Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding

Marco Baiesi¹,², Enzo Orlandini¹,², Flavio Seno¹,² & Antonio Trovato¹,²

Proteins must fold quickly to acquire their biologically functional three-dimensional native structures. Hence, these are mainly stabilized by local contacts, while intricate topologies such as knots are rare. Here, we reveal the existence of specific patterns adopted by protein sequences and structures to deal with backbone self-entanglement. A large scale analysis of the Protein Data Bank shows that loops significantly intertwined with another chain portion are typically closed by weakly bound amino acids. Why is this energetic frustration maintained? A possible picture is that entangled loops are formed only toward the end of the folding process to avoid kinetic traps. Consistently, these loops are more frequently found to be wrapped around a portion of the chain on their N-terminal side, the one translated earlier at the ribosome. Finally, these motifs are less abundant in natural native states than in simulated protein-like structures, yet they appear in 32% of proteins, which in some cases display an amazingly complex intertwining.

Theoretical and experimental efforts of the last two decades have established that kinetic and thermodynamic properties of the protein folding process can be inferred by some spatial features of the native structure itself¹–³. For instance, the contact map of the native state⁴–⁶, the matrix indicating which pairs of residues are close in space, determines the folding nucleus, i.e. the group of residues whose interaction network is essential for driving the folding. Similarly, the loops formed between residues in contact have an average chemical length, the contact order, which is strongly correlated with the folding time⁷–⁹. However, some proteins, being extremely self-entangled in space, are characterized by a folding process that cannot be simply rationalized by local (contact) properties. Examples are proteins hosting knots¹⁰–¹⁷, slipknots¹⁸,¹⁹, lassos²⁰,²¹ and links²²,²³. These complex motifs were found in about 6% of the structures deposited in the protein data bank (PDB) and, although it is expected that their presence can severely restrict the available folding pathways¹⁷,²², it is not clear how proteins avoid the ensuing kinetic traps and fold into the topologically correct state.

A crucial question is whether and how these topologically entangled motifs affect the protein energy landscape. According to the well established paradigm of minimal frustration²⁴,²⁵, energetic interactions in proteins are optimized in order to avoid as much as possible the presence of unfavorable interactions in the native state. Although non optimized interactions may result in kinetic traps along the folding pathway, some amount of residual frustration has been detected and related to functionality and allosteric transitions²⁶.

A further issue is whether the effect of topology-induced traps depends on the folding direction along the chain: if proteins fold cotranslationally when they are being produced at the ribosome, one then expects sequential folding pathways proceeding from the N-terminus to be less hindered by such traps.

To understand the relevance of topological motifs in proteins, here we quantify the amount of self-entanglement in native structures by computing the Gaussian entanglement (GE)⁹,²³, a generalization of the Gauss integrals used to compute the linking number²⁷ (see Materials and Methods for details). Indeed, if integrals were computed for two closed curves, e.g. loops in proteins closed by disulphide bridges or by any other form of covalent bond²⁸–²² (Fig. 1b), the result would be the (integer) linking number²⁷. By applying the method to open
Figure 1. Sketches of proteins with (a) a knot, (b) two linked loops with cysteine closures (magenta dots), (c) two linked loops with virtual non-covalent closures (yellow and green dots form two different contacts), and (d) a loop (red) intertwined with an open chain portion (blue) - a “thread”. (e) A configuration with the loop (γ_i) closer to the C-terminus than the thread (γ_j), and (f) one with the loop closer to the N-terminus. In the two latter pictorial representations of non-structured proteins, we also show the loop-thread sequence separation s and the loop length m.

Results and Analysis

Proteins with entangled loops are not rare. We use the contact GE parameter G_c to find protein domains with at least one loop γ, intertwined with a “thread” γ', which is another portion of the protein (Fig. 1d–f). More precisely, we can associate G_c(i, j) to a given loop-thread pair by using the Gauss double integral described in Materials and Methods. By maximizing |G_c(i, j)| over all the possible threads γ', we assign an entanglement score G_c(i) to the loop and, by further maximizing |G_c(i)| over all loops γ, we find the entanglement G_c of the protein. At variance with similar quantities defined for closed curves, G_c is a real number. Yet, we define a loop γ in a configuration as in Fig. 1d to be entangled if |G_c(i)| ≥ 1. Such a threshold is natural because a linking number |L| = 1 is the minimum value that guarantees that two closed curves are linked. In a data set of 16968 protein domains, 5375, the 31.7%, host at least one entangled loop. We also monitor the value L' of the linking entanglement (LE) for a single protein, defined as G_c, for two subchains that are both loops, as in Fig. 1c. In Fig. 2 we show five examples of “entangled” protein domains, along with their respective values of G_c and L'.

The non-trivial entanglement features of protein structures, when analyzed with GE and LE, are apparent in Fig. 3A, where protein domains are represented in the sequence space. All the points lie in the region |L'| ≤ |G_c| because the latter quantity is defined as an extremum over a wider subset.

A typical example with G_c ≈ L' ≈ 1 is shown in Fig. 2a. As expected, however, there are cases of proteins with at least one entangled loop (|G_c| ≥ 1) and all pairs of loops with negligible |L'|. These proteins correspond to the...
**Figure 2.** Some examples of protein domains with nontrivial entanglement, in which one notes a looped portion (red, with yellow ends, following the color code of Fig. 1d) intertwined with another portion of the protein, a thread (blue). (a) protein 2bjua02, with $L' = 1.145$ (also the blue chain contains a loop) and $G'_c = 1.285$ of similar magnitude; (b) protein 3hitA01, with significant entanglement ($G'_c = -1.56$) but without two linked loops ($L' = -0.34$); (c) protein 3tnxA00, with large $G'_c = -3.07$, while $L' = -1.31$ is much smaller and with the same sign; (d) protein 2i06A01, with $L' = 0.74 \approx 1$, partitioned in the two corresponding linked loops and thus following the color code of Fig. 1c (green ends of the blue loop); (e) again 2i06A01, with $G'_c = -1.26$, highlighting the related loop-thread partition. In the last two points one notes that the sign of $L'$ is opposite to the sign of $G'_c$. It is an example of the coexistence of different forms of entanglement in the same protein domain. (f) protein 1otjA00, one of the protein domains with largest (absolute) Gaussian Entanglement, with $G'_c = -3.24$ and $L' = -3.02$. The red loop, with yellow ends, is extremely entangled with the blue portion (which in this case also contains a loop).

