Uniwersytet Warszawski
Szkoła Doktorska Nauk Ścisłych i Przyrodniczych

Bartosz Ambroży Greń
Nr albumu: 320358

Topological analysis of biomolecules

Rozprawa doktorska
w dziedzinie nauk ścisłych i przyrodniczych
w zakresie dyscypliny nauk fizycznych

Pod kierownictwem:
dr. hab. Joanny Idy Sułkowskiej prof. UW
Centrum Nowych Technologii
Uniwersytet Warszawski

Warszawa, Wrzesień 2025

Keywords
knot theory, tangles, topologically complex protein, lasso protein, Brunnian topology, RNA knot, structural biology, protein folding

Słowa kluczowe
teoria węzłów, splątki, topologicznie złożone białko, białko lassowe, topologia brunnowska, zawęźlone RNA, biologia strukturalna, zwijanie białek

Tytuł pracy w języku polskim
Analiza topologiczna biomolekuł

Acknowledgements

I am deeply grateful to all the chaperones whose guidance and care assisted me through every step of my folding journey.

Abstract

There is no known life without proteins. Proteins are indispensable for virtually all biologically meaningful processes, functioning as structural components, transporters, containers, receptors, signaling molecules, regulatory factors, and, above all, enzymes. They are the basic building blocks of life, and understanding their properties is essential for advancing the frontiers of medicine. However, proteins themselves cannot exist without RNA—another fundamental biopolymer. Notably, recent advances have made the study of RNA increasingly parallel to that of proteins. Both proteins and RNAs occupy the boundary between the micro- and macro-worlds, attracting the attention of researchers from biology, chemistry, biophysics, bioinformatics, and pharmacy. While most existing drugs act by modulating protein activities, the importance of RNA as a target in medical research is rapidly growing.

A central theme in protein science is the relationship between sequence, structure, and function. Proteins are biopolymers composed of amino acids, assembled by ribosomes (large RNA-protein complexes) within our cells. The resulting chain of amino acids can (ideally) fold into a functional protein in a physics-driven process. While the amino acid sequence determines the structure and dynamics of the folded protein, the resulting structure and dynamics ultimately govern its biological functions.

The significance of structural protein science was underscored by last year's Nobel Prize in Chemistry, awarded "for computational protein design" (David Baker) and "for protein structure prediction" (Demis Hassabis and John Jumper, representing the AlphaFold project). The AI revolution has fundamentally transformed the field: from roughly 100,000 experimentally determined protein structures, we now have accurate structural predictions for all known amino acid sequences—over 200 million in total—allowing unprecedented opportunities for large-scale protein analysis.

Classically, protein structures are described in terms of four levels of organization, but this overlooks a crucial property influencing the entropic freedom of protein chains: topology. For the past 25 years, it has been recognized that some proteins are intrinsically knotted—a property not merely incidental, but one that dramatically affects their stability and dynamics. More recently, a wider variety of topologically complex motifs have been discovered, including knots, slipknots, lassos, θ-curves, and handcuffs—discoveries largely inspired by the application of knot theory to protein science.

In this thesis, I present my contributions to knot theory and the structural biology of topologically complex proteins and other biopolymers. I first describe a solution to a 50-year-old problem posed by John Conway: the classification of algebraic tangles. Algebraic tangles provide the foundation for classifying spatial graphs, themselves a basis for recognizing newly discovered, topologically complex protein folds. Secondly, I present the detailed characterization of lasso proteins and slipknotted transmembrane proteins. Furthermore, I describe the discovery of new knotted folds: a 6_3 protein knot, a Brunnian handcuff in protein, and—most notably—the first naturally occurring RNA molecule with a trefoil (3_1) knot. Additionally, I introduce practical tools for the broader scientific community, including AlphaKnot, AlphaLasso, and the Topoly Python package, which facilitate topological analysis of biopolymers, even for those without advanced training in topology or programming.

Contents

Chapter 1

Introduction

1.1. Knot theory

1.1.1. Some context

Knot theory is a branch of both algebraic topology and geometric topology. The interests of these three disciplines are similar: the development of methods to distinguish between different continuous spaces and the classification of these spaces. However, broadly speaking, the notion of what is "different" or "equivalent" is what distinguishes these disciplines. The application of knot theory in the study of biopolymers, particularly proteins and RNA, is a relatively new approach.

Algebraic topology is centered around homotopy – an equivalence relation that does not distinguish between objects that can be continuously deformed into each other. Geometric topology is more strict and centered around homeomorphisms; it distinguishes objects that cannot be mapped onto each other bicontinuously (by a bijective, continuous function that has a continuous inverse). As an example, a ball, a disk, and a point are homotopy equivalent but are not homeomorphic. The same relation exists between a loop, a Möbius strip, and the side area of a cylinder: all are deformable into one another but cannot be bijectively mapped onto each other.

Knot theory uses an equivalence relation that is more demanding than homotopy and homeomorphism – an ambient isotopy. Ambient isotopy is a homotopy that is homeomorphic at each moment of deformation. More intuitively, if two objects can be deformed into each other without gluing or tearing, they are ambient isotopic. Knot theory blends abstract algebra tools with geometric intuitions about space, bringing it closer to the natural sciences.

According to Józef Przytycki [1], the origin of modern knot theory should be attributed to 19th-century physicists: James Clark Maxwell, Hermann von Helmholtz, Peter Guthrie Tait, and William Thompson (Lord Kelvin) [2, 3, 4]. Maxwell's remarks about knots in his study on electricity and magnetism, and Helmholtz's article on vortex motion, led Lord Kelvin to the idea that different chemical elements are, in fact, differently knotted vortices in the aether. This theory is known as the "Theory of vertex atoms", and it motivated Tait to understand the structure of knots. Tait, during his work, made three conjectures, all of which were solved, the last one in 1991. However, the central problems are still unresolved: how to

algorithmically and efficiently determine if two knot diagrams are the same or different; or more specifically, how to recognize an unknot.

Although originally, knot theory revolved around knots (an S^1 loop embedded in $\mathbb{R}^3$) and links (collections of S^1 loops embedded in $\mathbb{R}^3$), its current interests are broader. Knot theory has been extended in a few directions, e.g.:

- higher dimensions – embeddings of n-dimensional manifolds into $(n+2)$-dimensional space, just like regular knots are embeddings of 1-dimensional manifolds into 3 dimensional space;
- open structures (with vertices of degree 1) – open knots, slipknots, lassos, knotoids;
- branching (addition of vertices with a degree > 2) – spatial graphs and bonded knots;
- collections of open strings anchored in a fixed boundary – tangles, braids.

Here, I will introduce a few topologies that are important for the rest of the thesis to describe the topology of proteins and RNA [Article 1-7]. The majority of the mentioned methods are available in the Topoly Python package [Article 8].

1.1.2. Knots

Knot invariants are maps from a knot to some mathematical objects (e.g. numbers, polynomials). Note that there is no requirement for a knot invariant to have a unique value for each knot; this is a property of a perfect invariant. No computable perfect knot invariant is known.

Many knot invariants do not operate directly on a knot but on a **knot diagram**. A knot diagram is a projection of a knot onto a $\mathbb{R}^2$ surface, preserving information that allows one to clearly restore the original knot. Every point of projection should be a place of crossing of at most two strands and should have a finite number of crossings. The strand that goes underneath the crossings is usually annotated by breaking the line in the diagram.

Reidemeister moves are a complete set of transformations of a knot diagram that preserve a knot type. By complete, I mean that any two knot diagrams that represent the same knot are related by a finite sequence of Reidemeister moves. To show that a mathematical object is a knot invariant, it is sufficient that this object does not change under Reidemeister moves. Using Reidemeister moves, a knot diagram can be transformed into a **minimal** one, in which a diagram has a minimal number of crossings. The minimal number of crossings of a knot is a knot invariant.

Knot polynomials are stronger knot invariants than numerical ones. The simplest and oldest discovered knot polynomial is the Alexander polynomial $\Delta(t)$ [5]. Knot polynomials can be computed recursively using a skein relation for the Alexander polynomial:

$$\Delta(t; \langle \times \rangle) - \Delta(t; \langle \times \rangle) = \left(t^{\frac{1}{2}} - t^{-\frac{1}{2}}\right) \Delta(t; \langle \,) \, (\, \rangle) \,, \tag{1.1}$$

where $\langle \times \rangle$, $\langle \times \rangle$, and $\langle \,) \, (\, \rangle$ are identical knot diagrams, except for a small part of the diagram where they resemble the figures indicated. The polynomial is normalized in such a way that the polynomial of an unknot is equal to 1:

$$\Delta(t; \bigcirc) = 1. \tag{1.2}$$

The Alexander polynomial does not detect chirality (if one swaps all  and crossings, the polynomial remains the same). However, there are stronger polynomials that detect chirality, e.g., the HOMFLY-PT polynomial [6, 7]. In one of its parameterizations $\mathcal{P}(a, z; K)$, its skein relation is defined as:

$$a\mathcal{P}(a, z; \langle \times \rangle) - a^{-1}\mathcal{P}(a, z; \langle \times \rangle) = z\mathcal{P}(a, z; \langle)(\rangle)\,, \tag{1.3}$$

and normalized to an unknot as well:

$$\mathcal{P}(a, z; \bigcirc) = 1. \tag{1.4}$$

For $a = 1, z = t^{\frac{1}{2}} - t^{-\frac{1}{2}}$, the HOMFLY-PT polynomial reduces to the Alexander polynomial.

Since the (naive) calculation of a polynomial of a knot diagram with N crossings requires the generation of 2^N diagrams, its time complexity is exponential $\mathcal{O}\big(2^N\big)$, which complicates their usage. However, an Alexander polynomial can be computed from the determinant of an $(N-1) \times (N-1)$ Alexander matrix. Entries of the Alexander matrix are t, $1-t$, -1, 1, and 0; if t is a number, then its determinant can be computed in $\mathcal{O}(N^3)$ time (using Gaussian elimination). In such a case, the Alexander polynomial is usually computed at $t = 1$ or $t = -1$. We have found that the Alexander polynomial at $t = 12$ is nearly as strong as the Alexander polynomial computed for the variable t in distinguishing knots up to 12 crossings, and it distinguishes all knots up to 7 crossings. Therefore, it is very useful for identifying simple knots found in proteins.

Since the time complexity of knot polynomials is so large, it is a good idea to simplify the chain first. One can perform Reidemeister moves, but the chain can be simplified before a projection as well. It can be done by a method proposed by Kleanthes Koniaris and Murugappan Muthukumar (**KMT algorithm**) [8]; it removes any (non-first and non-last) two neighboring segments $v_{n,n+1}$ and $v_{n+1,n+2}$ with a segment $v_{n,n+2}$, but only if a triangle created by these three segments is not pierced by any part of the chain.

The complete **classification of knots** is not known [9], and it seems that there is no simple rule to describe their complexity. By the classification of knot complements (in S^3 space), we know that each knot is either a torus knot, a hyperbolic knot, or a satellite knot [10]. Torus knots $T_{p,q}$ are fully classified – each one is uniquely specified (up to chirality) by a pair of co-prime integers $p, q \geq 2$, which can be determined, e.g., by the computation of the Alexander polynomial, which is unique in the set of torus knots [11]:

$$\Delta(t; T_{p,q}) = \frac{1}{t^g}\frac{(t^{pq}-1)(t-1)}{(t^p-q)(t^q-1)};\ g = \frac{(p-1)(q-1)}{2}. \tag{1.5}$$

However, torus knots represent only a small subset of all knots. The number of knot types grows superexponentially with an increasing minimal number of crossings of a knot. Knots that can be represented as a connected sum of two non-trivial knots are called composite knots. Non-composite knots are called prime knots and have been fully classified up to 19 crossings [9].

The process of knot classification (up to N crossings) can be divided into two stages: the generation of an initial set of knots and the rejection of duplicates. The first step is

"enormously" easier than the second one [9]. The naive way to perform the first step is to create all non-isomorphic graphs of order up to N, with 4-degree vertices connected by any number of edges; then substitute vertices with crossings ($\times$ and $\times$) in all possible ways. This sequence of operations creates a set of all possible knot diagrams up to N crossings; however, this set is much larger than the set of unique knots. Therefore, any rational strategy that limits the set of pre-generated knots is very desirable.

A more optimal strategy for the classification of knots has been introduced by John Conway [12], which I described in subsection 1.1.6. Specifically for this task, Conway created an algebra of **algebraic tangles** and found a method for the classification of rational tangles (a subset of algebraic tangles). I have extended his work by creating a method for the classification of all algebraic tangles, which is described extensively in [Preprint].

1.1.3. Open knots and slipknots

In a strictly mathematical sense, "we may be unable to define a knot in an open path, but we know one when we see one" [13]. While an "open string knot" could always be unknotted, we still intuitively feel that an open string can be knotted. It is especially justified if some physics are involved, which makes unknotting difficult, e.g., because of the stiffness of a chain or some stabilizing forces. In the case of proteins, which are brush-like polymers due to the side chains of amino acids, the snake-like movements are hindered, and at a specific temperature, it may be very unlikely for a protein to (un)knot, especially if it requires unthreading a loop.

A broadly adopted method of measuring how knotted a chain is involves probabilistic methods. One can extend the ends of an open curve by projecting them in random directions onto a large sphere surrounding the curve and joining them by a geodesic on the large sphere. The random projections are performed many times (100-200); if a majority (plus some margin of error) of closed loops are trivial, consider the chain to be unknotted; otherwise, consider it to be knotted.

As we are discussing open chains, we can take a step further and consider **slipknots** [14, 15]. They are (nontrivially) knotted subchains, but in a different way than the whole chain. An example is a tied shoelace – by pulling apart the ends, the shoelace forms an unknot, but by pulling the loops, the knot will be formed. To analyze slipknots in an open chain, a knot map is useful; an exemplary knot map is presented on the Figure 1.1.

A knot map presents the knot probability of all subchains of a given chain. On Figure 1.1 a knot map of a 492 residue-long protein is presented. Each point below the diagonal represents a specific subchain; to find the termini of a subchain, one needs to project a point onto the axes. Each connected area colored with a specific color represents a specific slipknot (or a knot if it contains the bottom left corner, which represents the whole chain). The higher the color intensity, the greater the probability of a given knot type for a specific subchain. We can characterize a slipknotting pattern of a specific chain by listing the present slipknots; this is called a knot fingerprint [15]; knot map on Figure 1.1 presents a protein with a $3_1 4_1 3_1$ knot fingerprint.

By analyzing a knot map, one can distinguish different chain areas (from the perspective of a chosen slipknot): a knot core and a knot tail; additionally, a slipknot tail and a slipknot

loop in the case of a slipknot. A knot core is the shortest subchain ($\pm\epsilon$) that preserves a specific type of knot. There are two remaining parts on each side of a knot core; they are called knot tails. A knot tail of a slipknot can be broken down further: a slipknot loop is the longest extension of a knot core, which preserves its knot type; a slipknot tail is the remaining part between the slipknot loop and the terminus.

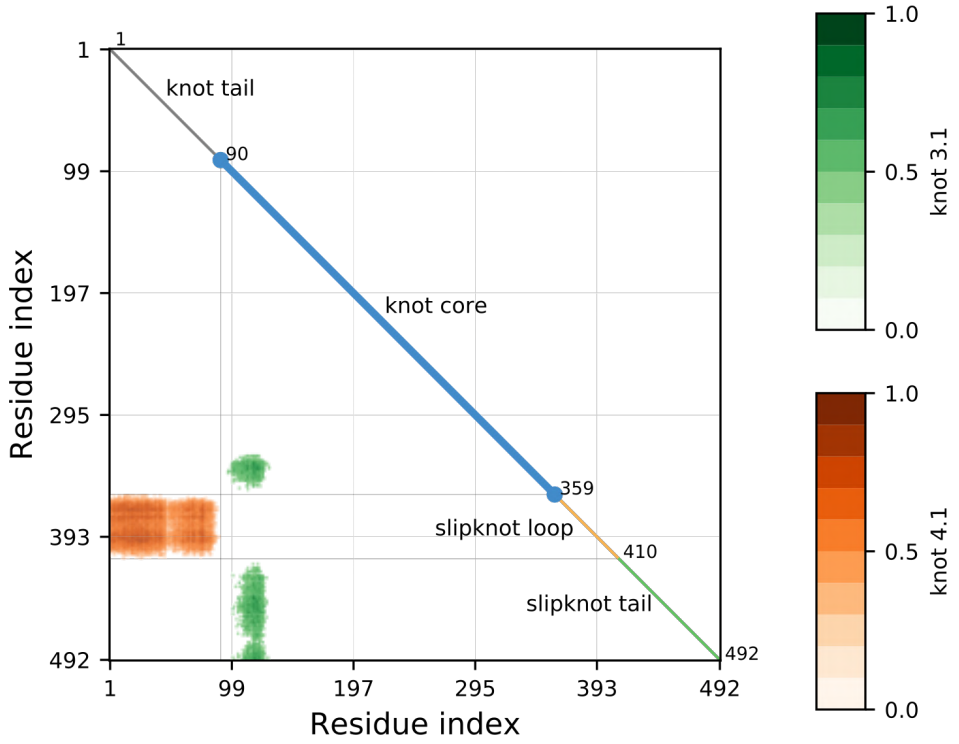


Figure 1.1: Knot map of Glycine betaine transporter (PDB id:4LLH, chain:B) with a knot fingerprint $3_1 4_1 3_1$. Here we can distinguish 3 slipknots: 3_1 (core 130-327, loops 80-130, 327-359), 4_1 (core 80-359, loop 359-410), second 4_1 (core 130-410, loop 80-130). The knot core, (slip)knot tail, and slipknot loop are marked for the 4_1 slipknot.

1.1.4. Lassos

Lasso is a closed loop with a tail connected to it. Complex lasso is a lasso that has a tail piercing the lasso loop [16, 17]. Lasso can have more than one tail. Each tail can pierce a loop more than once, which is a basis for complex lasso classification; moreover, the side from which the piercing occurs may also be taken into account. We assign "+" and "−" signs to opposite sides of the loop. In general, it is arbitrary which side is "+". Lasso type can be fully described by a sequence of "+" and "−" (one sequence per tail). If two signs in a row are the same, it means that the tail wound around a loop and pierced a loop again from the same side; we say that such a lasso is *supercoiled* [17, 16].

A linear polymer can form a lasso topology if it can form an extra intra-chain bond called the bridge. If the polymer is naturally oriented (like protein, RNA, DNA), we can unambiguously assign signs to both sides of a loop using the right-hand rule and multiplying

a bridge vector by a backbone vector. More precisely, for a loop closed by a bridge between beads (/atoms/monomers/residues/nucleotides) i and $j > i$, with coordinates $\vec{v}_i$ and $\vec{v}_j$, the orientation of a surface is defined by the vector $\vec{v}_{i,j} = (\vec{v}_j - \vec{v}_i) \times (\vec{v}_{i+1} - \vec{v}_i)$. In the case of proteins, a bridge is usually a covalent bond (cysteine or isopeptide bond); in the case of RNA or DNA, it is a hydrogen bond.

So far, two methods for lasso detection have been developed; both are implemented in our Python package, Topoly [Article 8]. The first one calculates a globally **minimal surface** spanned by the lasso loop (like a soap film) and identifies where the surface piercings occur [17]. This method has the advantage that the surface can be easily visualized. However, in some cases, this surface may be ill-defined – when a loop has more than one locally minimal surface with a similar area, small changes in chain arrangement may change which surface is globally minimal and alter the lasso type.

The other method is based on the Gauss linking integral [18], which is an extension of the Gauss linking number **(GLN)** and measures how many times two curves Γ_1 and Γ_2 wind around each other:

$$\int_{\Gamma_2}\int_{\Gamma_1} \frac{\gamma_1 - \gamma_2}{|\gamma_1 - \gamma_2|^3}\, d\gamma_1 \times d\gamma_2.$$

In the case of a lasso, curve Γ_1 is circular (lasso loop), and curve Γ_2 is open (lasso tail). The result can be computed for all subchains of the tail Γ_2 and represented as a matrix to study the lasso type; one can investigate how the GLN value changes with different subchain borders or by finding **maxGLN** (the maximal absolute value of the GLN matrix). This method is faster than the minimal surface method and removes any ambiguities, as continuous changes in input provide continuous changes in output; however, the results are harder to interpret.

1.1.5. Spatial graphs; handcuffs and θ-curves

Linear polymers can have additional intra-chain bonds, which alter their topology. A suitable description of their variety is possible using spatial graphs. Spatial graphs are a generalization of planar graphs, enabling all possible embeddings of graphs into $\mathbb{R}^n, n > 2$. If we treat a knot as a circular graph embedded in a $\mathbb{R}^3$ space, then spatial graphs generalize knots by enabling vertices with a degree higher than 2. Here, we are interested only in graphs with vertices of degree 2 and 3. The simplest examples of the generalization of knots into spatial graphs are handcuffs and θ-curves; they have exactly two 3-degree vertices. Topologies with a larger number of 3-degree vertices have not been named or classified.

Hiromasa Moriuchi classified handcuffs and θ-curves up to 7 crossings [19] using a similar approach to that used by Conway for the classification of knots, utilizing tangles (described in subsection 1.1.6). Firstly, Moriuchi needed to classify tangles, as Conway had not provided any classification of tangles. Moriuchi managed to classify tangles of up to 7 crossings [20]. Unfortunately, he skipped 4 tangles; in [Preprint] I show a different way to classify tangles.

I wanted to extend Moriuchi's classification by the addition of handcuffs and θ-curves with a higher number of crossings, and for spatial graphs with more 3-degree vertices. However, first, I needed to extend a classification of tangles, which appeared to be a more complex problem than I initially thought. Fortunately, I extended the theory of tangles and created

a classification of tangles with up to 14 crossings. I wrote more about tangles and their classification in the next subsection and in the [Preprint].

Yamada polynomial

One cannot compute a HOMFLY-PT [6, 7] or Alexander polynomial [5] of spatial graphs; however, one can compute a polynomial proposed by Shûji Yamada in 1989 [21]. Yamada's polynomial is an invariant of flat graphs and of spatial, pliable graphs of degree lower than 4.

Yamada's polynomial $\mathcal{Y}(A)$ is defined recursively by a skein relation:

$$\mathcal{Y}(A; \langle \times \rangle) = A\mathcal{Y}(A; \langle)(\rangle) + A^{-1}\mathcal{Y}(A; \langle \asymp \rangle) + \mathcal{Y}(A; \langle \times\!\!\!\cdot \rangle) \tag{1.6}$$

and contraction-deletion formula:

$$\mathcal{Y}(A; \langle \succ\!\!-\!\!\prec \rangle) = \mathcal{Y}(A; \langle \succ \; \prec \rangle) + \mathcal{Y}(A; \langle \times\!\!\!\cdot \rangle)\,. \tag{1.7}$$

and normalized in a way that a bucket graph B_n (a singular vertex with n loops) is equal:

$$\mathcal{Y}(A; B_n) = (-1)^{n+1}(A + 1 + A^{-1})^n. \tag{1.8}$$

The (naive) calculation of Yamada's polynomial for a diagram with N crossings requires the creation of 3 new diagrams; therefore, its time complexity is even worse than that of the HOMFLY-PT polynomial: it is $\mathcal{O}(3^N)$. Unfortunately, it is the only known polynomial invariant of spatial graphs.

1.1.6. Tangles

Tangle is an embedding of two arcs (and any number of circles) into a ball B^3, with the arcs' endpoints anchored on the boundary of the ball at regular intervals (corners of a square). The basic tangles are:

- $\asymp$ – 0 tangle,
- $)($ – ∞ tangle,
- $\times$ – 1 tangle,
- $\times$ – -1 tangle.

John Conway created an algebra of algebraic tangles [12]. Tangles can be connected by sum operations: $\times + \times = \times\!\times$, to create more complex tangles. Tangles can be rotated (by multiples of $\pi/2$) and reflected (through horizontal, vertical, and diagonal axes) as well; these operations create the $\mathrm{D}_8 \times \mathbb{Z}_2$ group, and each tangle inherits a symmetry group, which is one of its subgroups [Preprint] (D_8 is the dihedral group of order 8). An algebraic tangle is any tangle that:

- is a basic tangle, or
- can be represented as a rotation or reflection of an algebraic tangle, or
- can be represented as the sum of two algebraic tangles.

A rational tangle is any tangle that:

- is a basic tangle, or
- can be represented as a rotation or reflection of a rational tangle, or

- can be represented as the sum of a rational tangle and 1 tangle.

All rational tangles belong to the set of algebraic tangles. Rational tangles are very unique mathematical objects in knot theory, since they possess a perfect invariant – a fraction $\frac{p}{q} \in \mathbb{Q} \cup \left\{\frac{1}{0}\right\}$; therefore, they are fully classified [12, 22].

A more precise description of algebraic and rational tangles can be found in my [Preprint], where I also present a method for the classification of algebraic tangles. I defined a unique, canonical representation of algebraic tangles, proposed an algorithm that transforms any algebraic tangle into its canonical representation, proved that it works, and classified tangles up to 14 crossings.

Tangles provide a more optimal method of classification of knots and spatial graphs than the method mentioned in subsection 1.1.2. Conway's method inserts algebraic tangles instead of crossings in place of graph vertices. Firstly, we can reject all graphs with bigons, since a bigon in which both nodes are substituted with tangles is equivalent to the sum of two tangles, which is a tangle as well. Secondly, algebraic tangles can be tabulated before replacing vertices with them. This method greatly decreases the number of generated knot / spatial graph diagrams before the next stage, which aims to reject all non-unique knots / spatial graphs.

One could ask, "why is the usage of algebraic tangles so important, if knots were already classified up to 19 crossings?" [9]. The main reason is that hyperbolic knots are easier to classify since their complements have finite volume and can be *canonically triangulated*. Canonical triangulation of the complement of a knot is its perfect invariant. At the same time, a majority of knots are hyperbolic – up to 19 crossings: 352 151 858 knots are hyperbolic, and only 394 are not (380 are satellite, and 14 are torus).

1.2. Topologically complex biopolymers

1.2.1. Knotted DNA

Leroy Liu has been a pioneer in the research of knotted biomolecules. In 1976, he discovered the first knotted DNA in long, circular chains of bacterial DNA of Escherichia coli [23]. Later, he studied the consequences of DNA knotting; the process of DNA recombination and how DNA topoisomerases relieve the tension during this process [24]. Topoisomerases are proteins that change the knot type of DNA by breaking and reconnecting DNA strands, e.g., during replication; topoisomerase inhibitors are effective apoptosis inducers and have been used as anti-tumor drugs [25, 26]. So far, the topic of knotted DNA and recombination has been deeply explored and understood [27], partially due to the successful application of the theory of tangles [28, 29].

1.2.2. Knotted proteins – historical context

The idea of research on topologically complex proteins can be attributed to Marc Mansfield. In 1994, he posted a letter to Nature's editor, *Are there knots in proteins?* [13], which began with the statement:

Sir — most biochemists would probably agree that proteins in the native state are not knotted. The protein folding mechanism is not perceived as including reptative, snake-like motions of the chain along its own contour. However, reptation would be necessary for the protein to knot itself. Most biochemists would also agree that the native state is the global energy minimum of the protein. This seems to be the most popular explanation for the ability of the protein to find the native state from among the large number of available conformations. These two perceptions may be inconsistent.

Mansfield had known from statistical knot theory that the probability of an ideal polymer being unknotted falls exponentially with increasing length; he also knew from everyday observations that strings and ropes easily knot themselves. Therefore, he had expected that at least some proteins would be knotted; otherwise, their dynamic would be non-ergodic.

He analyzed the structure of ~ 400 proteins to check whether they are knotted. To identify knots, he used the Alexander polynomial (computed at $t = 1$). To close each protein chain, he used a probabilistic method described in subsection 1.1.3. He found 3 proteins with a knotting probability $> 50\%$; however, after a visual inspection, he rejected them as artifacts of the projection. He concluded that protein folding mechanisms somehow avoid knotted conformations.

Since 1994, the number of available, experimentally solved protein structures has changed vastly: from over 1000 to over 100 000 [30]. Now we know that $\sim 1\%$ of (experimentally solved) proteins contain a knot [31] (however this number has not been updated for 6 years), and we would say that 2 of 3 protein structures rejected by Mansfield were indeed knotted – but shallowly. The first deeply knotted protein was discovered by Taylor [32, 33]. Moreover, with an increase in the availability of protein structures, new complex topology types have been discovered in proteins: slipknots, links, lassos, Θ-curves, and handcuffs.

Joanna Sułkowska participated (either as a first author during her postdoctoral research or as a lab leader) in many discoveries related to complex topologies in proteins:

- creation of a database of knotted proteins [34];
- first slipknots in proteins [15];
- first links in proteins [34, 35];
- first articles on lasso topology [17, 16] ;
- first Θ-curves in proteins [36];

which is why I decided to choose her as my supervisor. Together, we have published articles about (slip)knots [37, 38, 39] and lassos in proteins [40, 41, 42] and discovered the first Brunnian topology in proteins [unpublished]. Moreover, we have discovered the first knotted RNA [43]. I was also a developer of the Topoly Python package [44], which collects computational knot theory methods used in the analysis of biopolymers.

AI revolution

It is impossible to omit the fact that, during my PhD, a whole field of protein science changed drastically. In 2024, the AlphaFold group (represented by Demis Hassabis and John Jumper) was awarded the Nobel Prize in Chemistry "for protein structure prediction" using AI methods (note that the Prize in Physics was related to AI as well). The problem of protein

structure prediction from amino acid sequence has been a central problem of the discipline for decades. Every two years since 1994, the CASP competition (Critical Assessment of Structure Prediction) has been organized to stimulate research in this area. During the CASP14 competition (in 2020), the AlphaFold group presented an AI method that achieved experimental-like accuracy in the prediction of protein structures [45], for which they were awarded the aforementioned Nobel Prize.

Today, the AlphaFold database contains over 200 million structure predictions [46], based on nearly all known protein sequences [47]. Moreover, AlphaFold is not the only available method with experimental-like accuracy; RoseTTAFold [48] and ESMFold [49] are publicly available as well. Unfortunately, our results show that AI methods are not very effective at predicting the structures of RNA molecules [50].

The AI revolution has opened up new possibilities for understanding the topological complexity of proteins. It may help to answer questions about which topology types can be found in proteins, how common they are, and whether they are conserved among families. This motivated us to create AlphaKnot and AlphaLasso servers and databases [37, 42]. They enable other scientists to perform topological analysis with our tools using easy-to-use server. Moreover, databases parts let us test our tools and make first conclusions about knots and lassos in predicted protein structures. Most interestingly, knotted lassos have been found [51].

1.2.3. Diversity of knots in proteins

Before the explosion of predicted protein structures [46, 48, 49], four different knot types (not counting 0_1, the unknot) have been observed in proteins: 3_1, 4_1, 5_2, and 6_1. Interestingly, all of them are twist knots – they can be easily obtained by linking together the opposite parts of a twisted loop. It was even proposed that this property may be crucial for the folding of knotted proteins [52].

It appears that structures predicted by AI methods can have hundreds of different knot types [Article 4]. Probably many of them are artifacts, but some are not; e.g., we precisely studied the human proteome and found a protein with 6_3 knot [Article 5], which has been successfully crystallized and whose knot type has been confirmed [X-ray structure deposited by the Sulkowska group]. Recently, other new knotted folds have also been found with the help of AI: 7_1 knot [53], $3_1\#3_1$ composite knot [54], and 8_3 knot [X-ray structure deposited by the Sulkowska group]. All of these newly found knots are not twist knots.

1.2.4. Structure and function of (slip)knotted proteins

Surprisingly, proteins can fold into a knotted conformation without the help of chaperones; however, they may speed up the process [55, 56, 57]. Experiments show that proteins that are knotted in their native state fold much slower (10–20 times); formation of a knotted conformation is a rate-limiting step in protein folding. It goes both ways – unfolding is slower as well. The transition between knotted and unknotted states requires overcoming a "topological energy barrier," which provides increased kinetic stability. Computational studies show that knots reduce the amount of available conformational space, which stabilizes

proteins by decreasing entropy. It provides higher resistance to mechanical, thermal, and chemical denaturation [58, 59, 60]. Resistance to enzymatic factors is discussed [61, 62].

Studies show that a great majority of homologs of knotted proteins are knotted as well [61, 15, 63]. Moreover, over 80% of knotted proteins are enzymes (with the active sites inside the knot core) compared to 64% of regular proteins deposited in the Protein Data Bank [64]. Decreased conformational freedom may be beneficial for precise protein activity, e.g., at the active site of enzymes or for signal transduction. Moreover, slipknots (being more flexible than knots) may aid particular mechanical functions. Notable examples are:

- UCH-L3 (ubiquitin carboxyl-terminal hydrolase) is an enzyme with a 5_2 knot that cleaves ubiquitin from proteins tagged for degradation with ubiquitin. During its activity, the enzyme is covalently bonded with ubiquitin, which makes it susceptible to degradation. It is hypothesized that this enzyme avoids degradation due to the presence of a knot, which acts like a plug, protecting the enzyme from being pulled into the proteasome itself [61, 62].
- SAM-MTase (S-adenosylmethionine-dependent methyltransferase) is a protein superfamily that can be divided into 5 classes with distinct folds [65]. One of them is the biggest known group of knotted proteins, known as "α/β knot". Despite the low sequence similarity of members of this family (below 20%), a 3_1 knot is very well conserved [66, 67]. One of its members, TrmD, is critical for bacterial metabolism; it has been shown that its knot performs an important role in restricting the conformation of the bounded SAM molecule and energy transmission from SAM to the active site where the methyl transfer occurs [68, 69]. It appears that Trm5, the eukaryotic version of TrmD, catalyzes the same chemical reaction, but their structures are fundamentally different (e.g., Trm5 is unknotted). Therefore, TrmD is an attractive target for rational drug design and the creation of a selective antibiotic with minimal side effects [70].
- A lot of mammalian secondary active transporters have a (slip)knot (compared to regular proteins: 25% vs 1%). Slipknots are highly conserved among 16 families of secondary active transporters. It appears that slipknot topology may be important for their function; alternating between inward-open and outward-open conformations correlates with changes in slipknot loop conformation, and hinge points coincide with the border between the knot core and the lasso loop. Moreover, the change between the inward- and outward-open states can be monitored in proteins from the Sodium Dicarboxylate Family by tracking the slipknot fingerprint. These results are part of the [Article 3]: *Conservation of knotted and slipknotted topology in transmembrane transporters.*

1.2.5. First knotted RNA molecule

For a long time, no knotted RNA molecule has been found in nature. It was surprising, since knots are present in proteins and in DNA. Moreover, a RNA molecule which spontaneously folds into a knotted native state has been artificially synthesized [71], therefore knotted RNA molecules are certainly possible. Furthermore, a human enzyme with RNA topoisomerase activity (an enzyme which cuts RNA strand and glues it back together, but with changed knot type) has been discovered [72] – what would be the purpose of such activity if there were

no knotted RNA molecules [73]? It has been theorized that maybe topological simplicity of RNA is favored for translocation through the nuclear pores [74].

In [Article 6] *Discovery of a trefoil knot in the RydC RNA: Challenging previous notions of RNA topology*, we announced the discovery of a knotted RNA molecule. Although such RNA molecules have finally been found, the question remains, but slightly altered: why is there only one known knotted RNA molecule?

The 15th edition of CASP in 2022 was the first one to evaluate predictions of RNA structures in addition to protein molecules. Following the insight from the AlphaKnot database [Article 4] that many predicted molecules are artificially knotted, we performed an independent analysis of predictions of RNA molecule structures. We found that AI-based methods predicted knots much more often than classical methods (and all target molecules were unknotted) [Article 7]. This shows that AI-methods should be improved to take topology into account to a greater extent.

1.2.6. Diversity of lassos in proteins

Lasso proteins are the most common topologically nontrivial motif found in protein structures [17]. In the set of proteins predicted by AlphaFold $\sim 16\%$, the lasso proteins are complex [Article 2]; all of them can be found in the LassoProt database [16]. Every lasso protein consists of at least one intra-chain bond; commonly, it is a disulfide bond, but others can also be found, such as amide (iso-peptide), ester, and thioester. An intra-chain bond creates a loop, with (maximally) two tails. These tails are called the N-tail and C-tail according to the side of the lasso loop on which they are positioned.

Lasso type can be fully described by two sequences of "+" and "−", each sequence corresponding to one tail, e.g., `N:[-,+,-];C:[+,+,+]`. Usually, we describe a lasso type using one symbol; in this case, $LL_{3-+-,3+++}$, but the notation can be simplified for simple lassos. The details of lasso classification are explained in the original method articles [17, 16], and in my [Article 1]. In [Article 1], we also presented statistics about 11 known lasso motifs. AI predictions show that the universe of lasso types may be much richer, predicting nearly 300 types of new lasso motifs; most interestingly, knotted lassos have been found [51].

1.2.7. Structure and function of lasso proteins

The antibacterial properties of the 21 amino acid Microcin j25 have been known for at least 25 years [75, 76, 77]. In 2003, it was discovered that Microcin j25 is lasso shaped, contrary to what was believed before: that it is circular [78, 79]. It appears that such lasso peptides are vulnerable to temperature and pH, making them candidates for future micromolecular machines [80, 81, 82]. Moreover, these features make them possible tools for the creation of fusion proteins [83] and a new generation of drugs [84]. Not long ago, it was shown that the lasso structure is, in fact, responsible for the antimicrobial activity of Microcin J25 [85]. In the meantime, a similar lasso topology has been discovered in a much bigger molecule: leptin [86]. It has been shown that lasso topology affects the dynamics of this hormone, which influences signal transduction to the leptin receptor, interestingly having no effect on binding

affinity [87, 88]. Moreover, the four-helix bundle fold of leptin is supported by unusually weak interactions. does the lasso topology compensate for this weakness?

To understand the role of lasso motif structures in proteins, we attempted several approaches: searching for correlations, characterizing their geometry, associating lasso types with their functions, and analyzing the conservation of lasso among protein families.

We found that lasso piercings closest to the bridge (distance measured along the backbone) tend to be negatively oriented. It may be attributed to the chirality of the chain, since this preference diminishes with increasing distance from the bridge [Article 1]. No preference has been observed between loops pierced by the N-tail and those pierced by the C-tail.

In my pre-PhD article [40], we performed a statistical analysis of ideal lasso chains (pure equilateral polygons, no physical interactions, no volume). We have found the expression for the probability of a lasso to be complex, as a function of tail and loop length. We compared this expression with statistical results in real proteins and found that a surprisingly large number of lassos with short tails form a complex lasso. It indicates that enthalpic interactions in these proteins favor the formation of complex lasso, which suggests that complex lasso may be crucial for their activity. Therefore, we performed an analysis of complex lasso proteins, with a special interest in families of short-tailed proteins [Article 1]. It appeared that antimicrobial proteins (lasso peptides and proteins with a so-called "cysteine knot") have the same lasso type (L_1; they are pierced once), which may be crucial in their antimicrobial activity as molecular plugs [85, 89]. We have found a greater stabilization effect near the lasso piercing in these small antimicrobial proteins, which supports the hypothesis regarding a functional role of their lassos. Both groups of antimicrobial proteins have proven therapeutic significance [90, 91, 92]. Moreover, a new topic of non-covalent lassos has recently become popular [93], which emerged from studies on lasso proteins.

1.2.8. Brunnian handcuff in proteins (unpublished)

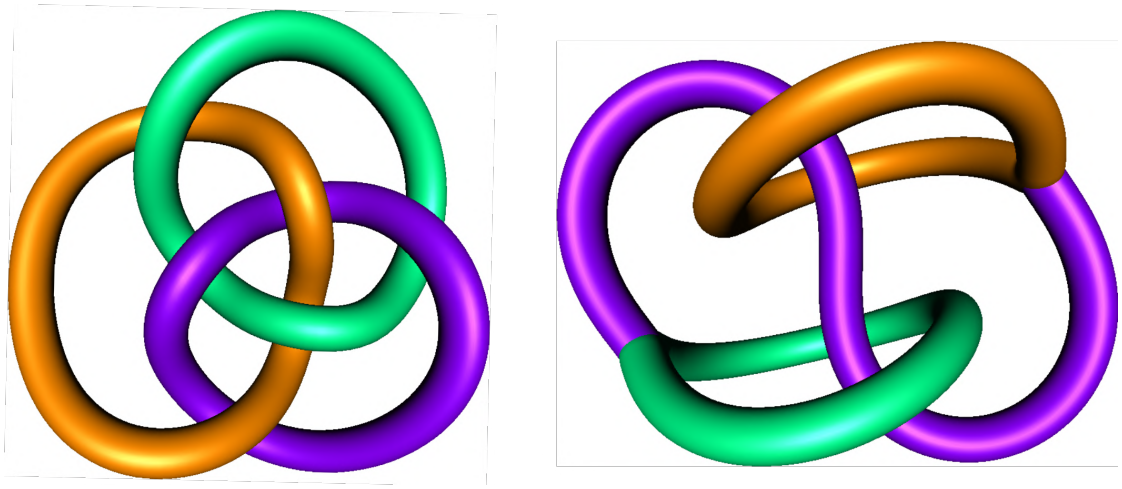


Figure 1.2: Borromean rings (left) and $h2_1\#_3h2_1$ handcuff (right) – examples of the simplest Brunnian topologies. Figure prepared using KnotPlot [94].

