Do Newly Born Orphan Proteins Resemble Never Born Proteins? A Study using Deep Learning Algorithms

Jing Liu[1,2*], Rongqing Yuan[3,*], Wei Shao[4], Jitong Wang[3], Israel Silman[5] and Joel L. Sussman[6]

[1]Department of Biotechnology and Food Engineering, Guangdong Technion-Israel Institute of Technology, Shantou, 535063, China

[2]Faculty of Biotechnology and Food Engineering, Technion-Israel Institute of Technology, Haifa, 32000, Israel

[3]Department of Chemistry, Tsinghua University, Beijing 100084, China

[4]School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

[5]Department of Brain Sciences, The Weizmann Institute of Science, Rehovot 7610001, Israel

[6]Department of Chemical and Structural Biology, The Weizmann Institute of Science, Rehovot 7610001, Israel

[*]Equal contribution

Authors for correspondence Israel.Silman@weizmann.ac.il, Orchid - 0000-0003-1923-0829; Joel.Sussman@weizmann.ac.il, Orchid - 0000-0003-0306-3878.

Running title: New Born and Never Born Proteins

Keywords: AlphaFold2; RoseTTAFold; orphan protein; molten globule; intrinsically disordered protein; taxonomic restriction

Conflict of interest: The authors declare no conflicts of interest.

Data Availability Statement: All data are presented in the paper, on in pointers to UniProt (https://www.uniprot.org) or the PDB (https://www.rcsb.org).

Acknowledgement: The Israel tutors and the Chinese students are all grateful to the YutChun-Weizmann Program that enabled this study.

ABSTRACT

Newly Born proteins, devoid of detectable homology to any other proteins, known as orphan proteins, occur in a single species or within a taxonomically restricted gene family. They are generated by expression of novel Open Reading Frames, and appear throughout evolution. We used the recently developed programs for predicting protein structures, RoseTTAFold and AlphaFold2, to compare such *Newly Born* proteins to random polypeptides generated by shuffling sequences of native proteins, which have been called '*Never Born*' proteins. The two programs were used to compare the structures of two sets of four *Never Born* proteins, one set that had been expressed and shown to be intrinsically disordered, and a second set that had been shown experimentally to possess substantial secondary structure. Since the programs rely to a large extent on multisequence alignment, the models generated were scored as being of low quality. However, a significant pattern emerged when the models generated by RoseTTAFold were examined. Specifically, all four members of Group 1 were shown to be very extended, as would be expected for intrinsically disordered proteins. In contrast, all four members of Group 2 appeared to be compact, and possessed substantial secondary structure. As a further control, both programs predicted unfolded structures for three well characterized intrinsically disordered proteins. The two programs were used to predict the structures of two orphan proteins whose crystal structures have been solved, both of which display novel folds. RoseTTAFold predicted both structures very well, whereas AlphaFold2 predicted only one well. The two programs were used to predict the structures of five orphan proteins with well-identified biological functions, one of which is predicted to be intrinsically disordered, and four to be folded. Both programs displayed the intrinsically disordered protein as an unfolded structure. RoseTTAFold displayed all four of those predicted to be folded as compact folded structures, with apparent novel folds, as determined by Dali and Foldseek. It is plausible that new biological functions may be implemented by orphan proteins due to their novel folds.

1. INTRODUCTION

The accepted view is that protein sequences have evolved so as to incorporate the features required for optimal folding and function[1]. Specific amino acid or oligopeptide patterns appear to yield insights into phylogenetic differences between the three kingdoms: prokaryotes, archaea, and eukaryotes[2]. Surprisingly, however, evidence has been presented that 'from a sequence similarity perspective, *real unrelated* proteins are indistinguishable from random amino acid sequences'[3], This, at first sight, seems to be counterintuitive, because it might be anticipated that natural sequences would differ from random sequences in their folding characteristics. Indeed, this conclusion has been challenged[4].

Two studies have shown that the sequences obtained by random shuffling of native protein sequences can be expressed, and that in many cases the expressed polypeptide chains of these *'Never Born'* proteins[5] fold in aqueous solution into compact structures that display resistance to proteolysis[6], or substantial secondary structure elements[7]. For earlier studies using randomized sequences see[8-10].

Recently, more than 129 random sequences, of 100 amino acids each, were used to generate 3D models via the RoseTTAFold tool. These initial models were optimized by Monte Carlo sampling in amino acid sequence space to yield "novel proteins spanning a wide range of sequences and predicted structures"[11]. These proteins were then expressed in *E coli*; 27 of them yielded monodisperse species with circular dichroism spectra consistent with a native structure. The 3D structures of three of them were determined, and all three displayed novel folds.

The folding propensity of random sequences is of relevance in the context of the issue of orphan genes, and of the proteins that they express, *viz.*, *'Newly Born'* proteins, devoid of detectable homology to any other proteins, which occur in a single species or within a taxonomically restricted gene (TRG) family[12-18]. Such a possibility was considered to be impossible by François Jacob[19]. However, to quote – "The origin of novel protein-coding genes was once considered so improbable as to be impossible. In the last decade, and especially in the last five years, this view has been overturned by extensive evidence from diverse eukaryotic lineages"[15]. Both the term orphan gene, and the term orphan protein, are often loosely used in more than one context. In the present study, the term orphan gene refers to a gene for which evidence has been presented that it has arisen from what was previously a non-coding DNA sequence, and is expressed as an open reading frame (ORF). Thus, the orphan protein for which it codes is seen only in a single species or in one that is closely related taxonomically, *i.e.*, in a TRG family. The general contention is thus that new genes may appear out of previously non-coding genomic regions, a process known also as exonization[20], and code for novel protein sequences[21]. The question that then arises is how new functional protein domains might evolve out of such random sequences[12]? It is fair to say that this is still an open question, and that much more experimental data are required. Of particular interest are studies of higher primates in which novel genes were identified that are shared by chimpanzees, gorillas and humans, whereas other *de novo* genes may, for example, be restricted to humans[22-27]. It should further be mentioned that it has been suggested that novel protein sequences may also be generated by ORFs present in long non-coding RNAs (lncRNAs)[13,28].

Very recently, the field of structural biology has undergone a revolution due to the development of the deep-learning-based protein structure prediction programs, AlphaFold2 (AF2)[29] and RoseTTAFold (RTF)[30]. Both these algorithms have been shown to predict 3D structures for many natural sequences that closely resemble the experimental structures deposited in the PDB[31-33]. The major breakthrough yielding higher quality compared to previous methods was the novel way in which multisequence analysis was employed[29]. We were curious as to whether these powerful new tools for protein structure prediction might be of value for distinguishing natural from random sequences. Here, we use these AI programs to predict the structures of several natural sequences, and of random sequences generated by shuffling them, as well as those of '*Newly Born*' orphan proteins. They are also used to predict the structures of the random sequences of '*Never Born*' proteins expressed by Tretyachenko *et al.*[7], some of which these authors had shown to fold into compact structures, others to belong to the category of intrinsically disordered proteins (IDPs)[34].

2. MATERIALS AND METHODS

2.1. Protein Sequences

Protein sequences for crystal structures were retrieved from the PDB (https://www.rcsb.org), for the three IDPs, sequences from UniProt (https://www.uniprot.org), for the eight '*Never Born*' proteins, from the supplementary information associated with the study of Tretyachenko *et al.*[7] (https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-017-15635-8/MediaObjects/41598_2017_15635_MOESM1_ESM.pdf) and for the '*Newly Born*' proteins from UniProt (https://www.uniprot.org). All of these sequences are listed in Table 1.

Table 1

Category/Protein	**Amino Acid Sequence**
Crystal Structures	
Human carbonic anhydraese	>6PEA Carbonic anhydrase 2 Homo sapiens (9606) MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPLSVSYDQATSLRILNNGHAFNVEFDDSQDKAVLKGGPLDGTYRLIQFHFHWGSLDGQGSEHTVDKKKYAAELHLVHWNTKYGDFGKAVQQPDGLAVLGIFLKVGSAKPGLQKVVDVLDSIKTKGKSADFTNFDPRGLLPESLDYWTYPGSLTTPPLLECVTWIVLKEPISVSSEQVLKFRKLNFNGEGEPEELMVDNWRPAQPLKNRQIKASFK
E. coli adenine phosphoribosyltransferase	>2DY0 Adenine phosphoribosyltransferase E. coli K12 (83333) GSSGSSGMTATAQQLEYLKNSIKSIQDYPKPGILFRDVTSLLEDPKAYALSIDLLVERYKNAGITKVVGTEARGFLFGAPVALGLGVGFVPVRKPGKLPRETISETYDLEYGTDQLEIHVDAIKPGDKVLVVDDLLATGGTIEATVKLIRRLGGEVADAAFIINLFDLGGEQRLEKQGITSYSLVPFPGH
S. cerevisiae ribosome anti- association factor EIF6	>1G62 RIBOSOME ANTI-ASSOC FACTOR EIF6 S. cerevisiae (4932) MATRTQFENSNEIGVFSKLTNTYCLVAVGGSENFYSAFEAELGDAIPIVHTTIAGTRIIGRMTAGNRRGLLVPTQTTDQELQHLRNSLPDSVKIQRVEERLSALGNVICCNDYVALVHPDIDRETEELISDVLGVEVFRQTISGNILVGSYCSLSNQGGLVHPQTSVQDQEELSSLLQVPLVAGTVNRGSSVVGAGMVVNDYLAVTGLDTTAPELSVIESIFRL
IDPs	
Drosophila gliotactin cytoplasmic domain	>Q7KT70 Gli-Cyt C-Term 207aa of UniProt Q7KT70 RNAKRQSDRFYDEDVFINGEGLEPEQDTRGVDNAHMVTNHHALRSRDNIYEYRDSPSTKTLASKAHTDTTSLRSPSSLAMTQKSSSQASLKSGISLKETNGHLVKQSERAATPRSQQNGSIAKVASPPVEEKRLLQPLSSTPVTQLQAEPAKRVPTAASVSGSSRSTTPVPSARSTTTHTTTATLSSQPAAQPRRTHLVEGVPQTSV
human CDN1C-Cyclin-dependent kinase inhibitor	>spP49918 CDN1C_HUMAN Cyclin-dependent kinase inhibitor 1 MSDASLRSTSTMERLVARGTFPVLVRTSACRSLFGPVDHEELSRELQARLAELNAEDQNRWDYDFQQDMPLRGPGRLQWTEVDSDSVPAFYRETVQVGRCRLLLAPRPVAVAVAVSPPLEPAAESLDGLEEAPEQLPSVPVPAPASTPPPVPVLAPAPAPAPAPVAAPVAAPVAVAVLAPAPAPAPAPAPAPAPVAAPAPAPAPAPAPAPAPAPAPDAAPQESAEQGANQGQRGQEPLADQLHSGISGRPAAGTAAASANGAAIKKLSGPLISDFFAKRKRSAPEKSSGDVPAPCPSPSAAPGVGSVEQTPRKRLR
human osteopnotin	>spP10451 OSTP_HUMAN Osteopontin OX=9606 GN=SPP1 PE=1 SV=1 MRIAVICFCLLGITCAIPVKQADSGSSEEKQLYNKYPDAVATWLNPDPSQKQNLLAPQNAVSSEETNDFKQETLPSKSNESHDHMDDMDDEDDDDHVDSQDSIDSNDSDDVDDTDDSHQSDESHHSDESDELVTDFPTDLPATEVFTPVVPTVDTYDGRGDSVVYGLRSKSKKFRRPDIQYPDATDEDITSHMESEELNGAYKAIPVAQDLNAPSDWDSRGKDSYETSQLDDQSAETHSHKQSRLYKRKANDESNEHSDVIDSQELSKVSREFHSHEFHSHEDMLVVDPKSKEEDKHLKFRISHELDSASSEVN
'Never Born' Proteins - Group 1	
1856	>1856 109aa MYQIEKADFTFDVRRRTAATDIENHAFNMVWLQSWCDVSIIKRTLDAYDEAYDAAFQRLKPAEWAIDDWVASIQRRRRHYVAYNLSKIKLPVRLEKLSGTTLEHHHHHH
6387	>6387 109aa MIEHCYSKTVYYNLEQEKYDLEVTHIEGWMRAGRKDLADNLLEDSGHVFIPEVALQENHYREVHAKIGDAEMRVYKRELFPEPQIVEVLETPSQLFFAEIELEHHHHHH
4090	>4090 109aa MVERDKPPNIWVYDAEPLQQGGIVWVHLAALYCANVDDYAPQDHLDITMYGFDHQKTNILSFEDESVNAQSYWQYGIIFVKSHWGEDLQGAIVESWRDSRSLEHHHHHH
2298	>2298 109aa MKWYGRGREDFGSPDVDVEKNCEGEVIYGTSQELYSNVVFDWWAGISEQPTIFIGSLTTPNTKDDMLWYRNDAKNPGHSILYNLINDYWEATEVSGIGNVVLEHHHHHH
'Never Born' Proteins - Group 3	
665	>665 109aa MATKGADHGLAAPQPHAKWDTQIPAEGADREHRSGGGNERRFYNEGAKHAQATWAIPDEPAFHLQPAVGEGATTDQAGSLEDQWVRSLNNNDVDPTQADETLEHHHHHH
8667	>8667 109aa MPQSLACAGTATRESQNDQLLDGHQPETYLDRMFPEELDEIDGVIPNYDMEWGKKSGLEV

