Definition and classification of evaluation units for tertiary structure prediction in CASP12 facilitated through semi-automated metrics

Luciano A. Abriata1 | Lisa N. Kinch | Giorgio E. Tamò1 | Bohdan Monastyrskyy3 | Andriy Kryshtafovych3 | Matteo Dal Peraro1

1Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
2Howard Hughes Medical Institute, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas
3Genome Center, University of California, Davis, California

Abstract
For assessment purposes, CASP targets are split into evaluation units. We herein present the official definition of CASP12 evaluation units (EUs) and their classification into difficulty categories. Each target can be evaluated as one EU (the whole target) or/and several EUs (separate structural domains or groups of structural domains). The specific scenario for a target split is determined by the domain organization of available templates, the difference in server performance on separate domains versus combination of the domains, and visual inspection. In the end, 71 targets were split into 96 EUs. Classification of the EUs into difficulty categories was done semi-automatically with the assistance of metrics provided by the Prediction Center. These metrics account for sequence and structural similarities of the EUs to potential structural templates from the Protein Data Bank, and for the baseline performance of automated server predictions. The metrics readily separate the 96 EUs into 38 EUs that should be straightforward for template-based modeling (TBM) and 39 that are expected to be hard for homology modeling and are thus left for free modeling (FM). The remaining 19 borderline evaluation units were dubbed FM/TBM, and were inspected case by case. The article also overviews structural and evolutionary features of selected targets relevant to our accompanying article presenting the assessment of FM and FM/TBM predictions, and overviews structural features of the hardest evaluation units from the FM category. We finally suggest improvements for the EU definition and classification procedures.

KEYWORDS
critical assessment of structure prediction, residue coevolution, contact prediction, homology modeling, sequence alignment, structural bioinformatics

1 INTRODUCTION
The biannual Critical Assessment of Structure Prediction (CASP) aims to provide an objective evaluation of state-of-the-art methodologies in protein structure prediction.1 Participants submit models for targets whose structures are unknown or withheld from public release during the experiment; then independent teams of assessors evaluate the submitted models along different tracks. In this article, we present the CASP12 targets, the procedure to split them into evaluation units (EU) against which models are assessed, and the classification of these units into one of three difficulty categories.

In this round, CASP organizers secured 71 experimental structures (Table 1) for predictors to model, thanks to contributions from multiple groups as listed in Table 2. Based on a number of metrics and protocols further described in this article, the 71 targets were split into 96 EUs. We introduced in this CASP a way of assigning difficulty to the EUs based on objective metrics, so as to semi-automate the process. Our
metrics, available for future CASPs at the Prediction Center,² include one that captures how well the automated servers perform, another that measures sequence-level similarity to Protein Data Bank (PDB) entries that could be used as templates, and a third one that captures structural similarity to PDB entries disregarding sequence similarity. These metrics enabled easy discrimination of easy and hard targets, allowing experts to focus on targets of intermediate difficulty.

We herein cover the definition of EUs and their classification into three difficulty categories based on combined metrics, with a focus on the sequence/structure relationships of the intermediate cases, which were harder to classify. We overview coarse structural features of hard and intermediate targets, show examples of the hardest targets, and describe their structural and evolutionary features. Supporting Information Table S1 provides an overview of the EUs, depicting their CASP12 classification and their relationship to existing PDB structures according to concepts outlined in the evolutionary classification of structure domains (ECOD) database.³

2 | MATERIALS AND METHODS

All GDTTS, HHpred, LGA, HHblits, Grishin plot, and Neff calculations used in this work were performed by the Prediction Center; thus they will be readily available in future CASP rounds.²

2.1 | Difficulty metrics

In this round of CASP, we introduced a combination of metrics that facilitate classification of the evaluation units. These metrics are the HHpred score, which measures the sequence-level similarity to PDB entries, and the LGA score, which measures structural similarity to PDB entries.

