

RESEARCH

Open Access



# Role of environmental specificity in CASP results

Irena Roterman<sup>1\*</sup>, Katarzyna Stapor<sup>2</sup> and Leszek Konieczny<sup>3</sup>

\*Correspondence:  
myroterm@cyf-kr.edu.pl

<sup>1</sup> Department of Bioinformatics and Telemedicine, Jagiellonian University - Medical College, Medyczna 7, 30-688 Krakow, Poland

<sup>2</sup> Faculty of Automatic, Electronics and Computer Science, Department of Applied Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

<sup>3</sup> Jagiellonian University - Medical College, Kopernika 7, 31-034 Krakow, Poland

## Abstract

**Background:** Recently, significant progress has been made in the field of protein structure prediction by the application of artificial intelligence techniques, as shown by the results of the CASP13 and CASP14 (Critical Assessment of Structure Prediction) competition. However, the question of the mechanism behind the protein folding process itself remains unanswered. Correctly predicting the structure also does not solve the problem of, for example, amyloid proteins, where a polypeptide chain with an unaltered sequence adopts a different 3D structure.

**Results:** This work was an attempt at explaining the structural variation by considering the contribution of the environment to protein structuring. The application of the fuzzy oil drop (FOD) model to assess the validity of the selected models provided in the CASP13, CASP14 and CASP15 projects reveals the need for an environmental factor to determine the 3D structure of proteins. Consideration of the external force field in the form of polar water (Fuzzy Oil Drop) and a version modified by the presence of the hydrophobic compounds, FOD-M (FOD-Modified) reveals that the protein folding process is environmentally dependent. An analysis of selected models from the CASP competitions indicates the need for structure prediction as dependent on the consideration of the protein folding environment.

**Conclusions:** The conditions governed by the environment direct the protein folding process occurring in a certain environment. Therefore, the variation of the external force field should be taken into account in the models used in protein structure prediction.

**Keywords:** Protein folding, Folding simulation in Silico, Folding environment, CASP

## Introduction

The protein folding problem has a long history of identification of the phenomenon in which proteins adopt a 3D structure in an explicitly defined and reproducible way [1, 2]. Monitoring the progress in this area is made possible by the CASP (“Critical Assessment Structure Prediction”) project, which has been held biannually since 1994 [3–12]. The project organiser provides participants with the amino acid sequence of a protein with a 3D structure (the target) known only to the organisers. Based on the protein sequence provided, CASP participants predict the 3D structure of the protein (model) using



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

methods they developed [3]. In addition to the traditionally used homology-based technique, *ab initio* techniques are also being developed. The homology-based technique consists in finding proteins with a sequence that has a sufficiently high degree of similarity and under the assumption that a similar sequence provides a similar structure (especially if homologous proteins are involved); it has traditionally provided better results in assessing the degree of the structural accuracy of the models [13–15]. The *ab initio* techniques seek a theoretical model without reference to known structures, and they develop tools – as one would expect – that reproduce the mechanism of the protein folding process. This is pursued by proposing different forms of force field notation, the presence of which for a given sequence directs the structuring towards the native structure [16–19]. These two techniques dominated the history of the CASP project until 2020, when the spectacular success of artificial intelligence (AI)-based technology was reported [20]. This technique provided models of protein structures (targets) and was ranked the best in CASP in all cases. This represents a significant advance from the previous editions of CASP [21, 22].

The AI method used provides a correct structure with a high accuracy score on the GDT\_TS (“Global Distance Test – Template Score”) scale used in CASP (it is the indicator used as a criterion for the degree of accuracy of model structure prediction against the target structure), which considers multiple similarity assessment criteria [3]. The AI technique uses a baseline in the form of maps of preferred inter-amino acid distances (contacts). Based on these, distances are reconstructed that are relevant to the given amino acid sequence.

However, the question of ‘Why proteins fold the way they do?’ remains unanswered. In this work, the degree of accuracy of the predicted structure was interpreted based on the consideration of the protein folding environment. The final structure of a given protein varies with the environment. Hence, using a steady-state internal force field (including preferred inter-amino acid distances) for all proteins may result in the success for one protein and a failure for another protein. The question arises as to why a jointly applied computational methods have not made any significant progress (WeFold project) [23, 24].

The Bhageerath-H server representing the hybrid *ab initio*/homology modelling reaching models of high accordance in CASP10 for water soluble monomeric proteins is freely available [25].

The three-track neural network linking a one-dimensional sequence level with the distance map, and the coordinate level expressing the integration of these three approaches was applied in CASP14 with good results [26].

Very popular recently techniques based on deep learning applied to the combination of template and multiple sequence alignment deliver some progress due to permanently significantly growing number of available structures making the comparable prediction more effective [27]. The interpretation of the physic-chemical 3D organisation of proteins in relation to structural and chemical classification of individual amino acids adopted to well defined homology (evolutionary) relationships is able to deliver the structures applicable for drug design techniques [28]. In the search for natural proteins sequence/structure relationships the nonnatural sequences are exploited allowing the distinguishing natural proteins on the basis of the resultant 3D structures [29]. The

combination of backbone dihedral angles and relative surface accessibility for tripeptides in relation to their N- and C-terminal relative positions in 3D structures applied to machine learning techniques tested on very large diverse proteins collection seems to be promising in the development of computational methods [30].

This approach is applied in the ProFitFun-Meta server, which is freely available [31, 32].

The analysis of the reasons for the variation in the degree of accuracy of the predicted structures was based on this work on the assumption of environmental contribution in the protein structuring process. A diverse environment governs the protein folding process. The absence of a diverse environment in the protein folding model eliminates a broad spectrum of structural forms. The possibility of considering conditions imposed by the environment is offered by the fuzzy oil drop (FOD) model, as well as its modified version (FOD-M) [33, 34]. The model quantitatively assesses the contribution of the polar water environment, as well as the contribution of the diverse environment affecting the protein structure formation. This differentiation in the environment, which affects the protein folding process in a different way, is assumed to answer the question as to why the results obtained are not uniformly correct, despite the use of programs (models) that perform positively in many cases.

The ab initio technique poses the challenge of finding the mechanism of the protein folding process that leads to the prediction of the correct structure by understanding such mechanism. This work demonstrates that this is not possible if the notation of the external force field of the environment origin is not considered. Environmental differentiation – as shown by previous analyses – has a significant impact on protein structuring [35, 36]. An averaged force field (averaged parameterisation) or any other criterion notation for protein structuring in an averaged form that does not take into account the effect of external factors cannot predict the diverse structural forms of proteins. The disadvantages resulting from the averaging of the parameterisation used can be demonstrated from the results of CASP competitions. The vast majority of the tools used in the CASP project provide results of similar status by eliminating certain specific groups of proteins, where, as it turns out, the introduction of a factor in the form of a differentiated external force field is needed.

## Materials and methods

### Data

The selection of proteins analysed is limited by the condition of the availability of the protein structures acting as targets in the CASP project and the models proposed by the CASP project participants. The subjective choice was also driven by the extremes: the best and the worst results.

A comparative analysis was carried out for the following:

1. The target structures (CASP13): T0953s2-D3 (PDB ID-6F45 [37]), T0990-D3 (PDB ID-6N9Y [38]), T1024 (PDB ID – 6T1Z [39]) and models for these targets [3]. The choice of these examples was driven by the extreme (high, medium and low) accordance of models in respect to target structures using the parameters based on the FOD model as the criteria to express the effect of the environment.

2. Proteins with an all helical structure are considered to be easy. However several other factors contribute to the computational structural model ability expressed by structural difficulty (SD) taking into account secondary structures, homology and physicochemical features of protein [40]. Also the availability of suitable template structure(s) influence the quality of prediction especially in comparative modelling.
3. Proteins with the code PDB ID—6POO [41] (GDT\_TS=65.90) and the PDB target ID—6UF2 [42] (GDT\_TS=45.8), whose structure was predicted with a low level of accuracy during CASP15 [3].
4. A comparative analysis of all models provided by the same technique (Baker-Rosettaserver – participant No. 368 [43]) demonstrating its limited potential for highly accurate prediction levels. The reason for the individual failures was identified as a lack of dependence on the environment affecting the protein folding process.

