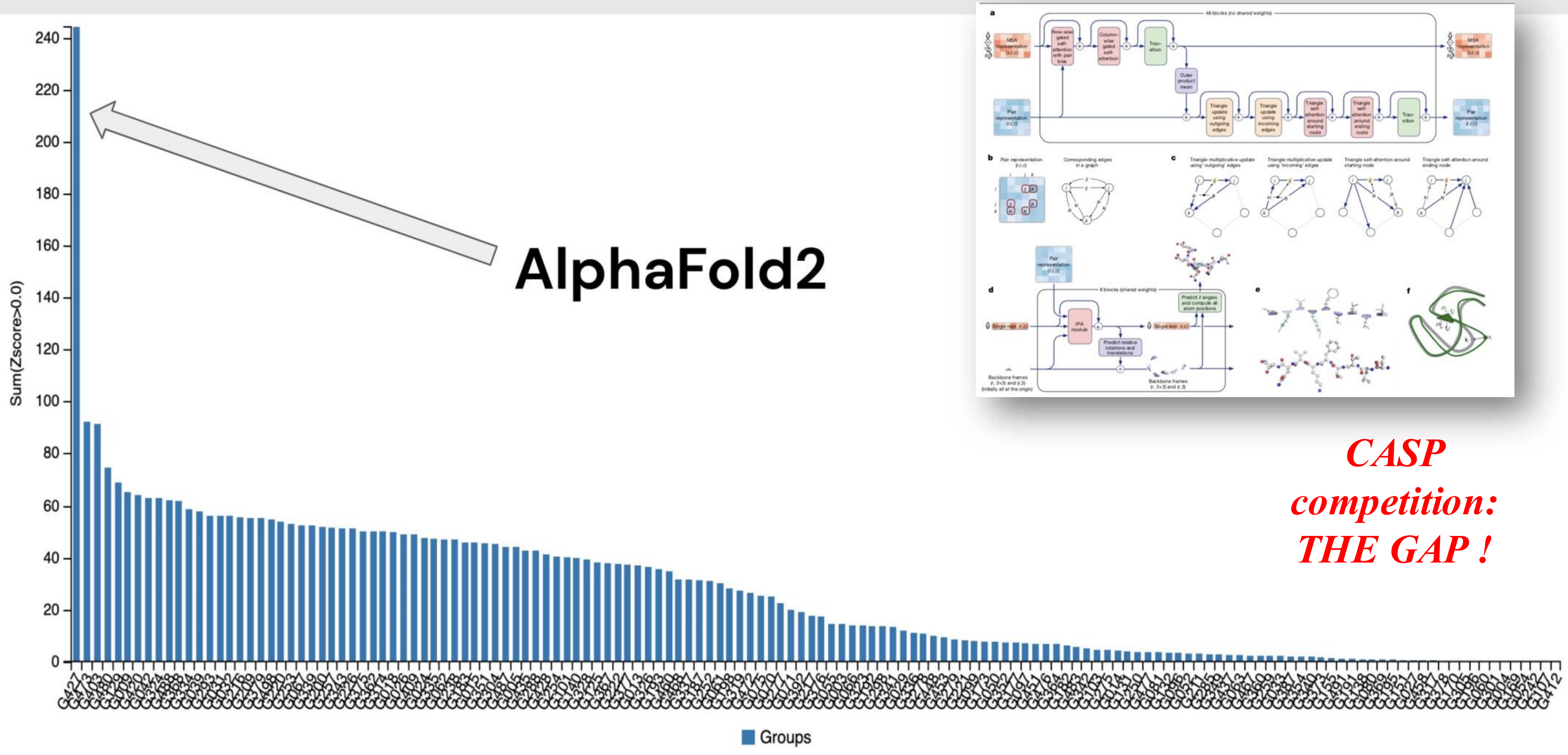
But everybody improves a little

AlphaFold2

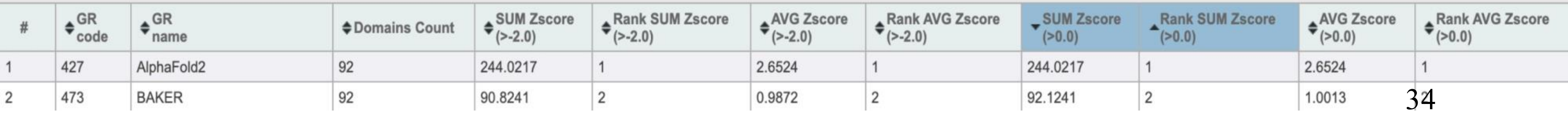


THE JUMP

AlphaFold2



CASP competition: THE GAP !



#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	427	AlphaFold2	92	244.0217	1	2.6524	1	244.0217	1	2.6524	1
2	473	BAKER	92	90.8241	2	0.9872	2	92.1241	2	1.0013	

AlphaFold2

- In all papers !! ➔ *Nature* 2021 (now > 30.000 citations)

  Breakthrough of the year *Science* 2021

  Method of the year *Nature Methods* 2021

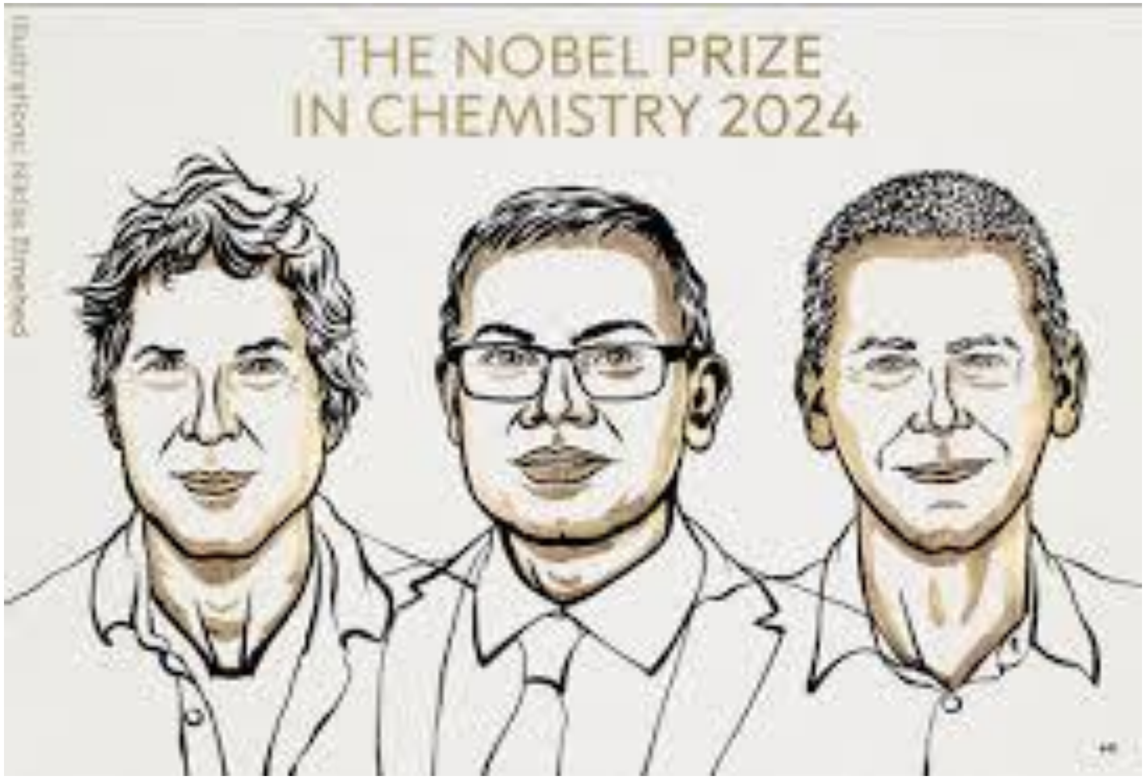
  Best invention of 2022 (*Life*)

  Prices ….

AlphaFold2

- And now, Chemistry Nobel prize 2024 (Demis Hassabis, born 1976 & John Jumper, born 1985)

DeepMind



David Baker   Demis Hassabis   John Jumper

AlphaFold2

- Deep Learning approaches

AF2 → close to LLM

Similarities to LLMs

Transformer Architecture:

AlphaFold 2's Evoformer is based on the Transformer architecture—the same core used in LLMs like GPT. It applies attention mechanisms to extract long-range dependencies, not across words but across residues and sequences in an MSA.

Sequence-based Learning:

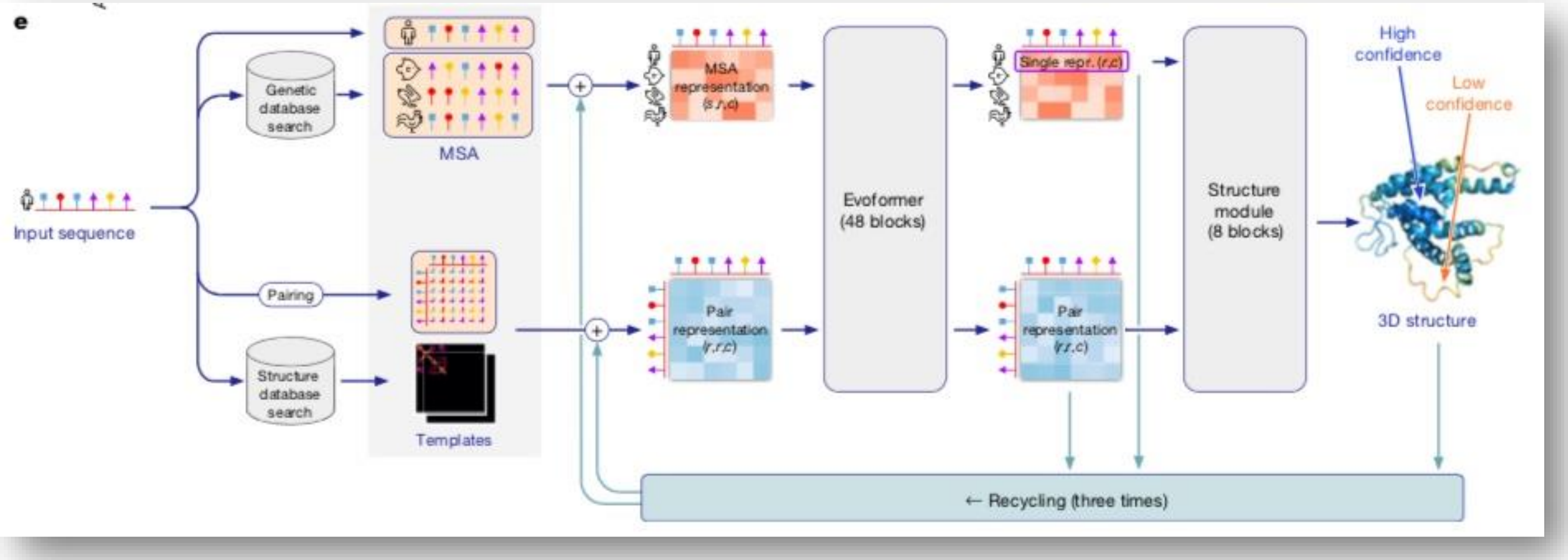
Like LLMs process text sequences, AlphaFold 2 processes biological sequences (protein sequences and their alignments). It captures contextual information about each amino acid based on its sequence and evolutionary context.

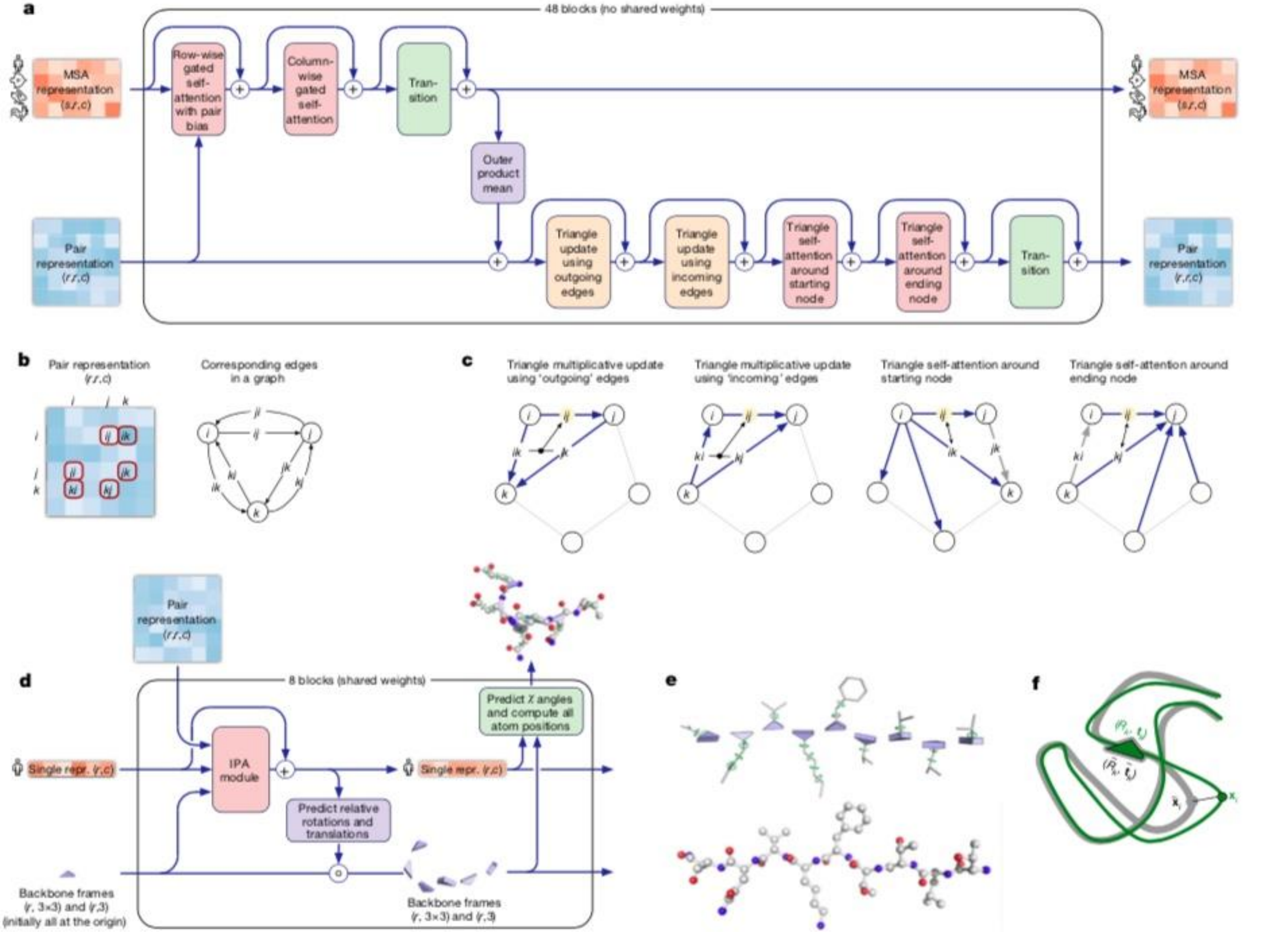
Representation Learning:

Both LLMs and AlphaFold 2 learn latent representations of input data: LLMs learn language semantics, while AlphaFold 2 learns structural constraints and relationships between residues.

AlphaFold2

- AlphaFold2 simplified architecture

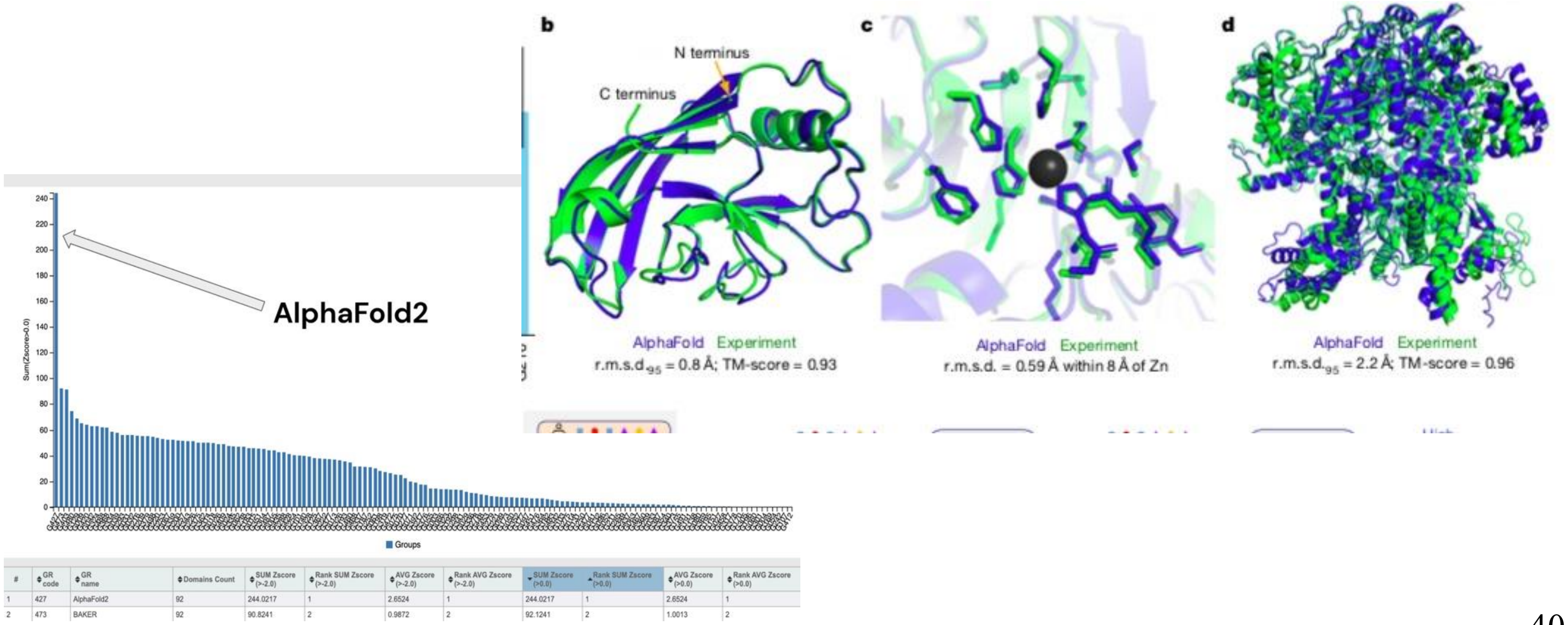

e
Input sequence
Genetic database search
Pairing
Structure database search
MSA
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)


a
48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
MSA representation (s,r,c)
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
Pair representation (r,r,c)
b
Pair representation (r,r,c)
i
j
k
ij
ik
ji
jk
ki
kj
Corresponding edges in a graph
c
Triangle multiplicative update using 'outgoing' edges
Triangle multiplicative update using 'incoming' edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Pair representation (r,r,c)
d
8 blocks (shared weights)
Single repr. (r,c)
IPA module
Predict relative rotations and translations
Single repr. (r,c)
Predict χ angles and compute all atom positions
Backbone frames (r, 3×3) and (r,3) (initially all at the origin)
Backbone frames (r, 3×3) and (r,3)
e
f
(R_k, t_k)
$(\tilde{R}_k, \tilde{t}_k)$
$\tilde{x}_i$
x_i

AlphaFold2

- Amazing results

Yes.



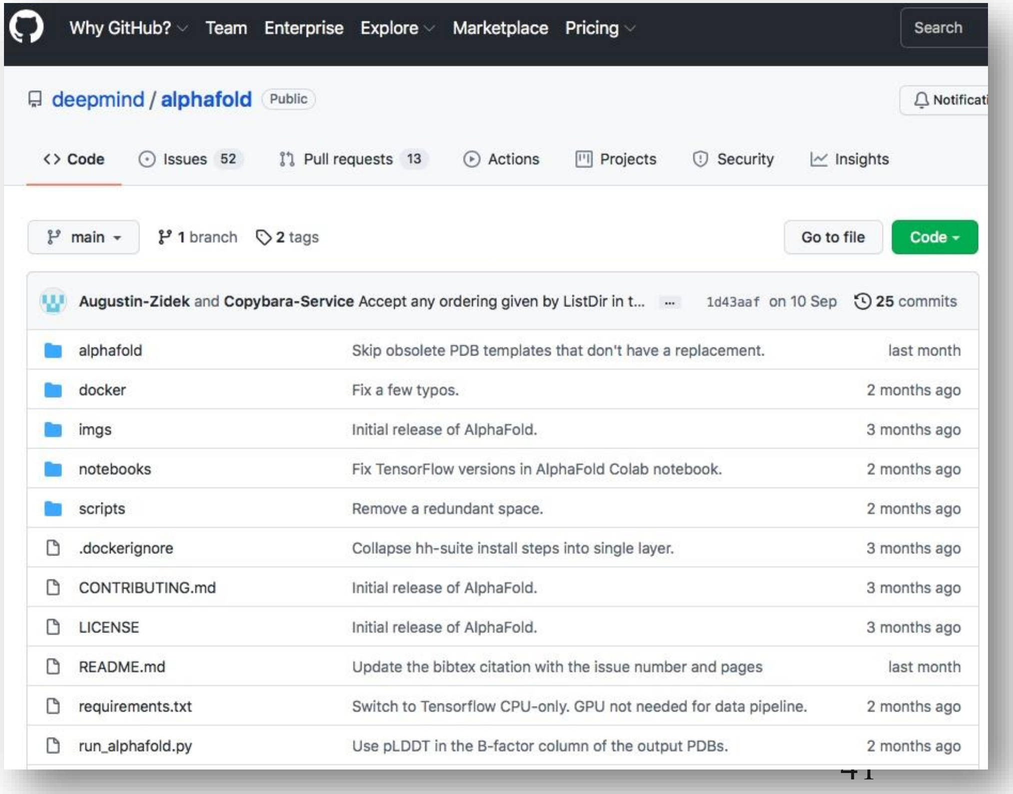
#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	427	AlphaFold2	92	244.0217	1	2.6524	1	244.0217	1	2.6524	1
2	473	BAKER	92	90.8241	2	0.9872	2	92.1241	2	1.0013	2

AlphaFold2

- You can use it at home

Algorithm is published and entirely avalaible (was not the case for v1)

https://github.com/deepmind/alphafold



Augustin-Zidek and Copybara-Service Accept any ordering given by ListDir in t... ... 1d43aaf on 10 Sep 25 commits

alphafold	Skip obsolete PDB templates that don't have a replacement.	last month
docker	Fix a few typos.	2 months ago
imgs	Initial release of AlphaFold.	3 months ago
notebooks	Fix TensorFlow versions in AlphaFold Colab notebook.	2 months ago
scripts	Remove a redundant space.	2 months ago
.dockerignore	Collapse hh-suite install steps into single layer.	3 months ago
CONTRIBUTING.md	Initial release of AlphaFold.	3 months ago
LICENSE	Initial release of AlphaFold.	3 months ago
README.md	Update the bibtex citation with the issue number and pages	last month
requirements.txt	Switch to Tensorflow CPU-only. GPU not needed for data pipeline.	2 months ago
run_alphafold.py	Use pLDDT in the B-factor column of the output PDBs.	2 months ago

AlphaFold2

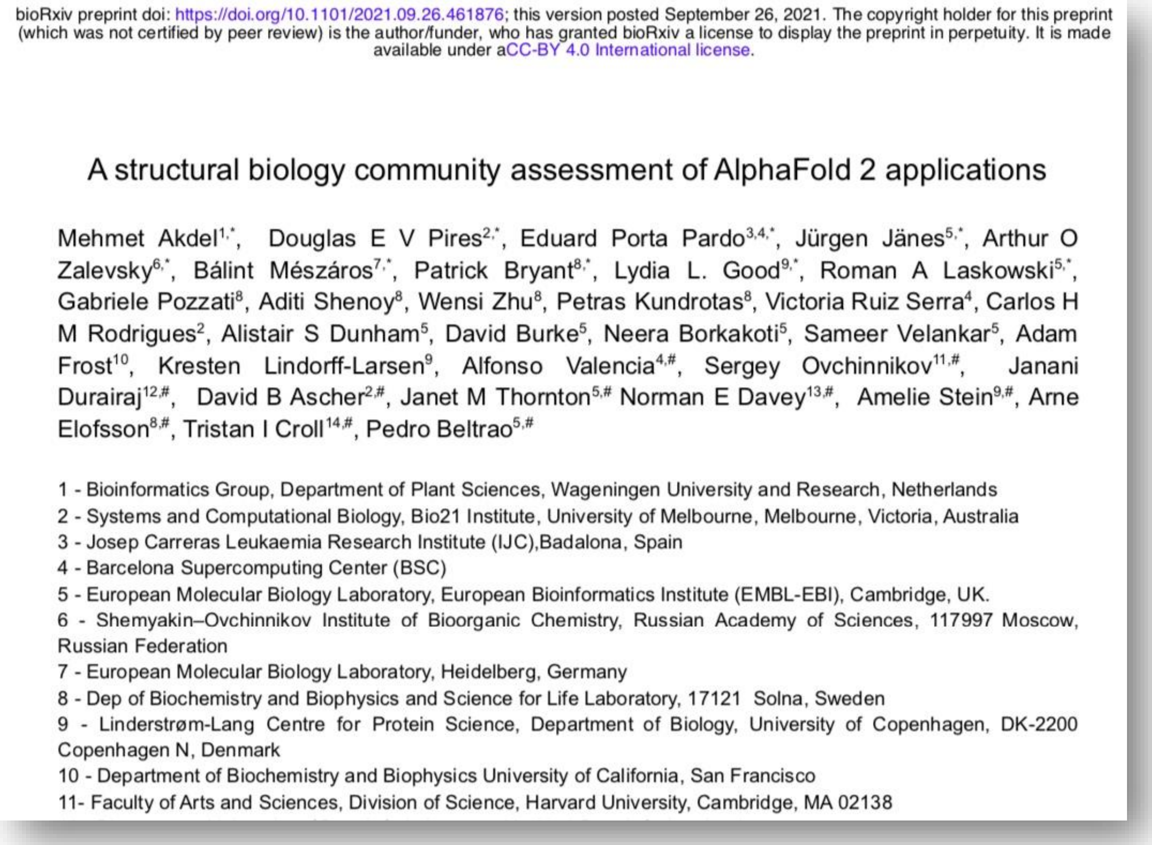
- You can use it at home

So people have used it.