The HHpred score for an EU is defined as a product of the raw HHpred probability of the top hit obtained in a HHpred⁴ run of its protein sequence against the sequences of all PDB entries available during the target prediction window and the percentage of the query sequence covered by this hit. The LGA score for an EU is the LGA_S score of the highest-scoring structure in structure-independent LGA⁵ runs of the target structure against all PDB entries available before closing the prediction window for the target. Both metrics range from 0 for no sequence or structure similarity to 100% for perfect sequence (HHpred score) or structure (LGA score) match. The combined metric is simply an average of the HHpred and LGA scores:

\[
\text{Combined metric} = \frac{\text{HHpred score} + \text{LGA score}}{2}
\]

Besides the difficulty metrics based on sequence- and structure-level matches to the PDB, part of the classification into difficulty categories is based on the actual performance of the top 20 server predictions. The metric used to quantify the quality of individual models was the Global Distance Test Total Score (GDTTS),⁶ which reports an average of the maximum number of residues that can be superimposed under cutoffs of 1, 2, 4, and 8 Å, normalized by the number of residues in the target. The GDTTS score is defined in the 0–100 range. Structurally wrong models usually score below 20 GDTTS points, while a perfect model that matches the full structure within 1 Å deviation in the Cα coordinates of all residues scores 100 GDTTS points. The GDTTS measure is historically the main metric for global assessment of tertiary structure predictions in CASP.⁷

2.2 | Grishin plots for objective evaluation of domain splits

An evaluation scheme to help objectively decide if individual proposed EUs corresponding to structure domains should be kept separate or merged into a single EU was introduced in CASP9.⁸ The so-called “Grishin plot” (called after the name of the CASP9 assessor, who introduced it) allows easy comparison of the weighted sum of GDTTS scores for each individual domain versus the GDTTS score for the

### TABLE 1 CASP12 Tertiary Structure Prediction Track in Numbers

| Number of targets | 71 |
|-------------------|----|
| Evaluation units  | 96 |
| TBM               | 38 |
| FM                | 39 |
| FM/TBM            | 19 |

### TABLE 2 Origin of CASP12 Targets

| CSIRO (Australia) | U. Milano (Italy) | Argonne Lab (USA) |
|-------------------|-------------------|-------------------|
| U Campinas (Brasil) | B.R.C.S., RigaLatvia | U.C. Irvine (USA) |
| U. Kobenhavn (Denmark) | N.C.I., Amsterdam (Netherlands) | B.C.M., Houston (USA) |
| U. Marseille (France) | U. Lisbon (Portugal) | U.C. San Francisco (USA) |
| U. Lyon (France) | K.A.U.S.T. (Saudi Arabia) | U. Rice (USA) |
| Max Planck (Germany) | C.S.I.S. (Spain) | U.C. Los Angeles (USA) |
| Tübingen (Germany) | U. Basel (Switzerland) | Columbia U (USA) |
| U. Limerick (Ireland) | U. Birmingham (UK) | U. Washington (USA) |
| Weizmann (Israel) | U. Newcastle (UK) | U.C. San Francisco (USA) |
| i.e.O. (Italy) | U. Oxford (UK) | U. Maryland (USA) |
| S.G.V. (Canada) | | |
merged domains. If predictions on individual domains perform better than on the merged domains, then dots fall above the diagonal, indicative that the domains should be considered separate EUs (example in Figure 1A). If predictions on separate domains are not better than the performance on the merged domain, then dots fall on the diagonal, in which case we do not split (example in Figure 1D). This way one keeps EUs as large as possible, so that if predictors got large portions correctly modeled we would eventually reward them upon assessment and also place accent on predicting correct inter-domain orientation.

2.3 | Evaluation of alignment depths

Given that the assessment of tertiary structure prediction revealed that deeper alignment depths (Neff) tend to correlate with better accuracy of models, we briefly investigated here whether Neff can be used as an additional metric for defining difficulty. The Neff values for this article were computed with the HHblits program on the uniprot20 database with an E-value threshold of $10^{-3}$, three iterations and a minimum coverage of 60%.

3 | RESULTS AND DISCUSSION

3.1 | Definition of evaluation units from target structures

The tertiary structure prediction track of CASP traditionally evaluates targets that are split into evaluation units (EUs) rather than the whole structures. The splitting minimizes penalties that arise from differences
in relative domain orientation and separates regions of different prediction difficulty. However, it is important to keep EUs as large as possible, such that when predictors get large portions of the target correctly modeled (spawning what we would in principle propose as multiple EUs) the assessors can reward such predictors.