The database from which the results for this analysis were taken is available at <https://predictioncenter.org/> (accessed: April 26, 2023).

### Description of the FOD model

The external force field generated by the aqueous environment becomes apparent in the micellization process of bi-polar molecules, which, avoiding the entropically unfavourable contact of hydrophobic parts with polar water, form structures with an exposed polar surface, isolating hydrophobic fragments concentrated in the central part of the micelle. The description of such a hydrophobicity distribution is expressed by a 3D Gaussian function that spans the protein body:

$$H_i^T = \frac{1}{H_{sum}^T} \exp\left(\frac{-(x_i - \bar{x})^2}{2\sigma_x^2}\right) \exp\left(\frac{-(y_i - \bar{y})^2}{2\sigma_y^2}\right) \exp\left(\frac{-(z_i - \bar{z})^2}{2\sigma_z^2}\right) \tag{1}$$

By varying the magnitudes of the parameters  $\sigma_x$ ,  $\sigma_y$  and  $\sigma_z$ , it is possible to describe globular forms of arbitrary size and shape. ‘ $H_i$ ’ expresses the idealised hydrophobicity value assuming a micelle-like system. This value assigned to the position of the effective atom (the averaged position of the atoms that make up a given amino acid) is referred to as ‘ $T_i$ ’ (theoretical).

The actual hydrophobicity distribution resulting from the inter-amino acid interaction (which depends on the distance of the interacting residues and on their intrinsic hydrophobicity) is, to a varying degree, aligned with the idealised distribution expressed by the 3D Gaussian function [33]. The determination of the actual level of hydrophobicity constituting the interaction effect is expressed by an equation proposed in [44].

$$H_i^O = \frac{1}{H_{sum}^O} \sum_j \begin{cases} (H_i^r + H_j^r) \left(1 - \frac{1}{2} \left(7 \left(\frac{r_{ij}}{c}\right)^2 - 9 \left(\frac{r_{ij}}{c}\right)^4 + 5 \left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right) & \text{for } r_{ij} \leq c \\ 0, & \text{for } r_{ij} > c \end{cases} \tag{2}$$

where  $r_{ij}$  is the distance between the positions of the interacting amino acids,  $c$  is the cutoff distance, and  $H^r$  is the intrinsic hydrophobicity. The value of the observed level of hydrophobicity (assigned to the position of the effective atom) is referred to as ‘ $O_i$ ’.

The first factor in both expressions introduces the normalisation of the distributions. The T, O and R profiles are shown in Fig. 1A.

Quantitatively, the differences between the idealised distribution, T and the actual distribution, O can be assessed using divergence entropy [45]

$$D_{KL}(P|Q) = \sum_{i=1}^N P_i \log_2 \frac{P_i}{Q_i} \tag{3}$$

where P denotes the distribution under analysis (here, O) and Q denotes the reference distribution (here, T).

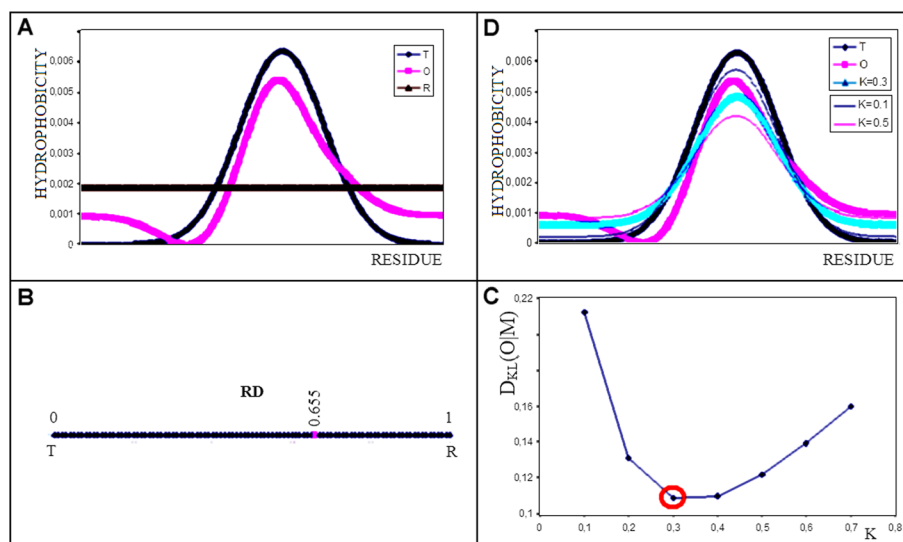
The  $D_{KL}$  value cannot be interpreted quantitatively. Therefore, a second reference distribution, R was introduced, where  $R_i = 1/N - N$  is the number of amino acids in the chain.

The R distribution, in contrast to the T distribution, represents a state with a uniformly distributed level of hydrophobicity (there is no hydrophobic core).

By comparing the values of  $D_{KL}(O|T)$  and  $D_{KL}(O|R)$ , the ‘proximity’ of the two distributions being compared can be determined. The relation  $D_{KL}(O|T) < D_{KL}(O|R)$  indicates the presence of a hydrophobic core. To avoid using two values to describe the same object, the quantity ‘RD’ (Relative Distance) is introduced:

$$RD = \frac{D_{KL}(O|T)}{D_{KL}(O|T) + D_{KL}(O|R)} \tag{4}$$

An RD value  $< 0.5$  indicates the presence of a hydrophobic core (Fig. 1B).



**Fig. 1** Graphical visualisation of the FOD-M model assumptions. **A** An example set of T (blue), O (pink) and R (brown) distributions. **B** The determined RD value for the example in (A) is 0.633, as shown on the axis. This value is calculated to express the difference expressed by  $D_{KL}(O|T)$  and  $D_{KL}(O|R)$  according to Eq. 4. **C** Determination of optimum value for K – the minimum  $D_{KL}$  value for different K values results in best fit. The  $K=0.3$  appears to be the optimal one for presented example. **D** The juxtaposition of the T (blue), O (pink) and M (cyan) distributions for  $K=0.3$  illustrates the interpretation of the M distribution, which most closely replicates the O distribution in the example in question. Additional (thin lines) represent the M distributions for  $K=0.1$  (thin line – pink) and M distribution for  $K=0.5$  (thin line – blue). The comparison of M distributions visualises the best fit for M distribution for  $K=0.3$

The deviations (the O distribution versus the T distribution) identified in protein distributions can be localised, where single and easily identifiable residues show a deficit or excess of hydrophobicity, as the case may be. If located in the common region of the protein molecule, residues with a hydrophobicity deficit in most cases are components of the active centre. Hydrophobicity deficits are often cavities ready to bind the ligand or the substrate (in the case of enzymes) [46]. A local hydrophobicity excess suggests the site of complex formation of a different protein [47]. Proteins with an O distribution similar to that of the T distribution were also identified. These are proteins with micelle-like structuring: down-hill, fast-folding, ultra-fast-folding and antifreeze type II [48].

The aqueous environment is not the only environment for protein activity.

Membrane proteins exhibit activity in a hydrophobic environment. Their stabilisation in this environment is ensured by the exposure of the hydrophobic residues to the outside (providing a preferable system with the hydrophobic membrane). If, in addition, the protein acts as an ion channel it has a concentration of polar residues in the central part (in particular). It is therefore an ‘inverted’ system in relation to proteins active in aqueous environments. Therefore, the idealised hydrophobicity distribution for the membrane environment is expressed by the complement of the 3D Gaussian function according to the equation below:

$$M_i = [T_{Max} - T_i]_n \tag{5}$$

where  $T_{Max}$  is the maximum value for the 3D Gaussian distribution and  $n$  is the normalisation of the resulting distribution.