	SEKCFDPWFNYPTYEETPSDMSGPFKNALMKYRPTQNARPDLEHHHHHH
3703	>3703 109aa MSLYKFGQRRAVDPLPRQCQRDKDYDAFIGGEQNCDNELSKSFPIVVMSVFLYDPTYNVDSEAQDNKLDHHGSEPTHGDTPTTSEDTRPGSDRVMRDVPQTLEHHHHHH
933	>933 109aa MVREIDDKTISDYLARGADEGTTAYSLKIPTDKCLFAPTKKHLHGGDKSQEADPPTKSPMVEHQFGHEPDFPSCREPEDYPGSPLVELTGLNRLTQEPNEELEHHHHHH
Orphan protein crystal structures	
T. maritima TM0875	>1O22 orphan protein TM0875 T. maritima (2336) MGSDKIHHHHHHMRLMDILEILYYKKGKEFGILEKKMKEIFNETGVSLEPVNSELIGRIFLKISVLEEGEEVPSFAIKALTPKENAVDLPLGDWTDLKNVFVEEIDYLDSYGDMKILSEKNWYKIYVPYSSVKKKNRNELVEEFMKYFFESKGWNPGEYTFSVQEIDNLF
H. influenzae Hypothetical protein HI1480	**>**1MW5 HYPOTHETICAL PROTEIN HI1480 H. influenzae (727) GSHMSETDLLMKMVRQPVKLYSVATLFHEFSEVITKLEHSVQKEPTSLLSEENWHKQFLKFAQALPAHGSASWLNLDDALQAVVGNSRSAFLHQLIAKLKSRHLQVLELNKIGSEPLDLSNLPAPFYVLLPESFAARITLLVQDKALPYVRVSMEYWHALEYKGELNDPAANKARKEAELAAATAEQ
'Newly Born' Proteins - well characterized	
Human PBOV1- tumor-specific gene	>spQ9GZY1 PBOV1_HUMAN Prostate & breast cancer overexpressed MRAFLRNQKYEDMHNIIHILQIRKLRHRLSNFPRLPGILAPETVLLPFCYKVFRKKEKVKRSQKATEFIDYSIEQSHHAILTPLQTHLTMKGSSMKCSSLSSEAILFTLTLQLTQTLGLECCLLYLSKTIHPQII
Human FLJ33706 expressed in neurons (alt gene name C20orf203)	>spQ8NBC4 CT203_HUMAN Uncharacterized protein C20orf203 MFPRPVLNSRAQAILLPQPPNMLDHRQWPPRLASFPFTKTGMLSRATSVLAGLTAHLWDLGGGAGRRTSKAQRVHPQPSHQRQPPPPQHPGPYQERIWVGGEGWGEVGGLRLSKVGRRDREVGRGLRAPAGRGRAMGGMPRMGTVGDFGQALSSLAWTSTCFQDFCLPSLPGKLPAPLISKQQFLSNSSRSLFN
Human NCYM - DNA binding transcriptional activator homolog	>spP40205 NCYM_HUMAN N-cym protein MQHPPCEPGNCLSLKEKKITEGSGGVCWGGETDASNPAPALTACCAAEREANVEQGLAGRLLLCNYERRVVRRCKIAGRGRAPLGTRPLDVSSFKLKEEGRPPCLKINK
Mouse Gm13030 involved in regulating the pregnancy cycle	trA2APQ6 A2APQ6_MOUSE Predicted gene 13030 MCRFHLLQAIKPPEKQMEQKSSALGSIMKLSQSHATETTWVLPSQGLRDYLLHPACFHHFRKEGRPDCRPANMIYGFDKTHPRRCCTDLLFQPRLLMLSRVLGPEQLQELLQIPDDLTSPSLSYGSNQNLSQALNFPKHVHTG
Wheat - TaFROG - an IDP	>trA0A0K1YY56 A0A0K1YY56_WHEAT Fusarium resistance orphan MVWSTSKQQGGEREESKQHKMVKEVKTPIFTHQLSFHSLPLNKVKNIEVDRLRLSFTTPKNSTLVPVDSGSDEESDEDRGCSDIDSNKPMDEGLDHICSGLHAIPRKNKARSAKKRSHKISSRKFYKIFS

2.2. AlphaFold2 Predictions

The protein AlphaFold2 predictions were performed by the AlphaFold2_advanced Colab[35] at https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/beta/AlphaFold2_advanced.ipynb. The defaults used were:

- multisequence alignment, mmseq2,
- template protein structures were not used.

2.3. RoseTTAFold predictions

The program RoseTTAFold predictions were performed by the Robetta server using the RoseTTAFold option[30] at: https://robetta.bakerlab.org. The defaults for this server were employed.

2.4. Natural protein sequences

We selected for this study three native proteins whose crystal structures had been experimentally determined. The sequence information for these proteins was obtained from the UniProt database (https://www.uniprot.org), and the crystal structures from the PDB [https://www.rcsb.org]. (1) Carbonic anhydrase (Human) (P00918, 6pea); (2) Ribosome anti-association factor EIF6 (*Saccharomyces cerevisiae*) (Q12522, 1g62); (3) Adenine phosphoribosyltransferase from *E. coli* (P69503, 2dyo).

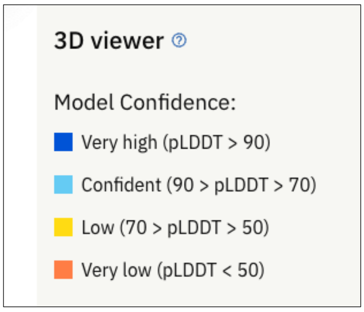
2.5. Randomized sequence generation

Three to five randomized sequences were generated from each natural sequence, maintaining the original amino acid composition of each protein, using the tool for Scrambling Protein or Peptide Sequences (https://peptidenexus.com/article/sequence-scrambler).

2.6. Structure prediction and comparison

Each of the natural and randomized sequences was submitted to the AlphaFold2 Colab (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/beta/AlphaFold2_advanced.ipynb) and RTF (https://robetta.bakerlab.org) servers for structure prediction. For each sequence, five most probable models were predicted by each server. The predicted structures for natural sequences were aligned with the experimental structures using PyMol [The PyMOL Molecular Graphics System, Version 2.1 ATI-4.8.101, Schrödinger, LLC]. The models of each of the randomized sequences were aligned with the model with the highest rank predicted by AF2, or with the number 1 model predicted by RTF.

AF2 produces a per-residue estimate of its confidence on a scale of 0-100. This confidence measure is called pLDDT, and corresponds to the model's predicted score on the IDDT-Cα metric[36]. It is stored in the B-factor fields of the PDB files available. pLDDT is also used to color-code the residues of the model in the 3D structure viewer (Fig. 1).



3D viewer

Model Confidence:

- 🟦 Very high (pLDDT > 90)
- 🟦 Confident (90 > pLDDT > 70)
- 🟨 Low (70 > pLDDT > 50)
- 🟧 Very low (pLDDT < 50)

Fig. 1 Color coding scheme based on the PDBe AlphaFold Database [https://alphafold.ebi.ac.uk]

For models predicted by RTF, the RMSD values are inserted in place of B-factors in the PDB files generated. These were converted into pLDDT values using a Python program that we wrote, based on the formulae described by[37] and as seen at https://phenix-online.org/version_docs/dev-4380/reference/process_predicted_model.html.

$$\mathrm{RMSD} = 1.5 * \mathrm{exp}(4*(0.7-\mathrm{LDDT}/100)) \qquad [1]$$

$$\mathrm{LDDT} = 100*((7-(\ln(\mathrm{RMSD}) - \ln\ (1.5))/4)) \qquad [2]$$

To keep the coloring scheme consistent for all structures shown, B-factors, in the original PDB files for 3D experimentally determined structures, were converted to pLDDT using the following equations:

$$\mathrm{B} = (\mathrm{rmsd}^2)*((8*(\pi^2))/3.0) \qquad [3]$$

$$\mathrm{rmsd} = \mathrm{sqrt}((3*\mathrm{B}/(8*\pi^2)) \qquad [4]$$

$$\mathrm{LDDT} = 100 * ((7- (\ln\ (\mathrm{sqrt}((3*\mathrm{B}/(8*\mathrm{p}^2))) - \ln\ (1.5))/4)) \quad [5]$$

For all structures, we then applied the color scheme used in the PDBe AlphaFold2 database. [https://alphafold.ebi.ac.uk] shown in Fig. 1, via the *coloraf.py* plugin for PyMol (https://github.com/cbalbin-bio/pymol-color-alphafold).