Results from a big consortium

“For 11 proteomes, an average of 25% additional residues can be confidently modelled when compared to homology modelling”
➔Automatic homology modelling …

Akdel et al (2021) *bioRxiv*
=> (2022) *Nat Struct Biol*



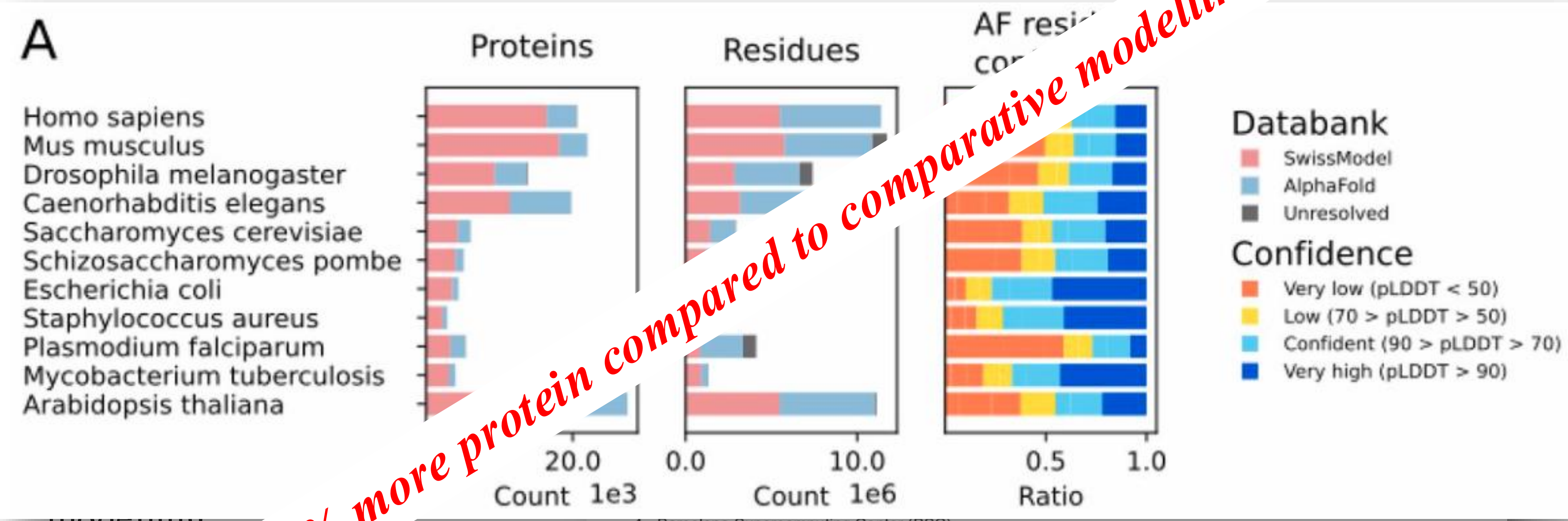
A structural biology community assessment of AlphaFold 2 applications

Mehmet Akdel[1,*], Douglas E V Pires[2,*], Eduard Porta Pardo[3,4,*], Jürgen Jänes[5,*], Arthur O Zalevsky[6,*], Bálint Mészáros[7,*], Patrick Bryant[8,*], Lydia L. Good[9,*], Roman A Laskowski[5,*], Gabriele Pozzati[8], Aditi Shenoy[8], Wensi Zhu[8], Petras Kundrotas[8], Victoria Ruiz Serra[4], Carlos H M Rodrigues[2], Alistair S Dunham[5], David Burke[5], Neera Borkakoti[5], Sameer Velankar[5], Adam Frost[10], Kresten Lindorff-Larsen[9], Alfonso Valencia[4,#], Sergey Ovchinnikov[11,#], Janani Durairaj[12,#], David B Ascher[2,#], Janet M Thornton[5,#] Norman E Davey[13,#], Amelie Stein[9,#], Arne Elofsson[8,#], Tristan I Croll[14,#], Pedro Beltrao[5,#]

1 - Bioinformatics Group, Department of Plant Sciences, Wageningen University and Research, Netherlands
2 - Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia
3 - Josep Carreras Leukaemia Research Institute (IJC),Badalona, Spain
4 - Barcelona Supercomputing Center (BSC)
5 - European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK.
6 - Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, 117997 Moscow, Russian Federation
7 - European Molecular Biology Laboratory, Heidelberg, Germany
8 - Dep of Biochemistry and Biophysics and Science for Life Laboratory, 17121 Solna, Sweden
9 - Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark
10 - Department of Biochemistry and Biophysics University of California, San Francisco
11- Faculty of Arts and Sciences, Division of Science, Harvard University, Cambridge, MA 02138

AlphaFold2

- You can use it at home



Only 10% more protein compared to comparative modelling!

modelling

➔ Autom ology
mod

4 - Barcelona Supercomputing Center (BSC)
5 - European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK.
6 - Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, 117997 Moscow, Russian Federation
7 - European Molecular Biology Laboratory, Heidelberg, Germany
8 - Dep of Biochemistry and Biophysics and Science for Life Laboratory, 17121 Solna, Sweden
9 - Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark
10 - Department of Biochemistry and Biophysics University of California, San Francisco
11- Faculty of Arts and Sciences, Division of Science, Harvard University, Cambridge, MA 02138

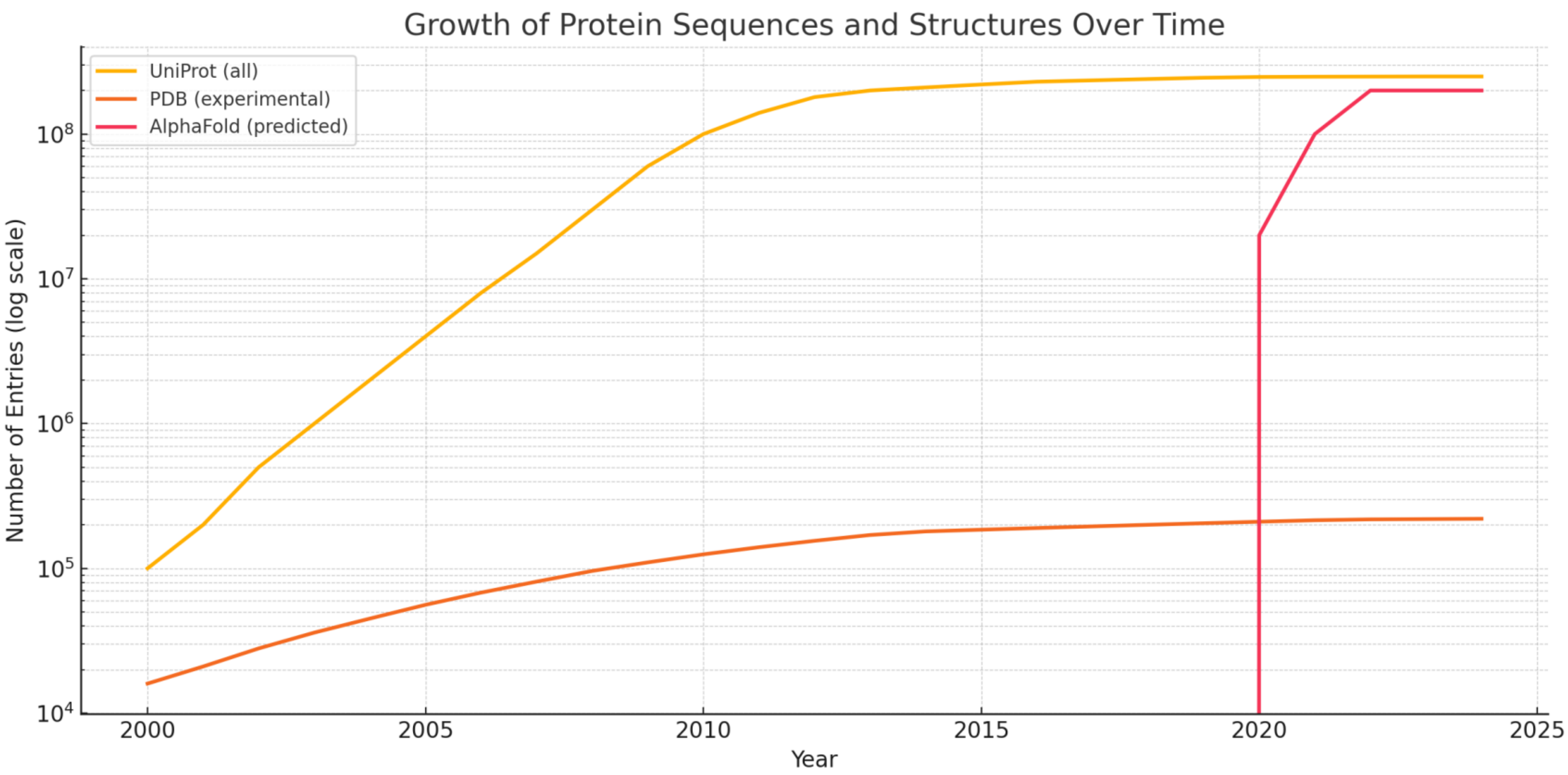
Akdel et al (2021) *bioRxiv*
=> (2022) *Nat Struct Biol*

AlphaFold2

- There is a database of already done model

EBI: h

Tunyas
596(78



AlphaFold2

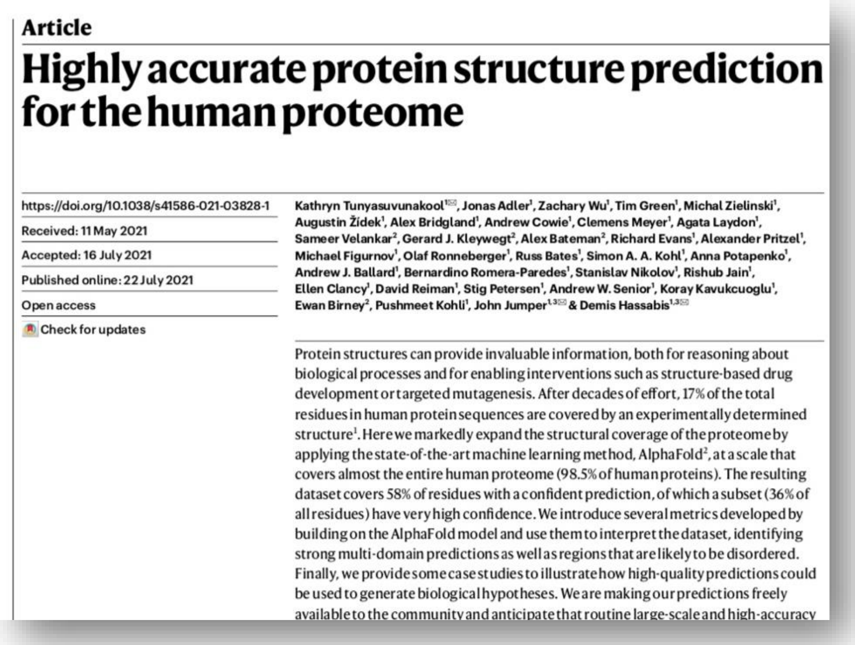
- ➢ There is a database of already done model

EBI: https://www.alphafold.ebi.ac.uk

AlphaFold2, at a scale that covers .. 98.5% of human proteins. The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence.

➔ 36% for drug design

Tunyasuvunakool K, et al (2021), *Nature*. 596(7873):590-596.



Article

Highly accurate protein structure prediction for the human proteome

https://doi.org/10.1038/s41586-021-03828-1
Received: 11 May 2021
Accepted: 16 July 2021
Published online: 22 July 2021
Open access
Check for updates

Kathryn Tunyasuvunakool[1]✉, Jonas Adler[1], Zachary Wu[1], Tim Green[1], Michal Zielinski[1], Augustin Žídek[1], Alex Bridgland[1], Andrew Cowie[1], Clemens Meyer[1], Agata Laydon[1], Sameer Velankar[2], Gerard J. Kleywegt[2], Alex Bateman[2], Richard Evans[1], Alexander Pritzel[1], Michael Figurnov[1], Olaf Ronneberger[1], Russ Bates[1], Simon A. A. Kohl[1], Anna Potapenko[1], Andrew J. Ballard[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Rishub Jain[1], Ellen Clancy[1], David Reiman[1], Stig Petersen[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Ewan Birney[2], Pushmeet Kohli[1], John Jumper[1,3]✉ & Demis Hassabis[1,3]✉

Protein structures can provide invaluable information, both for reasoning about biological processes and for enabling interventions such as structure-based drug development or targeted mutagenesis. After decades of effort, 17% of the total residues in human protein sequences are covered by an experimentally determined structure[1]. Here we markedly expand the structural coverage of the proteome by applying the state-of-the-art machine learning method, AlphaFold[2], at a scale that covers almost the entire human proteome (98.5% of human proteins). The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence. We introduce several metrics developed by building on the AlphaFold model and use them to interpret the dataset, identifying strong multi-domain predictions as well as regions that are likely to be disordered. Finally, we provide some case studies to illustrate how high-quality predictions could be used to generate biological hypotheses. We are making our predictions freely available to the community and anticipate that routine large-scale and high-accuracy

AlphaFold2

- ➢ There is a database of already done model

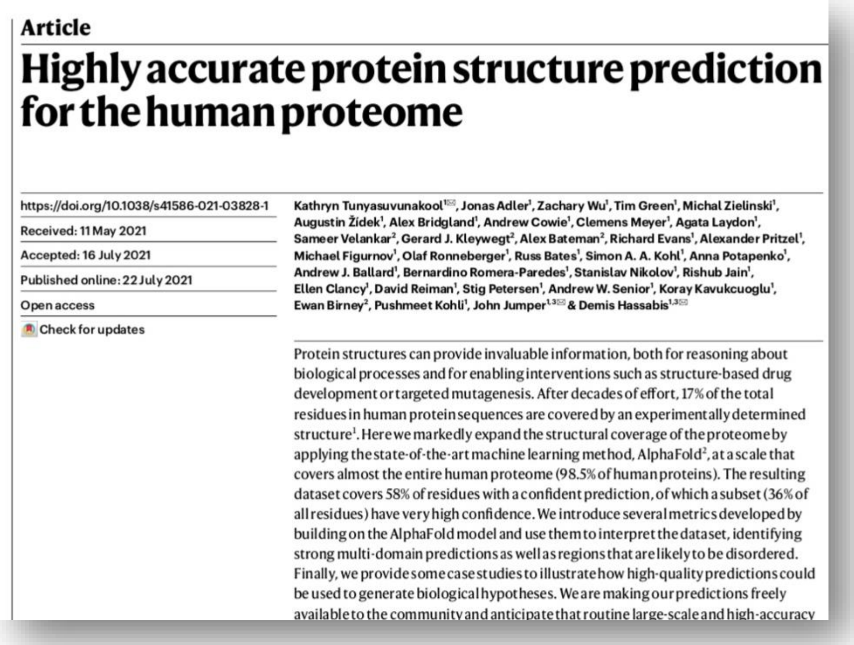
EBI: https://www.alphafold.ebi.ac.uk

AlphaFold2, at a scale that covers .. 98.5% of human proteins. The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence.

➔ 36% for drug design

➔ 42% question about fold

Tunyasuvunakool K, et al (2021), *Nature*. 596(7873):590-596.



Article

Highly accurate protein structure prediction for the human proteome

https://doi.org/10.1038/s41586-021-03828-1

Received: 11 May 2021

Accepted: 16 July 2021

Published online: 22 July 2021

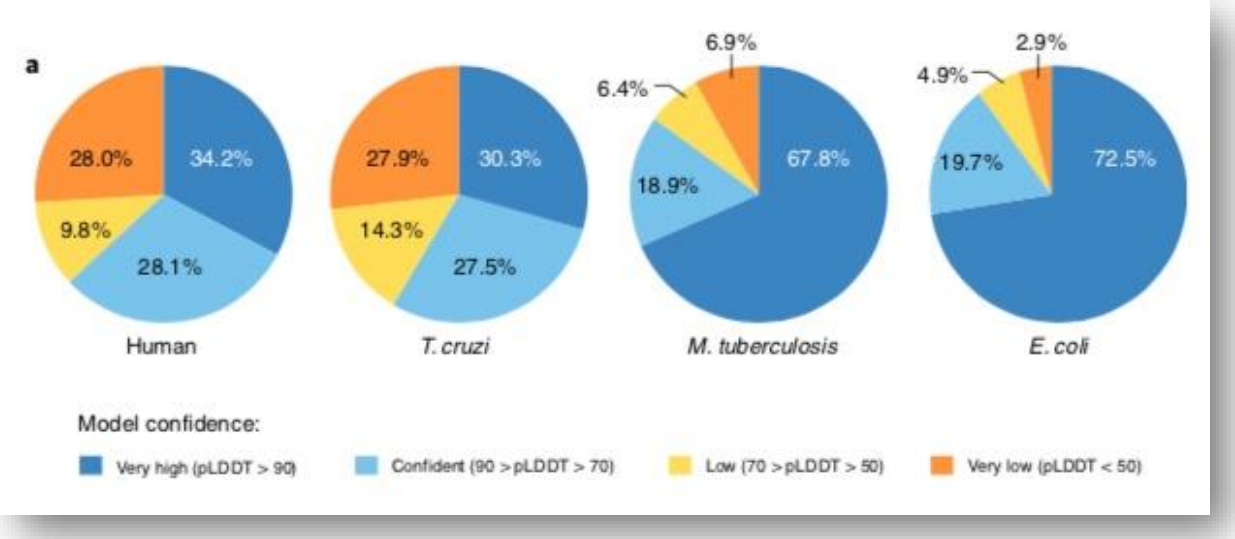
Open access

Check for updates

Kathryn Tunyasuvunakool[1✉], Jonas Adler[1], Zachary Wu[1], Tim Green[1], Michal Zielinski[1], Augustin Žídek[1], Alex Bridgland[1], Andrew Cowie[1], Clemens Meyer[1], Agata Laydon[1], Sameer Velankar[2], Gerard J. Kleywegt[2], Alex Bateman[2], Richard Evans[1], Alexander Pritzel[1], Michael Figurnov[1], Olaf Ronneberger[1], Russ Bates[1], Simon A. A. Kohl[1], Anna Potapenko[1], Andrew J. Ballard[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Rishub Jain[1], Ellen Clancy[1], David Reiman[1], Stig Petersen[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Ewan Birney[2], Pushmeet Kohli[1], John Jumper[1,3✉] & Demis Hassabis[1,3✉]

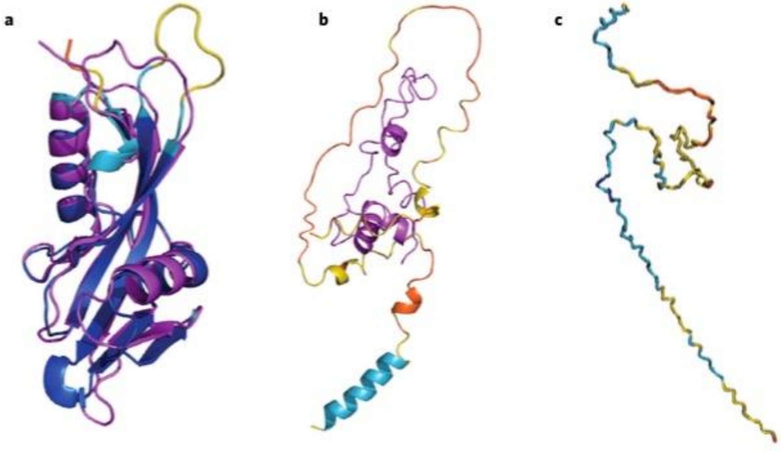
Protein structures can provide invaluable information, both for reasoning about biological processes and for enabling interventions such as structure-based drug development or targeted mutagenesis. After decades of effort, 17% of the total residues in human protein sequences are covered by an experimentally determined structure[1]. Here we markedly expand the structural coverage of the proteome by applying the state-of-the-art machine learning method, AlphaFold[2], at a scale that covers almost the entire human proteome (98.5% of human proteins). The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence. We introduce several metrics developed by building on the AlphaFold model and use them to interpret the dataset, identifying strong multi-domain predictions as well as regions that are likely to be disordered. Finally, we provide some case studies to illustrate how high-quality predictions could be used to generate biological hypotheses. We are making our predictions freely available to the community and anticipate that routine large-scale and high-accuracy

AlphaFold2



- *Protein structures predicted using artificial intelligence will aid medical research, but the greatest benefit will come if clinical data can be similarly used to better understand human disease.*

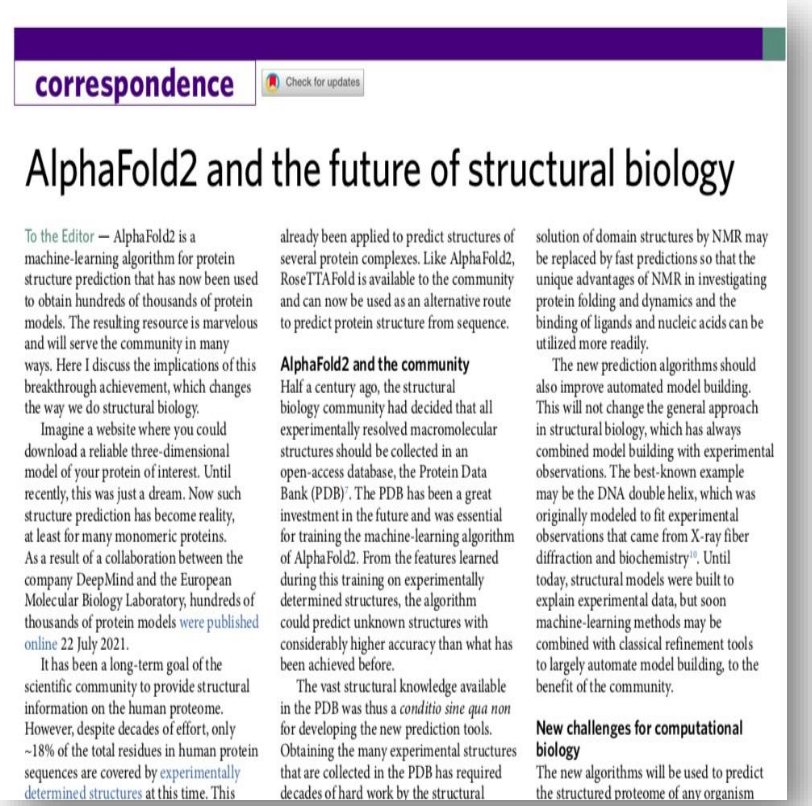
Janet M. Thornton, Roman A. Laskowski and Neera Borkakoti. (2021) *Nat Med*. 27:1666-1671.



The good, the bad and the ugly'

AlphaFold2

- The new prediction algorithms do not solve the protein folding problem in the sense that they do not reveal how a sequence encodes three-dimensional structure.
- However, they do solve the problem in practical terms, as they can reliably predict structure from sequence, ***at least in many cases.***
- ***Although only time will tell***, this advance is expected to represent a breakthrough in structural biology that is comparable to previous major advances,



correspondence

Check for updates

AlphaFold2 and the future of structural biology

To the Editor — AlphaFold2 is a machine-learning algorithm for protein structure prediction that has now been used to obtain hundreds of thousands of protein models. The resulting resource is marvelous and will serve the community in many ways. Here I discuss the implications of this breakthrough achievement, which changes the way we do structural biology.

Imagine a website where you could download a reliable three-dimensional model of your protein of interest. Until recently, this was just a dream. Now such structure prediction has become reality, at least for many monomeric proteins. As a result of a collaboration between the company DeepMind and the European Molecular Biology Laboratory, hundreds of thousands of protein models were published online 22 July 2021.

It has been a long-term goal of the scientific community to provide structural information on the human proteome. However, despite decades of effort, only ~18% of the total residues in human protein sequences are covered by experimentally determined structures at this time. This

already been applied to predict structures of several protein complexes. Like AlphaFold2, RoseTTAFold is available to the community and can now be used as an alternative route to predict protein structure from sequence.

AlphaFold2 and the community

Half a century ago, the structural biology community had decided that all experimentally resolved macromolecular structures should be collected in an open-access database, the Protein Data Bank (PDB)[7]. The PDB has been a great investment in the future and was essential for training the machine-learning algorithm of AlphaFold2. From the features learned during this training on experimentally determined structures, the algorithm could predict unknown structures with considerably higher accuracy than what has been achieved before.

The vast structural knowledge available in the PDB was thus a *conditio sine qua non* for developing the new prediction tools. Obtaining the many experimental structures that are collected in the PDB has required decades of hard work by the structural

solution of domain structures by NMR may be replaced by fast predictions so that the unique advantages of NMR in investigating protein folding and dynamics and the binding of ligands and nucleic acids can be utilized more readily.

The new prediction algorithms should also improve automated model building. This will not change the general approach in structural biology, which has always combined model building with experimental observations. The best-known example may be the DNA double helix, which was originally modeled to fit experimental observations that came from X-ray fiber diffraction and biochemistry[10]. Until today, structural models were built to explain experimental data, but soon machine-learning methods may be combined with classical refinement tools to largely automate model building, to the benefit of the community.

New challenges for computational biology

The new algorithms will be used to predict the structured proteome of any organism

Cramer P. (2021) *Nat Struct Mol Biol*. 28(9):704-705.

AlphaFold2

Biochimie 207 (2023) 11–19

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

ELSEVIER

An agnostic analysis of the human AlphaFold2 proteome using local protein conformations

Alexandre G. de Brevern*

Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM UMR_S 1134, BIGR, DSIMB Bioinformatics team, F-75014, Paris, France

ARTICLE INFO

Article history:
Received 26 April 2022
Received in revised form
14 October 2022
Accepted 17 November 2022
Available online 20 November 2022

Keywords:
Secondary structure
Helix
Sheet: turn
polyproline II
Structural alphabet: protein structure
Deep learning

ABSTRACT

Knowledge of the 3D structure of proteins is a valuable asset for understanding their precise biological mechanisms. However, the cost of production of 3D structures and experimental difficulties limit their obtaining. The proposal of 3D structural models is consequently an appealing alternative. The release of the AlphaFold Deep Learning approach has revolutionized the field. The recent near-complete human proteome proposal makes it possible to analyse large amounts of data and evaluate the results of the approach in greater depth. The 3D human proteome was thus analysed in light of the classic secondary structures, and many less-used protein local conformations (PolyProline II helices, type of γ-turns, of β-turns and of β-bulges, curvature of the helices, and a structural alphabet). Without questioning the global quality of the approach, this analysis highlights certain local conformations, which maybe poorly predicted and they could therefore be better addressed.

1. Introduction

Proteins are essential constituents of cells, one of the major macromolecules of life. Composed of 20 amino acids, proteins ensure a variety of essential biological functions. The dogma of molecular biology emphases the link between the nucleic sequence and the protein sequence to arrive at protein structures and their functions. Nowadays, access to these sequences has become particularly cheap [1]. Databases contain millions of protein sequences, while the three-dimensional structures of proteins are much more difficult to obtain experimentally [2].

Hence for more than 30 years, different computational approaches have been implemented to propose three-dimensional (3D) structural models of proteins from their amino acid sequence [3]. The classic categories of these approaches include homology or comparative modelling, threading, *ab initio*, *de novo* approaches, and meta-servers; these last combined several approaches. It is possible to notice the best-known tools such as Modeller [4], the golden standard of homology/comparative modelling and *de novo* most recent approaches Rosetta and I-TASSER [5–7]. These two last have won numerous Critical Assessment for Protein Structure Prediction (CASP) meetings [8,9].

Arrived at the CASP13 (2018), the company DeepMind presented its new Deep Learning approach named AlphaFold [10]. It won the Free Modelling category, i.e. the prediction of novel protein folds [11], whereas template-Based Modelling category, i.e. protein folds already found in the Protein Data Bank, was won by Zhang's group [12]. Two years later, AlphaFold version 2 obtained particularly remarkable at CASP14 (2020) [13,14]; some models were within the uncertainties of the experimental resolution, an impressive result. This improvement combined the delicate use of evolution, contacts within proteins, and large GPU computing power that allowed the implementation of a particularly complex and elegant architecture [15,16].

AlphaFold 2 was a hot topic for 2020 and 2021 [17–20], leading to a revolution in protein structural model building [21,22] and opening potential new opportunities, e.g. new drug design researches [23,24]. Three points can be noticed (i) the code can be downloaded freely on GitHub (https://github.com/deepmind/alphafold) and is really useable [25], (ii) different online notebooks for non-specialists are easy to use(e.g. https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/batch/AlphaFold2_batch.ipynb) [26], and (iii) EBI provides structural model databases [27]. Indeed, model building is expensive in

* INSERM UMR_S 1134, DSIMB Bioinformatics team, Paris Cité, 8, rue Maria Helena Vieira da Silva, 75014, Paris, France.
E-mail address: alexandre.debrevern@u-paris.fr.

https://doi.org/10.1016/j.biochi.2022.11.009
0300-9084/© 2022 Elsevier B.V. and Société Française de Biochimie et Biologie Moléculaire (SFBBM). All rights reserved.