Inspired by research on θ-curves in proteins [36], I decided to analyze proteins from the Protein Data Bank with respect to the presence of handcuffs. Since handcuffs are two loops joined by a linker, many of them have been studied before as linked loops [34, 35]. Therefore, I was interested in those handcuffs, which are topologically complex but have unlinked loops; the simplest example is the $h2_1\#_3h2_1$ handcuff. Such topology types are called "Brunnian". The most recognizable example of Brunnian topology is the Borromean rings – three linked loops, where no two loops are linked. After the removal of any loop from the Borromean rings (or the linker from $h2_1\#_3h2_1$), the remaining parts become unentangled.

Using the Yamada polynomial, I have found one family with a conserved $h2_1\#_3h2_1$ Brunnian handcuff: ENPP (Ecto-nucleotide pyrophosphatase/phosphodiesterase) proteins. ENPP proteins are mostly transmembrane, with crucial functional domains on the extracellular side. The exception is ENPP2 (also known as Autotaxin), which is a secreted protein, and ENPP6, which is a GPI-anchored protein. ENPP proteins are involved in key biological processes, including nucleotide and phospholipid signaling. According to the review [95], they can be divided into two groups with different domain architectures:

- ENPP1-3, composed of 4 domains in sequence:
  - two SMB domains (Somatomedin B, PF01033),
  - PDE domain (Type I phosphodiesterase / nucleotide pyrophosphatase, PF01663),
  - and NUC domain (DNA/RNA non-specific endonuclease, PF01223) encircled by a connector which joins PDE and NUC domains together.
- ENPP4-7, with only the PDE domain.

Interestingly, the authors of the review [95] do not mention secreted snake venom phosphodiesterases, which have the same domain architecture as ENPP1-3 proteins.

Brunnian handcuff $h2_1\#_3h2_1$ is present in all ENPP1-3 (and snake venom phosphodiesterase) proteins deposited in PDB. If this topology performs an important role in protein function or stability, we could expect a high conservation of the cysteines that constitute the bridges of handcuff loops. To check this, I performed a sequence alignment of all 934 sequences from the Pfam database that have the same domain architecture as ENPP1-3. Over $\sim 90\%$ of protein sequences conserved all four cysteines responsible for the covalent closure of the loops of the Brunnian handcuff. Moreover, structures predicted (by AlphaFold) from these sequences show $\sim 70\%$ conservation of the Brunnian handcuff. It cannot be rejected that the Brunnian handcuff could play a significant role in ENPP1-3-like proteins.

A majority of the first handcuff loop is located mostly inside the N-terminal side of the PDE domain. The second loop is closed by the bridge that joins the C-terminal side of the PDE domain with the C-terminal side of the NUC. While many amino acids from the PDE domain interact at the interface between the PDE and NUC, only the C-terminal region of the NUC interacts [96]. The area between these domains varies among ENPP1-3; however, the domains are tightly attached in all of them, and the disulfide bridge is highly conserved [96, 97]. It has been shown that PDE and NUC domains strongly interact, which contributes to the stability and enzymatic activity of the PDE domain [98]. Since it is impossible to spread apart loops of $h2_1\#_3h2_1$ handcuff further than $\frac{1}{3}$ linker length (Figure 1.2), could the handcuff be responsible for the entropic stabilization of the PDE and NUC domains at their

interface?

Chapter 2

Summary

2.1. The aim of the thesis

The aim of my PhD is twofold. The first goal is oriented toward biophysical research. I attempt to deepen our knowledge and understanding of different topologically complex motifs in proteins and RNA. Motifs that are highly conserved among families may be interesting for further investigation into the roles they play in the function and dynamics of biopolymers. On the other hand, topological differences between homologous biopolymers may be exploited for drug design. The second goal is method oriented. Secondly, I want to develop mathematical and computational methods and tools in knot theory. It aims to find new, interesting topological motifs for future analysis; as well as to provide other scientists with practical tools that allow them to perform topological analyzes independently.

Firstly, I present a [Preprint] *Classification of algebraic tangles.* Although it is still under review, it is the article of which I am most proud. In the article, I present a theory of algebraic tangles – introduced by John Conway [12] over 50 years ago as a method to study the most elementary entanglements. Conway created this algebraic theory to establish a method for the classification of knots. He found a way to classify rational tangles – a subset of algebraic tangles. I extended his theory and created a method for the classification of all (prime) algebraic tangles, and I tabulated them up to 14 crossings. This way, I extended the old classification from 201 tangles [20] to over 2 million. My motivation was similar to Conway's – to classify spatial graphs, which are a generalization of knots. Future classification of spatial graphs will enable the study of more complex topological folds in proteins.

Secondly, I present the studies on lasso proteins – the most common topologically complex motif in proteins [17]. In [Article 1] *Lasso proteins – unifying cysteine knots and miniproteins* , I performed a statistical characterization of the lasso motif in experimentally solved protein structures. In [Article 2] *AlphaLasso – a web server to identify loop and lasso motifs in the 3D structure of biopolymers* , I present a database of all high-quality AlphaFold-predicted protein structures with lasso motifs. The database extends the number of protein lasso motifs from 11 present to 300 potentially present. Moreover, we found there a first knotted lasso. AlphaLasso is a server that enables non-programming scientists to conduct lasso research with their own AI-predicted structures.

Thirdly, I focus on knots and slipknots in proteins. In [Article 3] *Conservation of knotted and slipknotted topology in transmembrane transporters* , we present results on the study of membrane proteins, where slipknotted topological motifs are the most prevalent. In [Article 4] *AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods*, we present a tool similar to AlphaLasso [Article 2] but dedicated to knotted motifs. Some knotted motifs present in AlphaKnot were already discovered in recently crystallized protein structures, e.g., 6_3 knot, which was mentioned for the first time in our screening of the predicted human proteome in [Article 5] *AlphaFold predicts novel human proteins with knots.*

Fourthly, I present two articles related to RNA topology. Contrary to other scientists [74, 73] who previously declared that there are no knots in RNA, we have found the first example of a knotted RNA molecule, which we present in [Article 6] *Discovery of a trefoil knot in the RydC RNA: Challenging previous notions of RNA topology.* Still, other topologically complex motifs are not known to be present in RNA molecules. In [Article 7] *Knotted artifacts in predicted 3D RNA structures*, we show that AI methods for RNA structure prediction are many times worse than classical energy-based methods at predicting the correct topology.

Finally, nearly all the mentioned research would not be possible without appropriately implemented knot theory tools. We shared our implementations of computational methods in the form of "Topoly" Python library. I am one of the developers of this library. In [Article 8] *Topoly: Python package to analyze topology of polymers*, we advertised the release of the first version of our package.

2.2. Conclusions

In this dissertation, I present a comprehensive topological analysis of biomolecules and provide new tools and methods. Most importantly, I extended John Conway's theory of tangles, which has consequences for both knot theory and structural biology [Preprint]. Extended classification of tangles provides an opportunity to expand the classification of spatial graphs, which can be further studied as potential protein folds.

I co-created tools for the topological analysis of biopolymers: Topoly, AlphaKnot, and AlphaLasso [37, 42, 44]. Topoly enables researchers to study different topological folds using a user-friendly Python library, while AlphaKnot and AlphaLasso do not require any programming to check for the presence of knots and lassos in specific molecules. AlphaKnot and AlphaLasso are provided with databases containing precomputed results for all high-quality AlphaFold-predicted structures, as well as servers adapted for the analysis of structures predicted by AlphaFold and other AI methods. These tools give scientists—who may not be experts in topology, such as chemists and biologists—the opportunity to understand the complexity of topological folds.

Presented research pushes the limits of the topological complexity of biopolymers by discovering new folds: 6_3 the non-twist knot in the human *Von Willebrand factor A domain-containing protein 5A* [39], the first knotted RNA molecule *RydC RNA* with a trefoil knot [43], and the ENPP family of proteins with Brunnian handcuff topology [unpublished]. The studies on slipknotted membrane transporters and the characterization of lasso proteins show

possible functional advantages of topologically complex folds.

The results I obtained show that topological analysis adds new dimensions to biomolecular research and enables the detection of potentially functional motifs overlooked by conventional sequence- or structure-related methods. Moreover, my results enable the research of even more topologically complex folds described by spatial graphs. All of these results show that it may be advantageous to introduce a new way of protein classification that goes beyond the standard primary, secondary, tertiary, and quaternary levels of organization and to include global topological features as well.

Chapter 3

Brief description of published articles

Preprint: *Classification of algebraic tangles*

> **Greń, B. A.**, Sulkowska, J. I., & Gabrovŝek B. (2025). Classification of algebraic tangles. Under review & arXiv.

The goal of this article was to find an invariant of algebraic tangles and extend the classification of algebraic tangles. The motivation for that was to extend the classification of handcuffs and θ-curves, and the creation of first classification of more complicated spatial graphs, which later could be used as a basis to classify new topological motifs in proteins.

In the article, firstly I provide an introduction to tangle algebra, establish the notation, and represent isotopy preserving moves using tangle algebra: elementary moves I and II, the twist, the flype, and the ring move.

Secondly, I extend the notation using binary expression trees with integral and rational tangles represented as tree leaves, and show the advantages of the notation. Then I propose an algebraic tangle invariant in the form of the canonical tree, propose an algorithm, which transforms any tree to the canonical one and prove it. The algorithm traverses the tree and performs isotopy preserving moves whenever possible. Isotopy preserving moves on a tree are suitable combinations of algebraic identities and the aforementioned moves (elementary moves, twist, flype and ring move), which utilize tree notation advantages.

Thirdly, I explain the process of classification of tangles up to 14 crossings, and list missing tangles with 7 crossings from the previous classification [20]. Missing tangles were mutants of other tangles, present in the classification, which shows an advantage of the canonical tree method.

Finally, I study symmetry groups of algebraic tangles as subgroups of $D_8 \times \mathbb{Z}_2$ group. I tabulate the results with respect to the group symmetries and prove why absent groups are impossible. We also provide a database of tangles `https://tangleinfo.cent.uw.edu.pl`.

I am responsible for the majority of the article. Many insightful discussions with Gabrovŝek B., a mathematical expert in the field, helped me to explore different approaches in tackling the problem, improving the theorems and proofs in the article. Moreover, he wrote a program for drawing tangles.

Article 1: *Lasso proteins – unifying cysteine knots and miniproteins*

> **Greń, B. A.**, Dabrowski-Tumanski, P., Niemyska, W., & Sulkowska, J. I. (2021). Lasso proteins—unifying cysteine knots and miniproteins. **Polymers**, 13(22), 3988.

The goal in this article, was to investigate relation between proteins with specific lasso types, and function they perform. Is there a functional role of specific lasso topology types? Are there common features of miniproteins, cysteine knots and lassos? We started with an observation about similar properties of lasso peptides and proteins with a so called "cysteine knot" (which is not a knot from a knot theory's point of view). From a structural point of view, both have at least one covalent loop pierced by a part of the chain, therefore they can be treated as a subset of complex lasso proteins. From a functional point of view, they are antimicrobial, which may be explained by their "molecular plug" structure. But these antimicrobial proteins are only a small subset of all complex lasso proteins, which we studied in the article. PD-T and I performed (independently) whole computational and bioinformatical analysis. Here I describe results which I perceive as the most important.

We have studied all structures present in Protein Data Bank with a non-backbone intra-chain covalent bond (LINK line in the PDB file); each intra-chain bond was treated as a closure of a lasso loops (a bridge). For each lasso we calculated a minimal surface and performed a statistical analysis. To remove redundancy, we performed a sequential alignment for each of 734 homology clusters with $\geq 40\%$ sequential homology; we reduced structures with $\geq 95\%$ sequential homology into one representative structure. Sequential alignment was done using a ClustalΩ server.

We have found that 55% homology clusters (405/734) contain only one representative structure, which means that a majority of complex lasso proteins are unique (singular in their homology cluster). It happens (according to intuition) that more complicated lassos are less common. Moreover, there is no preference for N-tailed or C-tailed complex lassos; however, there is a preference for the side from which the loop is pierced. It appears that piercings closest (along the sequence) to the bridge tend to be negative; it may be attributed to chain chirality. The effect diminishes with the sequential distance from the bridge, which we should expect if the chirality of the chain is the reason.

How often may a lasso topology be functional? To answer this question, we tried two approaches. The first approach was to check if lasso topology is conserved among protein families. Since function generally is conserved among families, no conservation would mean no correlation with lasso topology type; and lassos topology cannot be conserved, if cysteine resides are not conserved. Therefore, we performed an analysis of the conservation of cysteine residues crucial for the formation of a lasso bridge. We treated cysteine residues as conserved between two sequences if their positions were shifted maximally by one after the alignment. Interestingly, crucial cysteines of non-singular homology clusters (45% of clusters, 329/734) were highly conserved ($\geq 80\%$ conservation) in only 54% of cases, which does not differ from regular proteins.

The second approach was to investigate conservation of lasso topology among protein do-

mains (grouped by PFAM). Protein domains are considered the smallest stable, functional protein units. We checked which motif (complex lasso, trivial lasso, no lasso) can be found in a majority of structures among different PFAM domains with at least one complex lasso topology protein. It happened that among 275 different PFAM domains, complex lasso topology dominated in 97 domains (35% of cases), trivial lasso in 66 domains (24 %), and no lasso in the rest 112 (41%).

These results suggest, that although in some cases complex lasso may be functional (like in the case of miniproteins and cysteine knot proteins), it is certainly not a rule.

Article 2: *AlphaLasso – a web server to identify loop and lasso motifs in 3D structure of biopolymers*

> Rubach, P., Płonka, J., **Gren, B. A.**, Bruno da Silva, F., Korpacz, M., & Sulkowska, J. I. (2025). AlphaLasso – a web server to identify loop and lasso motifs in 3D structure of biopolymers. **Nucleic Acids Research**, 53(W1), W11-W19.

While the analysis which we performed in the [Article 1], gave us some information about lasso proteins, the set which we investigated is biased. We investigated proteins from PDB, which stores experimentally solved proteins, mainly those which can be crystallized. Especially underrepresented are intrinsically disoredered and membrane proteins. AI method give us opportunity to study them as well.

In this article we present the AlphaLasso server and database dedicated for identification of lasso types in protein structures – experimentally solved and predicted ones (taking into account their quality e.g. pLDDT). After the explosion of number of AI-predicted structures a tool for automated analysis of lasso proteins was needed. Besides the server which enables one to identify lasso type using both minimal surface method and GLN, AlphaLasso contains a database with precomputed results of all AlphaFold v4 predicted structures. The tool is available at `https://alphalasso.cent.uw.edu.pl`.

Since AI predicted structures do not provide information about intra-chain covalent bonds, lasso analysis requires an algorithm which searches/predicts such bonds. In AlphaLasso a user can choose a number of methods to detect lasso bridges of an uploaded structure:

- automatic detection of `CYS-CYS` disulfide bond in two ways: by searching for `S-S` bonds shorter than 3Å, and by pretrained SVM model which takes geometry into account;
- by inserting specific amino acids which constitute the bridge pair e.g. to search for lassos closed by salt bridges one can insert `LYS` for the first amino acid and `ASP,GLU` for the second amino acid;
- be inserting specific indices of bridge creating amino acids.

Moreover, the server generates a list of similar structures present in the database (using Foldseek and MMseqs2). Finally, uploaded structures are checked for steric clashes, and mean pLDDT of crucial parts: lasso loop, N-tail, C-tail.

Structures in the database show much greater variety and complexity of possible lasso motifs, compared to experimentally solved proteins. In experimental data 11 different lasso

types have been found, with the most complicated L_6, $LL_{4,3}$, and LS_4. AlphaLasso contains nearly 300 different lasso types with most complicated L_{30}, $LL_{5,9}$, LS_{10} (all pLDDT > 90!). Moreover, the database enables cross-reference checks with AlphaKnot, for search of knotted lassos – we have found such examples, e.g. L_1 lasso tangles with a 3_1 knot in a sodium/calcium exchanger protein (UniProtKB id: P32418).

My main contribution have been before the idea of server/database appeared. I wanted to find interesting cases of complex lasso proteins among AlphaFold predicted structures. I created a pipeline that identified lasso structures and calculated maxGLN to select out those with high absolute value, which correspond to supercoiled lassos. This way I could reduce the number of structures to analyze. These selected structures were used for further full analysis using minimal surface. PR (the first author) used his kafka-slurm-agent [99] to spread job computations between workstations and computer clusters. Later, this pipeline was used as a part of the server and to analyze all AlphaFold structures (with pLDDT > 70) which filled the database. JP (the second author) created the server and the database. My other contribution was a creation of a tool for visual comparison of ranges of lasso loop, and participation in writing the documentation.

Article 3: *Conservation of knotted and slipknotted topology in transmembrane transporters*

> Zayats, V., Sikora, M., Perlinska, A. P., Stasiulewicz, A., **Gren, B. A.**, & Sulkowska, J. I. (2023). Conservation of knotted and slipknotted topology in transmembrane transporters. **Biophysical Journal**, 122(23), 4528-4541.

Proteins with slipknots in their structure are another group of proteins in which a chain has to be threaded through the loop. Because a majority of slipknots were found in membrane proteins, we decided to study them for the presence of knots and slipknots.

Analysis of all proteins in OPM database have shown, that (slip)knots are present only in transmembrane proteins; more precisely, only in 4 out of 156 superfamilies (140 at the time of publication) of α-helical polytopic class. These 4 superfamilies divide into 17 families:

- APC (amino acid-polyamine-organocation) with 14 families, which depending on slipknot type can be divided into two groups of families
  - LeuT-like proteins (11 families) with $S3_14_13_1$ slipknots;
  - NCS2-like proteins (3 families) with $S3_13_1$ slipknots;
- DAACS (dicarboxylate/amino acid: cation symporter) superfamily, with a single family with proteins which change slipknot type between $S3_13_1$ and $S3_1$ depending on a conformation;
- CPA (monovalent cation-proton antiporter) with a single family with $S3_1$ slipknots;
- NCX (sodium/calcium exchanger) with a single family with $K3_1$ knots.

All of them are secondary active transporters (cation symporters and antiporters). It appears that 25% of all secondary active transporters in mammals possess a (slip)knot topology (compared to 1% for all PDB proteins).

Structural analysis shows, that division of slipknot into characteristic fragments (slipknot loop and tails, knot core) coincide with protein division into functional fragments. For example, in proteins of LeuT family (APC superfamily) change of conformation of a slipknot tail is responsible for switching the opening of the transporter gate to the inside/outside. Moreover, the border between a slipknot tail and a knot core is a hinge point. Another example, in proteins of SDF family (DAACS superfamily) change of conformation which between inward-open and outward-open can be tracked by change of slipknot fingerprint between $S3_13_1$ and $S3_1$.

These findings show, that slipknot complex topology was beneficial for coordination of complex motions of transporters' structural elements. It may be utilized in the future for creation of drugs which interfere with topology.

In this article, I performed an analysis of the largest groups of proteins with the most complicated topology (LeuT-like with $S3_14_13_1$ slipknot). I computed an Alexander polynomial for all subchains of each protein chain to identify slipknots. Then I algorithmically identified knot cores and slipknot loops of each of three constituent slipknots. Additionally, I computed knot types of proteins containing PFAM domain PF01699 (NCX superfamily) predicted by AlphaFold.

Article 4: *AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods*

> Niemyska, W., Rubach, P., **Gren, B. A.**, Nguyen, M. L., Garstka, W., Bruno da Silva, F., Rawdon, E. J., & Sulkowska, J. I. (2022). AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods. **Nucleic acids research**, 50(W1), W44-W50.

As mentioned in the summary of [Article 2], the set of protein structures available in the Protein Data Bank is biased by the capabilities of experimental methods. Studies on proteins from PDB show [100] that $\sim$1% of protein structures are knotted (however, this number has not been updated for 6 years), but how does this number compare for the entire set of all proteins? Moreover, if they are conserved among families, what are the most complicated knotted folds present in proteins. To answer these questions, we have created a public server and database to allow anyone to study knotted proteins predicted by AI methods.

AlphaKnot is a server and database dedicated for the recognition of entanglement in protein structures. If a structure contains a per-residue measure of local confidence (like pLDDT for AlphaFold predicted proteins), then it is used (together with identified knot type) to judge the reliability of the structure. Database collected all knotted AlphaFold v2 structures ($\sim$360 K structures); since the release of AlphaKnot in 2024 database collects all knotted structures predicted by AlphaFold v4 ($\sim$200 KK structures). The tool is available at `https://alphaknot.cent.uw.edu.pl`.

We performed a similar analysis as in the [Article 5] and deposited the results in a database. We have found many structures with suspiciously complex knots; therefore, we classified (by

visual inspection) predicted knotted structures into three categories:

- "knotted", highly confident structures;
- "unsure", structures with some issues;
- "artifact", structures where entanglement is probably due to issues with the model.

The aforementioned classification does not depend on pLDDT score, a good example is a prediction of a protein (UniProtKB ID: Q54P92) that contains 10_{146} knot, with high pLDDT (88/100), which we classified as an artifact.

Our server performs similar analysis of uploaded structures as in the case of structures deposited in a database. Additionally, it automatically performs an artifact check, which takes into account knotting complexity, steric clashes, and calculates a separate mean pLDDT check for a knot core and both tails. Moreover, if few structures are uploaded at the same time, the server enables graphical comparison of detected knot cores.

PR, WN and I were responsible for the majority of the software. PR created a system for the distribution of computing jobs between computer clusters [99], which greatly increased our computational capabilities. My responsibility (besides the discussion of the methods used) was to prepare a pipeline of jobs to distribute. There were two types of jobs, with separate pipelines:

- Determination of a knot type of a structure ($\sim$360 K structures).
- Calculation of Alexander map of each structure in which knot have been found (3342 structures).

Moreover, I was responsible for writing most of the online documentation ("Read more"), and for programming of a tool for a graphical comparison of detected knot cores.

Article 5: *AlphaFold predicts novel human proteins with knots*

> Perlinska, A. P., Niemyska, W. H., **Gren, B. A.**, Bukowicki, M., Nowakowski, S., Rubach, P., & Sulkowska, J. I. (2023). AlphaFold predicts novel human proteins with knots. **Protein Science**, 32(5), e4631.

In this article we performed an analysis of all human protein structures predicted by AlphaFold v2 for the presence of knots. We analyzed 23391 structures (20504 proteins) using HOMFLY-PT polynomial, from which 340 happened to be knotted. We treated a structure as knotted, if a majority of its (200) closures have given a knot. Then, a knot core for each knotted structure have been identified by computation of Alexander polynomial for all possible subchains.

These 340 structures were further analyzed by more detailed analysis including a visual inspection. Structures with obvious problems were classified as artifacts. We identified over 75 % of these 340 structures to be artifacts. Many of them have low pLDDT ($<$50), and over 1/3 of them are fragments of large proteins ($>$2700 amino acids), which were modeled by multiple overlapping structures. These structures were excluded from further analysis.

Remaining 69 structures were further analyzed in the context of their homologue structures. Homology clusters were created using sequence clustering (CD-hit with 60 % sequence identity cutoff). Proteins with a very low pLDDT were remodeled using locally installed AlphaFold v2.1.1 by checking topology of homologue protein structures. If a given knot was

present in over 80% of homologue structures and in a model generated by RoseTTaFold, then the was treated as knotted (51), otherwise it was treated as potentially knotted (18).

A majority of knotted proteins (41), were already known as knotted, either because they were solved experimentally in the past, or they are part of an established family of knotted proteins. The 10 remaining knotted proteins and all potentially knotted proteins, are considered a new observations. Few of them are worth additional attention:

- Calcium-activated chloride channel regulator 1 (UniProtKB ID: A8K7I4) is a partially solved 914 amino acid long protein. Its the only crystallized domain forms an unknot, but full structure predicted by AlphaFold contains a 3_1 knot. This example shows how it is important to have a full protein structure.
- Von Willebrand factor A domain-containing protein 5A (UniProtKB ID: O00534), contains brand new 6_3 knot, which haven't been observed in any experimentally solved structure. It is especially interesting, since all knots found in proteins are twist knots, however 6_3 is not a twist knot. Knot is deep and structure is highly confident due to pLDDT score, but it is not conserved among other species. Recently crystallized structure shows, that this proteins truly contains a 6_3 knot [unpublished].

My responsibility was to prepare a pipeline for identification of knot type of each structure and its knot core. Firstly, I checked all structures from human proteome (23391 structures) with the computation of HOMFLY-PT polynomial. It let me find 340 knotted structures. For each one I computed an Alexander polynomial for all possible continuous subchains, which enabled us to find knot core of each knotted structure (shortest subchain which is still knotted). The database is available at `https://knotprot.cent.uw.edu.pl/alphafold`.

Article 6: *Discovery of a trefoil knot in the RydC RNA: Challenging previous notions of RNA topology*

> Niemyska, W., Mukherjee, S., **Gren, B. A.**, Niewieczerzal, S., Bujnicki, J. M., & Sulkowska, J. I. (2024). Discovery of a trefoil knot in the RydC RNA: challenging previous notions of RNA topology. **Journal of Molecular Biology**, 436(6), 168455.

Knots have been found in proteins and DNA, and now we have found knotted RNA as well – contrariwise to what other scientists said about their absence [74, 73]. We checked the knot type of all RNA structures in the BGSU database with a resolution quality better than 4Å. One of them appeared to be knotted – RydC which is a small, non-coding RNA involved in the regulation of gene expression. Unfortunately, no homologs of RydC were characterized experimentally; therefore, we couldn't check if the knot type is conserved in a whole family.

I performed a screening of all high-quality RNA structures in search of knotted ones. Firstly, I obtained a list of all RNA representative chains from the BGSU RNA database with a 4 Å resolution or better. The representative chain is a chosen structure which represents a set of structures with the same sequence of nucleic acids. Some RNA molecules are composed of few RNA chains – I joined them if their ends were closer than 4 Å, and appended them to the list.

Secondly, I removed all atoms which do not constitute the backbone of a RNA molecule, therefore each nucleic acid was represented by P, O5′, C5′, C4′, C3′, O3′ atoms. In procedure described in a previous survey [74] each nucleotide was reduced into a single P atom. I realized that this method is too coarse-grained and can change the knot type of a RNA molecule.

Finally, I closed chains of RNA molecules and performed calculations of the Alexander polynomial to determine the knot type of each simplified closed chain. This way I found exactly one knot: -3_1 in chain Q of 4V2S PDB structure. To determine its chirality I used HOMFLY-PT polynomial.

The knotted structure of 4V2S has missing nucleotides (part of 19, and whole 20 and 21) which were modeled as a straight line, however beyond reasonable doubt it cannot alter the topology of this molecule. The missing part is an artifact of X-ray crystallography. The missing nucleotides in the structure file are due to the unstructured part of the chain in the investigated crystal. I have shown that there is no available conformation of the missing part that could affect topology of the whole molecule, even after approximating a missing chain by an infinitely-elastic string.

Additionally, we performed coarse-grained folding dynamics of RydC with a gaussian potential between native contacts and a repulsive Lennard-Jones potential between non-native contacts. Each nucleotide was represented by two beads: one for the backbone, and one for the sidechain. We found out that folding of RydC is protein-like – it is steered kinetically, follows a folding funnel, and is stabilized by an intermediate state.

Article 7: *Knotted artifacts in predicted 3D RNA structures*

> **Gren, B. A.**, Antczak, M., Zok, T., Sulkowska, J. I., & Szachniuk, M. (2024). Knotted artifacts in predicted 3D RNA structures. **PLOS Computational Biology**, 20(6), e1011959.

AlphaFold group during CASP14 has shown great, experimental-like accuracy in the prediction of protein structures, for which it has been awarded the Nobel Prize in Chemistry. After that, other new AI methods emerged; not only for the prediction of protein chains, but for the prediction of RNA chains as well. It motivated us to check how good these RNA structure predicting methods are compared to traditional, energy-based methods.

In this article, we analyzed all RNA models submitted to the CASP15 competition (Critical Assessment of Techniques for Protein Structure Prediction). CASP15 was the first edition in which RNA models, besides protein models, were evaluated. The goal of our analysis was to evaluate predicted structures from a topology viewpoint – which lacks in CASP scoring functions. The article is a manifest to increase awareness of topological irregularities in 3D RNA predictions.

In our analysis, we searched for incorrect topology at two levels: global, by identification of knots (JS and my responsibility), and local, by identification of entanglement of structural elements (MS, AM, and TZ responsibility). Note that all target molecules were trivial in these aspects. Any entanglement of structured elements (which from a topological point of view relates to non-trivial lassos and links) is considered to be impossible in real RNA

molecules; however, only recently the first knotted RNA molecule has been identified by me [Article 6]. Analysis at both levels (global and local) led to quantitatively slightly different, but qualitatively the same results.

I downloaded all target and predicted RNA molecules, extracted their backbone (P-O5′-C5′-C4′-C3′-O3′ trace) and calculated the Alexander polynomial after closing the chains in a probabilistic way (default way) or directly (if I decided by visual inspection that molecule termini are close enough). Then I performed a statistical analysis.

First part of main result refers to the methods used by groups participating in the competition; we divided methods into machine learning methods (ML), and classical methods (non-ML). We have found out that ML methods, compared to non-ML methods, were a few times more susceptible to predict topologically incorrect models. Knots were 7 times more present in structures modeled using ML methods (1.2% vs 8.2%), while entanglements of structural elements were 4 times more present (3.9% vs 16.2%). Moreover, only structures predicted by ML methods had knots complicated that much, we were unable to identify their type (0% vs 2.5%).

Second part of main result refers to types of modeling targets which had to be predicted; we divided targets between artificially designed RNAs and natural RNAs. They had 10 times more knots and 5 times more entanglements of structural elements. Moreover, all knots among predictions of natural RNAs were trefoils – the simplest non-trivial knots; however, in the case of artificial RNAs nearly half of knots were more complicated.

Article 8: *Topoly: Python package to analyze topology of polymers*

> Dabrowski-Tumanski, P., Rubach, P., Niemyska, W., **Gren, B. A.**, & Sulkowska, J. I. (2021). Topoly: Python package to analyze topology of polymers. **Briefings in Bioinformatics**, 22(3), bbaa196.

Other mentioned [Articles 1-7] could have been created because of a dozen of methods for topological analysis of (bio)polymers we created in our lab. A majority of them we released in the form of "Topoly", a Python software library. The idea of package is to be powerful and simple to use at the same time; e.g. simple command `topoly.alexander("protein.cif")`:

- loads a structure saved as `protein.cif`,
- using probabilistic closure creates 200 of closed chains;
- reduces complexity of 200 chains by performing KMT reduction, which preserves the type of a knot;
- computes Alexander polynomial for 200 closed chain and finds the corresponding topologies and returns the results.

By passing non-default values to the function variables, users can change: closure method, number of performed closures, output format. Moreover, users can choose to perform computation for all (or some) subchains of the inputted structure. Topoly has been supported since the release and is still in development by Paweł Rubach, Wanda Niemyska, and I. Topoly is the main software behind our AlphaKnot and AlphaLasso servers and databases.

The main functionality of Topoly can be divided into 3 parts:

- identification of a type of knot/link/handcuff/Θ-curve in (bio)polymers using knot polynomials: Alexander, HOMFLY-PT, Yamada, Conway, Jones, Kauffman;
- identification of lasso types in (bio)polymers using minimal surface, and calculation of Gaussian Linking Number;
- generation of random, ideal, equilateral polymers: open, closed, lasso, handcuff, Θ-curve.

My main contribution was implementation of aforementioned generation of random ideal, equilateral polymers of various topologies. Circular elements of these polymers were implemented using a quadratic time complexity method proposed by Jason Cantarella et al. [101] for sampling of equilateral polygons embedded in 3D. After personal consultation with Jason Cantarella [102, 103] I improved the method for generation of Θ-curves to make it more general; letting the generation of n polygonal, equilateral curves of arbitrary lengths with shared endpoints (for $n = 2$ it gives a circle, for $n = 3$ it gives a Θ-curve).

My secondary contribution was wrapping our C++ code (for computation of minimal surface and GLN) into a Cython. Moreover, I wrote down documentation available at `https://topoly.cent.uw.edu.pl/`, and programmed a tutorial with examples on how to use Topoly `https://github.com/ilbsm/topoly_tutorial/`.

In subsequent years after the publication of the article, few of my ideas were added to the Topoly :

- improvement of computation time complexity of maxGLN algorithm (inspired by Kadane algorithm used in bioinformatics) from $\mathcal{O}(m^2n^2)$ to $\mathcal{O}(mn^2)$ (where m and n are lengths of lasso loop and lasso tail). It enables finding of part of lasso tail which return maksimal absolute value of GLN without computation of GLN for all subchains;
- calculation of Alexander polynomial in point $\Delta(t = 12)$, which is as fast as popular computation at $t = \pm 1$ using Alexander matrix (with $\mathcal{O}(n^3)$, where n is number of crossings), but nearly as strong as $\Delta(t)$ (with $\mathcal{O}(2^n)$). This improvement is especially important for computation of matrices, for slipknot/knot core identification.
- extension of Topoly dictionary (which translates polynomials into specific knot types) from structures with up to 8 crossings, to structures with up to 12 crossings.

Bibliography

[1] J. H. Przytycki, "Classical roots of knot theory," *Chaos, Solitons & Fractals*, vol. 9, no. 4-5, pp. 531–545, 1998.

[2] H. v. Helmholtz, "Über integrale der hydrodynamischen gleichungen, welche den wirbelbewegungen entsprechen.," *Crelle's Journal für Mathematik*, 1858.

[3] J. C. Maxwell, *A treatise on electricity and magnetism*, vol. 1. Clarendon press, 1873.

[4] W. Thomson, "Ii. on vortex atoms," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 34, no. 227, pp. 15–24, 1867.

[5] J. W. Alexander, "Topological invariants of knots and links," *Transactions of the American Mathematical Society*, vol. 30, no. 2, pp. 275–306, 1928.

[6] P. Freyd, D. Yetter, J. Hoste, W. R. Lickorish, K. Millett, and A. Ocneanu, "A new polynomial invariant of knots and links," 1985.

[7] J. H. Przytycki and P. Traczyk, "Invariants of links of conway type," *arXiv preprint arXiv:1610.06679*, 2016.

[8] K. Koniaris and M. Muthukumar, "Self-entanglement in ring polymers," *The Journal of chemical physics*, vol. 95, no. 4, pp. 2873–2881, 1991.

[9] B. A. Burton, "The next 350 million knots," in *36th International Symposium on Computational Geometry (SoCG 2020)*, pp. 25–1, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020.

[10] W. P. Thurston, "Three dimensional manifolds, kleinian groups and hyperbolic geometry," 1982.

[11] J. Baldwin and S. Sivek, "On the complexity of torus knot recognition," *Transactions of the American Mathematical Society*, vol. 371, no. 6, pp. 3831–3855, 2019.

[12] J. H. Conway, "An enumeration of knots and links, and some of their algebraic properties," in *Computational problems in abstract algebra*, pp. 329–358, Elsevier, 1970.

[13] M. L. Mansfield, "Are there knots in proteins?," *Nature structural biology*, vol. 1, no. 4, p. 213, 1994.

[14] N. P. King, E. O. Yeates, and T. O. Yeates, "Identification of rare slipknots in proteins and their implications for stability and folding," *Journal of molecular biology*, vol. 373, no. 1, pp. 153–166, 2007.

[15] J. I. Sułkowska, E. J. Rawdon, K. C. Millett, J. N. Onuchic, and A. Stasiak, "Conservation of complex knotting and slipknotting patterns in proteins," *Proceedings of the National Academy of Sciences*, vol. 109, no. 26, pp. E1715–E1723, 2012.

[16] P. Dabrowski-Tumanski, W. Niemyska, P. Pasznik, and J. I. Sulkowska, "Lassoprot: server to analyze biopolymers with lassos," *Nucleic acids research*, vol. 44, no. W1, pp. W383–W389, 2016.

[17] W. Niemyska, P. Dabrowski-Tumanski, M. Kadlof, E. Haglund, P. Sułkowski, and J. I. Sulkowska, "Complex lasso: new entangled motifs in proteins," *Scientific reports*, vol. 6, p. 36895, 2016.

[18] W. Niemyska, K. C. Millet, and J. I. Sulkowska, "Gln: a method to reveal unique properties of lasso type topology in proteins," *Scientific Reports*, vol. 10, p. 15186, 2020.

[19] H. Moriuchi, "A table of θ-curves and handcuff graphs with up to seven crossings," in *Noncommutativity and Singularities: Proceedings of French–Japanese symposia held at IHÉS in 2006*, pp. 281–290, Mathematical Society of Japan, 2009.

[20] H. Moriuchi, "Enumeration of algebraic tangles with applications to theta-curves and handcuff graphs.," *Kyungpook Mathematical Journal*, vol. 48, no. 3, 2008.

[21] S. Yamada, "An invariant of spatial graphs," *Journal of graph theory*, vol. 13, no. 5, pp. 537–551, 1989.

[22] L. H. Kauffman and S. Lambropoulou, "On the classification of rational tangles," *Advances in Applied Mathematics*, vol. 33, no. 2, pp. 199–237, 2004.

[23] L. F. Liu, R. E. Depew, and J. C. Wang, "Knotted single-stranded dna rings: A novel topological isomer of circular single-stranded dna formed by treatment with escherichia coli ω protein," *Journal of molecular biology*, vol. 106, no. 2, pp. 439–452, 1976.

[24] L. F. Liu and J. C. Wang, "Supercoiling of the dna template during transcription.," *Proceedings of the National Academy of Sciences*, vol. 84, no. 20, pp. 7024–7027, 1987.

[25] O. Sordet, Q. A. Khan, K. W. Kohn, and Y. Pommier, "Apoptosis induced by topoisomerase inhibitors," *Current Medicinal Chemistry-Anti-Cancer Agents*, vol. 3, no. 4, pp. 271–290, 2003.

[26] X. Liang, Q. Wu, S. Luan, Z. Yin, C. He, L. Yin, Y. Zou, Z. Yuan, L. Li, X. Song, *et al.*, "A comprehensive review of topoisomerase inhibitors as anticancer agents in the past decade," *European journal of medicinal chemistry*, vol. 171, pp. 129–168, 2019.

[27] D. Sumners, "The role of knot theory in dna research," in *Geometry and topology*, pp. 297–318, CRC Press, 2020.

[28] C. Ernst and D. Sumners, "Solving tangle equations arising in a dna recombination model," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 126, pp. 23–36, Cambridge University Press, 1999.

[29] D. Buck and C. V. Marcotte, "Tangle solutions for a family of dna-rearranging proteins," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 139, pp. 59–80, Cambridge University Press, 2005.

[30] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[31] J. I. Sulkowska, "On folding of entangled proteins: knots, lassos, links and θ-curves," *Current Opinion in Structural Biology*, vol. 60, pp. 131–141, 2020.

[32] W. R. Taylor, "A deeply knotted protein structure and how it might fold," *Nature*, vol. 406, no. 6798, p. 916, 2000.

[33] W. R. Taylor and K. Lin, "Protein knots: a tangled problem," *Nature*, vol. 421, no. 6918, pp. 25–25, 2003.

[34] P. Dabrowski-Tumanski, A. I. Jarmolinska, W. Niemyska, E. J. Rawdon, K. C. Millett, and J. I. Sulkowska, "Linkprot: A database collecting information about biological links," *Nucleic acids research*, p. gkw976, 2016.

[35] P. Dabrowski-Tumanski and J. I. Sulkowska, "Topological knots and links in proteins," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3415–3420, 2017.

[36] P. Dabrowski-Tumanski, D. Goundaroulis, A. Stasiak, E. J. Rawdon, and J. I. Sulkowska, "Theta-curves in proteins," *Protein science*, vol. 33, no. 9, p. e5133, 2024.

[37] W. Niemyska, P. Rubach, B. A. Gren, M. L. Nguyen, W. Garstka, F. Bruno da Silva, E. J. Rawdon, and J. I. Sulkowska, "Alphaknot: server to analyze entanglement in structures predicted by alphafold methods," *Nucleic acids research*, vol. 50, no. W1, pp. W44–W50, 2022.

[38] V. Zayats, M. Sikora, A. P. Perlinska, A. Stasiulewicz, B. A. Gren, and J. I. Sulkowska, "Conservation of knotted and slipknotted topology in transmembrane transporters," *Biophysical Journal*, vol. 122, no. 23, pp. 4528–4541, 2023.

[39] A. P. Perlinska, W. H. Niemyska, B. A. Gren, M. Bukowicki, S. Nowakowski, P. Rubach, and J. I. Sulkowska, "Alphafold predicts novel human proteins with knots," *Protein Science*, vol. 32, no. 5, p. e4631, 2023.

[40] P. Dabrowski-Tumanski, B. Gren, and J. I. Sulkowska, "Statistical properties of lasso-shape polymers and their implications for complex lasso proteins function," *Polymers*, vol. 11, no. 4, p. 707, 2019.

[41] B. A. Greń, P. Dabrowski-Tumanski, W. Niemyska, and J. I. Sulkowska, "Lasso proteins—unifying cysteine knots and miniproteins," *Polymers*, vol. 13, no. 22, p. 3988, 2021.

[42] P. Rubach, J. Płonka, B. A. Gren, F. Bruno da Silva, M. Korpacz, and J. I. Sulkowska, "Alphalasso—a web server to identify loop and lasso motifs in 3d structure of biopolymers," *Nucleic Acids Research*, vol. 53, no. W1, pp. W11–W19, 2025.