2.7. Detection of novel and unique folds

Dali [38,39] (at http://www2.ebi.ac.uk/dali) and Foldseek[40] (at https://search.foldseek.com/search) were used to check if the folds were novel.

2.8. Morphs for the 5 top models

In order to help compare the 5 top models from AF2 and RTF, a morph was generated via PyMol after the 5 top models had been aligned on top of each other via the PyMol align command.

2.9. Prediction of intrinsically disordered regions

The prediction of intrinsically disordered regions was carried out by use of (1) FoldIndex [41], using: https://fold.proteopedia.org/cgi-bin/findex]; (2) NetSurfP3[42], using: https://services.healthtech.dtu.dk/service.php?NetSurfP-3.0]; (3) IUPRED3[43], using: https://iupred.elte.hu.

2.10. Amino Acid Compositions

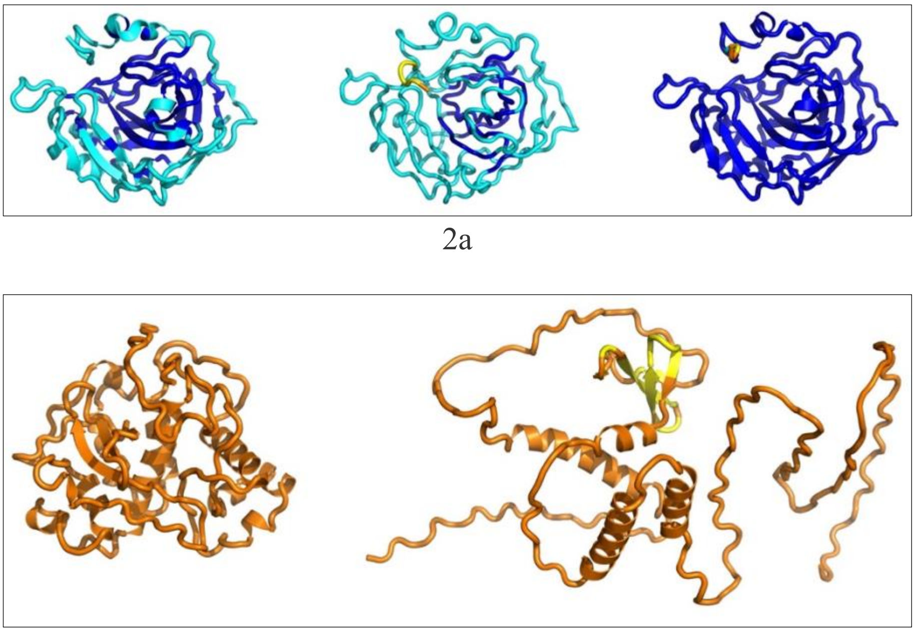
Amino acid compositions were calculated using the Expassy ProtParam tool [https://web.expasy.org/protparam].

2.11. BLASTP Sequence Search

BLASTP sequence searches were done using the NIH-NLM site [https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins].

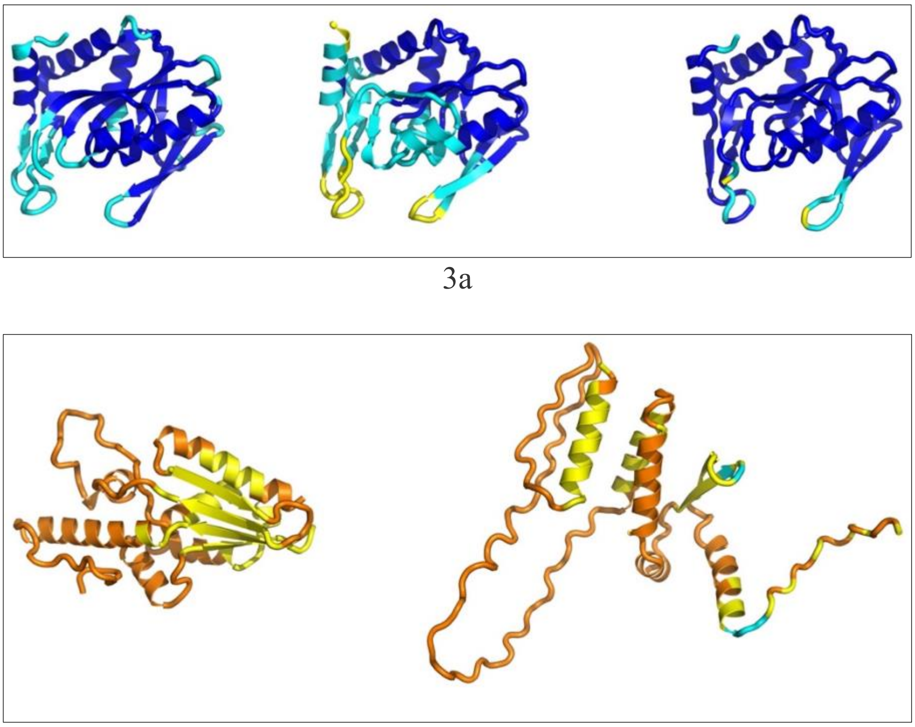
3. RESULTS

As already mentioned in the Introduction, several studies have shown that the sequences obtained by random shuffling of native protein sequences can be expressed, and that in many cases the expressed polypeptide chains fold in aqueous solution into compact structures that display resistance to proteolysis[6] or substantial secondary structure elements[7]. We considered that it would be of interest to see how the deep-learning-based protein structure prediction programs, RTF and AF2, would predict the structures of such shuffled sequences. Fig. 2a displays the crystal structure of human carbonic anhydrase (pdb 6pea), together with the structures predicted by AF2 and RTF. Both algorithms predict structures very similar to that of the experimental structure, with a high pLDDT score. Fig. 2b shows the results of applying the two algorithms to a shuffled 6pea sequence. In both cases the pLDDT score is low. Nevertheless, RTF predicts a compact structure, containing substantial secondary structure elements, with dimensions not much larger than those of the native structure, whereas AF2 predicts a considerably more unfolded structure, with a much lower percentage of secondary structure elements (Table 2). Figs 3 and 4, respectively, show similar representations for adenine phosphoribosyltransferase (*E coli* K12) (PDB-ID 2dy0), and for ribosome anti-association factor EIF6 from *Saccharomyces cerevisiae* (PDB-ID 1g62).


2a

2b

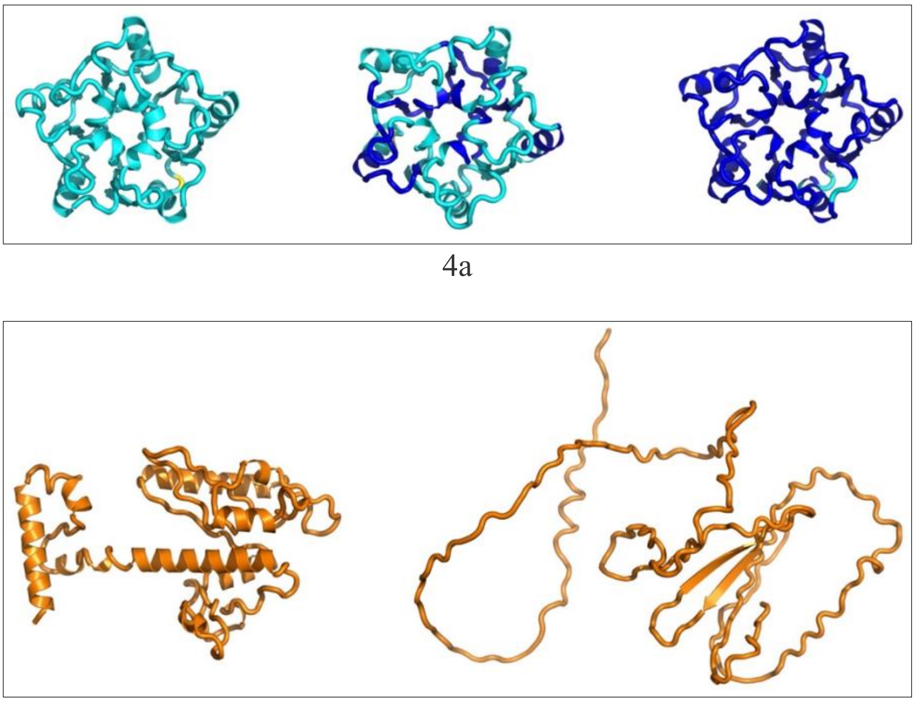
Fig. 2 **a.** Crystal structure of human carbonic anhydrase, pdb 6pea (left), and structures predicted by RTF (center) and AF2 (right); **b.** Structure of a randomized sequence of pdb 6pea as predicted by RTF (left) and AF2 (right).



3a

3b

Fig. 3 **a.** Crystal structure of adenine phosphoribosyltransferase from *E. coli* K12, pdb 2dy0 (left), and structures predicted by RTF (center) and AF2 (right); **b.** Structure of a randomized sequence of 2dy0 as predicted by RTF (left) and AF2 (right).