DSIMB
DYNAMICS OF STRUCTURES AND INTERACTIONS OF BIOLOGICAL MACROMOLECULES

Not all local conformations are properly predicted !

PPIIs are not good

γ-turns are not good

Cis-ω are not good

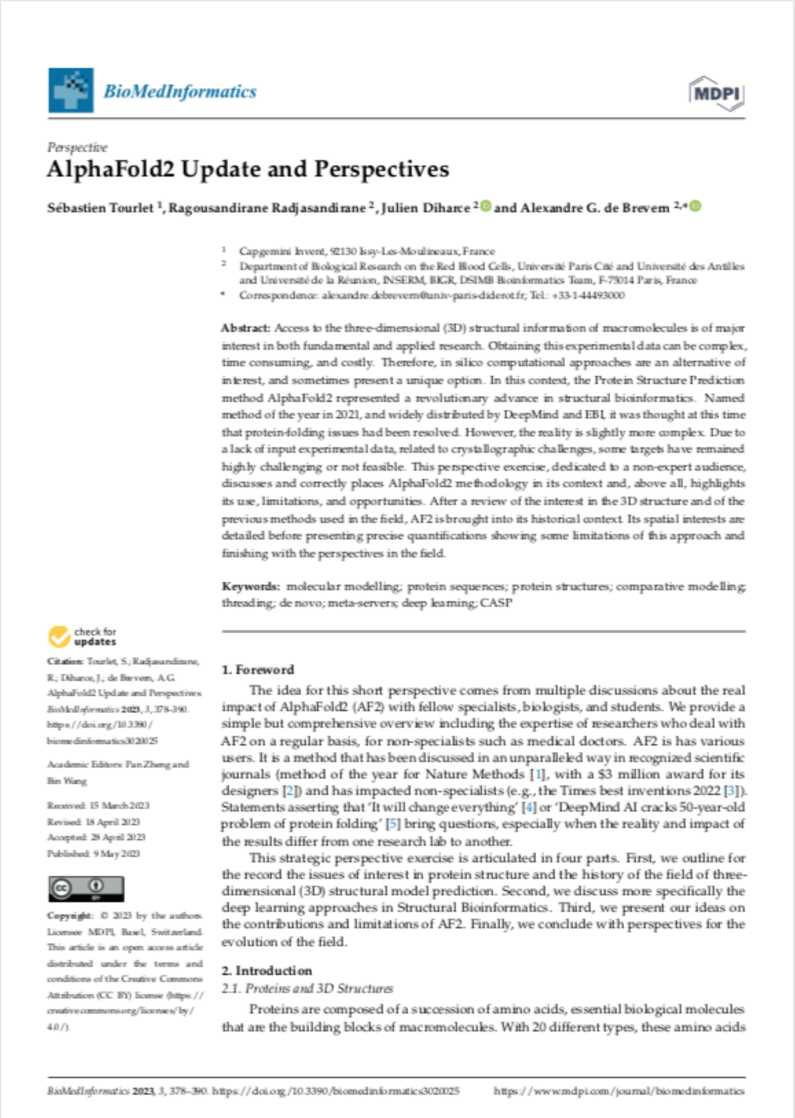
β-sheets are in limited number …

de Brevern A.G. An agnostic analysis of the human AlphaFold2 proteome using local protein conformations. *Biochimie* (2023) **207**:11-19.

Inserm
Institut national
de la santé et de la recherche médicale

AlphaFold2

SiMB
Bioinformatique & Interactions



BioMedInformatics MDPI

Perspective

AlphaFold2 Update and Perspectives

Sébastien Tourlet [1], Ragousandirane Radjasandirane [2], Julien Diharce [2] and Alexandre G. de Brevern [2,*]

[1] Capgemini Invent, 92130 Issy-Les-Moulineaux, France
[2] Department of Biological Research on the Red Blood Cells, Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM, BIGR, DSIMB Bioinformatics Team, F-75014 Paris, France
* Correspondence: alexandre.debrevern@univ-paris-diderot.fr; Tel.: +33-1-44493000

Abstract: Access to the three-dimensional (3D) structural information of macromolecules is of major interest in both fundamental and applied research. Obtaining this experimental data can be complex, time consuming, and costly. Therefore, in silico computational approaches are an alternative of interest, and sometimes present a unique option. In this context, the Protein Structure Prediction method AlphaFold2 represented a revolutionary advance in structural bioinformatics. Named method of the year in 2021, and widely distributed by DeepMind and EBI, it was thought at this time that protein-folding issues had been resolved. However, the reality is slightly more complex. Due to a lack of input experimental data, related to crystallographic challenges, some targets have remained highly challenging or not feasible. This perspective exercise, dedicated to a non-expert audience, discusses and correctly places AlphaFold2 methodology in its context and, above all, highlights its use, limitations, and opportunities. After a review of the interest in the 3D structure and of the previous methods used in the field, AF2 is brought into its historical context. Its spatial interests are detailed before presenting precise quantifications showing some limitations of this approach and finishing with the perspectives in the field.

Keywords: molecular modelling; protein sequences; protein structures; comparative modelling; threading; de novo; meta-servers; deep learning; CASP

check for updates

Citation: Tourlet, S.; Radjasandirane, R.; Diharce, J.; de Brevern, A.G. AlphaFold2 Update and Perspectives. *BioMedInformatics* **2023**, 3, 378–390. https://doi.org/10.3390/biomedinformatics3020025

Academic Editors: Pan Zheng and Bin Wang

Received: 15 March 2023
Revised: 18 April 2023
Accepted: 28 April 2023
Published: 9 May 2023

1. Foreword

The idea for this short perspective comes from multiple discussions about the real impact of AlphaFold2 (AF2) with fellow specialists, biologists, and students. We provide a simple but comprehensive overview including the expertise of researchers who deal with AF2 on a regular basis, for non-specialists such as medical doctors. AF2 is has various users. It is a method that has been discussed in an unparalleled way in recognized scientific journals (method of the year for Nature Methods [1], with a $3 million award for its designers [2]) and has impacted non-specialists (e.g., the Times best inventions 2022 [3]). Statements asserting that 'It will change everything' [4] or 'DeepMind AI cracks 50-year-old problem of protein folding' [5] bring questions, especially when the reality and impact of the results differ from one research lab to another.

This strategic perspective exercise is articulated in four parts. First, we outline for the record the issues of interest in protein structure and the history of the field of three-dimensional (3D) structural model prediction. Second, we discuss more specifically the deep learning approaches in Structural Bioinformatics. Third, we present our ideas on the contributions and limitations of AF2. Finally, we conclude with perspectives for the evolution of the field.

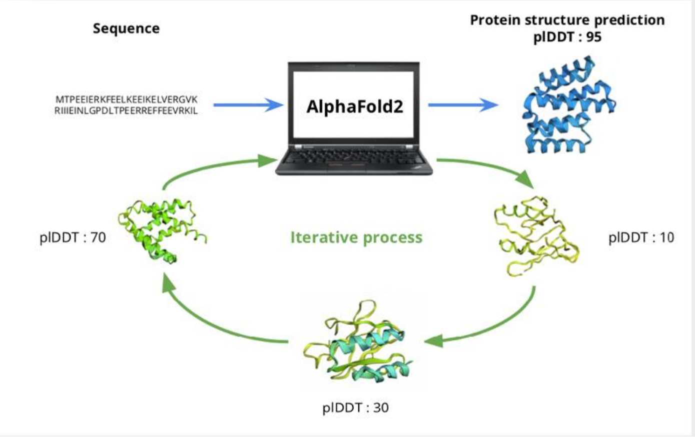
2. Introduction

2.1. Proteins and 3D Structures

Proteins are composed of a succession of amino acids, essential biological molecules that are the building blocks of macromolecules. With 20 different types, these amino acids

DSIMB
DYNAMICS OF STRUCTURES AND INTERACTIONS OF BIOLOGICAL MACROMOLECULES

Analyses of the impact of AlphaFold2 on the daily life of a Structural Bioinformatics lab.


Sequence
MTPEEIERKFEELKEEIKELVERGVK
RIIIEINLGPDLTPEERREFFEEVRKIL
AlphaFold2
Protein structure prediction
pIDDT : 95
pIDDT : 70
Iterative process
pIDDT : 10
pIDDT : 30

Tourlet S., Radjasandirane R., Diharce J., de Brevern A.G. AlphaFold2 Update and Perspectives. *BioMedInformatics* (2023) **3**(2), 378-390.

AlphaFold2

What I was doing before AlphaFold2

(a)

- **Protocol:**

  protein properties (S2, disorder, PTMs,...)
  PSI-BLAST, HMM, … searching in databases
  Looking for evolution
  Comparative modelling if possible (Modeller)
  Tools and webservers:
  comparative, e.g. SwissModel,
  threading, e.g. Phyre
  de novo, e.g. I-Tasser, Rosetta

- Analyses

Tourlet S., Radjasandirane R., Diharce J., de Brevern A.G. AlphaFold2 Update and Perspectives. *BioMedInformatics* (2023) **3**(2), 378-390.

AlphaFold2

What I was doing before AlphaFold2

(a)

- **Protocol:**

  protein properties (S2, disorder, PTMs,...)
  PSI-BLAST, HMM, … searching in databases
  Looking for evolution
  Comparative modelling if possible (Modeller)
  Tools and webservers:
  comparative, e.g. SwissModel,
  threading, e.g. Phyre
  de novo, e.g. I-Tasser, Rosetta

- Analyses

What I am doing now

(b)

- **Protocol:**

  protein properties (S2, disorder, PTMs,...)
  PSI-BLAST, HMM, … searching in databases
  Looking for evolution
  Comparative modelling if possible (Modeller)
  Tools and webservers:
  comparative, e.g. SwissModel,
  threading, e.g. Phyre
  de novo, e.g. I-Tasser, Rosetta
  Deep learning, e.g. AlphaFold2

- Analyses

Tourlet S., Radjasandirane R., Diharce J., de Brevern A.G. AlphaFold2 Update and Perspectives. *BioMedInformatics* (2023) **3**(2), 378-390.

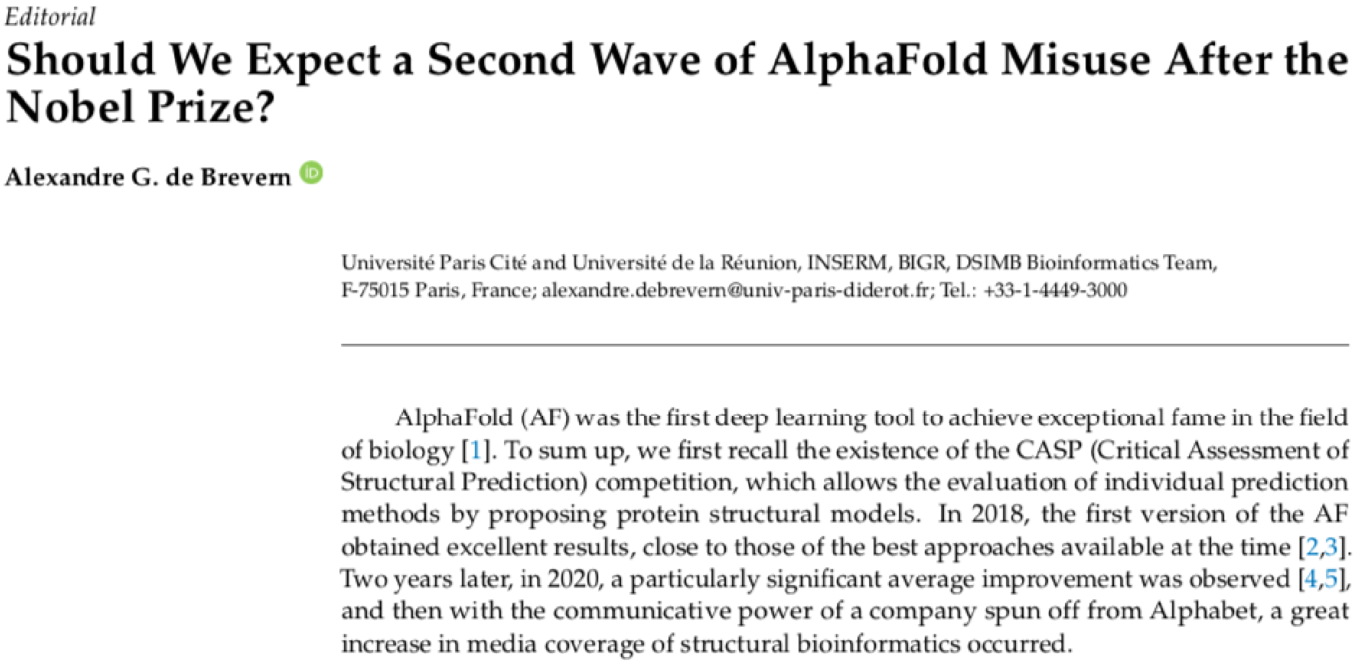
Inserm
Institut national
de la santé et de la recherche médicale

AlphaFold2

DSIMB Bioinformatique & Interactions

- *Editorial :* **Should We Expect a Second Wave of AlphaFold Misuse After the Nobel Prize?**

DSIMB
DYNAMICS OF STRUCTURES AND INTERACTIONS OF BIOLOGICAL MACROMOLECULES



Editorial

Should We Expect a Second Wave of AlphaFold Misuse After the Nobel Prize?

Alexandre G. de Brevern

Université Paris Cité and Université de la Réunion, INSERM, BIGR, DSIMB Bioinformatics Team, F-75015 Paris, France; alexandre.debrevern@univ-paris-diderot.fr; Tel.: +33-1-4449-3000

AlphaFold (AF) was the first deep learning tool to achieve exceptional fame in the field of biology [1]. To sum up, we first recall the existence of the CASP (Critical Assessment of Structural Prediction) competition, which allows the evaluation of individual prediction methods by proposing protein structural models. In 2018, the first version of the AF obtained excellent results, close to those of the best approaches available at the time [2,3]. Two years later, in 2020, a particularly significant average improvement was observed [4,5], and then with the communicative power of a company spun off from Alphabet, a great increase in media coverage of structural bioinformatics occurred.

AlphaFold2

- Yes, AlphaFold is a revolution, because now Deep Learning (Artificial Intelligence) is everywhere in Biology

- But it is methodological revolution, for Structural Bioinformatics it is only an evolution

- It only provides 10% more proteins and 25 residues per protein on average in regards to comparative modeling

- Not so simple to be highly sure

AlphaFold2

- An example:

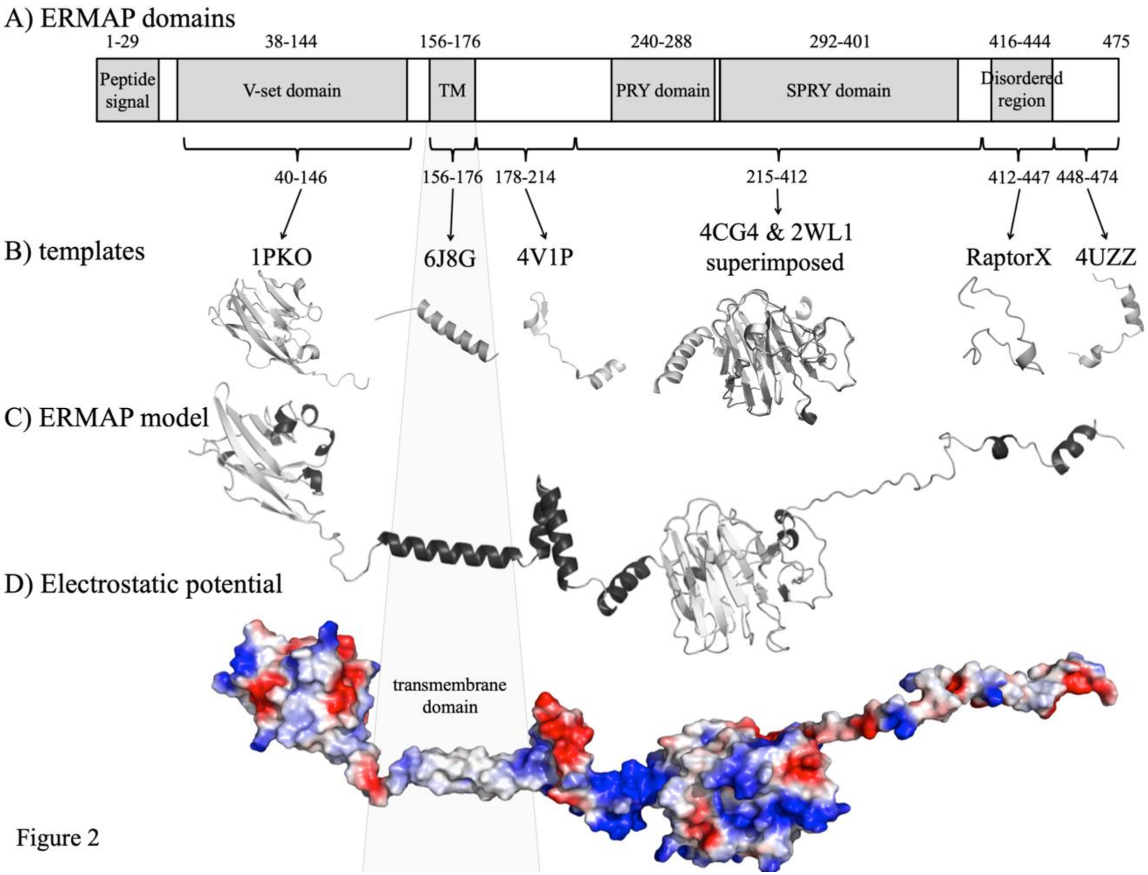


Figure 2

Floch A, Lomas-Francis C, Vege S, Burgos A, Hoffman R, Cusick R, de Brevern AG, Westhoff CM. *Transfusion*. 2023, **63**(1):230-238.

AlphaFold2

- An example:

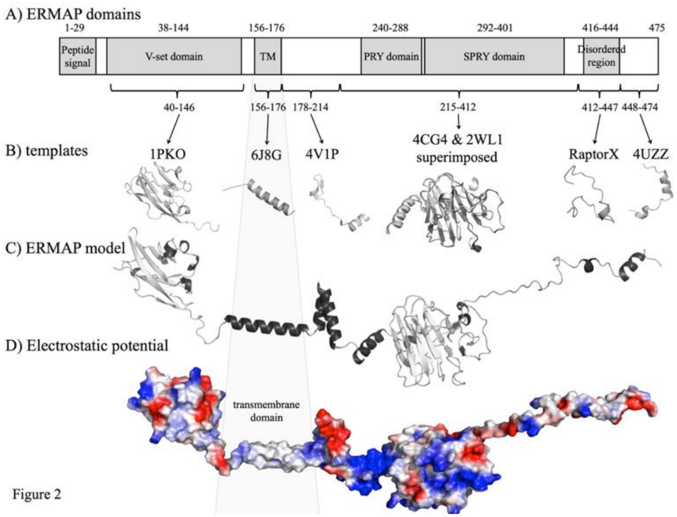
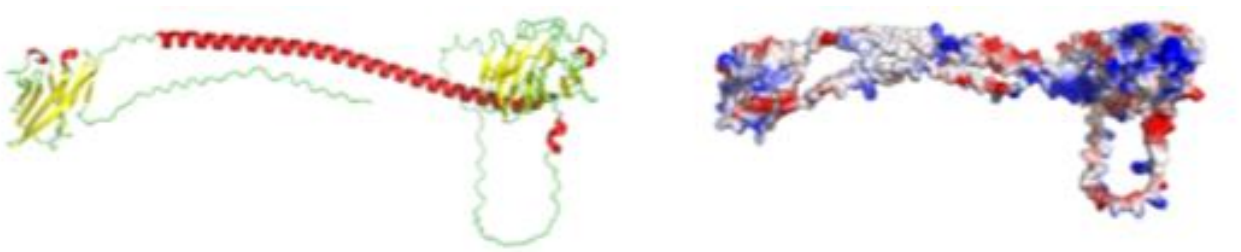


Figure 2

AF2:
wrong model



AlphaFold2

- An example:

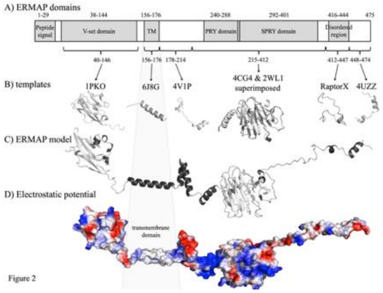


Figure 2

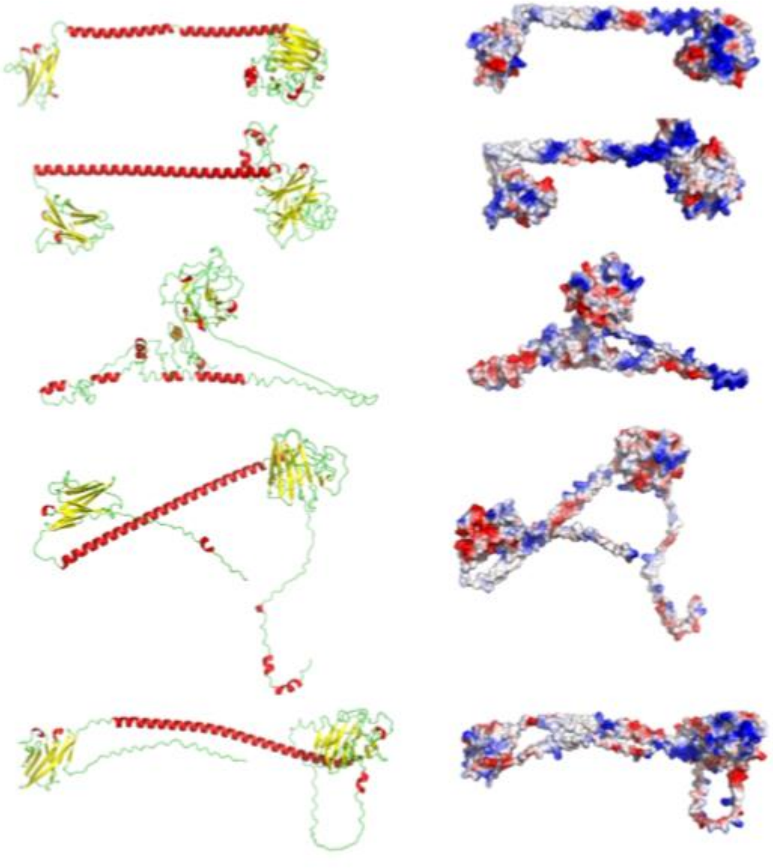
Raptor X

trRosetta

I-Tasser

RoseTTAFold

AF2

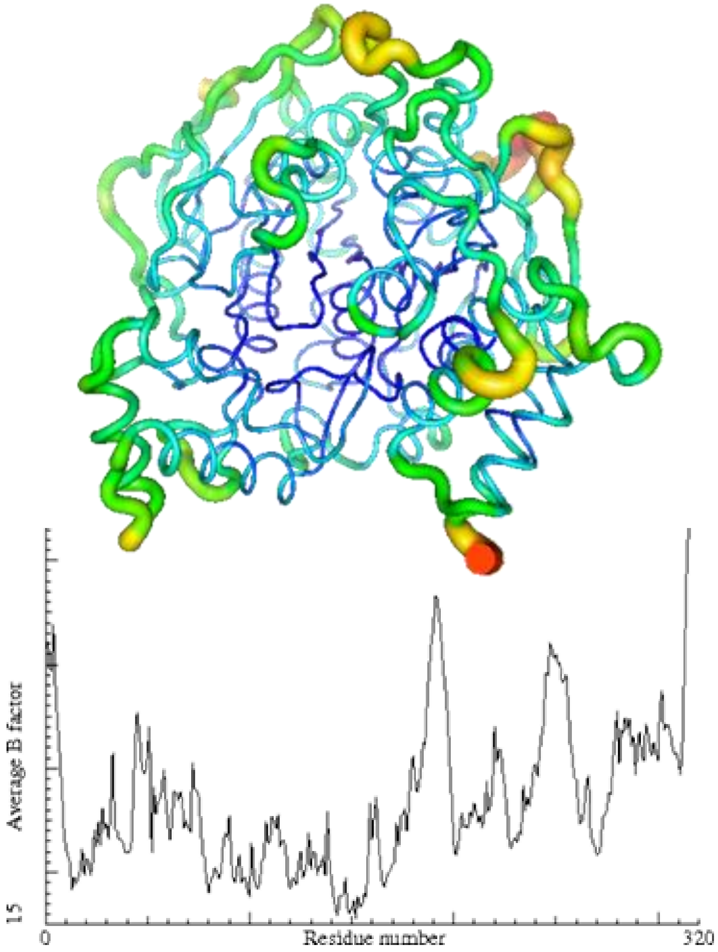


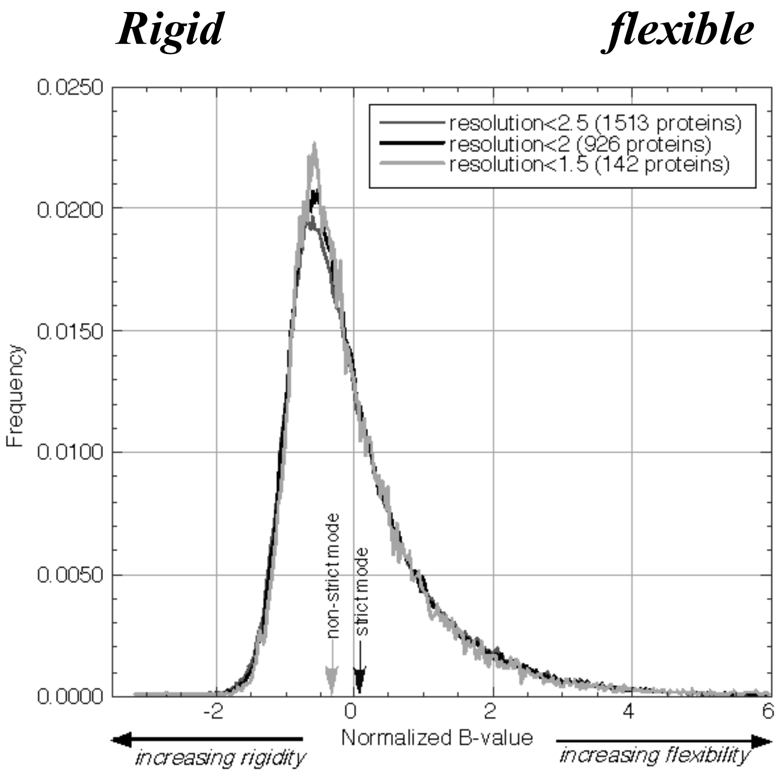
When it is complicated, it is complicated

3. PROTEIN FLEXIBILITY PREDICTION

Protein flexibility prediction

- Flexibility: Distribution of B-factors

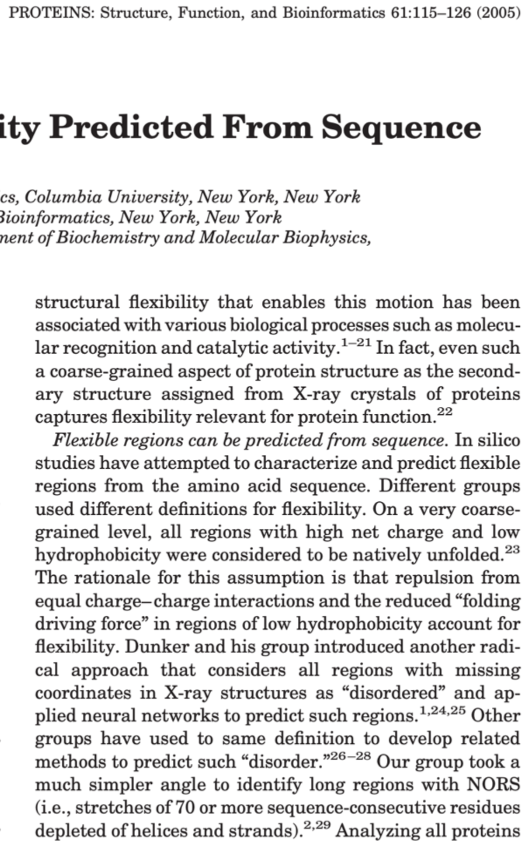




Protein flexibility prediction

2005

- Schlessinger & Rost



PROTEINS: Structure, Function, and Bioinformatics 61:115–126 (2005)