**Figure 3.** (A) Plot of $L'$ vs $G'_c$ for each protein in the CATH database; the five proteins shown in Fig. 2a–f are highlighted with the corresponding letter. (B) Smoothed histogram of data with significant linking ($|L'| > 1/2$). The highest probability is around $G'_c \approx L' \approx 1$. The data with the values of $L'$ and $G'_c$ computed for each protein in the CATH database are available at [http://researchdata.cab.unipd.it/id/eprint/123](http://researchdata.cab.unipd.it/id/eprint/123).
It is interesting to observe that in several other cases the GE has a different sign with respect to the LE. This may take place if the chain winds around itself with opposite chiralities in different portions of the same protein. An example is shown in Fig. 2d,e. One of the most entangled structures found in the database, with $G'_i \approx L'_i \approx 3$, is shown in Fig. 2f.

Figure 3A shows that the GE is distributed over a broad spectrum of values and that the threshold $|G'_i| \geq 1$ for entangled loops is conservative enough. Clusters emerge in the density plot of Fig. 3B, where the majority low LE points are removed by excluding data with $|L'_i| < 1/2$ (see also Fig. S1, which is an enlargement of Fig. 3A). The clusters are found around $L'_i \approx 1 \pm 1$ vs $G'_i \approx 1$. In particular the most populated region has $G'_i \approx L'_i \approx 1$.

For further analysis, we consider only the GE indicator, which captures more varieties of entangled motifs than the LE (e.g. winding without linking, see Fig. 2b).

Entangled loops are found more frequently on the C-terminal side of the corresponding intertwining segment. In the definition of $G'_i(i,j)$, the role of the loop $\gamma_i$ cannot be exchanged with that of the other chain portion $\gamma_j$. This feature of the GE may be exploited to detect possible asymmetries in the respective location along the backbone of $\gamma_i$ and $\gamma_j$. The score $G'_i(i)$ associates to a given loop $\gamma_i$ the open arm with which it is mostly entangled. We consider separately the following two cases: when the threading arm is between the N-terminus and the loop (N-terminal thread, see Fig. 1e) or when it is between the loop and the C-terminus (C-terminal thread, Fig. 1f).

We now focus on the properties of entangled loops (3.75% of the total) and we count how many of them are classified as N-terminal threads or C-terminal threads. In principle, there is no reason to expect one of the two classes to be more populated than the other.

In order to discuss carefully the statistical significance of possible asymmetries, we need to take into account that some degree of correlation occurs in the counting of entangled loops. In fact, different loops can belong to essentially the same topological configuration, for example when a protein arm intertwines both with the loop formed between amino acids $i_1$ and $i_2$ and with the one formed between $i_1 + 1$ and $i_2$. Thus, we employ a clustering procedure based on a pairwise distance defined between loops (see Materials and Methods for details).

By using the effective countings given by the clustering procedure, we find that the fraction of N-terminal (C-terminal) threads within entangled loops is 0.55 (0.45). The highlighted bias in favor of the N-terminal threads (Fig. 1e) against the C-terminal ones (Fig. 1f) is statistically significant at the level of 14 standard deviations. A somewhat similar result was found by studying topological barriers in protein folding.

One may ask whether N-terminal threads are favored, simply, because entangled loops are by themselves closer to the C-terminus, without the need of considering the $|G'_i|$ maximization that selects the intertwining thread. In order to check this, we consider a random reference case, whereby one putative threads is sampled randomly for each entangled loop, with uniform probability across all possible segments non overlapping with it. As a matter of fact, putative N-terminal threads are not favored in the reference case. We find instead a small bias in the opposite direction; namely, the fraction of putative N-terminal (C-terminal) threads within entangled loops is 0.487 (0.513). This small but statistically significant (3.5 standard deviations) imbalance, suggests that entangled loops are slightly, if any, located closer to the N-terminus, thereby highlighting even more that the favored placement of the intertwining thread to the entangled loop N-terminal side is a genuine effect.

Entangled loops favor positive chirality only for N-terminal threads. The formation of an entangled structure is not simple, as it requires a non local concerted organization of the amino acids in space, where a crucial role is played by the order of formation of different native structural elements along the folding pathway.

A misplaced nucleation event in the early stages of the folding pathway might prevent the protein from folding correctly. Dealing with spontaneous “in vitro” refolding, there is no reason to expect the folding order of different elements to be related to a preferential specific direction along the chain.