As previous analyses showed, the proteins do not demonstrate an arrangement that follows the given distribution (Eq. 5). It appears that the proteins represent a structure that is a sort of consensus between the two forms, the  $T_i$ -compatible form and the  $M_i$ -compatible form. Therefore, the final distribution is determined by the following equation:

$$M_i = [T_i + K * (T_{Max} - T_i)]_n \tag{6}$$

The  $K$ -factor indicates the degree of contribution of the ‘inverted’ distribution to the distribution expressed by the 3D Gaussian function (denoted here as  $T$ ). This parameter expresses the strength with which a given environment modifies the system resulting from the polar water environment. The proper  $K$  value for particular set of  $T$  and  $O$  profiles is found as expressing the lowest  $D_{KL}(O|M)$  value (Fig. 1C). Finally the representation of  $M$  profile for  $T$  and  $O$  is shown in Fig. 1D.

More on membrane protein structure analysis is provided in [34, 49].

The graphical visualisation of the model in question (Fig. 1) illustrates the significance of the individual parameters and their interpretation.

The  $M$  distribution is therefore considered to be an expression of the effect of the environment on the protein folding process, in which the protein adapts to the conditions imposed by the environment.

If a juxtaposition of the distributions (Fig. 1) represented a hypothetical protein, this protein would be classified as deprived of a hydrophobic core.

It is also possible to perform an operation to eliminate the positions with the greatest differences between  $T_i$  and  $O_i$ . A multiple step-wise elimination of these positions allows

the identification of that part of the protein that exhibits  $RD < 0.5$ , thus identifying the part of the protein with a micelle-like organisation responsible for the solubility of the protein in question.

The summaries of successive editions of the CASP project show a split between ‘easy’ and ‘hard’ proteins. This distinction applies to all calculation techniques used by the participants. Traditionally, a high degree of prediction difficulty is associated with the presence of a beta-structure, which is more challenging (‘harder’) for obvious reasons (a long-range interaction). However, there are examples of proteins entirely representing the helical structure with a low prediction accuracy score. The search for an answer to the question posed earlier as to why very good force fields that provide models with a high degree of similarity to the target fail for other proteins is done based on the identification of the environmental differentiation of protein folding conditions.

### Programs used

The potential user has two possible ways to access the program:

The program allowing the calculation of RD as well as T and O distribution is accessible upon request on the CodeOcean platform:

<https://codeocean.com/capsule/3084411/tree>. Please contact the corresponding author to get access to your private program example.

The application—implemented in collaboration with the Sano Centre for Computational Medicine (<https://sano.science>) and running on resources contributed by ACC Cyfronet AGH.

(<https://www.cyfronet.pl>) in the framework of the PL-Grid Infrastructure (<https://plgrid.pl>)—provides a web wrapper for the abovementioned computational component and is freely available at <https://hphob.sano.science>.

The VMD program was used to present the 3D structures [50, 51].

## Results

### Selected models in the CASP13 project

The selection of examples for a detailed analysis was driven by the highest, medium and lowest model-to-target fit that ranked at the top of the list (per GDT\_TS). The evaluation criterion used in the current work is the parameters of the FOD model (Table 1).

The interpretation of the respective sets as given in the columns is as follows:

- The TARGET column: the parameter values characterise the structure that is available in the PDB. According to the interpretation based on the FOD model, the RD value reveals the degree of organisation of the hydrophobicity distribution against the micelle-like distribution. The K parameter indicates the contribution of non-aqueous factors that affect the formation of the structure of the protein in question.
- The BEST MODEL column represents the status of the structure of the model by assessing the extent to which the T distribution is reproduced by the O distribution as represented by the model structure.
- The correlation coefficient: the relationship of the GDT\_TS rating value to the status of all models expressed by the RD parameter values.

**Table 1** Values of RD and K parameters for models that had the top position on the ranking list (the GDT\_TS values)

TARGET			BEST MODEL			Correlation coefficient
ID	RD	K	K	RD	GDT_TS	
T0990-D3	0.528	0.5	0.6	0.575	48.71	-0.275
T1024	0.648	0.8	0.4	0.540	63.30	-0.466
T0953s-D3	0.286	0.1	0.3	0.468	43.01	-0.785

The determination of RD and K values for the target – the T and O distributions obtained for the structures as given in PDB. The correlation coefficient expresses the relationship of the rating (per GDT\_TS) of the model to the RD parameter value, which determines the degree of adaptation of the protein structure to environmental conditions

In the presented system, T0953s-D3 shows a structuring consistent with a micelle-like distribution (very low RD and K values). The structure of this protein is distinguished by the presence of a hydrophobic core and a polar surface. The contribution of non-aquatic factors is negligible. The values describing the status of the model ranked at the top of the GDT\_TS classification also show a high adaptation to a micelle-like system, albeit significantly lower than the target status. This difference turns out to be significant, as the model is very poor in the assessment, despite its top position on the ranking list.

The model structure obtained for the target T1065s2-D1 turns out to be very close to that of the protein. Here, the best model scored very highly on the GDT\_TS scale. The assessment based on the RD and K parameters also shows considerable similarity.

The target T0990-D3 shows structuring above the cut-off level (RD=0.5) to a small extent (K=0.5). The status of the proposed model appears to represent a distribution further away from the micelle-like system (higher RD and K values for the model against the target status). This difference results in a relatively low rating on the GDT\_TS scale.

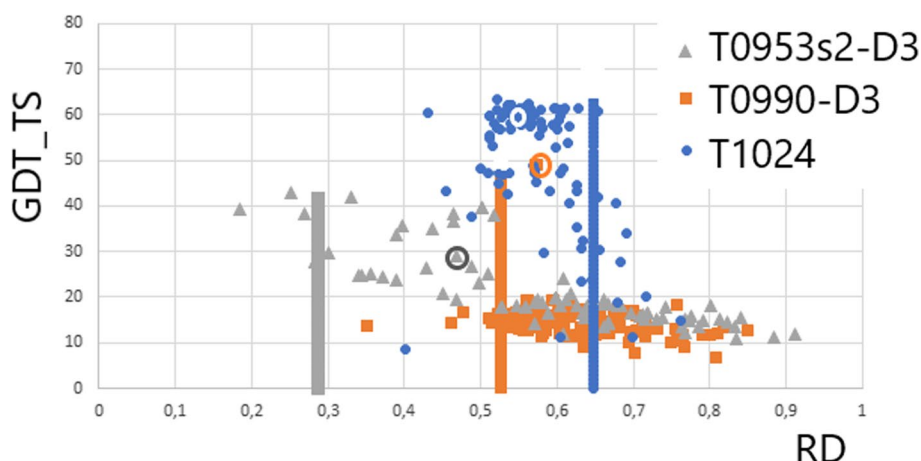
The target T1024 shows the highest variation against the top model in terms of RD and K. The top model was ranked with a relatively high GDT\_TS value (63.3).

The meaning of the ‘Correlation Coefficient’ column (Table 1) is shown in Fig. 2 and Table 1 provides very similar RD and K values since the top models are presented. The spread of the analysis of the model statuses (on the RD scale) varies widely across all participants of the CASP. A dependence with a negative correlation coefficient value is revealed, expressing a decreasing score (GDT\_TS) with an increasing RD that describes the status of the target. This means a lower rating (GDT\_TS) for models that do not take into account deviations from the FOD-ordered distribution. The correlation coefficient value is variable and depends on the status of the target.

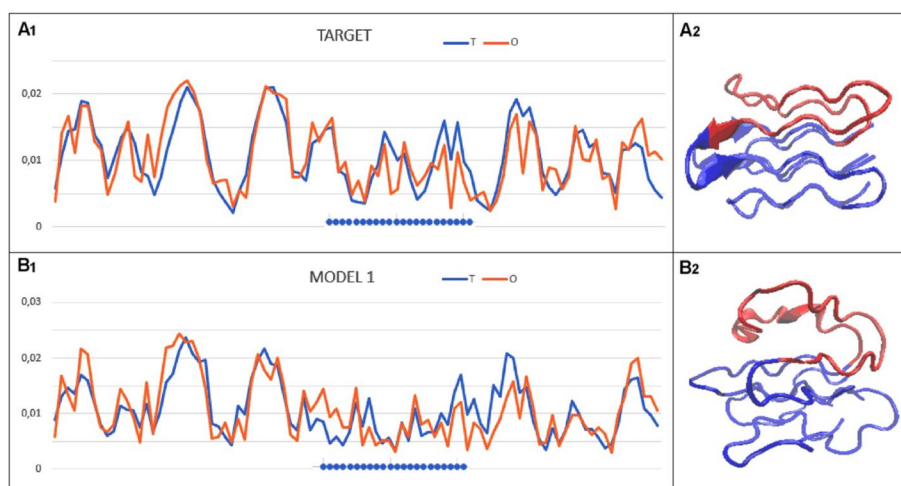
The classical GDT\_TS rating reduction relationship for RD values moving away from the target status is shown by the model obtained for the target T0953s2-D3 (Fig. 3), which is to the very low RD value that describes the status of the target. The relationship between the status expressed by RD and the GDT\_TS assessment is expressed by the correlation coefficient equal to  $CC = -0.785$  (Fig. 2). It means the larger the error in the RD status the lower the assessment of the model.