Protein Flexibility and Rigidity Predicted From Sequence

Avner Schlessinger[1,2] and Burkhard Rost[1–3*]
[1]*CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*
[2]*Columbia University Center for Computational Biology and Bioinformatics, New York, New York*
[3]*Northeast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*

ABSTRACT **Structural flexibility has been associated with various biological processes such as molecular recognition and catalytic activity. In silico studies of protein flexibility have attempted to characterize and predict flexible regions based on simple principles. B-values derived from experimental data are widely used to measure residue flexibility. Here, we present the most comprehensive large-scale analysis of B-values. We used this analysis to develop a neural network–based method that predicts flexible–rigid residues from amino acid sequence. The system uses both global and local information (i.e., features from the entire protein such as secondary structure composition, protein length, and fraction of surface residues, and features from a local window of sequence-consecutive residues). The most important local feature was the evolutionary exchange profile reflecting sequence conservation in a family of related proteins. To illustrate its potential, we applied our method to 4 different case studies, each of which related our predictions to aspects of function. The first 2 were the prediction of regions that undergo conformational switches upon environmental changes (switch II region in Ras) and the prediction of surface regions, the rigidity of**

structural flexibility that enables this motion has been associated with various biological processes such as molecular recognition and catalytic activity.[1–21] In fact, even such a coarse-grained aspect of protein structure as the secondary structure assigned from X-ray crystals of proteins captures flexibility relevant for protein function.[22]

Flexible regions can be predicted from sequence. In silico studies have attempted to characterize and predict flexible regions from the amino acid sequence. Different groups used different definitions for flexibility. On a very coarse-grained level, all regions with high net charge and low hydrophobicity were considered to be natively unfolded.[23] The rationale for this assumption is that repulsion from equal charge–charge interactions and the reduced "folding driving force" in regions of low hydrophobicity account for flexibility. Dunker and his group introduced another radical approach that considers all regions with missing coordinates in X-ray structures as "disordered" and applied neural networks to predict such regions.[1,24,25] Other groups have used to same definition to develop related methods to predict such "disorder."[26–28] Our group took a much simpler angle to identify long regions with NORS (i.e., stretches of 70 or more sequence-consecutive residues depleted of helices and strands).[2,29] Analyzing all proteins

Protein flexibility prediction

2005

- Schlessinger & Rost: rigid or flexible

A. SCHLESSINGER AND B. ROST

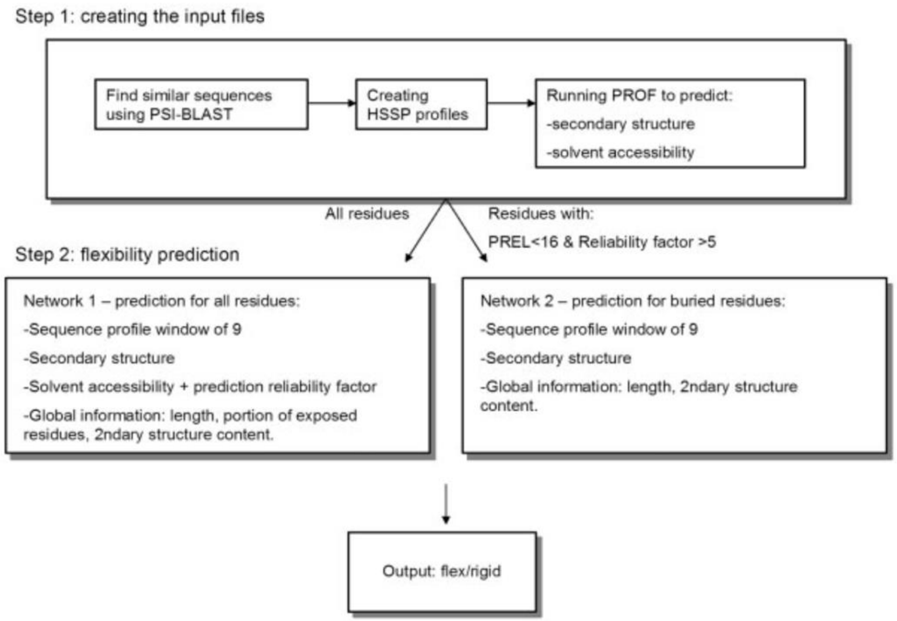

Step 1: creating the input files
Find similar sequences using PSI-BLAST
Creating HSSP profiles
Running PROF to predict:
-secondary structure
-solvent accessibility
All residues
Residues with:
PREL<16 & Reliability factor >5
Step 2: flexibility prediction
Network 1 – prediction for all residues:
-Sequence profile window of 9
-Secondary structure
-Solvent accessibility + prediction reliability factor
-Global information: length, portion of exposed residues, 2ndary structure content.
Network 2 – prediction for buried residues:
-Sequence profile window of 9
-Secondary structure
-Global information: length, 2ndary structure content.
Output: flex/rigid

Fig. 2. Prediction system. Step 1: Compile information used for neural network input. HSSP profiles were created using PSI-BLAST; these profiles are used to predict 1D structure (secondary structure and solvent accessibility) by PROFphd. Step 2: System of neural networks. Network 1 was trained on all residues with all input features, while network 2 was trained exclusively on reliably predicted buried residues. Residues that PROFacc predicted as buried with high reliability were passed to network 2; all others to network 1.

Protein flexibility prediction

2009

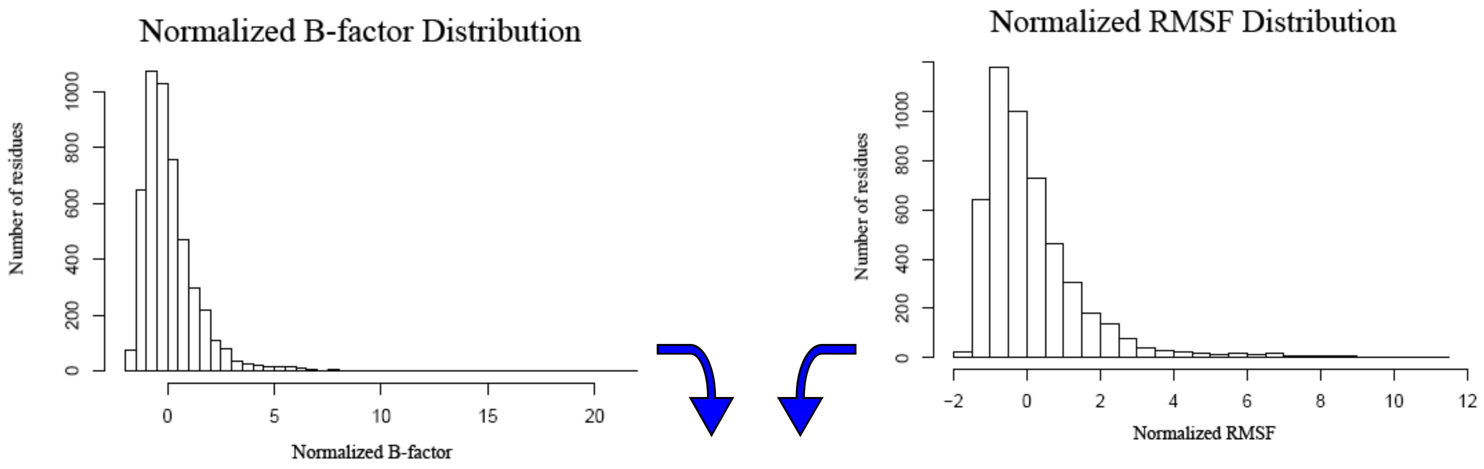
- PredyFlexy

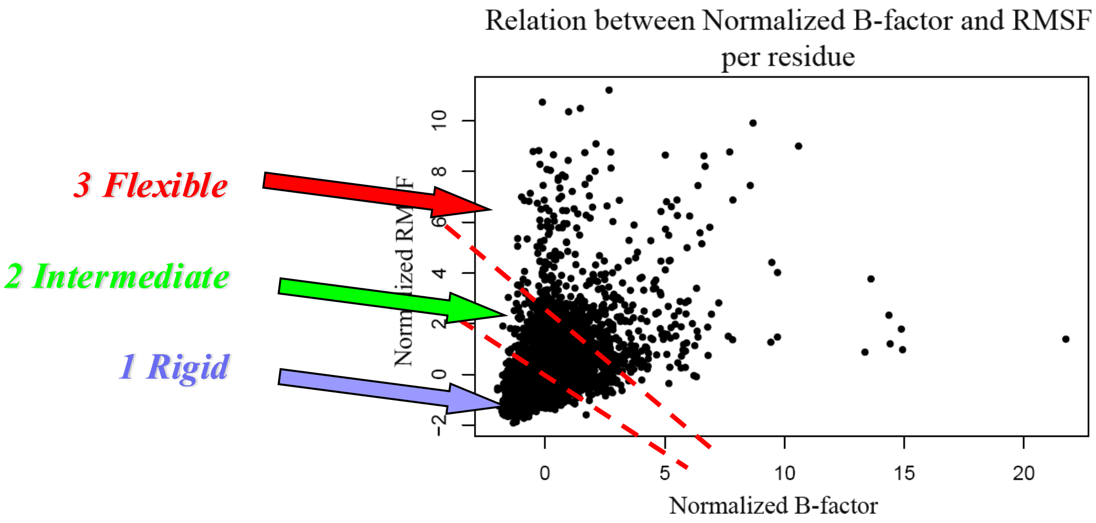
- Rigid / intermediate / flexible (from 2- to 3-states)

- Using experimental data (B-factors) and results from Molecular Dynamics (RMSF*)

**Root Mean Square Fluctuations*

Protein flexibility prediction





3 Flexible

2 Intermediate

1 Rigid

- *Experimental and simulation uncertainties*
- ***3 Flexibility classes***

Protein flexibility prediction

- PredyFlexy

A prediction of local protein conformations made with Support Vector Machines.

Protein flexibility prediction

- PredyFlexy

A prediction of local protein conformations made with Support Vector Machines.

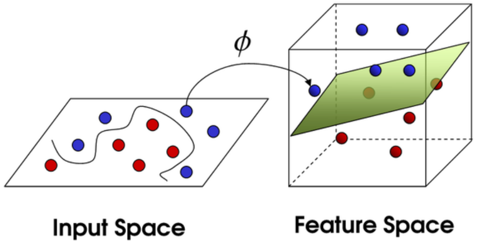
120 different Local Structure Prototypes (11 residues length), so 120 Support Vector Machines (each times 1 against the 119 others, defining the second class)

Protein flexibility prediction

- PredyFlexy

A prediction of local protein conformations made with Support Vector Machines.

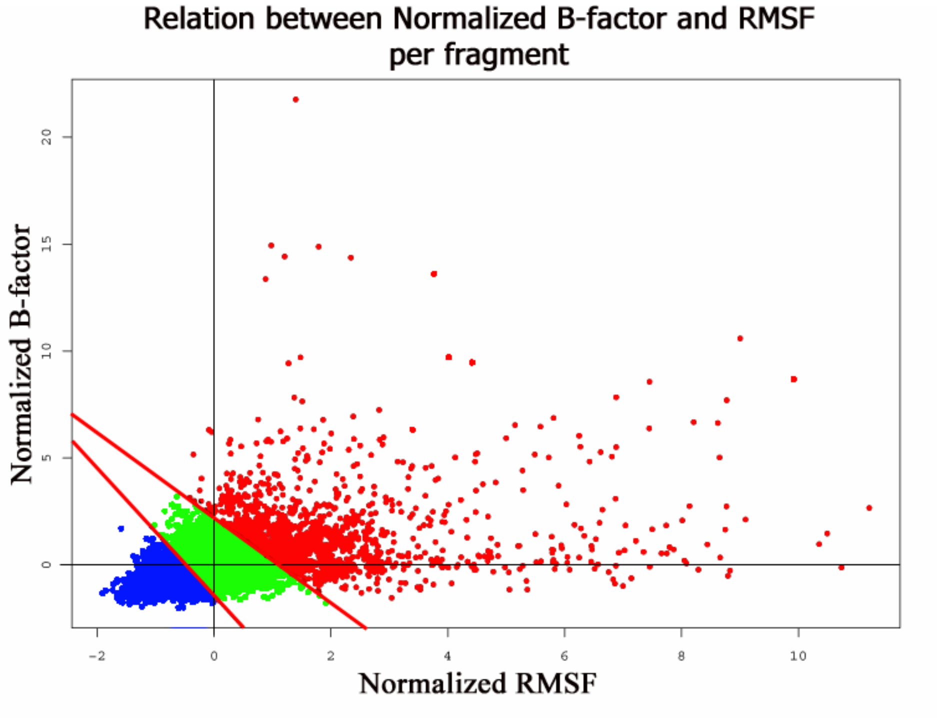
120 different Local Structure Prototypes (11 residues length), so 120 Support Vector Machines (each times 1 against the 119 others, defining the second class)



optimization with RBF (2 parameters tested in grid), i.e. 1000 simulations.

Protein flexibility prediction

- PredyFlexy



- Fragment repartition :
  - Rigid Class 1: 40.4 %
  - Intermediate Class 2 : 36.7 %
  - Flexible Class 3 : 22.9 %

- Limited confusion between extremely different classes :
  - Rigid Class normalized RMSF and Flexible Class normalized B-factor: ~ 2 %
  - Flexible Class normalized RMSF and normalized Rigid Class B-factor: ~ 2 %

Protein flexibility prediction

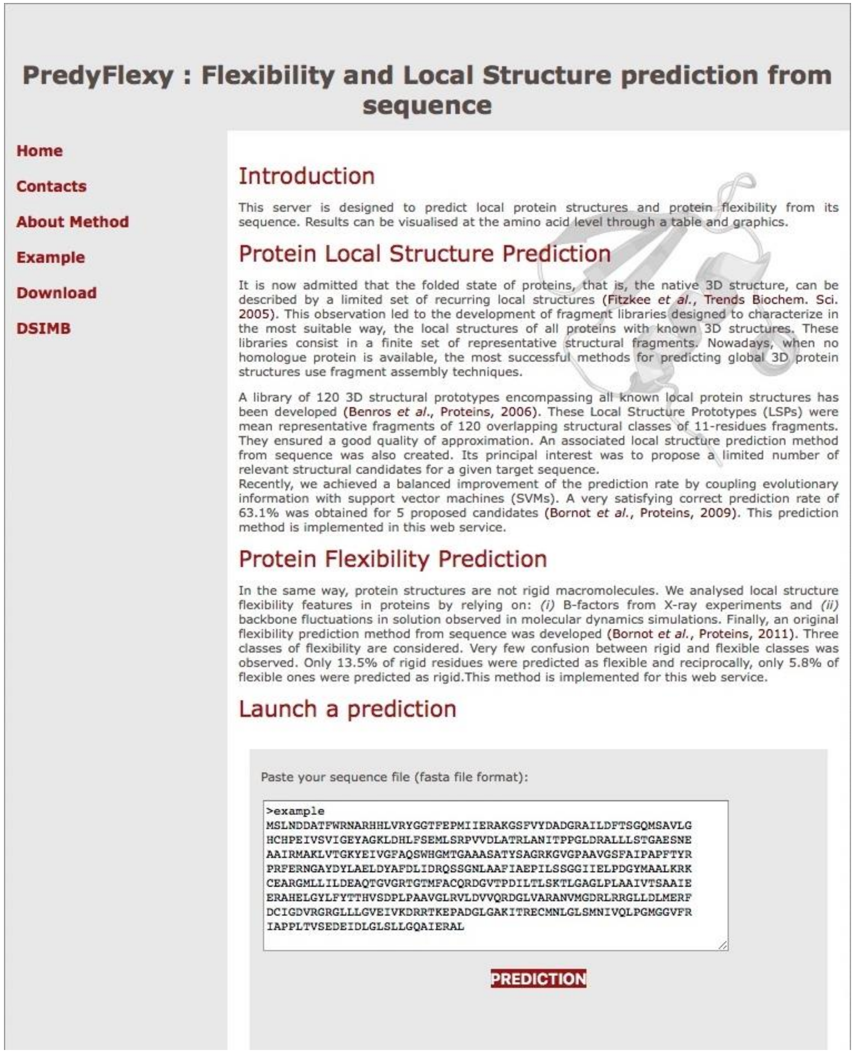
2012

DSIMB
DYNAMICS OF STRUCTURES AND INTERACTIONS OF BIOLOGICAL MACROMOLECULES

- PredyFlexy

Bornot A, Etchebest C, de Brevern AG (2011) Predicting Protein Flexibility through the Prediction of Local Structures. *Proteins*. **79**(3):839-52.

de Brevern AG, Bornot A, Craveur P, Etchebest C, Gelly J-C (2012) PredyFlexy: Flexibility and Local Structure prediction from sequence. *Nucleic Acid Res*. **40**:W317-22.



PredyFlexy : Flexibility and Local Structure prediction from sequence

Introduction

This server is designed to predict local protein structures and protein flexibility from its sequence. Results can be visualised at the amino acid level through a table and graphics.

Protein Local Structure Prediction

It is now admitted that the folded state of proteins, that is, the native 3D structure, can be described by a limited set of recurring local structures (Fitzkee *et al.*, Trends Biochem. Sci. 2005). This observation led to the development of fragment libraries designed to characterize in the most suitable way, the local structures of all proteins with known 3D structures. These libraries consist in a finite set of representative structural fragments. Nowadays, when no homologue protein is available, the most successful methods for predicting global 3D protein structures use fragment assembly techniques.

A library of 120 3D structural prototypes encompassing all known local protein structures has been developed (Benros *et al.*, Proteins, 2006). These Local Structure Prototypes (LSPs) were mean representative fragments of 120 overlapping structural classes of 11-residues fragments. They ensured a good quality of approximation. An associated local structure prediction method from sequence was also created. Its principal interest was to propose a limited number of relevant structural candidates for a given target sequence.
Recently, we achieved a balanced improvement of the prediction rate by coupling evolutionary information with support vector machines (SVMs). A very satisfying correct prediction rate of 63.1% was obtained for 5 proposed candidates (Bornot *et al.*, Proteins, 2009). This prediction method is implemented in this web service.

Protein Flexibility Prediction

In the same way, protein structures are not rigid macromolecules. We analysed local structure flexibility features in proteins by relying on: *(i)* B-factors from X-ray experiments and *(ii)* backbone fluctuations in solution observed in molecular dynamics simulations. Finally, an original flexibility prediction method from sequence was developed (Bornot *et al.*, Proteins, 2011). Three classes of flexibility are considered. Very few confusion between rigid and flexible classes was observed. Only 13.5% of rigid residues were predicted as flexible and reciprocally, only 5.8% of flexible ones were predicted as rigid.This method is implemented for this web service.

Launch a prediction

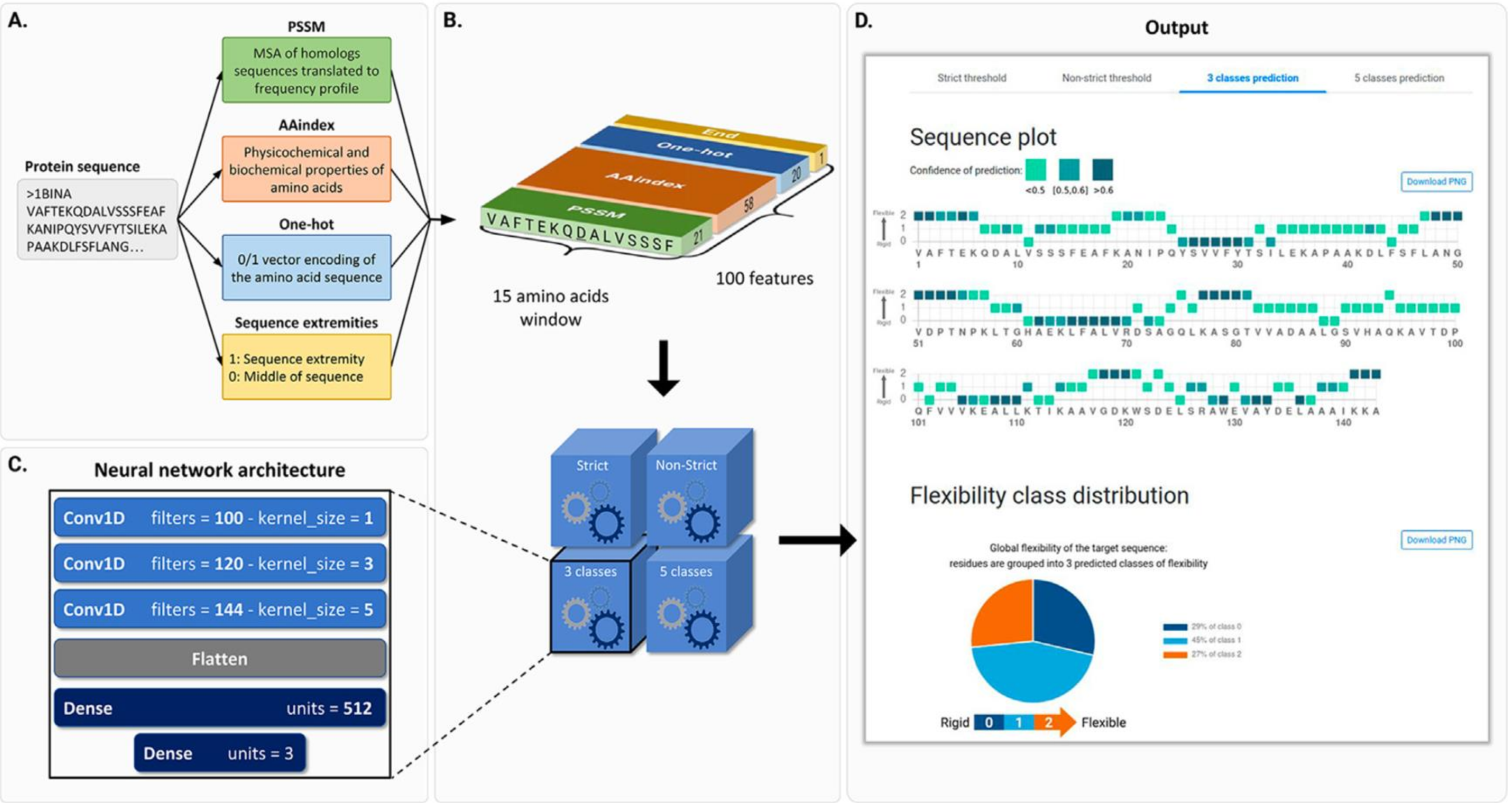
Paste your sequence file (fasta file format):

```
>example
MSLNDDATFWRNARHHLVRYGGTFEPMIIERAKGSFVYDADGRAILDFTSGQMSAVLG
HCHPEIVSVIGEYAGKLDHLFSEMLSRPVVDLATRLANITPPGLDRALLLSTGAESNE
AAIRMAKLVTGKYEIVGFAQSWHGMTGAAASATYSAGRKGVGPAAVGSFAIPAPFTYR
PRFERNGAYDYLAELDYAFDLIDRQSSGNLAAFIAEPILSSGGIIELPDGYMAALKRK
CEARGMLLILDEAQTGVGRTGTMFACQRDGVTPDILTLSKTLGAGLPLAAIVTSAAIE
ERAHELGYLFYTTHVSDPLPAAVGLRVLDVVQRDGLVARANVMGDRLRRGLLDLMERF
DCIGDVRGRGLLLGVEIVKDRRTKEPADGLGAKITRECMNLGLSMNIVQLPGMGGVFR
IAPPLTVSEDEIDLGLSLLGQAIERAL
```

Protein flexibility prediction

2021

- MEDUSA: Deep Learning approach



A.

Protein sequence

>1BINA
VAFTEKQDALVSSSFEAF
KANIPQYSVVFYTSILEKA
PAAKDLFSFLANG…

PSSM

MSA of homologs sequences translated to frequency profile

AAindex

Physicochemical and biochemical properties of amino acids

One-hot

0/1 vector encoding of the amino acid sequence

Sequence extremities

1: Sequence extremity
0: Middle of sequence

B.

15 amino acids window

100 features

C.

Neural network architecture

Conv1D filters = 100 - kernel_size = 1

Conv1D filters = 120 - kernel_size = 3

Conv1D filters = 144 - kernel_size = 5

Flatten

Dense units = 512

Dense units = 3

D.

Output

Sequence plot

Confidence of prediction:
<0.5
[0.5,0.6]
>0.6
Download PNG
Flexible
Rigid
2
1
0
V A F T E K Q D A L V S S S F E A F K A N I P Q Y S V V F Y T S I L E K A P A A K D L F S F L A N G
1
10
20
30
40
50
V D P T N P K L T G H A E K L F A L V R D S A G Q L K A S G T V V A D A A L G S V H A Q K A V T D P
51
60
70
80
90
100
Q F V V V K E A L L K T I K A A V G D K W S D E L S R A W E V A Y D E L A A A I K K A
101
110
120
130
140

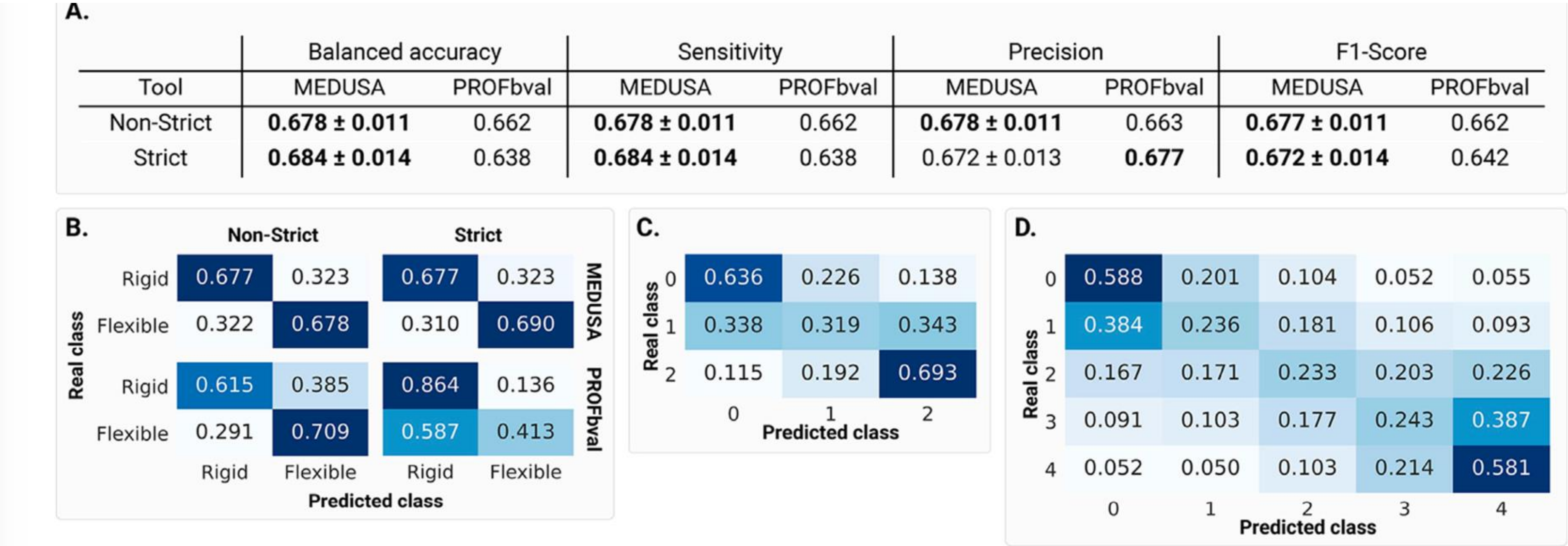
Flexibility class distribution

Global flexibility of the target sequence: residues are grouped into 3 predicted classes of flexibility

Protein flexibility prediction

2021

- MEDUSA: Deep Learning approach



A.