[43] W. Niemyska, S. Mukherjee, B. A. Gren, S. Niewieczerzal, J. M. Bujnicki, and J. I. Sulkowska, "Discovery of a trefoil knot in the rydc rna: challenging previous notions of rna topology," *Journal of Molecular Biology*, vol. 436, no. 6, p. 168455, 2024.

[44] P. Dabrowski-Tumanski, P. Rubach, W. Niemyska, B. A. Gren, and J. I. Sulkowska, "Topoly: Python package to analyze topology of polymers," *Briefings in Bioinformatics*, 2020.

[45] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, "Critical assessment of methods of protein structure prediction (casp)—round xiv," *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 12, pp. 1607–1617, 2021.

[46] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[47] The UniProt Consortium, "Uniprot: the universal protein knowledgebase in 2025," *Nucleic acids research*, vol. 53, no. D1, pp. D609–D617, 2025.

[48] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, *et al.*, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, 2021.

[49] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[50] B. A. Gren, M. Antczak, T. Zok, J. I. Sulkowska, and M. Szachniuk, "Knotted artifacts in predicted 3d rna structures," *PLOS Computational Biology*, vol. 20, no. 6, p. e1011959, 2024.

[51] F. B. da Silva, A. P. Perlinska, J. Płonka, E. Flapan, and J. I. Sulkowska, "Universe of lasso proteins: Exploring the limit of entanglement of protein predicted by alphafold," *Journal of Molecular Biology*, p. 169217, 2025.

[52] P. Dabrowski-Tumanski, M. Piejko, S. Niewieczerzal, A. Stasiak, and J. I. Sulkowska, "Protein knotting by active threading of nascent polypeptide chain exiting from the ribosome exit channel," *The Journal of Physical Chemistry B*, vol. 122, no. 49, pp. 11616–11625, 2018.

[53] M.-F. Hsu, M. K. Sriramoju, C.-H. Lai, Y.-R. Chen, J.-S. Huang, T.-P. Ko, K.-F. Huang, and S.-T. D. Hsu, "Structure, dynamics, and stability of the smallest and most complex 71 protein knot," *Journal of Biological Chemistry*, vol. 300, no. 1, p. 105553, 2024.

[54] A. P. Perlinska, M. L. Nguyen, S. P. Pilla, E. Staszor, I. Lewandowska, A. Bernat, E. Purta, R. Augustyniak, J. M. Bujnicki, and J. I. Sulkowska, "Are there double knots in proteins? prediction and in vitro verification based on trmd-tm1570 fusion from c. nitroreducens," *Frontiers in Molecular Biosciences*, vol. 10, p. 1223830, 2024.

[55] A. L. Mallam and S. E. Jackson, "Folding studies on a knotted protein," *Journal of molecular biology*, vol. 346, no. 5, pp. 1409–1421, 2005.

[56] N. P. King, A. W. Jacobitz, M. R. Sawaya, L. Goldschmidt, and T. O. Yeates, "Structure and folding of a designed knotted protein," *Proceedings of the National Academy of Sciences*, vol. 107, no. 48, pp. 20732–20737, 2010.

[57] A. L. Mallam and S. E. Jackson, "Knot formation in newly translated proteins is spontaneous and accelerated by chaperonins," *Nature chemical biology*, vol. 8, no. 2, pp. 147–153, 2012.

[58] J. I. Sułkowska, P. Sułkowski, P. Szymczak, and M. Cieplak, "Stabilizing effect of knots on proteins," *Proceedings of the National Academy of Sciences*, vol. 105, no. 50, pp. 19714–19719, 2008.

[59] M. A. Soler and P. F. Faisca, "Effects of knots on protein folding properties," *PloS one*, vol. 8, no. 9, p. e74755, 2013.

[60] Y. Xu, S. Li, Z. Yan, Z. Luo, H. Ren, B. Ge, F. Huang, and T. Yue, "Stabilizing effect of inherent knots on proteins revealed by molecular dynamics simulations," *Biophysical journal*, vol. 115, no. 9, pp. 1681–1689, 2018.

[61] P. Virnau, L. A. Mirny, and M. Kardar, "Intricate knots in proteins: Function and evolution," *PLoS computational biology*, vol. 2, no. 9, p. e122, 2006.

[62] S. G. Ferreira, M. K. Sriramoju, S.-T. D. Hsu, P. F. Faisca, and M. Machuqueiro, "Is there a functional role for the knotted topology in protein uch-l1?," *Journal of Chemical Information and Modeling*, vol. 64, no. 17, pp. 6827–6837, 2024.

[63] M. Sikora, E. Klimentova, D. Uchal, D. Sramkova, A. P. Perlinska, M. L. Nguyen, M. Korpacz, R. Malinowska, S. Nowakowski, P. Rubach, *et al.*, "Knot or not? identifying unknotted proteins in knotted families with sequence-based machine learning model," *Protein Science*, vol. 33, no. 7, p. e4998, 2024.

[64] P. Dabrowski-Tumanski and J. Sulkowska, "To tie or not to tie? that is the question," *Polymers*, vol. 9, no. 9, p. 454, 2017.

[65] H. L. Schubert, R. M. Blumenthal, and X. Cheng, "Many paths to methyltransfer: a chronicle of convergence," *Trends in biochemical sciences*, vol. 28, no. 6, pp. 329–335, 2003.

[66] K. L. Tkaczuk, S. Dunin-Horkawicz, E. Purta, and J. M. Bujnicki, "Structural and evolutionary bioinformatics of the spout superfamily of methyltransferases," *BMC bioinformatics*, vol. 8, no. 1, p. 73, 2007.

[67] P. Dabrowski-Tumanski, P. Rubach, D. Goundaroulis, J. Dorier, P. Sułkowski, K. C. Millett, E. J. Rawdon, A. Stasiak, and J. I. Sulkowska, "Knotprot 2.0: a database of proteins with knots and other entangled structures," *Nucleic acids research*, vol. 47, no. D1, pp. D367–D375, 2019.

[68] T. Christian, R. Sakaguchi, A. P. Perlinska, G. Lahoud, T. Ito, E. A. Taylor, S. Yokoyama, J. I. Sulkowska, and Y.-M. Hou, "Methyl transfer by substrate signaling from a knotted protein fold," *Nature structural & molecular biology*, vol. 23, no. 10, pp. 941–948, 2016.

[69] A. P. Perlinska, A. Stasiulewicz, E. K. Nawrocka, K. Kazimierczuk, P. Setny, and J. I. Sulkowska, "Restriction of s-adenosylmethionine conformational freedom by knotted protein binding sites," *PLoS computational biology*, vol. 16, no. 5, p. e1007904, 2020.

[70] M. Y. Galperin, D. R. Walker, and E. V. Koonin, "Analogous enzymes: independent inventions in enzyme evolution," *Genome research*, vol. 8, no. 8, pp. 779–790, 1998.

[71] H. Wang, R. J. Di Gate, and N. C. Seeman, "An rna topoisomerase.," *Proceedings of the National Academy of Sciences*, vol. 93, no. 18, pp. 9477–9482, 1996.

[72] D. Xu, W. Shen, R. Guo, Y. Xue, W. Peng, J. Sima, J. Yang, A. Sharov, S. Srikantan, J. Yang, *et al.*, "Top3β is an rna topoisomerase that works with fragile x syndrome protein to promote synapse formation," *Nature neuroscience*, vol. 16, no. 9, pp. 1238–1247, 2013.

[73] A. S. Burton, M. Di Stefano, N. Lehman, H. Orland, and C. Micheletti, "The elusive quest for rna knots," *RNA biology*, vol. 13, no. 2, pp. 134–139, 2016.

[74] C. Micheletti, M. Di Stefano, and H. Orland, "Absence of knots in known rna structures," *Proceedings of the National Academy of Sciences*, vol. 112, no. 7, pp. 2052–2057, 2015.

[75] A. Blond, J. Péduzzi, C. Goulard, M. J. Chiuchiolo, M. Barthélémy, Y. Prigent, R. A. Salomón, R. N. Farías, F. Moreno, and S. Rebuffat, "The cyclic structure of microcin j25, a 21-residue peptide antibiotic from escherichia coli," *European journal of biochemistry*, vol. 259, no. 3, pp. 747–756, 1999.

[76] J. O. Solbiati, M. Ciaccio, R. N. Farías, J. E. González-Pastor, F. Moreno, and R. A. Salomoón, "Sequence analysis of the four plasmid genes required to produce the circular peptide antibiotic microcin j25," *Journal of bacteriology*, vol. 181, no. 8, pp. 2659–2662, 1999.

[77] R. W. Jack and G. Jung, "Lantibiotics and microcins: polypeptides with unusual chemical diversity," *Current opinion in chemical biology*, vol. 4, no. 3, pp. 310–317, 2000.

[78] K.-A. Wilson, M. Kalkum, J. Ottesen, J. Yuzenkova, B. T. Chait, R. Landick, T. Muir, K. Severinov, and S. A. Darst, "Structure of microcin J25, a peptide inhibitor of bacterial rna polymerase, is a lassoed tail," *Journal of the American Chemical Society*, vol. 125, no. 41, pp. 12475–12483, 2003.

[79] S. Rebuffat, A. Blond, D. Destoumieux-Garzón, C. Goulard, and J. Peduzzi, "Microcin J25, from the macrocyclic to the lasso structure: implications for biosynthetic, evolutionary and biotechnological perspectives," *Current Protein and Peptide Science*, vol. 5, no. 5, pp. 383–391, 2004.

[80] C. Clavel, K. Fournel-Marotte, and F. Coutrot, "A ph-sensitive peptide-containing lasso molecular switch," *Molecules*, vol. 18, no. 9, pp. 11553–11575, 2013.

[81] C. D. Allen and A. J. Link, "Self-assembly of catenanes from lasso peptides," *Journal of the American Chemical Society*, vol. 138, no. 43, pp. 14214–14217, 2016.

[82] C. Zong, M. J. Wu, J. Z. Qin, and A. J. Link, "Lasso peptide benenodin-1 is a thermally actuated [1] rotaxane switch," *Journal of the American Chemical Society*, vol. 139, no. 30, pp. 10403–10409, 2017.

[83] C. Zong, M. O. Maksimov, and A. J. Link, "Construction of lasso peptide fusion proteins," *ACS chemical biology*, vol. 11, no. 1, pp. 61–68, 2015.

[84] T. A. Knappe, U. Linne, L. Robbel, and M. A. Marahiel, "Insights into the biosynthesis and stability of the lasso peptide capistruin," *Chemistry & biology*, vol. 16, no. 12, pp. 1290–1298, 2009.

[85] N. R. Braffman, F. J. Piscotta, J. Hauver, E. A. Campbell, A. J. Link, and S. A. Darst, "Structural mechanism of transcription inhibition by lasso peptides microcin J25 and capistruin," *Proceedings of the National Academy of Sciences*, vol. 116, no. 4, pp. 1273–1278, 2019.

[86] E. Haglund, J. I. Sułkowska, Z. He, G.-S. Feng, P. A. Jennings, and J. N. Onuchic, "The unique cysteine knot regulates the pleotropic hormone leptin," *Plos one*, vol. 7, no. 9, p. e45654, 2012.

[87] E. Haglund, A. Pilko, R. Wollman, P. A. Jennings, and J. N. Onuchic, "Pierced lasso topology controls function in leptin," *The Journal of Physical Chemistry B*, vol. 121, no. 4, pp. 706–718, 2017.

[88] E. Haglund, L. Nguyen, N. P. Schafer, H. Lammert, P. A. Jennings, and J. N. Onuchic, "Uncovering the molecular mechanisms behind disease-associated leptin variants," *Journal of Biological Chemistry*, vol. 293, no. 33, pp. 12919–12933, 2018.

[89] K. Adelman, J. Yuzenkova, A. La Porta, N. Zenkin, J. Lee, J. T. Lis, S. Borukhov, M. D. Wang, and K. Severinov, "Molecular mechanism of transcription inhibition by peptide antibiotic microcin j25," *Molecular cell*, vol. 14, no. 6, pp. 753–762, 2004.

[90] D. Frechet, J. Guitton, F. Herman, D. Faucher, G. Helynck, B. Monegier du Sorbier, J. Ridoux, E. James-Surcouf, and M. Vuilhorgne, "Solution structure of rp 71955, a new 21 amino acid tricyclic peptide active against hiv-1 virus," *Biochemistry*, vol. 33, no. 1, pp. 42–50, 1994.

[91] D. J. Craik, N. L. Daly, and C. Waine, "The cystine knot motif in toxins and implications for drug design," *Toxicon*, vol. 39, no. 1, pp. 43–60, 2001.

[92] H. Kessler, "Introducing lasso peptides as molecular scaffolds for drug design: engineering of an integrin antagonist," *Angewandte Chemie International Edition*, 2011.

[93] V. Rana, I. Sitarik, J. Petucci, Y. Jiang, H. Song, and E. P. O'Brien, "Non-covalent lasso entanglements in folded proteins: Prevalence, functional implications, and evolutionary significance," *Journal of molecular biology*, vol. 436, no. 6, p. 168459, 2024.

[94] R. G. Scharein and K. S. Booth, "Interactive knot theory with knotplot," in *Multimedia tools for communicating mathematics*, pp. 277–290, Springer, 2002.

[95] R. Borza, F. Salgado-Polo, W. H. Moolenaar, and A. Perrakis, "Structure and function of the ecto-nucleotide pyrophosphatase/phosphodiesterase (enpp) family: Tidying up diversity," *Journal of Biological Chemistry*, vol. 298, no. 2, 2022.

[96] J. Hausmann, S. Kamtekar, E. Christodoulou, J. E. Day, T. Wu, Z. Fulkerson, H. M. Albers, L. A. Van Meeteren, A. J. Houben, L. Van Zeijl, *et al.*, "Structural basis of substrate discrimination and integrin binding by autotaxin," *Nature structural & molecular biology*, vol. 18, no. 2, pp. 198–204, 2011.

[97] S. Jansen, A. Perrakis, C. Ulens, C. Winkler, M. Andries, R. P. Joosten, M. Van Acker, F. P. Luyten, W. H. Moolenaar, and M. Bollen, "Structure of npp1, an ectonucleotide pyrophosphatase/phosphodiesterase involved in tissue calcification," *Structure*, vol. 20, no. 11, pp. 1948–1959, 2012.

[98] M. Koyama, H. Nishimasu, R. Ishitani, and O. Nureki, "Molecular dynamics simulation of autotaxin: roles of the nuclease-like domain and the glycan modification," *The Journal of Physical Chemistry B*, vol. 116, no. 39, pp. 11798–11808, 2012.

[99] P. Rubach, "Applying large-scale distributed computing to structural bioinformatics – bridging legacy hpc clusters with big data technologies using kafka-slurm-agent," in *Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration*, pp. 1–8, 2025.

[100] M. Jamroz, W. Niemyska, E. J. Rawdon, A. Stasiak, K. C. Millett, P. Sułkowski, and J. I. Sulkowska, "Knotprot: a database of proteins with knots and slipknots," *Nucleic acids research*, vol. 43, no. D1, pp. D306–D314, 2015.

[101] J. Cantarella, B. Duplantier, C. Shonkwiler, and E. Uehara, "A fast direct sampling algorithm for equilateral closed polygons," *Journal of Physics A: Mathematical and Theoretical*, vol. 49, no. 27, p. 275202, 2016.

[102] J. Cantarella and C. Shonkwiler, "The symplectic geometry of closed equilateral random walks in 3-space," *The Annals of Applied Probability*, vol. 26, no. 1, pp. 549–596, 2016.

[103] J. Cantarella, B. Duplantier, and C. Shonkwiler. personal communication.

Preprint

Classification of algebraic tangles

Classification of algebraic tangles

Bartosz A. Gren [1*], Joanna I. Sulkowska [1] and Boštjan Gabrovšek [2,3]

[1]University of Warsaw, Centre of New Technology, S. Banacha 2C, 02-097 Warsaw, Poland.
[2]University of Ljubjana, Faculty of Education, Kardeljeva ploščad 16, 1000 Ljubjana, Slovenia.
[3]Rudolfovo – Science and Technology Centre, Podbreznik 15, 8000 Novo mesto, Slovenia.

*Corresponding author(s). E-mail(s): b.gren@cent.uw.edu.pl;
Contributing authors: j.sulkowska@cent.uw.edu.pl;
bostjan.gabrovsek@pef.uni-lj.si;

Abstract

We study algebraic tangles as fundamental components in knot theory, developing a systematic approach to classify and tabulate prime tangles using a novel canonical representation. The canonical representation enables us to distinguish mutant tangles, which fills the gaps in previous classifications. Moreover, we increase the classification of prime tangles up to 14 crossings and analyze tangle symmetry groups. We provide a database of our results: https://tangleinfo.cent.uw.edu.pl.

Keywords: Tangles, Algebraic Tangles, Rational Tangles, Mutant knots, Knot theory, Canonical form

1 Introduction

The formal study of tangles was introduced by John Conway in the 1970s [1]. Conway's work revolutionized the way mathematicians approached knot theory, providing new methods for decomposing and analyzing knots. Conway showed that a knot or link can be decomposed into tangles, allowing us to represent knots and links in a concise and systematic way. A tangle consists of two arcs and a finite number of circles embedded

in a three-dimensional ball such that the endpoints of the arcs lie on the boundary of the ball.

A particularly important subclass of tangles are rational tangles, which, due to their simplicity, have been extensively studied in the literature, as the continued fraction associated with a rational tangle is a perfect invariant. One notable application of studying tangles is in the development of knot tables, since tangles offer a very compact and efficient way to encode a knot.

However, creating knot tables through the use of tangles involves the classification of algebraic tangles [1], which are much more difficult to study, and the theory of algebraic tangles remains undeveloped – a gap that this paper aims to address. Knot tangle decompositions were also studied in the context of arborescent knots by Bonahon and Siebenmann [2].

The concept of knot classification follows a straightforward approach: first, construct an initial set of knot diagrams that comprehensively includes all possible knots; then, systematically eliminate duplicate representations to obtain a distinct classification. The second step presents a computational bottleneck; therefore, it is advantageous to start with the smallest possible initial set. A naive approach to generate an initial set of knots with up to N crossings involves a generation of all possible 4-valent planar graphs (with at most N vertices) and then a replacement of the vertices with crossings. A more effective approach is to generate 4-valent polyhedral graphs and replace the vertices with algebraic tangles – using this method, Conway managed to tabulate knots with up to 11 crossings and links with up to 10 crossings by hand [1]. Due to the lack of a tabulation of algebraic tangles, later classifications were created using the naive approach, supplemented by tangle properties, which led to the construction of knot tables up to 13 crossings (in 1983), 16 crossings (in 1998), and 19 crossings (in 2012) [3–5]. With a tabulation of algebraic tangles, knot and link tables could be easily extended.

Beyond their role in knot classification, tangles have found diverse applications in knot theory; for instance, decomposing a knot into tangles is essential for faster computations of the computationally intensive Khovanov homology [6], a theory that provides a powerful categorification of the Jones polynomial and has advanced the study of knot invariants, 4-dimensional topology, and quantum algebra.

Our paper is highly motivated by the need to create tables of various entangled structures that appear in the study of biological structures. Knot tables initiated the study of entanglement in proteins [7–9]; later, proteins with lasso motifs were also studied [10–12]. Moriuchi's work on classifying Θ-curves and handcuff links up to seven crossings [13] was essential for the study of Θ-curves in proteins [14–16], and we wish to extend these studies to analyse protein as bonded knots [17–19], bonded links [20, 21], and bonded knotoids [22–24].

Beyond topological classification, tangle theory also plays a crucial role in understanding biological processes. For example, the study of topoisomerase enzymes [25], which manage the topology of DNA, illustrates how tangles are manipulated during cellular processes. Tangles are also used to understand the mechanisms of DNA recombination [26, 27].

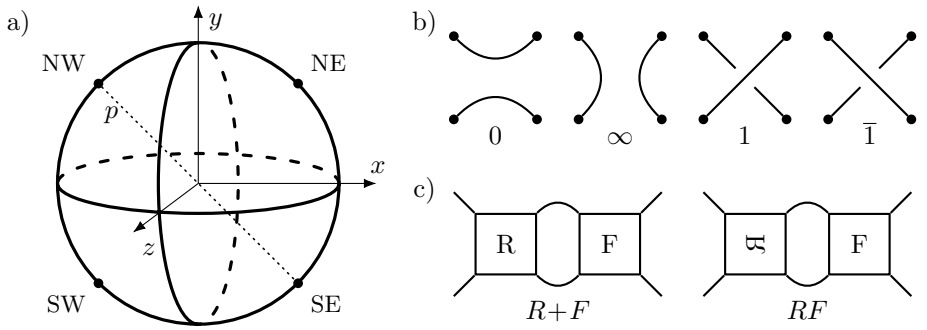


Fig. 1 a) A Conway's sphere, in which the tangle is embedded. b) Basic tangles which satisfy the boundary. c) Sum and product of tangles R and F.

This paper is structured as follows: In section 2, we provide the necessary background and key definitions to establish the scope of the paper. In section 3, we define isotopy preserving moves, up to which we provide a classification. In section 4, we introduce a binary tree notation as the foundation for a canonical representation of algebraic tangles, the central objective of this article. We further propose an algorithm to construct this representation and establish its uniqueness. In section 5, we present our tangle classification, which is further expanded in section 6 where we study tangle symmetries.

2 Preliminaries

Definition 1 (Tangle)**.** A *tangle* is an embedding of two arcs (homeomorphic to the interval $[0,1]$) and circles into the unit 3-ball $B^3 = \{(x,y,z) \in \mathbb{R}^3 \mid x^2+y^2+z^2 \leq 1\}$, where only the endpoints of the two arcs lie on the boundary of the ball, ∂B^3, and are located at the four points

$$\text{NW} = \left(\tfrac{-1}{\sqrt{2}}, \tfrac{1}{\sqrt{2}}, 0\right),\ \text{NE} = \left(\tfrac{1}{\sqrt{2}}, \tfrac{1}{\sqrt{2}}, 0\right),\ \text{SW} = \left(\tfrac{-1}{\sqrt{2}}, \tfrac{-1}{\sqrt{2}}, 0\right),\ \text{and SE} = \left(\tfrac{1}{\sqrt{2}}, \tfrac{-1}{\sqrt{2}}, 0\right).$$

The sphere ∂B^3, which the tangle intersects at these four points, is referred to as the *Conway sphere*. Furthermore, we define the *principal diagonal* axis as the line $p = \{(x,-x,0) \in \mathbb{R}^3 \mid x \in \mathbb{R}\}$, which connects the NW-SE boundary points, see Figure 1a.

Definition 2 (VHX)**.** We classify tangles based on how the open arcs connect the boundary points, grouping them into three distinct classes:

- V-type tangle: open arcs connect points vertically (NW-SW and NE-SE points),
- H-type tangle: open arcs connect points horizontally (NW-NE and SW-SE points),
- X-type tangle: if open arcs connect points diagonally (NW-SE and SW-NE points).

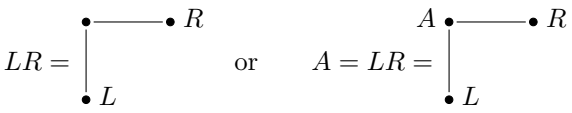
Definition 3 (Basic tangle)**.** The tangles 0, ∞, 1, $\overline{1} \in \mathcal{T}_{\mathbb{1}}$ depicted in Figure 1b are called *basic tangles*.

Basic tangles are building blocks for constructing more complex tangles, which can be formed by joining simpler ones through the identification of arc boundaries.

Definition 4 (Sum and product of tangles)**.** The *sum* $A+B$ and *product* AB of two tangles A and B are the operations depicted on Figure 1c.

Definition 5 (Algebraic tangle)**.** An *algebraic tangle* is a tangle $A \in \mathcal{T}_{\mathbb{A}}$ which can be represented as any composition of sums abd products of basic tangles.

Since every algebraic tangle can be expressed as a product of two subtangles $A = LR$, we can represent algebraic tangles as binary expression trees:



As a tangle can be represented in many different ways, our first goal is to determine whether two algebraic tangles are the same or different. We define "sameness" in two distinct ways: as isotopic or equivalent, as outlined below.

Definition 6 (Isotopy)**.** Two tangles A and B are *isotopic* if there exists an orientation-preserving self-homeomorphism $h : (B^3, A) \to (B^3, B)$ that is the identity map on the boundary ($h|_{\partial B^3} = \mathrm{Id}$). By slight abuse of notation, we will consider two isotopic tangles equal and will use the notation $A = B$.

Definition 7 (Equivalence)**.** Two tangles A and B are *equivalent*, denoted by $A \sim B$, if there exists an self-homeomorphism $h : (B^3, A) \to (B^3, B)$.

Note that equivalence is weaker than isotopy, since equivalence does not require the four endpoints NW, NE, SE, and SW to remain fixed.

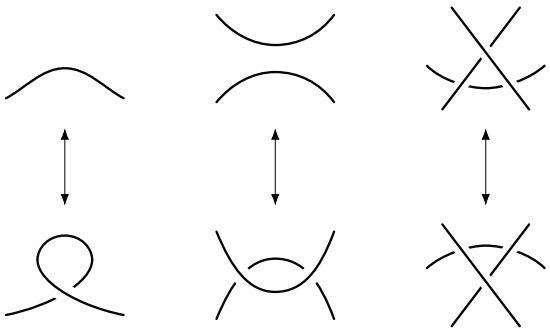


Fig. 2 Reidemeister moves of Type I, II, and III.

Definition 8 (Tangle diagram)**.** A *diagram* of a tangle is the regular projection of the tangle to the disk D obtained by the intersection of B^3 and the plane $z = 0$. By a general position argument, we will assume that the only singularities are a finite collection of double points, called *crossings*, which hold also over/under information about the projection.

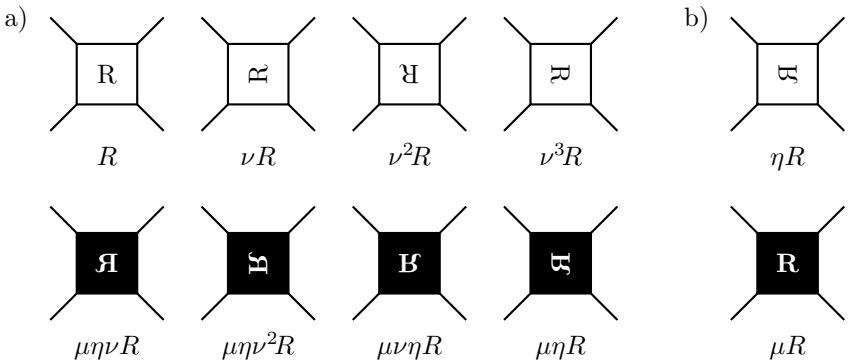


Fig. 3 Tangle R transformed by rotations and reflections represented by a two-faced square with letter "R". a) Set of all tangles equivalent to tangle R ordered by rotations: ν (horizontally), ρ_y (vertically). b) Reflections of tangle R: $\eta R \equiv R0$, and $\mu R \equiv \overline{R}$.

As in the case of knots, isotopy is generated by planar isotopy of Int D and the three Reidemeister moves depicted in Figure 2.

Theorem 1 ([1])**.** *Two tangles are isotopic if and only if their diagrams differ by planar isotopy and a finite sequence of Reidemeister moves.*

Naturally, we can generate new tangles by reflecting or rotating existing ones (see Figure 3). We will study the symmetry group of each tangle by examining its invariance under the following operations.

Definition 9 (Rotations and reflections)**.** The transformation μ is the reflection through the xy plane, η is the reflection through the zp plane, ρ_i are rotations around the axis $i \in \{x, y, z\}$ by angle π, and ν is the counterclockwise rotation around the z axis by angle $\pi/2$. Explicitly:

$$\begin{aligned} &\mu : (x, y, z) \to (x, y, -z)\,, && \rho_x = \mu\nu\eta : (x, y, z) \to (x, -y, -z)\,, \\ &\eta : (x, y, z) \to (-y, -x, z)\,, && \rho_y = \mu\eta\nu : (x, y, z) \to (-x, y, -z)\,, \\ &\nu : (x, y, z) \to (-y, x, z)\,, && \rho_z = \nu^2 \quad : (x, y, z) \to (-x, -y, z)\,. \end{aligned}$$

3 Notation and isotopy preserving moves

We use the following special notations for the reflection operators μ and η:

$$\mu(A) \equiv \overline{A} \equiv -A, \qquad \eta(A) \equiv A0. \tag{1}$$

It is easy to check graphically that rotation ρ_x is additive and rotation ρ_y is antiadditive in the following sense:

$$[A+B]_x = A_x + B_x, \qquad [A+B]_y = B_y + A_y. \tag{2}$$

It is also easy to check graphically that $\rho_x \eta = \eta \rho_y$, from where we obtain the following rules for multiplication:

$$[A0]_x = A_y 0, \qquad [A0]_y = A_x 0, \tag{3}$$

$$[AB]_x = A_y B_x, \qquad [AB]_y = (B_y 0)(A_x 0). \tag{4}$$

Any tangle can be represented using sum and product operations (Definition 4), and these operations can be used interchangeably, each with its own strengths and weaknesses. These operations are related through the following equalities:

$$A{+}B = A0B, \qquad AB = A0{+}B. \tag{5}$$

Addition can be fully expressed by multiplication, but the contrary is not true. The fact that 0 is the neutral tangle under addition translates in multiplicative form as follows:

$$0{+}A = A = A{+}0 \quad \Longleftrightarrow \quad 00A = A = A00 \tag{6}$$

Associativity of addition translates to the following "bracket juggling" multiplication rule:

$$(A{+}B){+}C = A{+}(B{+}C) \quad \Longleftrightarrow \quad (A0B)0C = A0(B0C), \tag{7}$$

and more general rule from the multiplication representation perspective:

$$(A0{+}B){+}C = A0{+}(B{+}C) \quad \Longleftrightarrow \quad AB0C = A(B0C), \tag{8}$$

which are exceptions to the left-associativity of tangle multiplication: $ABC := ((AB)C)$.

Definition 10 (Elementary moves)**.** Elementary moves I and II (Figure 4) are two tangles moves which are directly translated from I and II Reidemeister moves (Theorem 1). Algebraically, we represent them as follows (note that $00 = \infty$):

$$\infty{+}1 \overset{\text{I}}{=} \infty \quad \Longleftrightarrow \quad 01 \overset{\text{I}}{=} 00$$

$$1 + \overline{1} \overset{\text{II}}{=} 0 \overset{\text{II}}{=} 1{+}\overline{1}, \quad \Longleftrightarrow \quad 10\overline{1} \overset{\text{II}}{=} 0 \overset{\text{II}}{=} \overline{1}01.$$

If we have a sum of 1 and $\overline{1}$ tangles, we can remove any pairs of neighboring 1 and $\overline{1}$ tangles due to the IInd Reidemeister move and transform it to a tangle consisting only of 1's or $\overline{1}$'s. We call such tangles integral tangles.

Definition 11 (Integral tangle)**.** An *integral tangle* $n \in \mathcal{T}_{\mathbb{Z}}$ represented by an integer is a tangle created by sum of 1 and -1 tangles as follows:

$$n = \sum_{i=1}^{n} 1 = \underbrace{1{+}\,1{+}\cdots{+}1}_{n} \overset{(7)}{=} 1010\cdots01; \qquad -n = \sum_{i=1}^{n} \overline{1} = \underbrace{\overline{1}{+}\,\overline{1}{+}\cdots{+}\overline{1}}_{n} \overset{(7)}{=} \overline{1010\cdots01}$$

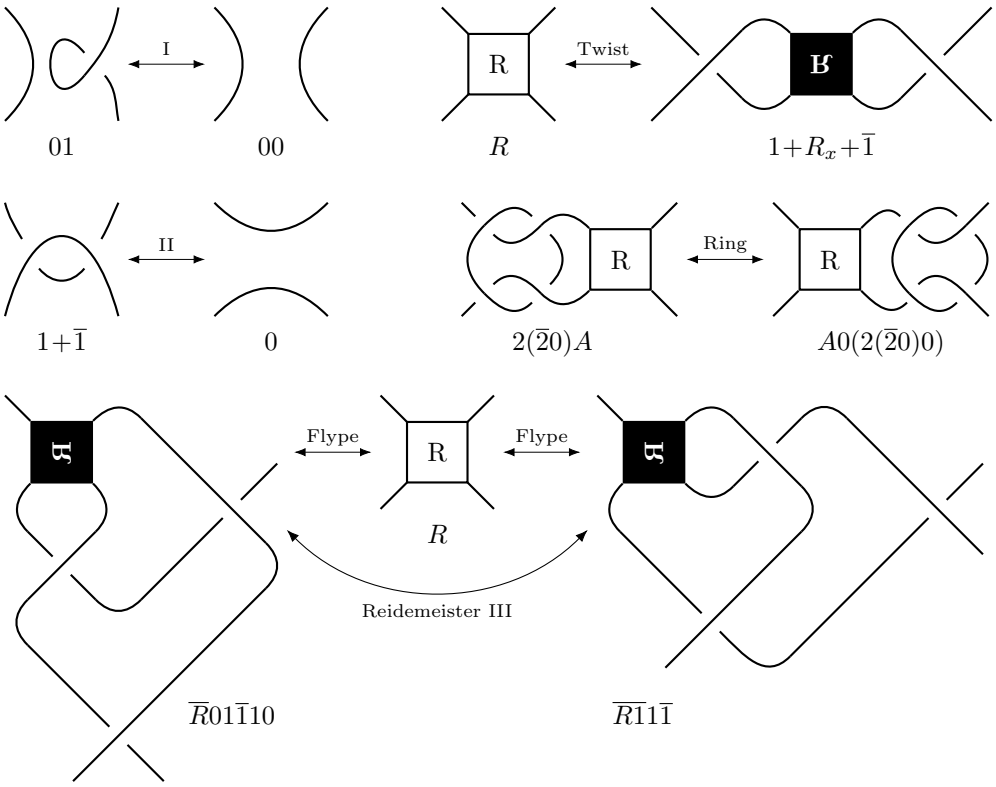


Fig. 4 Tangle moves: elementary moves I and II (special cases of Reidemeister I and II moves), twist, ring and flype moves. Two presented flypes are related by Reidemeister III move.

Next, we define three higher-order moves, namely the *twist*, the *flype*, and the *ring move* (Figure 4), which can be expressed as sequences of Reidemeister II and III moves. As we will see, these moves have convenient algebraic and geometric interpretations.

Definition 12 (Twist)**.** The *twist move* is a isotopy move obtained by rotating a subtangle around its x-axis by π. Algebraically, the twist can be formulated as:

$$1 + A_x + \overline{1} \overset{\mathrm{T}}{=} A \overset{\mathrm{T}}{=} \overline{1} + A_x + 1 \quad \Longleftrightarrow \quad 10A_x0\overline{1} \overset{\mathrm{T}}{=} A \overset{\mathrm{T}}{=} \overline{1}0A_x01.$$

In practice, we will use twist move to move integral tangles from one side of the tangle to the other:

$$n + A \overset{\mathrm{T}}{=} A_{x^n} + n \quad \Longleftrightarrow \quad n0A \overset{\mathrm{T}}{=} A_{x^n}0n,$$

where R_{x^n} denotes the tangle R to be rotated around the x-axis n times, i.e. $R_{x^n} = R_x$ if n is odd and $R_{x^n} = R$ if n is even.

Definition 13 (Flype)**.** A *flype move* is a tangle isotopy move obtained by the rotation of a subtangle around its principal diagonal p by π. There are 4 formulas for flypes, two for each direction of rotations (right- and left-handed). In both of them, the SE

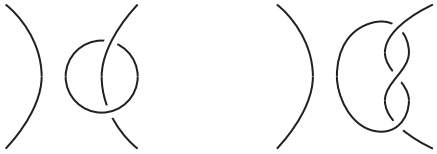


Fig. 5 Examples of non-prime tangles 0(20) and 0(30) (respectively).

arc can arrange itself in two different ways related by Reidemeister III move:

$$A \overset{\text{F}}{=} \overline{A1}1\bar{1} \overset{\text{III}}{=} \overline{A}01\bar{1}10 \overset{\text{F}}{=} A \overset{\text{F}}{=} \overline{A01}1\bar{1}0 \overset{\text{III}}{=} \overline{A}1\bar{1}1 = A.$$

Now, using flype, we can show that $1 = 10$ (with a consequence that $1n = n{+}1$):

$$1 \overset{\text{F}}{=} \bar{1}01\bar{1}10 \overset{\text{II}}{=} 0\bar{1}10 \overset{\text{I}}{=} 0010 = 10 \tag{9}$$

Definition 14 (Ring move)**.** The *ring move* (Figure 4) is a tangle isotopy move obtained by pushing the ring over a neighboring subtangle:

$$(20 + \bar{2}0)0 + A \overset{\circ}{=} A + (20{+}\bar{2}0)0 \quad \Longleftrightarrow \quad 2(\bar{2}0)A \overset{\circ}{=} A0(2(\bar{2}0)0).$$

For the tangles $2(20)\bar{1}A$ (and $\overline{2(20)}1A$) the a possibility of performing the ring move is hidden, but combination of flypes and twists reveals it:

$$\overline{2(20)}1A \overset{\text{F}}{=} 2(2\bar{1})10A \overset{\text{F}}{=} 2(\bar{2}0)10A \overset{\text{T}}{=} 2(\bar{2}0)A_x01 \overset{\circ}{=} A_x0(2(\bar{2}0)0)01. \tag{10}$$

Definition 15 (Prime tangle)**.** A tangle T in the ball B^3 is called prime (or non-composite) if there does not exist a 2-sphere $S^2 \subset B^3$ that intersects T transversely in exactly two points, such that the part of T inside of the S^2 sphere is not trivial, i.e. if we close the inner part of T by a geodesic on S^2, we do not obtain the unknot, see examples in the Figure 5.

A subtangle inside a 2-sphere (from definition above) can be moved anywhere along the the arc. Every tangle of the form $0A$, where A is a tangle which cannot be reduced into 0, is a composite tangle. Essentially, $0A$ is tangle A with connected NW and SW ends with an additional open arc. A simple example of such non-prime tangle is a $0(2n)$ tangle, where n is any integral tangle:

$$0(2n) = 0(20)0n \overset{\text{T}}{=} 0n0(20) \overset{\text{I}}{=} 0(20) \tag{11}$$

In the case of rational tangles, it was shown in [28] that the twist, flype, and IInd elementary move generate isotopy. In the case of prime algebraic tangles, we state the following conjecture:

Conjecture 1. *Two prime algebraic tangles are isotopic if and only if they are related by a finite sequence of twist, flype, ring, and elementary moves.*

4 An algebraic tangle as a tree

4.1 Tree notation

Due to the fact that the multiplication is left-associative, left factors and right factors behave differently. This clearly affects the behavior of tangles; however, for more complex tangles, the abundance of brackets makes the algebraic representation difficult to read. In such cases, the expression tree representation shows its advantage. For example, every algebraic tangle A can be represented as a product of two subtangles, $A = LR$, where both subtangles can similarly be decomposed into a product of sub-subtangles. This process can be repeated indefinitely (in fact, ad infinitum, since $1 = 10$):

$$A = LR = L_L R_L (L_R R_R) = L_{L_L} R_{L_L} (L_{R_L} R_{R_L}) \big(L_{L_R} R_{L_R} (L_{R_R} R_{R_R})\big) = \cdots \tag{12}$$

For clarity, one can represent such algebraic tangle as a binary expression tree, see Figure 6.

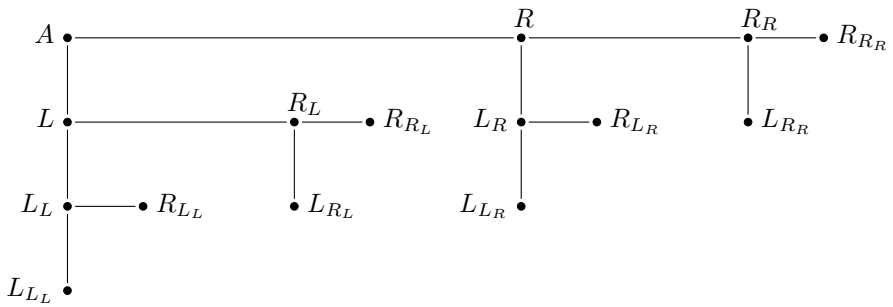


Fig. 6 Tangle $A = L_{L_L} R_{L_L} (L_{R_L} R_{R_L}) \big(L_{L_R} R_{L_R} (L_{R_R} R_{R_R})\big)$ represented as an expression tree. Every node represents a tangle which is a product of its left-child by its right-child.

In such an expression tree, every node represents some subtangle of tangle A and a product of its children. A node with no children, resulting from stopping the division of the tree at some point, is called a *leaf*.

If we expand only left children at each level, our representation stays clearer and we can write the tangle concisely in the multiplication representation:

$$A = LR = L_L R_L R = L_{L_L} R_{L_L} R_L R = L_{L_{L_L}} R_{L_{L_L}} \cdots R_L R := L_N \prod_{i=N}^{1} R_i. \tag{13}$$

Note that if every R_i term represents an integral tangle, and L_N represent a rational tangle, then the whole tangle is rational.