The positions circled in Fig. 2 are the results obtained with the AI method. It is apparent that the environment should also be considered in this method. This is particularly evident for the target T0953s2-D3.





**Fig. 2** Dependence of the GDT\_TS scale score on the status of the model protein structure expressed on the RD scale. The vertical lines are the RD values for the targets. The height of the vertical lines is the maximum score level on the GDT\_TS scale. The encircled positions are the results obtained with AlphaFold



**Fig. 3** Example of incompatibility for T0953s2-D3. The blue dots on the x-axis identify the area that does not reproduce the arrangement present in the target protein. The 3D presentation with a highlighted section of the chain where a significant discrepancy between the top-ranked model against the target is present

As an example of the type of incompatibility between a model and a target in terms of the FOD model, the set of T and O profiles for T0953s2-D3 representing the lowest GDT\_TS score is illustrated. Here, the degree of reproduction in the top-ranked model micelle-like system proved to be too poor. The central section of the chain (highlighted in red in the 3D image) in the proposed model contributes too little to the hydrophobic core.

To make the list of discussed examples complete, the structure of target T1065s2-D1 described by parameters: RD = 0.594 and K = 0.5 appeared to be very well predicted with GDT\_TS score = 98.47. This best model for this target represents the structure

of the status expressed by parameters  $RD = 0.578$  and  $K = 0.5$ . This example proves applicability of RD and K parameters as possible criteria for structure comparison.

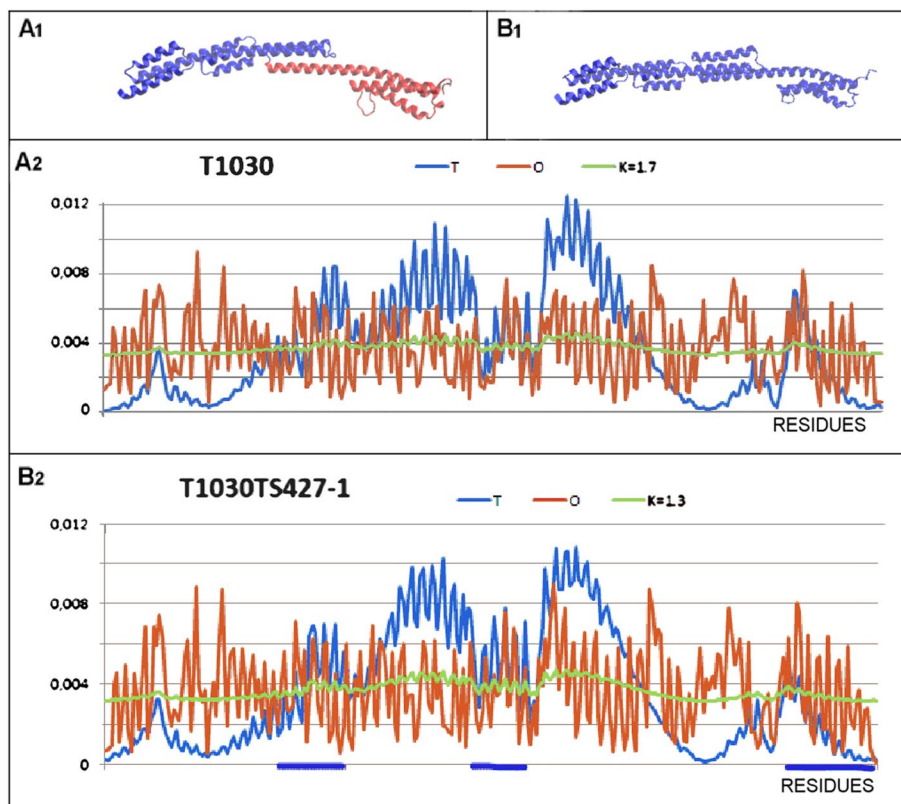
### Analysis of examples taken from CASP14 and CASP15

This analysis demonstrates the dependence of the result obtained (the top position on the ranking list) on the RD value of the target structure.

An extreme case of a globally distinct structure is a pathogenicity protein, putative from *Streptococcus agalactiae* serotype V from the *Streptococcus agalactiae* (PDB ID—6POO – target T1030 in CASP 14) [41] (Fig. 4).

The protein entirely represents an example of a helical protein with a structure far from globular. Therefore, a hydrophobic core is not expected to be present here. The structure of the protein (target) is described by the values:  $RD = 0.786$  with  $K = 1.7$  and, to determine the status of distinguished domains: (1–154)  $RD = 0.674$  and  $K = 0.8$ , while for the domain (155–273),  $RD = 0.658$  and  $K = 0.6$ .

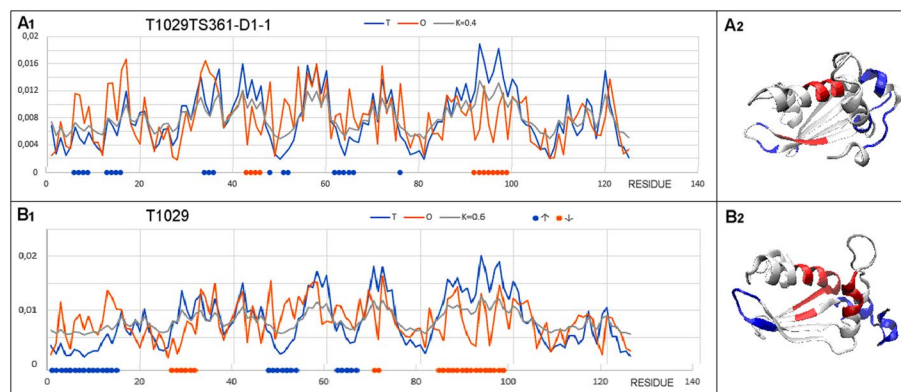
High RD values indicate a structuring devoid of a hydrophobic core, while the K values suggest a significant contribution of factors that are not water, driving the structuring. A compilation of the T, O and M distributions for the protein in question shows a nearly linear distribution (similar to the R distribution). This type of distribution is interpreted



**Fig. 4** Characteristics of the N-terminal helical domain of biba, a group b *Streptococcus* immunogenic bacterial adhesin (PDB ID—6POO). **A** 3D structure left – the target T1030; right – the model T1030TS427-1 **B** set of T, O and M profiles for the corresponding K representing the target T1030 **C** set of T, O and M profiles for the corresponding K, representing the model T1030TS427-1

**Table 2** Set of parameters based on the FOD model for the target T1030 and the model T1030TS427-1

	GDT_TS	RD-target/RD-model	K-target/K-model
D1	78.73	0.674/0.671	0.8/0.9
D2	89.5	0.658/0.640	0.6/0.6
CHAIN	63.0	0.786/0.757	1.7/1.3


**Fig. 5** Set of T, O and M profiles for the corresponding K, representing **A** Model T1029TS361-D1-1 **B** Target T1029

to be a result of conditions where there is no impact of an aquatic environment. The closer to the horizontal line the M distribution is, the lower the influence the aqueous environment has on the structuring. A protein with this type of M distribution can be thought of as a protein folding effect in a specific ‘water void’ environment.

A summary of the results (Table 2) shows the lowest GDT-TS score for the complete chain. Considering the nearly complete helical structure of this protein, the result turns out to be very poor. The very high values of the K parameter with large differences for the model and the target reveal the need to diversify the environment (its influence on the target is much higher than that obtained for the model).