	Balanced accuracy		Sensitivity		Precision		F1-Score	
Tool	MEDUSA	PROFbval	MEDUSA	PROFbval	MEDUSA	PROFbval	MEDUSA	PROFbval
Non-Strict	**0.678 ± 0.011**	0.662	**0.678 ± 0.011**	0.662	**0.678 ± 0.011**	0.663	**0.677 ± 0.011**	0.662
Strict	**0.684 ± 0.014**	0.638	**0.684 ± 0.014**	0.638	0.672 ± 0.013	**0.677**	**0.672 ± 0.014**	0.642

B.

	Real class	Non-Strict		Strict	
		Rigid	Flexible	Rigid	Flexible
MEDUSA	Rigid	0.677	0.323	0.677	0.323
MEDUSA	Flexible	0.322	0.678	0.310	0.690
PROFbval	Rigid	0.615	0.385	0.864	0.136
PROFbval	Flexible	0.291	0.709	0.587	0.413

Predicted class

C.

Real class	0	1	2
0	0.636	0.226	0.138
1	0.338	0.319	0.343
2	0.115	0.192	0.693

Predicted class

D.

Real class	0	1	2	3	4
0	0.588	0.201	0.104	0.052	0.055
1	0.384	0.236	0.181	0.106	0.093
2	0.167	0.171	0.233	0.203	0.226
3	0.091	0.103	0.177	0.243	0.387
4	0.052	0.050	0.103	0.214	0.581

Predicted class

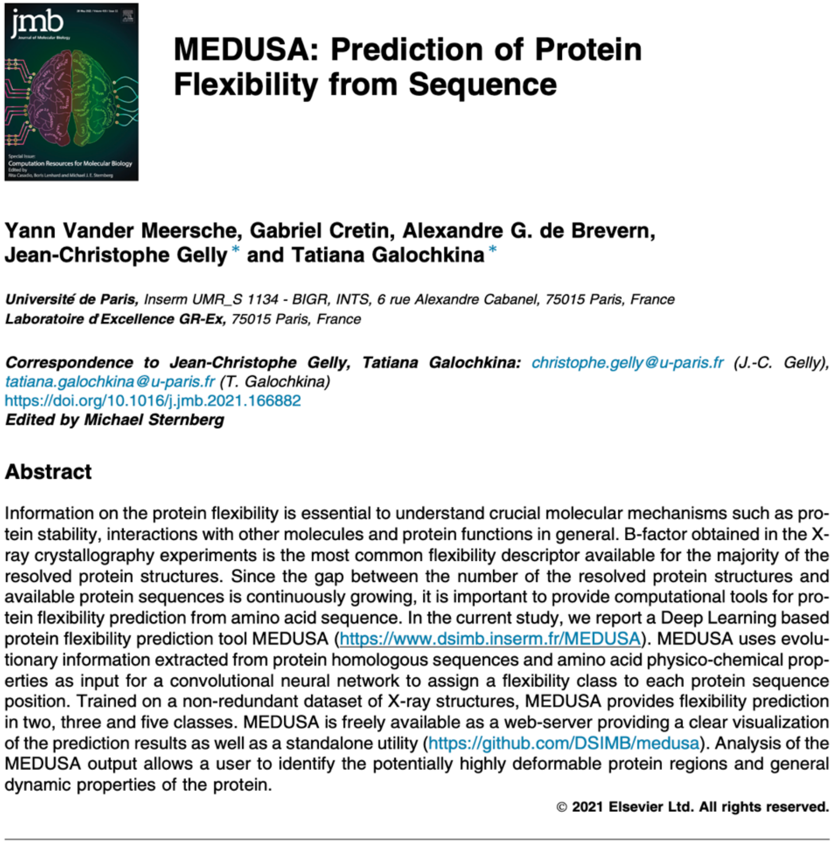
- Excellent results: 2-, 3- and 5-states
- with B-factors

Protein flexibility prediction

2021

- MEDUSA: Deep Learning approach

DSIMB
DYNAMICS OF STRUCTURES AND INTERACTIONS OF BIOLOGICAL MACROMOLECULES



MEDUSA: Prediction of Protein Flexibility from Sequence

Yann Vander Meersche, Gabriel Cretin, Alexandre G. de Brevern, Jean-Christophe Gelly * and Tatiana Galochkina *

***Université de Paris,** Inserm UMR_S 1134 - BIGR, INTS, 6 rue Alexandre Cabanel, 75015 Paris, France*
***Laboratoire d'Excellence GR-Ex,** 75015 Paris, France*

***Correspondence to Jean-Christophe Gelly, Tatiana Galochkina:** christophe.gelly@u-paris.fr (J.-C. Gelly), tatiana.galochkina@u-paris.fr (T. Galochkina)*

https://doi.org/10.1016/j.jmb.2021.166882
Edited by Michael Sternberg

Abstract

Information on the protein flexibility is essential to understand crucial molecular mechanisms such as protein stability, interactions with other molecules and protein functions in general. B-factor obtained in the X-ray crystallography experiments is the most common flexibility descriptor available for the majority of the resolved protein structures. Since the gap between the number of the resolved protein structures and available protein sequences is continuously growing, it is important to provide computational tools for protein flexibility prediction from amino acid sequence. In the current study, we report a Deep Learning based protein flexibility prediction tool MEDUSA (https://www.dsimb.inserm.fr/MEDUSA). MEDUSA uses evolutionary information extracted from protein homologous sequences and amino acid physico-chemical properties as input for a convolutional neural network to assign a flexibility class to each protein sequence position. Trained on a non-redundant dataset of X-ray structures, MEDUSA provides flexibility prediction in two, three and five classes. MEDUSA is freely available as a web-server providing a clear visualization of the prediction results as well as a standalone utility (https://github.com/DSIMB/medusa). Analysis of the MEDUSA output allows a user to identify the potentially highly deformable protein regions and general dynamic properties of the protein.

Protein flexibility prediction

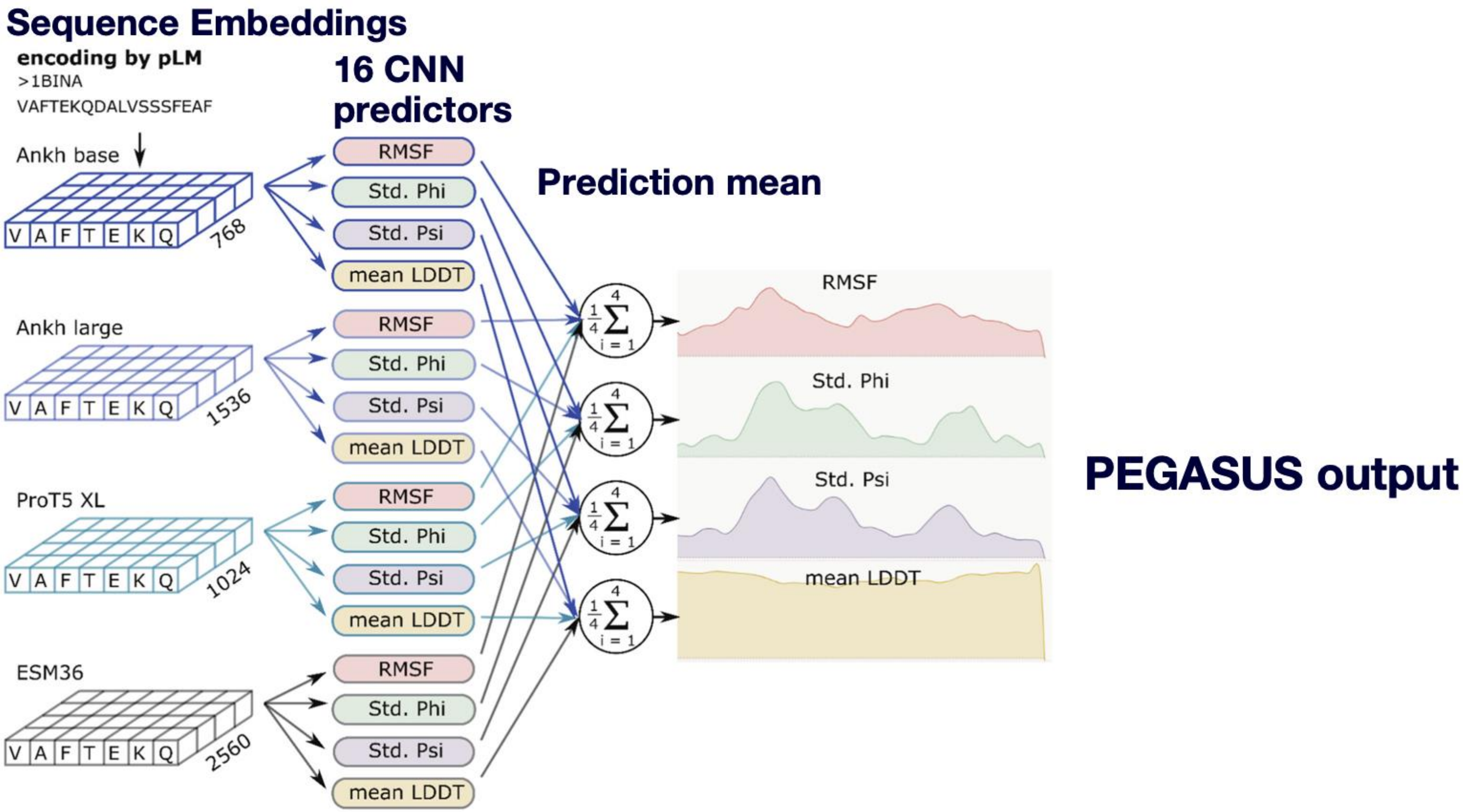
2021

- MEDUSA: Deep Learning approach
- Excellent results: 2-, 3- and 5-states
- A dedicated webserver
- ~150,000 to 350,000 trainable parameters

Protein flexibility prediction

2025

- To predict RMSF

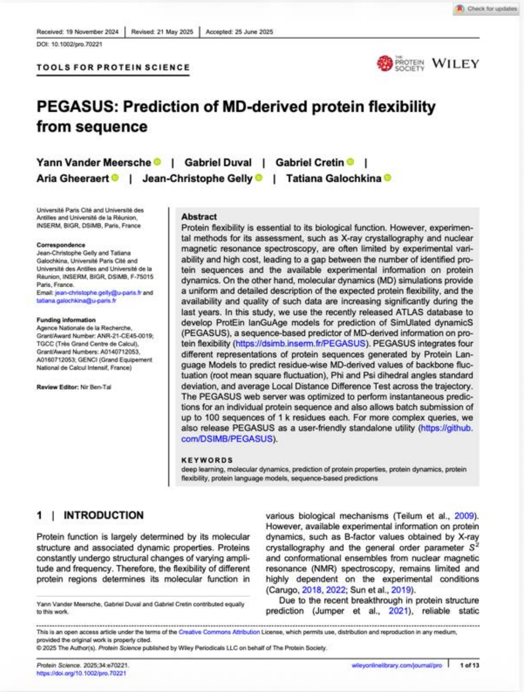


Protein flexibility prediction

2025

- To predict RMSF

- Flexibility prediction in 2 classes (rigid/flexible): F1 score of 0.71 vs. 0.65
- Better predictions despite a 7 × smaller training set
- Embeddings more informative than classical evolutionary/physicochemical descriptors
- Capable of detecting changes in flexibility induced by point mutations



Check for updates

Received: 19 November 2024 | Revised: 21 May 2025 | Accepted: 25 June 2025
DOI: 10.1002/pro.70221

TOOLS FOR PROTEIN SCIENCE

THE PROTEIN SOCIETY WILEY

PEGASUS: Prediction of MD-derived protein flexibility from sequence

Yann Vander Meersche | Gabriel Duval | Gabriel Cretin | Aria Gheeraert | Jean-Christophe Gelly | Tatiana Galochkina

Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM, BIGR, DSIMB, Paris, France

Correspondence
Jean-Christophe Gelly and Tatiana Galochkina, Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM, BIGR, DSIMB, F-75015 Paris, France.
Email: jean-christophe.gelly@u-paris.fr and tatiana.galochkina@u-paris.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-21-CE45-0019; TGCC (Très Grand Centre de Calcul), Grant/Award Numbers: A0140712053, A0160712053; GENCI (Grand Equipement National de Calcul Intensif, France)

Review Editor: Nir Ben-Tal

Abstract
Protein flexibility is essential to its biological function. However, experimental methods for its assessment, such as X-ray crystallography and nuclear magnetic resonance spectroscopy, are often limited by experimental variability and high cost, leading to a gap between the number of identified protein sequences and the available experimental information on protein dynamics. On the other hand, molecular dynamics (MD) simulations provide a uniform and detailed description of the expected protein flexibility, and the availability and quality of such data are increasing significantly during the last years. In this study, we use the recently released ATLAS database to develop ProtEin lanGuAge models for prediction of SimUlated dynamicS (PEGASUS), a sequence-based predictor of MD-derived information on protein flexibility (https://dsimb.inserm.fr/PEGASUS). PEGASUS integrates four different representations of protein sequences generated by Protein Language Models to predict residue-wise MD-derived values of backbone fluctuation (root mean square fluctuation), Phi and Psi dihedral angles standard deviation, and average Local Distance Difference Test across the trajectory. The PEGASUS web server was optimized to perform instantaneous predictions for an individual protein sequence and also allows batch submission of up to 100 sequences of 1 k residues each. For more complex queries, we also release PEGASUS as a user-friendly standalone utility (https://github.com/DSIMB/PEGASUS).

KEYWORDS
deep learning, molecular dynamics, prediction of protein properties, protein dynamics, protein flexibility, protein language models, sequence-based predictions

1 | INTRODUCTION

Protein function is largely determined by its molecular structure and associated dynamic properties. Proteins constantly undergo structural changes of varying amplitude and frequency. Therefore, the flexibility of different protein regions determines its molecular function in various biological mechanisms (Teilum et al., 2009). However, available experimental information on protein dynamics, such as B-factor values obtained by X-ray crystallography and the general order parameter S^2 and conformational ensembles from nuclear magnetic resonance (NMR) spectroscopy, remains limited and highly dependent on the experimental conditions (Carugo, 2018, 2022; Sun et al., 2019).

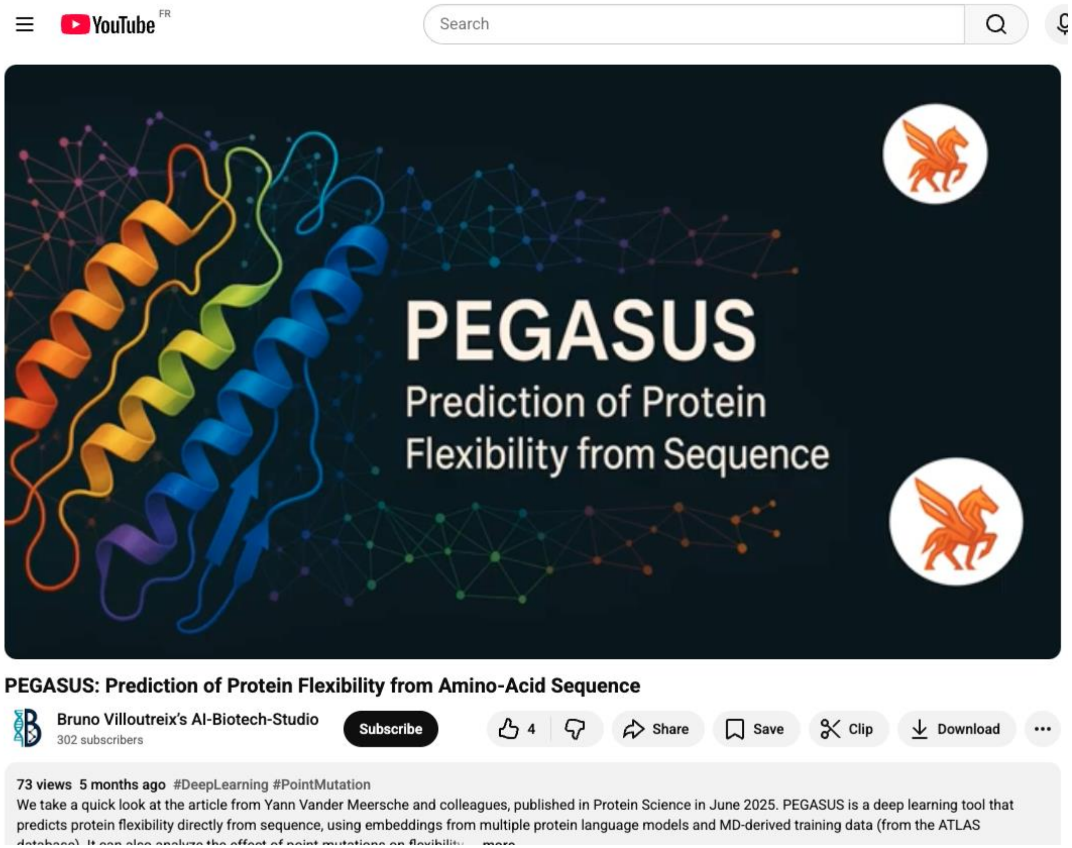
Due to the recent breakthrough in protein structure prediction (Jumper et al., 2021), reliable static

Yann Vander Meersche, Gabriel Duval and Gabriel Cretin contributed equally to this work.

Protein Science. 2025;34:e70221. https://doi.org/10.1002/pro.70221 wileyonlinelibrary.com/journal/pro 1 of 13

Protein flexibility prediction

2025



PEGASUS: Prediction of Protein Flexibility from Amino-Acid Sequence

Bruno Villoutreix's AI-Biotech-Studio
302 subscribers

73 views 5 months ago #DeepLearning #PointMutation
We take a quick look at the article from Yann Vander Meersche and colleagues, published in Protein Science in June 2025. PEGASUS is a deep learning tool that predicts protein flexibility directly from sequence, using embeddings from multiple protein language models and MD-derived training data (from the ATLAS database). It can also analyze the effect of point mutations on flexibility ...more

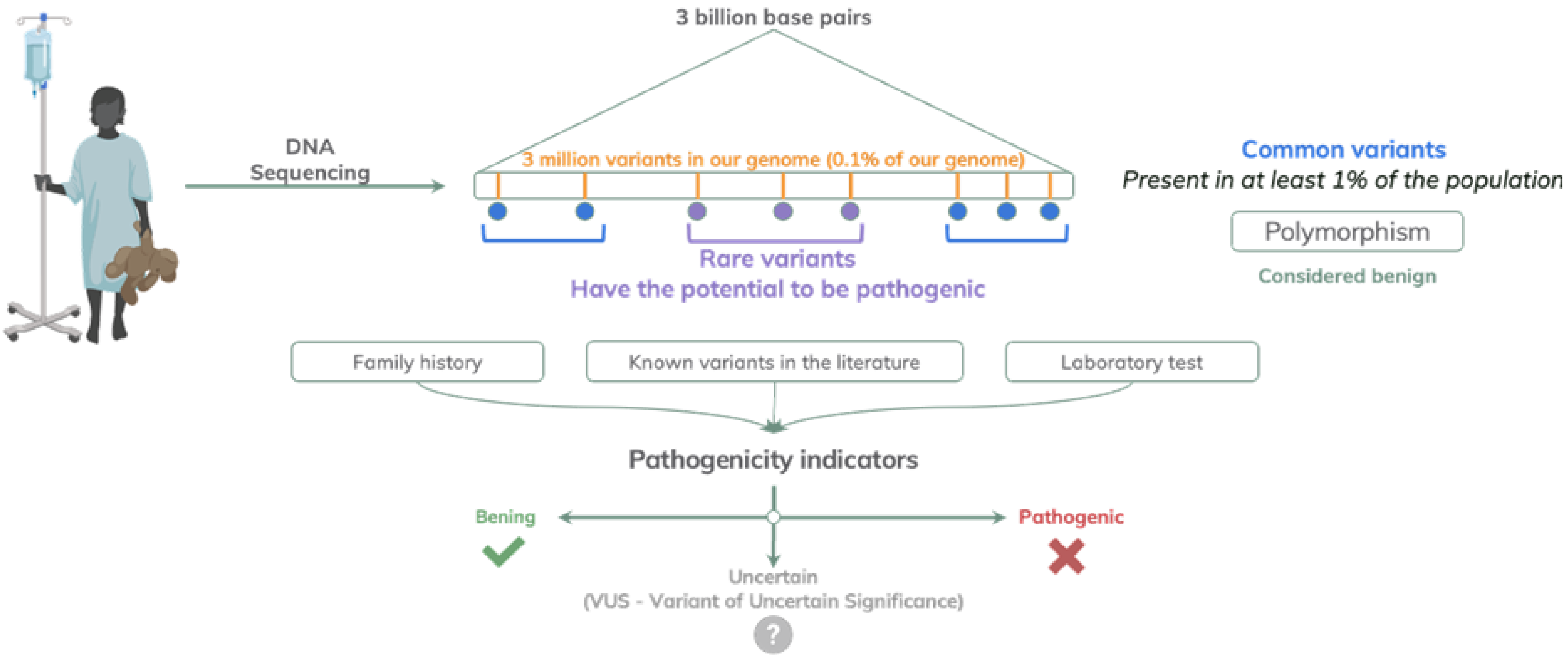
Bruno Villoutreix’s AI-Biotech-Studio

https://www.youtube.com/watch?v=tXu53l1K7h8

4. PATHOLOGY PREDICTION

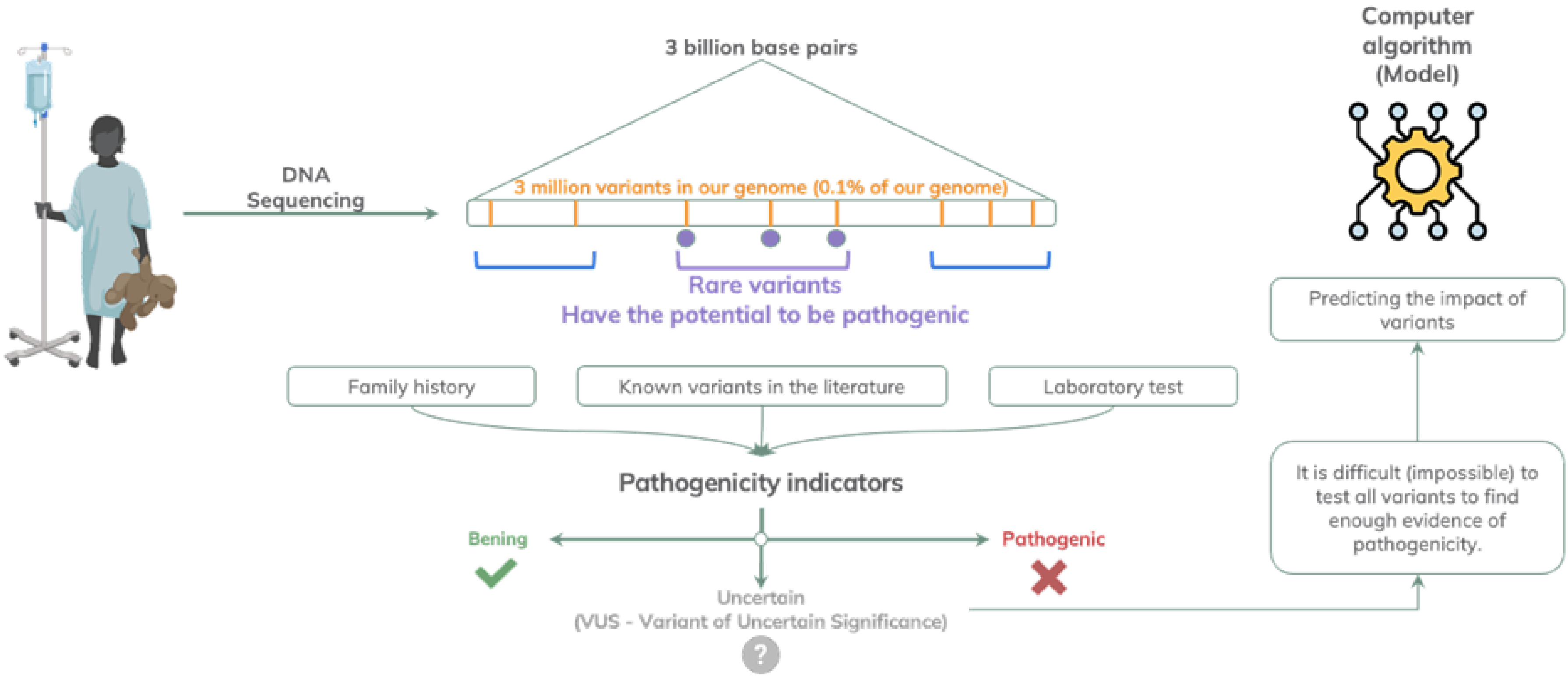
Pathology prediction

- Pathology prediction: the question


3 billion base pairs
DNA Sequencing
3 million variants in our genome (0.1% of our genome)
Rare variants
Have the potential to be pathogenic
Common variants
Present in at least 1% of the population
Polymorphism
Considered benign
Family history
Known variants in the literature
Laboratory test
Pathogenicity indicators
Bening
Pathogenic
Uncertain
(VUS - Variant of Uncertain Significance)
?