As in previous CASP editions,\textsuperscript{8–10} splitting of target structures into EUs was carried out through consensus among multiple human inspectors, including the authors of this work and the quaternary structure assessment team. The splitting procedure starts from running automatic programs for structural domain partition (DomainParser\textsuperscript{11} and DDomain\textsuperscript{12}) and checking the agreement between the suggested domain boundaries. These runs (followed by a visual check) allowed us to quickly identify trivial one-domain cases. Next, we searched for target templates using the sequence-based PSI-BLAST\textsuperscript{13} and HHpred\textsuperscript{4} programs. The identified templates were checked against the targets for regions of sequence match, compactness of secondary structure elements and hydrophobic cores, presence of internal repeats and continuity of sequence. Results of the template analysis were confronted with the results of automatic structural domain partitions to verify supposed domain composition of the targets and suggest their split. Each proposed split was tested through Grishin plots,\textsuperscript{8} which compare server performance on separate domains to server performance on merged domains (for details see the Materials and Methods section). This test is especially important when a split was apparent but server performance was high on joint domains or full structures. In such cases, by not splitting the domain into EUs we favor predictions that are good over larger portions of the target.

Figure 1 shows some of the examples of the EU definition. Target T0893 is a straightforward case. Both visually and from the Grishin plots, it consists of two evaluation domains. Target T0886 is also relatively straightforward. It displays three clear domains, but the third one consists of a single α-helix extended away from the globular core, and as such is of little interest to a tertiary structure assessment, so it was not considered for the evaluation. Regarding the two other domains, visual inspection and Grishin plot agree that they should be two separate EUs. We note that domain 1 constitutes one of the very few examples where definition of an EU does not imply sequence continuity but rather structural continuity. Targets T0920 and T0861 involved more discussion when defining their domains. T0920 visually looks like having two domains, but Grishin plots are ambiguous, as a group of predictions reached GDTTS scores of about 40 if both domains are considered as a single EU, but other better predictions reached GDTTS around 50 if the domains are considered as two separate EUs. Mostly based on the visual inspection and templates, this target was split in two EUs. Last, target T0861 visually looks like having two domains that could be assigned different EUs, but the Grishin plot indicates that most predictions are very good even if the target is considered as a whole, therefore it was kept as a single EU.

### 3.2 Classification of evaluation units into difficulty categories

Splitting of each target into EUs was followed by assignment of each unit to a difficulty category. Initially, all EUs were divided into two broad difficulty categories based on the average GDTTS score from the top 20 server predictions. The EUs with scores above 50 were preliminarily classified as not requiring detailed manual assessment (easier targets) and those with the score below 50 were preliminarily classified as requiring such an assessment. The split along the server performance lines traditionally corresponds to classification into template-based and free modeling targets, with commonly noted exceptions as in CASP11.\textsuperscript{9} Obviously, the correlation between the performance-based classification and the template-based classification is least reliable for the EUs with the scores around the division line. To help us in decision making, we introduced plots that correlate the average GDTTS score of the top 20 server predictions against the HHpred and LGA scores (Figure 2 and interactive version online at http://predictioncenter.org/casp12/domains_summary_paper.html). The HHpred score measures how likely it is to find a PDB template of high sequence similarity and coverage to the target sequence (Figure 2A); whereas the LGA score measures how similar the EU is to the closest PDB match regardless of sequence identity (Figure 2B). Panel A advises weak (arbitrary) clustering of EUs into two main groups depicted as boxes with gray dashed borders. One group encompasses EUs with average server GDTTS scores between 30 and 100 and template HHpred scores ranging from 40 to 100; the other group embraces EUs for which sequence matches are worse than 40 and average server GDTTS scores are below 60. Such a trend readily highlights the need for good PDB matches at the sequence level to achieve good 3D models, on average for all participating servers, but also indicates cases of quite successful modeling even in the absence of strong sequence matches (points with GDTTS score close to 60 despite very low HHpred scores). Panel B shows that having template structures that look like the target results in better models, with the caveat that many of the server models are not as good as expected from the close structural similarity of available targets (points at high template LGA scores but average model GDTTS scores barely reaching 45).

### 3.3 Three difficulty levels: Template-based modeling (TBM), free modeling (FM), and special cases of intermediate difficulty (FM/TB)

Based on the plots of average server performances against HHpred and LGA scores in Figure 2A, B, we reasoned that both the existence of good sequence and structure matches to the PDB provide metrics to quantify target difficulty. We therefore combined the HHpred and LGA scores as an average, obtaining a new plot (Figure 2C) with a quite smooth correlation against average server GDTTS. From this plot we set up boundaries (boxes with gray dashed borders in Figure 2C) from which we defined the easier, template-based modeling (TBM) EUs as those for which the combined HHpred-LGA score was higher than 60 and the average server GDTTS was above 50 (at which level the global topology usually begins to be visually evident, red points in Figure 2C). These EUs were considered as suitable for the evaluation not requiring heavy engagement of human assessors. We next defined the more difficult, free-modeling (FM) EUs as those for which the combined HHpred-LGA score was lower than 60 and the average server
prediction was worse than 50 (blue points in Figure 2C). These were the EUs that definitely required expert judgment on the submitted models.