An example that also requires a detailed discussion is the unknown function target category protein, T1029 (PDB ID—6UF2 [42]).

Complete with respective 3D presentations. Highlighted in red – fragments showing a hydrophobicity deficit; blue – hydrophobicity excess, as shown in the diagrams.

The selected example represents the model with the lowest GDT\_TS score in the set under discussion (Fig. 5). The parameters based on FOD-M for the target molecule are  $RD=0.622$  and  $K=0.6$  and  $GDT-TS=45.8$ , with  $RD=0.622$  and  $K=0.4$  for the model. The assessment of the accuracy of the model structuring according to the criteria based on the FOD model is higher against a very low score on the GDT\_TS scale. Rather, the identical RD values and the low variance of the K values suggest the accuracy of the model structure. When analysing the T, O and M profiles, a peculiar notation of property variation is seen, suggesting possible biological activity. It is indeed possible to identify a chain fragment whose status, revealing a local hydrophobicity deficit, suggests the presence of a cavity (the red fragments in Fig. 5) ready to interact with the ligand. It is

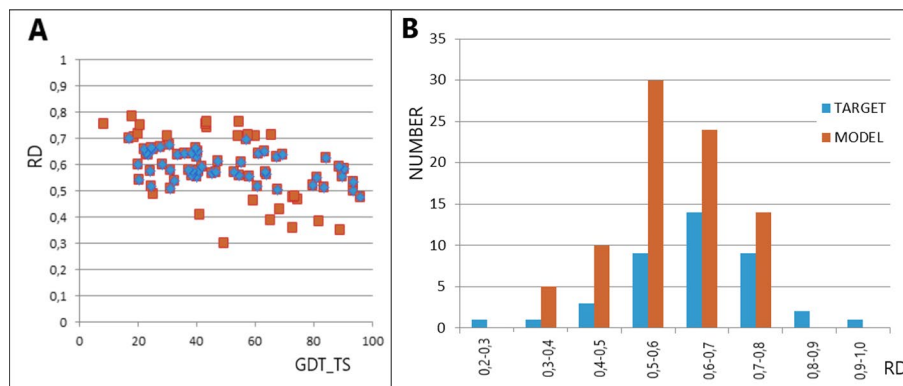
feasible to speculate on a possible interaction with another protein via an N-terminal chain fragment showing excess hydrophobicity. High RD and K values may suggest the need for other factors besides water to contribute. The location of sections of the O distribution significantly diverging from the T distribution suggests a similar design of the potential ligand binding cavity.

**Comparative analysis of results obtained by the same force field – Baker-Rosettaserver (participant No. 368) (CASP13)**

The model proposed in a group of programs whose history began with the ROSETTA program [52] is represented in a subsequent version modified in the form available with the Baker-Rosettaserver. The force field used in this program package is known very well. Alongside numerous successes, the group also delivered poorly rated results. It is therefore possible to pose the question as to why a very good force field fails in some cases. The search for an answer to this question was based on the analysis of a set of results in the form of models provided under ID 368 Baker-Rosettaserver in the CASP13 project [43].

The results of the evaluation of the models obtained with this server are presented in Fig. 6.

Analysis of the set of results (Fig. 6A) shows a correlation coefficient of -0.562. This is a result of very low scores for target protein status models with high RD values and two relatively good scores for low RD values. The summary in Fig. 6B reveals the status of the target proteins and the status of the models provided. The most numerous group of models represents the RD range of 0.5 < RD < 0.7, while the 0.6–0.7 range is the most numerous in the target group. This abundant presence is representative of the entire pool of proteins (an opinion expressed on the basis of the analysis of numerous proteins available in the publications of the I. Roterman team – the results have not been published). It can be speculated that the parameterisation used in the programs in the ROSETTA group (in particular the one used in the Baker-Rosettaserver – participant No. 368) was determined based on the analysis of multiple proteins, which justifies representing only this range of RD values.



**Fig. 6** Analysis of the models obtained with the Baker-Rosettaserver **A** relationship of the RD value to the GDT\_TS score. The blue dots represent the range of RD of highest representation as shown in **B**. **B** number of available targets and delivered models for the RD value ranges

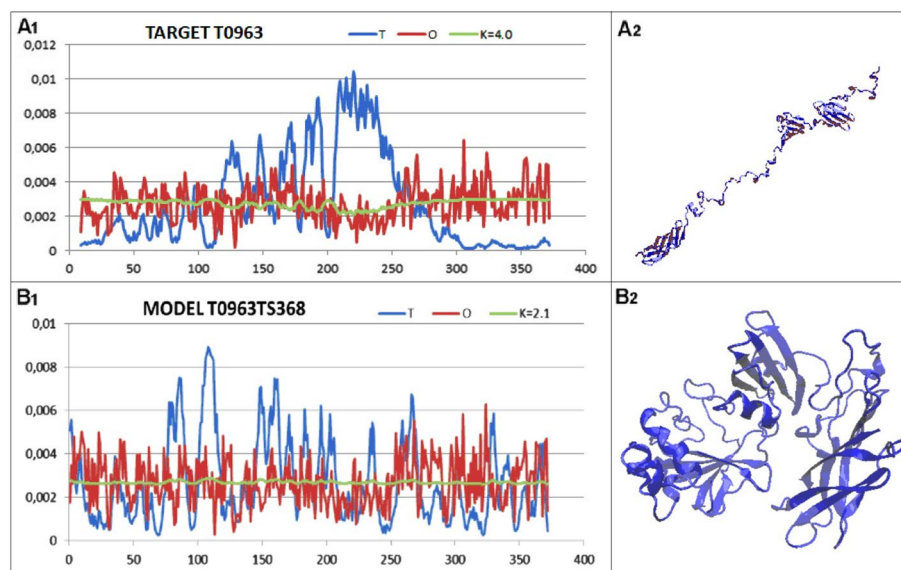
The summary in Fig. 6B reveals the absence of models with extreme statuses – low and high RD values. This observation reveals the need to vary the force field used to predict protein structures.

Of particular note are two structures (PDB ID–6CL6 [53] and PDB ID 6F45 [37]) whose native form shows extreme values of  $K=4.0$  for 6CL6 and  $K=0.1$  for 6F45. These examples represent cases that are in the target form but absent in the models provided (Fig. 6B) (with an RD-based classification).

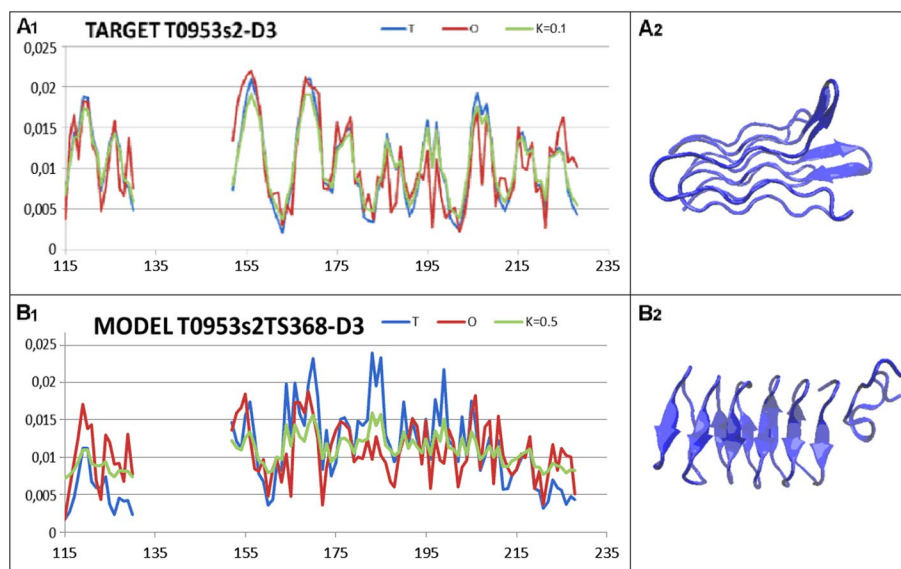
Analysis of the T, O and M profiles reveal the cause of the different structures of the target T0963, and the model provided by the server in question (Fig. 7). Notable is the fundamentally different RD value, with the target protein structure showing a significantly higher value. In contrast, a significantly higher value for the K parameter reveals the role of the environment. The target structure requires a suitable environment to stabilise this non-globular structure. It is clear that this structure would not be formed in an water environment. This is particularly evident in the N-terminal sections and particularly in the C-terminal sections, which show significantly higher hydrophobicity than the superficial location of these sections would suggest. This represents a preparation for interactions with other chains, which is in fact what happens when studying the biological activity of this protein [53]. This example clearly reveals the need of considering the presence of a certain external ‘rack’ in the form of non-aqueous factors stabilising this structural form that is far from globular.