An analogous situation occurs when we expand only the right children at each level, allowing us to concisely represent the tangle using an additive representation:

$$\begin{aligned} A = LR = L(L_R R_R) = L\big(L_R(L_{R_R} R_{R_R})\big) = L\Big(L_R\big(\cdots\big(L_{R_{R_R}} R_{R_{R_R}}\big)\cdots\big)\Big) = \\ = L0 + \Big(L_R 0 + \Big(\cdots + \big(L_{R_{R_R}} 0 + R_{R_{R_R}}\big)\cdots\Big)\Big) = \\ = L0 + L_R 0 + \cdots + L_{R_{R_R}} 0 + R_{R_{R_R}} := \left(\sum_{i=1}^{N} L_i 0\right) + R_N. \end{aligned} \tag{14}$$

In [1] Conway used special comma notation for such tangles if $R_N = n \in \mathcal{T}_{\mathbb{Z}}$: $(L_1, L_2, \cdots, L_N, +++/---)$, where the number of +/- symbols is equal to the value of n ("+" for positive, "−" for negative integrals).

Sometimes we want to represent only a part of an expression tree. In this case it may be important if the top of the subtree is a left-child or a right-child. It appears that tree-tops behave exactly as right-children. Therefore, we introduce some extra notation for edges presented in Figure 7.

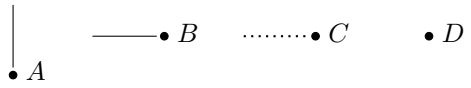


Fig. 7 Subtangle A is a left-child, subtangle B is a right-child, subtangle C is either a tree-top or a right-child, and subtangle D is either a child or not.

Definition 16 (Power notation)**.** We will use power notation A^B to emphasize that B is a subtangle of the tangle A. The subtangle B may be equal to A, a right-child of A, or a right-child of the right-child of A, and so on.

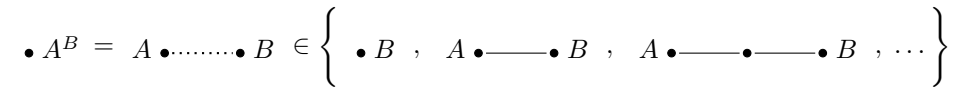


Fig. 8 Tree representation of power notation A^B. We use dotted-edge notation, to skip part of the tree. Note that A may be equal to B, then $A^B = B$.

Note that A may be equal to B, then $A^B = B$. Moreover, when A is an integral tangle $A = n$, then $A^B = n^B = B$.

4.2 Right leaves – integral tangles

Integral tangles (Figure 10a) appear to be ideal candidates for the right leaves of algebraic tangles, since it is possible to use a twist move to change their position. Moreover, they have high symmetry, since rotations around x, y, z axes do not affect

them: $n = n_x = n_y = n_z$. Indeed,

$$n_x = \left[\sum_{i=1}^{n} 1\right]_x = \sum_{i=1}^{n} 1_x = \sum_{i=1}^{n} 1 = n; \qquad n_y = \left[\sum_{i=1}^{n} 1\right]_y = \sum_{i=n}^{1} 1_y = \sum_{i=n}^{1} 1 = n. \quad (15)$$

It is convenient to represent each tangle A using power notation A^n to emphasize its rightmost integral tangle n. Let us show some properties of A^n notation.

Lemma 1. *An algebraic tangle A^n can be decomposed as $A^n = A^0 0 n$.*

Proof. If $A^n = n$, then the statement is trivial. Otherwise, let us expand the A^n as in Equation 14 and rearrange the brackets (due to associativity of addition):

$$\begin{aligned} A^n &= a_1\Big(a_2\big(\cdots(a_N n)\cdots\big)\Big) = \left(\sum_{i=1}^{N} a_i 0\right) + n = a_1 0 + \Big(a_2 0 + \big(\cdots + (a_N 0)\cdots\big)\Big) + n \\ &= a_1\Big(a_2\big(\cdots(a_N 0)\cdots\big)\Big) 0 n = A^0 0 n. \end{aligned}$$

□

Lemma 2. *A rightmost integral tangle n of tangle A^n can stay rightmost during rotations:*

$$[A^n]_x = [A_x]^n = A_x^n$$
$$[A^n]_y = \begin{cases} [A_y]^n, \text{ if } n \text{ is even}, \\ [A_z]^n, \text{ if } n \text{ is odd}, \end{cases}$$
$$[A^n]_z = \begin{cases} [A_x]^n, \text{ if } n \text{ is even}, \\ [A_z]^n, \text{ if } n \text{ is odd}. \end{cases}$$

Proof. We rotate the tangle A (Equation 2,3) after expanding its right children (Equation 14):

$$[A^n]_x = \left[\left(\sum_{i=1}^{N} a_i 0\right) + n\right]_x = \left(\sum_{i=1}^{N} [a_i]_y 0\right) + n = [A_x]^n = A_x^n,$$

$$[A^n]_y = \left[\left(\sum_{i=1}^{N} a_i 0\right) + n\right]_y = n + \sum_{i=N}^{1} [a_i]_x 0 \overset{\mathrm{T}}{=}$$

$$\overset{\mathrm{T}}{=} \begin{cases} \left(\sum_{i=N}^{1} [a_i]_x 0\right) + n, \text{ if } n \text{ is even}, \\ \left(\sum_{i=N}^{1} [a_i]_z 0\right) + n, \text{ if } n \text{ is odd}, \end{cases} = \begin{cases} [A_y]^n, \text{ if } n \text{ is even}, \\ [A_z]^n, \text{ if } n \text{ is odd}, \end{cases}$$

$$[A^n]_z = \left[\left(\sum_{i=1}^{N} a_i 0\right) + n\right]_z = n + \sum_{i=N}^{1} [a_i]_z 0 \overset{\mathrm{T}}{=}$$

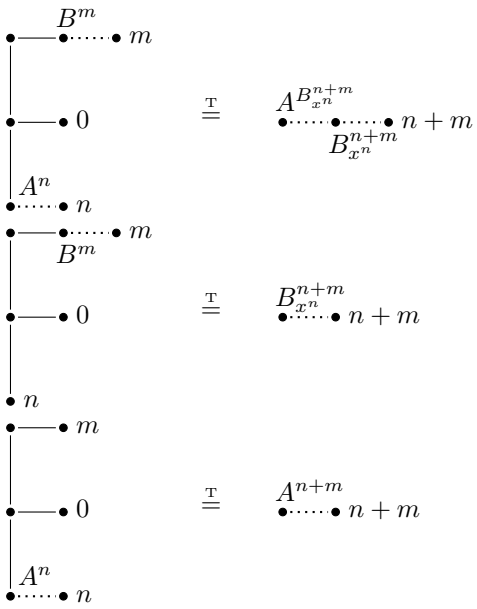


Fig. 9 Isotopy preserving move $A^n{+}B^m \overset{\mathrm{T}}{=} A^{B^{n+m}_{x^n}}$ represented on a tree. Note that if $A^n = n$ or $B^m = m$ the result simplifies to $B^{n+m}_{x^n}$ or A^{n+m} respectively.

$$\overset{\mathrm{T}}{=} \begin{cases} \left(\sum_{i=N}^{1}[a_i]_z 0\right) + n, \text{ if } n \text{ is even,} \\ \left(\sum_{i=N}^{1}[a_i]_x 0\right) + n, \text{ if } n \text{ is odd.} \end{cases} = \begin{cases} [A_x]^n, \text{ if } n \text{ is even,} \\ [A_z]^n, \text{ if } n \text{ is odd.} \end{cases}$$

As we see, rotation around y or z axis reverses order of left children, but the rightmost integral n can stay rightmost, at the expense of rotating left children around y axis if n is odd. □

Lemma 3. *If two tangles A^n and B^m are added together, $B^{n+m}_{x^n}$ becomes the right subtangle of A (Figure 9):*

$$A^n{+}B^m \overset{\mathrm{T}}{=} A^{B^{n+m}_{x^n}}.$$

Proof.

$$A^n{+}B^m = A^0{+}n{+}B^0{+}m \overset{\mathrm{T}}{=} A^0{+}B^0_{x^n}{+}n{+}m = A^0{+}B^{n+m}_{x^n} =$$
$$= \left(\sum_{i=1}^{N} a_i 0\right) + \left[\sum_{i=1}^{M} b_i(n+m)\right]_{x^n} = A^{B^{n+m}_{x^n}}.$$

□

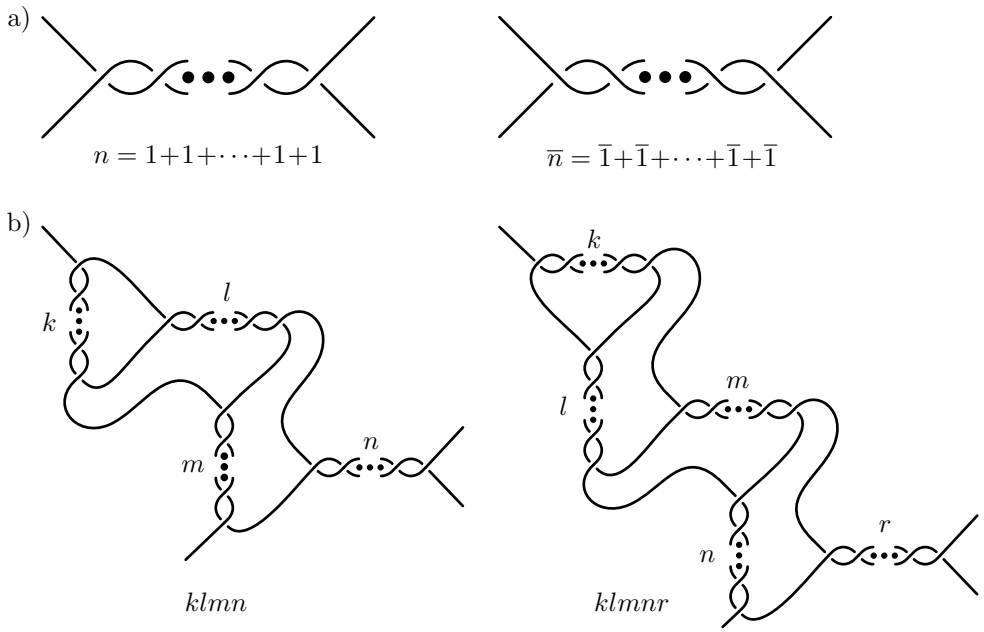


Fig. 10 a) Integral tangles n as a sum of 1 tangles, and $\overline{n}$ as sum of $\overline{1}$ tangles. v) Rational tangles $klmn$ and $klmnr$ in a standard form: as a multiplication of integral tangles k, l, m, n, and r.

4.3 Left leaves – rational tangles

Now we show that rational tangles are ideal candidates for left leaves. Let us first recall some known results for rational tangles, which have been thoroughly studied in [1, 28, 29].

Definition 17 (Rational tangle)**.** A *rational tangle* (Figure 10b) is a tangle $a \in \mathcal{T}_{\mathbb{Q}}$ that can be expressed as a multiplication of integral tangles: $kl \cdots mn$. If a rational tangle is written as such a product, we say that the rational tangle is in a *standard form*.

Like integral tangles, rational tangles are invariant to rotations around x, y, z axes, which can be shown by recurrent treatment:

$$[kl \cdots mn]_x = [kl \cdots m]_y n \tag{16}$$

$$[kl \cdots mn]_y = n0[kl \cdots m0]_y \overset{\mathrm{T}}{=} [kl \cdots m0]_{yx^n} 0n = [kl \cdots m]_{xy^n} n. \tag{17}$$

More importantly, there exists a perfect invariant of rational tangles – the continued fraction. Unlike many other knot invariants, the continued fraction uniquely determines a rational tangle, providing a simple method to distinguish them.

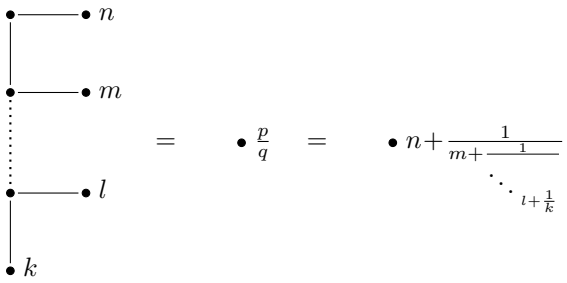


Fig. 11 Contraction of a rational tangle by replacing it with fraction of the rational tangle, represented by a move on a tree. On this figure dotted edge means 1 or more edges separated by nodes which right-children are integral tangles.

Definition 18. The fraction p/q, where $p, q \in \mathbb{Z}$ (including $1/0 = \infty$), of a rational tangle in a standard form $kl \cdots st$, is defined as the continued fraction:

$$\mathrm{Frac}(kl \cdots mn) = n + \cfrac{1}{m + \cfrac{1}{\ddots_{\,l+\frac{1}{k}}}} = p/q.$$

Theorem 2 (Conway [1])**.** *The fraction of a rational tangle is a perfect invariant – two rational tangles are isotopic if and only if they have the same fraction.*

The proofs can be found in [30], [29] p.196 and [31].

Note that all properties of rational tangles also apply to integral tangles, since integral tangles are a subset of rational tangles with an integer fraction.

Definition 19 (Canonical form)**.** Rational tangle is in *canonical form* when all of the following conditions hold:

- it is in standard form,
- all its terms are all positive or all negative except the last one, which can be 0,
- its first term is 1 or $\overline{1}$ only if it is the only term,
- its first term is 0 only if it is tangle 0 or 00.

Theorem 3 (Kauffman [28])**.** *Every rational tangle can be brought to a canonical form by isotopy.*

Theorem 4 (Kauffman [28])**.** *Every rational tangle has exactly one unique canonical form.*

Note that the canonical form, defined as in Definition 19, matches the output of Euclid's algorithm (cf. decomposition on Figure 12). Note that Kauffman used a slightly different definition of the canonical form but followed the same principles. The difference is that we used the relation $1 = 10$ to keep the minimal number of integral tangles, while Kauffman preferred to keep their number odd.

4.4 Fraction of an algebraic tangle

Here we generalize the fraction of rational tangles to the fraction of non-rational algebraic tangles, although, the generalized fraction loses the property of being a perfect invariant.

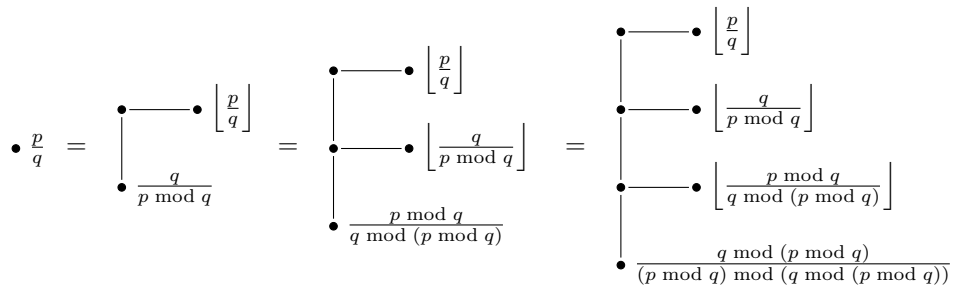


Fig. 12 Decomposition of a positive rational tangle represented with a tangle tree. Left-leaf can be recurrently decomposed until its fraction is integral. Note that the decomposition is equivalent to Euclid's algorithm of finding $\gcd(p, q)$ – consecutive generated integral tangles (right-children) are remainders of the operation.

Definition 20. The fraction p/q, where $p, q \in \mathbb{Z}$ (including $1/0 = \infty$), of a tangle $A = LR$, A,L,$R \in \mathcal{T}_{\mathbb{A}}$ is defined as follows:

$$\mathrm{Frac}(A) = \mathrm{Frac}(LR) = \mathrm{Frac}(R) + \frac{1}{\mathrm{Frac}(L)} = p/q.$$

Which, for chain-multiplication of tangles (Equation 13), coincides with continued fraction:

$$\mathrm{Frac}(A) = \mathrm{Frac}\left(\prod_{i=1}^{N} a_i\right) = \mathrm{Frac}(a_N) + \cfrac{1}{\mathrm{Frac}(a_{N-1}) + \cfrac{1}{\ddots_{\ \mathrm{Frac}(a_2) + \frac{1}{\mathrm{Frac}(a_1)}}}} = p/q, \quad (18)$$

and is linear with respect to addition (Equation 14):

$$\mathrm{Frac}(A) = \mathrm{Frac}\left(\sum_{i=1}^{N} a_i\right) = \sum_{i=1}^{N} \mathrm{Frac}(a_i) = p/q. \quad (19)$$

Lemma 4. *The fraction of an algebraic tangle is invariant to rotations around x, y, z axes.*

Proof. For a rotations of tangle $A = LR$ (Equation 3) we obtain:

$$\mathrm{Frac}(A_x) = \mathrm{Frac}(L_y R_x) = \mathrm{Frac}(R_x) + \frac{1}{\mathrm{Frac}(L_y)};$$
$$\mathrm{Frac}(A_y) = \mathrm{Frac}\big((R_y 0)(L_x 0)\big) = \frac{1}{\mathrm{Frac}(L_x)} + \frac{1}{1/\mathrm{Frac}(R_y)} = \mathrm{Frac}(R_y) + \frac{1}{\mathrm{Frac}(L_x)}.$$

Rotating a tangle around x or y axis does not change the expansion form of the fraction. After repeatedly expanding the tangle into rational or integral components, one can see that the structure of the rotated and non-rotated tangles remains the

same; only the rotation-indices are different. Since rational tangles are invariant under rotation, the indices can be ignored. □

Lemma 5. *Switching all signs of an algebraic tangle changes the sign of its fraction:* $\text{Frac}(\overline{A}) = -\text{Frac}(A)$, $A \in \mathcal{T}_{\mathbb{A}}$.

Proof. By similar argument as in Lemma 4. Since the statement is holds for rational tangles, it also holds for algebraic tangles. □

Lemma 6. *The fraction of an algebraic tangle is invariant to the elementary moves, flypes, twists, and ring moves.*

Proof. The proofs for elementary are trivial. We show proofs for a flype $A \overset{\text{F}}{=} \overline{A}1\overline{1}$, a twist $A \overset{\text{T}}{=} \overline{1}0A_x01$ a ring move, as all other cases follow analogously:

$$\begin{aligned}
\text{Frac}(\overline{A}1\overline{1}) &= -1 + \frac{1}{1 + \frac{1}{-1 + \frac{1}{\text{Frac}(\overline{A})}}} = -\text{Frac}(\overline{A}) = \text{Frac}(A). \\
\text{Frac}(10A_x0\overline{1}) &= -1 + \frac{1}{0 + \frac{1}{\text{Frac}(A_x) + \frac{1}{0+\frac{1}{1}}}} = \text{Frac}(A_x) = \text{Frac}(A). \\
\text{Frac}(A0(2(\overline{2}0)0)) &= \frac{1}{-\frac{1}{2} + \frac{1}{2}} + \frac{1}{0 + \frac{1}{\text{Frac}(A)}} = \text{Frac}(A) + \frac{1}{-\frac{1}{2} + \frac{1}{2}} = \text{Frac}(2(\overline{2}0)A)
\end{aligned}$$

□

Therefore, the generalized fraction is invariant to all isotopy-preserving moves which we defined in section 3, but it is not a perfect invariant, e.g. $2(20) \neq 1$ but $\text{Fraction}(2(20)) = \text{Fraction}(1)$.

4.5 Canonical representation

Here we define the canonical representation of algebraic tangles and show how to obtain it. The goal is to obtain a form in which twist, flype, ring and elementary moves are "fixed" in a specific position, and provide an algorithm of obtaining such a form.

While the canonical form of a rational tangle is alternating (see Definition 19 and Theorem 3), making it an ideal candidate, this property does not extend to all tangles. For instance, the ring $2(\overline{2}0)$ lacks any alternating form. Furthermore, prime non-rational tangles that do admit alternating forms have at least four such representations, with the number increasing rapidly as crossings grow. Thus, a unique representation must rely on other rules.

Definition 21 (Canonical representation of prime algebraic tangles)**.** A prime algebraic tangle A is in canonical representation when its tree $T(A)$ obeys following restrictions:

- on each leaf:
  - all right-leaves are integral tangles;
  - all left-leaves are positive rational tangles (represented by fractions) with fractions larger than 1;
- on each node, which is left-child:

  - it cannot be a parent of two leaves;
  - if its right-child is a leaf, then this leaf must be positive;
  - if its right-child is a 1 tangle, then the right-child of its left-child must be a leaf.
- on each node with a ring ($2(\overline{2}0)$, or in canonical form, $2(2\overline{1})$) as a left-child:
  - its right-child must be a leaf, or a parent of a ring as a left-child as well.

Forbidden subtress which follow from these rules are presented in Figure 13.

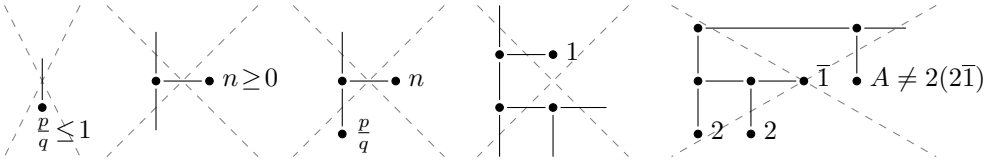


Fig. 13 Tangle trees in canonical representation must have left-leaves represented as rational tangles $\frac{p}{q}$, right-leaves represented as integral tangles n, and cannot have any subtangles are listed above.

Theorem 5. *The canonical form of a prime algebraic tangle is unique up to twist, flype, ring, and elementary moves*

We will proof the theorem by an algorithm (Algorithm 1) that puts a prime tangle into its canonical form and show that the algorithm returns the same result after performing twist, flype, ring and/or elementery moves.

To introduce the algorithm, we need to define ordering of tree elements, and the theory of binary trees already gives us such tools. We will define both a top-down ordering and a bottom-up ordering. The top-down ordering results from a pre-order traversal, e.g. for tree from Figure 6, the top-down ordering is $[A, L, L_L, L_{L_L}, R_{L_L}, R_L, L_{R_L}, R_{R_L}, R, L_R, L_{L_R}, R_{L_R}, R_R, L_{R_R}, R_{R_R}]$. The bottom-up ordering is the top-down ordering reversed.

Algorithm 1 transforms an algebraic tangle, represented by a binary tree, into its canonical representation (Lemma 9). In general, when the tree is traversed, and when a node is visited, possible isotopy preserving moves are performed on the node. The algorithm is divided into three parts, which differ how the tree is traversed and which moves $\alpha_{\mathbb{Z}}$, $\alpha_{\mathbb{Q}}$, $\beta_{\mathbb{Q}}$, β_0, β_-, β_1, γ, presented in Table 1, are possible to perform:

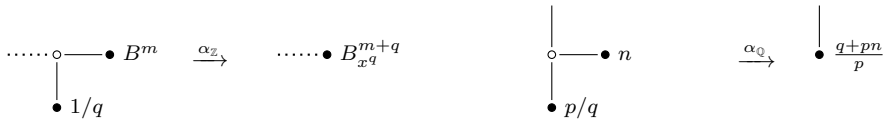
1. In the first part, we represent rational tangles in a possibly concise way and ensure that e.g., a subtangle 0 is represented as 0, not as $2(2\overline{1}0)1\overline{21}$. This prepares the algorithm for the second part to let it correctly recognize subtree patterns. At the same time, the twists are performed to keep twistable integral tangles as right leaves. The tree is traversed bottom-up once, and each node $\alpha_{\mathbb{Z}}$ (twist: $q{+}B \overset{\mathrm{T}}{=} B_{x^q}{+}q$) and $\alpha_{\mathbb{Q}}$ (contraction: Figure 11) moves are performed if possible.
2. In the second part, we perform the flypes required to follow the canonical representation. The tree is traversed top-down, and each time a move is performed, the traversal starts again from the top. The move β_0 (Lemma 3) is equivalent to bracket juggling and twisting integral tangles, to remove 0 subtangles when possible. The moves $\beta_{\mathbb{Q}}$, β_-, β_1 are flypes followed by twists. Sometimes both β_0 and β_- could be performed, but β_0 has higher priority, to ensure that $A^n 0 \overline{k} B^m$

Algorithm 1 Algorithm operates on objects N of a NODE class. Every NODE object has attributes which point to its *parent*, *left_child* and *right_child* and knows if it *is_a_left_child* of its *parent*. If a NODE represents a rational tangle, then it has a non-`None` value of a *fraction*. FIXLOCAL performs one of $\alpha_{\mathbb{Z}}$, $\alpha_{\mathbb{Q}}$, $\beta_{\mathbb{Q}}$, β_0, β_-, β_1, γ moves on a node N if it is possible (moves are presented in Table 1).

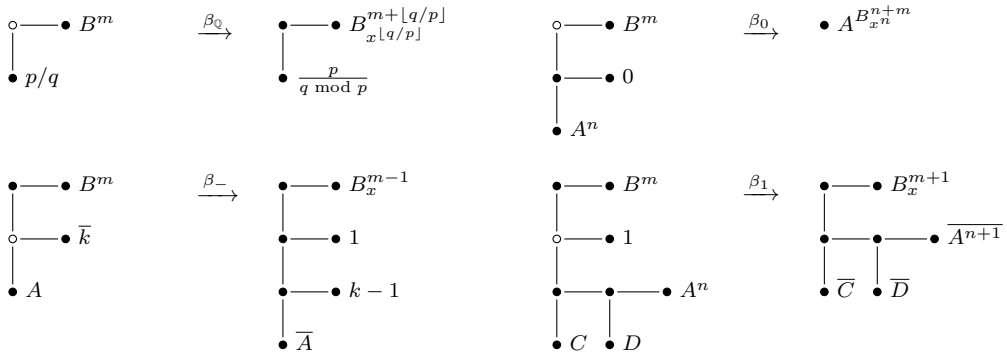
```
1: function GetCanonicalTree(T)                    ▷ T is address to a tree-top Node
2:     nodes ← Reversed(PreOrderTraversal(T))                          ▷ 1st part
3:     for N in nodes do
4:         FixLocal(N,[α_ℤ, α_ℚ])
5:     end for
6:     repeat                                                          ▷ 2nd part
7:         changed ← False
8:         nodes ← PreOrderTraversal(T)
9:         for N in nodes do
10:            changed ← FixLocal(N,[β_ℚ, β_0, β_−, β_1])
11:            if changed = True then
12:                break the for loop
13:            end if
14:        end for
15:    until changed = False
16:    nodes ← Reversed(PreOrderTraversal(T))                          ▷ 3rd part
17:    for N in nodes do
18:        FixLocal(N,[γ])
19:    end for
20:    return T
21: end function

22: function FixLocal(N, moves)
23:    for move in moves do
24:        if move on node N is possible then
25:            perform move on a node N
26:            return True
27:        end if
28:    end for
29:    return False
30: end function

31: function PreOrderTraversal(N)
32:    if N.fraction = None then
33:        return [N]
34:    else
35:        l ← PreOrderTraversal(N.left_child)
36:        r ← PreOrderTraversal(N.right_child)
37:        return joined lists l and r
38:    end if
39: end function
```

Table 1 Possible moves for traversals of the tree-fix algorithm. Empty dot represent a node visited by an algorithm. m and n represent integral tangles, while k represents strictly positive integral tangle.

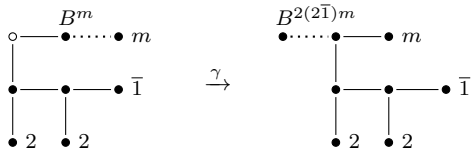
Part 1. moves:



Part 2. moves:



Part 3. move:



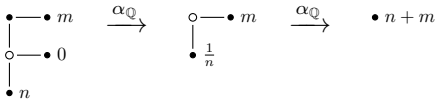
is twisted into $A^{n-k}B^m$, instead of being flyped into $\overline{A}1B^{m-1}_x$. The algorithm follows "twists before flypes" rule. The priority couldn't be ensured by adding β_0 to the first part of the algorithm, since 0-right leaf can appear as a result of both β_- and β_1 moves.

3. After the second part, the only missing thing is correct positioning of rings $2(\overline{2}0)$ (which in canonical representation are represented by $2(2\overline{1})$). Their position is fixed in the third part, by bottom-up traversal and the γ move.

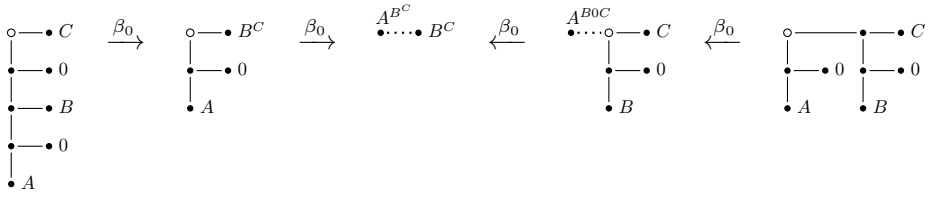
Lemma 7. *Algorithm 1 gives the same output for tangles in a tree form that differ by a twist, ring, elementary moves, or bracket juggling.*

Proof. The ring move switches horizontally two neighboring subtrees (one of which is a ring $(2(2\overline{1}))$). The third part of the algorithm "pushes" all rings to the right, fixing positions of all $2(2\overline{1})$ subtangles.

The first part of the algorithm collects bottom-up rational tangle subtangles into a single fraction, which ensures that every rational tangle is in its simplest form before the second part of the algorithm.



The second part of the algorithm returns the same output for both $A0B0C$ and $A0(B0C)$ tangles. Anything that happens to tangle A^{B0C} before the traversal reaches $B0C$, has the same impact on the A^{B^C} tangle. Note that if the tangle A is integral, then both $A0B0C$ and $A0(B0C)$ tangles become identical earlier after the first traversal. Twist moves, bracket juggling, and the II elementary move are special cases of this example.



Rotations were omitted for a clarity of the representation.

□

Lemma 8. *Algorithm 1 gives the same output for tangles in a tree form that differ by flype moves.*

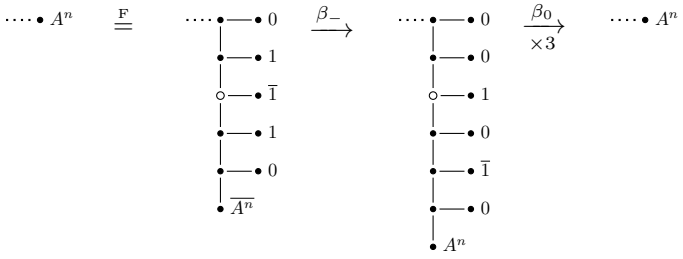
Proof. We will check all possible positions of a tangle A in a tree, perform a flype at each position, and show that algorithm always returns the same output. These positions can be divided into 4 cases:

1. A is a right-child/tree-top;
2. A is a left-child of a right-child/tree-top;
3. A is a left-child of a left-child of a right-child/tree-top;
4. A is a left-child of a left-child of a left-child, which collects all other cases.

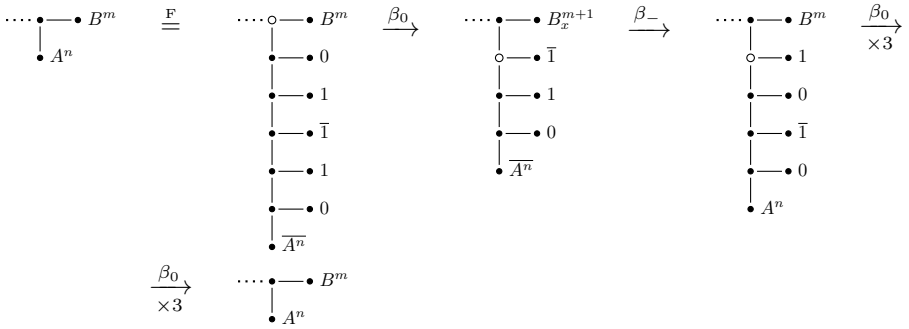
In all cases, A is a non-rational tangle. If A is a rational tangle, then it is represented by a fraction after a the first part of the algorithm, which is invariant to flypes.

There are 4 types of flypes; we will provide a proof for only one of them ($A \overset{\text{F}}{=} \overline{A}01\bar{1}10$), as proofs for other 3 flypes are analogous or simpler.

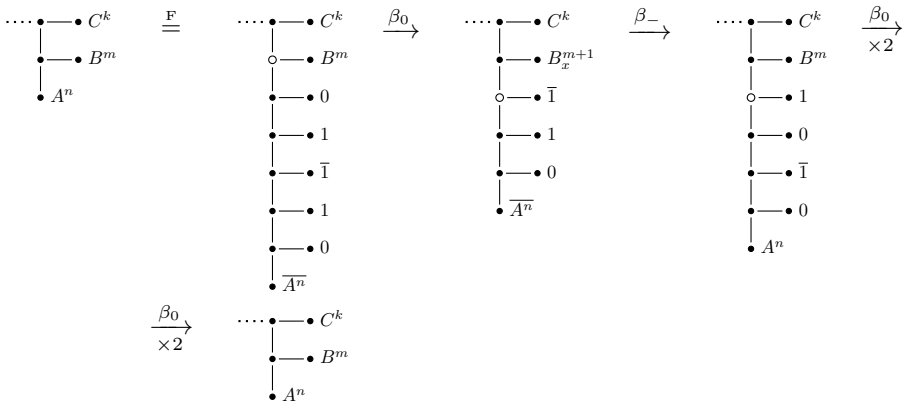
Case 1. A is a right-child/tree top.



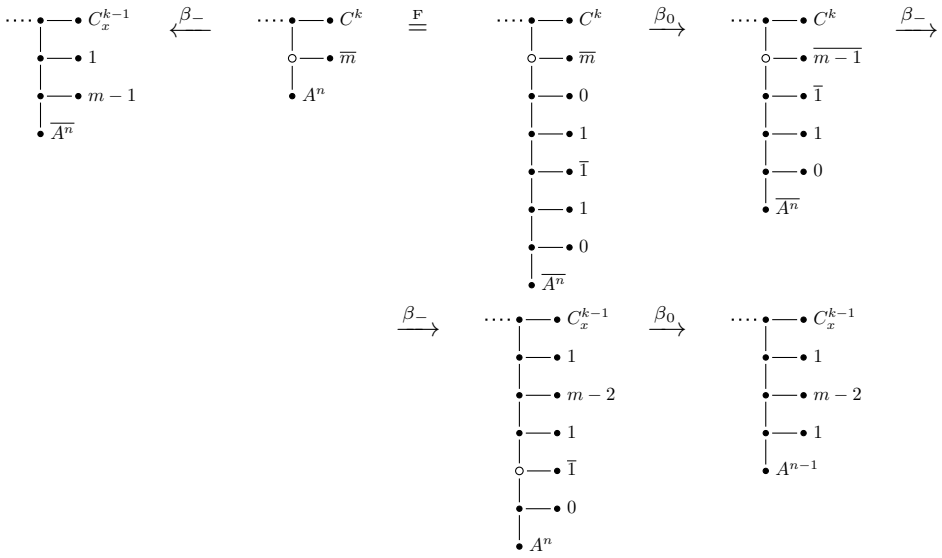
Case 2. A is a left-child of a right-child/tree top (B).



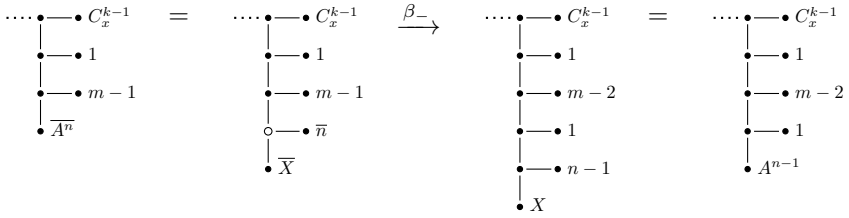
Case 3. A is a left child of a left-child (B) of a right-child/tree-top (C). There are five distinct subcases:
3.1) $B \notin \{0, -1, -2, \cdots\}$;
3.2) $B = \overline{m}$; $m \geq 3$;
  3.2.1) $A = Xn$; $n \in \{1, 2, \cdots\}$; $X \notin \mathcal{T}_{\mathbb{Q}}$;
  3.2.2) $A = X\overline{n}$; $n \in \{0, 1, 2, \cdots\}$; $X \notin \mathcal{T}_{\mathbb{Q}}$;
  3.2.3) $A = X(YZ)$; $YZ \notin \mathcal{T}_{\mathbb{Z}}$;
  3.2.4) $A = X0n$; $n \in \mathcal{T}_{\mathbb{Z}}$; $n \notin \mathcal{T}_{\mathbb{Z}}$;
3.3) $B = \overline{2}$;
3.4) $B = \overline{1}$;
3.5) $B = 0$.
We will prove subcases 3.1 and 3.2. Subcase 3.3 is analogous to 3.2, and subcases 3.4, 3.5 are simpler. In subcase 3.2.4, β_0 move is performed 0 from $A = X0n$, and we obtain one of other 3.2 cases.
Case 3.1. $B \notin \{0, -1, -2, \cdots\}$:



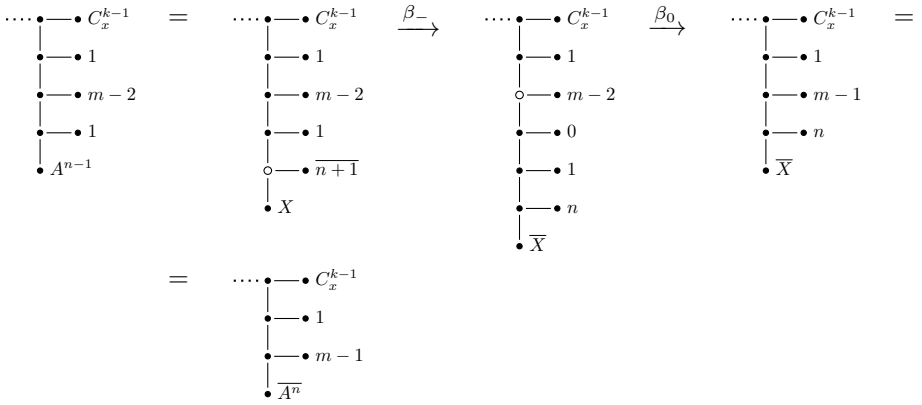
Case 3.2. $B = \overline{m}$; $m \geq 3$. In this case, the results do not match, and we need to consider further specific cases of A subtangle (Cases 3.2.1-3).



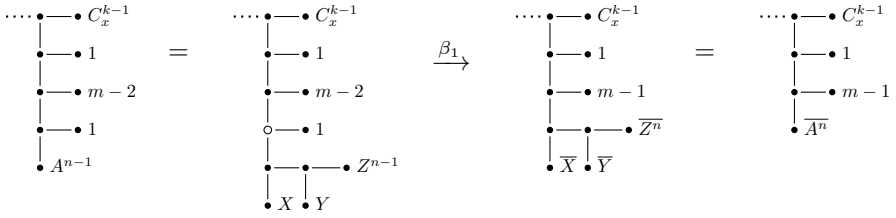
Case 3.2.1. Let the right-child of A be a positive integral tangle $A^n = Xn$. We show that algorithm acting on a non-flyped result of case 3.2 gives the same result as acting on the flyped result.



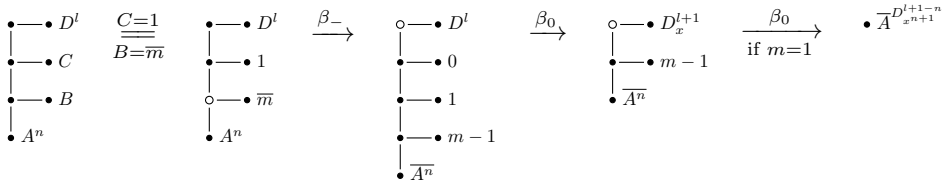
Case 3.2.2. Let the right-child of A be a non-positive integral tangle $A^{\overline{n}} = X\overline{n}$. We show that the algorithm acting on a flyped result of case 3.2 gives the same result as acting on the non-flyped result.



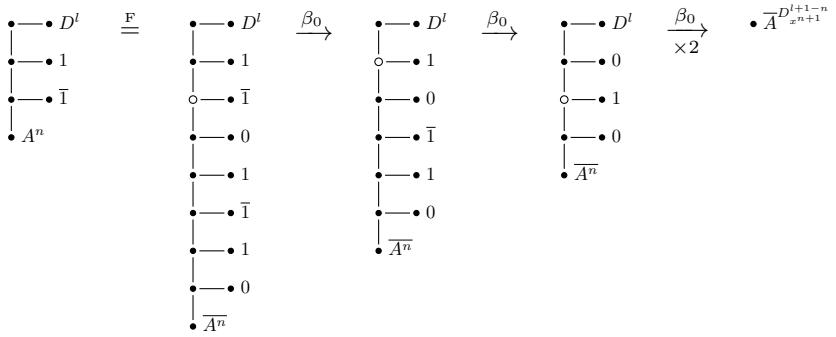
Case 3.2.3. Let the right-child of A be a non-rational tangle $A^n = X(YZ^n)$. We show that the algorithm acting on the non-flyped result of case 3.2 gives the same result as acting on the flyped result.