4a

4b

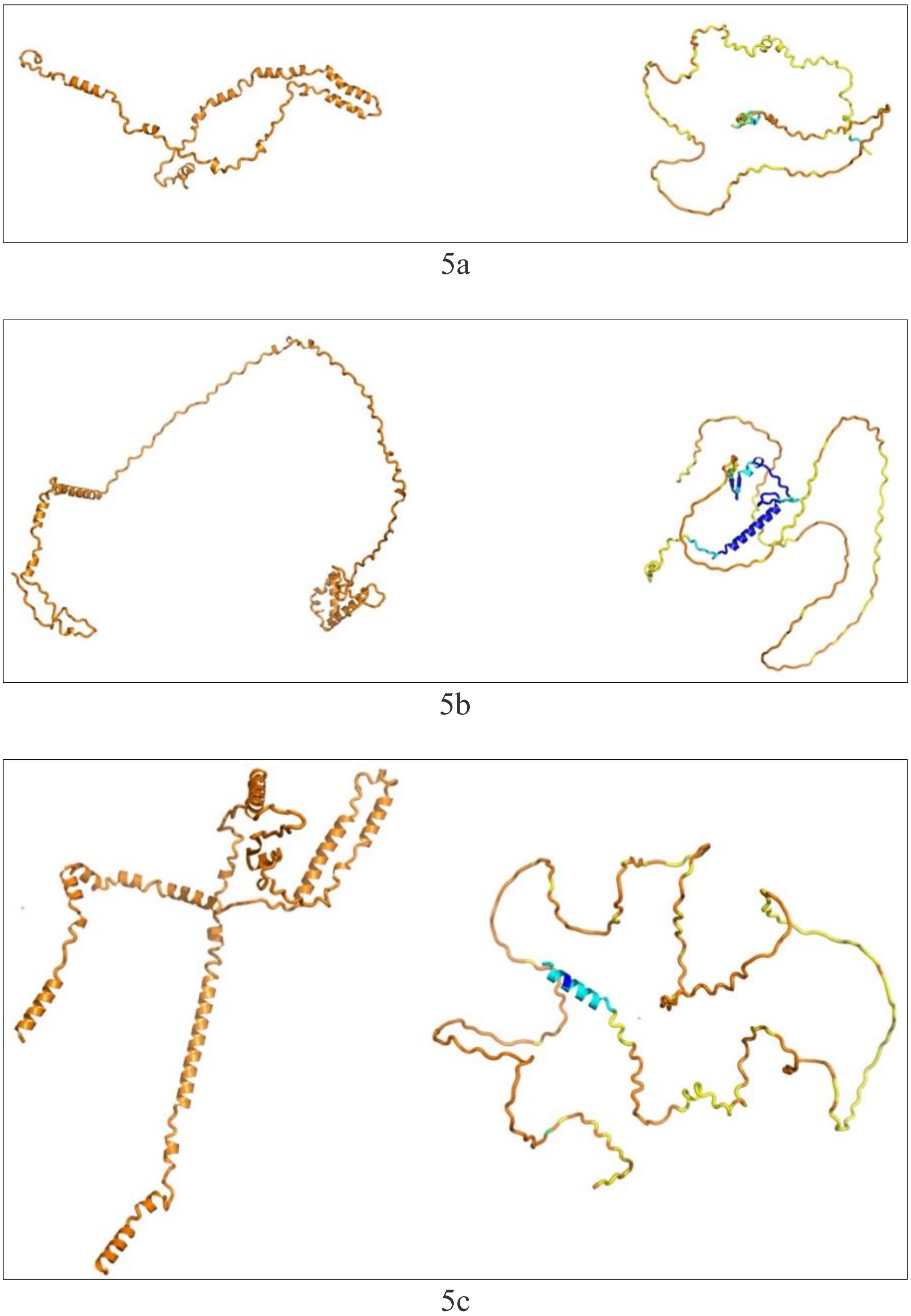
Fig. 4 **a.** Crystal structure of ribosome anti-association factor EIF6 from *Saccharomyces cerevisiae*, pdb 1g62 (left), and structures predicted by RTF (center) and AF2 (right); **b.** Structure of a randomized sequence of pdb 1g62 as predicted by RTF (left) and AF2 (right).

Table 2

Category/Protein	**Numb of AAs**	**ID[a]**	**pLDDT**		
			Xtal[b]	RTF[c]	AF2[d]
Crystal structures					
Human carbonic anhydraese	260	6pea/P00918	86	87	94
Human carbonic anhydrase Ran_01				29	20
E. coli adenine phosphoribosyltransferase	190	2dy0/P69503	92	86	96
E. coli adenine phosphoribosyltransferase Ran_01				42	46
S. cerevisiae ribosome anti-association factor EIF6	225	1g62/Q12522	81	88	96
S. cerevisiae ribosome anti-association factor EIF6 Ran_01				17	29
IDPs					
Drosophila gliotactin cytoplasmic domain[e]	207	Q7KT70		9	52
human CDN1C-Cyclin-dependent kinase inhibitor	316	P49918		11	58
human osteopnotin	314	P10451		6	50
'*Never Born*' Proteins - Group 1					
#1856	109			39	49
#6387	109			25	36
#4090	109			26	43
#2298	109			45	49
'*Never Born*' Proteins - Group 3					
#665	109			37	58
#8667	109			30	47
#3703	109			23	53
#933	109			25	53
Orphan protein crystal structures					
Thermatoga maritima TM0875	170	1o22/Q9WZX8	81	77	77
Thermatoga maritima TM0875 Ran_01				28	32
H. influenzae Hypothetical protein HI1480	187	1mw5/P44209	75	67	45
H. influenzae Hypothetical protein HI1480 Ran_01				24	48
'*Newly Born*' Proteins - well characterized					
Human PBOV1- tumor-specific gene	135	Q9GZY1		35	46
Human FLJ33706 expressed in neurons (alt gene symbol C20orf203)	194	Q8NBC4		29	51
Human NCYM - DNA binding transcriptional activator homolog	109	P40205		26	45
Mouse Gm13030 involved in regulating the pregnancy cycle	143	A2APQ6		26	37
Wheat - TaFROG -an IDP	130	A0A0K1YY56		29	60

[a]ID: 4 letter code is PDB ID (https://www.rcsb.org), while longer length ID corresponds to UniProt accession ID (https://www.uniprot.org
[b]Calculated from the pdb entry (https://www.rcsb.org)
[c]Calculated from RTF model 1
[d]Calculated from AF2-Colab highest ranked model
[3]C-term 207 residues of *Drosophila* gliotactin cytoplasmic domain

The fact that the shuffled protein sequences, as well as several of the '*Newly Born*' proteins (see below), were predicted by AF2 to display long unfolded stretches, prompted us to examine how RTF and AF2 would predict the structures of *bona fide* intrinsically disordered proteins (IDPs). Fig. 5 shows the predictions for three such proteins, the cytoplasmic domain of the ChE-like adhesion molecule (CLAM) from *Drosophila*, gliotactin[44], human CDN1C-cyclin-dependent kinase inhibitor[45], and human osteopontin[46]. In all three cases, both RTF and AF2 predict highly unfolded structures, as might be anticipated. However, in all the models substantial α-helical stretches are predicted, with their percentage in the RTF models being significantly higher than in the AF2 models.



5a

5b

5c

Fig. 5 3D structure predictions for 3 IDPs using RTF and AF2. a. Gli-Cyt, RTF (left), AF2 (right); b. CDN1C, RTF (left), AF2 (right); c. Osteopontin, RTF (left), AF2 (right).

We then turned our attention to modeling some of the '*Never Born*' proteins generated and expressed by[7]. As shown in Fig. 6, they predicted, using a repertoire of bioinformatic tools, that some of the sequences, which they subsequently expressed would be ordered/folded, with a high content of secondary structure elements (Group 1), whereas others would be disordered/unfolded, with a low content of secondary structure elements (Group 3).

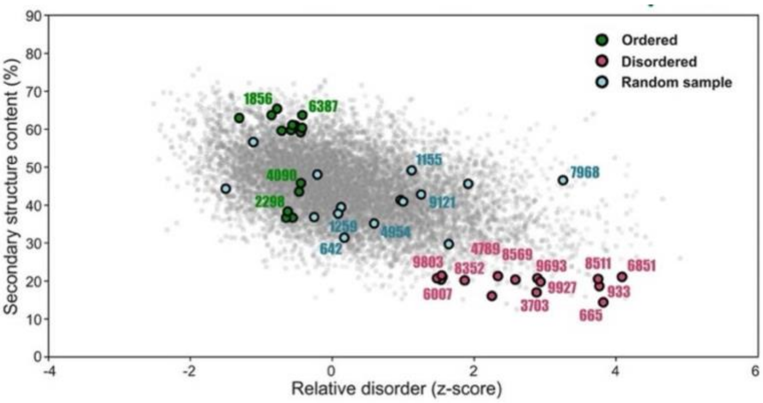


Fig. 6 Selection of sequences from the set of '*Never Born'* proteins taken for experimental characterization. Secondary structure is plotted on the y-axis vs relative disorder on the x-axis. Members of Group 1 (green circles) fall into the category of ordered/folded proteins, and members of Group 3 (red circles) fall into the category of disordered/unfolded proteins (taken with permission from the paper of Tretyachenko et al, 2017).

Again, we used both RTF and AF2 to model the structures; the data for four members of Group 1 are displayed in Fig. 7, and for four members of Group 3 in Fig. 8. For all four members of Group 1, RTF predicts compact structures, with a high percentage of secondary structure. The structures predicted by AF2 are more open, with a much higher percentage of disordered stretches. Both RTF and AF2 predict highly unfolded structures for all four members of Group 3 modelled, though significant amounts of helical elements are observed, especially in the models generated by RTF.