Nevertheless, an asymmetry can be envisaged if a protein folds cotranslationally, according to the following argument. For the C-terminal thread, the loop might be formed in the early folding stages, making it difficult for the rest of the protein to entangle with it and thus to reach the native conformation. Conversely, for the N-terminal thread, the loop could wrap more easily around the open threading arm, already folded in its native conformation, after ejection from the ribosome. If confirmed, this picture would explain the asymmetry we observe between N- and C-terminal threads. The latter could be anyway interpreted as a possible fingerprint of an evolutionary process, intimately related to entanglement regulation driven by cotranslational folding. Such conclusion is corroborated by looking, separately for C- and N-terminal threads, at the normalized distributions of loop-thread sequence separations $s$ plotted in Fig. 4a. The distributions for the random reference case (empty symbols) are very similar for N-terminal (triangles) and C-terminal (circles) threads, showing again that if the putative thread is chosen randomly no asymmetry is present. One would expect a uniform reference distribution for loops at a fixed distance from the relevant chain terminus (N for N-threads). The regular decay observed for increasing $s$ is due instead to the fact that different protein domains have different lengths and different loops are located differently along the backbone. On the other hand, the $|G'_i(i,j)|$ maximization leading to $G'_i(i)$ selects preferentially arms that start just after (or before) the loop, at a distance of one or few amino acids. This is similar to what already observed for pokes, and reflects the fact that a rapid turning of the protein chain is the simplest way for maximizing the mutual winding between two subchains. However, loop-thread pairs that are one unit distance apart are significantly more favored for C-terminal (squares) than for N-terminal (diamonds) threads.
highlights that the values around could in principle adopt. This ensemble is found in a poly-valine “V AL60” database, obtained with an all atom formation by a set of putative alternative compact conformations (i.e. rich in secondary structures) that a protein fraction is small or large we compare it with an analogous quantity computed in an unbiased reference state of the CATH domains there are 3617208 loops, of which 135530 (3.75%) are entangled. To assess whether this fact, we find C-terminal entangled loops to be balanced between positive and negative chiralities within the 0.3%, whereas N-thread entangled loops are highly biased (74%) in favor of positive chiralities. As a matter of fact, we find C-thread entangled loops to be balanced between positive and negative chiralities within the 0.3%, whereas N-thread entangled loops are highly biased (74%) in favor of positive chiralities, $G_c(i) > 0$.

Natural proteins are less entangled than protein-like compact conformations. In the ensemble of the CATH domains there are 3617208 loops, of which 135530 (3.75%) are entangled. To assess whether this fraction is small or large we compare it with an analogous quantity computed in an unbiased reference state formed by a set of putative alternative compact conformations (i.e. rich in secondary structures) that a protein could in principle adopt. This ensemble is found in a poly-valine “VAL60” database, obtained with an all atom simulation that accurately sampled the configurational space of a homopolypeptide formed by 60 valine amino acids (see Materials and Methods for details).

For a proper comparison with VAL60, we restrict our CATH database only to the proteins of comparable length, filtering out 772 proteins with length $n$ from $n = 55$ to $n = 64$ amino acids. In this reduced “CATH60” ensemble of natural proteins there are 47954 loops, of which 138 (0.3%) are entangled. There are 19 proteins (2.46%) hosting at least one entangled loop. These values are of course lower than those for the full CATH ensemble, in which longer proteins can host more entanglement. In VAL60 there are 2284693 loops, of which 57577 (2.46%) hosting at least one entangled loop. These values are of course lower than those for the full CATH ensemble, in which longer proteins can host more entanglement. In VAL60 there are 47954 loops, of which 138 (0.3%) are entangled. There are 19 proteins (2.46%) hosting at least one entangled loop. These values are of course lower than those for the full CATH ensemble, in which longer proteins can host more entanglement. In VAL60 there are 2284693 loops, of which 57577 (2.46%) hosting at least one entangled loop. These values are of course lower than those for the full CATH ensemble, in which longer proteins can host more entanglement.

However, it is known that, presumably for kinetic reasons, VAL60 is characterized by loops on average longer than those of natural proteins. Consequently, to avoid any possible bias in the comparison, we divide loops in classes of homogeneous length $m$. For some classes, the normalized histogram of the GE for CATH60 and VAL60 datasets are plotted in Fig. 5a–c. In all cases it is apparent that the range of $G_c(i)$ is wider for the VAL60 homopolypeptides than for the natural proteins. The deep difference between the two distributions can be appreciated in Fig. 5d, where the root mean squared $G_c(i)$ is plotted as a function of the loop length: the values for VAL60 are always significantly higher than those for natural proteins. Note that the root mean squared $G_c(i)$ increases with $m$ only up to half of the protein length. From there on, the remaining subchain starts getting too short to entangle.

In conclusion, we have a clear statistical evidence that entangled loops occur less frequently in natural proteins with respect to random compact protein-like structures.

Amino acids at the ends of entangled loops are frustrated. In the preceding sections we provided two independent evidences that, although entangled loops are not rare in natural protein structures, their occurrence and position along the backbone chain are kept under control. A possible reason is the need to limit potential kinetic traps in the folding process brought about by entangled loops, for example by deferring their formation to the latter stages of the folding pathway. Thus, we expect to detect a related fingerprint in the specific amino acids found in contact with each other at the end of entangled loops (“entangled contacts”). We check
whether such amino acids share the same statistical properties of the amino acids forming any possible contact ("normal contacts").

The frequency with which two amino acids are in contact is typically employed to estimate knowledge based potentials. In a nutshell, if two amino acids \(a\) and \(b\) occur to be in contact more frequently than on average, they are expected to manifest a mutual attraction and are therefore characterized by a negative effective interaction energy \(E_{\text{norm}}(a, b)\) (see Materials and Methods).

If effective interaction energies are computed by restricting the analysis only to the entangled contacts, a new set of entangled contact potentials \(E_{\text{GE}}(a, b)\) can be derived. The discrepancies between such potentials and the normal ones can be conveniently captured by an enrichment score \(\Delta E_{\text{enr}}(a, b)\). A negative enrichment score \(\Delta E_{\text{enr}}(a, b) < 0\) implies that \((a, b)\) are more frequently in contact when they are at the ends of entangled loops, and vice-versa for positive scores. Figure 6 shows that \(\Delta E_{\text{enr}}(a, b)\) anticorrelates with \(E_{\text{norm}}(a, b)\). This correlation is statistically significant. The Pearson correlation coefficient is \(r = -0.31\), with a \(P\)-value of \(2 \times 10^{-6}\). The Spearman rank correlation is \(\rho = -0.23\) with a \(P\)-value of \(6 \times 10^{-4}\).