EUs outside these boundaries exhibited significant deviations from expected performance and were classified as FM/TBM, while certain borderline cases within the TBM and FM definitions that presented sequence or structure deviations from possible templates were reclassified, after detailed inspection, from TBM (T0912D2 and T0945D1) or FM (T0874D1, T0896D2, T0909D1, and T0868D1) into FM/TBM as well (green in Figure 2A). This definition results in the FM/TBM EUs distributing nearly orthogonal to the main trend, such that those with higher average score are actually predicted worse by the servers. In fact, these FM/TBM targets are harder than TBM because of fold changes and other effects relative to PDB entries, as detailed below and exemplified in Figure 3. These targets were considered as potentially suitable for the automatic-only evaluation alongside the TBM targets, at the same time deserving a more rigorous human assessment typical for free modeling.

### 3.4 Difficult to classify FM/TBM domain examples

Several FM/TBM EUs exhibit high sequence-based scores, yet automated server models display relatively low performance (green dots with high HHpred score in Figure 2A: T0868D1, T0896D1, T0896D2, T0945D1, T0874D1, T0875D1, T0876D1, T0884D1, T0887D1, T0890D1, T0891D1, T0896D2, T0901D1, T0907D1, T0912D2, T0943D1, T0945D1, T0943D1).
T0896D2, T0898D2, T0901D1, T0909D1, and T0912D2). For these FM/TBM EUs, unsupervised choice of template and homology modeling by the servers leads to significant deviations relative to the top LGA target structures. The following examples highlight problems associated with each of these targets, including inclusions, deletions, and fold deviations with respect to the sequence-related template, variable multidomain associations between differing adjacent domains, choice among numerous sequence-related templates, presence of repeating units and complex multidomain topologies. These problems likely lead to lowered average GDTTS performance scores and require manual inspection of the results.

The restriction endonuclease-like T0868D1 target, which was crystallized in complex with its immunity protein (this is described later on from a different perspective as one of the interesting cases regarding our assessment of free modeling predictions) highlights problems with insertions/deletions and fold deviations with respect to sequence-related templates. When compared to the best HHpred template (4g6uA), T0868D1 presents different insertions, lengths of secondary structures and twists of the core \(\beta\)-sheet (Figure 3A). The variability between the restriction endonuclease-like T0868D1 target and the evolutionary related template (4g6u, LGA score 46.78) is highlighted by the fact that the top LGA template (2cw6D, LGA score 51.4) belongs to a completely different and unrelated TIM-barrel fold. Similar extensive insertion/deletions or fold deviations are observed in the soluble domains from T0875D1 and T0876D1 (Figure 3B, C), where either the C-terminal helix is flipped (T0875D1) with respect to the top HHpred template or the top HHpred template adopts a swapped dimer with respect to the monomeric target (T0876D1) leading to difficulty in template-based modeling.

Another interesting example is the T0896D1 SH3-like domain (Figure 3D), which is adjacent to a L,D-transpeptidase catalytic domain-like domains in the target and to a cysteine proteinase domain in the best
HHpred template. This alternate multidomain arrangement requires different interactions that are reflected in alternate SH3-like loop structures. Additionally, the sequence-related SH3-like templates for T0868D1 display a wide range of LGA scores, with the top scoring HHpred templates exhibiting the worst LGA scores, and vice versa. Both the altered multidomain interactions and the choice of sequence-related templates for T0896D1 potentially lead to lower automated prediction performance. A similar alternate multidomain arrangement arises for LD-transpeptidase catalytic domain-like T0896D2 with its best HHpred template (4xzzA), which forms extensive interactions with an adjacent cystatin-like fold. The prediction for T0896D2 is further complicated by multiple different insertions in the target structure with respect to the template.