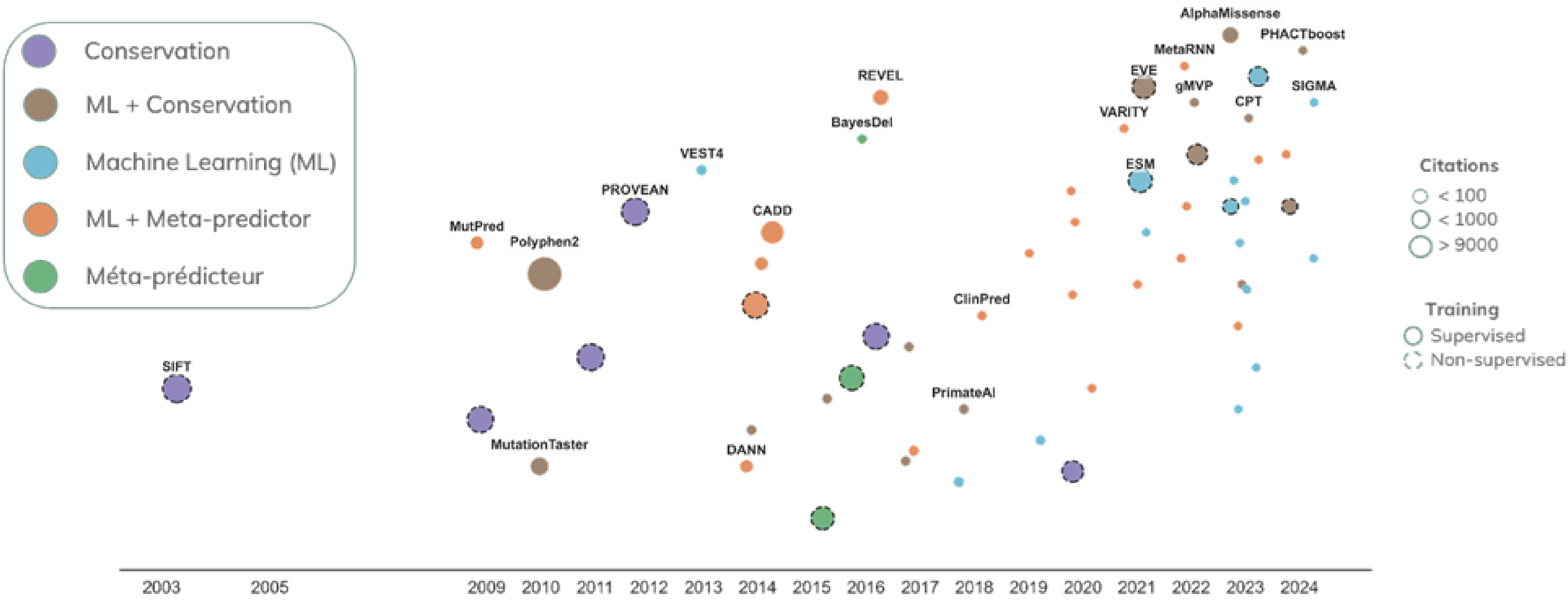
Pathology prediction

- Pathology prediction: model – Variant Effect Predictor


DNA
Sequencing
3 billion base pairs
3 million variants in our genome (0.1% of our genome)
Rare variants
Have the potential to be pathogenic
Computer
algorithm
(Model)
Predicting the impact of variants
Family history
Known variants in the literature
Laboratory test
Pathogenicity indicators
Bening
Pathogenic
Uncertain
(VUS - Variant of Uncertain Significance)
?
It is difficult (impossible) to test all variants to find enough evidence of pathogenicity.

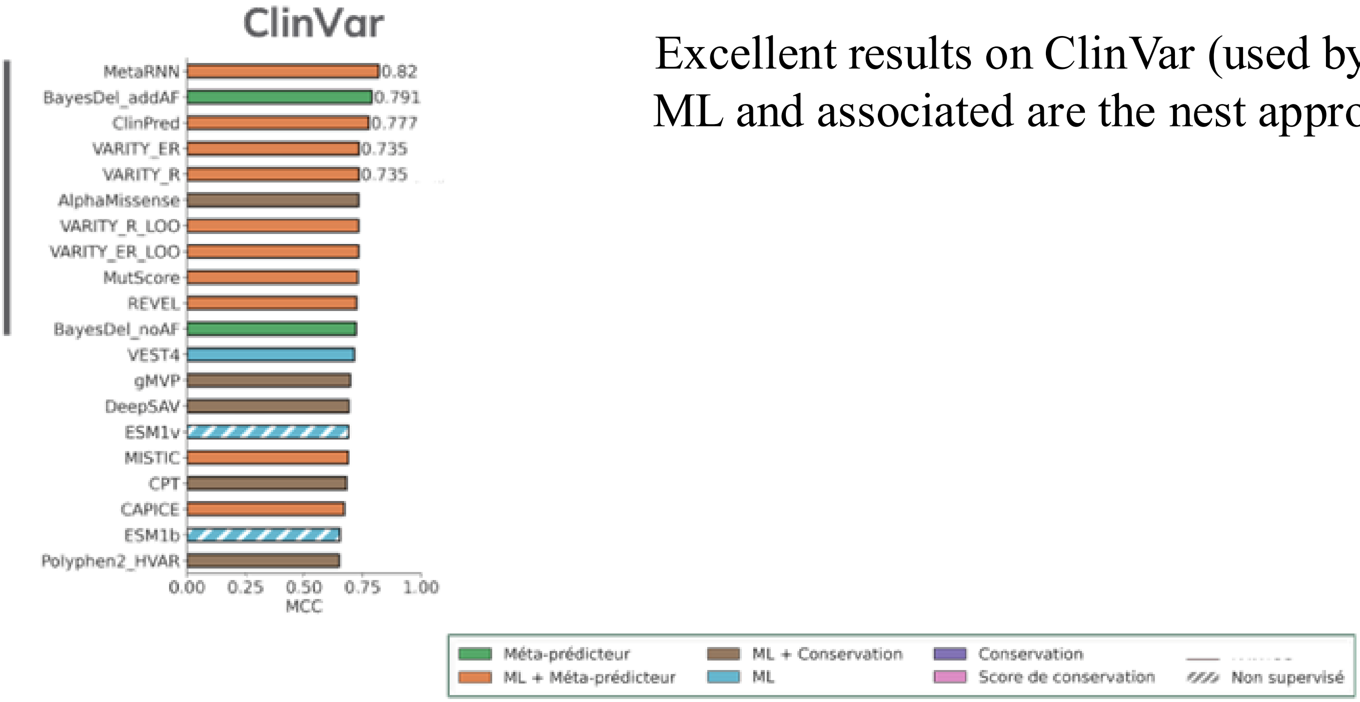
Pathology prediction

- ➢ Pathology prediction: VEP chronology



Pathology prediction

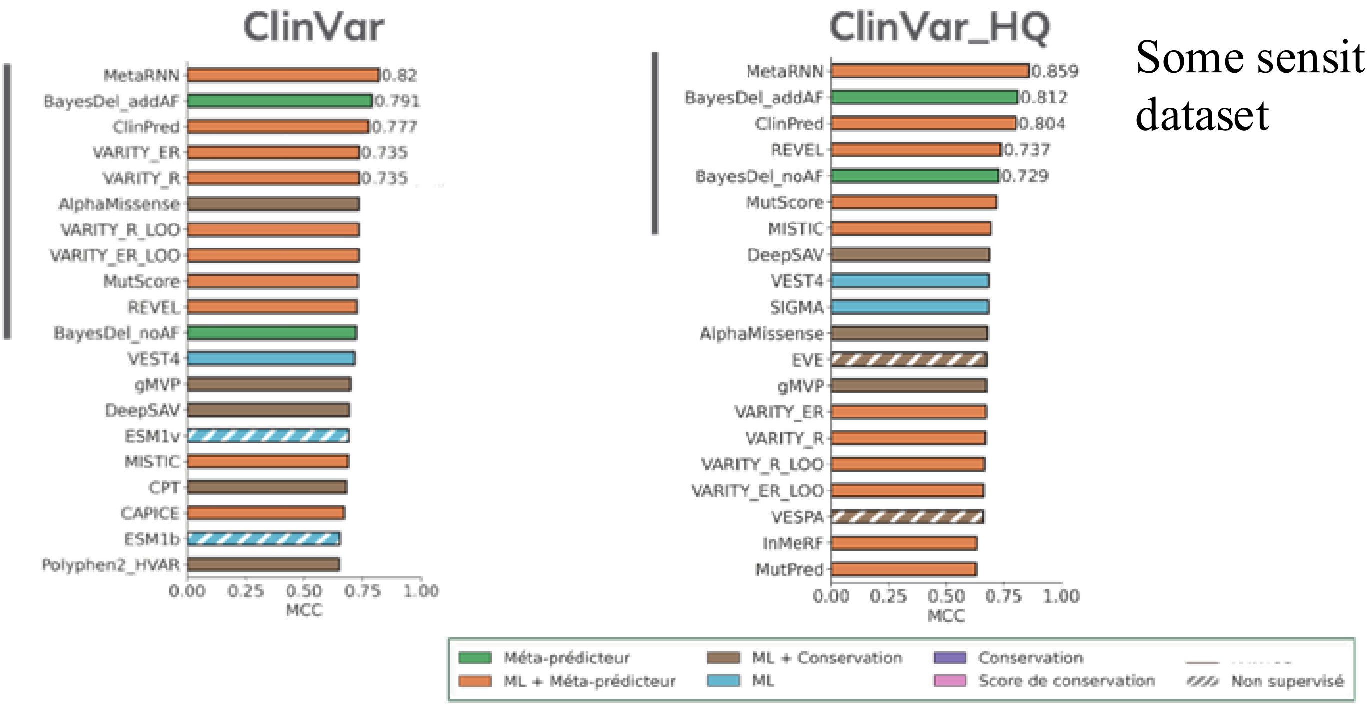
- ➢ Pathology prediction: benchmark (MCC)



Excellent results on ClinVar (used by most methods)
ML and associated are the nest approaches

Pathology prediction

- Pathology prediction: benchmark (MCC)

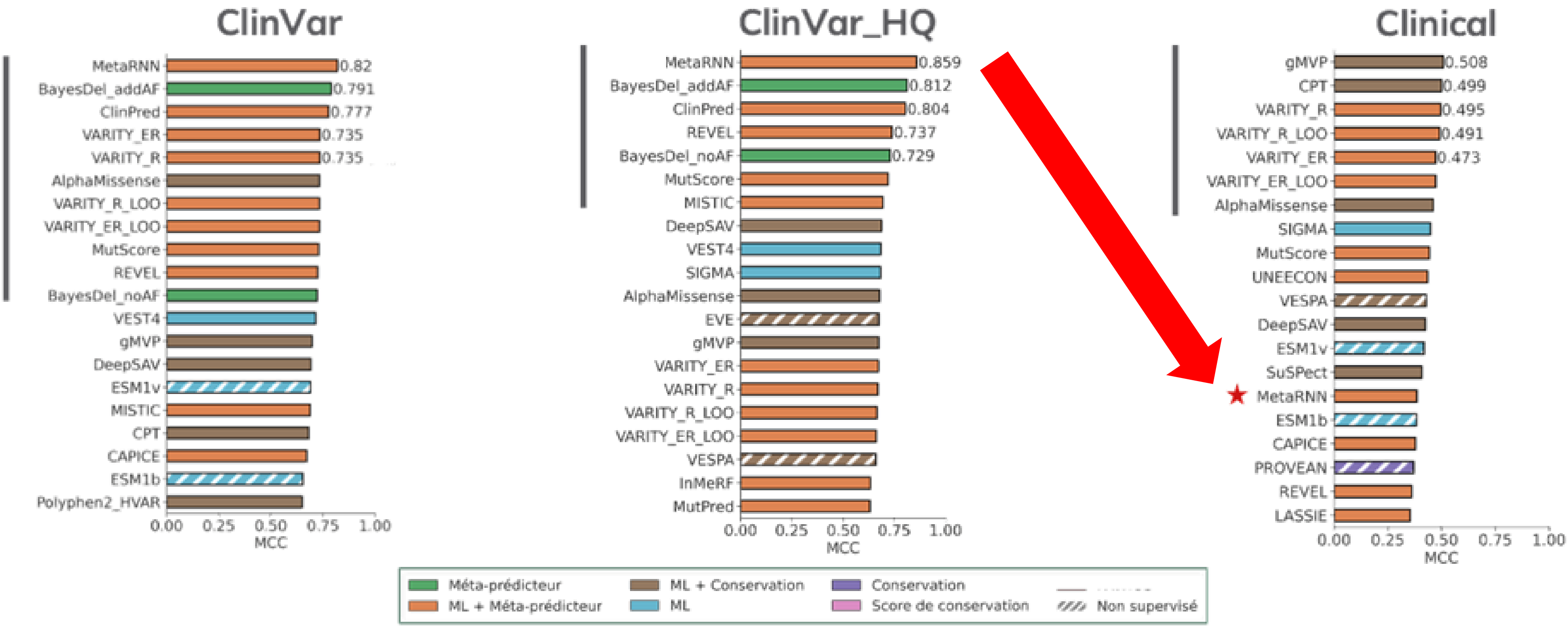


Some sensitivity on the test dataset

Pathology prediction

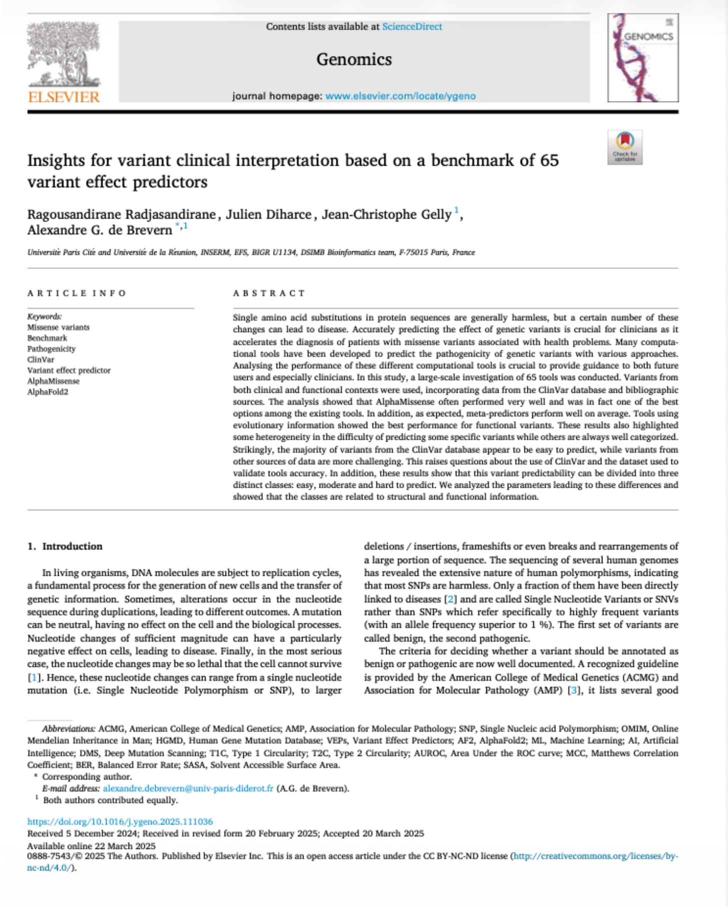
- Pathology prediction: benchmark (MCC)

In fact, terribly sensible on the dataset (different type of bias)



Pathology prediction

- Pathology prediction: benchmark



Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Insights for variant clinical interpretation based on a benchmark of 65 variant effect predictors

Ragousandirane Radjasandirane, Julien Diharce, Jean-Christophe Gelly[1], Alexandre G. de Brevern[*,1]

Université Paris Cité and Université de la Réunion, INSERM, EFS, BIGR U1134, DSIMB Bioinformatics team, F-75015 Paris, France

ARTICLE INFO

Keywords:
Missense variants
Benchmark
Pathogenicity
ClinVar
Variant effect predictor
AlphaMissense
AlphaFold2

ABSTRACT

Single amino acid substitutions in protein sequences are generally harmless, but a certain number of these changes can lead to disease. Accurately predicting the effect of genetic variants is crucial for clinicians as it accelerates the diagnosis of patients with missense variants associated with health problems. Many computational tools have been developed to predict the pathogenicity of genetic variants with various approaches. Analysing the performance of these different computational tools is crucial to provide guidance to both future users and especially clinicians. In this study, a large-scale investigation of 65 tools was conducted. Variants from both clinical and functional contexts were used, incorporating data from the ClinVar database and bibliographic sources. The analysis showed that AlphaMissense often performed very well and was in fact one of the best options among the existing tools. In addition, as expected, meta-predictors perform well on average. Tools using evolutionary information showed the best performance for functional variants. These results also highlighted some heterogeneity in the difficulty of predicting some specific variants while others are always well categorized. Strikingly, the majority of variants from the ClinVar database appear to be easy to predict, while variants from other sources of data are more challenging. This raises questions about the use of ClinVar and the dataset used to validate tools accuracy. In addition, these results show that this variant predictability can be divided into three distinct classes: easy, moderate and hard to predict. We analyzed the parameters leading to these differences and showed that the classes are related to structural and functional information.

1. Introduction

In living organisms, DNA molecules are subject to replication cycles, a fundamental process for the generation of new cells and the transfer of genetic information. Sometimes, alterations occur in the nucleotide sequence during duplications, leading to different outcomes. A mutation can be neutral, having no effect on the cell and the biological processes. Nucleotide changes of sufficient magnitude can have a particularly negative effect on cells, leading to disease. Finally, in the most serious case, the nucleotide changes may be so lethal that the cell cannot survive [1]. Hence, these nucleotide changes can range from a single nucleotide mutation (i.e. Single Nucleotide Polymorphism or SNP), to larger deletions / insertions, frameshifts or even breaks and rearrangements of a large portion of sequence. The sequencing of several human genomes has revealed the extensive nature of human polymorphisms, indicating that most SNPs are harmless. Only a fraction of them have been directly linked to diseases [2] and are called Single Nucleotide Variants or SNVs rather than SNPs which refer specifically to highly frequent variants (with an allele frequency superior to 1 %). The first set of variants are called benign, the second pathogenic.

The criteria for deciding whether a variant should be annotated as benign or pathogenic are now well documented. A recognized guideline is provided by the American College of Medical Genetics (ACMG) and Association for Molecular Pathology (AMP) [3], it lists several good

Abbreviations: ACMG, American College of Medical Genetics; AMP, Association for Molecular Pathology; SNP, Single Nucleic acid Polymorphism; OMIM, Online Mendelian Inheritance in Man; HGMD, Human Gene Mutation Database; VEPs, Variant Effect Predictors; AF2, AlphaFold2; ML, Machine Learning; AI, Artificial Intelligence; DMS, Deep Mutation Scanning; T1C, Type 1 Circularity; T2C, Type 2 Circularity; AUROC, Area Under the ROC curve; MCC, Matthews Correlation Coefficient; BER, Balanced Error Rate; SASA, Solvent Accessible Surface Area.

[*] Corresponding author.

E-mail address: alexandre.debrevern@univ-paris-diderot.fr (A.G. de Brevern).

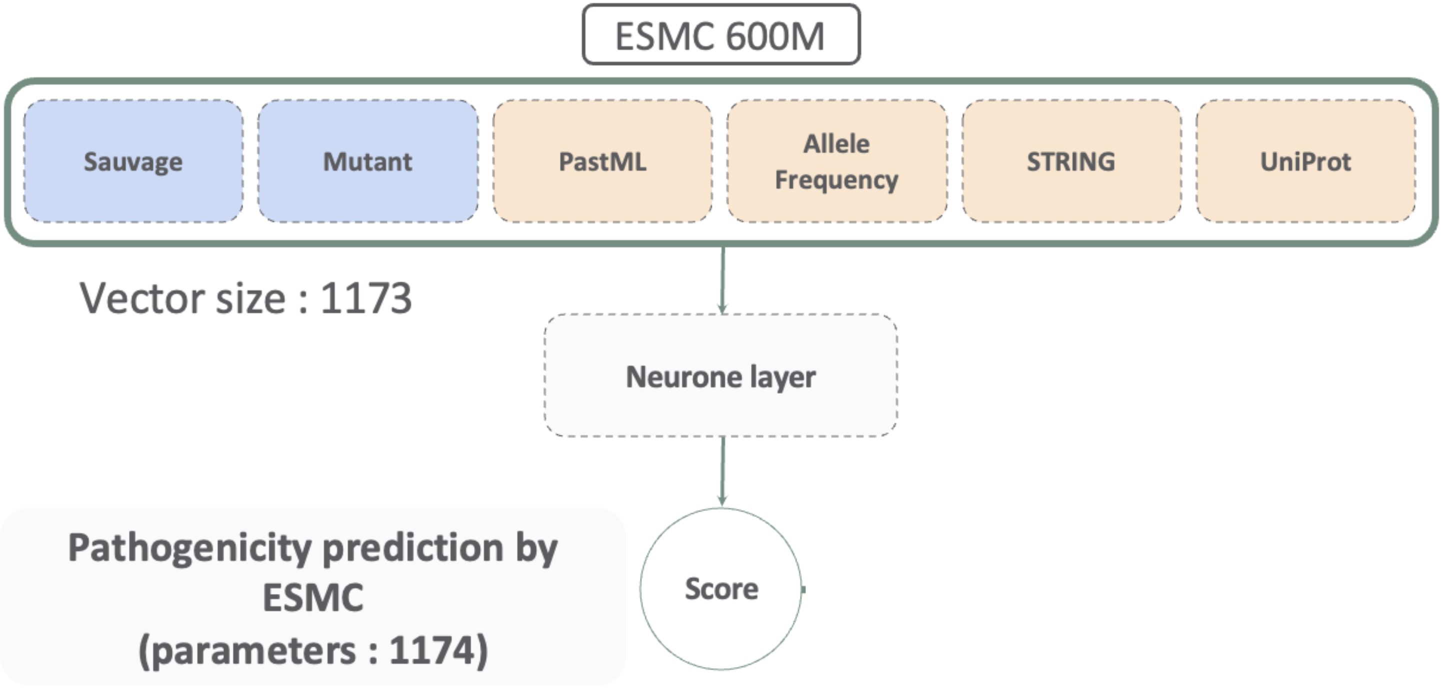
[1] Both authors contributed equally.

https://doi.org/10.1016/j.ygeno.2025.111036
Received 5 December 2024; Received in revised form 20 February 2025; Accepted 20 March 2025
Available online 22 March 2025

In fact, terribly sensible on the dataset (different type of bias)

Pathology prediction

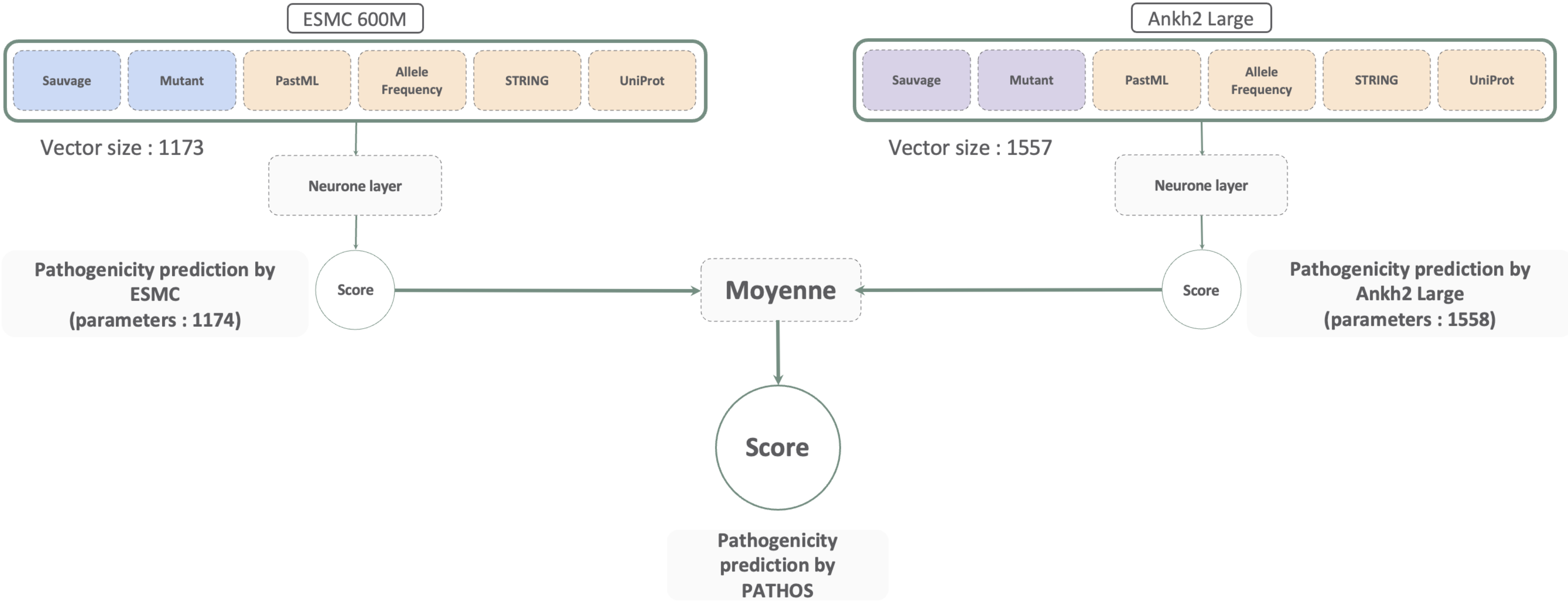
- Pathology prediction: a new approach (with proper data)



Pathogenicity prediction by ESMC (parameters : 1174)

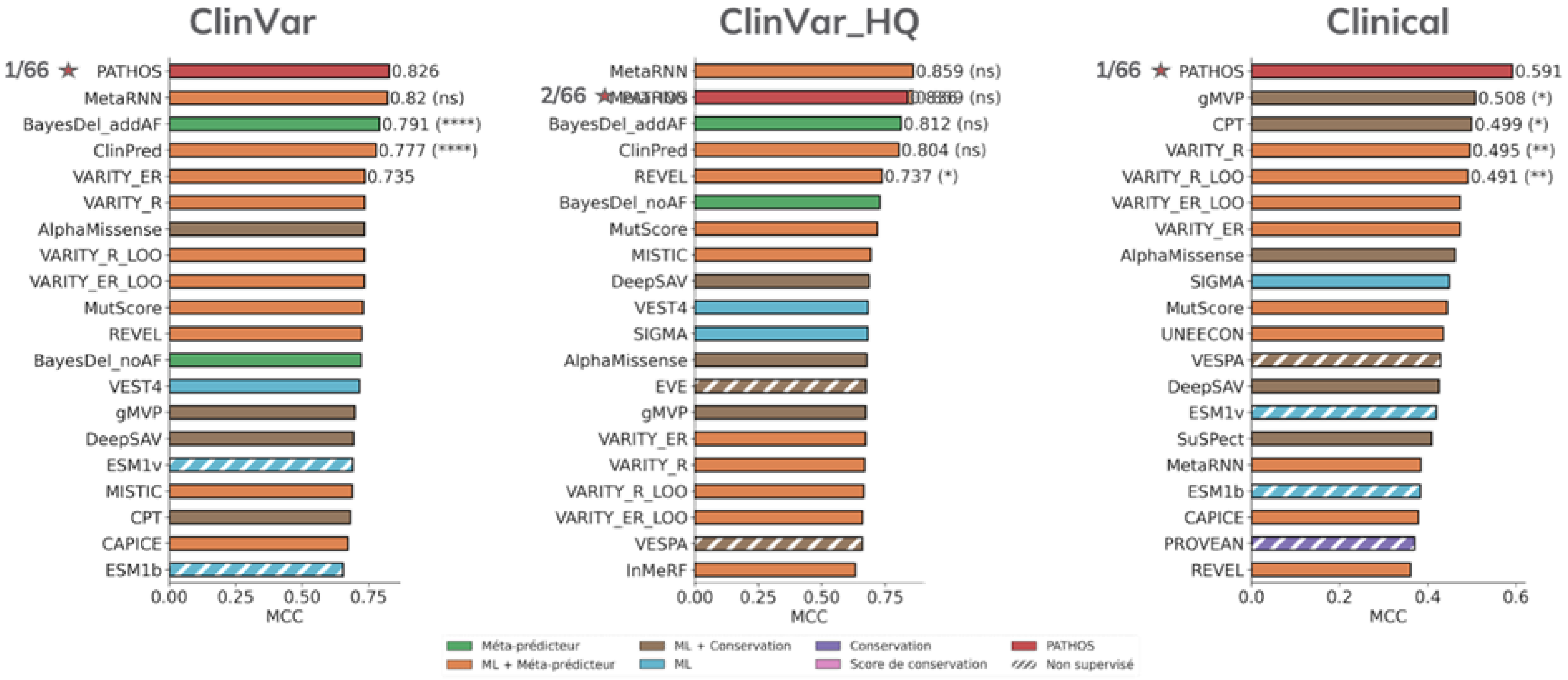
Pathology prediction

- Pathology prediction: PATHOS



Pathology prediction

- Pathology prediction: PATHOS

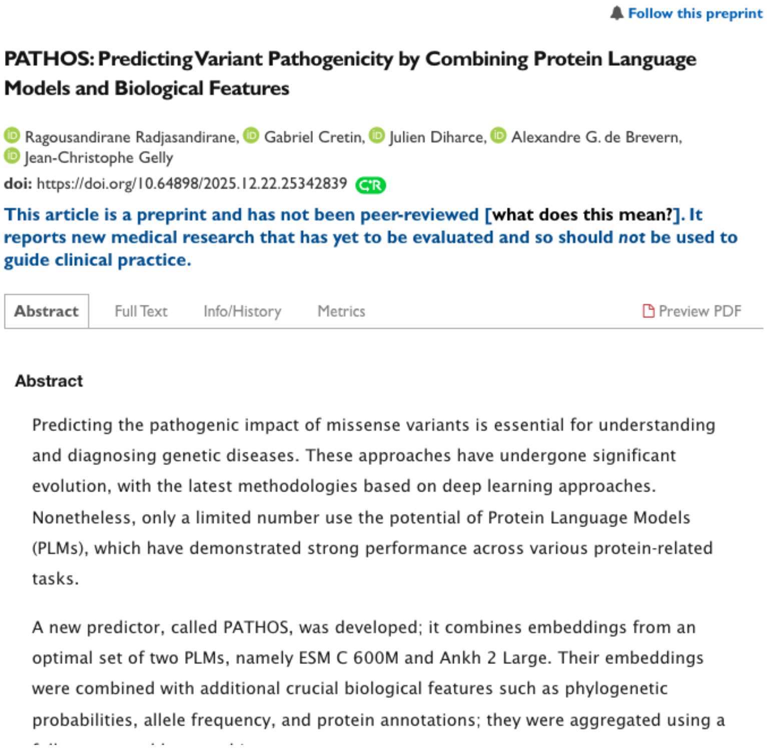


Pathology prediction

- Pathology prediction: PATHOS

H

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES



Follow this preprint

PATHOS: Predicting Variant Pathogenicity by Combining Protein Language Models and Biological Features