The anticorrelation of Fig. 6 has an important consequence: pairs of amino acids that in a globular protein interact strongly \((E_{\text{norm}}(a, b) < 0\), mainly hydrophobic amino acids) are present less often \((\Delta E_{\text{enr}}(a, b) > 0)\) in entangled contacts, while amino acids that typically interact weakly \((E_{\text{norm}}(a, b) > 0\), mainly polar and hydrophilic amino acids) are more abundant \((\Delta E_{\text{enr}}(a, b) < 0)\) at the ends of entangled loops. We checked that this result is not trivially due to entangled contacts being preferentially located on the protein surface, finding that residues involved in entangled contacts are even slightly more buried in the protein interior than those involved in normal contacts (see Fig. S2). The deep difference between the two set of scores \(E_{\text{norm}}(a, b)\) and \(E_{\text{GE}}(a, b)\) emerges clearly from the graphical representations in Fig. 7 of \(E_{\text{norm}}(a, b)\) and \(\Delta E_{\text{enr}}(a, b)\), in which positive and negative values are marked red and blue, respectively, whereas white boxes mark scores that are not significant within the related statistical uncertainty.

The blue spots in Fig. 7a represent interactions between amino acids that interact frequently with each other (mainly hydrophobic pairs), whereas the red area is populated by amino acids that are rarely in contact (mainly polar pairs).

In Fig. 7b, the blue spots highlight amino acids that have decreased their energy score and which are therefore more present at the ends of the entangled loops than in normal contacts. These include mainly polar amino acids. Note that proline is particularly enriched at the end of entangled loops. The red spots in Fig. 7b indicate amino acids which are less present at the ends of the entangled loops than in normal contacts. These include mainly hydrophobic ones. The case of cysteine self-interaction is pedagogical: the strongest attractive interaction between
amino acids turns out to be the more diminished one at the end of entangled loops (see also Fig. 6), consistently with the very low number of linked loops closed by disulphide bonds (Fig. 1c) that was found in the PDB39. Interestingly, the four aromatic amino acids (HIS, PHE, TRP, TYR) violate the general trend. Interactions between aromatic pairs are found in the bottom-left quadrant of Fig. 6. Despite being very frequent in normal contacts (all their mutual entries are dark blue in Fig. 7a), they become even more abundant when at the ends of entangled loops (still blue in Fig. 7b), highlighting a special role likely played by aromatic rings in such complex structures.

Figures 6 and 7b provide clear evidence for the existence of an evolutionary pressure shaping the amino acid sequences. This natural bias weakens energetically the contacts which close entangled loops, consistently with the argument that a too early stable formation of the loop could prevent the correct folding of the full protein.

These results are very robust to changes in the $G_c'$ threshold used to define entangled contacts, in the minimum length $m_0$ of the considered subchains, and to the introduction of a minimum loop-thread separation $s_0$, see Figs S3–S5.

Discussion and Conclusions
With the notion of Gaussian entanglement we extend the measure of mutual entanglement between two loops to any pair of open subchains of a protein structure. This allows us to perform an unprecedented large scale investigation of the self entanglement properties of protein native structures, through which we identify and locate a large variety of entangled motifs (Fig. 2), by focusing on the notion of “entangled loop”, a loop intertwining with another subchain (Fig. 1d). Different entangled motifs can coexist in the same protein domain, even with opposite chiralities, and few domains exhibit a pair of loops intertwining even thrice around each other (see the examples in Fig. 2d,f, and points in Fig. 3). Gaussian entanglement could be used to improve the classification of existing protein folds40, as previously done with Gauss integrals computed over the whole protein chain41.

Our analysis shows unequivocally (Fig. 5) that, although entangled motifs are present in a remarkably high fraction, 32%, of protein domains, these host a lower amount of entangled loops than protein-like decoys produced with molecular dynamic simulations35. The question is then why natural folds avoid overly entangled conformations with otherwise plausible secondary structure elements. Are entangled loops obstacles for the folding process? If yes, how does Nature cope with them when they are present?

To answer these questions, we recall that an efficient folding of proteins is fundamental for sustaining the biological machinery of cell functioning. The rate and the energetics of the protein folding process, which are defined by its energy landscape, are encoded in the amino acid sequence. Over the course of evolution, this landscape was shaped to allow and stabilize protein folding, avoiding possible slowdowns.

We find indeed two clear hallmarks suggesting that the entangled loops in proteins are kept under control: (i) a statistically significant asymmetry in their positioning with respect to the other intertwining chain portion, which is consistent with cotranslational folding promoting the presence of entangled loops on the C-terminal side of the intertwining thread (see Fig. 4a); (ii) weak non optimized interactions between the amino acids in contact at the end of entangled loops, an example of energetic frustration (see Figs 6 and 7). Both these findings suggest that the late formation of entangled loops along the folding pathway could be a plausible control mechanism to avoid kinetic traps. On the other hand, the additional stability that one expects to be provided by entangled and knotted structures can compensate for the presence of such weak interactions.

Figure 6. Scatter plot of the enrichment score $\Delta E_{enr}(a, b)$ vs normal contact potential $E_{norm}(a, b)$. Each point is for an amino acid pair $(a, b)$ and is colored according to amino acid types: black for pairs of aromatic residues (HIS, PHE, TRP, TYR); magenta for CYS-CYS; green for the rest. Error bars are computed with a bootstrapping procedure and we plot only errors for $\Delta E_{enr}$ as those for $E_{norm}$ are smaller than the symbol size.
Interestingly, interactions between aromatic amino acid pairs are promoted at the end of entangled loops (see Fig. 7b), suggesting that their presence could be related to the protein biological function. Whether entangled loops may have specific biological functions is an intriguing open question, as in the case of knots in protein domains17,42.

Finally, we detect a remarkable bias, favoring positive chiralities, that is present only for entangled loops on the C-terminal side of the intertwining thread (see Fig. 4b). This suggests that the observed chirality bias arises in the context of cotranslational folding. A simple possibility is that loop winding of the C-terminal part of the chain may have a preferred orientation when just released from the ribosome. Further work will be needed to test this speculation and to fully rationalize the chirality bias. As a matter of fact, the ribosome can discriminate the chirality of amino acids during protein synthesis43.

