HHS Public Access

Author manuscript

J Mol Biol. Author manuscript; available in PMC 2024 July 23.

Published in final edited form as:

J Mol Biol. 2024 March 15; 436(6): 168459. doi:10.1016/j.jmb.2024.168459.

Non-covalent Lasso Entanglements in Folded Proteins: Prevalence, Functional Implications, and Evolutionary Significance

Viraj Rana[1,†], **Ian Sitarik**[1,†], **Justin Petucci**[2], **Yang Jiang**[1], **Hyebin Song**[3,4], **Edward P. O'Brien**[1,2,3]

[1]-Department of Chemistry, Pennsylvania State University, University Park, PA, United States

[2]-Institute for Computational and Data Sciences, Pennsylvania State University, University Park, PA, United States

[3]-Bioinformatics and Genomics Graduate Program, The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, United States

[4]-Department of Statistics, Pennsylvania State University, University Park, PA, United States

Abstract

One-third of protein domains in the CATH database contain a recently discovered tertiary topological motif: non-covalent lasso entanglements, in which a segment of the protein backbone forms a loop closed by non-covalent interactions between residues and is threaded one or more times by the N- or C-terminal backbone segment. Unknown is how frequently this structural motif appears across the proteomes of organisms. And the correlation of these motifs with various classes of protein function and biological processes have not been quantified. Here, using a combination of protein crystal structures, AlphaFold2 predictions, and Gene Ontology terms we show that in *E. coli, S. cerevisiae* and *H. sapiens* that 71%, 52% and 49% of globular proteins contain one-or-more non-covalent lasso entanglements in their native fold, and that some of these are highly complex with multiple threading events. Further, proteins containing these tertiary motifs are consistently enriched in certain functions and biological processes across these organisms and depleted in others, strongly indicating an influence of evolutionary selection pressures acting positively and negatively on the distribution of these motifs. Together, these

Correspondence to Hyebin Song and Edward P. O'Brien: hps5320@psu.edu (H. Song), epo2@psu.edu (E.P. O'Brien).
[†]These authors contributed equally to this research project.

Credit authorship contribution statement

Viraj Rana: Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation. **Ian Sitarik:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Justin Petucci:** Methodology, Formal analysis. **Yang Jiang:** Resources, Methodology, Investigation. **Hyebin Song:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Formal analysis, Conceptualization. **Edward P. O'Brien:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2024.168459.

results demonstrate that non-covalent lasso entanglements are widespread and indicate they may be extensively utilized for protein function and subcellular processes, thus impacting phenotype.

Keywords

PROTEIN structure; entanglements; go; database; function

Introduction

The remarkable discovery[1] was made four years ago that one-third of protein domains reported in the CATH database[2] contain previously unrecognized tertiary structure elements referred to as non-covalent lasso entanglements. These tertiary structural motifs consist of a protein backbone loop closed by a non-covalent interaction between two residues that is pierced one or more times by a backbone threading segment that can coil around the loop (Figure 1). This class of entanglement is structurally and mathematically distinct from knots, slip-knots, and covalent lasso entanglements that has loops which are closed by covalent disulfide bonds.[3]

Most proteins contain two or more domains,[4,5] however, CATH only reports information on a perdomain basis. Thus, the frequency of non-covalent lasso entanglements in full length proteins has not been characterized across the proteome of globular proteins. Furthermore, the biological consequences, if any, of these widespread tertiary structural motifs are unknown.[1,6,7] This raises questions of whether nature has enriched or depleted these structural motifs in different functional classes of proteins or spatially near functional protein sites - a strong sign, if found, that natural selection is acting on them.

One biological effect that has been established is that the presence of non-covalent lasso entanglements, as measured by the minimum number of times the thread segment pierces the loop in a proteins most complex entanglement, anticorrelates with protein folding times[8] (Pearson $R = -0.74$ for multi-state folders). That is, the more complex this class of entanglement, the longer it takes to fold proteins containing them. As seen for other secondary and tertiary structural elements,[9–14] we hypothesize this class of structural motif can affect protein function and hence be over- or under-represented in biological processes within the cell.

A protein entanglement involves the intertwining of protein backbone loops and segments leading to one of five broad and mathematically distinct categories[3]: (1) a *knot* – in which pulling on both termini of the protein tighten the entanglement preventing the full extension of the protein backbone; (2) a *slipknot* – in which pulling on both termini disentangle the loops leading to full extension of the backbone, and no supercoiling occurs; (3) a *covalent lasso entanglement* – in which a backbone loop closed by a covalent (e.g., disulfide) bond is pierced by a backbone segment; (4) *deterministic links* – in which two backbone loops, each closed by a disulfide bond, pierce one other; and the recently identified category[1] of (5) *non-covalent lasso entanglements* – in which a non-covalent contact closes the loop, supercoiling of the threading segment around the loop can occur, and pulling the termini often results in disentanglement. The first four categories are estimated to be present in

6% of native protein structures[3]; while it is unknown how common non-covalent lasso entanglements are across the proteomes of organisms.

There are two reasons why non-covalent lasso entanglements were largely ignored by researchers until 2019. First, the mathematical algorithms for detecting these topological entanglements improved over time.[15,16] The algorithms have evolved from the "sphere-point" method introduced in 1994,[17] to a 'Knotoid' methodology[7,18,19] that projects the protein backbone coordinates onto a sphere in 2010, to the use of Gauss Linking Integrals in 2017.[8] With each improvement the ability to accurately detect lasso entanglements increased. Indeed, "Gaussian Linking Integrals reveal unique properties of lasso topology in proteins, which may lead to new biological and chemical discoveries" according to a recent paper.[15] Second, the field of protein folding has historically focused on entanglements involving disulfide bonds[15] because these covalent bonds are an energetically strong constraint and therefore more likely to influence the process of protein folding. Thus, the field neglected cataloging these non-covalent lasso entanglements until 2019.[1]

Here, we characterize the frequency of non-covalent lasso entanglements across the cytosolic proteomes of *E. coli*, *S. cerevisiae* and *H. sapiens*, and address other open questions related to their potential structural and biological function by creating a high-quality dataset containing information on globular proteins and the presence and location of non-covalent lasso entanglements in their native structures and cross-referencing it against other publicly available data.

Materials and Methods

Retrieval of genes and protein structures.

All reviewed genes and their corresponding protein crystal structures were obtained from the UniProt-Proteome and Protein Data Bank (PDB) databases for three species: *E. coli* K12 on February 15th, 2021 (1,568 genes, out of 4,389 in the genome, have at least one crystal structure), *S. cerevisiae* S288c on February 23rd, 2022 (1,893 out of 6,050 genes have crystal structures), and *H. sapiens* on April 14th, 2022 (7,423 out of 20,361 have crystal structures).[20] For protein structures, we retrieved the first biological assembly structures which are believed to be the main functional form of the molecule.[21] The files for biological assembly were available with the pdb1 file extension in the PDB. These PDB files were processed to remove any lines after the last termination line (denoted by TER in the PDB file) to remove all non-protein elements in the files. Small molecules, water molecules, and modified amino acids can be found after the last TER line and these non-protein elements reduce the accuracy of identifying protein entanglements. In addition, missing residues in each PDB structure file were identified by comparing the current PDB structure's residue IDs to the vector $(m, m+1, \ldots, n)$, where m is the starting residue ID, and n is the largest residue ID of the file. Lastly, we obtain globular proteins by removing intrinsically disordered proteins and membrane proteins as classified by UniProt.[17]

Selection of genes with high-quality protein structures.

To link sequence and structural information we used Protein BLAST to align the UniProt canonical sequences with the protein chain sequences from the corresponding PDB structures.[22] We then generated a dataset by selecting genes that had high-quality protein structures available. Specifically, to be included in our dataset, a gene had to meet the following criteria: it must have at least one corresponding PDB structure with a resolution of 3 or less, and at least one alignment between the gene's UniProt canonical sequence and the protein chain sequence from the corresponding PDB structure must satisfy the following conditions: a minimum of 95% identity and positive scores, less than 5% gaps, and an Expect score of less than or equal to 10^{-5}. For each gene with at least one high-quality protein structure (i.e., meets all the criteria described above), we selected a representative protein structure for the gene by choosing the chain sequence with the highest coverage of the UniProt canonical sequence. When the protein chain sequence extended beyond the UniProt canonical sequence, we truncated the corresponding sequence parts (and protein structures) to ensure that the representative protein structures only reflected structures that aligned with the UniProt canonical sequence. This allowed us to focus exclusively on the structural information that corresponded to the protein sequence of interest. In the end, our dataset includes 1,294 genes for *E. coli*, 1,023 genes for *S. cerevisiae*, and 5,190 genes for *H. sapiens*, where each gene is associated with the unique representative protein structure.

Entanglement identification.