A second example of the target whose status was not reproduced in the model is the target T0953s2-D2, due to its very low RD (target: 0.286) at RD (model: 0.522) (Fig. 8).

In the set of targets in the CASP13 edition, the example representing a micelle-like structure is the D3 domain (115–130, 152–228) (PDB ID 6F45). The RD value for the target is 0.286 (extremely low) with  $K=0.1$ , while the structure predicted by T0953s2TS368-D3 shows a structuring with  $RD=0.599$  and  $K=0.5$ . This is an example



**Fig. 7** Summary of the T, O and M profiles for the corresponding K values. **A** Target T0963 together with the 3D structure of the native form of the protein in question (PDB ID 6CL6) **B** – Model T0963TS368 together with the 3D structure proposed using the Baker-Rosettaserver (participant No. 368) [43]



**Fig. 8** LOW A set of the T, O and M profiles for the respective K values, together with a 3D presentation for: **A** target T0953s2-D3 **B** model T0953s2TS368-D3

of a structure determination by a force field, whose parameterisation is focused on the range of  $0.5 < RD < 0.7$ , as shown earlier. Here, the structure of the model could probably be reproduced by running simulations of the protein folding process in an aqueous environment that directs the process towards a micelle-like form [35].

### Discussion

The vast majority of programs in the ab initio category as well as the recently introduced AI-based method are driven by parameterisation resulting from the analysis of protein structures available in the PDB. The proteome of a particular organism is made up of proteins performing all the biological activities required to sustain life. Given the complexity of the system that is the world of living organisms, it is not surprising and indeed seems necessary to have a high degree of diversity in the tools that are proteins. Structural differentiation stems from the differences in amino acid sequences. In addition, the environment in which a protein performs its function is also a source of high diversity in the protein world. The environmental diversity has a very wide range, which provides a highly differentiating factor for the structures obtained. Therefore, a force field that does not consider external conditions or refers to ‘averaged’ characteristics remains deficient in relation to the great variety of tools and machines that are the proteins in every living organism.

The characterisation of the target T0953s2-D3 (CASP13), the structure of which no method could handle, shows a very low value of the K parameter = 0.1 in the native structure. This means that the force fields applied (which presumably are averaged) cannot reproduce the structure according to the micellization mechanism. As shown in Fig. 6 and Fig. 9, the range of characteristics best reproduced in the programs is between RD 0.5 and 0.7. According to the assessment to date, this is the range most highly represented by proteins of the proteome. Within this range are single-chain enzymes less

than 200 aa in length with a clearly localised incompatibility with the micelle-like system (an active centre). Enzymes with an extended structure (300–500 aa chains) and specific incompatibility with the micelle-like system are also in this range. The incompatibility consists of a scattering of small differences along the chain, without being able to distinguish the specific location of an explicit micelle-like incompatibility. This numerical predominance of proteins with the status expressed as  $0.5 < RD < 0.7$  indicates the reasons why the largest number of highly rated models precisely involve structures that, under conditions of activity, represent such a structure.

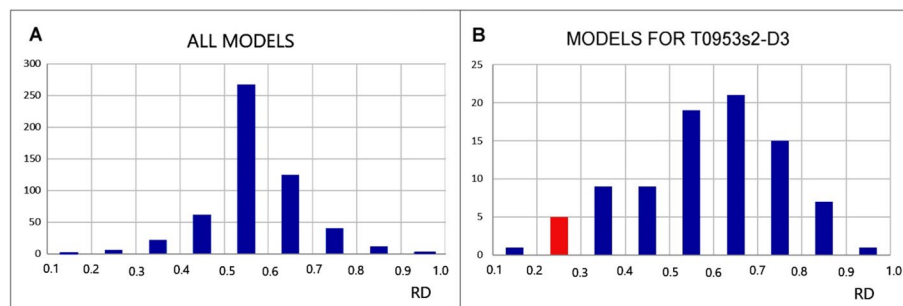
The need to consider environmental specificity for the folding protein is revealed by all amyloid proteins, whose change in structure is not caused by mutation but, it seems, solely by environmental effects.

The proposal to include a structure assessment in the RD and K category was already raised [54]. Structures rated highly by geometrical criteria (GDT\_TS) do not perform well in terms of specificity derived from the hydrophobicity system (RD and K), which is important for the biological functions performed by proteins.

A summary (Fig. 9) reveals the dominance of the parameterisation of the tools used over the range of variation in environmental conditions expressed as  $0.5 < RD < 0.7$ . According to analyses carried out by the authors of this work, this is the most common range of variation for this parameter. If the parameterisation for the tools used had been based on domain structures, the range would have been expressed with much lower RD values, as the domains show a large majority of structuring based on the presence of a hydrophobic core (low RD and low K values).

The summary of results (Fig. 9) reveals a dominant parameterisation for structures defined by the FOD-M model convention as representing a status with  $0.5 < RD < 0.7$ . This coverage is mostly present in large-scale analysis of proteins as available in PDB. The standardisation of force fields and other criteria for protein structure prediction seems to be common for all procedures independently of the method applied. The data base in the form of domains as the test set for parameterisation shall deliver the results from the range  $0.0 < RD < 0.4$  since this range has been identified for large scale analysis of domains as they are available in PDB.

In conclusion: the protein structures do not follow the averaged model. The differentiation of force fields expressing the external conditions shall be present in models oriented on protein structure prediction.



**Fig. 9** Summary of the results based on the examples discussed. The vertical axis expresses the number of models provided for a given range of RD values. **A** all examples discussed in this work. **B** target T0953s2-D3 representing the lowest RD=0.286 for the target status – the range distinguished by the red bar

The proteins presented in this analysis were selected subjectively, focusing mainly on examples that show significant differentiation in performance between groups and examples that score low, despite the use of models that perform very well for other targets.

The analysis of two examples with extreme RD values (0.286 and 0.918), whose structures were predicted to have RD values of 0.599 and 0.783 (with GDT\_TS scores of 43.01 and 24.66, respectively), demonstrate the validity of the thesis assumed here about the necessity of the environmental factor.

## Conclusion

As long as protein structure prediction-oriented programs do not take into account the presence of environmental conditions, the prediction is only reliable for a limited group of proteins (here defined by parameters based on the FOD model at  $0.5 < RD < 0.7$ ). The use of an averaged parameterisation determined from the analysis of known structures cannot lead to the prediction of protein structures, which are themselves differentiated as products of the contribution of the environment directing the protein folding process.

This work proposes a model that takes into account the environmental specificities to be considered in the construction of the external force field.

An external field for successive values of  $K$  (from  $K=0$  to  $K=1.0$  and even  $K>1.0$ ) applied to an energy optimisation procedure extended by optimisation due to the presence of an environment that directs the protein folding process appears to be necessary. An unambiguous example is amyloid proteins, which, with an unchanged sequence, acquire very different structural forms depending on the environment (shaking as the experimental procedure for amyloid production). These conclusions are also confirmed by the WeFold project, where combining the tools of the leading groups did not significantly change the validity of the results obtained [23]. The example shown in [35] also supports this conclusion.

## Abbreviations

PDB	Protein data bank
CASP#	Critical assessment of structure prediction with appropriate number of project edition
RD	Relative distance
GDT_TS	The name of scale applied by CASP model to express the degree of model-target similarity—Global Distance Test – Template Score

## Acknowledgements

Many thanks to Anna Śmietańska and Zdzisław Wiśniowski for technical support. This research was partially supported by the European Union's Horizon 2020 program under grant Sano No 857533. and Sano project carried out within the International Research Agendas program of the Foundation for Polish Science. co-financed by the European Union under the European Regional Development Fund.