Stemming from works on glassy transitions44,45, the concept of minimal frustration between the conflicting forces driving the folding process is a well established paradigm24–26 in protein physics. It has been further argued26 that frustration is an essential feature for the folding dynamics and that it can give surprising insights into how proteins fold or misfold.

Is it possible to reconcile the frustration detected at the ends of entangled loops with the minimal frustration principle? Let us assume that a non optimal ordering of the events along the folding pathways (for example, the formation of a loop which has then to be threaded by another portion of the protein to form an entangled structure) is highly deleterious. In order to prevent this, it could indeed be preferable to select suitable sub-optimal interactions. In fact, this would be a remarkable example of minimal frustration in action, having to compromise between topological and energetic frustration.

Obviously, other data will be needed to confirm this proposed mechanism for the folding process, from both simulations and experiments. In either case, a simple protocol could consist in mutating into cysteine both

![Figure 7](https://www.nature.com/scientificreports/)

**Figure 7.** (a) Normal contact potential $E_{\text{norm}}$; amino acids are ranked from left to right (top to bottom) with increasing average $E_{\text{norm}}$ (over row/column). (b) Enrichment score $\Delta E_{\text{enr}}$ for entangled contacts. Different backgrounds are used for highlighting negative and positive values: blue for $E < E_0$, light blue for $-E_0 \leq E \leq 0$, pink for $0 < E \leq E_0$, and red for $E > E_0$ with $E_0 = 35$. White is used for scores that differ from zero less than the corresponding statistical uncertainty, computed by means of a bootstrapping procedure.
residues at the ends of an entangled loop, provided no other cysteines are present in the sequence, and in assessing whether the folding is then hindered by the formation of a disulfide bridge in oxidizing conditions. In the context of knotted proteins, single molecule force spectroscopy techniques were shown to be particularly useful in controlling the topology of the unfolded state 26. Similarly, both “in vivo” folding experiments 27 and appropriate simulation protocols 28–30 could be employed to test the possible role of cotranslational folding in determining the patterns detected for entangled motifs: double cysteine mutants would then be predicted to be more deleterious for the folding of C-terminal threads with respect to N-terminal threads. In all cases, it is essential to gather statistics over several different proteins before validating or rejecting our hypothesis; the signals that we reveal in this contribution are statistical in nature; therefore we do not expect all entangled loops to form late in the folding process nor all C-terminal threads to be cotranslationally disfavored. For example, it has been recently proposed, in the context of deeply knotted proteins, that loops formed by a synthesized earlier portion of the same protein can be actively thread by nascent chains at the ribosome 31. However, this is not in contradiction with our findings, since knotted proteins are much less frequent than the general entangled motifs discussed here.

Materials and Methods

CATH database. We use the v4.1 release of the CATH database for protein domains, with a non-redundancy filter of 35% homology 25. To avoid introducing entanglement artificially for proteins with big gaps in their experimental native structures, we do not consider any protein in the CATH database that presents a distance $> 10$ Å between subsequent $C_\alpha$ atoms in the coordinate file. We find that this selection keeps $N_{prot} = 16968$ out of the available 21155 proteins. CATH domain names such as 2bjuA02 refer to the 2nd domain from chain A with PDB code 2bju. The CATH database is available at https://download.cathdb.info/cath/releases/all-releases/v4_1_0/.

Poly-valine database. The VAL60 database is an ensemble of 30064 structures obtained by an exhaustive exploration of the conformational space of a 60 amino acid poly-valine chain described with an accurate all-atom interaction potential 35. The exploration was performed with molecular dynamics simulations using the AMBER03 force field 26 and the molecular dynamics package GROMACS 34 and by exploiting a bias exchange metadynamics approach 36 with 6 replicas. The simulation was performed in vacuum at a temperature of 400 K. The conformations have been selected as local minima of the potential energy with a secondary structure content of at least 30% and a small gyration radius. The protein-like character of VAL60 conformations was successfully tested by using different criteria commonly employed to assess the quality of protein structures 35. The stability of a small subset of VAL60 structures was successfully tested even after mutation of all residues to Alanines. Crucially, it was observed that the VAL60 database contains almost all the natural existing folds of similar length 35. The stability of a small subset of VAL60 structures was successfully tested even after mutation of all residues to Alanines. Crucially, it was observed that the VAL60 database contains almost all the natural existing folds of similar length 35. However, these known folds form a rather small subset of the full ensemble, which can be thought as an accurate representation of the universe of all possible conformations physically attainable by polypeptide chains of length around 60. A repository for the VAL60 database is available at https://doi.org/10.5061/dryad.1922.

Mathematical definition of the linking number and its computational implementation. The linking number between two closed oriented curves $\gamma_i = \{\mathbf{r}^{(i)}\}$ and $\gamma_j = \{\mathbf{r}^{(j)}\}$ in $\mathbb{R}^3$ may be computed with the Gauss double integral

$$ G \equiv \frac{1}{4\pi} \oint_{\gamma_i} \oint_{\gamma_j} \frac{\mathbf{r}^{(i)} - \mathbf{r}^{(j)}}{|\mathbf{r}^{(i)} - \mathbf{r}^{(j)}|} \cdot (d\mathbf{r}^{(i)} \times d\mathbf{r}^{(j)}) $$

(1)

It is an integer number and a topological invariant 27. If computed for open curves, it becomes a real number $G'(\mathbf{r})$ (the GE) that quantifies the mutual entanglement between the curves $\gamma_i$ and $\gamma_j$ 9,23,28–30. In proteins, piece-wise linear curves join the coordinates of subsequent $C_\alpha$ atoms. In particular, $\gamma_i$ is an open subchain joining $C_\alpha$ atoms from index $i_1$ to $i_2$ and similarly $\gamma_j$ is another nonoverlapping subchain from $j_1$ to $j_2$.