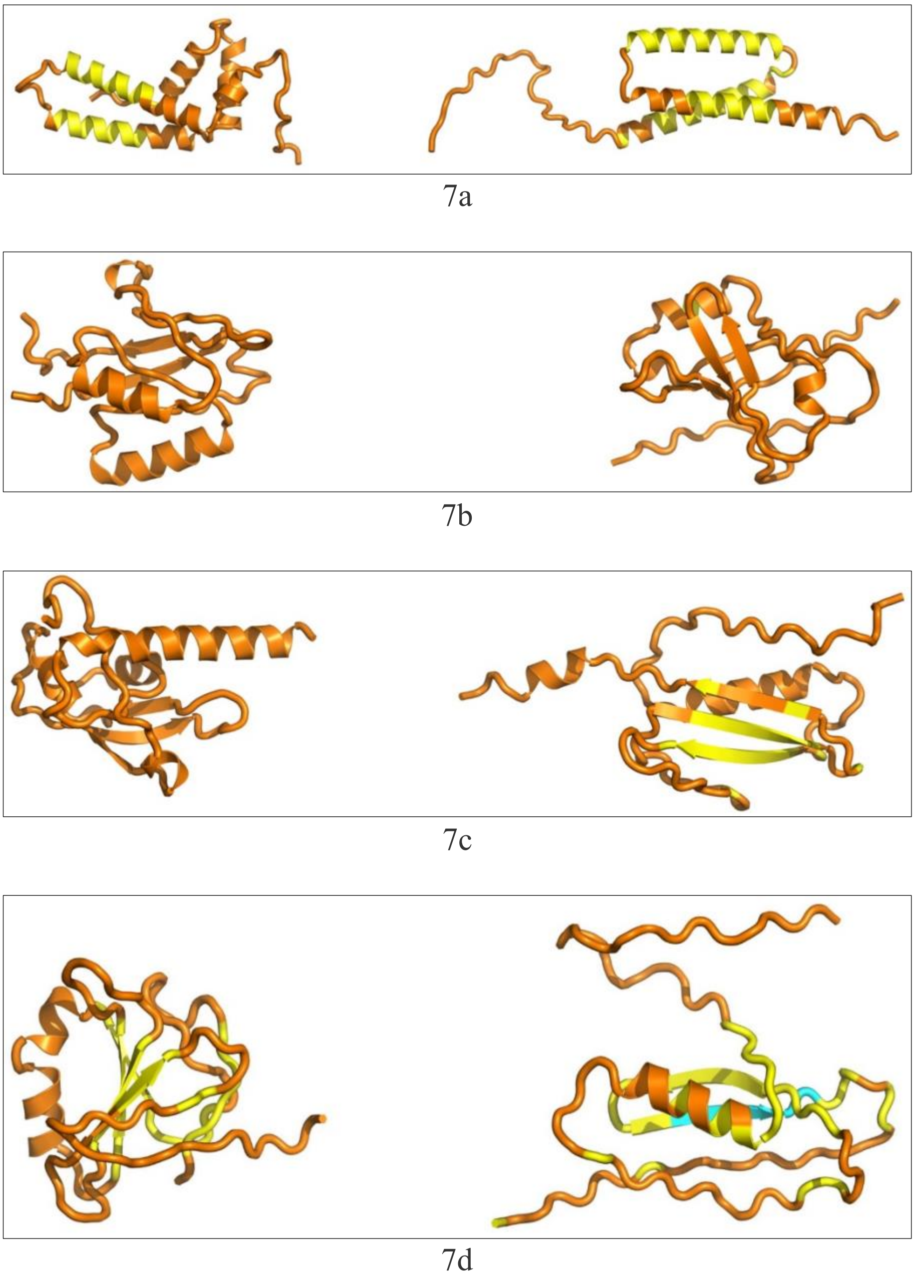
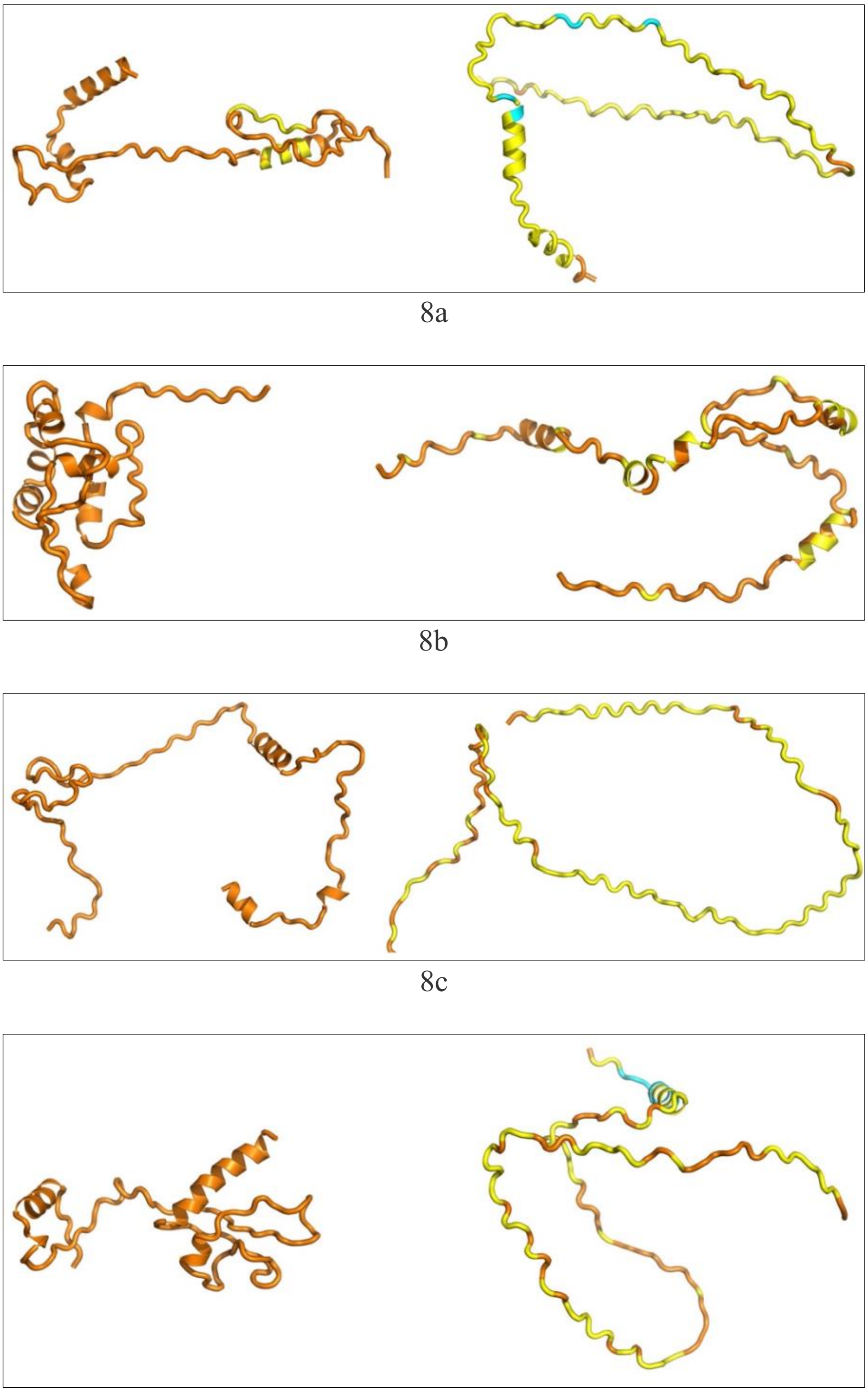


Fig. 7 3D structure predictions for 4 members of Group 1 of *'Never Born'* proteins. **a.** #1856, RTF (left), AF2 (right); **b.** #6387, RTF (left), AF2 (right); **c.** #4090, RTF (left), AF2 (right); **d.** #2298, RTF (left), AF2 (right).



8a

8b

8c

8d

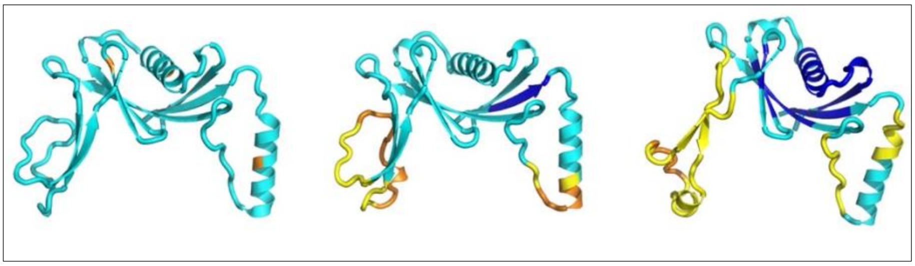
Fig. 8 3D structure predictions for 4 members of Group 3 of *'Never Born'* proteins. a. #665, RTF (left), AF2 (right); **b.** #8667, RTF (left), AF2 (right); **c.** #3703, RTF (left), AF2 (right); **d.** #933, RTF (left), AF2 (right).

Despite the fact that research on orphan proteins is a hot topic, largely due to its evolutionary implications, we were able to identify only three crystal structures of orphan proteins in the PDB. Fig 9a shows the crystal structure of orphan protein TM0875 from *Thermatoga maritima*

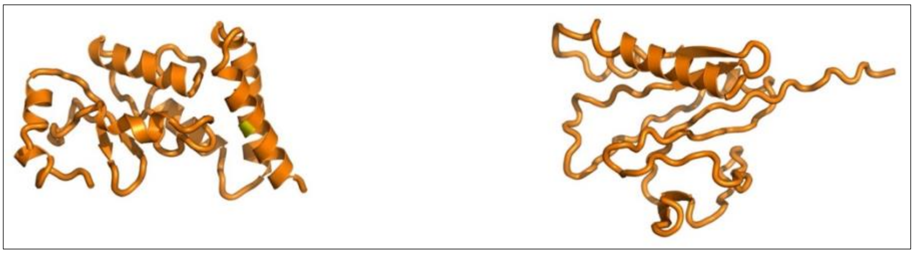
(PDB 1o22)[47], alongside high-quality predictions of its structure by both RTF and AF2. The authors pointed out that this is a novel and unique fold, and application of the Dali [38,48] and Foldseek[40] servers reveals that it still maintains this status based on a much large number of experimental structures in the PDB, as well as in the entire AlphaFold Database (https://alphafold.ebi.ac.uk). Fig. 9b shows the results of applying the two prediction algorithms to a shuffled 1o22 sequence. In both cases the pLDDT score is low. Nevertheless, RTF predicts a compact structure, containing substantial secondary structure elements, with dimensions not much larger than those of the native structure, whereas, again, AF2 predicts a more open structure.

Fig 10a shows the crystal structure of the other orphan protein deposited in the PDB, that of the hypothetical protein HI1480 from *Haemophilus influenzae* (PDB 1mw5)[49], alongside the structures predicted by RTF and AF2. In this case, too, the authors pointed out that this is a novel and unique fold, and application of the Dali and Foldseek servers again reveals that it still maintains this status based on the much large number of experimental structures now available, as well as the entire AlphaFold Database. In this case, however, whereas RTF predicts a large portion of the crystal structure fairly well, AF2 generates a very poor prediction, with a pLDDT of 45. Fig 10b shows the results of applying the two algorithms to a shuffled 1mw5 sequence. In both cases the pLDDT score is low. Nevertheless, RTF predicts a compact structure, containing substantial secondary structure elements, with dimensions about the same size as the native structure, whereas AF2 predicts a somewhat more open structure, with little secondary structure.

Solution of the crystal structure of Cthe_2751, a third singleton from *Clostridium thermocellum*, revealed an all α-helical topology similar to those observed for nucleic acid processing proteins[50]. Thus, perhaps not surprisingly, not all orphan proteins display novel folds.

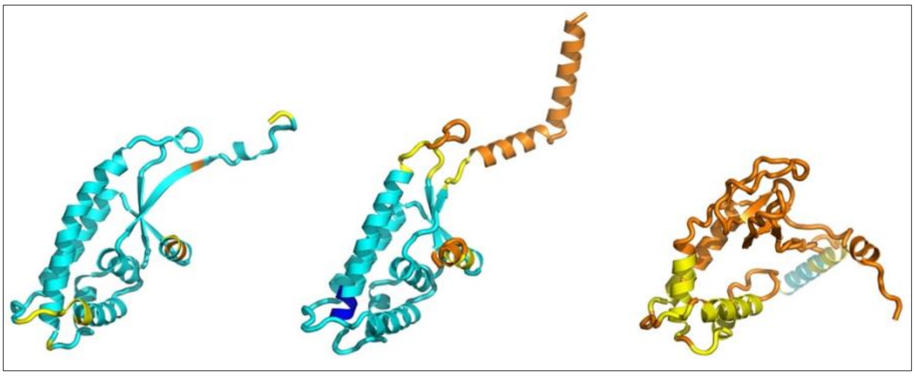


9a

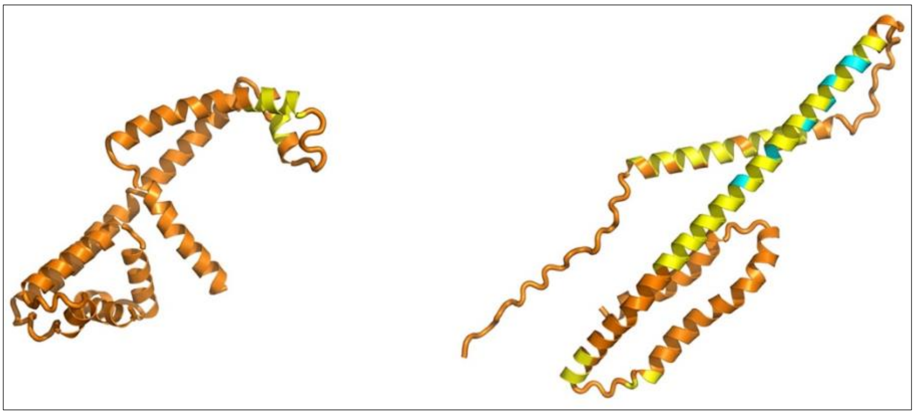


9b

Fig. 9 **a.** Crystal structure of the Orphan protein TM0875 from *Thermatoga maritima*, pdb 1o22 (left), and structures predicted by RTF (center) and AF2 (right); **b.** Structure of a randomized sequence of pdb 1o22 as predicted by RTF (left) and AF2 (right).