Last, cases like T0909D1 and T0912D2 (Figure 3E, F) are difficult because of the presence of repeats, in these cases pectin lyase-like β duplicates, in the full targets. The T0909D1 pectin lyase-like EU is further affected by a different twist of the β-sheet scaffold and different insertions relative to the top HHpred template (2iuA). Alternately, the immunoglobulin-related β sandwich T0912D2 is inserted into a pectin lyase-like domain and is further affected by complex packing of this domain against two other domains, one of which is actually an FM target (T0912D3).

The other group of FM/TBM EUs, T0892D1, T0907D1, T0907D2, T0894D2, T0890D1, and T0943D1, all have very low HHpred scores (green dots on the left in Figure 2A) but LGA scores above 70 (Figure 2B) indicating that despite their low sequence match to the PDB, similar structures exist that could be used to model them if detected. Thus, these represent potentially informative, good templates that are probably hard to find, at least through simple sequence-level searches. We note however that all these EUs have server GDT-TS scores better than 50.

The T0892D1 domain structure adopts an insertion subdomain in DsbA-like fold that is not detected by sequence. However, the adjacent T0892D2 domain adopts a thioredoxin-like fold that detects a related template (1x6m), albeit with a mild HHpred score of 59. The detected thioredoxin-like template also includes an insertion subdomain in DsbA-like fold, suggesting that the target template alignment could be extended to include both domains. The immunoglobulin-like domain repeats in T0907D1 and T0907D2 exhibit a complex multimeric assembly with swapped secondary structure elements that do not exist in the current PDB. Similar to the T0892D1 multidomain target, T0907D1 weakly detects correct immunoglobulin-like templates by sequence (2k7p, probability 22.8), but T0892D1 did not, and the template includes a immunoglobulin-like fold duplication that could potentially be extended to the other domain. T0894D2 adopts a HEEE subunit half of an EndoU-like fold (T0894D1 adopts the other half). Although this target is homologous to EndoU (2c1wB), the sequence has diverged significantly, preventing detection by HHpred. However, a weakly detected template (4ntQA, probability 33.4) includes an analogous HEEE subunit that could be used as a template for modeling, indeed detecting by one of the top performing servers. We note here that the relative simplicity of the HEEE fold might also help contribute to the server performance. Similarly, the relatively simple T0890D1 HHH spectrin repeat-like fold, which weakly detected an analogous sequence template (2ilkA, probability 20.2) covering two of the helices, and the simple T0943D1 R3H domain HEHEE fold might allow elevated server performance.

### 3.5 | Final set of 96 evaluation units for the tertiary structure prediction track of CASP12

The overall procedure of EU definition and classification led to 96 EUs, of which 38 are TBM, 39 are FM, and 19 are FM/TBM. Key global structural features of these domains are shown in Figure 4, qualitatively revealing that domains defined as FM or FM/TBM have shape distributions similar to those of TBM domains, are overall smaller (possibly because they arise from smaller portions of the targets that cannot be easily matched to PDB entries) and include larger fractions of residues in disordered conformations (reasonable considering they are difficult to model).

Figure 5 shows pictures of selected FM targets, to exemplify that the most complex topologies in this CASP round include α-helical bundles and superhelices, β-barrels and β-sandwiches with multipledecorations, Rossman-like folds, and cases of long extended regions with no definite clear structure but orderly wrapped around more globular domains. We note several cases of extremely difficult targets that classify as new folds with no reasonable predictions, for which the best models have very low GDT-TS. Among them, extreme cases shown in Figure 5 are an integral membrane STRA6 receptor for a retinol-binding protein (split into T0863D1 and T0863D2, PDB ID 5sy114), a protein with a cavity made up of a long twisted β-sheet and four α-helices with two additional long helices sticking out whose function remained undisclosed (T0923D1), a protein consisting in a long β-sheet packed with shorter β-sheets and α-helices and a large fraction of loops, whose function also remained confidential (T0941D1). Additional new folds that displayed slightly higher average server performance levels include the α complex topology T0914D1, the α + β complex topology T0914D2, and the α bundle T0898D1. Finally, a domain with limited secondary structure but compactly wrapped around other domains (T0896D3) likely represents an extension of the T0896D2 fold.

### 3.6 | Some interesting cases regarding FM and FM/TBM predictions

This section describes examples of FM and FM/TBM EUs whose predictions are discussed in our assessment article illustrating successful cases of special interest. The structures of these EUs are shown and compared to the best available templates in Figure 6.