We specialize to the configurations studied in ref. 9, in which $i_1$ and $j_1$ amino acids are required to be in contact. In this study, the contact is present if any of the heavy (non hydrogen) atoms of residue $i_1$ is near any of the heavy atoms of residue $i_2$, namely they are at a distance at most $d = 4.5$ Å. The “contact” Gaussian entanglement of these configurations (sketched in Fig. 1d–f) is named $G'(i,j)$. Since proteins are thick polymers and bonds joining $C_\alpha$ atoms are quite far from each other (compared to their length), we may approximate the integral with a discrete sum. Given the coordinates $\mathbf{r}_i$ of $C_\alpha$’s, the average bond positions $R_j \equiv \frac{1}{2}(\mathbf{r}_{j_1} + \mathbf{r}_{j_2})$ and the bond vectors $\Delta \mathbf{R}_j = \mathbf{r}_{j+1} - \mathbf{r}_j$ enter in the estimate of $G'(i,j)$ for $\gamma_i$ and $\gamma_j$,

$$ G'_j(i,j) \equiv \frac{1}{4\pi} \sum_{i_1 = i-1}^{i_2 - 1} \sum_{i = i_1}^{j-1} \left| \frac{R_j - R_i}{|R_j - R_i|} \right| \cdot (\Delta \mathbf{R}_j \times \Delta \mathbf{R}_i). $$

(2)

We then associate a contact entanglement $G'_j(i,j)$ to a “loop” $\gamma_i$ as the extreme (i.e. with largest modulus) $G'_j(i,j)$, for all “threads” $\gamma_j$ with $j_1 - j_2 \geq m_0$ ($m_0 = 10$). Finally, the contact entanglement $G'_j$ of a protein is the extreme of $G'_j(i,j)$ for all loops of length $m = i_2 - i_1 \geq m_0$. The linking entanglement $L$ is equal to $G'_j$ for configurations with two loops as in Fig. 1c. It is not exactly the linking number $L$ because the two closures between contacts are not performed.

Clustering procedure for counting effectively independent loops. Each entangled loop $\gamma$ is characterized by five numbers, its two indices $(i_1, i_2)$, the indices of the threading portion $(j_1, j_2)$, and the corresponding Gaussian entanglement $G'_j(i,j)$. It is thus natural to define a distance between two entangled loops $\gamma^A$ and $\gamma^B$ as
\[ d_{AB} = \sqrt{\left( t_i - t_{ij}^A \right)^2 + \left( t_i - t_{ij}^B \right)^2 + \left( t_i - t_{ij}^A \right)^2 + \left( t_i - t_{ij}^B \right)^2 + w_g G'_i(i^A, j^A) - G'_c(i^B, j^B) \}^2 }. \]  

where \( w_g \) is a weight to be defined. In order to count the effectively independent loops we used the following procedure within each protein in the CATH database: first we selected the loop with the largest number of neighbors, namely with the largest number of loops at a distance smaller than a threshold \( d^* \). We assign the selected loop and all its neighbors to the same cluster, removing them from the running list of loops. We iterate this procedure until the running list is empty, so that each loop \( \gamma \) belongs to a cluster with a given number of members \( N_\gamma \). Each loop is then included in all the statistics and distributions reported in the main text, with an effective counting weight \( 1/N_\gamma \). By using \( d^* = 20 \), \( w_g = 10^4 \), we find an overall effective counting of 18041 independent entangled loops, a 13\% of the original 135530 countings. The results reported in the main text are qualitatively robust against reasonable variations of the \( d^* \), \( w_g \) parameters. The data with the values of \( G'_c \) computed for each loop in each protein in the CATH database, and grouped after clustering, are available at http://researchdata.cab.unipd.it/id/eprint/123.

**Inference of statistical potentials.** In order to estimate effective interactions between amino acids in protein structures, we use an established knowledge based approach \(^{38}\). Pairwise potentials can be obtained by analyzing databases of known protein conformations\(^{56}\). These potentials are derived measuring the probability of an observable, such as the formation of a contact, relative to a reference unbiased state\(^{37}\). The conversion of the probability in an energy is done by employing Boltzmann’s law\(^{57}\).

The first step includes characterizing the reference null space of possible pairs of amino acids. All amino acid pairs within each protein sum up to a grand total of \( N \) generic pairs (i.e. just combinatorial pairings not necessarily related to a spatial contact) in our ensemble of protein structures. In the same way, given two amino acid kinds \( a \) and \( b \), one sums up the occurrence of \( a-b \) pairs within each protein to a grand total of \( N(a, b) \) pairs in the ensemble.

To quantify energies of “normal” contacts \( E_{\text{norm}}(a, b) \) between amino acids of type \( a \) and \( b \), we consider two amino acids to be in contact if any inter-residue pair of their side chain heavy atoms is found at a distance lower than 4.5 Å. By considering only the ensemble of amino acids which are in contact within each protein, their total counting results in \( N \) generic contacts. Similarly, the specific contacts between amino acids of kind \( a \) and \( b \) are summed up to a total \( N_{(a, b)} \).

The statistical potentials for normal contacts are defined by comparing the frequencies\(^{37,58}\)

\[ f(a, b) = N(a, b)/N, \quad f_{\text{c}}(a, b) = N_{(a, b)}/N, \]  

within the ensemble of all pairs or the ensemble of contacts, respectively. If \( f(a, b) \) is relatively high compared to \( f_{\text{c}}(a, b) \), it means that chemistry and natural selection favored the organization of native protein structures toward configurations where \( a \) and \( b \) types were in contact. Thus, the argument is that a lower potential energy should be associated to such contacts; the normal contact potentials are therefore given by

\[ E_{\text{norm}}(a, b) = -\tau \log \frac{f(a, b)}{f_{\text{c}}(a, b)}, \]  

where we introduced a parameter \( \tau = 100 \) for the convenience of rescaling numbers and rounding them off to readable integers.