We first define a few terms. A loop is a segment of the protein backbone between a pair of residues in a native contact. For the residues i, j, we define the spatial distance between i, j as $d(i,j) = min[\parallel b(h) - b(h') \parallel \forall h \in H(i), h' \in H(j)]$ where $b\left(h\right) = (b_x(h), b_y(h), b_z(h)) \in \mathrm{R}^3$ represents the x, y, z coordinates of a heavy atom h and $H(i)$ refer to the set of heavy atoms (non-hydrogen atoms) of residue i. We say a residue pair i, j is a native contact if the spatial distance between i and j, is less than or equal to 4.5 (i.e. $d(i,j) \leq 4.5$). We use the first location in a PDB entry if a residue has multiple alternate orientations. A loop spans the primary structure between i, j that form a native contact: $[i, i+1, \ldots, j-1, j]$. Outside this loop we have an N-terminal segment, composed of residues 1 through $i-1$, and a C-terminal segment, composed of residues $j+1$ through N, where N is the protein length. An entanglement occurs when an N-terminal or C-terminal segment pierces the loop. The residue(s) that pierces the plane of the loop is referred to as the crossing residue(s).

Our overall workflow is illustrated in Figure S4. To identify non-covalent lasso entanglements, we begin by identifying all loops with crossing residues, where either the N-terminal or C-terminal segment pierces through the loop at least once. To do this, we first find all loops with potential crossing events using Gauss linking numbers[1,8] and refine these results using a minimal surface analysis.[15,16] We then employ a clustering method to group together loops with crossing events that are likely part of the same entanglement to avoid duplicates in the final set of identified entanglements. This second step ensures that each entanglement is represented only once in the final analysis. The Gauss linking number characterizes the degree of entanglement between two curves.[8] For two oriented closed curves in 3D space, for example, the Gauss linking number is an integer value that

describes how many times one curve winds around the other.[8] It was previously shown that the Gauss linking number is a useful metric for detecting native self-entanglement events between two sub-segments of a protein backbone.[8] The minimal surface analysis is based on triangulating a loop surface into small triangles, which are termed minimal surfaces, and determining whether or not a segment pierces one of the triangle planes.[16] While minimal surface analysis tends to be more accurate in identifying entanglements due to its verification of the actual piercing events, but it is more computationally intensive compared to the Gauss linking number method. Therefore, in our analysis we use the Gauss linking number method as a pre-liminary screening tool to initially identify all loops with potential crossings and then verify if they have actual piercings in their triangulated surfaces.

$$\left\{ g_N(i,j) = \frac{1}{4\pi} \sum_{m=6}^{i-6} \sum_{n=i}^{j-1} \frac{R(m) - R(n)}{\| R(m) - R(n) \|^3} \cdot (dR(m) \times dR(n)) \right. \tag{1}$$

$$\left\{ g_C(i,j) = \frac{1}{4\pi} \sum_{m=i}^{j-1} \sum_{n=j+6}^{N-6} \frac{R(m) - R(n)}{\| R(m) - R(n) \|^3} \cdot (dR(m) \times dR(n)) \right. \tag{2}$$

Specifically, in the first step, we identify loops closed by native contacts in each protein. We then compute the Gauss linking number for each loop and N/C-terminal segment using Eqs. (1) and (2).[23,24] We note that at this stage we consider all loops and terminal segments regardless of whether they contain missing residues or not. Identified entanglements that contain a certain number or proportion of missing residues are filtered at a later stage. $c\left(I\right) \in R^3$ represents the $C\alpha$ coordinates of the residue I, and we define the mid-point $R(I)$ between $c(I), c(I+1)$ as $R\left(I\right) = \frac{1}{2}\left(c\left(I\right) + c\left(I+1\right)\right)$ and $dR(I) = c(I+1) - c(I)$. In Eqs. (1) and (2) the first five residues of the N-terminal tail, the last five residues of the C-terminal tail, and six residues from either side of the loop are excluded to reduce errors associated arising from flexible regions. Such flexible regions can lead to situations where the crossing residue(s) fail to pierce the loop surface completely. Loop (i, j) has potential crossing event (s) if $|g_N(i,j)| \geq 0.6$ or $|g_C(i,j)| \geq 0.6$. To confirm whether a thread pierces the triangulated loop surface we perform a minimal surface analysis on all loops with potential crossing events. We use the Python package Topoly[16,25] and its lasso type function for this analysis, with the options density parameter = LOW and precision parameter = HIGH.

Finally, to focus on non-covalent lasso entanglements, and avoid false positives we exclude several cases from the detected set of loops with crossing events, including: (1) Loops containing a pair of cysteine residues forming a native contact; (2) Loops from PDB structures that are known or estimated to have knots or slipknots; (3) Loops with more than 5% of missing residues or three consecutive missing residues; and (4) any entanglement with missing residues within 10 residues of the piercing location on the thread. Proteins were

removed from our dataset if criteria 1 or 2 were met, while only specific entanglements were removed if criteria 3 or 4 were met. The first and second criteria were chosen to eliminate proteins containing any covalent lasso entanglements, knots, and slipknots from the detected set of entanglements. The list of PDB structures containing knots was acquired on February 16, 2023, from the KnotProt 2.0 database.[7] The presence of knots in PDB structures that are not in the KnotProt 2.0 database was also estimated using the Alexander Polynomial algorithm from the Topoly Python module.[16,25] If a given PDB structure indicated a knot type other than the unknot, then the structure was considered to have a knot and removed from our dataset. Knots were classified as slipknots using the KnotProt 2.0 database. The third and fourth criteria were implemented because we found the identification of crossing events inaccurate when there are missing residues in the loop or near the piercing location. This may be a result of distorted loop triangulated surfaces, uncertainty in thread location, or inherent errors in the loop closing method used to calculate the Gauss linking numbers. (Note that there is one crossing residue in an entanglement in gene product P13382 to which Topoly assigns two different chiralities. We removed this gene from our analyses.).

With this set of loops with crossing events, we ran a clustering algorithm to cluster loops with crossing events that are likely part of the same entanglement. The idea is to cluster loops with crossing events that are spatially close, and the chirality of piercings is shared. The exact algorithm, ALG1, that we used is provided in Supplementary information. As a result of the clustering algorithm, we have a representative entanglement consisting of a loop and residues on the N/C threads at the point of piercing events per cluster, which we refer to as the *crossing residues*.

Detection of Non-Covalent Lasso Entanglements in AlphaFold-Predicted Protein Structures.

Version 4 of AlphaFold F1 model PDB files were downloaded for *E. coli*, *S. cerevisiae*, and *H. sapiens*.[26,27] We performed the entanglement identification procedure described above to identify the set of non-covalent lasso entanglements in the AlphaFold-predicted structures. When executing the procedure, we used the AlphaKnot database[28] to obtain the list of PDB structures which contain knots or slipknots. Globular proteins were obtained by removing membrane proteins and/or proteins containing any disordered regions, as identified by UniProt.[20]

AlphaFold Entanglement Prediction.

Each identified entanglement was transformed into a 1-D entanglement vector, with a size equal to the length of the corresponding PDB chain sequence (given by the experimental or AlphaFold predicted PDB). The vector elements are binary (0 or 1), where nonzero elements represent crossing residues as well as thread or loop residues located within 4.5 of heaving atoms of a crossing residue. All representative entanglements corresponding to a given PDB were combined into a single PDB-level entanglement vector by applying the logical OR operator element-wise.

To allow for direct comparison, the mapping between the UniProt canonical sequences and the PDB chain sequences was used to make the experimental and AlphaFold PDB-

level entanglement vectors the same length. This was accomplished by first extracting the list of residues in the experimental PDB chain that maps to the UniProt canonical sequence. For each residue that was successfully mapped, the corresponding element from the experimental and AlphaFold PDB-level entanglement vectors was selected. This process produces new experimental and AlphaFold PDB-level entanglement vectors that are the same length; equal to the length of the corresponding mapped PDB chain sequence.

To estimate the performance of AlphaFold in predicting which residues are associated with entanglement, the PDB-level entanglement vectors were compared element wise (using the experimental entanglement vector as ground-truth). Given the imbalance between the number of 0 and 1 elements in the entanglement vectors, which have a 0 : 1 ratio of 8.2 across all species, the balanced accuracy, defined by,

$$Balanced\,Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{3}$$

was used as the performance metric; where $TP = \text{True Positive}$, $FN = \text{False Negative}$, and $FP = \text{False Positive}$, $TN = \text{True Negative}$. The following number of PDBs were not included in the analysis: *E. coli*: 2, *H. sapiens*: 34. These were excluded due to: (1) Not being present in the AlphaFold Database, (2) having an AlphaFold PDB chain sequence different from the UniProt canonical sequence, or (3) it was computationally expensive to complete the clustering phase because of the sheer number of higher order entanglements.

Gene Ontology (GO) Statistical Association Analysis.