T0866 comes from a 183 residue long mammalian cell entry (MluD) protein that forms hexamers, crystallized as a construct that lacks an N-terminal helix and a small C-terminal periplasmic domain. The solved structure presents one clear EU that is 104 residues long. Sequence-level search on the PDB does not lead to obvious confident templates (maximum HHpred score is 8.2), while a structure-level search identifies an analogous pseudorabies virus protease as the best
template (PDB 4cx8) of modest LGA score = 44.7. This top template adopts a β-barrel architecture with similar topology as the MlaD target, with the modest structure similarity score arising from helical insertions in the template loops as well as alternate orientations of the C-terminal strands. Considering these metrics and that the average server GDTTS is at 39.1, T0866-D1 is classified as an FM target; however, one server and several manual predictions reach high GDTTS scores. These scores might reflect a sequence-level template (cell shape determining protein MreC, PDB ID 2qf4) of low confidence that adopts a six-stranded rift-related barrel of the cradle-loop metafold. The MlaD target could also

**FIGURE 4**  Structural properties of TBM (top) and FM + FM/TBM (bottom) EUs. Shape factors are the ratios of the first-to-second and second-to-third largest inertia components (main panel and inset, respectively)

**FIGURE 5**  Examples of folds among CASP12’s FM targets, with information about average server GDTTS (“ave GDTTS”), maximum GDTTS among all predictions (“max GDTTS”), and HHpred and LGA scores of top sequence- and structure-level matches to PDB entries
be considered as a cradle-loop barrel with a flipped and circularly permuted C-terminal hairpin with respect to the MreC template. The sequence alignment between the MlaD target and the MreC template correctly covers the remaining four cradle-loop barrel strands (colored in rainbow from blue to yellow), suggesting a possible homology relationship between the two. Numerous substantiated homologs with different barrel topologies belong to the cradle-loop metafold, including other examples of homologs with flipped hairpins and circular permutations.\textsuperscript{16} Given the nature of this metafold, the MreC template probably represents a homology relationship to the T0866 MlaD target, despite its classification as FM. Accordingly, several of the lower-scoring structure-level templates adopt a cradle-loop barrel (that is, PDB 18dB, rank 3, LGA score 43.5).

Target T0886, representing flagellar hook-associated protein FliD (PDB ID 5fhy\textsuperscript{17}) comes from a 346 residue long protein for which a construct consisting of residues 21–249 was crystallized. This target was split in two domains, with the first one spanning two discontinuous sequence segments. Both domains were classified as FM, as the average server GDTTS is below 40 for both domains and the HHpred score is close to 0, despite the presence of potentially useful templates in the PDB with LGA scores of 62.4 and 48.5 for domains 1 and 2, respectively. The closest identified structures to both domains also correspond to flagellar proteins, including a continuous \(\beta\)-sandwich domain in FlgK (2d4y) with no insertions that is related to the discontinuous FliD target \(\beta\)-sandwich domain (T0886-D1) and a two-layer \(\alpha + \beta\) fold insertion domain in FliC (4nx9) related to the insertion domain in the FliD target (T0886-D2). Being components of the same flagellar nanomachine, these domains probably represent homologs resulting from duplications, domain rearrangements and fast evolution. The relatively low average server GDTTS for both domains (29.5 and 23.9, respectively) reflects the lack of sequence signal and distant relationship of the homologous domains, resulting in their classification as FM.

T0868 is part of a heterodimer together with T0869, (PDB ID 5j4a\textsuperscript{18}) The heterodimer pair represents a bacterial CdiA trRNase toxin domain (T0868) in complex with its immunity protein CdiI (T0869). The T0868 trRNase toxin comes from a 161 residue long protein, of which the segment 46–161 is present in the target structure. T0868 adopts a three-layer \(\alpha/\beta\) sandwich fold with a restriction endonuclease-like topology. The T0869 CdiI comes from a 120 residue long protein, with segment 3–106 present in the structure. Both targets were treated as single domains for evaluation, but predictors were asked to submit models in the same reference frame so that the complex be assessed in the quaternary structure prediction track.

T0868 was classified as FM/TBM as described in the section devoted to FM/TBM targets above. Briefly recapping from that section, the main difficulty of this target probably arises from the top-scoring sequence relationship to a ribonuclease structure of alternate topology with the second hit to another CdiA toxin template with the correct restriction endonuclease topology (4g6vA). Regarding the T0869 immunity protein, it was classified as FM due to its very poor sequence match to the PDB despite a reasonable structural match (HHpred score = 12.2 and LGA score = 54.2) and a low average server GDTTS of 31.4. The top structure-level template (2I19) is unrelated to the target, while the next best template represents a structure analog with similar overall topology of the helical portion of the target structure (3ajB).