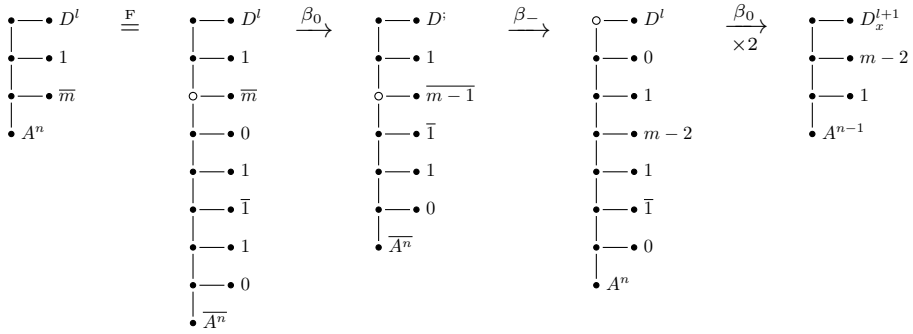
Case 4. In this case the algorithm already reaches a C^k subtangle, meaning that both C and B are not the 0 tangle, and C cannot be a negative integral tangle. In most cases the output of algorithm is the same as in case 3), with the exception when $C = 1$ and $B = \overline{m}$, $m > 0$, since C^k will become $C_x^{k-1} = 0$ after the β_- move will act on B. In this case, the algorithm performs a β_0 move on C's parent (D).



If we flype subtangle A, there are 3 subcases to consider:
4.1) $m = 1$;
4.2) $m = 2$;
4.3) $m \geq 3$;
We do not evaluate subcase 4.2, as it is analogous to 4.3.
Case 4.1. $C = 1$, $B = \overline{1}$.



Case 4.3. $C = 1$, $B = \overline{m}$, $m \geq 3$.



Results of subcases 4.2 and 4.3 after the flype differ from the result of case 4 (without flyping), but in the same way as in the subcases 3.2. and 3.3, which were shown to finally give the same result regardless if the flype was performed or not. □

Lemma 9. *Algorithm 1 transforms an algebraic tangle into its canonical representation while preserving its isotopy class.*

Proof. All restricted subtangle states (Figure 13) are handeled by corresponding algorithm moves (Table 1) which correct the subtree from the restricted state. The algorithm terminates when the tangle reaches its canonical form. It is finite: the first and third parts both have only one iteration; the second part of algorithm can require multiple traversals, however, with (nearly) every iteration the traversal goes further, the only exception (case 4 in Lemma 8) can occur only a finite number of times, since the tree is finite. Therefore, the algorithm terminates. □

We are now ready to prove Theorem 5.

Proof of the Theorem 5. Lemma 9 shows Algorithm 1 transforms a prime tangle into its canonical form. Lemma 7 and Lemma 8 state that the algorithm is invariant to twists, flypes, ring and elementary moves . □

5 Classification of prime tangles

In addition to classifying tangles up to isotopy and equivalence, we also classify them by their orbit class.

Definition 22 (Tangle orbit)**.** The *orbit of a tangle* is a set of tangles generated by all possible compositions of μ, ν, η acting on the tangle.

Note that μ, η, ν are generators of the $\mathbf{D}_8 \times \mathbf{Z}_2$ group. This topic is extend in section 6.

The classifications of tangles up to 14 crossings and up to isotopy, equivalence, and orbit classes, is presented in Table 2. The classification of tangles up to 10 crossings up to the orbit classes (and their diagrams) can be found in the supplementary material [32].

5.1 Generating tangles and the classification methods

In this subsection we describe how we generate tangles and what methods we used to classify them (generate the tangle tables).

Table 2 Number of all distinct tangle orbits, divided into subgroups by number of crossings ("#cross") and number of extra closed components ("#closed components"). Tangles in each column/row are added up ("Orbit" column/row). Additionally tangles are counted and added up to equivalence ("Equiv.") and isotopy ("Isotopy").

#cross	#closed components							Total		
	0	1	2	3	4	5	6	Orbit	Equiv.	Isotopy
0	1	-	-	-	-	-	-	1	1	2
1	1	-	-	-	-	-	-	1	1	2
2	1	-	-	-	-	-	-	1	2	4
3	2	-	-	-	-	-	-	2	4	8
4	4	2	-	-	-	-	-	6	11	22
5	11	3	-	-	-	-	-	14	28	68
6	30	12	2	-	-	-	-	44	87	236
7	86	46	4	-	-	-	-	136	270	880
8	267	152	37	3	-	-	-	459	912	3442
9	844	608	129	4	-	-	-	1585	3168	13900
10	2910	2202	611	81	3	-	-	5807	11595	57488
11	10102	8742	2803	286	5	-	-	21938	43864	242206
12	37115	34853	11913	2021	178	4	-	86084	172091	1035696
13	139778	141102	56073	9831	573	5	-	347362	694690	4482956
14	539872	592627	249371	50979	5961	328	4	1439142	2877981	19602964
Orbit	731024	780349	320943	63205	6720	337	4	1902582		
Equiv.	1461922	1560597	641753	126347	13408	670	8		3804705	
Isotopy	10192496	10480208	3994874	705590	64048	2642	16			25439874

Generation of alternating tangles

The initial set of all alternating algebraic tangles up to N crossings is generated using a standard integer partition algorithm, that is, generate sequences $(p_i \in \mathbb{N}_0)_i$, satisfying $\sum_i p_i = N$, without consecutive zeroes, and nest the partitions in all possible ways. Using integer partitions, our goal is to generate all tangles up to N crossings and up

to μ, η reflections. These reflected tangles can be obtained by changing signs and/or by adding 0 at the end of existing tangles (see (1)). Such generated non-negative partitions represent alternating tangles in multiplication form. Not all non-negative partitions are needed, since many representations are redundant:

- zeroes can appear only at the end of the bracket – zeroes at the beginning of bracket generate non-prime tangles, which we omit in the classification, and zeroes inside the bracket are redundant: $n0m = (n+m)$.
- consecutive opening brackets are redundant (left associativity of multiplication): $((A \cdots B)C \cdots) = (A \cdots BC \cdots)$
- first integral tangle after the opening bracket is ambiguous: $(1 \cdots) = (10 \cdots) \implies (1n \cdots) = ((n+1) \cdots)$.

Next, we remove tangles that end with 0. At this stage, 1,527,810 tangles up to 14 crossings are obtained.

Generation of non-alternating tangles

Tangles containing 0 at the end of any bracket are used as a template for generation of non-alternating tangles. The 0's inside tangles are then replaced with negative integers $-1, \ldots, -k$, where k is the level of nestedness, defined as the number of closing brackets to the right of the 0. At this stage, 25,267,083 tangles are obtained.

Determination of tangles' minimal set and symmetries

The main algorithm translates each tangle into a binary tree. We perform the following set of operations on each tangle in a binary tree form:

- generation of the remaining 15 tangles in each the tangle's orbit,
- transforming all 16 tangles into their canonical forms, and determination of a tangle symmetry group (by equality of canonical forms),
- all 16 tangles are minimized to determine the number of crossings in their minimal representations, here the minimization only of one of the orbits' representative is sufficient.

If a tangle is found in the orbit of any other tangle, then it is rejected. In each orbit, we choose the lexicographically minimal tangle as the representative. Finally, we obtain 1,902,582 unique tangle representatives and their symmetry groups, which translates into 3,804,705 unique tangles up to equivalence, and 25,439,874 unique tangles up to isotopy.

Minimization algorithm

In order to find the tangle form with a minimal number of crossings, we first modify the tangle in its canonical form by twisting the negative rightmost tangles, to incorporate them into rational tangles (this way the number of crossings of each rational tangle does not change, only their sign, but the number of crossings of the rightmost integral tangle is lowered). Next, we search for possible flypes that could lower the number of crossings (between two right leaves with opposite signs, which are neighbors or separated by one 0 right-leaf) and perform these flypes on separate copies of the tangle. Then we recursively repeat this process until no possibilities are left.

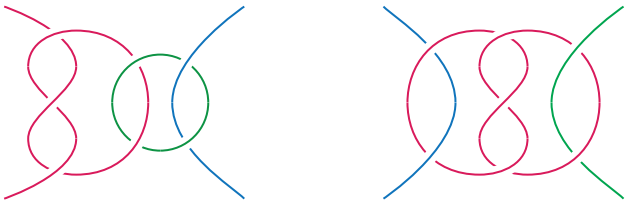


Fig. 14 Example of mutant tangles: left 3(2(20)), right: 2(3(20)).

Table 3 4 tangles missing from classification [13] are mutants of other classified tangles. Moriuchi used Conway's comma notation $(A, B, C) = A0{+}B0{+}C0 = A(B(C0))$.

In Moriuchi's table		Missing mutant tangle	
our notation	Conway's notation	our notation	Conway's notation
$3(2(20))$	$(3,2,2)$	$2(3(20))$	$(2,3,2)$
$3(2(\bar{2}0))$	$(3,2,\bar{2})$	$2(3(\bar{2}0))$	$(2,3,\bar{2})$
$3(\bar{2}(\bar{2}0))$	$(3,\bar{2},\bar{2})$	$\bar{2}(3(\bar{2}0))$	$(\bar{2},3,\bar{2})$
$21(2(20))$	$(21,2,2)$	$2(21(20))$	$(2,21,2)$

5.2 Distinguishing mutant tangles

Identifying and distinguishing mutant knots presents significant challenges in knot theory, as they share several of the same invariants (including the hyperbolic volume and the HOMFLYPT polynomial), similarly, this difficulty extends to tangles. Notable examples are tangles with 7-crossings, which are missing in the table of algebraic tangles in [13], where mutants are present at the very end of the classification. The methods used in [13] could not distinguish these 4 pairs, since after closures (numerator, denominator, double) they form the same links.

In Table 3, we list pairs of mutants, and in Figure 14 we show diagrams of one of these pairs. Moreover, missing tangles have higher symmetry and belong to $z\rho$ group, when their mutants belong to one of ρ groups (see next section).

6 Tangle symmetry groups

Operators ν, η, μ from Definition 9 are generators of $\mathbf{D}_8 \times \mathbf{Z}_2$ group:

$$\mathbf{D}_8 \times \mathbf{Z}_2 = \langle \mu, \nu, \eta \mid \mu^2 = \nu^4 = \eta^2 = e, [\nu, \eta] = \nu^2, [\nu, \mu] = [\eta, \mu] = 0 \rangle \tag{20}$$

which is the symmetry group of the Conway sphere. Figure 15 presents tangles related by ν, η, μ operations (orbits), which in general generate the $\mathbf{D}_8 \times \mathbf{Z}_2$ group. Furthermore, we study symmetry groups of tangles in terms of subgroups of $\mathbf{D}_8 \times \mathbf{Z}_2$. Figure 16 presents the lattice diagram of subgroups of $\mathbf{D}_8 \times \mathbf{Z}_2$.

Some subgroups of $\mathbf{D}_8 \times \mathbf{Z}_2$ cannot be the symmetry group of a tangle (e.g. $\mathbf{D}_8 \times \mathbf{Z}_2$ itself), since no tangle can be invariant to ν, ν^3, $\mu\eta$, and $\mu\eta\nu^2$ operations, which is stated in Theorem 6. All other symmetry groups have been observed and they are

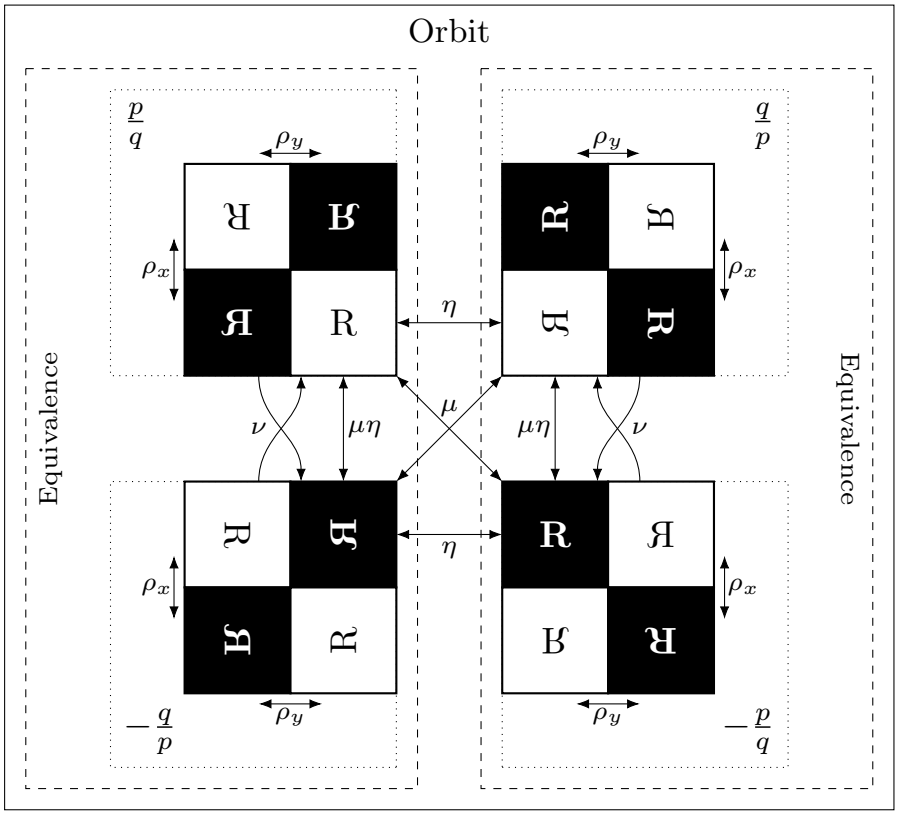


Fig. 15 The orbit of a tangle R. The tangle R and its transformations are represented by two-faced square tiles. The orbit of a tangle collects all tangles obtainable by μ, ν and η operations. Tangles in the orbits have up to 4 distinct fraction values, which are their own inverses and/or opposites: $\frac{p}{q}$, $\frac{q}{p}$, $-\frac{p}{q}$, $-\frac{q}{p}$.
Equivalent tangles are related by rotations, ν and $\mu\eta$, and their compositions, $\rho_x = \mu\eta\nu^3$, $\rho_y = \mu\eta\nu$, $\rho_z = \nu^2$. If a tangle has no mirror symmetries, then its orbit splits into two sets of equivalent tangles.

listed in Table 4 together with operations they are invariant to. The remaining Tables 5 – 8 count tangles up to 14 crossings split into subgroups based on the numbers of crossings and tangle symmetry groups. Tables 6 and 8 count only tangles with no additional closed components.

Table 7 shows that tangles with mirror symmetries are very rare. Note that, while in general tangles generated by ν, μ and η create two groups of equivalent tangles, if any mirror symmetry is present, all tangles generated by ν, μ and η, are equivalent. Interestingly, mirror symmetry groups are split between X-tangles and H/V-tangles (Tables 4, 7, and 8). Moreover, presence of mirror symmetries for a specific number of crossings (in minimal representation) is restricted to either X-tangles or H/V-tangle (Table 7) – at least for tangles up to 14 crossings. It remains an open question if this is generally true for all algebraic tangles – with any number of crossings.

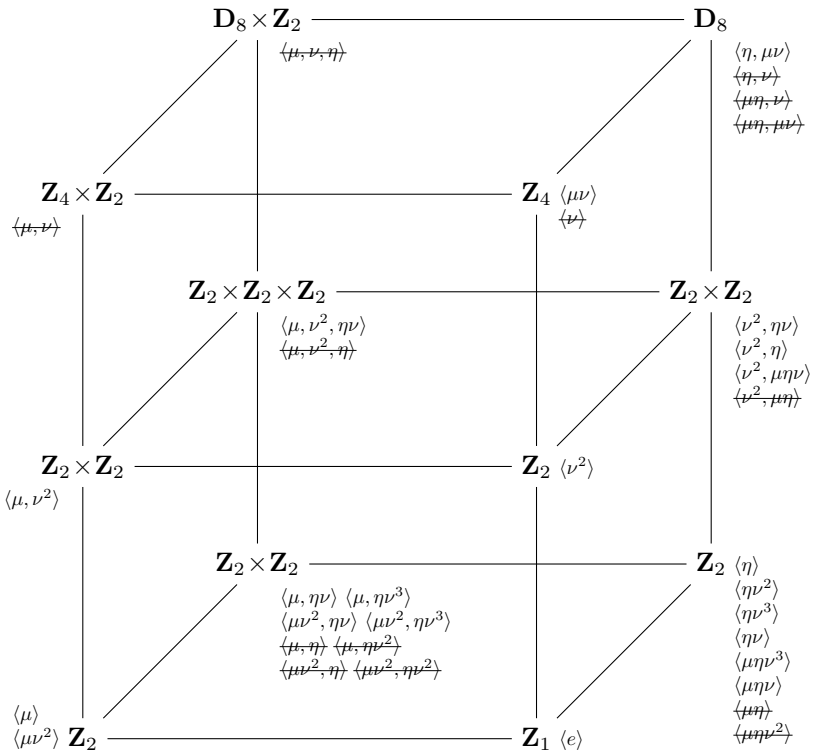


Fig. 16 The lattice of subgroups of $\mathbf{D}_8 \times \mathbf{Z}_2$. Next to each group there are listed possible sets of generators. Sets of generators which are ~~crossed out~~ cannot generate the tangle symmetry group, since they include at least one of $\nu, \nu^3, \mu\eta, \mu\eta\nu^2$ (Theorem 6). All remaining generating sets create symmetry groups, which were all observed in the tangles.

Theorem 6. *No tangle is isotopic to itself after ν, ν^3, $\mu\eta$ or $\mu\eta\nu^2$ transformations.*

Proof. The proof follows from the properties of the fraction (Definition 20 and Lemma 4,5). Let us assume that the fraction of tangle A is w. From $\rho_x \overset{\text{def.}}{=} \mu\nu\eta$, $\rho_z \overset{\text{def.}}{=} \nu^2$, $\eta R \overset{\text{def.}}{=} R0$, $\mu R \overset{\text{def.}}{=} \overline{R}$ we obtain:

1) $\text{Frac}(\mu\nu\eta R) = \text{Frac}(\nu^2 R) = w$,
2) $\text{Frac}(\mu R) = -w$,
3) $\text{Frac}(\eta R) = 1/w$.

1)+2)+3) imply that $\text{Frac}(\nu R) = \text{Frac}(\nu^3 R) = \text{Frac}(\mu\eta R) = \text{Frac}(\mu\eta\nu^2 R) = -1/w$. Invariance to $\nu, \nu^3, \mu\eta$, or $\mu\eta\nu^2$ implies that $w = -1/w$, which is a contradiction in $\mathbb{Q} \cup \{\infty\}$. □

Table 4 Possible symmetry groups of tangles. For each symmetry group, the invariant transformations are listed. Each group has 1, 2, 4 or 8 invariant operations (including Id, which is not listed). Invariance to $\mu\nu^3$ is not listed, since it is an inverse of $\mu\nu$, and other listed transformations are their own inverses. Some equivalence groups of tangles collect tangles from two symmetry groups – such conjugated groups are marked with curly braces. In the first column an abbreviation is assigned to each group (or to a pair of conjugated groups).

Symmetry group			Is invariant to?									Type
Abb.	Type	Generators	ρ_z	ρ_y	ρ_x	$\mu\rho_z$	$\mu\rho_y$	$\mu\rho_x$	μ	η	$\eta\rho_z$	$\mu\nu$	VHX
$\mathfrak{1}$	$\mathbf{D}_8$	$\langle \eta, \mu\nu \rangle$	✓	✓	✓					✓	✓	✓	X
$\overline{\nu}$	$\mathbf{Z}_4$	$\langle \mu\nu \rangle$	✓									✓	X
ηz	$\mathbf{Z}_2 \times \mathbf{Z}_2$	$\langle \eta, \nu^2 \rangle$	✓							✓	✓		X
η \	$\mathbf{Z}_2$	$\langle \eta \rangle$								✓			X
η /	$\mathbf{Z}_2$	$\langle \eta\nu^2 \rangle$									✓		X
$\mathfrak{o}$	$\mathbf{Z}_2 \times \mathbf{Z}_2 \times \mathbf{Z}_2$	$\langle \mu, \nu^2, \eta\nu \rangle$	✓	✓	✓	✓	✓	✓	✓				VH
μz	$\mathbf{Z}_2 \times \mathbf{Z}_2$	$\langle \mu, \nu^2 \rangle$	✓			✓			✓				VH
$\mu\rho$ μy	$\mathbf{Z}_2 \times \mathbf{Z}_2$	$\langle \mu, \eta\nu \rangle$		✓			✓		✓				VH
μx	$\mathbf{Z}_2 \times \mathbf{Z}_2$	$\langle \mu, \eta\nu^3 \rangle$			✓			✓	✓				VH
$\overline{xy}$	$\mathbf{Z}_2 \times \mathbf{Z}_2$	$\langle \eta\nu, \nu^2 \rangle$	✓				✓	✓					VH
$\overline{z\rho}$ $\overline{zx}$	$\mathbf{Z}_2 \times \mathbf{Z}_2$	$\langle \mu\nu^2, \eta\nu^3 \rangle$		✓		✓		✓					VH
$\overline{zy}$	$\mathbf{Z}_2 \times \mathbf{Z}_2$	$\langle \mu\nu^2, \eta\nu \rangle$			✓	✓	✓						VH
$\overline{z}$	$\mathbf{Z}_2$	$\langle \mu\nu^2 \rangle$				✓							VH
$\overline{\rho}$ $\overline{y}$	$\mathbf{Z}_2$	$\langle \eta\nu \rangle$					✓						VH
$\overline{x}$	$\mathbf{Z}_2$	$\langle \eta\nu^3 \rangle$						✓					VH
μ	$\mathbf{Z}_2$	$\langle \mu \rangle$							✓				VH
$z\rho$	$\mathbf{Z}_2 \times \mathbf{Z}_2$	$\langle \nu^2, \mu\eta\nu \rangle$	✓	✓	✓								VHX
z	$\mathbf{Z}_2$	$\langle \nu^2 \rangle$	✓										VHX
ρ y	$\mathbf{Z}_2$	$\langle \mu\eta\nu \rangle$		✓									VHX
x	$\mathbf{Z}_2$	$\langle \mu\eta\nu^3 \rangle$			✓								VHX
e	$\mathbf{Z}_1$	$\langle e \rangle$											VHX

Table 5 Number of all tangle orbits with non-mirror symmetries split by the number of crossings ("#cross"). The properties of each symmetry group are listed in Table 4. Tangles in these group can be of any V-, H-, or X-type. Tangles in each column/row are summed up ("Orbit" column/row), additionally, tangles are counted and summed up to equivalence ("Equiv.") and isotopy ("Isotopy").

#cross	Non-mirror symmetry group				Total		
	$z\rho$	z	ρ	e	Orbit	Equiv.	Isotopy
0	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-
2	1	-	-	-	1	2	4
3	2	-	-	-	2	4	8
4	5	-	-	-	5	10	20
5	11	-	3	-	14	28	68
6	28	4	11	-	43	86	232
7	61	15	52	6	134	268	876
8	154	64	183	52	453	906	3424
9	352	227	675	329	1583	3166	13888
10	858	827	2289	1814	5788	11576	57384
11	1968	2798	7833	9327	21926	43852	242152
12	4754	9510	25961	45782	86007	172014	1035296
13	11049	31482	86257	218540	347328	694656	4482748
14	26467	104385	283466	1024521	1438839	2877678	19601012
Orbit	45710	149312	406730	1300371	1902123		
Equiv.	91420	298624	813460	2600742		3804246	
Isotopy	182840	1194496	3253840	20805936			25437112

Table 6 Same as in Table 5, but only for tangles with no extra closed components.

#cross	Non-mirror symmetry group				Total		
	$z\rho$	z	ρ	e	Orbit	Equiv.	Isotopy
0	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-
2	1	-	-	-	1	2	4
3	2	-	-	-	2	4	8
4	4	-	-	-	4	8	16
5	8	-	3	-	11	22	56
6	19	4	6	-	29	58	156
7	38	8	34	6	86	172	584
8	83	41	101	41	266	532	2124
9	173	120	332	218	843	1686	7796
10	373	391	1027	1110	2901	5802	30596
11	781	1201	3075	5044	10101	20202	118036
12	1688	3562	9156	22688	37094	74188	471504
13	3551	10551	26999	98674	139775	279550	1893388
14	7621	31091	78626	422447	539785	1079570	7667372
Orbit	14342	46969	119359	550228	730898		
Equiv.	28684	93938	238718	1100456		1461796	
Isotopy	57368	375752	954872	8803648			10191640

Table 7 Number of all tangle orbits with mirror symmetries divided by number of crossings ("#cross"). Mirror symmetries are divided into two groups, available either only for X-type or VH-type tangles.
Properties of each symmetry group are listed in Table 4. Tangles in each column/row are added up ("Orbit" column/row). Additionally tangles are counted and added up to equivalence ("Equiv.") and isotopy ("Isotopy").

#cross	X-only				VH-only								Total		
	$\mathfrak{1}$	$\overline{\nu}$	ηz	η	$\mathfrak{o}$	μz	$\mu\rho$	$\overline{xy}$	$\overline{z\rho}$	$\overline{z}$	$\overline{\rho}$	μ	Orbit	Equiv.	Isotopy
0	-	-	-	-	1	-	-	-	-	-	-	-	1	1	2
1	1	-	-	-	-	-	-	-	-	-	-	-	1	1	2
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	-	-	-	-	1	-	-	-	-	-	-	-	1	1	2
5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	1	-	-	-	1	1	4
7	2	-	-	-	-	-	-	-	-	-	-	-	2	2	4
8	-	-	-	-	3	-	-	-	3	-	-	-	6	6	18
9	-	1	-	1	-	-	-	-	-	-	-	-	2	2	12
10	-	-	-	-	-	1	2	-	9	3	4	-	19	19	104
11	3	3	3	3	-	-	-	-	-	-	-	-	12	12	54
12	-	-	-	-	10	3	3	1	32	12	15	1	77	77	400
13	4	8	2	20	-	-	-	-	-	-	-	-	34	34	208
14	-	-	-	-	-	10	17	2	89	77	98	10	303	303	1952
Orbit	10	12	5	24	15	14	22	3	134	92	117	11	459		
Equiv.	10	12	5	24	15	14	22	3	134	92	117	11		459	
Isotopy	20	48	20	192	30	56	88	12	536	736	936	88			2762

Table 8 Same as in Table 7, but only for tangles with no extra closed components.

#cross	X-only		VH-only				Total		
	$\mathfrak{1}$	$\mu\nu$	$\mathfrak{o}$	$\overline{z}\rho$	$\overline{z}$	$\overline{\rho}$	Orbit	Equiv.	Isotopy
0	-	-	1	-	-	-	1	1	2
1	1	-	-	-	-	-	1	1	2
2	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-
6	-	-	-	1	-	-	1	1	4
7	-	-	-	-	-	-	-	-	-
8	-	-	-	1	-	-	1	1	4
9	-	1	-	-	-	-	1	1	4
10	-	-	-	3	3	3	9	9	60
11	-	1	-	-	-	-	1	1	4
12	-	-	-	8	7	6	21	21	136
13	-	3	-	-	-	-	3	3	12
14	-	-	-	17	36	34	87	87	628
Orbit	1	5	1	30	46	43	126		
Equiv.	1	5	1	30	46	43		126	
Isotopy	2	20	2	120	368	344			856

Declarations

Funding

This work was supported by the National Science Centre (#2021/43/I/NZ1/03341 to JIS), University of Warsaw IDUB grant (BOB-IDUB-622-823/2023 to BAG) and the Slovenian Research and Innovation Agency grant N1-0278 to BG.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Data and code availability

Database of tangles is available on the website: https://tangleinfo.cent.uw.edu.pl [33]. Programs that run the classification and draw tangles, as well as supplementary material with classification of tangles up to 10 crossings, are available in the GitHub repository: https://github.com/baaagr/Classification_of_algebraic_tangles [32].

Author contribution

All authors: conceptualization, funding, writing. Bartosz A. Gren: data curation, formal analysis, methodology, software, validation, visualisation. Boštjan Gabrovšek: methodology, software, supervision, validation, visualisation. Joanna I. Sulkowska: project administration, supervision.

References

[1] Conway, J.H.: An enumeration of knots and links, and some of their algebraic properties. In: Computational Problems in Abstract Algebra, pp. 329–358 (1970). Elsevier

[2] Bonahon, F., Siebenmann, L.: New geometric splittings of classical knots and the classification and symmetries of arborescent knots. preprint (2010)

[3] Dowker, C.H., Thistlethwaite, M.B.: Classification of knot projections. Topology and its Applications **16**(1), 19–31 (1983)

[4] Hoste, J., Thistlethwaite, M., Weeks, J.: The first 1,701,936 knots. Math. Intelligencer **20**(4), 33–48 (1998)

[5] Burton, B.A.: The next 350 million knots. In: 36th International Symposium on Computational Geometry (SoCG 2020) (2020). Schloss-Dagstuhl-Leibniz Zentrum für Informatik

[6] Bar-Natan, D.: Fast khovanov homology computations. Journal of Knot Theory and Its Ramifications **16**(03), 243–255 (2007)

[7] Mansfield, M.L.: Are there knots in proteins? Nature structural biology **1**(4), 213 (1994)

[8] Taylor, W.R.: A deeply knotted protein structure and how it might fold. Nature **406**(6798), 916 (2000)

[9] Sulkowska, J.I., Rawdon, E.J., Millet, K.C., Onuchic, J.N., Stasiak, A.: Conservation of complex knotting and slipknotting patterns in proteins. Biophysical Journal **102**(3), 253 (2012)

[10] Niemyska, W., Dabrowski-Tumanski, P., Kadlof, M., Haglund, E., Sułkowski, P., Sulkowska, J.I.: Complex lasso: new entangled motifs in proteins. Scientific reports **6**, 36895 (2016)

[11] Dabrowski-Tumanski, P., Niemyska, W., Pasznik, P., Sulkowska, J.I.: Lassoprot: server to analyze biopolymers with lassos. Nucleic acids research **44**(W1), 383–389 (2016)

[12] Greń, B.A., Dabrowski-Tumanski, P., Niemyska, W., Sulkowska, J.I.: Lasso proteins—unifying cysteine knots and miniproteins. Polymers **13**(22), 3988 (2021)

[13] Moriuchi, H.: Enumeration of algebraic tangles with applications to theta-curves and handcuff graphs. Kyungpook Mathematical Journal **48**(3), 337–357 (2008)

[14] Dabrowski-Tumanski, P., Goundaroulis, D., Stasiak, A., Rawdon, E.J., Sulkowska, J.I.: Theta-curves in proteins. Protein science **33**(9), 5133 (2024)

[15] Silva, F., Gabrovšek, B., Korpacz, M., Luczkiewicz, K., Niewieczerzal, S., Sikora, M., Sulkowska, J.I.: Knots and θ-curves identification in polymeric chains and native proteins using neural networks. Macromolecules **57**(9), 4599–4608 (2024)

[16] Sulkowska, J.I.: On folding of entangled proteins: knots, lassos, links and θ-curves. Current opinion in structural biology **60**, 131–141 (2020)

[17] Gabrovšek, B.: An invariant for colored bonded knots. Studies in Applied Mathematics **146**(3), 586–604 (2021)

[18] Adams, C., Devadoss, J., Elhamdadi, M., Mashaghi, A.: Knot theory for proteins: Gauss codes, quandles and bondles. Journal of Mathematical Chemistry **58**, 1711–1736 (2020)

[19] Dabrowski-Tumanski, P., Rubach, P., Goundaroulis, D., Dorier, J., Sułkowski, P., Millett, K.C., Rawdon, E.J., Stasiak, A., Sulkowska, J.I.: Knotprot 2.0: a database of proteins with knots and other entangled structures. Nucleic acids research **47**(D1), 367–375 (2019)

[20] Dabrowski-Tumanski, P., Jarmolinska, A.I., Niemyska, W., Rawdon, E.J., Millett, K.C., Sulkowska, J.I.: Linkprot: a database collecting information about biological links. Nucleic acids research, 976 (2016)

[21] Dabrowski-Tumanski, P., Sulkowska, J.I.: Topological knots and links in proteins.

Proceedings of the National Academy of Sciences **114**(13), 3415–3420 (2017)

[22] Goundaroulis, D., Dorier, J., Stasiak, A.: Knotoids and protein structure. Topol. Geom. Biopolym **746**, 185 (2020)

[23] Gabrovšek, B., Gügümcü, N.: Invariants of multi-linkoids. Mediterranean journal of mathematics **20**(3), 165 (2023)

[24] Gügümcü, N., Gabrovsek, B., Kauffman, L.H.: Invariants of bonded knotoids and applications to protein folding. Symmetry **14**(8), 1724 (2022)

[25] Cameron, D.P., Grosser, J., Ladigan, S., Kuzin, V., Iliopoulou, E., Wiegard, A., Benredjem, H., Jackson, K., Liffers, S.T., Lueong, S., *et al.*: Coinhibition of topoisomerase 1 and brd4-mediated pause release selectively kills pancreatic cancer via readthrough transcription. Science Advances **9**(41), 5109 (2023)

[26] Ernst, C., Sumners, D.W.: A calculus for rational tangles: applications to dna recombination. Mathematical Proceedings of the Cambridge Philosophical Society **108**(3), 489–515 (1990) https://doi.org/10.1017/S0305004100069383

[27] Sumners, D.W.: Dna, knots and tangles. In: The Mathematics of Knots: Theory and Application, pp. 327–353. Springer, Berlin, Heidelberg (2011)

[28] Kauffman, L.H., Lambropoulou, S.: On the classification of rational tangles. Advances in Applied Mathematics **33**(2), 199–237 (2004)

[29] Burde, G., Zieschang, H.: Knots vol. 5. de Gruyter Stud. Math., Berlin, New York (1986)

[30] Montesinos, J.M.: Revêtements ramifiés de noeuds, espaces fibrés de Seifert et scindements de Heegaard. Universidad de Zaragoza. Seminario Matematico Garcia de Galdeano (1984)

[31] Goldman, J.R., Kauffman, L.H.: Rational tangles. Advances in Applied Mathematics **18**(3), 300–332 (1997)

[32] Classification of algebraic tangles – code and supplement (2025). https://github.com/baaagr/Classificationofalgebraictangles

[33] TangleInfo (2025). https://tangleinfo.cent.uw.edu.pl

Article 1

Lasso proteins – unifying cysteine knots and miniproteins

polymers

MDPI

Article

Lasso Proteins—Unifying Cysteine Knots and Miniproteins

Bartosz Ambroży Greń [1,2,†], Pawel Dabrowski-Tumanski [1,†], Wanda Niemyska [3] and Joanna Ida Sulkowska [1,*]

[1] Centre of New Technologies, University of Warsaw, 02-097 Warsaw, Poland; b.gren@cent.uw.edu.pl (B.A.G.); p.dabrowski-tumanski@uksw.edu.pl (P.D.-T.)
[2] Faculty of Physics, University of Warsaw, 02-093 Warsaw, Poland
[3] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, 02-097 Warsaw, Poland; wanda@mimuw.edu.pl
* Correspondence: jsulkowska@cent.uw.edu.pl
† These authors contributed equally to this work.

Abstract: Complex lasso proteins are a recently identified class of biological compounds that are present in considerable fraction of proteins with disulfide bridges. In this work, we look at complex lasso proteins as a generalization of well-known cysteine knots and miniproteins (lasso peptides). In particular, we show that complex lasso proteins with the same crucial topological features—cysteine knots and lasso peptides—are antimicrobial proteins, which suggests that they act as a molecular plug. Based on an analysis of the stability of the lasso piercing residue, we also introduce a method to determine which lasso motif is potentially functional. Using this method, we show that the lasso motif in antimicrobial proteins, as well in that in cytokines, is functionally relevant. We also study the evolution of lasso motifs, their conservation, and the usefulness of the lasso fingerprint, which extracts all topologically non-triviality concerning covalent loops. The work is completed by the presentation of extensive statistics on complex lasso proteins to analyze, in particular, the strange propensity for "negative" piercings. We also identify 21 previously unknown complex lasso proteins with an ester and a thioester bridge.

Keywords: topology; lasso; proteins

Citation: Greń, B.A.; Dabrowski-Tumanski, P.; Niemyska, W.; Sulkowska, J.I. Lasso Proteins—Unifying Cysteine Knots and Miniproteins. *Polymers* **2021**, *13*, 3988. https://doi.org/10.3390/polym13223988

Academic Editor: Ivan Gitsov

Received: 8 October 2021
Accepted: 9 November 2021
Published: 18 November 2021

1. Introduction

Unification is a key concept in which the most important properties of a given phenomena are extracted. A major goal in physics is the unification of fundamental interactions. In mathematics, common concepts are generalized in frames of category theory. In biology, homological sequences are clustered, leading to the notion of highly conserved (hence functionally important) residues. The spotted similarities spotted may guide the fundamental principles in various fields of science.

During the last 30 years, though not expected, more entangled structures have constantly been discovered. The most well-known examples include knotted [1,2] or slipknotted [3,4] proteins, links [5,6], and chain mail structures of bacteriofag capsid [7]. In 2019, θ-curves were found [8]. In the 1990s, two other complex structures were discovered—cysteine knots [9–13] and lasso peptides (or lariat protoknots) [14–19]. To avoid confusion, throughout this paper, we call these miniproteins. Both types have proven therapeutic significance [11,20,21], both featuring the presence of at least one covalent loop (formed by the main chain closed by an amide or disulfide bridge), which is pierced by some portion of the chain or its bridge (see Figure 1). Moreover, some cysteine knots and miniproteins fulfill a similar function.

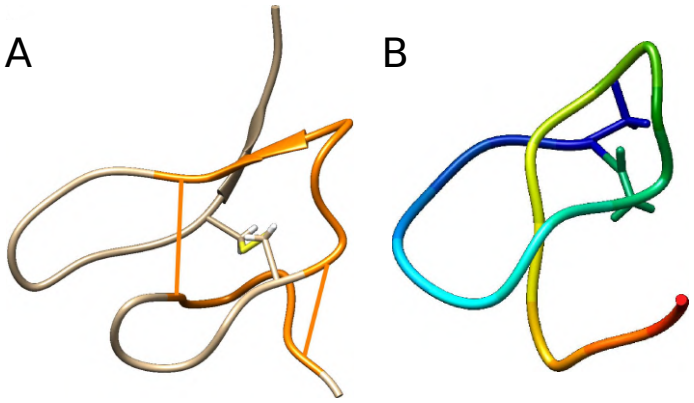


Figure 1. (**A**): Cysteine knot motif. The closed loop (orange) is formed by two disulfide bonds (straight stripes) and the main chain. The third cysteine bond, which pierces the closed loop, is shown explicitly. (**B**): Schematic depiction of the miniprotein Microcin J25. The loop is closed by an amide bridge (to facilitate the view, only bridge-forming residues are shown in atomic resolution).

Is this only an accident or is there any direct correspondence between the entanglement type and the fulfilled function? In 2016, a similar new entanglement type was observed—the complex lasso proteins [22,23]—in which the covalent loop closed by a disulfide (or other) bridge is pierced by one or two tails (Figures 2 and 3). The complex lasso extracts the most important structural properties of cysteine knots and miniproteins, and is more general (18% of proteins with an intrachain cysteine bridge exhibit complex lasso entanglement [23]). This has forced us to think about both cysteine knots and lasso peptides as cases of the same general topological motif.

Such a viewpoint has introduced a whole new chapter to the study of complex topology proteins and has led to many new questions. For example, is it indeed the topology that determines the function? If so, are there any other functional entangled motifs? Do we know all types of complexity in proteins? Are there any non-disulfide-based complex lasso proteins other than miniproteins?

Some of those questions have been at least partially answered so far. In a previous survey [23], we reported the existence of six major classes of complex lasso proteins (with disulfide loops):

- L_1, (L_2, L_3, L_6)—with one, (two, three, six) piercing(s) of the covalent loop, where each subsequent piercing is from a different side of the loop's surface,
- LS_n—with n piercings, but at least two consecutive ones from the same side of the loop's surface (the tail winds around the covalent loop),
- $LL_{i,j}$—with i piercings performed by the N-terminal tail and j piercings by the C-terminal tail.

This classification is completed with the trivial class L_0, which gather closed loops that are not threaded (see Figure 2). A detailed description of the lasso type naming is contained in the Supplementary Materials.

The question of the lasso function may be approached with statistical analysis. However, to obtain reliable results, one has to also take into account the piercing tail (N- or C-terminal) and the piercing direction (see Figure 3). Observe, e.g., that all known miniproteins have the same piercing direction when orienting as shown in Figure 1B. If the loop begins "behind" the plane, the chain pierces always from "above" the covalent loop. Moreover, one has to be aware that one protein chain can have more then one pierced covalent loop.