To obtain a comprehensive set of functional annotations for each gene in our dataset, we extracted Gene Ontology (GO) annotations from UniProt reviewed entries.[20,29] These annotations provide information about the biological processes, molecular functions, and cellular components associated with each gene. However, since the annotations obtained from UniProt are highly granular, we utilized the QuickGO database[30] and Kahn's algorithm for topological sorting as implemented in the NetworkX python package[31,32] to cluster the UniProt annotations into broader categories. This involves using the ancestral chart of each annotation, represented as a directed acyclic graph, to transform it into level 2 and level 3 categories.[33,34] The levels indicate the position of an annotation in the chart relative to the root GO class.

We tested whether there is any association between the existence of an entanglement in a protein structure and a specific GO annotation at each level. For each annotation, we created a 2-by-2 contingency table where the rows represented the presence or absence of genes with non-covalent entanglements, while the columns represented the presence or absence of a specific GO annotation. We used a two-sided Fisher's exact test to determine the statistical association between the presence of entanglement and a specific GO annotation. To determine enrichment or depletion for significant annotations, we used a one-sided Fisher's exact test.

The odds ratio, the two-tail p-values, and enrichment/depletion p-values were obtained for all annotations in a GO class. The two-tail p-values were adjusted using the Benjamini-Hochberg false discovery rate correction at a significance level of 0.05. The adjusted two-tail p-values were used to determine whether an annotation was significant, and the other two p-values determined whether the significant annotation was enriched or depleted. Excel files (Supplementary Files 1–3) were generated for *E. coli* (November 15, 2022), *S. cerevisiae* (November 22, 2022), and *H. sapiens* (November 23, 2022). Each contains three worksheets representing a GO class. Columns represent: 1) GO term; 2) Odds Ratio; 3) Enriched p-value; 4) Depletion p-value; 5) Two-tail p-value; 6) Adjusted two-tail p-value; 7) Number of entangled proteins with the annotation; 8) Number of entangled proteins without the annotation, 9) Number of non-entangled proteins with the annotation; 10) Number of non-entangled proteins without the annotation; 11) Number of high-quality proteins with the annotation; 12) Number of high-quality proteins without the annotation; and 12) whether the annotation was significant after hypothesis correction.

Statistical analysis for structural enrichment of non-covalent lasso entanglements near functional sites.

We conduct hypothesis tests to determine if entanglements tend to be spatially located close to or far from functional sites of a protein for each species. We say that entanglements are *enriched* near a functional site F in gene-product g if crossing residues occur more frequently near functional site F, or equivalently, if crossing residues are located closer to F compared to a randomly chosen residue. Conversely, we say entanglements are *depleted* from a functional site F in g if the crossing residues are located far from the functional site F on average when compared to randomly chosen residues.

In our analysis, we focused on several types of functional sites, including active sites, zinc fingers, DNA binding sites, RNA-binding sites, protein–protein interfaces, small molecule binding sites, and metal binding sites. The active sites and zinc fingers sites were directly retrieved from UniProt. UniProt defines active sites as residues directly involved in catalysis and zinc fingers as residues that coordinate zinc ions.[20] The location of these sites for each gene were mapped to the corresponding residues in the representative chain sequence using the top Protein BLAST alignment result. For the other functional sites, we used the representative protein structure of each gene product and identified the sites using heavy atoms (non-hydrogen atoms) and a 4.5 spatial cut-off between the coordinates of heavy atoms in the bound molecule(s) or ion(s) and those in the protein. We consider the protein as the standard 20 amino acids and those modified amino acids that are covalently attached to the protein backbone. The standard four nucleotides were used to identify the bound molecule as DNA or RNA. The bound ion is a metal if it is part of groups 1–12 (not including hydrogen), metalloids, post-transition metals, and the f-block in the periodic table. The metals were referenced against the'PDB Ligand Summary Page' for their existence in PDB structures. Small molecules here refer to free ligands non-covalently attached to the protein.[21] They were obtained using a spatial cutoff of 1.97 between the atom and heteroatom records that do not contain hydrogens. This criterion takes advantage of the hydrogen bond length in water to find small molecules. Metal binding sites were

identified using a 4.5, cut-off, excluding active sites and zinc fingers sites. We also removed non-functional common solvents found in the crystal structures for each proteome.

For each given gene, we use a permutation test to test whether entanglements are enriched or depleted near a specific functional site F. In a permutation test, one compares the test statistic computed from the observed data with the test statistics that are recomputed over certain permutations of the observed data. The null hypothesis is rejected if the observed test statistic is sufficiently extreme compared to the test statistic from the permuted data. Here, the key condition is that one should use the set of permutations that, under the null hypothesis, do not make the distribution of the permuted data different from the original data distribution. For instance, suppose one wants to test whether observations from two groups $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_n)$ are from the same distribution or not using a permutation test. Here the total data set is $Z = (X_1, \ldots, X_m, Y_1, \ldots, Y_n) = (Z_1, \ldots, Z_N)$, where the first m observations correspond to the elements in the first group and $N = m + n$. One can obtain the k^{th} permuted data set $Z^{(k)} = \left(Z_1^{(k)}, \ldots, Z_N^{(k)}\right)$ by choosing m elements randomly from the observed dataset Z where the first m elements of the $Z^{(k)}$ correspond to the m randomly chosen elements and the last n elements are those remaining in Z. Note the distribution of the permuted data $Z^{(k)}$ is the same as the distribution of Z under the null hypothesis. Once K permuted datasets are obtained, one can compare the observed test statistic $T(Z)$ with the test statistics from different permutations $T\left(Z^{(k)}\right)$ for $k = 1, \cdots, K$.

Returning to the original testing problem, we consider a test that assesses whether the minimum distances from residues to the functional site F follow the same distribution for crossing and non-crossing residues. First, for each residue i in the gene g, let us define the minimum distance of the residue i to the functional site F by $d_F\left(i\right) = \min_{j \epsilon F} \| c\left(i\right) - c\left(j\right) \|$ where $c(i)$ refers to the C_α coordinates of the residue i, $c(j)$ is the C_α coordinates of the residue j and minimum is taken over all residue j in functional site F. In this analysis, we use the *ranks* of the minimum distances to F instead of the original distances. Use of ranks helps to obtain a more robust result, which is less likely to be influenced by outliers. The data is $Z = (Z_1, \ldots, Z_N)$ where Z_i corresponds to the i^{th} residue's rank of the minimum distances. For example, if the residue i has the smallest minimum distance, $Z_i = 1$. Ties receive a rank equal to the average of the ranks they span. The null hypothesis is the distribution of Z_i from the crossing residue group is the same as the distribution of Z_i from the non-crossing residue group. We use the sum of the ranks of the residues in the crossing residue group, i.e., $T(Z) = \sum_{i \epsilon C} Z_i$ as the test statistic, where C is the set of crossing residues.

Similar to the example above, for the data $Z = (Z_1, \cdots, Z_N)$, a permuted data $Z^{(p)}$ from the permutation p can be obtained by generating a random subset I_p of size m from $\{1, \cdots, N\}$ and reordering Z so that $\{Z_i; i \in I_p\}$ are in C, where $m = |C|$ is the number of crossing residues from gene g. Here $\{Z_i; i \in I_p\}$ are the ranks of the crossing residues in $Z^{(p)}$. The test statistic for the permuted data $Z^{(p)}$ is $T\left(Z^{(p)}\right) = \sum_{i \epsilon I_p} Z_i$, i.e., the sum of ranks of the crossing residues in the permuted data $Z^{(p)}$. However, unlike the previous two sample permutation example, a

significant complication in this analysis is that not all permutations of the data would satisfy the topological constraint of C_α coordinates. In particular, there are positions $I_p \subseteq \{1, \cdots, N\}$ where no rewiring of the amino acid residues would result in entanglements in I_p, making the corresponding permutation p invalid. Therefore, the set of possible permutations P has to be limited to those whose permutations are compatible with the given C_α coordinates and distances from the population. If we can uniformly sample permutations $p_1, \ldots, p_B$ from such set of valid permutations P, we can construct the p-value for the tests under the null hypothesis of no association. For example, the p-value for testing whether the crossing residues are spatially enriched near F can be constructed by[35] :

$$p_{enr} = \frac{1}{B+1}\left[1 + \sum_{k=1}^{B} I\left\{T\left(Z^{(k)}\right) \geq T\left(Z\right)\right\}\right] \tag{4}$$

where $Z^{(k)}$ is the generated data from the permutation p_k. The test which compares p_{enr} with the pre-specified significance level α is a valid level α test, i.e., the type I error probability is less than or equal to α.

Unfortunately, enumerating the set of valid permutations P is difficult due to the computational intractability of modeling all possible 3-dimensional structures of permuted sequences subject to given C_α coordinates. One necessary condition for a crossing residue is that it has to be located in a buried part of the protein, due to the topological properties of entanglements. We use the normalized surface accessible solvent area (SASA) of a residue,[36,37] defined as the surface accessible solvent area of the residue divided by the average surface accessible solvent area of the protein, as a measure of the degree of burial of the residue in the protein. Assuming that crossing residues from the permuted data $Z^{(k)}$ with any valid permutation $p_k \in P$ share a similar degree of burial as the observed crossing residues, we select crossing residues for permuted data in a way that the resulting distribution of normalized SASA is similar to the observed distribution of the normalized SASA from the crossing residues in the proteome. The exact algorithm for sampling crossing residues is given as ALG2 in the Supplementary information.