We can compute a similar kind of potentials \( E_{\text{GE}}(a, b) \) for “entangled” contacts, just restricting the analysis to the subset of contacts between amino acids that are at the end of an “entangled loop”, defined as a loop \( \gamma \), for which \( | G'_c(i, j) | \geq 1 \), that is a loop for which at least one thread \( \gamma \) exists such that the corresponding \( | G'_c(i, j) | \geq 1 \). Within all proteins, in total we count \( N \) such contacts while the specific ones are \( N_{(a, b)} \), and hence

\[ f_{\text{c}}^G(a, b) = N_{(a, b)}/N, \]  

\[ E_{\text{GE}}(a, b) = -\tau \log \frac{f_{\text{c}}(a, b)}{f_{\text{c}}^G(a, b)}, \]  

To easily capture dissimilarities between \( E_{\text{norm}}(a, b) \) and \( E_{\text{GE}}(a, b) \), we introduce an enrichment score defined as

\[ \Delta E_{\text{en}}(a, b) = -\tau \log \frac{f_{\text{c}}^G(a, b)}{f_{\text{c}}(a, b)} = E_{\text{GE}}(a, b) - E_{\text{norm}}(a, b) \]  

In all cases we have imposed a constraint on the pairs of amino acid considered: the two amino acids \( i \) and \( j \) in contact must have indices difference \( i - j \geq m_0 = 10 \). This threshold \( m_0 \) removes any eventual bias in comparing potentials due to the entanglement constraint, which requires entangled loops and threads of a minimal length to be present.

We computed all statistical potentials, together with the related uncertainties, by using a bootstrapping procedure with 101 independent resamplings. The countings and the statistical scores obtained for each amino-acid pair are available at http://researchdata.cab.unipd.it/id/eprint/123.
The raw data required to reproduce these findings are available at http://download.cathdb.info/cath/releases/all-releases/v4_1_0/ and https://doi.org/10.5061/dryad.1922. The processed data required to reproduce these findings are available at http://researchdata.cab.unipd.it/id/eprint/123.