## Author contributions

Conceptualization: IR and LK; methodology: KS; software: KS; validation: KS, formal analysis: IR, LK; investigation: IR; resources: IR; data curation: KS; writing — original draft preparation: KS and IR; writing — review and editing: KS and IR; visualization: IR; supervision: IR; project administration: IR; funding acquisition: IR. All authors have read and agreed to the published version of the manuscript.

## Funding

This research was funded by Jagiellonian University Medical College grant number N41/DBS/000722.

## Data availability

All data can be available on request addressed to the corresponding author. The program allowing calculation of RD is accessible on GitHub platform: <https://github.com/KatarzynaStapor/FODmodel> and on the platform <https://hphob.sano.science>.



## Declarations

### Ethics approval and consent to participate

Neither human nor animals data are used in the presented research.

### Consent to publication

Not applicable.

### Competing interests

The authors declare no conflict of interest.

Received: 13 September 2023 Accepted: 6 November 2023

Published online: 11 November 2023

## References

1. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. *Science*. 2012;338(6110):1042–6. <https://doi.org/10.1126/science.1219021>.
2. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved? *Curr Opin Struct Biol*. 2007;17(3):342–6. <https://doi.org/10.1016/j.sbi.2007.06.001>.
3. <https://predictioncenter.org/> (accessed Aug 7, 2023)
4. MacCallum JL, Pérez A, Schnieders MJ, Hua L, Jacobson MP, Dill KA. Assessment of protein structure refinement in CASP9. *Proteins*. 2011;79(Suppl 10):74–90. <https://doi.org/10.1002/prot.23131>.
5. Runthala A. Protein structure prediction: challenging targets for CASP10. *J Biomol Struct Dyn*. 2012;30(5):607–15. <https://doi.org/10.1080/07391102.2012.687526>.
6. Taylor TJ, Bai H, Tai CH, Lee B. Assessment of CASP10 contact-assisted predictions. *Proteins*. 2014;82(Suppl 2):84–97. <https://doi.org/10.1002/prot.24367>.
7. Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV. Assessment of CASP11 contact-assisted predictions. *Proteins*. 2016;84(Suppl 1):164–80. <https://doi.org/10.1002/prot.25020>.
8. Kryshtafovych A, Moulton J, Baslé A, Burgin A, Craig TK, Edwards RA, Fass D, Hartmann MD, Korycinski M, Lewis RJ, Lorimer D, Lupas AN, Newman J, Peat TS, Piepenbrink KH, Prahlad J, van Raaij MJ, Rohwer F, Segall AM, Seguritan V, Sundberg EJ, Singh AK, Wilson MA, Schwede T. Some of the most interesting CASP11 targets through the eyes of their authors. *Proteins*. 2016;84(Suppl 1):34–50. <https://doi.org/10.1002/prot.24942>.
9. Park H, DiMaio F, Baker D. CASP11 refinement experiments with ROSETTA. *Proteins*. 2016;84(Suppl 1):314–22. <https://doi.org/10.1002/prot.24862>.
10. Kryshtafovych A, Monastyrskyy B, Fidelis K, Moulton J, Schwede T, Tramontano A. Evaluation of the template-based modeling in CASP12. *Proteins*. 2018;86(Suppl 1):321–34. <https://doi.org/10.1002/prot.25425>.
11. Cheng J, Choe MH, Elofsson A, Han KS, Hou J, Maghrabi AHA, McGuffin LJ, Menéndez-Hurtado D, Olechnovič K, Schwede T, Studer G, Uziela K, Venclovas Č, Wallner B. Estimation of model accuracy in CASP13. *Proteins*. 2019;87(12):1361–77. <https://doi.org/10.1002/prot.25767>.
12. Simpkin AJ, Sánchez Rodríguez F, Mesdaghi S, Kryshtafovych A, Rigden DJ. Evaluation of model refinement in CASP14. *Proteins*. 2021;89(12):1852–69. <https://doi.org/10.1002/prot.26185>.
13. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018;46(W1):W296–303. <https://doi.org/10.1093/nar/gky427>.
14. Bitencourt-Ferreira G, de Azevedo WF. Homology modeling of protein targets with MODELLER. *Methods Mol Biol*. 2019;2053:231–49. [https://doi.org/10.1007/978-1-4939-9752-7\\_15](https://doi.org/10.1007/978-1-4939-9752-7_15).
15. Studer G, Tauriello G, Bienert S, Biasini M, Johnner N, Schwede T. ProMod3-A versatile homology modelling toolbox. *PLoS Comput Biol*. 2021;17(1):e1008667. <https://doi.org/10.1371/journal.pcbi.1008667>.
16. Chivian D, Robertson T, Bonneau R, Baker D. Ab initio methods. *Methods Biochem Anal*. 2003;44:547–57. <https://doi.org/10.1002/0471721204.ch27>.
17. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol*. 2007;5:17. <https://doi.org/10.1186/1741-7007-5-17>.
18. Hardin C, Pogorelov TV, Luthey-Schulten Z. Ab initio protein structure prediction. *Curr Opin Struct Biol*. 2002;12(2):176–81. [https://doi.org/10.1016/s0959-440x\(02\)00306-8](https://doi.org/10.1016/s0959-440x(02)00306-8).
19. Bonneau R, Baker D. Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct*. 2001;30:173–89. <https://doi.org/10.1146/annurev.biophys.30.1.173>.
20. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577(7792):706–10. <https://doi.org/10.1038/s41586-019-1923-7>.
21. Quraishi M. AlphaFold at CASP13. *Bioinformatics*. 2019;35(22):4862–5. <https://doi.org/10.1093/bioinformatics/btz422>.
22. Callaway E. Deepmind's AI predicts structures for a vast trove of proteins. *Nature*. 2021;595:635. <https://doi.org/10.1038/d41586-021-01968-y>.
23. Khoury GA, Liwo A, Khatib F, Zhou H, Chopra G, Bacardit J, Bortot LO, Faccioli RA, Deng X, He Y, Krupa P, Li J, Mozolewska MA, Sieradzian AK, Smadbeck J, Wirecki T, Cooper S, Flatten J, Xu K, Baker D, Cheng J, Delbem AC, Floudas CA, Keasar C, Levitt M, Popović Z, Scheraga HA, Skolnick J, Crivelli SN, Foldit Players. WeFold: a cooperation for protein structure prediction. *Proteins*. 2014;82(9):1850–68. <https://doi.org/10.1002/prot.24538>.