After generating the set of $B = 50,000$ permuted data set $Z^{(1)}, \ldots, Z^{(B)}$ from ALG2, we compute the p-value to test the enrichment p_{enr} using Equation (3). We compute the p-value for the depletion p_{dep} similarly, with the direction of inequality reversed. For the two-tailed test, we compute the p-value as $2 \cdot min(p_{enr}, p_{dep})$. To control the false discovery rate in each species, since we are testing whether entanglements are enriched or depleted in each species, we apply the Benjamini-Hochberg false discovery correction to the p-values from each gene.[38]

Results

49% to 71% of globular native structures contain non-covalent lasso entanglements.

We analyzed all unique globular proteins in the three species (*E. coli*, *S. cerevisiae*, and *H. sapiens*) that have a high-resolution crystal structure in the Protein Data Bank (n = 1,294 for *E. coli*, n = 1,023 for *S. cerevisiae*, n = 5,190 for *H. sapiens*) and maximally cover the canonical protein sequence. In *E. coli* we find 71% (921 out of 1,294) of these protein structures contain one or more non-covalent lasso entanglements. The median number of entanglements per protein is two, with a mode of one (Figure 2a). And the maximum number observed in any one protein is 28 (Gene P09152, PDB 1Q16 Chain A, interactive visualization at https://obrien-lab-psu.github.io/Non-covalent-Lasso-Entanglements-in-Folded-Proteins-Prevalence-Functional-Implications-and-Evolut/). 80% of the entanglements contain a single crossover residue (i.e., one piercing event of the plane of the loop by the threading segment), 10% have only two crossover residues (two piercing events), and 6% have only three crossovers (Figure 2c). The maximum number of cross-over residues observed in any single entanglement is seven residues in gene-products P21179 (PDB 4BFL, Chain B) and P24171 (PDB 1Y79 chain 1, interactive visualization available). Loop sizes in these entanglements range from 9 to 812 residues in length, with a median of 72 and mode of 26 residues (Figure 2b).

In *S. cerevisiae*, 52% (538 out of 1,023) of native structures have non-covalent lasso entanglements. The median number of lasso entanglements per protein is two, with a mode of one (Figure 3a). And the maximum number observed in any one protein is 15 (Gene P22138, PDB 4C2M Chain B, interactive visualization available). 83% of the entanglements have only one crossover residue, 8% have two crossover residues, and 6% have three crossover residues (Figure 3c). The maximum number of cross-over residues observed in any entanglement is seven residues in gene-products P39958 (PDB 2BCG, Chain G), P14743 (PDB 1A4E Chain A, interactive visualization available), and P14743 (PDB 2P6E Chain A). Loop sizes in these entanglements range from 9 to 699 residues in length, with a median of 69 and mode of 41 residues (Figure 3b).

And for the human proteome 49% (2,564 out of 5,190) of globular proteins have non-covalent lasso entanglements. The median number of lasso entanglements per protein is two, with a mode of one (Figure 4a). And the maximum number observed in any one protein is 19 (Gene P31327, PDB 5DOU Chain B and Gene P47989, PDB 2E1Q Chain B, interactive visualization available). 84% of entanglements have only one crossover residue, 8% have two crossover residues, and 5% have three crossover residues (Figure 4c). The maximum number of cross-over residues observed in any entanglement is nine residues in gene-product P14735 (PDB 2G54, Chain A, and Q96HY7 (PDB 6U3J Chain B, interactive visualization available). Loop sizes in these entanglements range from 8 to 1,067 residues in length, with a median of 59 and mode of 37 residues (Figure 4b).

We emphasize that the aforementioned statistics on non-covalent lasso entanglements do not include any contributions from knots, slip-knots, nor covalent lasso entanglements as proteins that contain these were removed from our dataset. We find knots, slip-knots, and covalent lasso entanglements only rarely co-occur in the same proteins that non-covalent

lasso entanglements are present in. They co-occur in 2.8%, 3.5%, and 8.4% of the non-covalent-lasso-entanglement-con taining proteins, respectively, in *E. coli*, *S. cerevisiae*, and *H. sapiens*.

These results demonstrate that across these three organisms non-covalent lasso entanglements occur in the majority or near majority of globular proteins, they vary widely in their number per protein and complexity in terms of the number of times the threading segment pierces the plane of the loop.

AlphaFold2 structures yield similar results.

We next asked whether AlphaFold2 accurately predicts the location of non-covalent lasso entanglements in proteins by comparing our dataset based on PDB crystal structures to AlphaFold2 predictions for the same proteins. We find the predictions yield balanced-accuracy values of 0.91, 0.90, and 0.88 for E.coli, *S. cerevisiae*, and *H. sapiens*, respectively. Indicating that AlphaFold structures correctly predict the location of entanglements in approximately 90% of instances.

Because of this accuracy we next used the AlphaFold2 predictions to estimate the fraction of the globular proteome that contains non-covalent lasso entanglements including those proteins that have yet to have a resolved crystal structure. We applied the same analysis to AlphaFold2 predicted structures of globular proteins in *E. coli*, *S. cerevisiae*, and *H. sapiens* that do not contain knots, slipknots, or covalent lasso entanglements. We find 65.8% of *E. coli* globular proteins (2,302 out of 3,498), 62.3% of *S. cerevisiae* globular proteins (2, 953 out of 4,737), and 54.0% of *H. sapien* globular proteins (8,420 out of 15,591) contain one or more entanglements. When membrane and intrinsically disordered proteins are included these percentages remain similar at 65.9% (2,874 out of 4,361), 62.4% (3,766 out of 6,039), and 54.2% (10,899 out of 20,118) in, respectively, E.coli, *S. cerevisiae*, and *H. sapiens* (see Methods). These results are consistent with a recent finding that approximately 60% of membrane protein-domains have entanglements based on our Gaussian linking cutoff of 0.6 (Figure 2b from reference 37).[39] Thus, whether using structures from the PDB or predicted via AlphaFold2, our conclusion is the same: a majority or near majority of globular proteins contain non-covalent lasso entanglements.

Entangled proteins are enriched in some molecular functions, depleted in others.

Next, we examined whether this class of entanglement is enriched or depleted in specific molecular function classes using Level 2 and Level 3 categories defined and assigned by the EMBL-EBI QuickGO database.[30] Molecular functions are defined as “molecular activities of individual gene products” according to the Gene Ontology Consortium.[29] To do this, we use 2-by-2 contingency tables categorizing each protein as either having one-or-more entanglements or not, and having a particular molecular function or not. For example, in our *E. coli* protein dataset, the molecular function category “small molecule binding” contains 76 proteins that have entanglements and this molecular function, 845 protein that have entanglements but do not have this function, 7 proteins that do not have an entanglement but do have this function, and 366 proteins that neither have an entanglement nor this function (Figure 5). This leads to an odds-ratio (OR) of 4.70 (enriched p-value = 2.95 ×

10^{-6}, Fisher's Exact Test), meaning that there is a strong and significant association (an enrichment) between the presence of entanglements in proteins and the tendency for that protein to carry out the function of small molecule binding.

Odds-ratio results for other *E. coli* molecular functions are provided in SI File 1. For brevity we display in Figure 5 the odds-ratios for the top 5 enriched, and top 5 depleted molecular functions in *E. coli*. We find statistically significant associations between the *presence* of entanglements and the molecular functions 'ligase activity' (OR = 5.45), 'catalytic activity acting on a nucleic acid' (OR = 5.04), 'catalytic activity' (OR = 4.81), 'small molecule binding' (OR = 4.70), and 'isomerase activity' (OR = 2.84). And a significant association between the *absence* of entanglements and the molecular functions 'structural molecule activity' (OR = 0.235), 'transcription regulator activity' (OR = 0.231), 'molecular transducer activity' (OR = 0.182), and 'DNA-Binding transcription factor activity' (OR = 0.139).

We report the same results for *S. cerevisiae* and *H. sapien* proteins in Figure 5. (All odds-ratio results for molecular functions are reported in SI Files 2 and 3) Taking the intersection across these three species, we find that the functions that are universally enriched with proteins containing non-covalent lasso entanglements are small molecule binding, lyase activity, transferase activity, oxidoreductase activity, ligase activity, catalytic activity on nucleic acid, catalytic activity, ion binding, isomerase activity and hydrolase activity. And universally depleted are structural molecule activity, DNA-binding transcription factor activity, and transcription regulator activity. Thus, there are both organism specific associations between entanglements and molecular functions, and what appear to be conserved associations across species.

Entangled proteins are enriched in some biological processes, depleted in others.