Ragousandirane Radjasandirane, Gabriel Cretin, Julien Diharce, Alexandre G. de Brevern, Jean-Christophe Gelly

doi: https://doi.org/10.64898/2025.12.22.25342839

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

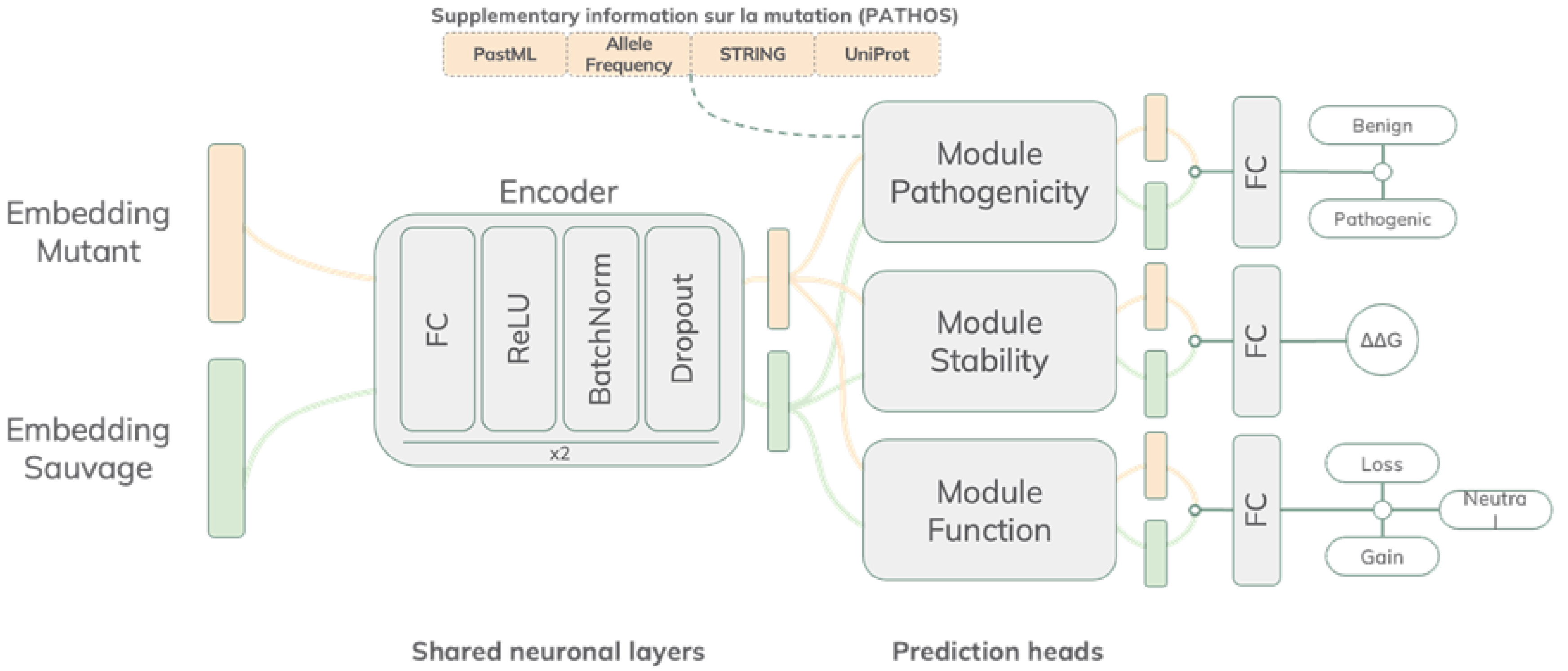
Abstract

Predicting the pathogenic impact of missense variants is essential for understanding and diagnosing genetic diseases. These approaches have undergone significant evolution, with the latest methodologies based on deep learning approaches. Nonetheless, only a limited number use the potential of Protein Language Models (PLMs), which have demonstrated strong performance across various protein-related tasks.

A new predictor, called PATHOS, was developed; it combines embeddings from an optimal set of two PLMs, namely ESM C 600M and Ankh 2 Large. Their embeddings were combined with additional crucial biological features such as phylogenetic probabilities, allele frequency, and protein annotations; they were aggregated using a

Pathology prediction

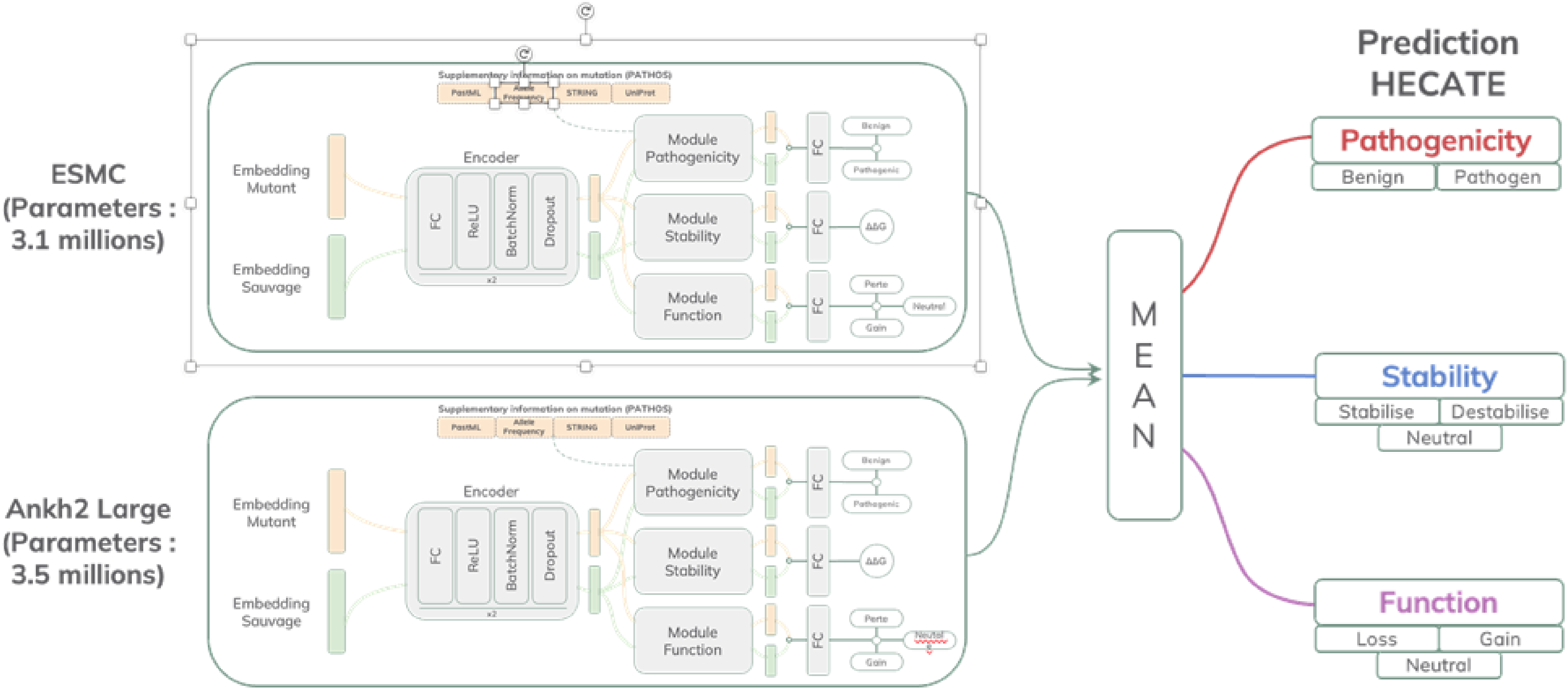
- Pathology prediction ++ … : HECATE, a Siamese model



FC: Fully connected layer
(Layer of artificial neurons)

Pathology prediction

- Pathology prediction ++ … : HECATE, a Siamese model

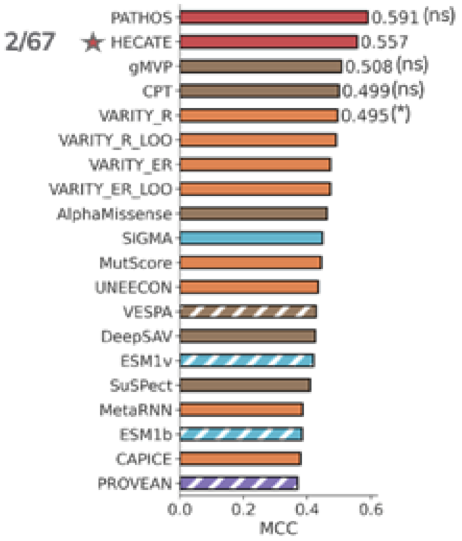


Pathology prediction

➢ HECATE results

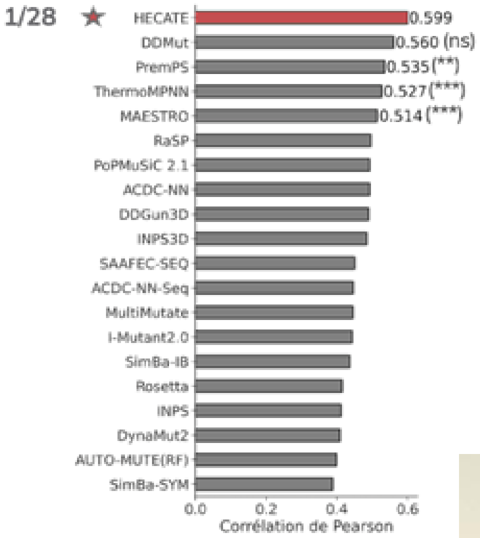
Pathogenicity

Data : Clinical, used in our comparative evaluation



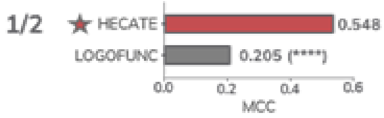
Stability

Comparative evaluation published by Zheng et al.
27 tools tested on 4038 mutations



Gain/Loss of Function

Dataset published by Flanagan et al. 133 mutations from the literature and patient cohorts



HECATE performs well in all areas

Pathology prediction

- Pathology prediction

Terribly sensitive to the data curation

Improvement can be done

Architecture can be used for other property researches

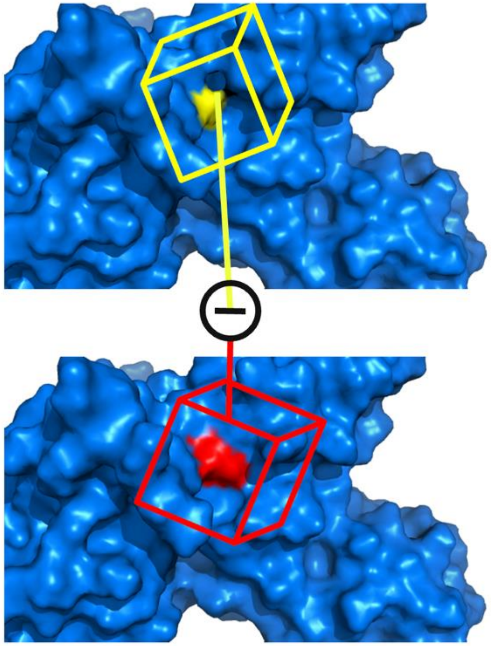
5. RECENT DEVELOPMENTS

Recent developments

➢ Prediction of the effects of mutations on protein interactions

Alessandra Carbone

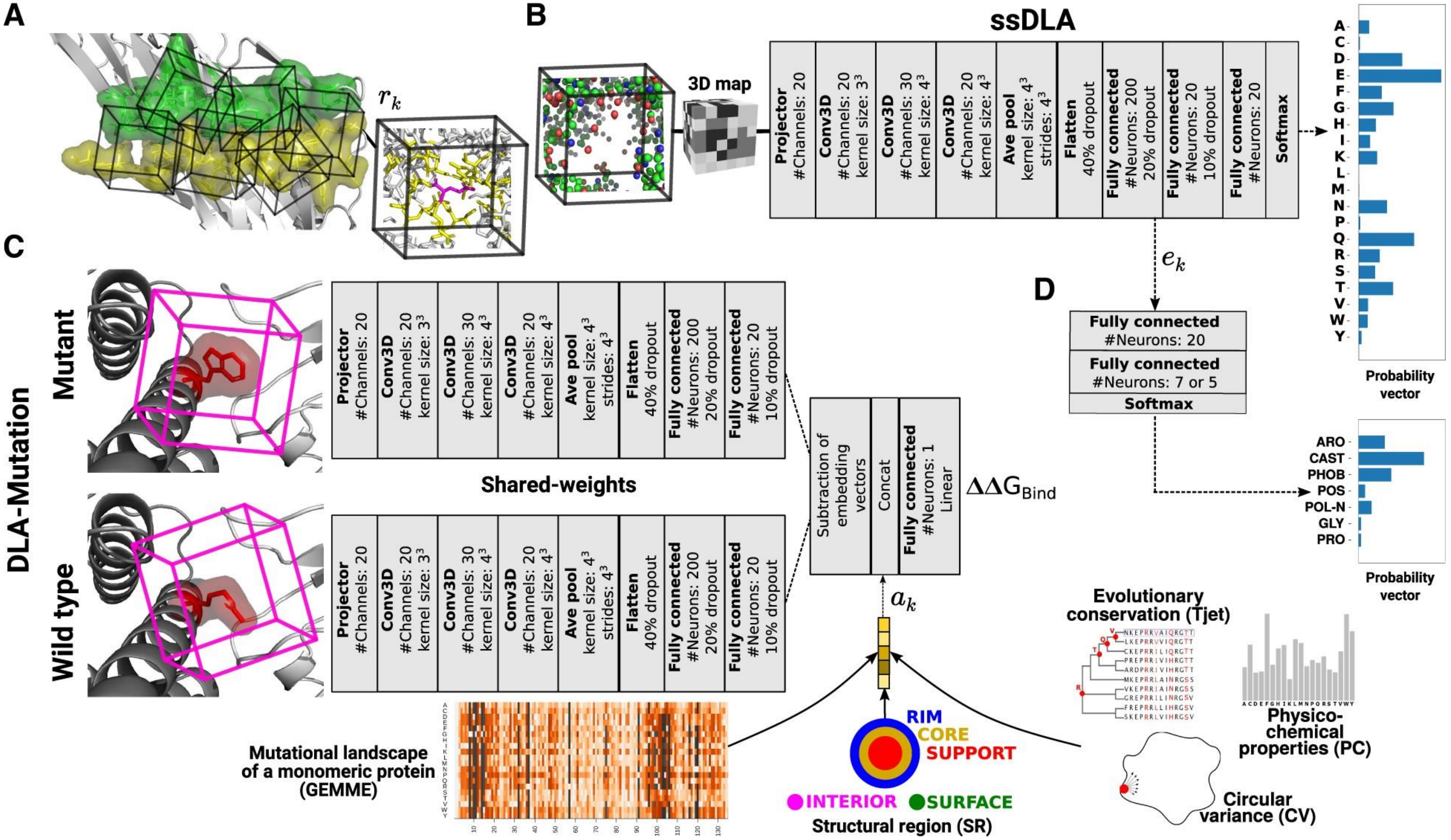
Élodie Laine



- A deep learning framework
- Mutations on the interface of PPI
- Direct prediction of ΔΔGBind
- Local changes
- Leverage the knowledge of protein-protein interactions
- Transfer the knowledge in an end-to-end architecture

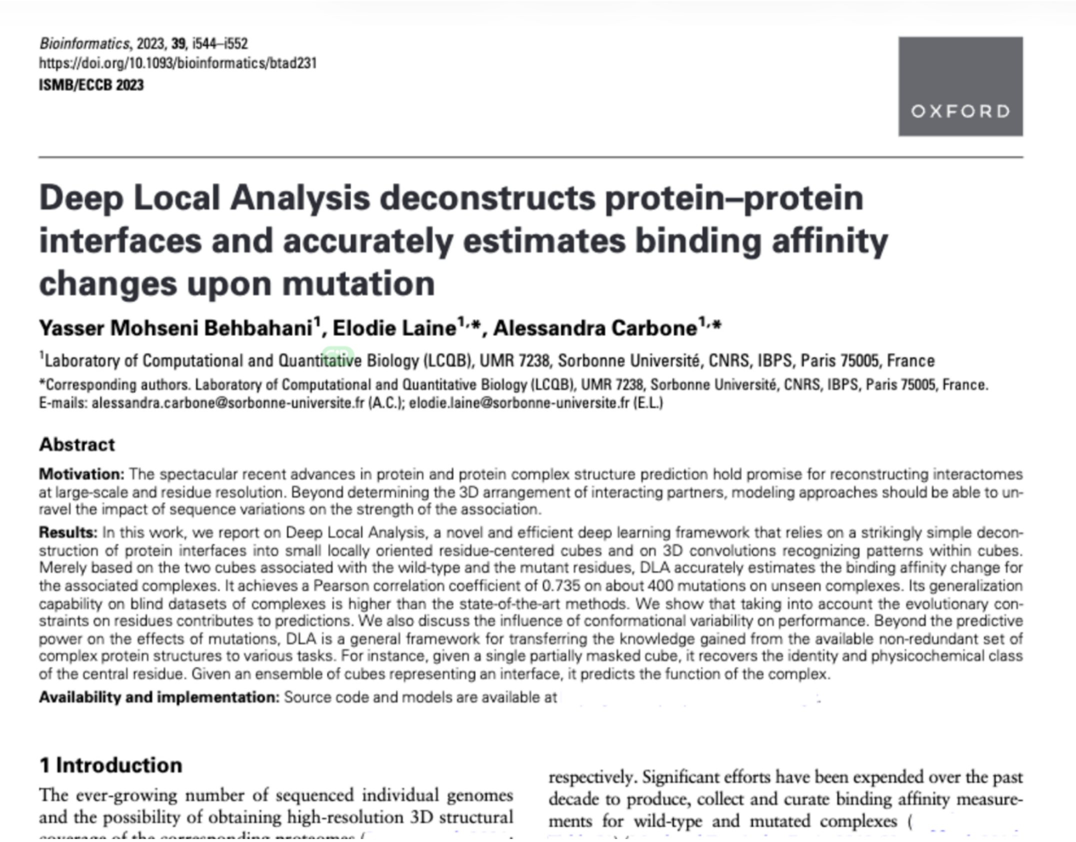
Recent developments

- ➢ Prediction of the effects of mutations on protein interactions


A
r_k
B
3D map
ssDLA
Projector #Channels: 20
Conv3D #Channels: 20 kernel size: 3^3
Conv3D #Channels: 30 kernel size: 4^3
Conv3D #Channels: 20 kernel size: 4^3
Ave pool kernel size: 4^3 strides: 4^3
Flatten 40% dropout
Fully connected #Neurons: 200 20% dropout
Fully connected #Neurons: 20 10% dropout
Fully connected #Neurons: 20
Softmax
A C D E F G H I K L M N P Q R S T V W Y
Probability vector
e_k
C
D
Fully connected #Neurons: 20
Fully connected #Neurons: 7 or 5
Softmax
ARO CAST PHOB POS POL-N GLY PRO
Probability vector
DLA-Mutation
Mutant
Wild type
Shared-weights
Subtraction of embedding vectors
Concat
Fully connected #Neurons: 1 Linear
$\Delta\Delta G_{Bind}$
a_k
Evolutionary conservation (Tjet)
Physico-chemical properties (PC)
Circular variance (CV)
RIM
CORE
SUPPORT
INTERIOR
SURFACE
Structural region (SR)
Mutational landscape of a monomeric protein (GEMME)

Recent developments

- Prediction of the effects of mutations on protein interactions



Bioinformatics, 2023, **39**, i544–i552
https://doi.org/10.1093/bioinformatics/btad231
ISMB/ECCB 2023

OXFORD

Deep Local Analysis deconstructs protein–protein interfaces and accurately estimates binding affinity changes upon mutation

Yasser Mohseni Behbahani[1], Elodie Laine[1,*], Alessandra Carbone[1,*]

[1]Laboratory of Computational and Quantitative Biology (LCQB), UMR 7238, Sorbonne Université, CNRS, IBPS, Paris 75005, France

*Corresponding authors. Laboratory of Computational and Quantitative Biology (LCQB), UMR 7238, Sorbonne Université, CNRS, IBPS, Paris 75005, France.
E-mails: alessandra.carbone@sorbonne-universite.fr (A.C.); elodie.laine@sorbonne-universite.fr (E.L.)

Abstract

Motivation: The spectacular recent advances in protein and protein complex structure prediction hold promise for reconstructing interactomes at large-scale and residue resolution. Beyond determining the 3D arrangement of interacting partners, modeling approaches should be able to unravel the impact of sequence variations on the strength of the association.

Results: In this work, we report on Deep Local Analysis, a novel and efficient deep learning framework that relies on a strikingly simple deconstruction of protein interfaces into small locally oriented residue-centered cubes and on 3D convolutions recognizing patterns within cubes. Merely based on the two cubes associated with the wild-type and the mutant residues, DLA accurately estimates the binding affinity change for the associated complexes. It achieves a Pearson correlation coefficient of 0.735 on about 400 mutations on unseen complexes. Its generalization capability on blind datasets of complexes is higher than the state-of-the-art methods. We show that taking into account the evolutionary constraints on residues contributes to predictions. We also discuss the influence of conformational variability on performance. Beyond the predictive power on the effects of mutations, DLA is a general framework for transferring the knowledge gained from the available non-redundant set of complex protein structures to various tasks. For instance, given a single partially masked cube, it recovers the identity and physicochemical class of the central residue. Given an ensemble of cubes representing an interface, it predicts the function of the complex.

Availability and implementation: Source code and models are available at

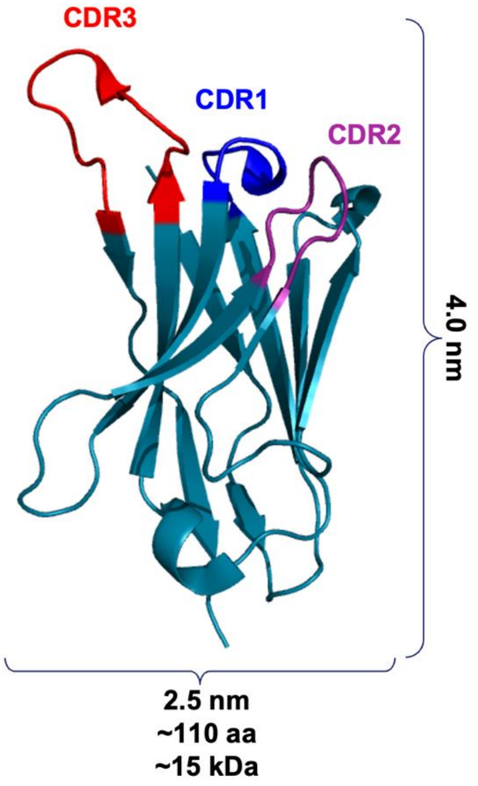
1 Introduction

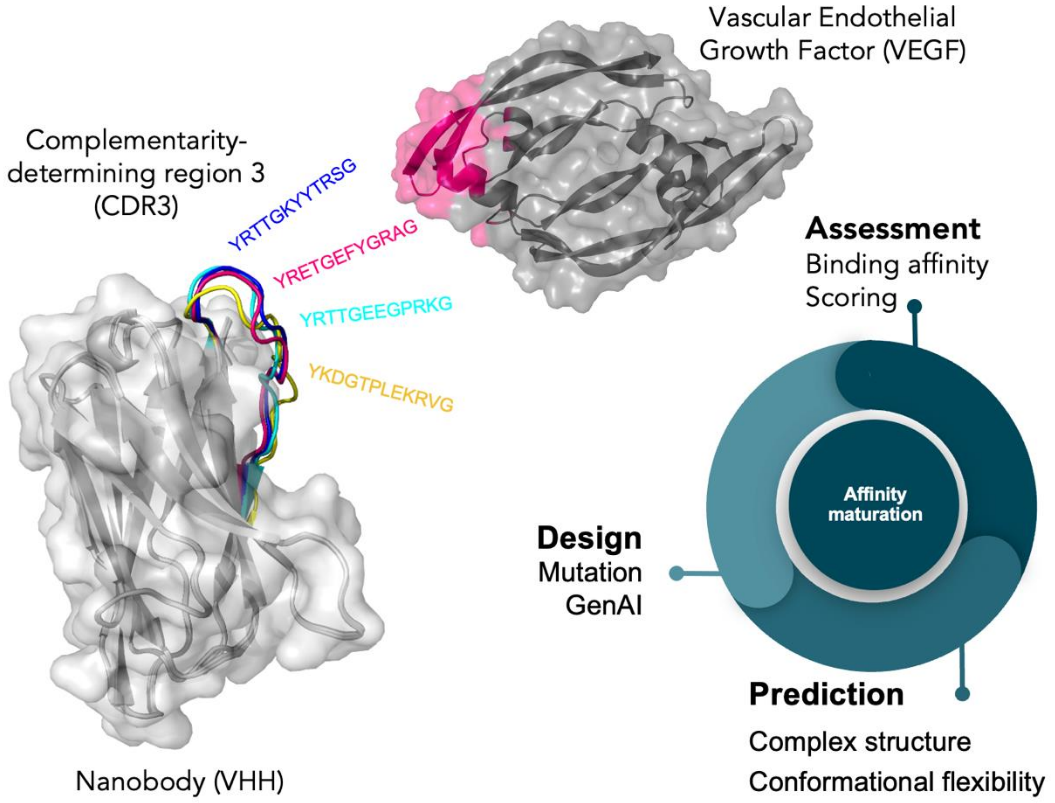
The ever-growing number of sequenced individual genomes and the possibility of obtaining high-resolution 3D structural coverage of the corresponding proteomes (

respectively. Significant efforts have been expended over the past decade to produce, collect and curate binding affinity measurements for wild-type and mutated complexes (