T0945 (PDB ID 5lev) is an integral membrane protein covered almost entirely in its X-ray structure. Although the Grishin plots and visual inspection might suggest splitting into two domains to separate an insertion in the transmembrane fold, the overall predictions and the good sequence match to PDB entries (top HHpred score is 75.6) led us
to leave it as a single EU of FM/TBM class (the closest structural match has an LGA score of 44.7 and the average server prediction is at 50.9). The top predictions recapitulate the transmembrane region of the protein very well, but no group captured the insertion relative to the best template. This insertion consists in three short β-sheets followed by an amphipathic helix that according to the experimental structure very likely sits horizontally on the membrane.

3.7 | Outlook for definition and classification of evaluation units in future CASP rounds—including alignment depth as a difficulty metric?

The metrics and plots for classification of EUs into difficulty categories help identify non-ambiguous classification cases allowing assessors to devote more time to the analysis of the borderline FM/TBM cases. However, this being the first time that semi-automated metrics are used for EU classification, we anticipate there might be room for improvement. One specific case that suggests improvements in semi-automated classification is possible is EU T0866D1, a clear FM EU according to our HHpred, LGA and server GDTTS metrics, for which several predictions reached high GDTTS scores around 75–80. As we suggest in the assessment article, predictions for this target might have benefitted from good alignment-based contact predictions; the same might have occurred for other FM EUs like T0886D1 and D2 as also described therein.

Upon assessment of the tertiary structure predictions and considering the content of abstracts for the CASP12 meeting, we concluded that contact predictions from alignments might be driving actual improvement in prediction performance, at least for FM targets. Such methods are actually starting to have practical use to explore the topologies of proteins for which no clear templates are available for homology modeling.\textsuperscript{19–26}

Strictly speaking, our observation was that for FM targets, alignment depth (Neff, defined as the number of sequences in an alignment divided by the number of residues in the sequence) correlates positively with the GDTTS of top models, which might be for reasons other than improved contact predictions. For example, deeper alignments might facilitate the search of remote PDB templates. One way or the other, the observed correlation and examples like T0866D1 pose the question whether domain classification by difficulty should include alignment depths on top of the metrics for sequence and structure matches to the PDB.

A plot of average server GDTTS versus $\log_{10}(1 + \text{Neff})$ shows no strong trend other than a weak apparent lower limit for server prediction scores increasing slowly with alignment depth (Figure 7). This weak trend is actually clear almost exclusively for the FM models; and regarding FM/TBM targets, the plot does not suggest that alignment depth can be used to improve their classification into clear FM or TBM. This may suggest that although deeper alignments might help to make an FM target “easier”, the extent to which it is made easier is far less than the contributions from sequence and structural similarities to PDB entries.

Overall it therefore seems that alignment depth is currently not a good additional metric for difficulty among all targets, although it might help to more finely assign difficulty among FM targets. One way or the other, alignment depth should be kept in mind for EU classification in future CASPs when alignment-based methods might improve and sequence databases grow even further. Moreover, if alignment depths are eventually included for EU classification, they shall be considered during EU definition as well.

4 | CONCLUSION

We have presented here the CASP12 EUs and their classification into TBM, FM and FM/TBM categories, and described salient structural and evolutionary features of selected FM and FM/TBM target units. We further make the point that, although case-by-case expert analysis is still needed for EU definition and classification, the Prediction Center now offers an array of tools that facilitate both tasks, freeing time that human experts can dedicate to the most complicated cases.

ACKNOWLEDGMENTS

We would like to dedicate this work to Anna Tramontano, we thank her for her unstoppable dedication to the CASP initiative and for her continuous support and motivation during our assessment, and to Guido Capitani with whom we shared this experience as CASP assessors. We thank the CASP12 organizers for the invitation to participate in target domain definition and classification, and as assessors for the topology and data-assisted modeling evaluations; and to the groups that contributed structures and experimental data for CASP12.
CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest with the contents of this article.