We carried out the same analysis but for enrichment or depletion of non-covalent lasso entanglements in the GO ontologies associated with the biological processes individual proteins are involved in. A biological process is defined as "the pathways and larger processes to which the gene product's activity contributes" and is distinct from molecular function because biological process refers to a series of events driven by molecular interactions that contribute to a particular biological function.[29] These molecular interactions may refer to a specific biochemical pathway or larger-scale events such as "cell division".

In *E. coli* the strongest associations between the presence of entanglements and biological processes are organic substance metabolic process (OR = 2.54), primary metabolic process (OR = 2.44), metabolic process (OR = 2.28), catabolic process (OR = 2.24), small molecule metabolic process (OR = 1.72), and the absence of entanglements and biological processes are biological regulation (OR = 0.404), cellular localization (OR = 0.368), and cell aggregation (OR = 0.312); all of which are statistically significant (p-value < 0.05). Results for *S. cerevisiae* and *H. sapiens* proteins are reported in Figure 6 and Supplementary Files 2 and 3.

Cellular metabolic process, metabolic process, organic substance metabolic process, primary metabolic process, and small molecule metabolic process are enriched across all three

organisms, while cellular localization and localization are associated with depletion of entanglements.

Association with cellular components.

We repeated this process for the cellular component ontologies, where the cellular component label is defined as where inside a cell "the gene products are active."[29] In *E. coli* the strongest associations between the presence of entanglements and cellular components are intracellular anatomical structure (OR = 1.51), and the absence of entanglements and cellular components are cell periphery (OR = 0.54), organelle (OR = 0.52), and transcription regulator complex (OR = 0.11). Results for yeast and human proteins are reported in Figure 7 and Supplementary Files 2 and 3. Overall, there is no consistency across the three organisms for significant cellular component annotations.

Rarely, entanglements are spatially enriched near functional sites.

Next, we tested if non-covalent lasso entanglements are spatially enriched or depleted near functional residues within the native folds of globular proteins. Functional residues refer to specific amino acids within protein structures that play direct roles in biological functions, such as those involved in the binding of substrates to the protein. The functional categories we use in this analysis, which are different than the GO ontological terms, are 'protein – protein interface', 'DNA binding', 'RNA binding', 'Zinc finger region', 'active site', 'metal binding' and 'small molecules' - as defined in the Methods Section. Specifically, for a given protein's native structure we examined if the minimum rank ordered distance between the crossing residue(s) in an entanglement and a given category of functional residues is closer or further away than random chance (see Methods). If it was closer, then that entanglement was spatially enriched near the functional residue(s), while further away than random chance indicated depletion. This analysis also detects situations where there is neither enrichment nor depletion – that is, the relative locations are indistinguishable from random chance.

After correcting p-values for the false discovery rate (see Methods), we find that in *E. coli* (Table 1) three proteins exhibit a significant spatial enrichment of an entanglement near the 'Protein-protein interface' residues, another near 'active site' residues, and 15 proteins exhibit entanglement(s) statistically enriched near residues that bind 'small molecules.' We never observe spatial depletion of entanglements in these individual functional categories and in most cases there is neither enrichment nor depletion. Combining all functional categories together for this analysis – meaning we identify residues as functional, or not, without regard to their specific functional class - we find 23 are enriched, three are depleted, and 779 are neither enriched nor depleted.

In *S. cerevisiae* the numbers of proteins with spatially enriched entanglements near the residues involved in 'RNA binding' ($n = 1$), 'active site' ($n = 2$), 'protein–protein interfaces' ($n = 2$), and 'small molecules' ($n = 9$) functional classes (Table 1). Spatial depletion of entanglements is also observed in one protein for residues involved in 'Active site', 'Protein-protein interfaces' and another for residues that are part of a 'Zinc finger.' Two proteins are depleted in 'small molecules', and in most cases there is neither enrichment nor depletion.

Like in *E. coli*, combining all functional categories together for this analysis, we find 10 are enriched, two are depleted, and 425 are neither enriched nor depleted.

In *H. sapiens* six and 33 proteins, respectively, exhibit spatially enriched entanglements in residues involved in 'RNA binding' and binding 'small molecules' (Table 1). And the numbers for spatially depleted entanglements are one, one, and four for residues that are involved in RNA-binding, zinc-fingers and small molecules. In most cases there is neither enrichment nor depletion. Combining all functional categories together for this analysis, we find 38 are enriched, 6 are depleted, and 2,127 are neither enriched nor depleted.

These results indicate that for at least 97% of proteins that contain native, non-covalent lasso entanglements and functional residues that those entanglements exhibit no selection pressure to be structurally enriched or depleted near those functional sites. This suggests that for these proteins there is no positive or negative effect on protein function from the presence of non-covalent lasso entanglements. However, there is a small subset of proteins for which there is a bias greater than expected by random chance to find crossing residues at or near certain functional residues. And less frequent is an even smaller number of proteins where the crossing residues are further than would be expected. Thus, this combination of results points towards a selection benefit to the spatial placement of these entanglements in the native folds of a very small number of proteins.

Discussion

We have found that non-covalent lasso entanglements are common across the globular proteome of three diverse organisms, and that natural selection is acting on the distribution of entangled proteins occurring in different functional and biological process classes – enriching them in some and depleting them in others. In contrast to earlier work which found one-third of protein domains contain non-covalent lasso entanglements[1] we find 71%, 52%, and 49% of globular proteins have non-covalent lasso entanglements, respectively, in *E. coli*, *S. cerevisiae*, and *H. sapiens*. These higher percentages are expected as most proteins contain two or more domains. What is remarkable, however, is that so little is known about the functional and biological roles of this widespread class of tertiary structure.

Our statistical analyses of GO ontologies addressed this issue. Across these three very different species, natural selection has consistently enriched this type of tertiary structure in the biological processes of metabolic and cellular processes, and the functional classes of enzymes, ion-binding and small molecule binding activity. Natural selection has selected against the presence of these entanglements in biological regulation process related to transcription and functions related to the structural integrity of the cell. Carrying out a structural enrichment analysis, we observed that small molecules functional class is consistently enriched across these different organisms. This consistent enrichment indicates that non-covalent lasso entanglements are likely to be biologically beneficial across diverse evolutionary lineages and the environments in which these organisms exist.We observe that 2% (57/2,446) of proteins have entanglements that are spatially enriched near small molecule binding sites. This suggests there are rare situations where these motifs might directly influence functions involving small molecules. But in general, most non-covalent

lasso entanglements do not appear to be being utilized to directly influence function – in which case we would expect these entanglements and functional sites to spatially co-occur.

These two observations raise the question as to why proteins containing non-covalent lasso entanglements are strongly overrepresented in certain functional classes (such as in enzymes) but not spatially enriched close to the corresponding functional sites? One possibility is that these entanglements are being used for allosteric regulation of function. Since allostery can be indirect, influencing active site activity from a distance, if the entanglements have an allosteric role, they might not be spatially enriched near active sites. Or, it could be the case that instead of influencing protein function these entanglements are instead influencing other protein properties such as protein stability, lifetime, or dynamics. For example, there are limited reports of knotted proteins increasing protein thermodynamic stability,[40,41] the same might be true for proteins containing non-covalent lasso entanglements in their folded structure. And proteins containing non-covalent lasso entanglements might exhibit differential degradation rates.[8,42]

The derivative data sets we have created, that include detailed information on the location of entanglements in proteins and their associated functional enrichments provides a starting point for follow-up experimental and computational studies on specific proteins and the influence of their native entanglements on protein functions and biological processes. An interesting follow-up computational study would be to further categorize non-covalent lasso entanglements using advanced topological metrics like those applied to covalent lasso entanglements.[4]

In summary, our findings demonstrate that non-covalent lasso entanglements are widespread across the proteomes of various species and the effects of both positive and negative selection pressures have enriched and depleted these motifs in certain functional and biological classes. Disentangling their role in molecular, structural, biochemical, and cellular processes is likely to yield interesting and unexpected results in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

E.P.O. gratefully acknowledges support from the National Science Foundation (MCB-2031584) as well as from the National Institutes of Health (R35-GM124818). Portions of numerical computations and data analysis in this work have been carried out on high-performance computing collectively known as ROAR, which is operated by the Institute for Computational and Data Sciences at The Pennsylvania State University.

DATA AVAILABILITY

The data is provided as supplementary files, the code is shared on Github (https://github.com/obrien-lab-psu/Non-covalent-Lasso-Entanglements-in-Folded-Proteins-Prevalence-Functional-Implications-and-Evolut.git).