10a



10b

Fig. 10 **a.** Crystal structure of the orphan protein Hypothetical protein HI1480 from *Haemophilus influenzae*, pdb 1mw5 (left), and structures predicted by RTF (center) and AF2 (right); **b.** Structure of a randomized sequence of pdb 1mw5 as predicted by RTF (left) and AF2 (right).

We thought that it would be of interest to use RTF and AF2 to predict the structures of orphan proteins for which no experimental structure was available, in order to see whether they would yield novel folds. To this end we selected four proteins whose characterization as orphan proteins appeared to be robust, which have been shown to have clear biological functions, and for which the necessary sequence data are available. These proteins were:

- PBOV1, a human tumor-specific gene that mitigates the clinical outcome[51]
- FLJ33706 (alternative gene symbol C20orf203), expressed in neurons within the human brain[52]
- NCYM, a DNA-binding transcriptional activator homolog in *Homo sapiens*[53]
- GM13030, uncharacterized protein LOC105734733 from *Mus musculus*, involved in regulating the pregnancy cycle[54]

As an initial step in characterizing these four proteins, we utilized two software programs, IUPred3[43] and NetSurfP3[42], to investigate whether they were intrinsically disordered or folded. Fig. 11 shows data obtained using IUPred3. It can be seen that, except for short stretches at their N- and C-termini, three of the proteins are predicted to be completely folded, while FLJ33706 has a short, disordered stretch in the middle of its sequence. NetSurfP3, too, predicts these proteins to be folded, except at their extremities. It also indicates substantial α-helical content, and a few β-strands. The number of amino acids for these four orphan proteins ranges from 109 for NCYM to 194 for FLJ77306, and their Pi values range from 8.91 for Gm13030 to 11.74 for FLJ77306, which contains 10% Arg residues.

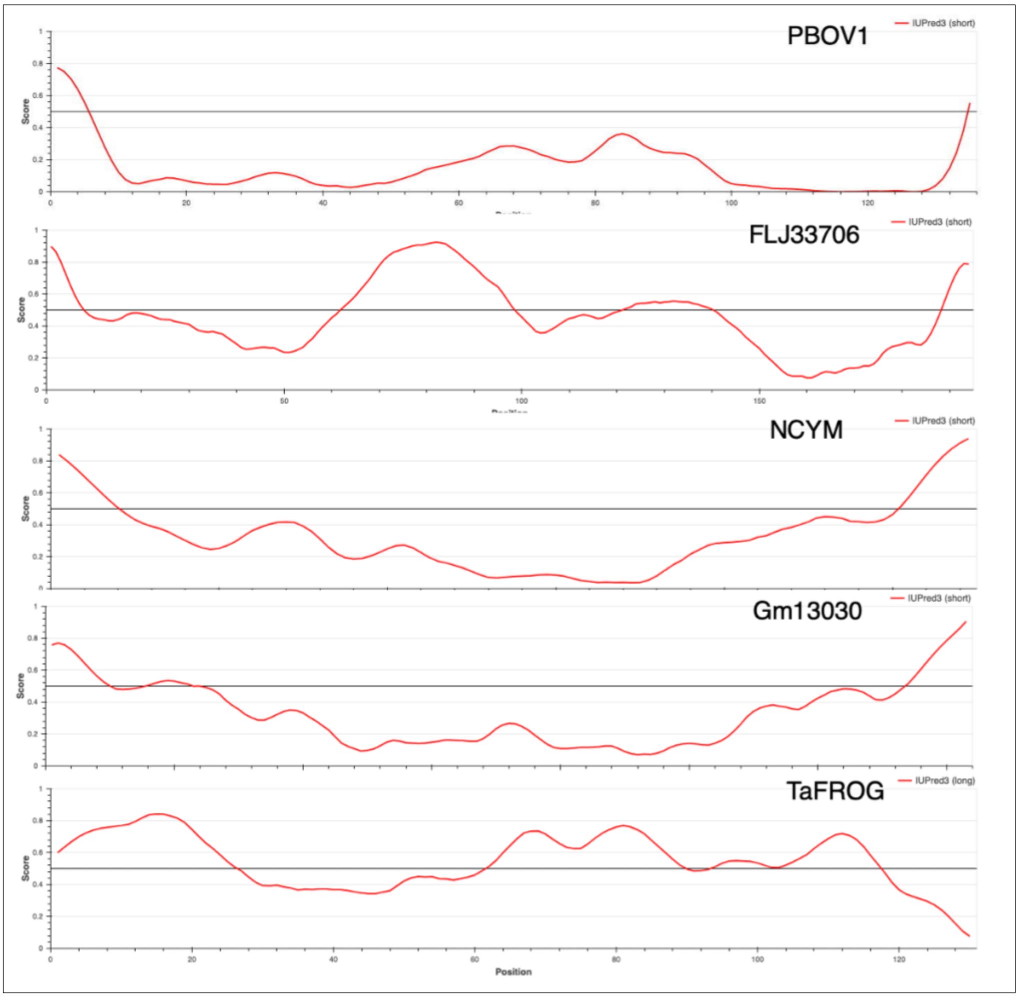
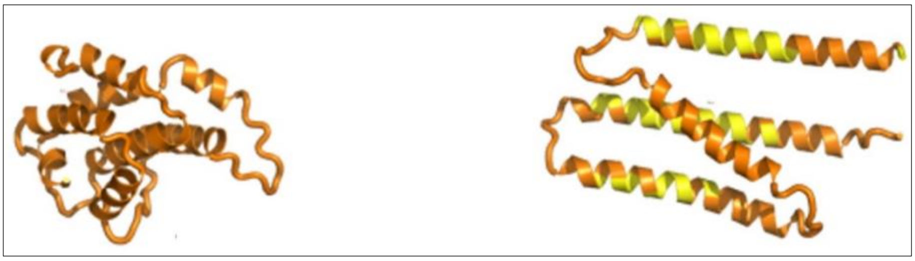
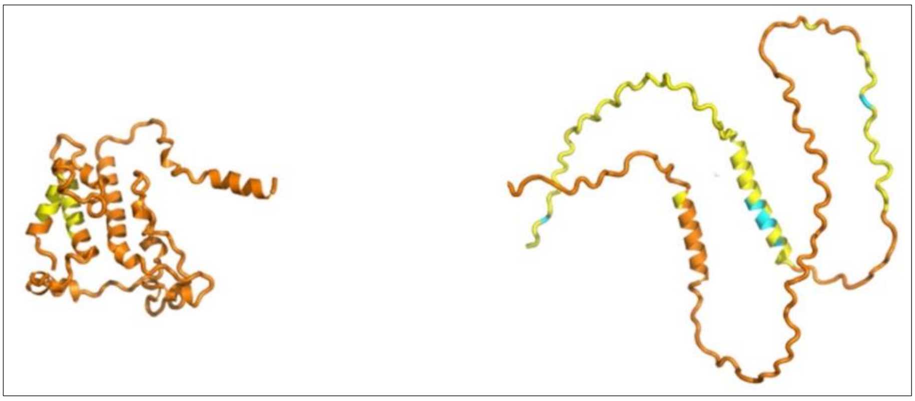


Fig. 11 IUPRED3 predictions of order/disorder in five well-studied Orphan proteins. PBOV1; FLJ33706; NCYM; GM13030; *Ta*FROG. Sequences above the horizontal line are classified as disordered, and below the line as ordered.

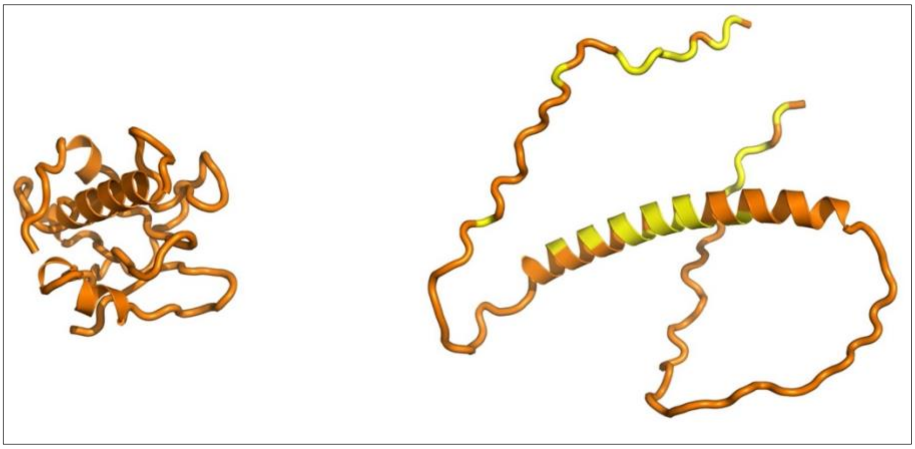
As seen in Fig. 12, RTF predicts compact structures, with substantial helical content, for all five proteins. AF2 predicts a compact structure, with substantial helical content, for PBOV1, which differs substantially from that predicted by RTF. For the other three proteins, AF2 predicts structures that are largely disordered, thus being in disagreement with the predictions of both IUPred3 and NetSurfP3. Finally, we examined how the two algorithms would model the structure of the orphan protein from wheat (*Triticum aestivum*), *Ta*FROG^{55}, which was predicted to be an IDP using FoldIndex^{41} (as seen at: https://fold.proteopedia.org/cgi-bin/findex), an assignment that we confirmed using both IUPred3 (Fig. 11) and NetSurfP3 (not shown). Both RTF and AF2 predicted unfolded structures that differ substantially from each other, with that predicted by AF2 being much more disordered (Fig. 12).