ORCID

Luciano A. Abriata http://orcid.org/0000-0003-3087-8677
Andriy Kryshtafovych http://orcid.org/0000-0001-5066-7178
Giorgio E. Tamò http://orcid.org/0000-0003-1634-9357

REFERENCES

[1] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (casp): progress and new directions in round XI. Proteins. 2016; 84: 4–14.

[2] Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP11 statistics and the prediction center evaluation system. Proteins. 2016;84(Suppl 1): 15–19.

[3] Cheng H, Schaeffer RD, Liao Y, et al. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014;10(12): e1003926

[4] Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005;33(W244–W248).

[5] Zemla A, LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003;31(13):3370–3374.

[6] Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins. 1999;(Suppl 3):22–29.

[7] Kinch LN, Wrabl JO, Krishna SS, et al. CASP5 assessment of fold recognition target predictions. Proteins. 2003;53(Suppl 6):395–409.

[8] Kinch LN, Shi S, Cheng H, et al. CASP9 target classification. Proteins. 2011;79(Suppl 10):21–36.

[9] Kinch LN, Li W, Schaeffer RD, et al. CASP 11 target classification. Proteins. 2016;84(Suppl 1):20–33.

[10] Taylor TJ, Tai C-H, Huang YJ, et al. Definition and classification of evaluation units for CASP10. Proteins. 2014;82(Suppl 2):14–25.

[11] Guo J, Xu D, Kim D, Xu Y. Improving the performance of Domain-Parser for structural domain partition using neural network. Nucleic Acids Res. 2003;31(3):944–952.

[12] Zhou H, Xue B, Zhou Y. DDOMAIN: dividing structures into domains using a normalized domain-domain interaction profile. Protein Sci Publ Protein Soc. 2007;16(5):947–955.

[13] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402.

[14] Chen Y, Clarke OB, Kim J, et al. Structure of the STRA6 receptor for retinol uptake. Science. 2016;353(6302):aad8266.

[15] Scorsato V, Lima TB, Righetto GL, et al. Crystal structure of the human Tip41 orthologue, TIPRL, reveals a novel fold and a binding site for the PP2Ac C-terminus. Sci Rep. 2016;6(1):30813.

[16] Alva V, Koretkė KK, Coles M, Lupas AN. Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. Curr Opin Struct Biol. 2008;18(3):358–365.

[17] Postel S, Deredge D, Bonson DA, et al. Bacterial flagellar capping proteins adopt diverse oligomeric states. eLife. 2016;5. pii: e18857.

[18] Johnson PM, Gucinski GC, Garza-Sánchez F, et al. Functional diversity of cytotoxic rRNAse/immunity protein complexes from Burkholderia pseudomallei. J Biol Chem. 2016;291(37):19387–19400.

[19] Wickles S, Singharyo A, Andreani J, et al. A structural model of the active ribosome-bound membrane protein insertase YidC. eLife. 2014;3:e03035

[20] Tian P, Boomsma W, Wang Y, Otzen DE, Jensen MH, Lindorff-Larsen K. Structure of a functional amyloid protein subunit computed using sequence variation. J Am Chem Soc. 2015;137(1):22–25.

[21] Kelly GS, Mukherjee R, Kilacsková E, et al. GtRA protein Rv3789 is required for arabinosylation of arabinogalactan in Mycobacterium tuberculosis. J Bacteriol. 2015;197(23):3686–3697.

[22] Abriata LA. Homology- and coevolution-consistent structural models of bacterial copper-tolerance protein CopM support a “metal sponge” function and suggest regions for metal-dependent protein-protein interactions. bioRxiv. 2016. https://doi.org/10.1101/013581.

[23] Antala S, Ovchinnikov S, Kamisetty H, Baker D, Dempski RE. Computation and functional studies provide a model for the structure of the zinc transporter hZIP4. J Biol Chem. 2015;290(29):17796–17805.

[24] Abriata LA. Structural models and considerations on the COA6, COX18 and COX20 factors that assist assembly of human cytochrome C oxidase subunit II. bioRxiv. 2017. https://doi.org/10.1101/013349.

[25] Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife. 2014;3:e02030

[26] Ovchinnikov S, Park H, Varghese N, et al. Protein structure determination using metagenome sequence data. Science. 2017;355(6322): 294–296.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Abriata LA, Kinch LN, Tamò GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Definition and classification of evaluation units for tertiary structure prediction in CASP12 facilitated through semi-automated metrics. Proteins. 2018;86:16–26. https://doi.org/10.1002/prot.25403