References

1. Baiesi M, Orlandini E, Seno F, Trovato A, (2019). Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding. Sci. Rep 9
2. Dawson NL et al. , (2017). CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 45, D289–D295. [PubMed: 27899584]
3. Sulkowska JI, (2020). On folding of entangled proteins: knots, lassos, links and θ-curves. Curr. Opin. Struct. Biol 60, 131–141. 10.1016/j.sbi.2020.01.007. [PubMed: 32062143]
4. Barrera A, Alastruey-Izquierdo A, Martín MJ, Cuesta I, Vizcaíno JA, (2014). Analysis of the protein domain and domain architecture content in fungi and its application in the search of new antifungal targets. PLoS Comput. Biol 10,1003733.
5. Vogel C et al. , (2004). Structure, function and evolution of multidomain proteins This review comes from a themed issue on Theory and simulation Edited. Curr. Opin. Struct. Biol 14, 208–216. [PubMed: 15093836]
6. Niemyska W et al. , (2016). Complex lasso: new entangled motifs in proteins. Sci. Rep 6
7. Dabrowski-Tumanski P et al. , (2019). KnotProt 2.0: a database of proteins with knots and other entangled structures. Nucleic Acids Res. 47, D367–D375. [PubMed: 30508159]
8. Baiesi M, Orlandini E, Seno F, Trovato A, (2017). Exploring the correlation between the folding rates of proteins and the entanglement of their native states. J. Phys. A Math. Theor. 10.1088/1751-8121/aa97e7.
9. Lu C-H et al. , (2006). The fragment transformation method to detect the protein structural motifs. Proteins Struct. Funct. Bioinf 63, 636–643.
10. White SW, Appelt K, Wilson KS, Tanaka I, (1989). A protein structural motif that bends DNA. Proteins Struct. Funct. Genet 5, 281–288. [PubMed: 2508086]
11. Dimitriou PS, Denesyuk AI, Nakayama T, Johnson MS, Denessiouk K, (2019). Distinctive structural motifs co-ordinate the catalytic nucleophile and the residues of the oxyanion hole in the alpha/beta-hydrolase fold enzymes. Protein Sci. 28, 344–364. [PubMed: 30311984]
12. Truong AB, Masters SC, Yang H, Fu H, (2002). Role of the 14-3-3 C-terminal loop in ligand interaction. Proteins Struct. Funct. Bioinf 49, 321–325.
13. Mackenzie CO, Grigoryan G, (2017). Protein structural motifs in prediction and design. Curr. Opin. Struct. Biol 44, 161–167. [PubMed: 28460216]
14. Bittrich S, Burley SK, Rose AS, (2020). Real-time structural motif searching in proteins using an inverted index strategy. PLoS Comput. Biol 16, e1008502. [PubMed: 33284792]
15. Niemyska W, Millett KC, Sulkowska JI, (2020). GLN: a method to reveal unique properties of lasso type topology in proteins. Sci. Rep 10
16. Dabrowski-Tumanski P, Rubach P, Niemyska W, Gren BA, Sulkowska JI, (2021). Topoly: Python package to analyze topology of polymers. Brief. Bioinform 22, 1–8. [PubMed: 33401308]
17. Mansfield ML, (1994). Are there knots in proteins? Nature Struct. Mol. Biol 1, 213–214.
18. Jamroz M et al. , (2015). KnotProt: a database of proteins with knots and slipknots. Nucleic Acids Res. 43, D306–D314. [PubMed: 25361973]
19. Turaev V, (2010). Knotoids. Osaka J. Math. (Wuhan) 49, 195–223.
20. The UniProt Consortium, (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47, D506–D515. [PubMed: 30395287]
21. Zardecki C et al. , (2022). PDB-101: Educational resources supporting molecular explorations through biology and medicine. Protein Sci. 31, 129–140. [PubMed: 34601771]
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, (1990). Basic local alignment search tool. J. Mol. Biol 215, 403–410. [PubMed: 2231712]
23. Jiang Y et al. , (2022). How synonymous mutations alter enzyme structure and function over long timescales. Nature Chem.. 10.1038/s41557-022-01091-z.
24. Nissley DA et al. , (2022). Universal protein misfolding intermediates can bypass the proteostasis network and remain soluble and less functional. Nature Commun. 13
25. Millett KC, Rawdon EJ, Stasiak A, Sułkowska JI, (2013). Identifying knots in proteins. Biochem. Soc. Trans 41,533–537. [PubMed: 23514149]

26. Jumper J et al. , (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. [PubMed: 34265844]
27. Varadi M et al. , (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 50, D439–D444. [PubMed: 34791371]
28. Niemyska W et al. , (2022). AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods. Nucleic Acids Res. 50, W44–W50. [PubMed: 35609987]
29. Ashburner M et al. , (2000). Gene ontology: tool for the unification of biology. Nature Genet. 25, 25–29. [PubMed: 10802651]
30. Binns D et al. , (2009). QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 25, 3045–3046. [PubMed: 19744993]
31. Kahn AB, (1962). Topological sorting of large networks. Commun. ACM 5, 558–562.
32. Hagberg AA, Schult DA & Swart PJ (2008). Exploring Network Structure, Dynamics, and Function using NetworkX.
33. Ventoso P et al. , (2019). RNA-Seq transcriptome profiling of the queen scallop (aequipecten opercularis) digestive gland after exposure to domoic acid-producing pseudonitzschia. Toxins (Basel) 11, 97 [PubMed: 30736356]
34. Diz AP, Romero MR, Pérez-Figueroa A, Swanson WJ, Skibinski DOF, (2018). RNA-seq data from mature male gonads of marine mussels Mytilus edulis and M. galloprovincialis. Data Brief 21, 167. [PubMed: 30364736]
35. Lehmann EL, Romano JP, (2005). Testing Statistical Hypotheses. Springer, New York 10.1007/0-387-27605-X.
36. McGibbon RT et al. , (2015). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys. J 109, 1528–1532. [PubMed: 26488642]
37. Shrake A, Rupley JA, (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol 79, 351–371. [PubMed: 4760134]
38. Korthauer K et al. , (2019). A practical guide to methods controlling false discoveries in computational biology. Genome Biol. 20, 118. [PubMed: 31164141]
39. Salicari L, Trovato A, (2023). Entangled Motifs in Membrane Protein Structures. Int. J. Mol. Sci 24, 9193, [PubMed: 37298146]
40. Zhao Y, Cieplak M, (2018). Stability of structurally entangled protein dimers. Proteins Struct. Funct. Bioinf 86, 945–955
41. King NP, Yeates EO, Yeates TO, (2007). Identification of rare slipknots in proteins and their implications for stability and folding. J. Mol. Biol 373, 153–166. [PubMed: 17764691]
42. Zhu M et al. , (2022). Pulse labeling reveals the tail end of protein folding by proteome profiling. Cell Rep. 40, 111096 [PubMed: 35858568]

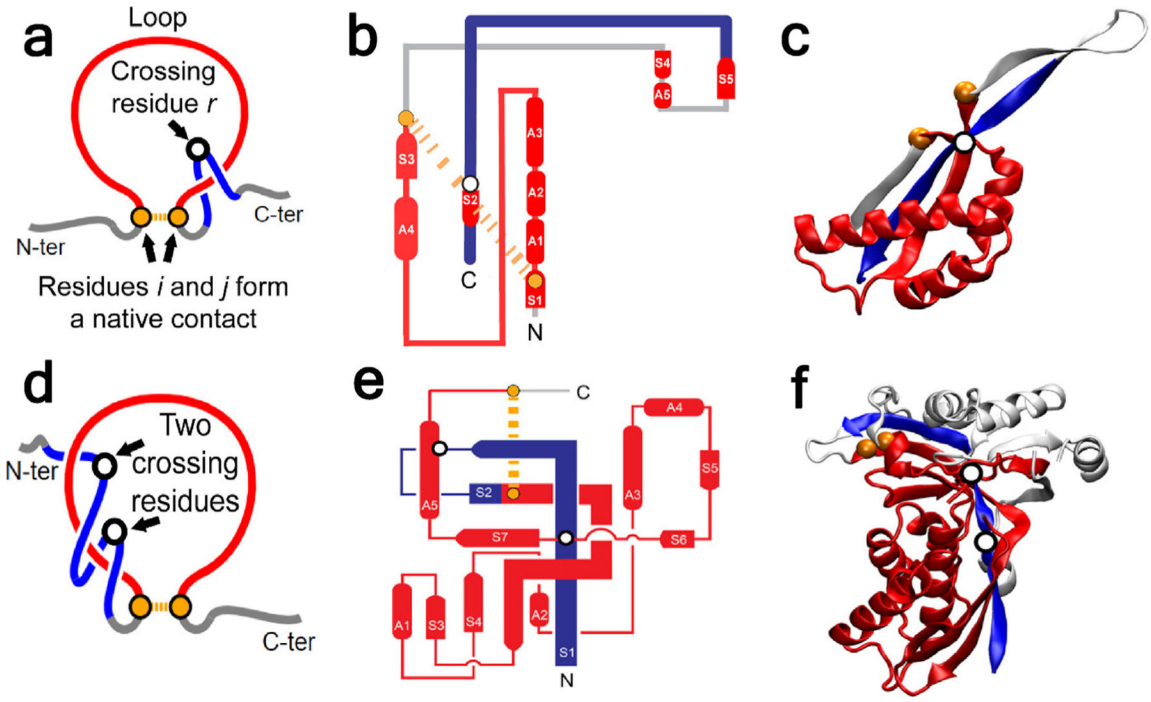

Figure 1. Non-covalent lasso entanglements in globular, folded protein structures.