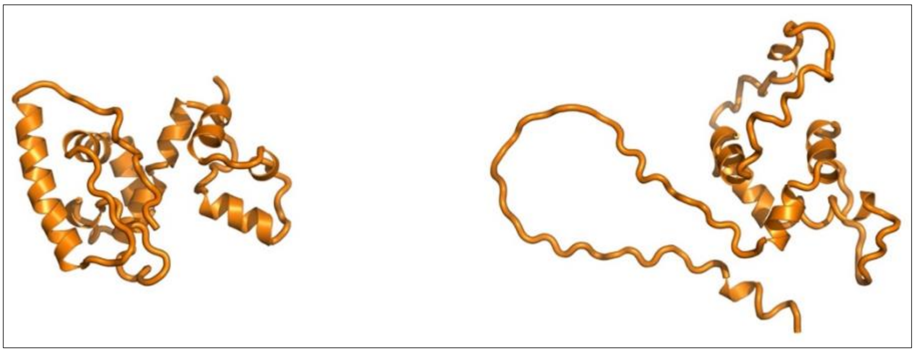
12a



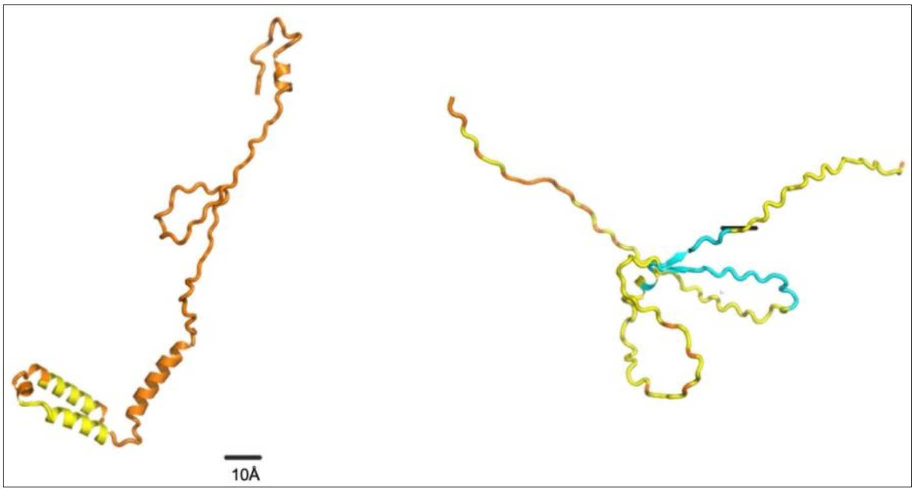
12b



12c



12d



12e

Fig. 12 3D structure predictions for 6 orphan proteins using RTF and AF2. **A.** PBOV1, RTF (left), AF2 (right); **b.** Q8NBC4, RTF (left), AF2 (right); **c.** NCYM, RTF (left), AF2 (right); **d.** Gm13030, RTF (left), AF2 (right); **e.** *Ta*FROG, RTF (left), AF2 (right).

4. DISCUSSION

Both AF2 and RTF use multiple sequence alignment of homologous proteins extensively as an important element for structure prediction[29,30]. The pLDDT values that they obtain for structures of proteins that lack homologs generally indicate lower reliability of the models[29,30]. We thus feared that the low pLDDT scores that we obtain for the '*Newly Born*' proteins and for the '*Never Born*' proteins, both of which are categories which are devoid of homologs, might lack significance. However, for the two experimentally determined orphan proteins in the PDB, 1o22 and 1mw5, RTF predicts both structures accurately, with relatively high pLDDT scores, and AF2, too, predicts 1o22 well, although it predicts 1mw5 rather poorly. It is plausible, however, that high pLDDT scores were obtained with RTF for orphan proteins for which structural data were available, while much lower scores were obtained for those for which structural data were lacking, is due to the fact that the structures that had already been experimentally determined were used in the training sets for the prediction algorithms.

We feel, therefore, that, despite the low pLDDT scores obtained, the prediction by RTF that all four orphan proteins are globular, in agreement with the assignments of both IUPRED3 and NetSurfP3, is significant, even though the details of the particular folds predicted may not be accurate. As a positive control, for the orphan protein, *Ta*FROG, which has experimentally been shown to be an IDP, both RTF and AF2 predict unfolded conformations.

Monzon *et al.*[56] recently examined 250 proteins sequence families in the AntiFam resource[57], which are thought to be *spurious* proteins. Specifically, they are believed to be ORFs either on the opposite strand or in a different, overlapping reading frame, with respect to the true protein-coding or non-coding RNA gene[57]**.** They conjectured that proteins belonging to these families would not fold into well folded globular structures. Using AF2 they confirmed this prediction with one exception. To the best of our knowledge these spurious proteins were not examined with RTF. The findings of Monzon *et al.*[56] are in contrast to our observations for the well characterized orphan proteins that were discussed above, which have been shown to have biological functions.

In the Introduction, we referred to a recent study from the Baker lab[11], in which 129 random sequences were modeled with RTF. These models, presumably with low reliability scores, served as useful starting models for optimization into folded proteins. Thus, although these models have low pLDDT scores, and thus most likely differ significantly in detail from the actual structures, they can provide good low-resolution starting points.

The experimental evidence demonstrates that all the 'Never Born' Group 1 proteins display significant secondary structure in solution, and this is corroborated by the RTF predictions for the four members of the group that we examined, though much less so by AF2. In contrast, for the 'Never Born' Group 3 proteins, which the physicochemical evidence classifies as IDPs, both RTF and AF2 predict them to be unfolded.

Thus, for the purposes of our study, despite the low pLDDT scores for the structures predicted by RTF for both the '*Newly Born*' and '*Never Born*' proteins, we feel that the conclusions that we draw are valid.

A principal conclusion that can be drawn from the data presented above is that orphan proteins often display novel folds, which do not overlap with folds already present in the PDB. Indeed, the total number of distinct protein folds has been the topic of heated controversy[58,59]. In the present study, novel folds are observed for two of the three orphan proteins for which crystal structures exist (Figs 9 and 10), and in the four orphan proteins that were predicted to be folded, using both IUPred3 and NetSurfP3. Thus, RTF predicted novel folds for all four. One of the structures predicted by AF2 was folded, and bore some rough similarity to that predicted by RTF, while the remaining three were predicted by AF2 to be largely unfolded, despite the predictions of IUPred3 and NetsurfP3.

One of the major reservations that has been made with respect to orphan proteins being '*Newly Born*' proteins, coded for by sequences that were previously non-coding sequences, is that the genes in question might have undergone such rapid evolution that their homology to their predecessors was no longer recognizable[60,61]. The fact that, with one exception, the orphan proteins studied here display novel folds, substantially weakens this argument.

Examination of the predicted structures of random sequences produced by shuffling of native sequences by Tretyachenko *et al.*[7] took advantage of the fact that these authors had already performed spectroscopic studies on a set of such polypeptides that they had expressed in *E. coli*, and had shown that one could distinguish categories corresponding to folded proteins (Group 1), and to IDPs (Group 3), as illustrated in Fig. 9.

We used RTF and AF2 to model 4 representatives of each group, as shown above. As expected, those classified as IDPs in Group 3 were indeed largely unfolded for both RTF and AF2, with the AF2 models being significantly larger. Those in Group 1 were in all cases much more compact, but those predicted by RTF were more compact than those predicted by AF2, and contained very few unfolded stretches, whereas substantial unfolded stretches were observed in three out of four of the structures modelled by AF2. It will be interesting to find out whether the folds predicted for the '*Never Born*' proteins in group 1 are novel folds rather than ones that already have appeared in nature.

In retrospect, it is not surprising that many random polypeptide sequences of a suitable amino acid composition, at a first approximation with a high content of hydrophobic residues, and a low net charge[62], will yield a compact structure containing substantial secondary structure motifs, as shown by[7]. It is possible that this is due to the fact that the sequences retained the amino acid compositions of the natural proteins from which they were generated, thus not being completely random.

The paradigm change introduced by Kuwajima and Ptitsyn in the 1980s[63,64] resulted in the realization that the newly synthesized polypeptide that emerges from the ribosome does not persist as an extended unfolded polypeptide, unless it is an IDP, but rather collapses to what is termed a 'Molten Globule' (MG), a compact structure somewhat larger than the fully folded native structure (Fig. 13). The MG contains substantial secondary structure elements, but lacks the precise tertiary interactions of the native structure. Small proteins may spontaneously undergo transition to the native state, whereas larger proteins may require the assistance of molecular chaperones to complete the folding process. The spectroscopic data of Tretyachenko *et al.*[7] only tell us that compact structures, with secondary structure elements, have been produced by their shuffled polypeptide sequences. When a native protein unfolds to a MG, or to some other partially unfolded species, hydrophobic amino acid side chains that are buried in the hydrophobic core become exposed. The degree of their exposure can be checked by use of the amphiphilic probe, 1-anilinonaphthalene-8-sulfonate (ANS), whose fluorescence is enhanced upon interaction with the hydrophobic residues[65]. It would, therefore, be interesting to compare the ANS fluorescence of the *'Never Born'* proteins generated by Tretyachenko *et al.*[7] to that of typical globular proteins in their native state. It is worth mentioning that in an early study on folding of polypeptides with random sequences of simplified amino acid composition, NMR data indicated loose packing of the folded state[8].

In any event, one can speculate that '*Newly Born*' proteins might, initially, assume a MG-like conformation that would resemble that of the '*Never Born*' proteins, and that mutations, coupled with natural selection, might convert some of them into 'native' orphan proteins with novel biological activities. This can be considered analogous to what occurred in the study in which the hallucinatory proteins were generated[11].

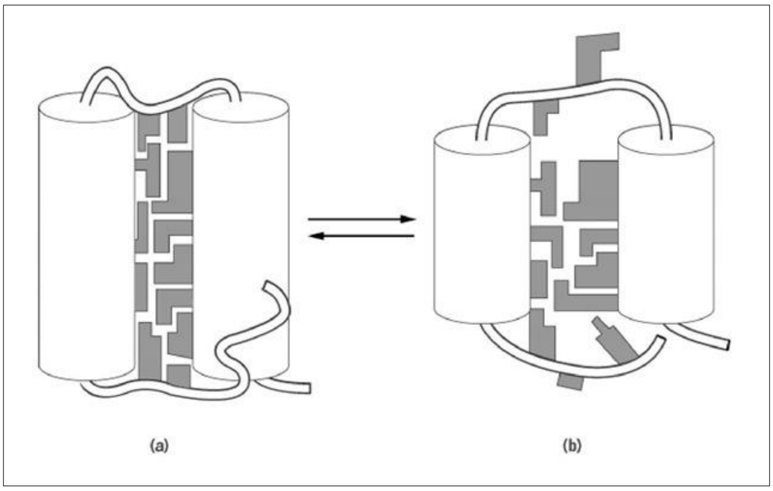


Fig. 13 A schematic model of the native (a) and molten globule (MG) (b) states of a protein molecule. For the sake of simplicity, only two α-helices are represented. According to this model, the MG preserves the mean overall structural features of the native protein, but differs from the native state mainly in being more loosely packed, and thus having a volume larger by *ca.* 10%, and exposing more hydrophobic surfaces (reproduced from http://what-when-how.com/molecular-biology/molten-globule-molecular-biology).