References

1. Baker, D. A surprising simplicity to protein folding. Nature 405, 39 (2000).
2. Dokholyan, N. V., Li, L., Ding, F. & Shakhnovich, E. I. Topological determinants of protein folding. Proceedings of the National Academy of Sciences 99, 8637–8641 (2002).
3. Dill, K. & MacCallum, J. The protein-folding problem. 50 years on. Science 338, 1042–1046 (2012).
4. Micheletti, C., Banavar, J. R., Maritan, A. & Seno, F. Protein structures and optimal folding from a geometrical variational principle. Physical Review Letters 82, 3372 (1999).
5. Muñoz, V. & Eaton, W. A. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proceedings of the National Academy of Sciences 96, 11311–11316 (1999).
6. Alm, E. & Baker, D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proceedings of the National Academy of Sciences 96, 11305–11310 (1999).
7. Plaxco, K. W., Simons, K. T., Ruczinski, I. & Baker, D. Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. Biochemistry 39, 11177–11183 (2000).
8. Dixit, P. D. & Weikl, T. R. A simple measure of native-state topology and chain connectivity predicts the folding rates of two-state proteins with and without crosslinks. PROTEINS: Structure, Function, and Bioinformatics 64, 193–197 (2006).
9. Baiesi, M., Orlandini, E., Seno, F. & Trovato, A. Exploring the correlation between the folding rates of proteins and the entanglement of their native states. J. Phys. A: Math. Theor. 50, 504001 (2017).
10. Taylor, W. R. A deeply knotted protein structure and how it might fold. Nature 406, 916 (2000).
11. Virnau, P., Mirny, L. A. & Kardar, M. Intricate knots in proteins: Function and evolution. PLoS computational biology 2, e122 (2006).
12. Luo, R. C. & Grosberg, A. Y. Statistics of knots, geometry of conformations, and evolution of proteins. PLoS computational biology 2, e45 (2006).
13. Sulkowska, J. I., Sulkowski, P., Szymczak, P. & Cieplak, M. Stabilizing effect of knots on proteins. Proceedings of the National Academy of Sciences 105, 19714–19719 (2008).
14. Bolinger, D. et al. A stevedores protein knot. PLoS Computational Biology 6 (2010).
15. Rawdon, E., Millett, K. & Stasiak, A. Subknots in ideal knots, random knots, and knotted proteins. Scientific Reports 5 (2015).
16. Goundaroulis, D., Dorier, J., Benedetti, E. & Stasiak, A. Studies of global and local entanglements of individual protein chains using the concept of knots. Scientific Reports 7 (2017).
17. Jackson, S. E., Suma, A. & Micheletti, C. How to fold intricately: using theory and experiments to unravel the properties of knotted proteins. Curr. Opinion Struct. Biol. 42, 6–14 (2017).
18. Sulkowska, J., Sulkowski, P. & Onuchic, J. Dodging the crisis of folding proteins with knots. Proceedings of the National Academy of Sciences of the United States of America 106, 3119–3124 (2009).
19. Sulkowska, J., Rawdon, E., Millet, K., Onuchic, J. & Stasiak, A. Conservation of complex knotting and slipknotting patterns in proteins. Proceedings of the National Academy of Sciences of the United States of America 109, E1715–E1723 (2012).
20. Prechet, D. et al. Solution structure of RP 71955, a new 21 amino acid tricyclic peptide active against HIV-1 virus. Biochemistry 33, 42–50 (1994).
21. Niemyska, W. et al. Complex lasso: new entangled motifs in proteins. Scientific Reports 6, 36895 (2016).
22. Dabrowski-Tumanski, P. & Sulkowska, J. I. To tie or not to tie? That is the question. Polymers 9, 454 (2017).
23. Baiesi, M., Orlandini, E., Trovato, A. & Seno, F. Linking in domain-swapped protein dimers. Sci. Rep. 6, 33872 (2016).
24. Bryngelson, J. D. & Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. Proceedings of the National Academy of Sciences 84, 7524–7528 (1987).
25. Frauenfelder, H., Slier, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. Science 254, 1598–1603 (1991).
26. Ferreiro, D. U., Komives, E. A. & Wolynes, P. G. Friction in biomolecules. Quarterly Reviews of Biophysics 47, 285–363 (2014).
27. Ricca, R. L. & Nipoti, B. Gauss’ linking number revisited. Journal of Knot Theory and Its Ramifications 20, 1325–1343 (2011).
28. Dui, M. & Edwards, S. The theory of polymer dynamics (Oxford University Press, 1988).
29. Panagiotou, E., Millet, K., Onuchic, J. & Stasiak, A. Conservation of complex knotting and slipknotting patterns in proteins. Proceedings of the National Academy of Sciences of the United States of America 109, E1715–E1723 (2012).
30. Panagiotou, E., Kroger, M. & Millet, K. C. WRite and mutual entanglement combine to give the entanglement length. Phys. Rev. E 88, 062604 (2013).
31. Connolly, M. L., Kuntz, I. & Crippen, G. M. Linked and threaded loops in proteins. Biopolymers 19, 1167–1182 (1980).
32. Janin, J. & Chothia, C. Domains in proteins: definitions, location, and structural principles. In Methods in enzymology, vol. 115, 420–430 (Elsevier, 1985).
33. Khatib, F., Kohl, C. A. & Karplus, K. Pokefind: a novel topological filter for use with protein structure prediction. Bioinformatics 25, i281–i288 (2009).
34. Bradley, P. et al. Rosetta predictions in casp5: successes, failures, and prospects for complete automation. Proteins: Structure, Function, and Bioinformatics 53, 457–468 (2003).
35. Cossio, P. et al. Exploring the universe of protein structures beyond the protein data bank. PLoS Computational Biology 6, e1000957 (2010).
36. Norcross, T. S. & Yeates, T. O. A framework for describing topological frustration in models of protein folding. Journal of molecular biology 362, 605–621 (2006).
37. Samudrala, R. & Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of molecular biology 275, 895–916 (1998).
38. Lazaridis, T. & Karplus, M. Effective energy functions for protein structure prediction. Current opinion in structural biology 10, 139–145 (2000).
39. Dabrowski-Tumanski, P. & Sulkowska, J. I. Topological knots and links in proteins. Proceedings of the National Academy of Sciences 106,15862 (2017).
40. Kolodny, R., Pereyaslavets, L., Samson, A. & Levitt, M. On the universe of protein folds. Annual Review of Biophysics 42, 559–582 (2013).
41. Rogan, P. & Fain, B. Automatic classification of protein structure by using Gauss integrals. Proceedings of the National Academy of Sciences of the United States of America 100, 119–124 (2003).
42. Dabrowski-Tumanski, P., Stasiak, A. & Sulkowska, J. In search of functional advantages of knots in proteins. Plos One 11 (2016).
43. Englander, M. & Tanford, C. The ribosome can discriminate the chirality of amino acids within its peptidyl-transferase center. Proceedings of the National Academy of Sciences 112, 6038–6043 (2015).
44. Derrida, B. Random-energy model: Limit of a family of disordered models. Physical Review Letters 45, 79 (1980).
45. Gross, D. J. & Mélard, M. The simplest spin glass. Nuclear Physics B 240, 431–452 (1984).
46. Ziegler, F. et al. Knotting and unknotting of a protein in single molecule experiments. Proceedings of the National Academy of Sciences of the United States of America 113, 7533–7538 (2016).
47. Wruck, F., Katranidis, A., Nierhaus, K., Bildt, G. & Hegner, M. Translation and folding of single proteins in real time. Proceedings of the National Academy of Sciences of the United States of America 114, E4399–E4407 (2017).
48. Chwastyk, M. & Cieplak, M. Cotranslational folding of deeply knotted proteins. Journal of Physics Condensed Matter 27 (2015).
49. Bui, P. & Hoang, T. Folding and escape of nascent proteins at ribosomal exit tunnel. Journal of Chemical Physics 144 (2016).
50. Trovato, F. & O’Brien, E. Insights into cotranslational nascent protein behavior from computer simulations. Annual Review of Biophysics 45, 345–369 (2016).
51. Dabrowski-Tumanski, P., Piejko, M., Niewieczerzal, S., Stasiak, A. & Sulikowska, J. I. Protein knotting by active threading of nascent polypeptide chain exiting from the ribosome exit channel. The Journal of Physical Chemistry B 122, 11616–11625 (2018).
52. Dawson, N. et al. Cath: An expanded resource to predict protein function through structure and sequence. Nucleic Acids Research 45, D289–D295 (2017).
53. Duan, Y. et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed phase quantum mechanical calculations. Journal of computational chemistry 24, 1999–2012 (2003).
54. Lindahl, E., Hess, B. & Van Der Spoel, D. Gromacs 3.0: a package for molecular simulation and trajectory analysis. Molecular modeling annual 7, 306–317 (2001).
55. Piana, S. et al. Predicting the effect of a point mutation on a protein fold: the villin and advillin headpieces and their pro62ala mutants. Journal of molecular biology 375, 460–470 (2008).
56. Miyazawa, S. & Jernigan, R. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18, 534–552 (1985).
57. Sippl, M. J. Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. Journal of molecular biology 213, 859–883 (1990).
58. Cossio, P., Granata, D., Laio, A., Seno, F. & Trovato, A. A simple and efficient statistical potential for scoring ensembles of protein structures. Scientific Reports 2, 351 (2012).

Acknowledgements
M.B. acknowledges support from Progetto di Ricerca Dipartimentale BIRD173122/17. F.S. and A.T. acknowledge R. Battistuta and G. Zanotti for fruitful discussions. We thank A. Kabakçoğlu for a careful reading of the manuscript.

Author Contributions
All authors wrote and reviewed the manuscript. M.B. and A.T. analysed data and prepared the figures. All authors contributed extensively to the work presented in this paper.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-44928-3.

Competing Interests: The authors declare no competing interests.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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 license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license 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 license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019