(a) Illustration of a non-covalent lasso entanglement with a single crossing residue. Loop is in red, native contact closing the loop in orange, treading segment in blue, and crossing residue is shown as a white sphere. Note, the loop is closed by a non-covalent interaction, *not* a disulfide bond. **(b)** The flattened topology representation of the 50S ribosomal protein L22 highlighting the loop (red), thread (blue), closing native contact, and crossing residue. **(c)** Crystal structure of 50S ribosomal protein L22 (PDB ID:6XZ7, chain S) with loop and thread highlighted. **(d)** The topology diagram of a non-covalent lasso entanglement with double crossings observed in 4-Diphosphocytidyl-2-C-Methyl-D-Erythritol Kinase. **(e)** The flattened topology plot for 4-Diphosphocytidyl-2-C-Methyl-D-Erythritol Kinase. **(f)** Crystal structure of 4-Diphosphocytidyl-2-C-Methyl-D-Erythritol Kinase (PDB ID: 2WW4).

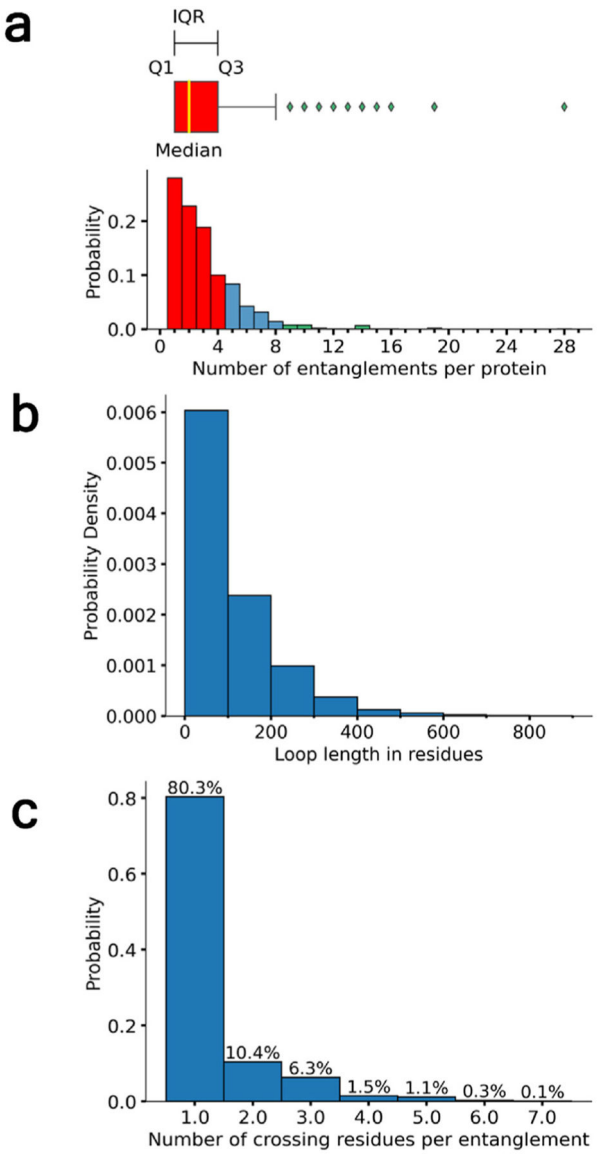

Figure 2. Analysis of proteome-wide Escherichia coli non-covalent lasso entanglements in cytosolic proteins.

a-c Different density plots characterizing the properties of non-covalent lasso entanglements. **(a)** The distribution of the number of entanglements per protein in the *Escherichia coli* proteome using a combined boxplot (top) and probability distribution (bottom). Interquartile range is calculated as Q3 – Q1 in red. Data after Q3 is shown in blue. Outliers, colored in green, are determined using the 1.5 × (IQR) method in green. **(b)** The probability distribution of loop length (number of residues composing the loop) of all non-covalent lasso entanglements. Loop length is defined as the difference in the native contact residues closing the loop, $|j - i|$. **(c)** The probability distribution of the number of crossing residues per entanglement.

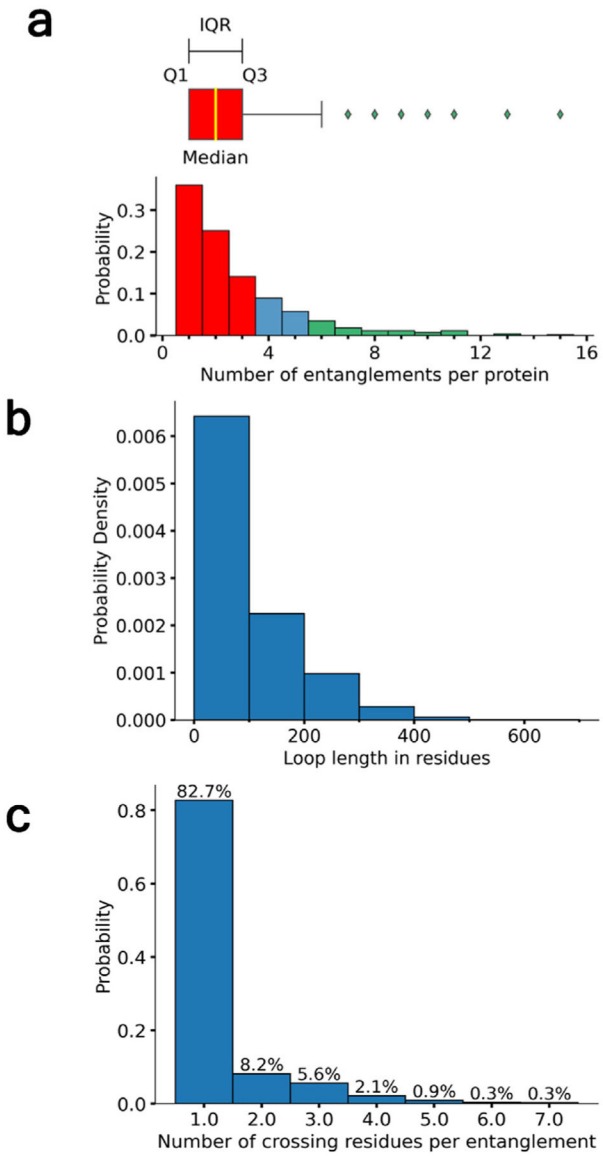

Figure 3. Analysis of proteome-wide Saccharomyces cerevisiae non-covalent lasso entanglements in cytosolic proteins.

a-c Same as Figure 2 but for *S. cerevisiae*.

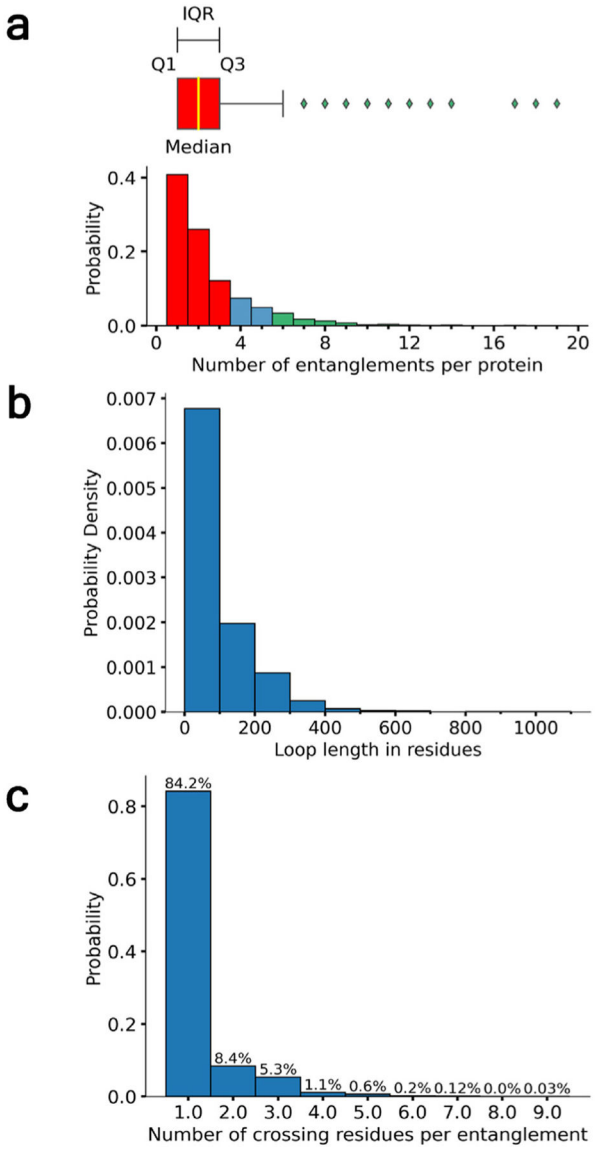

Figure 4. Analysis of proteome-wide Homo sapiens non-covalent lasso entanglements in cytosolic proteins.

a-c Same as Figure 2 but for *H. sapiens*.