Recent developments

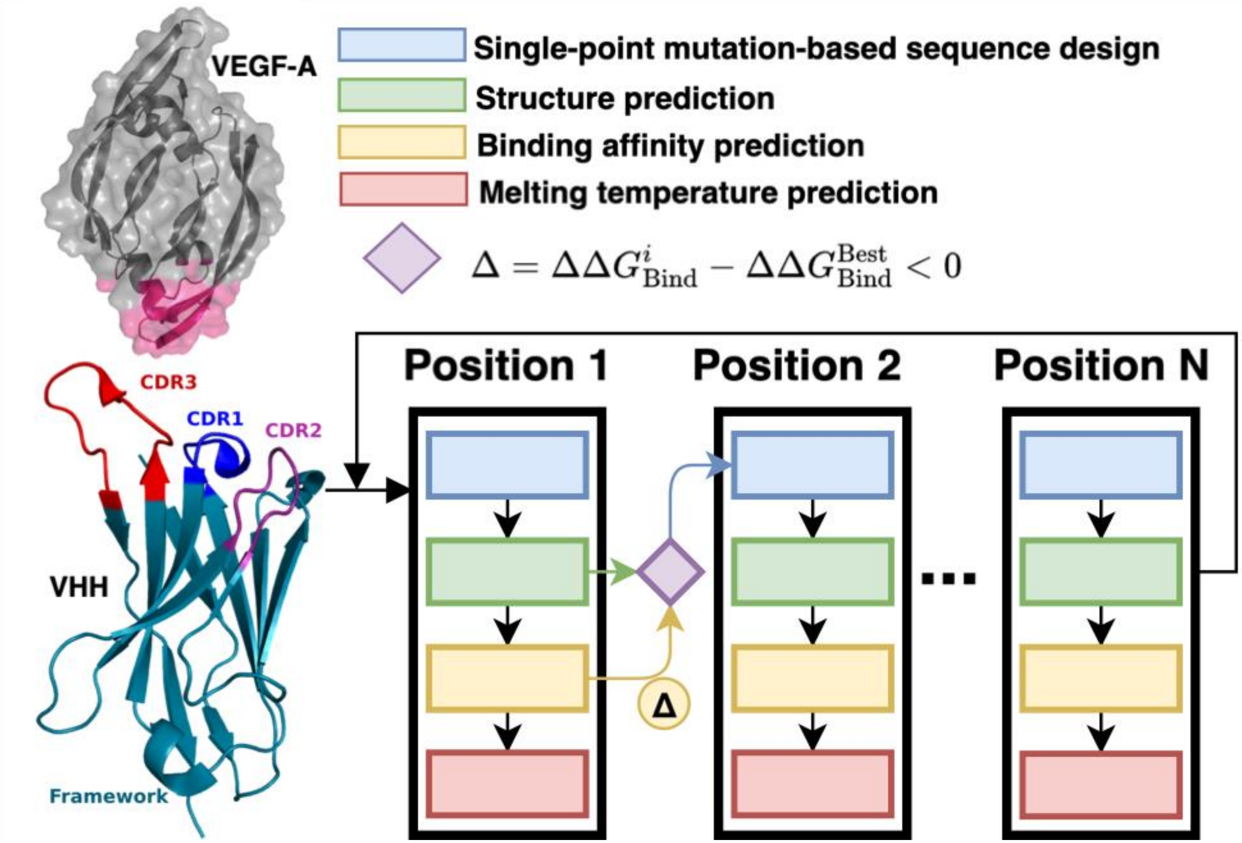
- Deep learning-based nanobody design





Recent developments

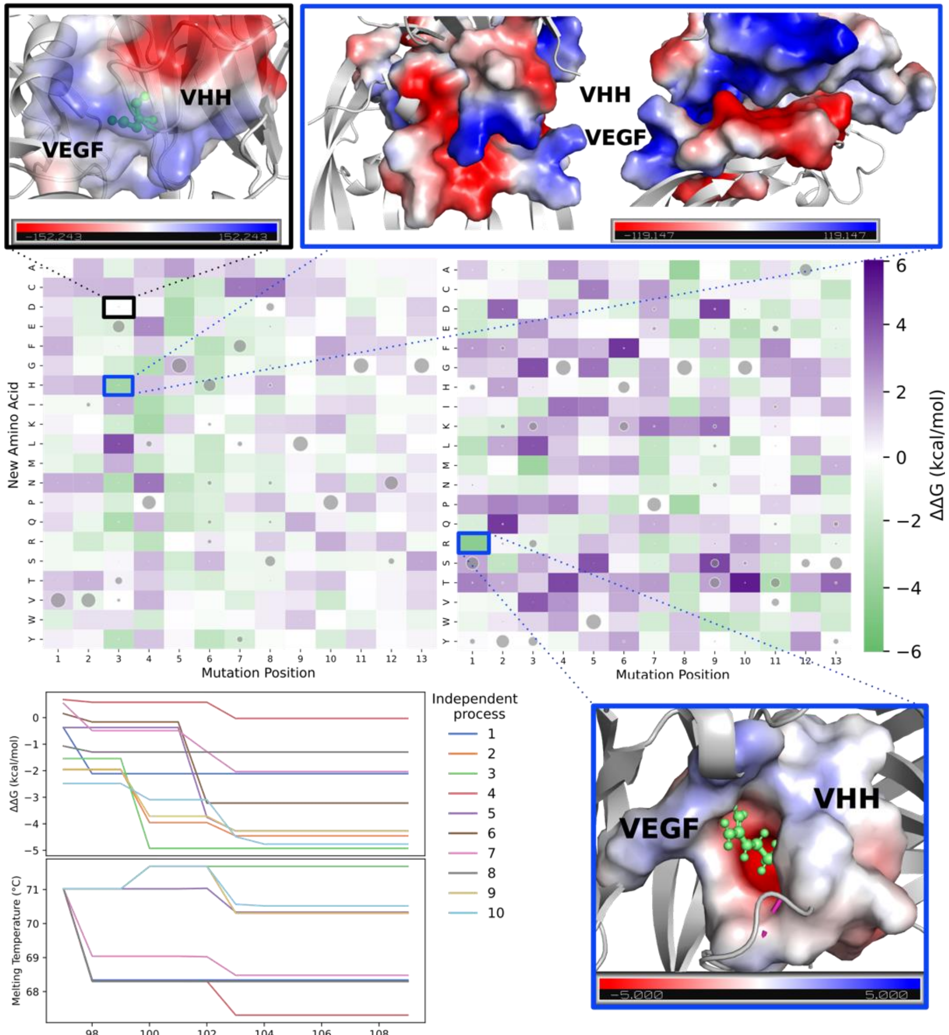
- ➢ Deep learning-based nanobody design

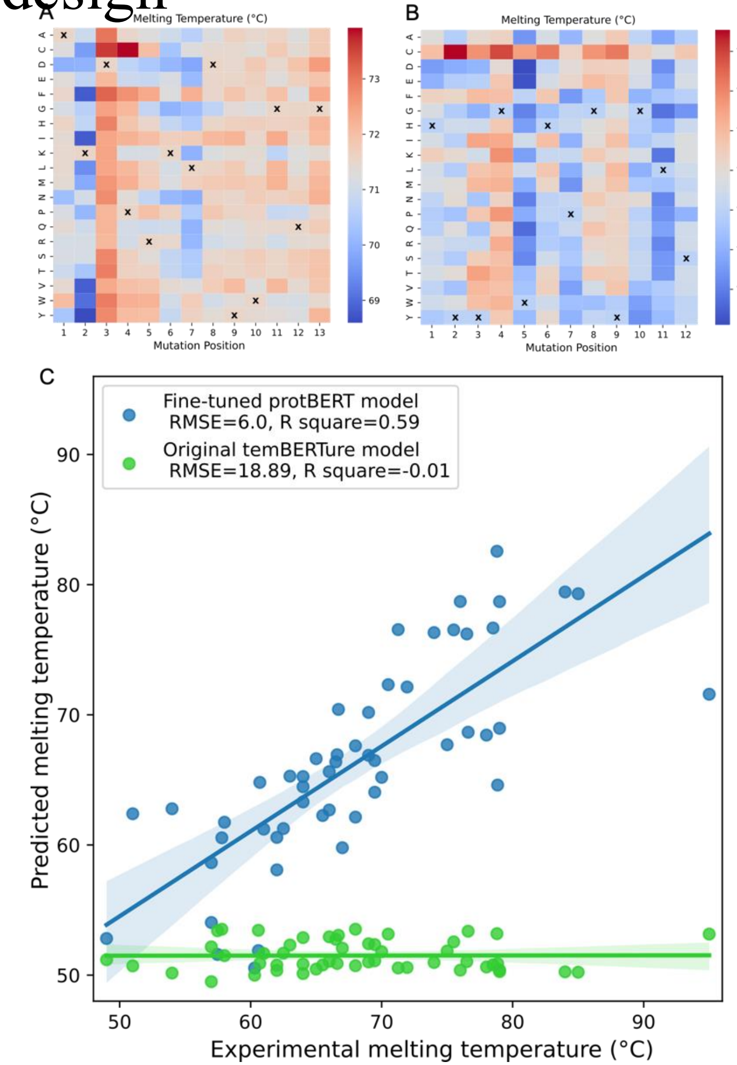

VEGF-A
Single-point mutation-based sequence design
Structure prediction
Binding affinity prediction
Melting temperature prediction
CDR3
CDR1
CDR2
VHH
Framework
Position 1
Position 2
Position N
...
Δ

$$\Delta = \Delta\Delta G^{i}_{\mathrm{Bind}} - \Delta\Delta G^{\mathrm{Best}}_{\mathrm{Bind}} < 0$$

Recent developments

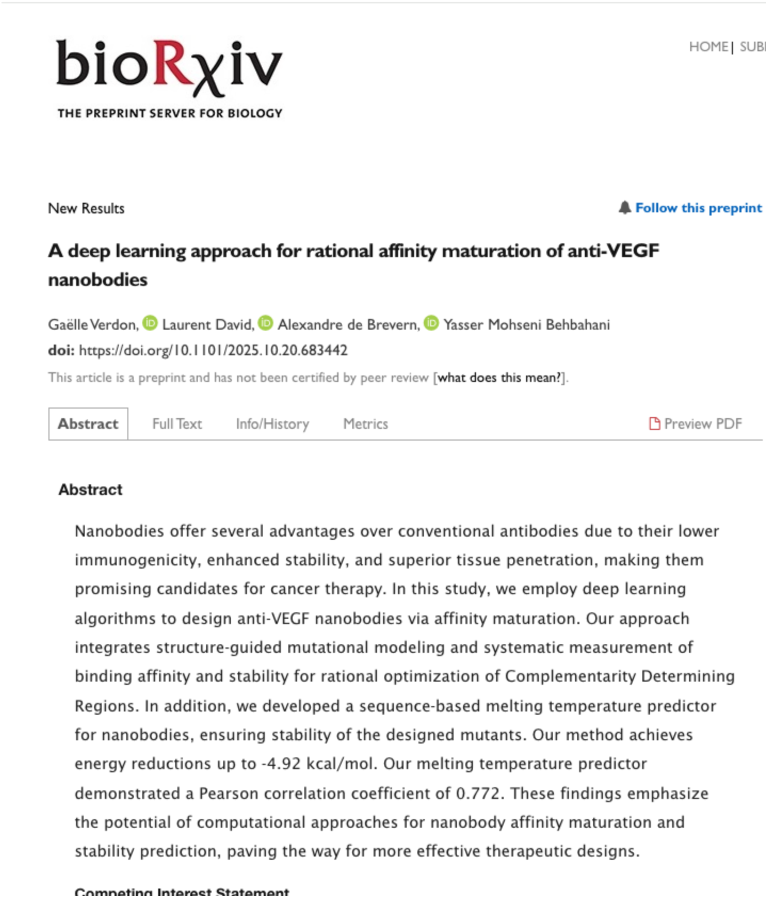
- Deep learning-based nanobody design





Recent developments

➢ Deep learning-based nanobody design



bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

A deep learning approach for rational affinity maturation of anti-VEGF nanobodies

Gaëlle Verdon, Laurent David, Alexandre de Brevern, Yasser Mohseni Behbahani

doi: https://doi.org/10.1101/2025.10.20.683442

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract

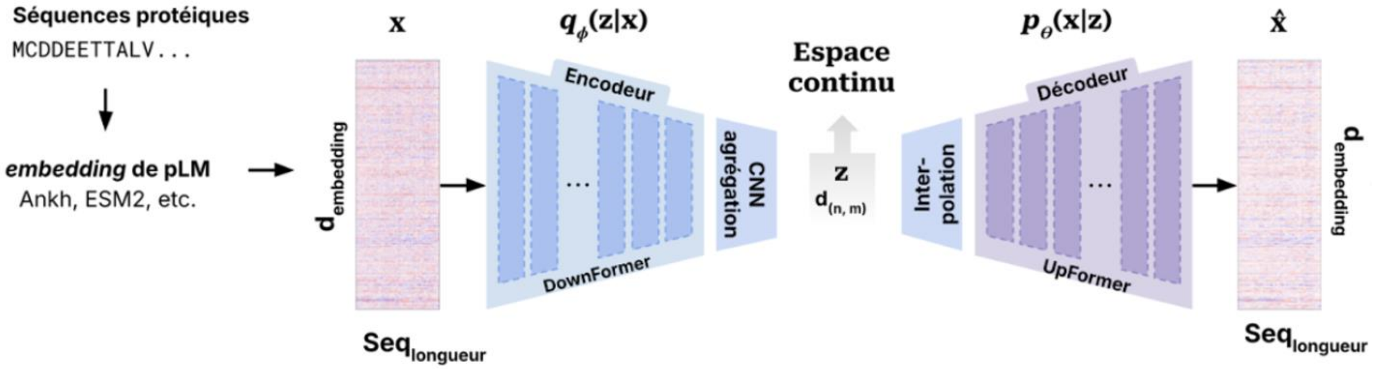
Nanobodies offer several advantages over conventional antibodies due to their lower immunogenicity, enhanced stability, and superior tissue penetration, making them promising candidates for cancer therapy. In this study, we employ deep learning algorithms to design anti-VEGF nanobodies via affinity maturation. Our approach integrates structure-guided mutational modeling and systematic measurement of binding affinity and stability for rational optimization of Complementarity Determining Regions. In addition, we developed a sequence-based melting temperature predictor for nanobodies, ensuring stability of the designed mutants. Our method achieves energy reductions up to -4.92 kcal/mol. Our melting temperature predictor demonstrated a Pearson correlation coefficient of 0.772. These findings emphasize the potential of computational approaches for nanobody affinity maturation and stability prediction, paving the way for more effective therapeutic designs.

Competing Interest Statement

Gaëlle Verdon

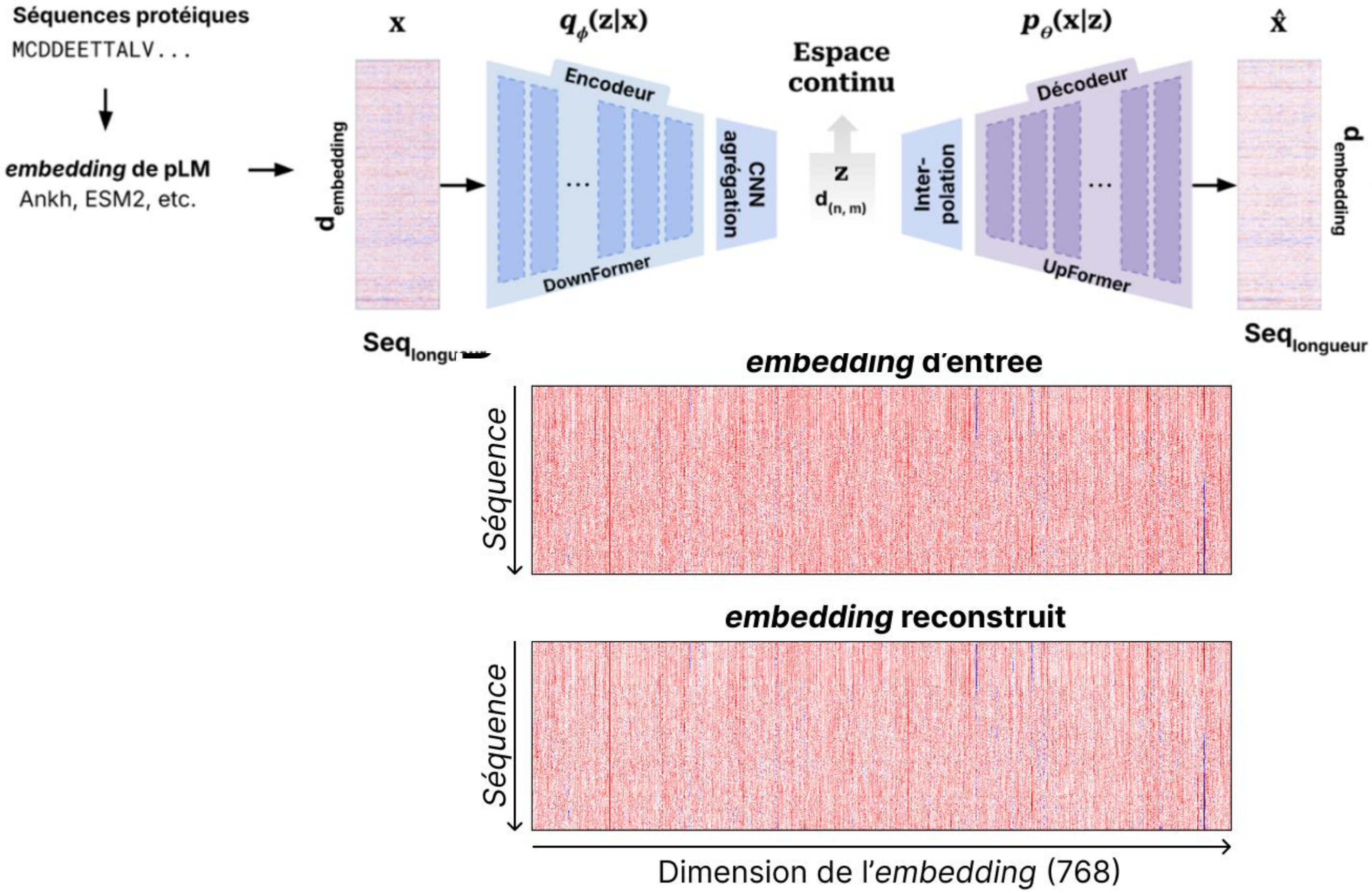
Recent developments

- Compress the embeddings:



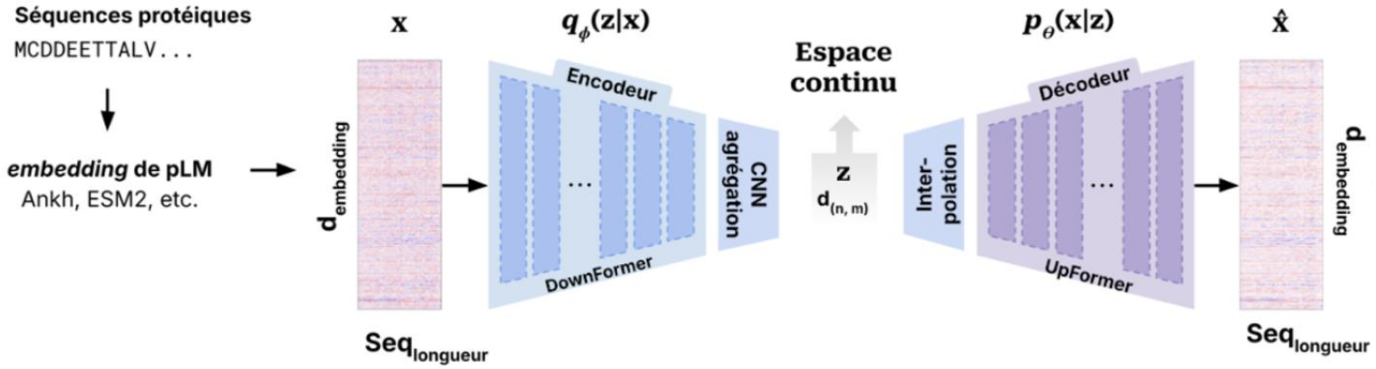
Recent developments

- Compress the embeddings: Faithful reconstruction despite x96 compression



Recent developments

- Compress the embeddings:



Compression down to x96, thus reducing storage costs and enabling learning for downstream task prediction.

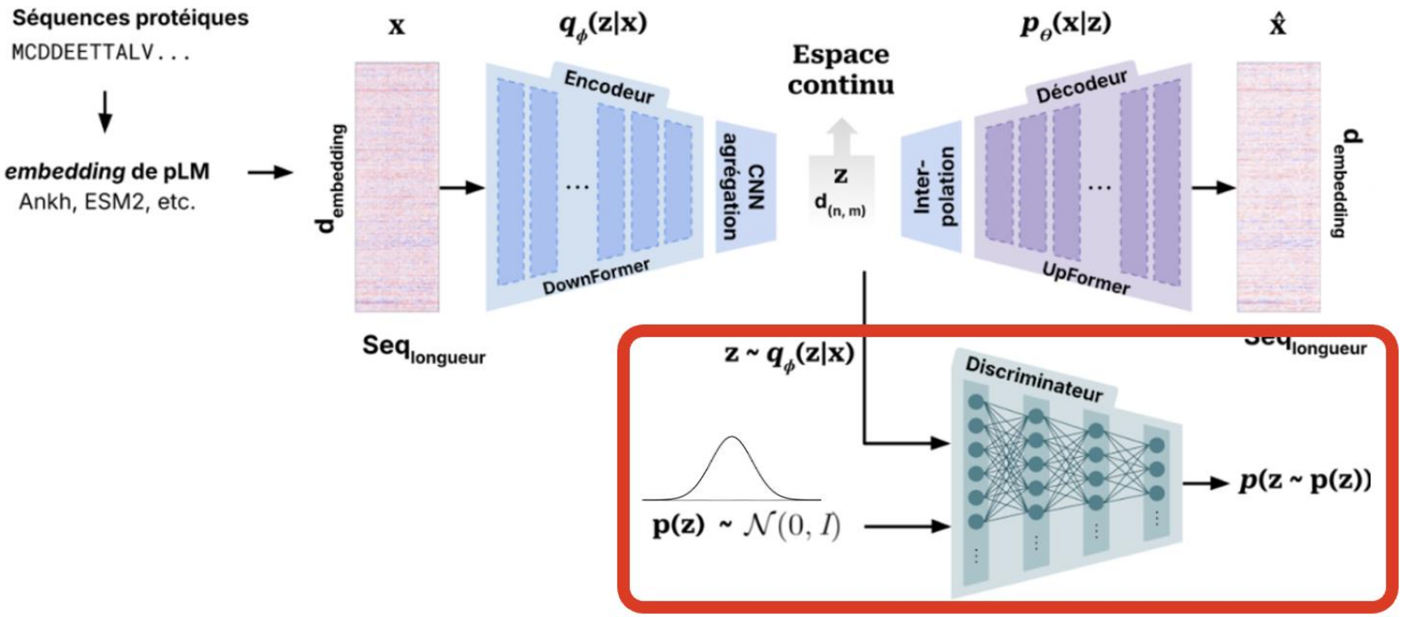
However, variable-sized compressed embeddings and latent space are still not continuous.

Recent developments

- Compress the embeddings: now x400 and more continuous

Compression to a fixed size of 64 x 32 (x400)

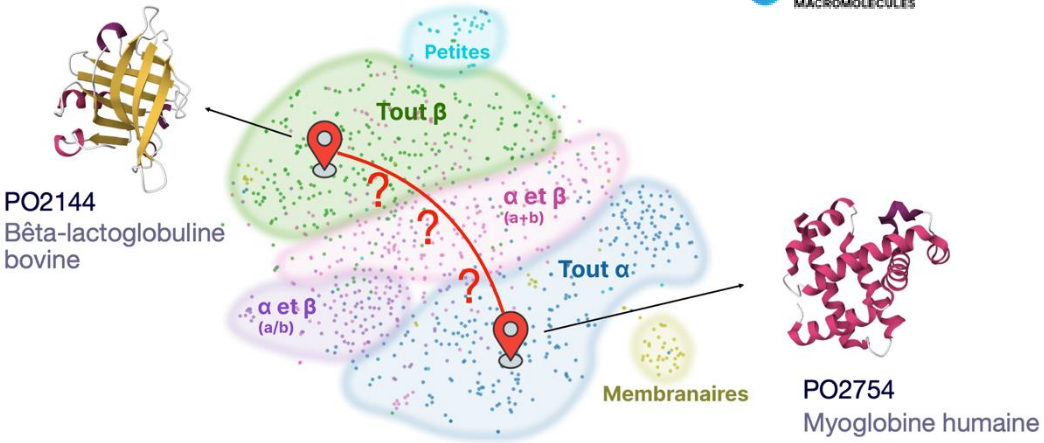
Sequence dimension Dimension **embedding**



Recent developments

- Compress the embeddings

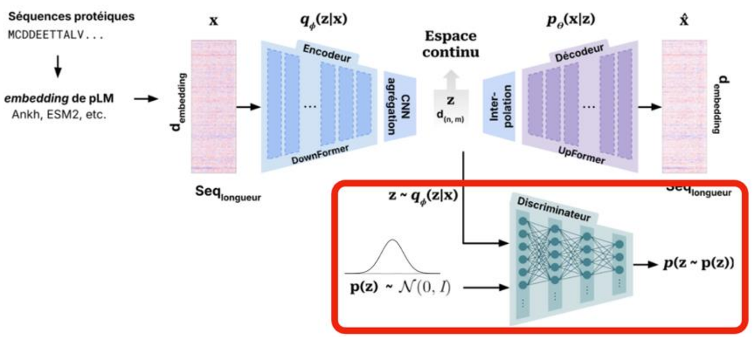
DSIMB
DYNAMICS OF STRUCTURES AND INTERACTIONS OF BIOLOGICAL MACROMOLECULES



Compression to a fixed size of 64 x 32 (x400)

Sequence dimension

Dimension **embedding**



6. CONCLUSION(S)

6. Conclusion(s)

- A methodological revolution

6. Conclusion(s)

- A methodological revolution
- An evolution in the field of Structural Bioinformatics

6. Conclusion(s)

- A methodological revolution
- An evolution in the field of Structural Bioinformatics
- Data, data… is always the most important

6. Conclusion(s)

- A methodological revolution
- An evolution in the field of Structural Bioinformatics
- Data, data… is always the most important
- Very different types of approaches, needs to be properly defined (size of the dataset)

6. Conclusion(s)

- A methodological revolution
- An evolution in the field of Structural Bioinformatics
- Data, data… is always the most important
- Very different types of approaches, needs to be properly defined (size of the dataset)
- Black boxes, no explanation

6. Conclusion(s)

- A methodological revolution
- An evolution in the field of Structural Bioinformatics
- Data, data… is always the most important
- Very different types of approaches, needs to be properly defined (size of the dataset)
- Black boxes, no explanation
- Lot of ‘experts’, difficult to asses all the new approaches and papers (example of protein design)

Pr. C. Etchebest
Pr. J.-C. Gelly
Pr. F. Cadet
Dr. J. Kozelka
Dr. F. Gardebien
Dr. Ph. Charton
Dr. Y. Mohseni Behbahani
Dr. J. Diharce
Dr. T. Galochkina
Dr. F. Guyon
Dr. G. Cretin

Dr. R. Radjasandirane
Pr. M. Ostuni (BIGR)

Thank you