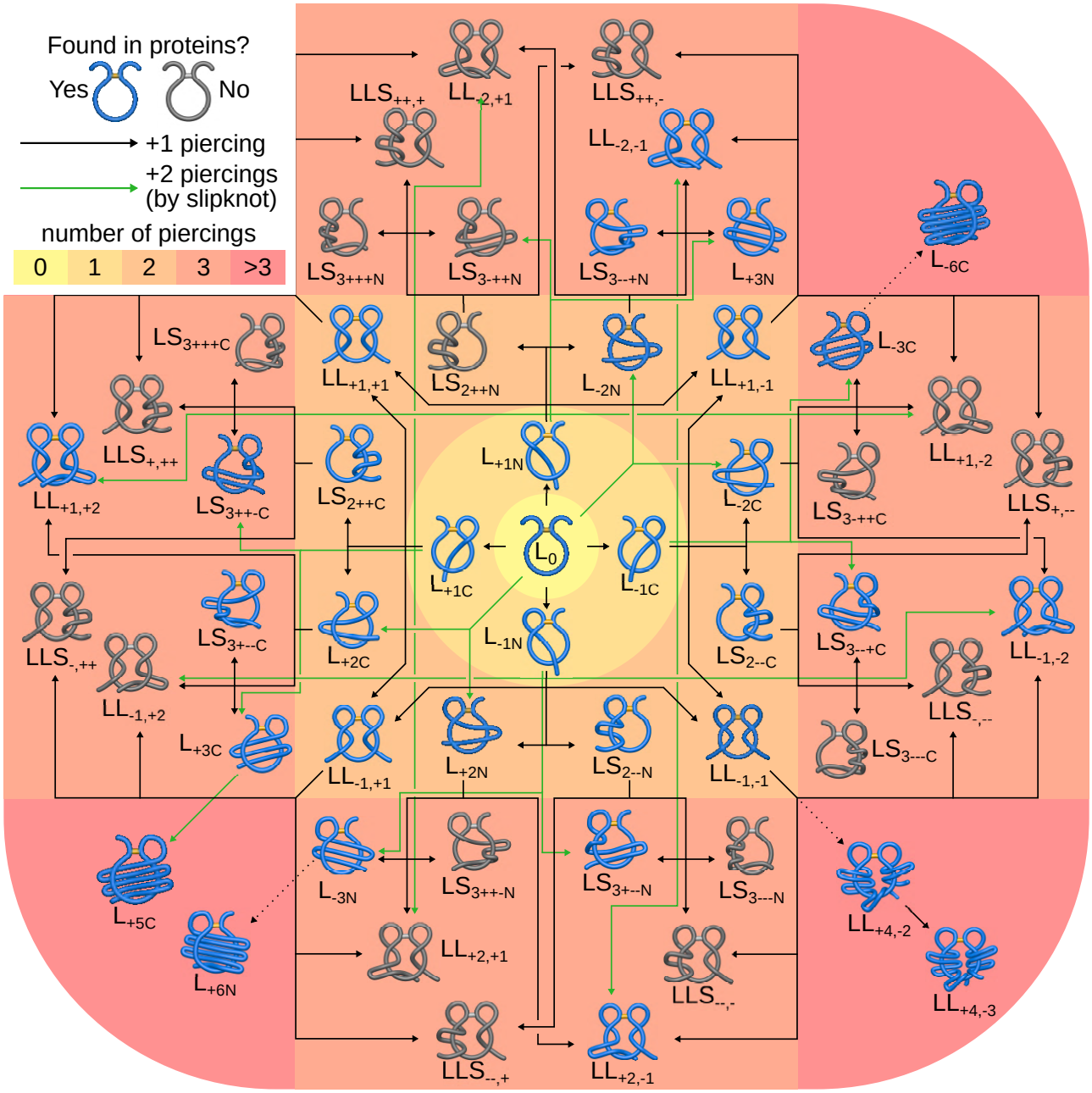


Figure 2. The "evolution flower" of lasso types. The figure shows all possible lasso motifs with up to three piercings and observed lassos with higher number of piercings. The structures observed in PDB are shown in blue, and those not observed (up to 3 piercings) are shown in grey. The structures are grouped according to the total number of piercings (shown by the colored background). The lasso types that can be formed by the addition or reduction of a terminal piercing are joined by a solid black line. Lasso types that can be formed by the addition or reduction of a terminal slipknot piercings are joined by a solid green line. Please note that some lasso motifs can be obtained in more than one way. The dotted lines denote the possible evolution pathway, for which however, some intermediate lasso types have not yet been identified.

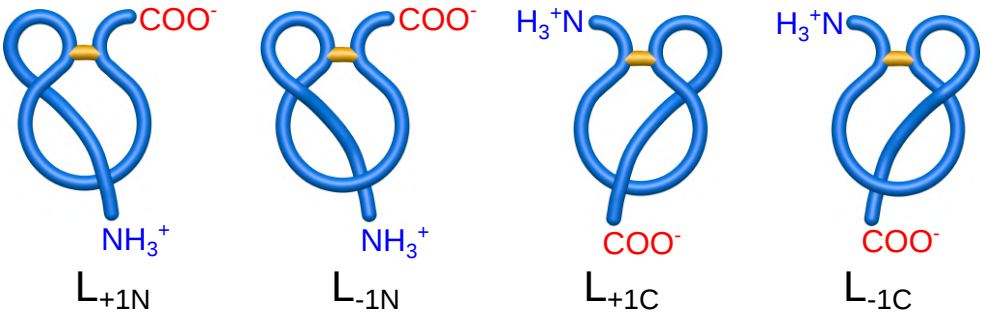


Figure 3. Subclasses of the L_1 major type. From left to right are L_{+1N}, L_{-1N}, L_{+1C}, and L_{-1C}.

To deal with those issues, we created a server and database called LassoProt to analyze complex lasso entanglement in proteins and arbitrary polymers [24]. This is available at lassoprot.cent.uw.edu.pl (accessed on 31 October 2021). This algorithm is also available in our Python package, which is called Topoly [25] topoly.cent.uw.edu.pl (accessed on 31 October 2021) and in the PyMol plug-in PyLasso [26] pylasso.cent.uw.edu.pl (accessed on 31 October 2021). Moreover, topological information is supplemented by basic biological and structural information and various filters, allowing a methodological statistical search to be conducted. Equipped with this tool, we now present a comprehensive analysis of complex lasso proteins from chemical, physical, biological, and also evolutional points of view.

Previously we conducted a statistical analysis of lassoed ideal chains (no volume, no interactions, fixed distance between neighboring atoms) [27]. We found the expression for the probability of a complex lasso formation as a function of loop and tail lengths and compared these results with the distribution of complex lassos in proteins. We found a few protein groups with short tails that have a larger probability of being complex than the ideal chains that the model provides. This means that enthalpic interactions in these proteins favor complex lasso formation. This is a strong indication that the lasso structure plays an important role in these proteins. Many of them are miniproteins or other small antimicrobial proteins with similar plug structures. Some others are enzymes, which may perform a similar function as in knots—making the active site more rigid [28].

Lasso proteins can serve as a molecular switch, as their topology depends on their oxidation potential (the first attempts towards such an application have been made so far [29]). If two stable conformations of the loop relative to the piercing chain exist, such proteins can be used as molecular machines or as a molecular memory [30], similarly to the utilization of the rotaxanes. Redox and pH-controlled lasso-based pseudorotaxanes have already been synthesized using chemical protocols [31,32]. After three decades of research, there is still keen interest in new synthesis methods for molecular rotaxanes [33–36]. Therefore, lasso peptides can provide biological inspiration for (bio)chemical design. In fact, lasso peptides were recently used as a template to obtain self-assembling protein rotaxane structures after enzymatic cleavage [37,38].

This paper is organized as follows: First we compare the results of a survey based on a non-redundant set with only the lasso types present in the entire PDB, taking into account the tail and direction of the piercing. We show that there is a strange propensity for "negative" piercings. Next, we move to the analysis of proteins comprising a few complex lassos, proteins with non-disulfide-based bridges, and structures classified, due to various reasons, as artifacts. In particular, we show that there are around 20 ester-based and thioester-based pierced loops. We also point out some gapped structures which, upon determining the whole structure, may reveal new lasso motifs.

After analyzing the structural data, we move to the biological results. We analyze the conservation of the lasso type and its function and evolution. To study the functional relevance, we introduce a method of analyzing the bulky residues and the flexibility of the chain based on the assumption that a functionally important lasso should be stabilized, as happens in the case of miniproteins. This approach allows us to show that the L_{-1C} lasso

type present in antimicrobial lasso proteins and the L_2 lasso type present in cytokines may be functionally relevant.

The final part of the paper is devoted to the discussion of the results, the possible utilization of complex lasso proteins, and the future research direction.

2. Materials and Methods

2.1. Protein Set

In our analysis, we used all the proteins deposited in the Protein Data Bank up to October 2020, i.e., 535,079 structures. The gaps were modeled as straight intervals. The proteins that did not pass the structure validity test described in [24] and the Supplementary Materials were counted as artifacts.

2.2. Lasso Type Assignment and Closed Loop Detection, Visualization

The topology assignment followed the algorithm developed in [23] with its extension introduced to deal with surface orientation [24]. The covalent loops were detected as described in [24]. Molecular graphics and analyses were performed with the UCSF Chimera package [39] and JSmol implemented in the LassoProt server [24]. Chimera was developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311).

2.3. Sequence Alignment and Bridge Conservation

Sequence alignment was done using the Clustal server via its Perl web client (based on the SOAP protocol). Alignment was done for all clusters with a 40% sequential homology, for which highly redundant structures (with sequential homology over 95%) were reduced. For each alignment, the conservation of the bridge was calculated as the conservation of both bridge-forming cysteines separately. The bridge was considered highly conserved if both cysteines were present in more than 80% of the aligned structures. The bridge was considered poorly conserved if at least one cysteine was present in less than 30% of the aligned structures. In the other case, the bridge was considered medium conserved.

2.4. Protein Characteristics

Extracellular locations were derived from the UNIPROT database. The B-factors, mean square deviations, and locations of bulky residues were determined from the appropriate PDB file. Detailed descriptions are provided in the Supplementary Materials.

3. Results

3.1. Analysis of the Entire Protein Data Bank

3.1.1. The Non-Redundant Set Is Sufficient

We analyzed all protein structures (535,079 chains in October 2020) deposited in the Proteins Data Bank, extracting those possessing a covalent loop (closed by any kind of a chemical interaction—78,723 chains). The chains were divided into topologically certain (52,832 chains) and the "artifact" class with an uncertain entanglement type (e.g., due to the gap size—25,891 chains) in accordance with to [24]. The details for the division methodology are given in the Supplementary Materials.

For each covalent loop, we determined the major lasso type (not taking the orientation and piercing tail into account) and minor lasso type (including the piercing tail and piercing orientation). The major lasso types found in the entire PDB were identical to those within the non-redundant set, except for L_5. All motifs present across the entire PDB were represented in the non-redundant set, except for five exceptional lasso types (see Table 1). These exceptional motifs arose from an additional or different bridge ($LL_{+1,+2}$ motif in glutaminase with PDB code 3a55 and L_{+5C} motif in glucosidase with PDB code 5yj7), truncation of the piercing tail ($LL_{+4,-2}$ in the chemokine inhibitor with PDB code 1cq3) or a slight change in the position of the chain, resulting in acceptance of the piercings,

which otherwise would be disregarded as too shallow (LS_{3++-C} and LS_{3--+N} motifs in transferrin with PDB code 1bp5).

Table 1. Statistics of occurrences of individual loop types—numbers of loops in a given type in all PDB and in the set of non-redundant structures. Note that, in some cases, more than one pierced covalent loop may occur in a given structure (see Lasso fingerprint section).

Class	# Loops	# Repr. Loops	Class	# Loops	# Repr. Loops	Class	# Loops	# Repr. Loops
L_{-1C}	1459	244	LS_{2--C}	60	8	$LL_{-1,-1}$	11	4
L_{+1C}	426	85				$LL_{-1,+1}$	34	6
L_{-1N}	2202	204	LS_{2++C}	95	4	$LL_{+1,-1}$	43	6
L_{+1N}	546	127				$LL_{+1,+1}$	1	1
L_{-2C}	504	60	LS_{2--N}	195	8	$LL_{-1,-2}$	1	1
L_{+2C}	56	12				$LL_{-2,-1}$	3	1
L_{-2N}	180	21	LS_{3--+C}	4	2	$LL_{-2,+1}$	3	1
L_{+2N}	31	8				$LL_{+2,-1}$	1	1
L_{-3C}	88	7	LS_{3+--C}	3	2	$LL_{+1,+2}$	2	0
L_{+3C}	5	1				$LL_{+4,-2}$	1	0
L_{-3N}	329	24	LS_{3++-C}	1	0	$LL_{+4,-3}$	3	1
L_{+3N}	11	3				$LLS_{2--,+2}$	3	1
L_{+5C}	3	0	LS_{3--+N}	12	0			
L_{-6C}	4	2						
L_{+6N}	1	1	LS_{3+--N}	9	3			
Sum	7845	799	Sum	379	27	Sum	106	23
	Total			All loops—6330 All chains—5559			Loops in representative set—849 Chains in representative set—732	

3.1.2. Piercings Tend to Arise from the Negative Side

By taking into account the piercing direction and the type of piercing tail (N-end or C-end), the major lasso type can be split into smaller (minor) subtypes. For example, the L_1 type is composed of four subtypes (L_{-1C}, L_{-1N}, L_{+1C}, L_{+1N}), with the direction of piercing described by a $+$ or $-$ sign and the crossing tail denoted by a capital N or C (Figure 3). For further details concerning the naming convention, see the Supplementary Materials.

The entire PDB was found to contain 35 subclasses—15 of the L_n type, 11 of the two-sided lasso (LL) type, 8 of the supercoiling lasso type (LS), and 1 of the supercoiling two-sided lasso type (LLS). Note that not all possible lasso types were represented—in particular, there was no protein belonging to the LS_{2++N} subclass, while there were examples of every other LS_2 lasso type: LS_{2--N}, LS_{2++C} and LS_{2--C}. The lasso subclasses found throughout the entire PDB are shown in Figure 2, and the numbers of covalent loops for each type are contained in Table 1.

The number of structures in each class was found to correspond to the complexity of the structure (Figure 4A). The most popularweare singly threaded lassos. As a rule of thumb, double sided lassos and supercoiling lassos are more complex and, hence, less frequent (with LS being more frequent than LL). This trend was preserved in both the entire PDB and in the non-redundant set, showing that the latter reflects the properties of the former well.

No clear preference for the piercing tail was shown, although piercing via the C-terminus was slightly more common (52%, Figure 4B). However, there was a strong tendency for "negative" piercings. Indeed, around 64% of pierced loops in the representative structures belonged to the class with a "negative" piercing closest to the bridge (Figure 4C), taking into account all but two-sided lassos (LL). The origin of this phenomenon is not clear. A possible explanation involves the bridge chirality. In this case, the chirality effect should decrease with the increasing of the distance from the bridge. Therefore, for the "positive" piercings, the mode of distance distribution would be shifted towards larger distances compared with the "negative" piercings. Such behaviour can be observed (Figure 4D).

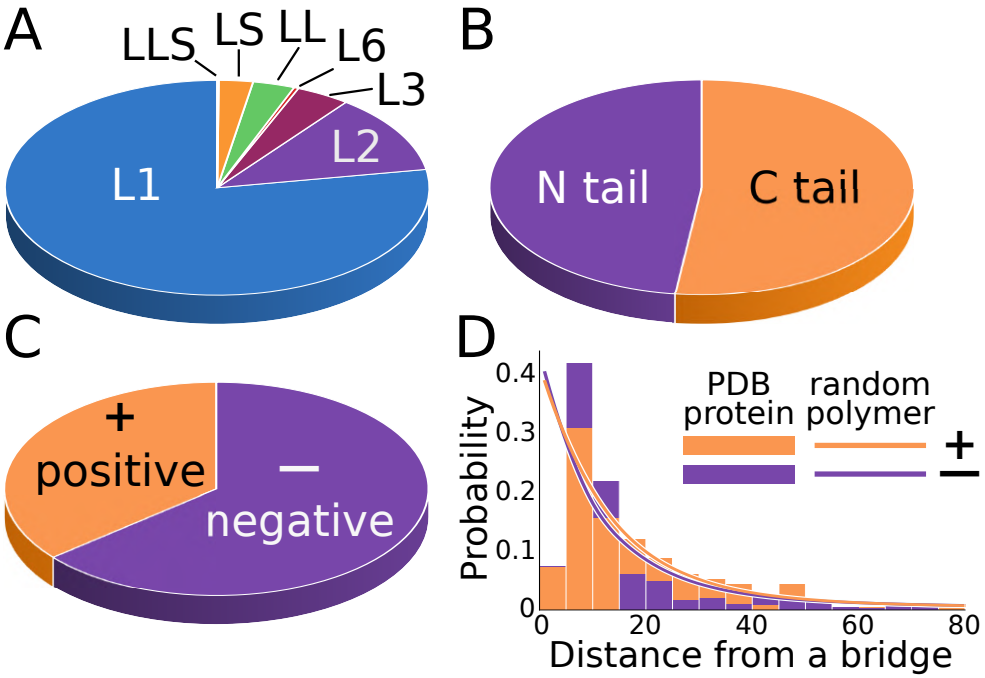


Figure 4. Statistics of loop-type ocurence among the non-redundant set of proteins. The pie charts present (**A**) major types, (**B**) piercing tails, and (**C**) the piercing direction. In (**B**,**C**), all lassos except the two-sided (*LL*) class are taken into account. (**D**) Histogram of the sequential distance between the bridge and the nearest piercing with distinction of the positive (+) and negative (−) piercings. Curves represent the theoretical prediction for a random polymer (based on the fit of the simulated data in [27]).

3.1.3. Lasso Fingerprint

In some cases, the protein chain possesses more than one pierced covalent loop. A description of the global entanglement, including all covalent loops is indispensable for a proper understanding of the structural influence on the biological and mechanical chain properties. To describe the overall topology of the protein, we introduced a "lasso fingerprint" [24], a notion similar to the knot fingerprint used in the case of knotted proteins [40]. The lasso fingerprint is built as a concatenation of lasso type symbols for all pierced loops in their sequential order (with the L_0 symbols, piercing tail, and piercing direction suppressed). Such a simple and intuitive notion extracts all non-trivial parts of the protein lasso entanglement (Figure 5).

Currently, proteins with 44 different lasso fingerprints are deposited in the entire PDB database. Most structures (over 88%) posses only one pierced loop; however, ca 9% of the structures posseses two pierced loops and two structures posses 5 pierced loops (Figure 6). The greatest variety of fingerprints occurs for proteins with two pierced loops (17 different fingerprints). There are 8 different fingerprints with 6 piercings in total and 2 different fingerprints with 7 and 8 piercings (e.g., $LS_3\ 2L_1\ LS_2$, $L_1\ L_2\ L_1\ 2L_2$). Moreover, one protein possesses two loops, the first classified as $LL_{4,3}$ and the second being of the L_2 type, making 9 piercings in total.

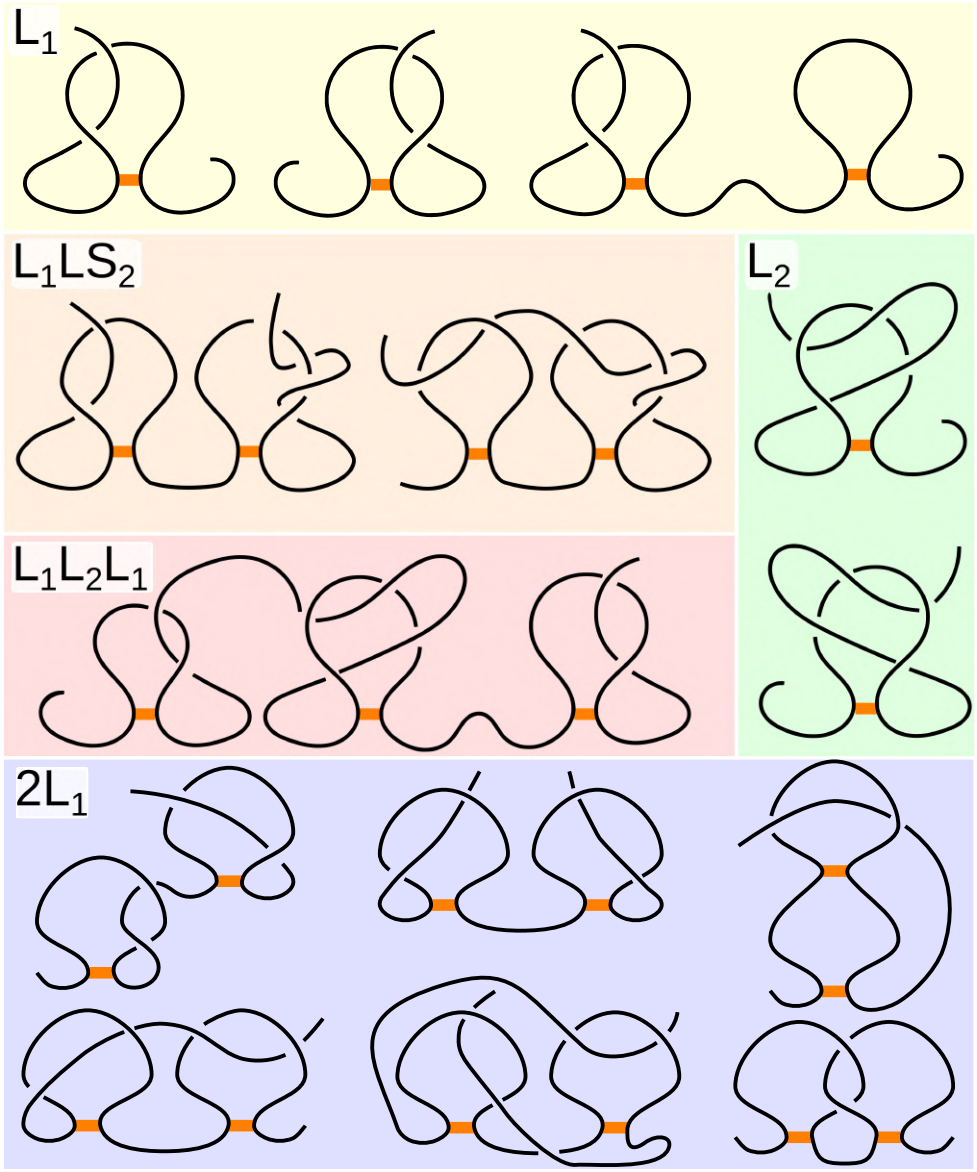


Figure 5. Exemplary lasso structures with fingerprints prescribed. The color boxes contain structures with the same fingerprint. The fingerprint is shown in the upper left corner of the box. The bridge is denoted with an orange strip. Note that by taking into account the piercing tail and orientation, the number of the structures in the Figure multiplies. Note also that the covalent loops in the bottom right structure are linked. Note that the structure over it has one big and one small covalent loop, both being pierced in the same place.

Although the lasso fingerprint is easy to construct and contains a lot of information characterizing the chain topology (e.g., number of pierced loops, their sequential order, etc), it does not characterize the chain topology unequivocally. Observe, for example, that there are over 20 different structures that could be categorized as the $2L_1$ lasso fingerprint (some of them are shown in Figure 5). The development of an easy method for unequivocal multiple lasso classification could be a challenge for mathematicians but has potential importance in biology. Note that, in particular, the $2L_1$ fingerprint can denote the linked covalent loops (forming a Hopf link—Figure 5, bottom right panel). Such structures possessing interesting properties that were identified in [22,28].

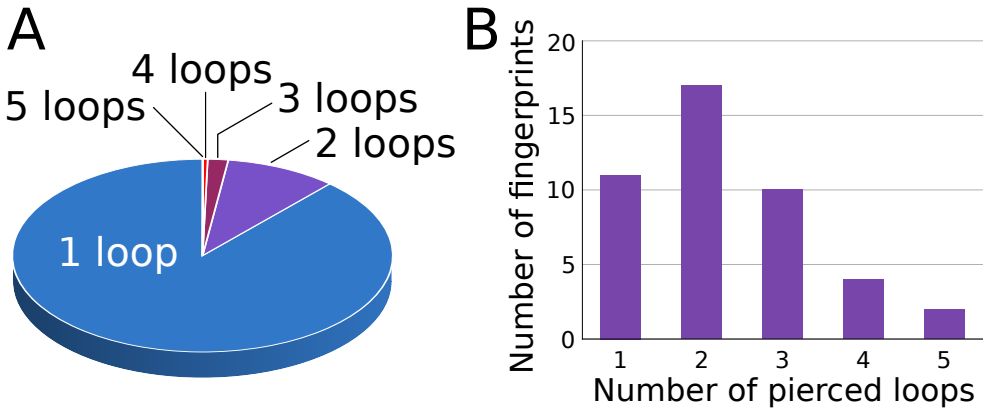


Figure 6. Distribution of pierced loops and lasso fingerprints in lasso proteins. (**A**) Number of pierced loops in a single chain. (**B**) Number of different lasso fingerprints depending on the number of pierced loops.

3.1.4. Non-Disulfide Bridges

The S-S bridge is beyond doubt the most abundant post-translational modification joining two residue side groups. Nevertheless, in some cases, the covalent loop is closed by another type of chemical bond. In the analysis of non-disulfide bonds formally present in the PDB file, one has to be very careful, as many such bonds are artificial. The details of the detection and treatment a of non-disulfide bonds is described in the Supplementary Materials. Across the entire PDB set, we identified four certain non-disulfide loop-forming bonds: N–C, C–O, C–S, and C–C (see Figure 7 and Tables S1 and S2).

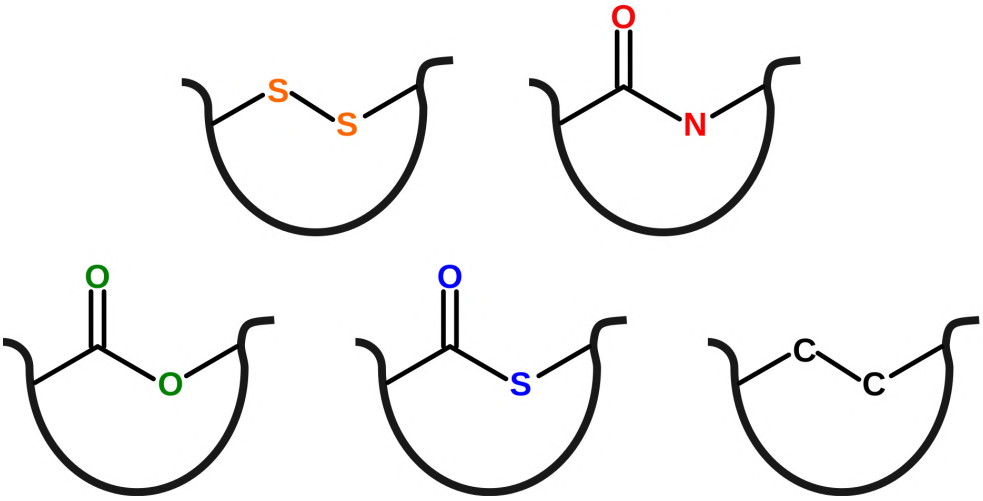


Figure 7. Loop-forming bonds found in PDB. From left to right, (**top** row) disulfide, amide (**bottom** row) ester, thioester, C–C. The amine, ether, and thioether bonds lack the keto-oxygen atom. The black curve denotes the protein chain.

Most covalent bonds between residues are a result of a post-translational enzymatic modification, although some can be an effect of an autocatalitic reaction during folding. N–C is a strong bond present in the amide linkage that is rarely found in intra-residue amines. As it is not prone to mild hydrolysis or changes in oxidation potential, it usually serves as a factor that stabilizes the structure. The C–O bond is present in ester linkages and intra-residue ether structures. The ester bond is more susceptible to hydrolysis than the amide linkage; therefore, it can serve as a kind of molecular switch between protein conformations. The C–S bond (present in thioesters and thioethers) is rather weak and can be regarded as a hidden thiol group. It can be important, e.g., in proteins with adhesive properties, which can be turned on/off depending on the conditions. The C–C bond can be formed in-between aromatic side groups, extending the system of conjugated bonds and

changing the spectral properties of the molecule. Hence, it can be found in photoactive proteins. As a very stable bond under biological conditions, the C–C interaction is also introduced artificially to stabilize the structure.

Across the entire PDB, there are three groups of non-disulfide-based complex lasso proteins (i.e., with non-trivial geometry of the covalent loop):

- amide based—22 chains from X miniproteins with the L_{-1C} lasso type,
- thioester based—13 chains from 10 proteins with 5 different lasso types (L_{-1N}, L_{-1C}, L_{+2C}, L_{+2N}, $LL_{+1,+2}$, $LL_{+1,-3}$),
- ester based—24 chains from 11 proteins with 7 different lasso types (L_{-1N}, L_{-1C}, L_{+1N}, L_{-2N}, L_{+2N}, $LL_{-1,+1}$, $LL_{+1,-1}$).

In the case of miniproteins, the amide bridge is well-known to be essential for the function (see Function section) [41]. On the other hand, there is no evidence of functional importance for the two other bridges (thioester in protein glutaminase and ester in viral protease). In particular, the thioester bridge is absent when the protein is crystallized under different conditions [42], while the C-terminus of the viral protease (forming the ester bridge in the structure with PDB code 3P06) is involved in the enzymatic reaction [43]. Identified protein structures with thioesters and ester-based loops are individual cases among all structures contained in these proteins (check the Table S2 for more details). For example, there are 120 known chain structures of M17 leucyl aminopeptidase (UniProt ID: Q8IL11), but only one of them has an ester bond. The only exception is "Myosin heavy chain, skeletal muscle, adult" (UniProt ID: P13538), which has 13 chains with an ester bridge (out of 64 chains). Hence, in these cases, the presence of the complex lasso structure is probably accidental, stemming from the conditions of crystallization.

3.2. Topology Conservation

In the case of known topological motifs (e.g., knots), the topology is better conserved than the sequence [40]. However, in the case of a complex lasso protein, a simple comparison of lasso fingerprints in the homology cluster can be misleading. The complex lasso motif can be introduced or removed upon mutation of only one crucial amino acid—the bridge-forming cysteine. In general, many factors contribute to the ostensible change of the topology (mutation, insufficient resolution etc), which are described in detail in the Supplementary Materials. Therefore, to study the lasso motif's conservation, we used two complementary approaches.

The first approach followed the procedure of [44], in which the conservation of the cysteine residues (not bridges) in the homology cluster is calculated. Bridges were considered conserved if at least 80% of non-unique structures in the homology cluster had a bridge (see the Supplementary Materials for details). Structures were treated as unique if they had a similarity of at least 95%. It turned out that 405 of 734 (55%) clusters had only one unique structure. This means that, in most cases, complex lasso proteins were unique. In the remaining set of 329 clusters, the disulfide bridges were exceptionally conserved, with 72% classified as highly conserved compared with only 54% in regular proteins [44]. This high fraction also stemmed from the fact that some proteins were found to possess over 10 (all highly conserved) disulfide bridges. However, when calculating the conservation for the pierced-loop-forming bridges only, the percentage of highly conserved bridges was equal to 54%.

The second approach used to determine the lasso conservation was to search for the conservation of a lasso type among PFAM domains. The complex lasso proteins represent 275 different PFAM domains. Among the representative structures of the checked PFAM domains, we aimed to determined what was more common out of a complex lasso structures, a trivial lassos, and a structures with no disulfide bridges. It turned out that, in over 35% of cases (97 distinct PFAM domains), the complex lasso structure was conserved among most representants of the domain. On the other hand, the trivial lasso structure dominated among the representants of 24% of the distinct PFAM domains (66 cases). The reason for nonconservation of the lasso structure could be the truncation of the chain or

the removal of the crucial bridge. Nevertheless, such proteins are perfect objects to study the biophysical importance of the pierced loop. The remaining 41% (112 cases) of PFAM domains most commonly had no bridge at all.

3.3. Complex Lasso Protein Function

The advantage of having a lasso motif is unclear. It was shown that the lasso structure can influence hormone action [45] or can serve as an on/off switch [29]. The lasso can also introduce additional stability to the protein [28]. Indeed, over 50% of the non-redundant set of lasso proteins for which a location is given in the UNIPROT database are secreted and, therefore, exposed to harsh conditions, a process that requires enhanced stability (for details see the Table S3). Moreover, a correlation was found between different lasso types and protein function [23]. In particular, most antimicrobial proteins were of the L_1 type. More specifically, the L_{-1C} lasso type of antimicrobial proteins resembling the (antimicrobial) lasso peptides structurally was found to be characteristic to all families from the PFAM Defensin Clan (CL0075 Defensin/myotoxin superfamily), 4 out of 12 families from the Cystine-knot (CL0079) clan, and 4 other PFAM families containing antimicrobial proteins. This and the resemblance to miniproteins suggest that they have a common mechanism of action as molecular plugs of appropriate channels [46]. Note that we only checked families with known protein structures.

3.3.1. Other Correlations

Apart from the L_{-1C} lasso type, other lasso types exhibit interesting correlations. Previously, it was shown that all eucariotic proteins from the Double Psi beta barrel glucanase PFAM Clan form a Hopf-linked $L_{+1C}L_{+1N}$ lasso structure. LS_{2--N} is characteristic for the PA14 superfamily, and the L_{-2C} type is characteristic for the TIMP-like superfamily. The L_2 type, in general, is also found in many transport or signaling proteins, such as cytokines.

3.3.2. Stabilization of the Lasso Motif as an Indirect Proof of Functionality

If the L_{-1C} antimicrobial lasso proteins indeed act as molecular plugs like the miniproteins, the functional lasso motif should be stabilized by inter-molecular interactions. In the case of miniproteins, stabilization is introduced by the presence of bulky residues surrounding the piercing residue, blocking its movement. In fact, the stabilization of the lasso and, in particular, the piercing residue is indirect proof of the functionality of the motif. Indeed, for small-loop lasso proteins (for which piercing the loop is the least probable due to entropic reasons), the piercing residue is also surrounded by large residues [23].

To quantify the stabilization of the piercing residues, we used two approaches. First, we concentrated on the mechanical blocking, known from miniproteins. In this case, the piercing residue surrounded by bulky residues is located in the piece of chain with a high concentration of large-volume amino acids. We, therefore, calculated the local concentration of bulky residues and identified its local maxima. Then, we calculated the sequential distance between the piercing and the (sequentially) closest maximum (for details see the Supplementary Materials). If the piercing is stabilized mechanically, the distance should be small compared to the average.

Secondly, the highly stable parts of a protein should be characterized by a low B-factor in the case of a crystal structure. In the case of an NMR structure, the analog of the B-factor can be the mean square deviation of positions of atoms in different models (for details, see the Supplementary Materials). We, therefore, identified the minima of the B-factor (or mean square deviation) and, similarly to the previous approach, we calculated the sequential distance between the piercing and the (sequentially) closest minimum. Again, the stabilized piercing should be characterized by a low distance compared with the average. This approach is further supported by the fact that, in many cases, the location of the piercing correlates well with the location of the calculated minimum of the B-factor [47], e.g., for lipocalin with PDB code 1A3Y (see Figure 8A). To determine the loop sizes for which the effect is still visible, we plotted both average distances (from the maximum

of bulky residues concentration and minimum of B-factor)—Figure 8B. To determine the possible importance of each class, we also compared the average for all proteins with average values for secreted antimicrobial L_{-1C} and cytokines with the L_2 lasso type.

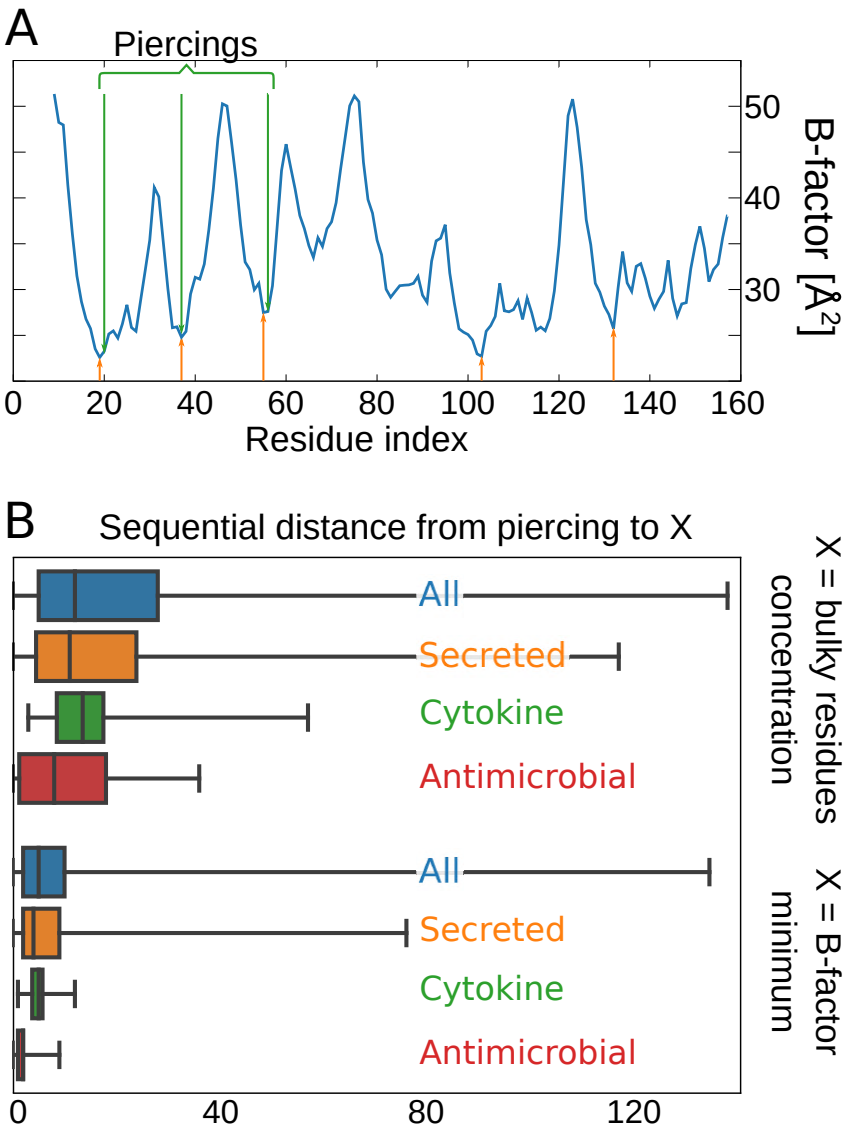


Figure 8. Quantifying lasso stabilization. (**A**) Comparison of the locations of piercings (green arrows) and minima of the B-factor (orange arrows) for lipocalin with PDB code 1A3Y. (**B**) Box plots of the mean distance from the local maximum of the bulky residues concentration and the mean distance from the local B-factor (or mean square deviation in the case of NMR structures) maximum. White dots represent the mean value of each protein family set. Box widths represent 1st-3rd quartile range with the horizontal line showing the 2nd quartile (median).

It turned out that, in both approaches, the mean value corresponding to both the antimicrobial L_{-1C} type and cytokines with the L_2 type was located below the mean value for all possible loop sizes. This means that, in both groups, the lasso motif is stabilized more than in an arbitrary chosen lasso structure. This implies that, in those groups, the lasso motifs can be important for the function. The mean value corresponding to the secreted proteins is closer to the mean value of the whole set, implying that, in many cases, the lasso structure may not be important for the function. and the stability of those proteins is maintained sufficiently solely by the existence of the disulfide bridges.

3.4. Lasso Evolution

The origin of the lasso motif and its evolution have not been studied thus far. There are two possible evolution pathways. Either the lasso covalent loop is primordial and the piercings were "added" during its evolution by extending the piercing tail (covalent-loop-based evolution), or the fold of the protein was optimized during evolution with the formation of a disulfide bridge as the last step (fold-based evolution). In this case, all piercings would have "appeared" at the same time, upon loop closure. Possibly, different mechanisms were involved for different proteins, and establishing such a mechanism for a particular group of proteins would require an in-depth bioinformatical analysis. However, covalent-loop-based evolution seems harder to perform, so the question we asked was, are there are any traces of such evolution? In other words, could the flower in Figure 2 really be an evolution flower?

To answer this question, we searched for (evolution-based) PFAM families containing structures with different lasso types, gathered together in the same (evolution-based) PFAM clan. Co-ocurence of different lasso types in one PFAM clan is a clue that such structures could evolve from the same ancestor. As a result, we found several pairs of lasso motifs joined by membership in the same clan. All possibly evolutionarily important connections (like $L_{-1C} \rightarrow L_{-2C}$) are depicted in Figure 2. Interestingly, there 5 proteins were identified which, depending on their structures, had different topologies: L_1 or L_3 (L_{+1C}/L_{+3C}, L_{-1C}/L_{-3C} and L_{-1N}/L_{-3N}). This proves that the triply pierced proteins may have evolved from singly pierced ones by adding two piercings "at once", i.e., the slipknot conformation, known from knotted protein folding [48].

3.5. Possible Applications

The observation that many well known proteins carry an intrinsic topological structure opens an entirely new chapter in protein application and classification. The stability introduced by lasso proteins and probably utilized by viruses (large fraction of lasso proteins was found in many different viruses [23]) and can be also used to construct proteins with enhanced stability. Additionally, lasso proteins may form a new group of drugs or antibiotics working as the plugs of an appropriate channel, utilizing the mechanism suspected for L_{-1C} type proteins. Another interesting application of (biologically optimized) lasso proteins is in nanomachinery. In general, exploring the biophysics of the lasso proteins depending on different oxidation potential of the solution may lead to many exciting discoveries.

4. Discussion

In this work, we have presented the results of a scrupulous analysis of all complex lasso proteins deposited in the entire RCSB PDB database. We showed that there is strong resemblance between complex lasso proteins of the L_{-1C} type, lasso peptides, and antimicrobial cysteine knots. This suggests that the piercing of the covalent loop is a key structural feature of such proteins and that they can share the same mechanism of action. From this perspective, the complex lasso proteins seem to be the generalization of such a concept, with different lasso types responsible for different functions (e.g., L_2 popular among cytokines). Moreover, although lasso proteins are usually disulfide based, we also identified ester- and thioester-based lasso proteins.