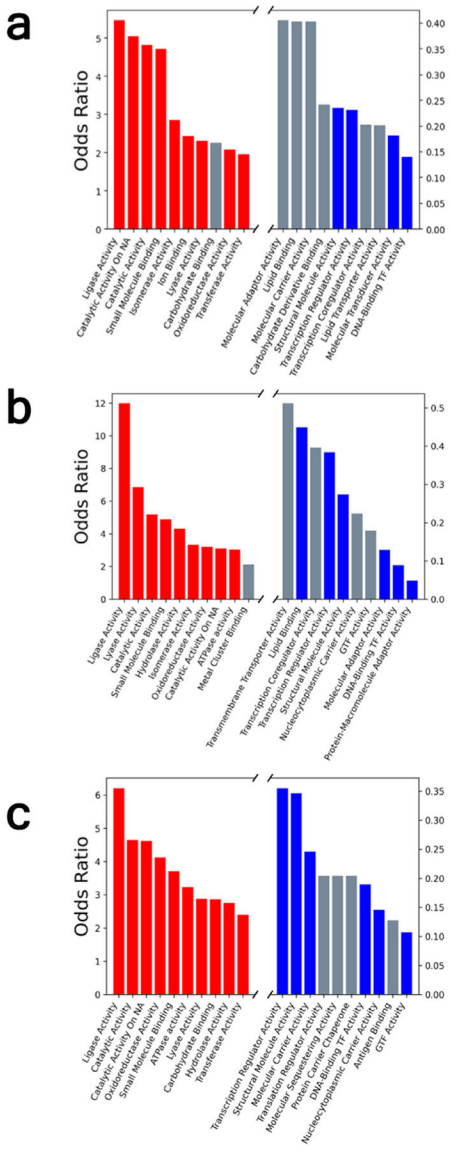

Figure 5. Top and bottom ten molecular function GO annotations ranked by odds ratio.

a-c The Gene Ontology analysis was conducted for (a) *E. coli*, (b) *S. cerevisiae*, and (c) *H. sapiens* level 2 and 3 Molecular Function Annotations. Red indicates enrichment, blue indicates depletion, and gray indicates non-significant annotations after Benjamini-Hochberg false discovery rate correction to the p-values.

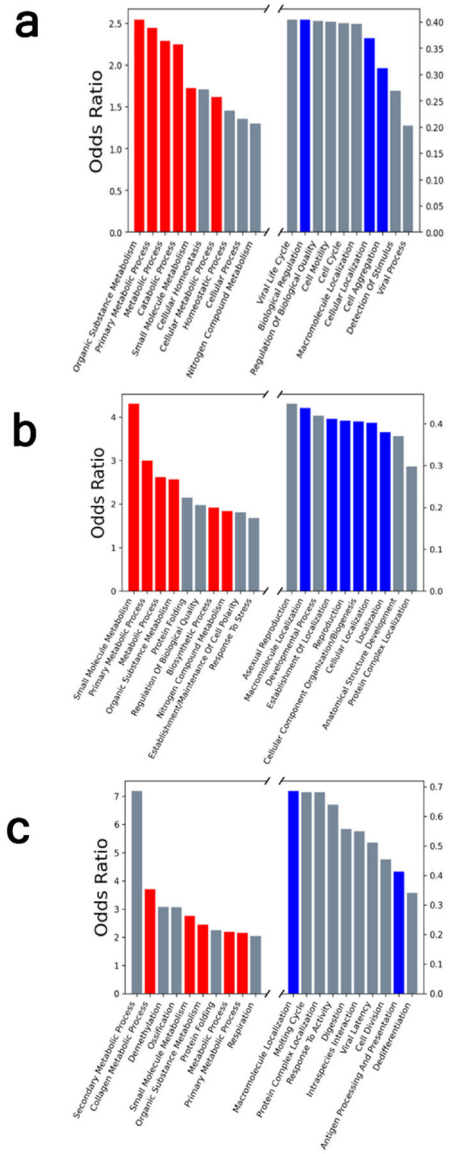

Figure 6. Top and bottom ten biological process GO annotations ranked by odds ratio.

a-c The Gene Ontology analysis was conducted for (a) *E. coli*, (b) *S. cerevisiae*, and (c) *H. sapiens* level 2 and 3 Biological Process Annotations. Red indicates enrichment, blue indicates depletion, gray indicates non-significant annotations after false discovery rate correction.

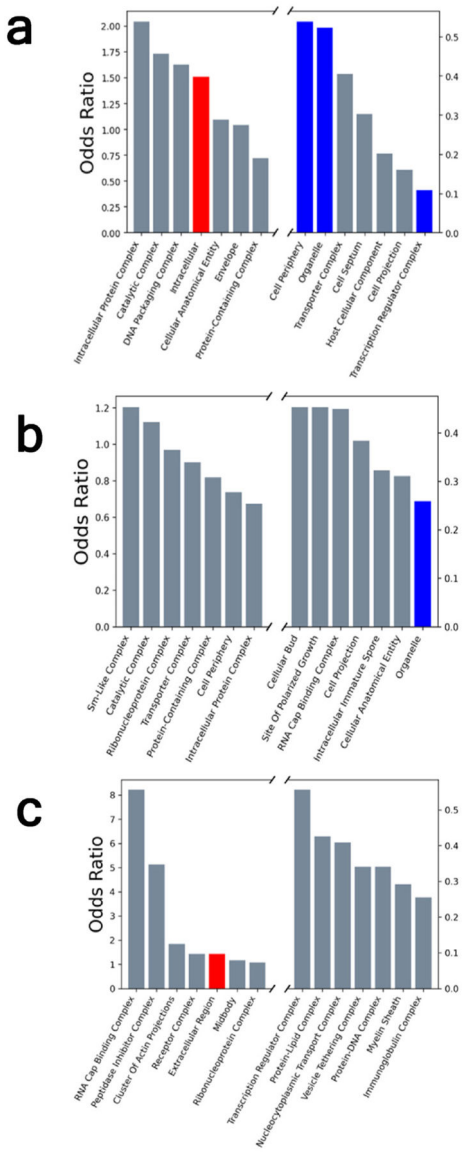

Figure 7. Top and bottom seven cellular component GO annotations ranked by odds ratio.

a-c The Gene Ontology analysis was conducted for (a) *E. coli*, (b) *S. cerevisiae*, and (c) *H. sapiens* level 2 and 3 Cellular Component Annotations. Red indicates enrichment, blue indicates depletion and gray indicates non-significant annotations after false discovery rate correction.

J Mol Biol. Author manuscript; available in PMC 2024 July 23.

Table 1

Spatial association analysis of protein entanglement crossing residues and functional residues.

a-c The analysis was conducted for (a) *E. coli*, (b) *S. cerevisiae*, and (c) *H. sapiens* non-covalent lasso entanglements. Enriched/depleted results were obtained after FDR hypothesis correction (q-values ≤ 0.05). Columns describe if entanglement crossing residues are located closer to functional residues than random chance (enriched), further from functional residues that random chance (depleted), or neither. The first seven rows represent different functional categories. The last row is for all functional categories combined. The genes representing the counts can be found in SI File 7.

Spatial Association Frequency Table for *Escherichia coli* Non-Covalent Lassos				
Functions	**Enrichment**	**Depletion**	**Neither**	**Total**
DNA binding	0	0	30	30
RNA binding	0	0	31	31
Zinc finger region	0	0	4	4
Active site	1	0	285	286
Protein - protein interfaces	3	0	463	466
Metal binding	0	0	161	161
Small molecules	15	0	561	576
All	23	3	779	805
Spatial Association Frequency Table for *Saccharomyces cerevisiae* Non-Covalent Lassos				
Functions	**Enrichment**	**Depletion**	**Neither**	**Total**
DNA binding	0	0	17	17
RNA binding	1	0	32	33
Zinc finger region	0	1	16	17
Active site	2	1	119	122
Protein - protein interfaces	2	1	260	263
Metal binding	0	0	87	87
Small molecules	9	2	260	271
All	10	2	425	437
Spatial Association Frequency Table for *Homo sapiens* Non-Covalent Lassos				
Functions	**Enrichment**	**Depletion**	**Neither**	**Total**
DNA binding	0	0	67	67
RNA binding	6	1	54	61
Zinc finger region	0	1	47	48

Active site	0	0	774	774
Protein - protein interfaces	0	0	1050	1050
Metal binding	0	0	429	429
Small molecules	33	4	1562	1599
All	38	6	2127	2171