Why does RTF do relatively well in predicting plausible compact structures for randomized sequences, whereas AF2 often makes predictions that are either clearly wrong or implausible? In a recent brief survey of the principles underlying AF2[66] it was emphasized that AF makes extensive use of detection of conserved interactions of residues that are remote from each other in the linear sequence. This approach was earlier proposed, and implemented with a certain degree of success, by Marks and Sander[67]. Obviously, such conserved interactions of distant residues would not exist in randomized sequences. Nor would such conserved interactions be available in orphan proteins, which lack ancestral homologs and, furthermore, mostly display novel folds. Apparently, even if RTF makes use of such conserved long-distance interactions, it is able to successfully model the overall shape of novel proteins consistently even in the absence of such information.

The principal observation in this study is the demonstration that, based on the RTF predictions, orphan proteins may have novel folds, derived from their unique sequences, which are associated with the appearance of novel biological functions. It will be interesting to express and purify more of these proteins, so as to determine their experimental structures.

5. REFERENCES

1. Bränden C, Tooze J. *Introduction to Protein Structure.* Second ed. New York: Garland Publishing, Inc.; 1999.
2. Pe'er I, Felder CE, Man O, Silman I, Sussman JL, Beckmann JS. Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. *Proteins.* 2004;54(1):20-40.
3. Lavelle DT, Pearson WR. Globally, unrelated protein sequences appear random. *Bioinformatics.* 2010;26(3):310-318.
4. De Lucrezia D, Slanzi D, Poli I, Polticelli F, Minervini G. Do natural proteins differ from random sequences polypeptides? Natural vs. random proteins classification using an evolutionary neural network. *PLoS One.* 2012;7(5):e36634.
5. Chiarabelli C, Vrijbloed JW, Thomas RM, Luisi PL. Investigation of de novo totally random biosequences, Part I: A general method for in vitro selection of folded domains from a random polypeptide library displayed on phage. *Chem Biodivers.* 2006;3(8):827-839.
6. Chiarabelli C, Vrijbloed JW, De Lucrezia D, et al. Investigation of de novo totally random biosequences, Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display. *Chem Biodivers.* 2006;3(8):840-859.
7. Tretyachenko V, Vymetal J, Bednarova L, et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci Rep.* 2017;7(1):15449.
8. Davidson AR, Lumb KJ, Sauer RT. Cooperatively folded proteins in random sequence libraries. *Nat Struct Biol.* 1995;2(10):856-864.
9. Keefe AD, Szostak JW. Functional proteins from a random-sequence library. *Nature.* 2001;410(6829):715-718.
10. Hecht MH, Das A, Go A, Bradley LH, Wei Y. De novo proteins from designed combinatorial libraries. *Protein Sci.* 2004;13(7):1711-1723.
11. Anishchenko I, Pellock SJ, Chidyausiku TM, et al. De novo protein design by deep network hallucination. *Nature.* 2021;600(7889):547-552.
12. Tautz D, Domazet-Loso T. The evolutionary origin of orphan genes. *Nat Rev Genet.* 2011;12(10):692-702.
13. Carvunis AR, Rolland T, Wapinski I, et al. Proto-genes and de novo gene birth. *Nature.* 2012;487(7407):370-374.
14. Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet.* 2013;9(10):e1003860.
15. McLysaght A, Guerzoni D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci.* 2015;370(1678):20140332.
16. Schmitz JF, Bornberg-Bauer E. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Res.* 2017;6:57.
17. Vakirlis N, Carvunis AR, McLysaght A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife.* 2020;9.

18. Li J, Singh U, Bhandary P, et al. Foster thy young: enhanced prediction of orphan genes in assembled genomes. *Nucleic Acids Res.* 2022;50(7):e37.
19. Jacob F. Evolution and tinkering. *Science.* 1977;196(4295):1161-1166.
20. Schmitz J, Brosius J. Exonization of transposed elements: A challenge and opportunity for evolution. *Biochimie.* 2011;93(11):1928-1934.
21. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* 2009;25(9):404-413.
22. Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. *Genome Res.* 2009;19(10):1752-1759.
23. Siepel A. Darwinian alchemy: Human genes from noncoding DNA. *Genome Res.* 2009;19(10):1693-1695.
24. Toll-Riera M, Bosch N, Bellora N, et al. Origin of Primate Orphan Genes: A Comparative Genomics Approach. *Mol Biol Evol.* 2009;26(3):603-612.
25. Wu DD, Irwin DM, Zhang YP. De novo origin of human protein-coding genes. *PLoS Genet.* 2011;7(11):e1002379.
26. Xie C, Zhang YE, Chen JY, et al. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* 2012;8(9):e1002942.
27. Dowling D, Schmitz JF, Bornberg-Bauer E. Stochastic Gain and Loss of Novel Transcribed Open Reading Frames in the Human Lineage. *Genome Biol Evol.* 2020;12(11):2183-2195.
28. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. *eLife.* 2014;3:e03523.
29. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596(7873):583-589.
30. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science.* 2021;373(6557):871-876.
31. Sussman JL, Lin D, Jiang J, et al. Protein data bank (PDB): a database of 3D structural information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr.* 1998;54(Pt 6 Pt 1):1078-1084.
32. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000;28(1):235-242.
33. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins.* 2021;89(12):1607-1617.
34. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol.* 2008;18(6):756-764.
35. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods.* 2022.
36. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics.* 2013;29(21):2722-2728.
37. Hiranuma N, Park H, Baek M, Anishchenko I, Dauparas J, Baker D. Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat Commun.* 2021;12(1):1340.

38. Holm L, Sander C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* 1998;26(1):316-319.

39. Holm L. DALI and the persistence of protein shape. *Protein Sci.* 2020;29(1):128-140.

40. van Kempen M, Kim S, Tumescheit C, Mirdita M, Soeding J, Steinegger M. Foldseek: fast and accurate protein structure search. *bioRxiv.* 2022:2022.2002.2007.479398.

41. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, et al. FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics.* 2005;21(16):3435-3438.

42. Høie MH, Kiehl EN, Petersen B, et al. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res.* 2022:gkac439.

43. Erdős G, Pajkos M, Dosztányi Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* 2021;49(W1):W297-W303.

44. Zeev-Ben-Mordehai T, Rydberg EH, Solomon A, et al. The intracellular domain of the Drosophila cholinesterase-like neural adhesion protein, gliotactin, is natively unfolded. *Proteins.* 2003;53(3):758-767.

45. Adkins JN, Lumb KJ. Intrinsic structural disorder and sequence features of the cell cycle inhibitor p57Kip2. *Proteins.* 2002;46(1):1-7.

46. Kurzbach D, Platzer G, Schwarz TC, Henen MA, Konrat R, Hinderberger D. Cooperative unfolding of compact conformations of the intrinsically disordered protein osteopontin. *Biochemistry.* 2013;52(31):5167-5175.

47. Bakolitsa C, Schwarzenbacher R, McMullan D, et al. Crystal structure of an orphan protein (TM0875) from Thermotoga maritima at 2.00-A resolution reveals a new fold. *Proteins.* 2004;56(3):607-610.

48. Holm L. Dali server: structural unification of protein families. *Nucleic Acids Res.* 2022;50(W1):W210-W215.

49. Lim K, Sarikaya E, Galkin A, et al. Novel structure and nucleotide binding properties of HI1480 from Haemophilus influenzae: a protein with no known sequence homologues. *Proteins.* 2004;56(3):564-571.

50. Cheng C, Shaw N, Zhang X, et al. Structural View of a Non Pfam Singleton and Crystal Packing Analysis. *PLOS ONE.* 2012;7(2):e31673.

51. Samusik N, Krukovskaya L, Meln I, Shilov E, Kozlov AP. PBOV1 is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. *PLoS One.* 2013;8(2):e56162.

52. Li CY, Zhang Y, Wang Z, et al. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput Biol.* 2010;6(3):e1000734.

53. Suenaga Y, Islam SM, Alagu J, et al. NCYM, a Cis-antisense gene of MYCN, encodes a de novo evolved protein that inhibits GSK3beta resulting in the stabilization of MYCN in human neuroblastomas. *PLoS Genet.* 2014;10(1):e1003996.

54. Xie C, Bekpen C, Kunzel S, et al. A de novo evolved gene in the house mouse regulates female pregnancy cycles. *eLife.* 2019;8.

55. Perochon A, Jianguang J, Kahla A, et al. TaFROG Encodes a Pooideae Orphan Protein That Interacts with SnRK1 and Enhances Resistance to the Mycotoxigenic Fungus Fusarium graminearum. *Plant Physiol.* 2015;169(4):2895-2906.

56. Monzon V, Haft DH, Bateman A. Folding the unfoldable: using AlphaFold to explore spurious proteins. *Bioinformatics Advances.* 2022;2(1):vbab043.

57. Eberhardt RY, Haft DH, Punta M, Martin M, O'Donovan C, Bateman A. AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database.* 2012;2012:bas003.

58. Rost B. Did evolution leap to create the protein universe? *Curr Opin Struct Biol.* 2002;12(3):409-416.

59. Coulson AF, Moult J. A unifold, mesofold, and superfold model of protein fold use. *Proteins.* 2002;46(1):61-71.

60. Light S, Basile W, Elofsson A. Orphans and new gene origination, a structural and evolutionary perspective. *Curr Opin Struct Biol.* 2014;26:73-83.

61. Bornberg-Bauer E, Hlouchova K, Lange A. Structure and function of naturally evolved de novo proteins. *Curr Opin Struct Biol.* 2021;68:175-183.

62. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins.* 2000;41(3):415-427.

63. Ptitsyn O. How molten is the molten globule? *Nat Struct Biol.* 1996;3(6):488-490.

64. Arai M, Kuwajima K. Role of the molten globule state in protein folding. *Adv Protein Chem.* 2000;53:209-282.

65. Dolginova EA, Roth E, Silman I, Weiner LM. Chemical modification of Torpedo acetylcholinesterase by disulfides: appearance of a "molten globule" state. *Biochemistry.* 1992;31(48):12248-12254.

66. Jumper J, Hassabis D. Protein structure predictions to atomic accuracy with AlphaFold. *Nat Methods.* 2022;19(1):11-12.

67. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. *Nat Biotechnol.* 2012;30(11):1072-1080.