The functional advantage of lasso proteins was tested based on the assumption that the functionally relevant lasso should be stabilized compared with the "average lasso". Following this approach, we showed that for the antimicrobial L_{-1C} lasso proteins and L_2 cytokines, the lasso motif should be functionally important. On the other hand, the lasso motif may arise accidentally in secreted proteins with large numbers of disulfide bridges. Indeed, over 51% of lasso proteins are secreted, almost 36% are membrane, 15% are virial, and only 32% are intercellular proteins, according to UNIPROT. In the case of secreted proteins, there is a smaller stabilization effect, meaning that, in many secreted proteins, the lasso may not be functional.

The secretion of lasso proteins leads to the problem of lasso protein folding. The secreted proteins fold under oxidative conditions, in which the disulfide bridges form during folding. Therefore, the folding of lasso proteins may require threading of a chain through the covalent loop. This process has not been investigated in general, although some preliminary studies have been conducted [45]. This is especially interesting in the case of the supercoiling lasso, for which two threadings from the same side are required, causing folding to be especially hard to complete. The analysis of the folding problem may give clues about the problem of knotted protein folding [48]. On the other hand, knowledge about the folding process may solve the conundrum of the significant prevalence of structures with "negative" piercings. Moreover, folding simulations and experiments with truncated or elongated chains may prove the "covalent-loop-based" mechanism of lasso protein evolution, in which the more complicated lassos are thought to have evolved from less complicated ones.

The evolution of lasso proteins was studied by the comparison of different PFAM families within one PFAM clan. In particular, it seems more probable that the triply pierced lassos evolved from singly pierced rather than from doubly pierced lassos. This may explain the lack of structures with three piercings in the supercoiling manner (tail winding two times around the covalent loop, effectively piercing a loop three times from the same side), as such lasso types cannot be formed by introducing a slipknot conformation into the chain.

In this work, we also analyzed the properties of the lasso fingerprint, which describes, in the most basic way, the nontriviality of the chain. We think that this could be a valuable extension of known classification methods of proteins based on the disulfide bond pattern [49,50], enabling us to prescribe a more detailed protein functions. However, we argue that this method cannot distinguish between all possible composite lasso structures. Full classification of possible lasso motifs, taking into account all protein loops, may require a sophisticated mathematical approach, similarly to how the knot theory entered biology through knotted proteins. However, no mathematical description of lasso-like structure exists as yet.

Regarding the mathematical classification, folding process, evolution, exact function and correlation with the lasso type, free energy landscape, exact influence of the lasso on the stability and conformational ensemble of the protein, much is still unknown in the field of lasso proteins. However, on the other hand, evolutionarily optimized nanomachinery, new therapeutically significant proteins, better understanding of the influence of topology, and much more can be achieved in the future. It is worth making the effort to understand lasso proteins. They have much to offer.

Significance

In this paper, we indicate that complex lasso proteins are a generalization of inhibitor cysteine knots and miniproteins (lasso peptides). We show that their key structural motif of loop piercing is also conserved in other antimicrobial proteins with a disulfide covalent loop. This enables us to identify the "antimicrobial lasso type". In depth study of this phenomena is desirable from the therapeutic point of view. Moreover, we correlate another lasso type with the signaling function and introduce a technique that enables us to indicate potentially functional lasso motifs in proteins. On the other hand, the tools used by us can serve as an additional, complimentary classification method to distinguish between detailed protein functions. We also observe the puzzling piercing direction propensity. Finally, we also investigate the conservation of cysteine bridges. We show that proteins with a complex lasso topology posses exceptionally stable bridges; therefore, the non-trivial lasso motif can be also regarded as a marker of bridge stability.

Supplementary Materials: The following are available at https://www.mdpi.com/article/10.3390/polym13223988/s1, Table S1: List of nontrivial loops closed by an amide bridge, Table S2: List of nontrivial loops closed by a thioester or ester bridge, Table S3: PFAM domain, CLAN and representative protein structure of proteins with complex lasso motifs.

Author Contributions: Conceptualization and methodology, P.D.-T. and J.I.S.; software, W.N. and J.I.S.; validation, B.A.G., P.D.-T. and J.I.S.; formal analysis, B.A.G., P.D.-T. and J.I.S.; investigation, B.A.G., P.D.-T. and J.I.S.; resources, B.A.G., P.D.-T. and J.I.S.; data curation, B.A.G., W.N. and J.I.S.; writing—original draft preparation, P.D.-T. and J.I.S.; writing—review and editing, B.A.G., P.D.-T., W.N. and J.I.S.; visualization, B.A.G., P.D.-T. and J.I.S.; supervision, J.I.S.; project administration, J.I.S.; funding acquisition, J.I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by MNiSW, Idea Plus 0003/ID3/2016/64 (J.I.S.), an EMBO Installation Grant #2057 (J.I.S.), and National Science Centre #2017/24/T/NZ1/00490 (P.D.-T.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within this article, its supplementary material and LassoProt database https://lassoprot.cent.uw.edu.pl.

Acknowledgments: The authors would like to acknowledge Paweł Rubach for text corrections. The authors would like to acknowledge the contribution of COST Action CA17139.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

COST	European Cooperation in Science and Technology
EMBO	European Life Scientist Organisation
NIGMS	National Institute of General Medical Sciences
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PFAM	Protein Families (database)
RCSB	Research Collaboratory for Structural Bioinformatics
UCSF	University of California, San Francisco

References

1. Taylor, W.R. A deeply knotted protein structure and how it might fold. *Nature* **2000**, *406*, 916–919. [CrossRef]
2. Bölinger, D.; Sułkowska, J.I.; Hsu, H.P.; Mirny, L.A.; Kardar, M.; Onuchic, J.N.; Virnau, P. A Stevedore's protein knot. *PLoS Comput. Biol.* **2010**, *6*, e1000731. [CrossRef]
3. Taylor, W.R. Protein knots and fold complexity: Some new twists. *Comput. Biol. Chem.* **2007**, *31*, 151–162. [CrossRef] [PubMed]
4. Yeates, T.O.; Norcross, T.S.; King, N.P. Knotted and topologically complex proteins as models for studying folding and stability. *Curr. Opin. Chem. Biol.* **2007**, *11*, 595–603. [CrossRef]
5. Yan, L.Z.; Dawson, P.E. Design and synthesis of a protein catenane. *Angew. Chem.* **2001**, *113*, 3737–3739. [CrossRef]
6. Boutz, D.R.; Cascio, D.; Whitelegge, J.; Perry, L.J.; Yeates, T.O. Discovery of a thermophilic protein complex stabilized by topologically interlinked chains. *J. Mol. Biol.* **2007**, *368*, 1332–1344. [CrossRef] [PubMed]
7. Helgstrand, C.; Wikoff, W.R.; Duda, R.L.; Hendrix, R.W.; Johnson, J.E.; Liljas, L. The refined structure of a protein catenane: The HK97 bacteriophage capsid at 3.44 Å resolution. *J. Mol. Biol.* **2003**, *334*, 885–899. [CrossRef]
8. Dabrowski-Tumanski, P.; Goundaroulis, D.; Stasiak, A.; Sulkowska, J.I. θ-curves in proteins. *arXiv* **2019**, arXiv:1908.05919.
9. McDonald, N.Q.; Hendrickson, W.A. A structural superfamily of growth factors containing a cystine knot motif. *Cell* **1993**, *73*, 421–424. [CrossRef]
10. Murray-Rust, J.; McDonald, N.Q.; Blundell, T.L.; Hosang, M.; Oefner, C.; Winkler, F.; Bradshaw, R.A. Topological similarities in TGF-β2, PDGF-BB and NGF define a superfamily of polypeptide growth factors. *Structure* **1993**, *1*, 153–159. [CrossRef]
11. Craik, D.J.; Daly, N.L.; Waine, C. The cystine knot motif in toxins and implications for drug design. *Toxicon* **2001**, *39*, 43–60. [CrossRef]
12. Postic, G.; Gracy, J.; Périn, C.; Chiche, L.; Gelly, J.C. KNOTTIN: The database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling. *Nucleic Acids Res.* **2017**, *46*, D454–D458. [CrossRef] [PubMed]
13. Dabrowski-Tumanski, P.; Rubach, P.; Goundaroulis, D.; Dorier, J.; Sułkowski, P.; Millett, K.C.; Rawdon, E.J.; Stasiak, A.; Sulkowska, J.I. KnotProt 2.0: A database of proteins with knots and other entangled structures. *Nucleic Acids Res.* **2018**, *47*, D367–D375. [CrossRef] [PubMed]
14. Blond, A.; Péduzzi, J.; Goulard, C.; Chiuchiolo, M.J.; Barthélémy, M.; Prigent, Y.; Salomón, R.A.; Farías, R.N.; Moreno, F.; Rebuffat, S. The cyclic structure of microcin J25, a 21-residue peptide antibiotic from *Escherichia coli*. *Eur. J. Biochem.* **1999**, *259*, 747–756. [CrossRef] [PubMed]

15. Rosengren, K.J.; Clark, R.J.; Daly, N.L.; Göransson, U.; Jones, A.; Craik, D.J. Microcin J25 has a threaded sidechain-to-backbone ring structure and not a head-to-tail cyclized backbone. *J. Am. Chem. Soc.* **2003**, *125*, 12464–12474. [CrossRef] [PubMed]
16. Zimmermann, M.; Hegemann, J.D.; Xie, X.; Marahiel, M.A. The astexin-1 lasso peptides: Biosynthesis, stability, and structural studies. *Chem. Biol.* **2013**, *20*, 558–569. [CrossRef]
17. Hegemann, J.D.; Zimmermann, M.; Xie, X.; Marahiel, M.A. Lasso peptides: An intriguing class of bacterial natural products. *Accounts Chem. Res.* **2015**, *48*, 1909–1919. [CrossRef]
18. Li, Y.; Zirah, S.; Rebuffat, S. Lasso Peptide Bioengineering and Bioprospecting. In *Lasso Peptides*; Springer: Berlin/Heidelberg, Germany, 2015.
19. Martin-Gómez, H.; Tulla-Puche, J. Lasso peptides: Chemical approaches and structural elucidation. *Org. Biomol. Chem.* **2018**, *16*, 5065–5080. [CrossRef] [PubMed]
20. Frechet, D.; Guitton, J.; Herman, F.; Faucher, D.; Helynck, G.; Monegier du Sorbier, B.; Ridoux, J.; James-Surcouf, E.; Vuilhorgne, M. Solution structure of RP 71955, a new 21 amino acid tricyclic peptide active against HIV-1 virus. *Biochemistry* **1994**, *33*, 42–50. [CrossRef]
21. Knappe, T.A.; Manzenrieder, F.; Mas-Moruno, C.; Linne, U.; Sasse, F.; Kessler, H.; Xie, X.; Marahiel, M.A. Introducing lasso peptides as molecular scaffolds for drug design: Engineering of an integrin antagonist. *Angew. Chem. Int. Ed.* **2011**, *50*, 8714–8717. [CrossRef]
22. Dabrowski-Tumanski, P.; Jarmolinska, A.I.; Niemyska, W.; Rawdon, E.J.; Millett, K.C.; Sulkowska, J.I. LinkProt: A database collecting information about biological links. *Nucleic Acids Res.* **2017**, *45*, D243–D249. [CrossRef] [PubMed]
23. Niemyska, W.; Dabrowski-Tumanski, P.; Kadlof, M.; Haglund, E.; Sułkowski, P.; Sulkowska, J.I. Complex lasso: New entangled motifs in proteins. *Sci. Rep.* **2016**, *6*, 36895. [CrossRef]
24. Dabrowski-Tumanski, P.; Niemyska, W.; Pasznik, P.; Sulkowska, J.I. LassoProt: Server to analyze biopolymers with lassos. *Nucleic Acid Res.* **2016**, *44*, W383–W389. [CrossRef] [PubMed]
25. Dabrowski-Tumanski, P.; Rubach, P.; Niemyska, W.; Gren, B.A.; Sulkowska, J.I. Topoly: Python package to analyze topology of polymers. *Briefings Bioinform.* **2019**, *22*, bbaa196. [CrossRef]
26. Gierut, A.M.; Niemyska, W.; Dabrowski-Tumanski, P.; Sułkowski, P.; Sulkowska, J.I. PyLasso: A PyMOL plugin to identify lassos. *Bioinformatics* **2017**, *33*, 3819–3821. [CrossRef] [PubMed]
27. Dabrowski-Tumanski, P.; Gren, B.; Sulkowska, J.I. Statistical properties of lasso-shape polymers and their implications for complex lasso proteins function. *Polymers* **2019**, *11*, 707. [CrossRef] [PubMed]
28. Dabrowski-Tumanski, P.; Sulkowska, J.I. Topological knots and links in proteins. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3415–3420. [CrossRef]
29. Haglund, E. Engineering covalent loops in proteins can serve as an on/off switch to regulate threaded topologies. *J. Phys. Condens. Matter* **2015**, *27*, 354107. [CrossRef]
30. Green, J.E.; Choi, J.W.; Boukai, A.; Bunimovich, Y.; Johnston-Halperin, E.; DeIonno, E.; Luo, Y.; Sheriff, B.A.; Xu, K.; Shin, Y.S.; et al. A 160-kilobit molecular electronic memory patterned at 1011 bits per square centimetre. *Nature* **2007**, *445*, 414–417. [CrossRef]
31. Clavel, C.; Fournel-Marotte, K.; Coutrot, F. A pH-sensitive peptide-containing lasso molecular switch. *Molecules* **2013**, *18*, 11553–11575. [CrossRef] [PubMed]
32. Schröder, H.V.; Wollschläger, J.M.; Schalley, C.A. Redox-controlled self-inclusion of a lasso-type pseudo [1] rotaxane. *Chem. Commun.* **2017**, *53*, 9218–9221. [CrossRef]
33. Wu, C.; Lecavalier, P.R.; Shen, Y.X.; Gibson, H.W. Synthesis of a rotaxane via the template method. *Chem. Mater.* **1991**, *3*, 569–572. [CrossRef]
34. Hübner, G.M.; Gläser, J.; Seel, C.; Vögtle, F. High-yielding rotaxane synthesis with an anion template. *Angew. Chem. Int. Ed.* **1999**, *38*, 383–386. [CrossRef]
35. Hänni, K.D.; Leigh, D.A. The application of CuAAC 'click'chemistry to catenane and rotaxane synthesis. *Chem. Soc. Rev.* **2010**, *39*, 1240–1251. [CrossRef] [PubMed]
36. Heard, A.W.; Goldup, S.M. Synthesis of a mechanically planar chiral rotaxane ligand for enantioselective catalysis. *Chem* **2020**, *6*, 994–1006. [CrossRef] [PubMed]
37. Allen, C.D.; Link, A.J. Self-assembly of catenanes from lasso peptides. *J. Am. Chem. Soc.* **2016**, *138*, 14214–14217. [CrossRef]
38. Schröder, H.V.; Zhang, Y.; Link, A.J. Dynamic covalent self-assembly of mechanically interlocked molecules solely made from peptides. *Nat. Chem.* **2021**, *13*, 850–857. [CrossRef] [PubMed]
39. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [CrossRef]
40. Sułkowska, J.I.; Rawdon, E.J.; Millett, K.C.; Onuchic, J.N.; Stasiak, A. Conservation of complex knotting and slipknotting patterns in proteins. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E1715–E1723. [CrossRef]
41. Bayro, M.J.; Mukhopadhyay, J.; Swapna, G.; Huang, J.Y.; Ma, L.C.; Sineva, E.; Dawson, P.E.; Montelione, G.T.; Ebright, R.H. Structure of antibacterial peptide microcin J25: A 21-residue lariat protoknot. *J. Am. Chem. Soc.* **2003**, *125*, 12382–12383. [CrossRef] [PubMed]
42. Hashizume, R.; Maki, Y.; Mizutani, K.; Takahashi, N.; Matsubara, H.; Sugita, A.; Sato, K.; Yamaguchi, S.; Mikami, B. Crystal structures of protein glutaminase and its pro forms converted into enzyme-substrate complex. *J. Biol. Chem.* **2011**, *286*, 38691–38702. [CrossRef]

43. Chung, I.Y.W.; Paetzel, M. Crystal Structure of a Viral Protease Intramolecular Acyl-enzyme Complex Insights into cis-Cleavage at the VP4/VP3 Junction of Tellina Birnavirus. *J. Biol. Chem.* **2011**, *286*, 12475–12482. [CrossRef]
44. Thangudu, R.R.; Manoharan, M.; Srinivasan, N.; Cadet, F.; Sowdhamini, R.; Offmann, B. Analysis on conservation of disulphide bonds and their structural features in homologous protein domain families. *BMC Struct. Biol.* **2008**, *8*, 55. [CrossRef] [PubMed]
45. Haglund, E.; Sułkowska, J.I.; He, Z.; Feng, G.S.; Jennings, P.A.; Onuchic, J.N. The unique cysteine knot regulates the pleotropic hormone leptin. *PLoS ONE* **2012**, *7*, e45654. [CrossRef]
46. Adelman, K.; Yuzenkova, J.; La Porta, A.; Zenkin, N.; Lee, J.; Lis, J.T.; Borukhov, S.; Wang, M.D.; Severinov, K. Molecular mechanism of transcription inhibition by peptide antibiotic Microcin J25. *Mol. Cell* **2004**, *14*, 753–762. [CrossRef] [PubMed]
47. Niemyska, W.; Millett, K.C.; Sulkowska, J.I. GLN—A method to reveal unique properties of lasso type topology in proteins. *Sci. Rep.* **2020**, *10*, 15186. [CrossRef] [PubMed]
48. Sułkowska, J.I.; Sułkowski, P.; Onuchic, J. Dodging the crisis of folding proteins with knots. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 3119–3124. [CrossRef]
49. Gupta, A.; Van Vlijmen, H.W.; Singh, J. A classification of disulfide patterns and its relationship to protein structure and function. *Protein Sci.* **2004**, *13*, 2045–2058. [CrossRef] [PubMed]
50. Mas, J.M.; Aloy, P.; Marti-Renom, M.A.; Oliva, B.; de Llorens, R.; Aviles, F.X.; Querol, E. Classification of protein disulphide-bridge topologies. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 477–487. [CrossRef] [PubMed]

Article 2

AlphaLasso – a web server to identify loop and lasso motifs in 3D structure of biopolymers

Nucleic Acids Research, 2025, gkaf375
https://doi.org/10.1093/nar/gkaf375
Web Server issue

AlphaLasso—a web server to identify loop and lasso motifs in 3D structure of biopolymers

Pawel Rubach [1,2,†], Jacek Płonka [2,3,†], Bartosz A. Gren [2], Fernando Bruno da Silva [2], Marta Korpacz [2], Joanna I. Sulkowska [2,*]

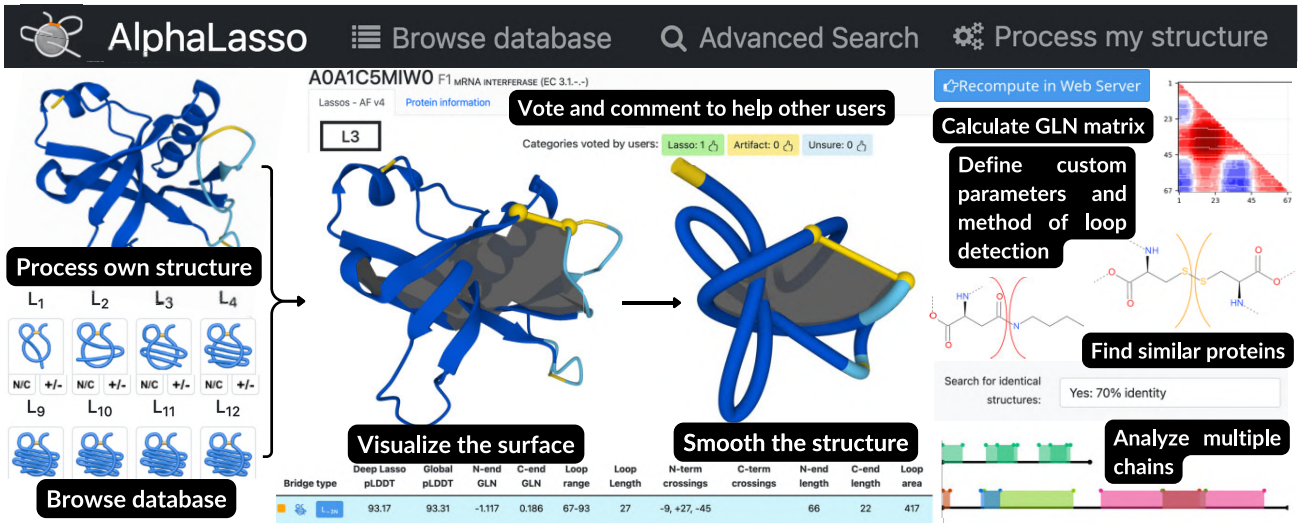
[1]Institute of Information Systems and Digital Economy, Warsaw School of Economics, Al. Niepodleglosci 162, 02-554 Warsaw, Poland
[2]Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland
[3]Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

[*]To whom correspondence should be addressed. Email: j.sulkowska@cent.uw.edu.pl
[†]The first two authors should be regarded as Joint First Authors.

Abstract

With the growing number of AI-predicted protein structures, automated methods of broad-scale analysis are required to parse this volume of data. The application of mathematically defined topologies to protein science enables such analysis. Building on the foundation of lasso peptides, complex lasso motifs are their macroscopic analogs in proteins, promising novel discoveries in drug design and the biopolymer industry. Here we present AlphaLasso, a web server designed to find and analyze lasso-type topologies in protein structures. It finds cysteine, amide, ester, and thioester or user-specified closing bridges. The modern visualization interface provides extensive capabilities to study lasso motifs, such as structure smoothing, creating topology maps, searching for similar proteins, in-depth model evaluation, and metadata annotation. This rich feature set makes AlphaLasso a powerful tool useful in biology, biophysics, chemistry, and mathematics. To enable large-scale analysis, we have precomputed the lasso topologies of high-quality models from the AlphaFold Database, finding >14 million proteins with lasso motifs closed by cysteine bridges, 2.2 million of which are complex lassos. Lasso motifs classified by complexity are available to users via an interactive website, supporting comparison with user-submitted structures. AlphaLasso is available at https://alphalasso.cent.uw.edu.pl/.

Graphical abstract



Introduction

The explosion of available protein structure data made available by large-scale machine learning models, such as AlphaFold [1], ESMFold [2], or RoseTTAFold [3], provided a valuable resource for the advancements in all research and applications that use protein structures. However, it is impossible to parse such an immense volume of data manually. Automated feature discovery is needed, but it can only be built when those features can be defined algorithmically.

Received: March 21, 2025. **Revised:** April 20, 2025. **Editorial Decision:** April 22, 2025. **Accepted:** April 23, 2025

Proteins with a complex lasso motif are an example of such a feature. Covalently closed loops (e.g. by a cysteine bridge) formed by a certain proportion of a protein's backbone with a minimal surface spanned on them can be pierced by the tails created by the loop's closure [4]. As was shown in the example of proteins with one or more closed loops by the presence of a cysteine bond, e.g. cysteine knots and cyclotides [5], such motifs increase the thermodynamic stability of the protein [6, 7]. Moreover, generally, the presence of cysteine bonds [8] improves this stability [8]; thus, the complex lasso motif can be a significant feature that increases the stability and flexibility of a protein [9]. On the other hand, even a non-covalent lasso motif can cause misfolding [10]. An analysis carried out in the Protein Data Bank (PDB) database found that at least 18% of proteins possess this topology [11] with the number of individual examples steadily growing, following the progressive deposition of new structures. With 39% of those found proteins being enzymes [4], it is expected that this motif influences the protein function, as it was shown for the hormone leptin [12] and folding pathway [13, 14]. However, the use of the PDB database creates an issue. The proteins deposited in the PDB are only a tiny fraction of all known protein sequences and consequently do not represent all possible folds and, as a result, do not allow for a comprehensive analysis of the world of proteins. Consequently, it is not known whether the complex lasso motif is common or not, whether it occurs in a particular group of proteins, whether it is related to the function of a protein, or perhaps where it resides in the cell [15, 16]. One would expect that if the lasso has a function, it is also conserved in a particular family, although precisely because of the lack of data, this is still unknown. Analysis of PDB data showed that there are relatively simple lasso motifs with up to four seeps, but there is also a supercoil motif [4]. Whether more complex motifs exist remains to be discovered [17].

Historically, the lasso motif was first found in peptides, and over time its unique biological and chemical properties have been demonstrated. Lasso peptides display antimicrobial, antiviral [18], peptide antagonist [18], and antitumor activities, making them promising candidates for medicinal applications [19]. In particular, Braffman *et al.* [20] showed that in the case of microcin j25 and capistruin, the presence of lassos is crucial for their antimicrobial role. Their conformation can be vulnerable to temperature and pH, making them possible blocks of future molecular micromachines [21–23]. Furthermore, their extraordinary features make them viable tools for the creation of fusion proteins [24] and a new generation of drugs [25]. This shows that there is a large gap between the understanding of the lasso motif in proteins and peptides.

Differences in understanding the lasso motif in proteins and peptides come not only from its absence in PDB structures but also from its complexity [17]. In the case of proteins, the lasso motif is too difficult to spot with the naked eye. Consequently, different algorithms have been developed to analyze them [26] and can be applied to perform a large-scale analysis of structures predicted by AI methods. Although at first, it may seem difficult to predict a new topology, recent data show that AlphaFold has correctly predicted several new knot types in proteins, such as the double-knotted topology [27] or a knot with seven crossings [28], other new types of knots are awaiting experimental verification [29].

In this study, we introduce the AlphaLasso web server, designed to identify the lasso topology in predicted protein structures. It is based on a combination of state-of-the-art methods for detecting the lasso topology [26] and tested on over 14 million proteins with such a motif. This service enables research of the lasso topology on a new scale, allowing for a comprehensive analysis of lasso proteins, peptides, and other biopolymers in a single website, and driving the pursuit of answers regarding this unique motif's origins, influence, and function.

Although AI methods make the 3D structure known, files created as the output do not include annotations for inter-residue bonds, such as disulfide or amide bonds. This poses a problem when determining whether the polypeptide chain forms a loop. The AlphaLasso web server automatically detects sites of link formation, supporting cysteine, amide, ester, and thioester bonds. Additionally, it provides a ready-to-use database of high-quality AlphaFold models for which a cysteine bond was predicted.

The AlphaLasso service is free and open to all users, and there is no log-in requirement.

Materials and methods

Detection of lassos

To determine the lasso motif, it is necessary to define a closing bridge along the polypeptide backbone. It can be identified by any pair of amino acids (non-covalent lasso) or by existing covalent bonds. A set of criteria was developed by analyzing the protein structures and existing links in the PDB database, allowing for accurate detection of cysteine, amide, ester, and thioester closing bridges. In its simplest implementation, a distance parameter is calculated between each atom capable of forming a specified covalent bond. If the distance is below the established threshold, the pair is determined to be capable of bonding. As seen in Fig. 1, the thresholds capture most of the bonds present in the PDB dataset. Performance metrics for this method are available in Table 1. While the LINK annotations in the PDB file format were used during this analysis, the AlphaLasso web server does not depend on this or any other annotation.

Even with high-quality models predicted by AlphaFold, the side-chain positions can be incorrect [30], and the lasso detection depends on this information. To battle this handicap for the most prevalent cysteine bond, an SVM (support vector machine) machine-learning model is used as an additional layer of validation for the detection results. In this approach, the detection is based on a feature vector comprising a flattened distance matrix computed among the key atoms (C, C_α, C_β, N, and S) three dihedral angles ($C_{\beta_1} - S_1 - S_2 - C_{\beta_2}$, $C_{\alpha_1} - C_{\beta_1} - S_1 - S_2$, $C_{\alpha_2} - C_{\beta_2} - S_2 - S_1$). The SVM was trained using `svm.SVC` from the scikit-learn 1.6.1 package [31] with hyperparameters `C = 10`, `kernel = rbf`, and `gamma = scale`. The results are presented in Table 2. By default, both mechanisms are used in the detection process. Each cysteine bond is then annotated to indicate which method found the bond.

Once the closing bond is found, the determination of the lasso topology begins. It is facilitated using the Topoly package [26]. A minimal surface is spanned on the loop closed by the detected bridge, and the lasso type is determined by the number and orientations of deep crossings made by the lasso's tails. Shallow (non-deep) crossings are those that are <3 residues from a chain terminus or a loop, or <10 residues

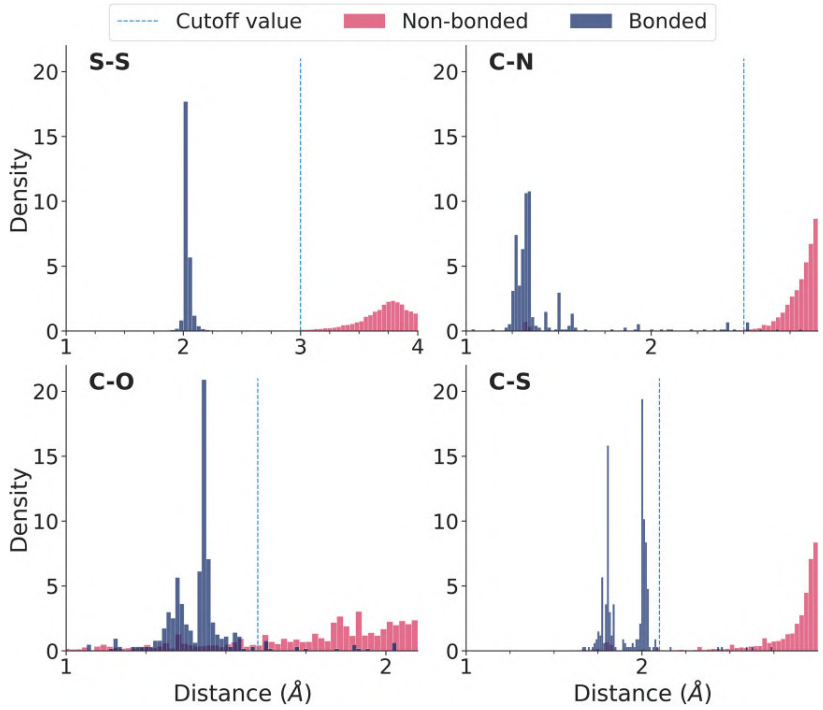


Figure 1. Distribution of distances between bonded and non-bonded atom pairs. An analysis of all protein structures deposited in the PDB was used to establish a distance threshold for the detection of cysteine (S–S), amide (C–N), ester (C–O), and thioester (C–S) bonds.

Table 1. Performance metrics for the distance method

Testing set				Performance metrics			
Bond type	#Samples	#Bonded	#Non-bonded	Accuracy	Recall	Specificity	Precision
Cysteine	739 528	476 221	263 307	0.99	1.00	0.98	0.99
Amide	52 765	416	52 349	0.95	0.98	0.95	0.12
Ester	8456	460	7996	0.73	0.96	0.72	0.16
Thioester	18 810	310	18 500	0.93	0.98	0.93	0.19

Samples were extracted from all protein structures available in the PDB. Presence of the bond is known based on the LINK annotation in the PDB file.

Table 2. Performance metrics for the SVM method

Set	Class	Count	Metrics
Training	#Bonded	8804	Accuracy: 0.99
	#Non-bonded	8804	Recall: 1.00
Testing	#Bonded	2201	Specificity: 0.99
	#Non-bonded	2201	Precision: 0.99

Samples were obtained from a non-redundant PDB database generated by Vector Alignment Search Tool (VAST) [32] at a *P*-value cutoff of 10^{-7}. Structures lacking annotated disulfide bonds were excluded to avoid annotation errors and ensure oxidative relevance. To balance the dataset, trivial non-bonded cysteine pairs (C_α distance >10 Å) were removed, and redundant pairs were randomly reduced.

from another crossing (unless both crossings are results of piercing the loop from the same side).

Additionally, Gauss linking number (GLN) values [33] are computed between the loop and each tail residue, creating a GLN map that is displayed on the web server interface. If the input file contains the predicted Local Distance Difference Test (pLDDT) data in the b-factor column (e.g. mmCIF files from the AlphaFold model), it is used for the assessment of the structure and the topological prediction validity. The topology is calculated based on the chain simplified to just the C_α atoms.

Lasso types

Based on the naming conventions described precisely in [4, 17], we divide lassos into four main types:

- one-sided lassos: L_1, L_2, L_3, …
- supercoiled one-sided lassos: LS_2, LS_3, LS_4, …
- two-sided lassos: $LL_{1,1}$, $LL_{1,3}$, $LL_{4,2}$, …
- supercoiled two-sided lassos: $LSL_{2,1}$, $LLS_{1,3}$, $LSLS_{4,2}$, …

Supercoiled lassos are created when a tail winds around the loop and crosses it from the same side (at least) twice in a row. One-sided lassos are pierced by only one tail, which can be specified by an additional "N" or "C" behind a number, e.g. L_{2N}, L_{4C}. In the case of two-sided lassos, the first number refers to the N-tail, and the other to the C-tail. LSL indicates the supercoiling of the N-tail, LLS supercoiling of the C-tail, and LSLS that both tails supercoil. Additionally, we distinguish loop sides, which induces crossing orientations: + and −. Therefore, the lasso type can be even more specific, by having a sequence of + and − behind a number, e.g. $LSLS_{3++-,2--}$.

Note that consecutive tail piercings of a non-supercoiled lasso must alternate; therefore, to shorten the notation, we leave only the first sign, e.g. $L_{3+-+N} \equiv L_{+3N}$.

Database of pre-calculated structures

The AlphaLasso server also includes a database of pre-analyzed AlphaFold models containing cysteine bridges. This database consists of all high-quality (pLDDT $\geq$ 70) AlphaFold-predicted protein structures for which cysteine bridges have been detected. Models obtained from AlphaFold were processed using the pipeline available on the submission server.

Graphical content

The web server uses Mol* [34] for visualization of the structure with a custom plugin allowing the display of the lasso surface. A clear interface is provided, facilitating operations on the structure visualization, changing the color scheme, representation style, and visibility of the lasso surface.

Using the Matplotlib [35] library for Python, a GLN map can be generated for each submitted structure, providing a detailed look into the lasso topology. The GLN map represents GLN values calculated between a loop and every possible subchain of each tail—map indices define the endpoints of each subchain. For a more exhaustive explanation, please check [33]. The map is interactive—by pressing a point on the map, a corresponding tail subchain is marked in the 3D structure view.

Server and database technical description

The server is developed in Python using the Flask framework, with server-side rendered HTML pages, and runs on Apache2 with WSGI. The data are stored using a PostgreSQL 14 database. Proteins' metadata are downloaded from PDB, UniProt, and InterPro using RESTful services. The whole service is installed on multicore Linux nodes and uses the *kafka-slurm-agent* [36] to run the lasso detection on a SLURM-managed Linux cluster as well as individual workstations. The same tool (*kafka-slurm-agent*) was used to scale the detection process and run it on around 178 million AlphaFold-predicted protein structures—the subset of the whole AlphaFold DB that has pLDDT >70.

Results

Submission server

The server performs an analysis of the entanglement of the input structure, which can be given as an mmCIF or a PDB file. All mmCIF/PDB files containing a single continuous polypeptide chain are accepted as valid inputs. Submitting files containing the pLDDT metric (e.g. generated by the AlphaFold model) will yield additional information accounting for this metric. The server facilitates the analysis of various types of closed loops: (i) the so-called non-covalent lassos [37] defined by the user, (ii) bonds between user-defined amino acids with a specified distance threshold, and (iii) covalent bonds detected automatically. Multiple files can be uploaded simultaneously, or, depending on their complexity, up to 15 models in a single file of the PDB format, creating a multi-model view enabling easier model comparison or study of molecular dynamics. Users can control the detection and analysis process by modifying the input parameters available on the submission page, as seen in Fig. 2. The features available there can be used to facilitate the analysis of non-biological polymers.

The output of each analysis is a detailed interactive web page with a variety of visualizations of the structure, including the 3D model, barycentric view, and a GLN map. This map provides insight into how the tail is positioned relative to the loop's surface. Each structure containing pLDDT information is passed through the *artifact check* functionality, resulting in a detailed summary of the prediction quality of each lasso feature. An overview of these features is presented in Fig. 3.

Users can also schedule a similarity check between proteins already available in the database. In that case, MMseqs2 [38] and Foldseek [39] are used to compute the sequential and structural identity (a user-specified parameter that defaults to 70%), and the user is presented with a list of similar lasso proteins from the AlphaLasso database along with their sequential/structural identity to the analyzed protein and links pointing to the respective subpages.

Database

The database part of AlphaLasso contains >14 million protein structures with cysteine bridges. This includes >2 million structures that possess nontrivial (non-L_0) topologies. The catalog provides an overview similar to that of a user-initiated analysis. It allows for quick access to the vast repertoire of AlphaFold-generated structures without the need to run individual tasks on the submission server. Each protein in the database can be easily recomputed to search for additional bridges, generate the GLN map, or find similar proteins. With such a vast amount of data, manual verification is not possible; thus, the service allows users to leave comments and votes (choosing "Lasso," "Artifact," or "Unsure") on each model. Our team also manually annotates the entries and includes them in the "Reviewed" category.

Each available model can be evaluated using the same functionality as for the user-provided structures on a page consisting of two tabs: one with the latest AlphaFold model (*Lassos – AF v4*) and the *Protein information* tab. The details of the model can be found in the *Model information* tab. The information tabs provide a single source of reference to the UniProt, InterPro, AlphaFold, and AlphaKnot databases. The AlphaKnot database [40] also provides links to AlphaLasso proteins that have both knots and lassos. There are currently 72 312 such protein models. In AlphaLasso, users can use the "Knotted" button on the *Browse database* page to view all knotted lasso proteins or use the "Knot type" in the *Advanced Search* to filter the results for specific knot types.

The structures already analyzed and present in the database section include new types of lassos divided into four categories, which are reflected in the letters in their names and the numbers that represent the piercing: simple lassos (L—one terminal is piercing the closed loop), two-sided (LL), supercoiling (LS), the two-sided + supercoiling (LLS, LSL, LSLS), and composition of lasso and knot topology (LK or L#K). In the database, there are 28 types of simple lassos (L_1–L_{26}, L_{28}, and L_{30}), which are found in 1 959 216 structures; 25 types of two-sided lasso with 123 179 structures; 25 types of supercoiling with 106 376 structures; and 230 types of two-sided + supercoiling with 22 715 structures. The user can also

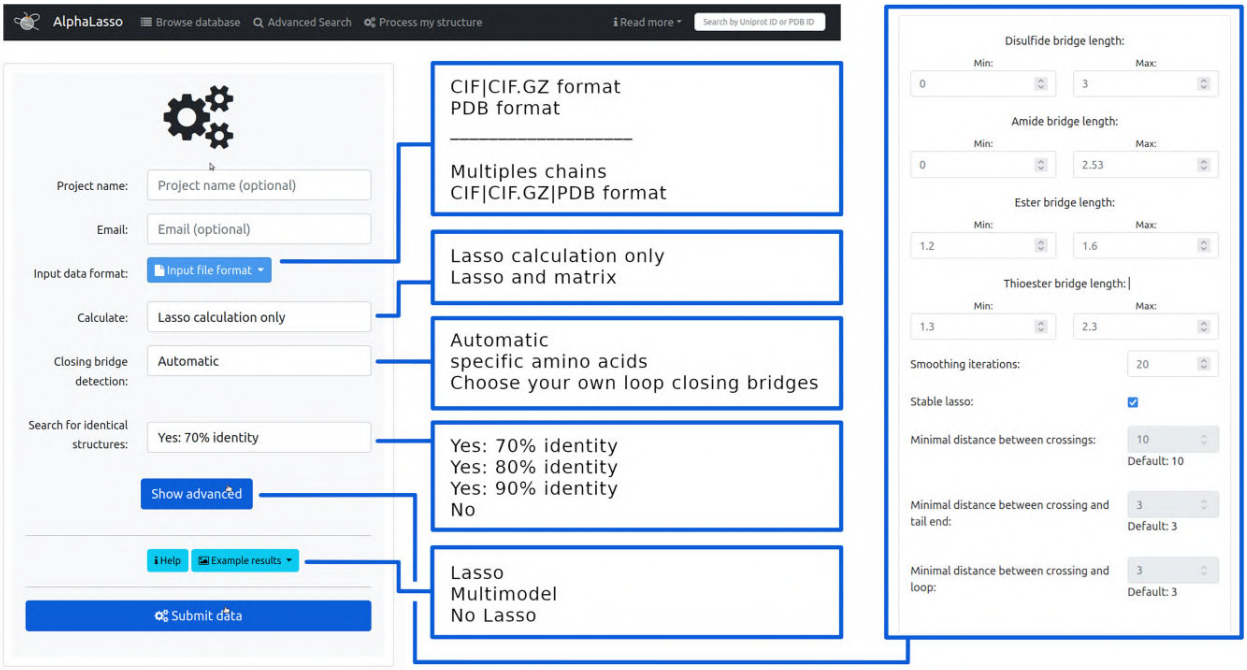


Figure 2. Submission options. Users can choose how AlphaLasso processes the structure, allowing for manual bridge positioning and identity search threshold specification. Advanced options bring the ability to change bridge detection thresholds, smoothing iterations, and give control over the lasso topology definition. Results of this analysis are shown in Fig. 3.

find 13 624 651 structures that we identify as trivial lassos, annotated by L_0.