24. Keasar C, McGuffin LJ, Wallner B, Chopra G, Adhikari B, Bhattacharya D, Blake L, Bortot LO, Cao R, Dhanasekaran BK, Dimas I, Faccioli RA, Faraggi E, Ganzynkowicz R, Ghosh S, Ghosh S, Gieldoń A, Golon L, He Y, Heo L, Hou J, Khan M, Khatib F, Khoury GA, Kieslich C, Kim DE, Krupa P, Lee GR, Li H, Li J, Lipska A, Liwo A, Maghrabi AHA, Mirdita M, Mirzaei S, Mozolewska MA, Onel M, Ovchinnikov S, Shah A, Shah U, Sidi T, Sieradzan AK, Ślusarz R, Ślusarz R, Smadbeck J, Tamamis P, Trieber N, Wirecki T, Yin Y, Zhang Y, Bacardit J, Baranowski M, Chapman N, Cooper S, Defelibus A, Flatten J, Koepnick B, Popović Z, Zaborowski B, Baker D, Cheng J, Czaplowski C, Delbem ACB, Floudas C, Kloczkowski A, Oldziej S, Levitt M, Scheraga H, Seok C, Söding J, Vishveshwara S, Xu D, Crivelli SN. An analysis and evaluation of the WeFold collaborative for protein structure prediction and its pipelines in CASP11 and CASP12. *Sci Rep*. 2018;8(1):9939. <https://doi.org/10.1038/s41598-018-26812-8>.
25. Jayaram B, Dhingra P, Mishra A, Kaushik R, Mukherjee G, Singh A, Shekhar S. Bhageerath-H: a homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. *BMC Bioinformatics*. 2014;15(Suppl 16):S7. <https://doi.org/10.1186/1471-2105-15-S16-S7>.
26. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, Avan Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 2021;373(6557):871–6. <https://doi.org/10.1126/science.abj8754>.
27. Wu F, Jing X, Luo X, Xu J. Improving protein structure prediction using templates and sequence embedding. *Bioinformatics*. 2023;39(1):btac723. <https://doi.org/10.1093/bioinformatics/btac723>.
28. Kaushik R, Singh A, Jayaram B. Where Informatics Lags Chemistry Leads. *Biochemistry*. 2018;57(5):503–6. <https://doi.org/10.1021/acs.biochem.7b01073>.
29. Kaushik R, Zhang KYJ. A protein sequence fitness function for identifying natural and nonnatural proteins. *Proteins*. 2020;88(10):1271–84. <https://doi.org/10.1002/prot.25900>.
30. Kaushik R, Zhang KYJ. ProFitFun: a protein tertiary structure fitness function for quantifying the accuracies of model structures. *Bioinformatics*. 2022;38(2):369–76. <https://doi.org/10.1093/bioinformatics/btab666>.
31. <http://github.com/KYZ-LSB/ProFitFun-Meta>
32. Kaushik R, Zhang KYJ. An integrated protein structure fitness scoring approach for identifying native-like model structures. *Comput Struct Biotechnol J*. 2022;20(20):6467–72. <https://doi.org/10.1016/j.csbj.2022.11.032.eCollection>.
33. Konieczny L, Roterman I. Description of the fuzzy oil drop model. In: Roterman-Konieczna I, editor. *From globular proteins to amyloids*. Elsevier; 2020. p. 1–12.
34. Roterman I, Stapor K, Fabian P, Konieczny L, Banach M. Model of environmental membrane field for transmembrane proteins. *Int J Mol Sci*. 2021;22(7):3619. <https://doi.org/10.3390/ijms22073619>.
35. Roterman I, Sieradzan A, Stapor K, Fabian P, Wesolowski P, Konieczny L. On the need to introduce environmental characteristics in ab initio protein structure prediction using a coarse-grained UNRES force field. *J Mol Graph Model*. 2022;114: 108166. <https://doi.org/10.1016/j.jmgm.2022.108166>.
36. Roterman I, Konieczny L. Protein is an intelligent Micelle. *Entropy (Basel)*. 2023;25(6):850. <https://doi.org/10.3390/e25060850>.
37. Dunne M, Denyes JM, Arndt H, Loessner MJ, Leiman PG, Klumpp J. Salmonella phage S16 tail fiber adhesin features a rare polyglycine rich domain for host recognition. *Structure*. 2018;26(12):1573–1582.e4. <https://doi.org/10.1016/j.str.2018.07.017>.
38. Kerviel A, Ge P, Lai M, Jih J, Boyce M, Zhang X, Zhou ZH, Roy P. Atomic structure of the translation regulatory protein NS1 of bluetongue virus. *Nat Microbiol*. 2019;4(5):837–45. <https://doi.org/10.1038/s41564-019-0369-x>.
39. Debruycker V, Hutchin A, Masureel M, Ficici E, Martens C, Legrand P, Stein RA, Mchaourab HS, Faraldo-Gómez JD, Remaut H, Govaerts C. An embedded lipid in the multidrug transporter LmrP suggests a mechanism for polyspecificity. *Nat Struct Mol Biol*. 2020;27(9):829–35. <https://doi.org/10.1038/s41594-020-0464-y>.
40. Kaushik R, Jayaram B. Structural difficulty index: a reliable measure for modelability of protein tertiary structures. *Protein Eng Des Sel*. 2016;29(9):391–7. <https://doi.org/10.1093/protein/gzw025>.
41. Manne K, Chattopadhyay D, Agarwal V, Blom AM, Khare B, Chakravarthy S, Chang C, Ton-That H, Narayana SVL. Novel structure of the N-terminal helical domain of BibA, a group B streptococcus immunogenic bacterial adhesin. *Acta Crystallogr D Struct Biol*. 2020;76(Pt 8):759–70. <https://doi.org/10.1107/S2059798320008116>.
42. Zhang N, Liwang AL.—PDB ID 6UF2
43. <https://robetta.bakerlab.org/> - accessed Aug 15, 2023
44. Levitt MA. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*. 1976;104(1):59–107. [https://doi.org/10.1016/0022-2836\(76\)90004-8](https://doi.org/10.1016/0022-2836(76)90004-8).
45. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat*. 1951;22(1):79–86. <https://doi.org/10.1214/aoms/1177729694>.
46. Banach M, Konieczny L, Roterman I. Ligand binding cavity encoded as a local hydrophobicity deficiency. In: Roterman-Konieczna I, editor. *From globular proteins to amyloids*. Elsevier; 2020. p. 91–4.
47. Banach M, Konieczny L, Roterman I. Protein-protein interaction encoded as an exposure of hydrophobic residues on the surface. In: Roterman-Konieczna I, editor. *From globular proteins to amyloids*. Elsevier; 2020. p. 79–90.
48. Banach M, Stapor K, Konieczny L, Fabian P, Roterman I. Downhill, ultrafast and fast folding proteins revised. *Int J Mol Sci*. 2020;21(20):7632. <https://doi.org/10.3390/ijms21207632>.
49. Roterman I, Stapor K, Fabian P, Konieczny L. The functional significance of hydrophobic residue distribution in bacterial beta-barrel transmembrane proteins. *Membranes*. 2021;11(8):580. <https://doi.org/10.3390/membranes11080580>.
50. <https://www.ks.uiuc.edu/Research/vmd/> - accessed Dec 2022
51. Humphrey W, Dalke A, Schulten K. VMD - visual molecular dynamics. *J Molec Graphics*. 1996;14:33–8.
52. Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, Aprahamian M, Baker D, Barlow KA, Barth P, Basanta B, Bender BJ, Blacklock K, Bonet J, Boyken SE, Bradley P, Bystruff C, Conway P, Cooper S, Correia BE, Coventry B, Das R, De Jong RM, DiMaio F, Dsilva L, Dunbrack R, Ford AS, Frenz B, Fu DY, Geniesse C, Goldschmidt L,

- Gowthaman R, Gray JJ, Gront D, Guffy S, Horowitz S, Huang PS, Huber T, Jacobs TM, Jeliaskov JR, Johnson DK, Kappel K, Karanicolas J, Khakzad H, Khar KR, Khare SD, Khatib F, Khramushin A, King IC, Kleffner R, Koepnick B, Kortemme T, Kuenze G, Kuhlman B, Kuroda D, Labonte JW, Lai JK, Lapidoth G, Leaver-Fay A, Lindert S, Linsky T, London N, Lubin JH, Lyskov S, Maguire J, Malmström L, Marcós E, Marcu O, Marze NA, Meiler J, Moretti R, Mulligan VK, Nerli S, Norn C, Ó'Conchúir S, Ollikainen N, Ovchinnikov S, Pacella MS, Pan X, Park H, Pavlovicz RE, Pethe M, Pierce BG, Pilla KB, Raveh B, Renfrew PD, Burman SSR, Rubenstein A, Sauer MF, Scheck A, Schief W, Schueler-Furman O, Sedan Y, Sevy AM, Sgourakis NG, Shi L, Siegel JB, Silva DA, Smith S, Song Y, Stein A, Szegegy M, Teets FD, Thyme SB, Wang RY, Watkins A, Zimmerman L, Bonneau R. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods*. 2020;17(7):665–80. <https://doi.org/10.1038/s41592-020-0848-2>.
53. Buth SA, Mikhail M, Shneider MM, Dean Scholl D, Petr G, Leiman PG. Structure and analysis of R1 and R2 pyocin receptor-binding fibers. *Viruses*. 2018;10(8):427. <https://doi.org/10.3390/v10080427>.
54. Gadzała M, Kalinowska B, Banach M, Konieczny L, Roterman I. Determining protein similarity by comparing hydrophobic core structure. *Heliyon*. 2017;3(2): e00235. <https://doi.org/10.1016/j.heliyon.2017>.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