Distinctive features of AlphaLasso

Visualization of a simplified structure

Since lasso topologies tend to hold complex and abstract shapes, a smoothing operation can be applied by the user to the protein structure, creating an easy-to-understand image of the arrangement of the lasso motif. Since structures vary in their complexity, users can fine-tune the smoothing process during the structure submission to obtain an image best suited to their needs. The Topoly Python package is utilized for this operation.

Analysis of topology of homologous proteins

A single amino acid mutation can drastically change the lasso motif. In contrast to the highly conserved knot motif [41], lassos are not strictly conserved in protein families. Proteins in a single family can have different lasso motifs or even lack them entirely. With the protein data already stored in the database, users can compare their structures to the already analyzed AlphaFold protein models featuring lasso topology. Additionally, by supplying multiple files with different structures, the user is presented with a graph comparing the lassos' locations and types.

Information about >14 million lasso AlphaFold models

Compared to the closest comparable database featured in the LassoProt web server [11], which currently holds 9000 lasso entries, the AlphaLasso expands the known universe of lasso proteins over 200 times. Such growth of available data improves the ability to study the origins and function of these topologies.

Extensive database filtering

The vast number of lasso models in the AlphaLasso database provides an opportunity for large-scale analyses of lasso proteins. For convenient access to the data we store, the user can specify precisely what they seek by using the extensive *Advanced search* functionality, which applies filters on the database entries and is also available as an Application Programming Interface (API). For example, the user can specify lasso types, pLDDT structure prediction quality metrics (for the whole chain or the deep or shallow lasso loop), or parts of the protein's name and customize the output columns, adding relevant information. The output format can be easily changed to produce a downloadable CSV file.

Comparison to other services

To the best of our knowledge, AlphaLasso is the only web server allowing for the analysis of AI-predicted biopolymer 3D structures that lack annotated information about inter-residue bonds. The algorithm searches for the possible closing bridges automatically.

Compared to the LassoProt web server [11] presented in 2016, the AlphaLasso Web Server analyzes new kinds of structural data. LassoProt depends on the annotations in PDB files provided by experimental methods. Additionally, while both servers provide the option to calculate the topology of a given protein structure, AlphaLasso boasts more options to choose from, including more detailed parameters, the incorporation

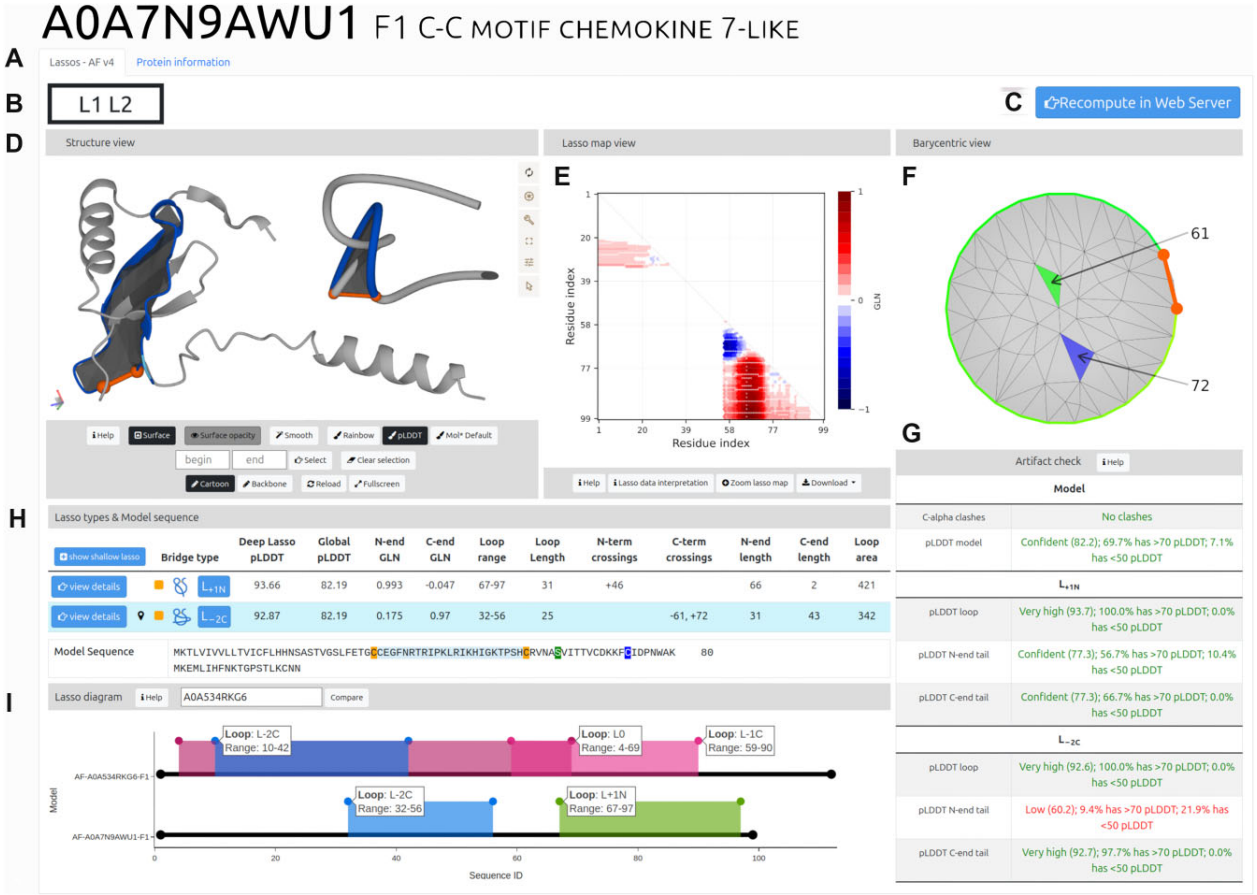


Figure 3. Protein analysis result. At the top of the page is the UniProtKB ID of the protein, model number, and protein name. (**A**) Button to alternate between lasso information and protein information. The protein information tab contains the organism and cross-references to other databases. (**B**) Lasso fingerprint of the model. (**C**) Button to recompute the structure with non-default settings. (**D**) Panel with a 3D "structure view" of the model. There is also a simplified, smoothed version of the 3D structure shown. (**E**) Map of the GLN values for the whole protein structure. (**F**) Barycentric view of the minimal surface of a lasso loop. (**G**) Artifact check for clashes and pLDDT scores of lassos. (**H**) Summary of found lassos and amino acid sequence of the modeled protein. (**I**) Tool to show and compare ranges of each lasso loop. One can add here another model (top) and compare it with a currently browsed model (bottom).

of the pLDDT metric, and the use of modern visualization tools.

The Topoly Python package, on which AlphaLasso is based, requires Python knowledge and is operated using a terminal, making access limited to skilled individuals. AlphaLasso eliminates this barrier by providing a user-friendly interface and adding several visualizations not included in this package.

We are aware of the existence of other tools facilitating disulfide bond discovery, such as DISULFIND [42], DiANNA [43], DBCP [44], and DisLocate [45]. However, these tools base their predictions on the amino acid sequence, while AlphaLasso predictions are based on the already available 3D structure. DISULFIND and DBCP are no longer available as web servers. AlphaLasso provides a modern interface with extensive capabilities supporting more types of bonds, surpassing the functionality of the aforementioned tools.

New lasso topologies

AlphaLasso has broadened the scope of lasso topologies. The new L_4 lasso topology in proteins, characterized by a loop pierced four times by a single tail, represents a significant expansion of known lasso structures [46]. Specifically, two distinct variations, L_{+4C} and L_{-4C} (Fig. 4A and B), have been identified by a comprehensive analysis of the AlphaFold protein models within the AlphaLasso database. The L_{+4C} topology is prominently observed in the glycosyl hydrolase family 13, particularly within bacterial alpha-amylases, where it appears to be a highly conserved characteristic. However, the L_{-4C} topology is found in a subset of xylosyltransferase proteins, predominantly within the eukaryota. This discovery shows us that AlphaLasso and AlphaFold are not just topological and structural predictors, but powerful tools for discovering entirely new protein topologies and structures.

Extending our investigation, one can find proteins with the intersection of lasso and knotted motifs (such structures are shown in the database under the keyword knot in the search section). Two protein families exhibit a combination of lasso and 3_1 knot topologies, the most common knot type. In the sodium/calcium exchanger family, a single-pierced lasso (L_{+1C}) is intertwined with a 3_1 knot within the transmembrane core domain (Fig. 4D), suggesting a cooperative structural role. In contrast, calcium-activated potassium channels have separate lasso (L_{+3C}) and knot domains (Fig. 4C), indicating independent structural contributions. Finally, the database also contains the most complex lasso motif (LS_4LS_3; Fig. 4)

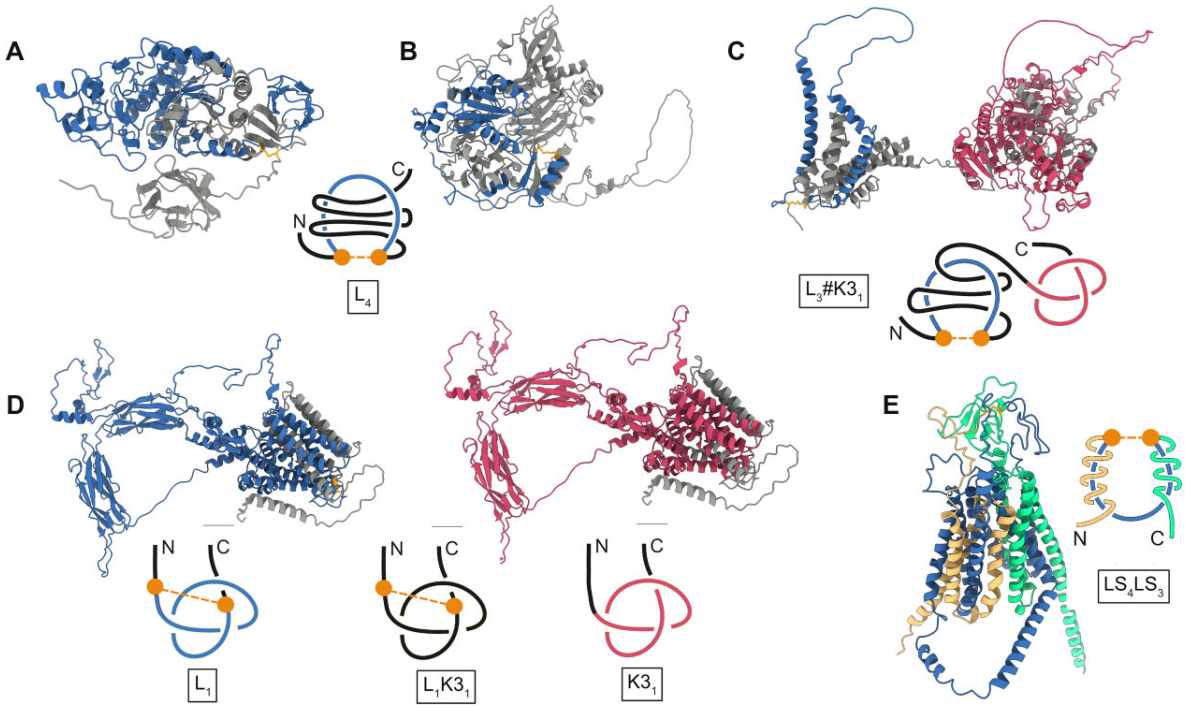


Figure 4. Proteins with a new lasso type. (**A**) Left: L_{+4C} lasso in alpha-amylase protein. Cartoon and schematic representation of the AlphaFold model of *Escherichia coli* alpha-amylase (UniProtKB ID: P25718). The blue color highlights the lasso loop. The orange beads and the line connecting them represent a cysteine bridge forming the lasso loop. (**B**) Right: L_{-4C} lasso in xylosyltransferase 1 protein from *Homo sapiens* (UniProtKB ID: Q86Y38). (**C**) L_{+3C} lasso and a 3_1-knot in a calcium-activated potassium channel. Cartoon and schematic representation of the AlphaFold model of the human calcium-activated potassium channel subunit alpha-1 (UniProtKB ID: Q12791). (**D**) L_{+1C} lasso intertwined with a 3_1 knot in a sodium/calcium exchanger protein. Cartoon and schematic representation of the AlphaFold model of the human sodium/calcium exchanger 1 protein (UniProtKB id: P32418). (**E**) Schematic representation of the protein with the most complex lasso topology (LS_4LS_3). AlphaFold model of Q9Y6L6 protein.

found in organic anion transporting polypeptides (OATPs, InterPro ID: IPR004156), a double supercoiled lasso. This motif, particularly in the human OATP1B1 protein (UniProtKB ID: Q9Y6L6), shows how AlphaFold and AlphaLasso can reveal intricate topological features, surpassing even the complexity observed in some experimental structures. Examples of the new motifs presented here are described in [46]. There are other, even more complex motifs in the database that require a detailed analysis to confirm their existence.

Summary

Currently available AI methods allow for the generation of thousands of new protein structures intended to enhance their physical and chemical properties, and biological function. Furthermore, with the improved stability of polymers featuring disulfide bonds and non-trivial topology, further study of those structures is a valid research direction, since understanding their stability and functional properties could lead to advancements in drug design and biomaterials. AlphaLasso is a web server designed to automatically analyze large groups of proteins, offering a detailed visualization of lasso-type entanglement in user-provided structures, as well as a database of those topologies in AlphaFold-predicted protein models. This service offers several new features that help in the analysis of the topology and aid the user evaluate the topological correctness of their query.

The web server gives detailed topological information on the structure (provided by the user as either an mmCIF- or a PDB-formatted file), as well as a visual guide to understand it with a detailed output provided by the Topoly package. Additionally, it is integrated with the AlphaLasso database to provide information about lasso types of proteins (from the database) sequentially similar to the query. It enables a deeper study of homologous proteins, giving insight into their conservation in the family. Most of the web server's output is available for all proteins present in the database, while all other features can be quickly computed via a two-click job submission button.

Compared to the already-known lasso motifs available in the PDB structures (simple lassos L_1, L_2, L_3, and L_6; two-sided lassos $LL_{1,1}$, $LL_{1,2}$, $LL_{2,1}$, and $LL_{4,3}$; and supercoiling LS_2, LS_3, and LS_4), AlphaLasso presents nearly 300 potentially new types of lasso motifs in AlphaFold-predicted proteins, opening up new potential for (i) understanding the impact of non-trivial topology on the physicochemical properties of proteins and the effect on biological function and (ii) applying these potentially nature-associated motifs to create artificial connections in proteins or other biopolymers to achieve the desired effect.

Acknowledgements

Author contributions: Pawel Rubach (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Writing—original draft, Writing—review & editing), Jacek Płonka (Conceptualization, Data curation, Formal analysis, Investigation, Method-

ology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Bartosz A. Gren (Data curation, Formal analysis, Investigation, Validation, Visualization, Writing—review & editing), Fernando Bruno da Silva (Data curation, Formal analysis, Investigation, Validation, Visualization, Writing—review & editing), Marta Korpacz (Formal analysis, Investigation, Methodology, Validation, Visualization, Writing—review & editing), and Joanna I. Sulkowska (Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing—review & editing).

Conflict of interest

None declared.

Funding

This work was supported by the National Science Centre 2022/47/B/NZ1/03480 (to J.I.S.). Funding to pay the Open Access publication charges for this article was provided by National Science Centre (2022/47/B/NZ1/03480).

Data availability

AlphaLasso is available at https://alphalasso.cent.uw.edu.pl/. It is free and open to all users without a log-in requirement.

References

1. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2
2. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574
3. Baek M, DiMaio F, Anishchenko I *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6. https://doi.org/10.1126/science.abj8754
4. Niemyska W, Dabrowski-Tumanski P, Kadlof M *et al.* Complex lasso: new entangled motifs in proteins. *Sci Rep* 2016;**6**:36895. https://doi.org/10.1038/srep36895
5. Wang CK, Kaas Q, Chiche L *et al.* CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Res* 2007;**36**:D206–10. https://doi.org/10.1093/nar/gkm953
6. Iwai H, Plückthun A. Circular β-lactamase: stability enhancement by cyclizing the backbone. *FEBS Lett* 1999;**459**:166–72. https://doi.org/10.1016/S0014-5793(99)01220-X
7. Daly NL, Craik DJ. Bioactive cystine knot proteins. *Curr Opin Chem Biol* 2011;**15**:362–8. https://doi.org/10.1016/j.cbpa.2011.02.008
8. Dombkowski AA, Sultana KZ, Craig DB. Protein disulfide engineering. *FEBS Lett* 2014;**588**:206–12. https://doi.org/10.1016/j.febslet.2013.11.024
9. Niewieczerzał S, Sulkowska JI. Supercoiling in a protein increases its stability. *Phys Rev Lett* 2019;**123**:138102. https://doi.org/10.1103/PhysRevLett.123.138102
10. Jiang Y, Deane CM, Morris GM *et al.* It is theoretically possible to avoid misfolding into non-covalent lasso entanglements using small molecule drugs. *PLoS Comput Biol* 2024;**20**:e1011901. https://doi.org/10.1371/journal.pcbi.1011901
11. Dabrowski-Tumanski P, Niemyska W, Pasznik P *et al.* LassoProt: server to analyze biopolymers with lassos. *Nucleic Acids Res* 2016;**44**:W383–9. https://doi.org/10.1093/nar/gkw308
12. Haglund E, Pilko A, Wollman R *et al.* Pierced lasso topology controls function in leptin. *J Phys Chem B* 2017;**121**:706–18. https://doi.org/10.1021/acs.jpcb.6b11506
13. Perego C, Potestio R. Searching the optimal folding routes of a complex lasso protein. *Biophys J* 2019;**117**:214–28. https://doi.org/10.1016/j.bpj.2019.05.025
14. Salicari L, Baiesi M, Orlandini E *et al.* Folding kinetics of an entangled protein. *PLoS Comput Biol* 2023;**19**:e1011107. https://doi.org/10.1371/journal.pcbi.1011107
15. Simien JM, Haglund E. Topological twists in nature. *Trends Biochem Sci* 2021;**46**:461–71. https://doi.org/10.1016/j.tibs.2020.12.004
16. Tubiana L, Alexander GP, Barbensi A *et al.* Topology in soft and biological matter. *Phys Rep* 2024;**1075**:1–137.https://doi.org/10.1016/j.physrep.2024.04.002
17. Greń BA, Dabrowski-Tumanski P, Niemyska W *et al.* Lasso proteins—unifying cysteine knots and miniproteins. *Polymers* 2021;**13**:3988. https://doi.org/10.3390/polym13223988
18. Maksimov MO, Pan SJ, Link AJ. Lasso peptides: structure, function, biosynthesis, and engineering. *Nat Prod Rep* 2012;**29**:996–1006. https://doi.org/10.1039/c2np20070h
19. Cheng C, Hua ZC. Lasso peptides: heterologous production and potential medical application. *Front Bioeng Biotechnol* 2020;**8**:571165. https://doi.org/10.3389/fbioe.2020.571165
20. Braffman NR, Piscotta FJ, Hauver J *et al.* Structural mechanism of transcription inhibition by lasso peptides microcin J25 and capistruin. *Proc Natl Acad Sci USA* 2019;**116**:1273–8. https://doi.org/10.1073/pnas.1817352116
21. Clavel C, Fournel-Marotte K, Coutrot F. A pH-sensitive peptide-containing lasso molecular switch. *Molecules* 2013;**18**:11553–75. https://doi.org/10.3390/molecules180911553
22. Allen CD, Chen MY, Trick AY *et al.* Thermal unthreading of the lasso peptides astexin-2 and astexin-3. *ACS Chem Biol* 2016;**11**:3043–51. https://doi.org/10.1021/acschembio.6b00588
23. Zong C, Wu MJ, Qin JZ *et al.* Lasso peptide benenodin-1 is a thermally actuated [1]rotaxane switch. *J Am Chem Soc* 2017;**139**:10403–9. https://doi.org/10.1021/jacs.7b04830
24. Zong C, Maksimov MO, Link AJ. Construction of lasso peptide fusion proteins. *ACS Chem Biol* 2016;**11**:61–8. https://doi.org/10.1021/acschembio.5b00745
25. Knappe TA, Linne U, Robbel L *et al.* Insights into the biosynthesis and stability of the lasso peptide capistruin. *Chem Biol* 2009;**16**:1290–8. https://doi.org/10.1016/j.chembiol.2009.11.009
26. Dabrowski-Tumanski P, Rubach P, Niemyska W *et al.* Topoly: Python package to analyze topology of polymers. *Brief Bioinform* 2021;**22**:bbaa196. https://doi.org/10.1093/bib/bbaa196
27. da Silva FB, Lewandowska I, Kluza A *et al.* First crystal structure of double knotted protein TrmD-Tm1570—inside from degradation perspective. bioRxiv, https://doi.org/10.1101/2023.03.13.532328, 14 March 2023, preprint: not peer reviewed.
28. Hsu MF, Sriramoju MK, Lai CH *et al.* Structure, dynamics, and stability of the smallest and most complex 71 protein knot. *J Biol Chem* 2024;**300**:105553. https://doi.org/10.1016/j.jbc.2023.105553
29. Perlinska AP, Niemyska WH, Gren BA *et al.* AlphaFold predicts novel human proteins with knots. *Protein Sci* 2023;**32**:e4631. https://doi.org/10.1002/pro.4631
30. Terwilliger TC, Liebschner D, Croll TI *et al.* AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat Methods* 2024;**21**:110–6. https://doi.org/10.1038/s41592-023-02087-4
31. Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
32. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol* 1996;**6**:377–385. https://doi.org/10.1016/S0959-440X(96)80058-3
33. Niemyska W, Millett KC, Sulkowska JI. GLN: a method to reveal unique properties of lasso type topology in proteins. *Sci Rep* 2020;**10**:15186. https://doi.org/10.1038/s41598-020-71874-2

34. Sehnal D, Bittrich S, Deshpande M *et al.* Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res* 2021;**49**:W431–7. https://doi.org/10.1093/nar/gkab314
35. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 2007;**9**:90–5. https://doi.org/10.1109/MCSE.2007.55
36. Rubach P. Applying large-scale distributed computing to structural bioinformatics – bridging legacy HPC clusters with big data technologies using kafka-slurm-agent. arXiv, https://arxiv.org/abs/2503.14806, 19 March 2025, preprint: not peer reviewed.
37. Jiang Y, Neti SS, Sitarik I *et al.* How synonymous mutations alter enzyme structure and function over long timescales. *Nat Chem* 2023;**15**:308–18. https://doi.org/10.1038/s41557-022-01091-z
38. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8. https://doi.org/10.1038/nbt.3988
39. Van Kempen M, Kim SS, Tumescheit C *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2024;**42**:243–6. https://doi.org/10.1038/s41587-023-01773-0
40. Rubach P, Sikora M, Jarmolinska A *et al.* AlphaKnot 2.0: a web server for the visualization of proteins' knotting and a database of knotted AlphaFold-predicted models. *Nucleic Acids Res* 2024;**52**:W187–93. https://doi.org/10.1093/nar/gkae443
41. Sulkowska JI, Rawdon EJ, Millet KC *et al.* Conservation of complex knotting and slipknotting patterns in proteins. *Proc Natl Acad Sci USA* 2012;**102**:E1715–23.
42. Ceroni A, Passerini A, Vullo A *et al.* DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res* 2006;**34**:W177–81.
43. Ferrè F, Clote P. DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine classification. *Nucleic Acids Res* 2006;**34**:W182–5.
44. Lin HH, Tseng LY. DBCP: a web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding state of cysteines. *Nucleic Acids Res* 2010;**38**:W503–7.
45. Savojardo C, Fariselli P, Alhamdoosh M *et al.* Improving the prediction of disulfide bonds in eukaryotes with machine learning methods and protein subcellular localization. *Bioinformatics* 2011;**27**:2224–30.
46. Bruno da Silva F, Perlinska AP, Płonka J *et al.* Universe of lasso proteins: exploring the limit of entanglement and folding landscape of proteins predicted by AlphaFold. bioRxiv, https://doi.org/10.1101/2025.03.21.644650, 22 March 2015, preprint: not peer reviewed.

Received: March 21, 2025. **Revised:** April 20, 2025. **Editorial Decision:** April 22, 2025. **Accepted:** April 23, 2025

Article 3

Conservation of knotted and slipknotted topology in transmembrane transporters

Biophysical Journal
Article

Biophysical Society

Conservation of knotted and slipknotted topology in transmembrane transporters

Vasilina Zayats,[1] Maciej Sikora,[1,2] Agata P. Perlinska,[1] Adam Stasiulewicz,[1,3] Bartosz A. Gren,[1] and Joanna I. Sulkowska[1,*]

[1]Centre of New Technologies, University of Warsaw, Warsaw, Poland; [2]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland; and [3]Department of Drug Chemistry, Faculty of Pharmacy, Medical University of Warsaw, Warsaw, Poland

ABSTRACT The existence of nontrivial topology is well accepted in globular proteins but not in membrane proteins. Our comprehensive topological analysis of the Protein Data Bank structures reveals 18 families of transmembrane proteins with nontrivial topology, showing that they constitute a significant number of membrane proteins. Moreover, we found that they comprise one of the largest groups of secondary active transporters. We classified them based on their knotted fingerprint into four groups: three slipknotted and one knotted. Unexpectedly, we found that the same protein can possess two distinct slipknot motifs that correspond to its outward- and inward-open conformational state. Based on the analysis of structures and knotted fingerprints, we show that slipknot topology is directly involved in the conformational transition and substrate transfer. Therefore, entanglement can be used to classify proteins and to find their structure-function relationship. Furthermore, based on the topological analysis of the transmembrane protein structures predicted by AlphaFold, we identified new potentially slipknotted protein families.

SIGNIFICANCE Membrane proteins are a vast and frequently analyzed group of proteins. However, the study of their topology rarely includes entanglements. Herein, we show that α-helical transmembrane proteins possess nontrivial topology: a slipknotted and knotted protein backbone. Although they constitute only 1% of the total number of transmembrane proteins, we show that they form about one-fourth of all known secondary active transporters in mammals. Moreover, the knotted region is often involved in biological function. We found that during the transport the slipknot motif undergoes conformational changes along with the protein. Therefore, we can classify the state of the proteins by analyzing their slipknot or knot topology. Overall, our results show that entanglement is important in any analysis of transmembrane proteins.

INTRODUCTION

Membrane proteins are an abundant group of proteins that constitute 30% of most of the proteomes; however, only 10% of them have a resolved structure (1,2). These proteins play different biological roles and are crucial targets for medical applications. Herein, we focus on their often overlooked aspect—the entanglement of the proteins' backbone (3), with the main focus on the transmembrane (TM) proteins. Based on proteins with structures determined experimentally (4) and predicted by machine learning (ML) methods (5,6), we reveal that a significant number of TM proteins are entangled, with their backbone forming a knot (3) or slipknot (7,8) (Fig. 1). A slipknot is created when a knot is formed by a part of a protein chain, after which the backbone doubles back so that the entire structure becomes unknotted (in a mathematical sense) (7). Herein, we aim to show that entanglement (nontrivial topology) can serve as a new descriptor to classify TM proteins and to find their structure-function relationship.

The simplest way to find a knot in a protein structure is to pull apart N- and C-termini in opposite directions. The resulting string will have an easily identifiable knot. When the protein is unknotted or possesses a slipknot, the line will be straight, showing that only knots can be found with this simple method. In order to find slipknots, we need to use the extension of this method, the so-called matrix approach (Fig. 1) (7,8). Given that a slipknot has an

Submitted February 21, 2023, and accepted for publication October 30, 2023.

*Correspondence: jsulkowska@cent.uw.edu.pl

Editor: Philip Biggin.

https://doi.org/10.1016/j.bpj.2023.10.031

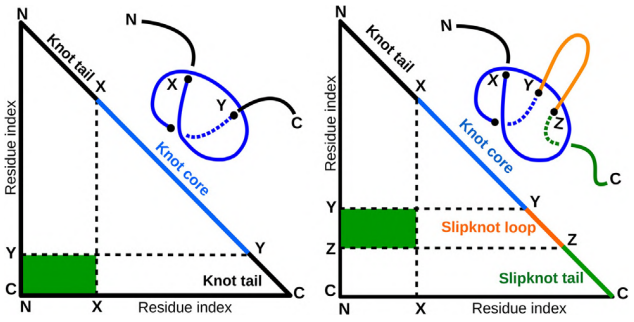


FIGURE 1 Matrix presentation of protein knotting fingerprint. Each entry in the matrix indicates the knot type formed by one continuous subchain shown by green color (or by orange for 4_1 knot in, e.g., Figs. 5 and 6). In each case, the subchain starts with the N-terminal amino acid at position x and ends with the C-terminal amino acid at position y, indicated on the horizontal and vertical axes, respectively. Equivalently, this subchain can be interpreted as a part of the diagonal, delimited by the corresponding coordinates x and y, where the entire diagonal corresponds to the entire protein backbone. Left, an example of knotted (a 3_1 knot) protein. The green entry in the lower left-hand corner, which corresponds to the entire protein, indicates that the corresponding chain or subchain is knotted. The knot core is defined as the shortest subchain that still forms a knot. The two remaining parts of the chain form knot tails. Right, an example of slipknotted protein. Notice that in the case of slipknots, the entire protein is unknotted (the element in the lower corner is not colored), but as one terminal is trimmed to some extent, the remaining fragment forms a trefoil knot, denoted $S3_1$. The corresponding subchain entries are therefore colored in the matrix. Above diagonal, schematic drawings of trefoil knot and slipknot illustrate which parts of the chains constitute the knot cores (*blue line*), the knot and slipknot tails (*black line*), and the slipknot loop (*orange line*). More details about the fingerprint can be found in (8). To see this figure in color, go online.

internal knot, the analysis of all of the protein's subchains is needed. If some of them are knotted but the whole structure is unknotted, the protein has a slipknot (for more details, see Fig. 1). This approach also helps to fully understand the knotting of the protein backbone. Identification of all knots along the protein backbone allows the assignment of a knot fingerprint to the protein (a pattern in a matrix diagram characterizing protein knotting) and comparison between different proteins (8). It was shown that homologous proteins (even with very low sequence similarity) possess exactly the same unique pattern of internal knots and that slipknots and knots are strongly conserved during evolution. These methods were also used to characterize knot behavior in proteins (9,10), polymers, and a synthetic molecular knot (11). In proteins, the knot's ends are found to jump to well-defined sequential locations along the protein backbone, whereas in homopolymers, they diffuse around and eventually slide off (12,13).

The topic of entanglement in proteins has already been studied from many angles; however, the studies rarely focus on membrane proteins even though they have been known for some time. When the presence of slipknots was first discovered in proteins in 2006, a membrane protein was one of them (7). It was a member of amino acid-polyamine-organocation (APC) transporters with the most complex fingerprint, composed of two trefoil (3_1) slipknots and one figure-eight (4_1) slipknot. In 2019, new slipknotted TM proteins as well as the first knotted TM proteins were discovered (3) (the information about the topology of these structures can be found in KnotProt 2.0 database (4)). This lack of thorough investigations is surprising given that the entanglement in globular proteins has been extensively explored for over a decade, including different typologies such as knots, slipknots, links (14), lassos (15), and θ-curves (16), which helped to understand their folding, degradation, and potential biological role of entanglement (17,18) (we refer those interested to recent review papers (19,20)). Membrane proteins with nontrivial topology, such as leucine transporter (LeuT) and uracil transporters, were analyzed from different points of view—folding and unfolding pathways, ions transport, up to the application as drug targets (21–25). However, none of the research took into account nontrivial topology, which can explicitly restrict the conformational freedom of a protein's backbone and thus influence its kinetics, thermodynamics, and biological function.

The existence of the entanglement in membrane proteins raises many intriguing questions: what is the function of nontrivial topology, and how can such proteins fold or be degraded? Also, why and how did the knots and slipknots emerge during evolution, whether the nontrivial topology is more typical for globular or TM proteins, and whether it is also conserved between different organisms. Herein, we show all membrane proteins with nontrivial topology discovered up to now (including predicted structures) to facilitate the process of finding answers to these questions. We also analyze them to understand the impact of entanglement on their structure and function. We identify a new protein family with a deep slipknot fingerprint—Sodium-transporting carboxylic acid decarboxylase protein family (Monovalent cation-Proton antiporter superfamily)—based on ML predictions. We classify all membrane proteins with nontrivial topology based on the knot fingerprint and then on the biological function. We are showing that in the case of the slipknot, the knot fingerprint is not always conserved and can be used to study structural changes related to the biological function.

MATERIALS AND METHODS

The topology detection

Topology information was collected from the KnotProt database (4) and AlphaKnot server (https://alphaknot.cent.uw.edu.pl). All of the structures predicted or deposited in the PDB were analyzed by computing the HOMFLY-PT polynomial for 100 random closures; when finding a nontrivial topology, 200 closures were used. The details of the method are explained in (46,47). The fingerprint method was used to determine the position of the knot core, knot tails, slipknot core, and loop based on methods explained here (48). Additionally, knot cores for structures were verified by calculating the Alexander polynomial. A structure is classified as knotted or slipknotted when random closures form a nontrivial knot more frequently than a trivial knot, the pLDDT score for the knot core is

above 70, there are no clashes in the knot core region (based on MolProbity tool (49)), the same knot type is found in more than 50% of the protein's homologs, and no suspicious geometry was found after visual inspection. Structures that did not pass visual inspection and those with obvious problems were rejected from our analysis. The knot fingerprint was determined (following the rules described above) based on the Topoly package (50).

All TM proteins in the Orientations of Proteins in Membranes (OPM) database were examined (26) and compared with the list of known nontrivial proteins to assign topology.

Data classification

Secondary active transporters were analyzed using solute carrier (SLC) tables. Matching between OPM results and SLC tables was made based on UniProt IDs as well as Pfam (51) family identifiers. This, however, is not pair matching, and certain OPM groups might have multiple or no SLC hits (as seen in Table 1).

From the matched SLC list, we extracted representative proteins. Each protein was then identified in the UniProt database and assigned various characteristics like transport types and family tags based on the Pfam database classification. Such a procedure allowed us to expand the so far limited data set for more reliable topology studies. In total, we found 61 Pfam families.

Data collection: PDB and AlphaFold

The topology assignment process requires a protein structure; thus, SLC-based data analysis consists of two separate parts: PDB structures and AlphaFold predictions. For each family, if PDB structures were present, we determined their topology as described above. Unfortunately, many families do not have any structures assigned, leaving an incomplete picture. For proteins with unknown structures, we used data deposited in the AlphaFold databases. We determined the topology of 8621 predictions representing all families from Fig. 3 with an extension of other members of the clan CL0062 (because of the large amount of slipknotted proteins). Those representatives were picked based on the following criteria. We performed sequential clustering with a 60% similarity threshold to remove redundancy and from each cluster picked one prediction with the highest value of the predicted local distance difference test (pLDDT) (but not lower than 70 pLDDT). Despite the filtering attempts, we report AlphaFold results separately because of the error-prone nature of artificial intelligence models.

Data evaluation

Data matching between databases and basic analysis were done using Python3. Sequence alignments were prepared with ClustalOmega (52) and ESPript (53). VMD, Pymol, and YASARA View were used for structure analysis and visualization (54,55).

RESULTS

Nontrivial topology only in α-helical transmembrane proteins

We used OPM (26) and KnotProt (4) databases to identify entangled structures among membrane proteins. The OPM database distinguishes three types of membrane proteins: TM, monotopic/peripheral, and peptides. We analyzed all resolved structures of these proteins available in PDB (4235 for TM, 1607 for monotopic/peripheral, and 726 for peptides) and found that entanglement is present only in the TM type. Moreover, within its three classes, we found entanglement in only one: α-helical polytopic class (Fig. 2). No slipknotted or knotted topologies were found among bitopic and β-barrel TM proteins (Fig. 2). Altogether, we identified 17 families with slipknot and one with knot topology (see details in Table 1).

To further classify these superfamilies, we used the knot fingerprint approach (defined in Fig. 1) and found four different motifs (Fig. 2): $K3_1$ (single trefoil knot), $S3_1$ (single trefoil slipknot), $S3_13_1$, and the most complex that involved a 4_1 knot, $S3_14_13_1$. Notice, that the notation $S3_13_1$ indicates that the entire protein chain forms a slipknot (denoted by S) with two disjoint 3_1 territories in the matrix presentation of the proteins (Fig. 1). The notation $S3_14_13_1$ indicates that the entire protein forms a slipknot with two 3_1 and one 4_1 territories in the matrix representation of the protein. Notice, that in the case of globular proteins, more complex 5_2 and 6_1 (27) knots exist.

Amino acid-polyamine-organocation (APC) superfamily (CL0062) is the largest group of TM proteins with slipknot topology. It gathers families with two different fingerprints ($S3_14_13_1$ and $S3_13_1$) that might indicate a common evolutionary origin for these two distinct slipknotted folds. On the other hand, the dicarboxylate/amino acid: cation symporter (DAACS) superfamily contains proteins with the same slipknot type but different fingerprints ($S3_1$ or $S3_13_1$) that depend on the protein's conformational state. Moreover, proteins from the superfamily of monovalent cation-proton antiporters possess the same single 3_1 slipknot, whose depth differs depending on the conformational state of the protein. To the best of our knowledge, the difference in the fingerprint motif across the same family was not observed before. This raises an intriguing question: how does the fingerprint correlate with the protein function?

All nontrivial transmembrane proteins are secondary active transporters

After finding that among membrane proteins only TM ones have nontrivial topology, we focused on their further classification. We established that all slipknotted and knotted TM proteins known to date are secondary active transporters. Given that it is a vast and diverse group of proteins, we analyzed how commonly the nontrivial topology occurs among them. For this analysis, we used SLC tables (the classification system of all known membrane-bound solute carrier transporter proteins in human and mammalian proteomes). The SLC lists 430 unique transporters at the date of publication (2017) (28). For each SLC transporter, we identified its family and clan based on the Pfam database. Fig. 3 shows all SLC transporters grouped by their clan.

The largest group of SLC is the major facilitator superfamily (CL0015). According to our analysis, all known protein structures from this clan possess a trivial topology. The

TABLE 1 OPM superfamilies with nontrivial topologies

OPM database			Pfam database				
Class	Superfamily	Family	Clan	Family	SLC database	Evidence	Knot fingerprint
α-Helical polytopic	Amino acid-polyamine-organocation (APC)	Neurotransmitter:Na^+ symporter	CL0062	PF00209	SLC6	experiment and AlphaFold	$3_1 4_1 3_1$ (LeuT)
		Anion exchanger family	CL0062	PF00955	SLC4	experiment and AlphaFold	$3_1 3_1$ (NCS2-like)
		Amino acid-polyamine-organocation (APC) family	CL0062	PF13520	SLC12	experiment and AlphaFold	$3_1 4_1 3_1$ (LeuT)
		Amino acid/auxin permease	CL0062	PF01490	SLC32 SLC36 SLC38	experiment and AlphaFold	$3_1 4_1 3_1$ (LeuT)
		Amino acid permease	CL0062	PF13520	SLC7 SLC12	experiment and AlphaFold	$3_1 4_1 3_1$ (LeuT)
		Solute:Na^+ symporter	CL0062	PF00474	SLC5	experiment and AlphaFold	$3_1 4_1 3_1$ (LeuT)
		Sulfate permease	CL0062	PF00916	SLC26	experiment and AlphaFold	$3_1 3_1$ (NCS2-like)
		Nucleobase:cation symporter-2	CL0062	PF00860	SLC23	experiment and AlphaFold	$3_1 3_1$ (NCS2-like)
		Divalent metal ion transporter (NRAMP)	CL0062	PF01566	SLC11	experiment and AlphaFold	$3_1 4_1 3_1$ (LeuT)
		Nucleobase:cation symporter-1	CL0062	PF02133	–	experiment and AlphaFold	$3_1 4_1 3_1$ (LeuT)
		Betaine/carnitine/choline transporters	CL0062	PF02028	–	experiment and AlphaFold	$3_1 4_1 3_1$ (LeuT)
		Cation-Cl^- cotransporter	CL0062	PF00324	SLC12	experiment and AlphaFold	$3_1 4_1 3_1$ (LeuT)
		K^+ uptake permease (KUP) family	CL0062	PF02705	–	AlphaFold	$3_1 4_1 3_1$ (LeuT)
		Alanine:cation symporter	CL0062	PF01235	–	experiment and AlphaFold	$3_1 4_1 3_1$ (LeuT)
	Dicarboxylate/amino acid:cation symporter (DAACS)	H^+:glutamate symporter	–	PF00375	SLC1	experiment and AlphaFold	3_1 $3_1 3_1$
	Monovalent cation-H^+ antiporter	2-Hydroxycarboxylate transporter	–	PF03390	–	experiment and AlphaFold	3_1
		Na^+ – H^+ antiporter 1	CL0064	PF00999	SLC9	experiment and AlphaFold	unknotted
		Bile acid:Na^+ symporter (BASS) family	CL0064	PF01758	SLC10	experiment and AlphaFold	unknotted
		Na^+ -transporting carboxylic acid decarboxylase	CL0064	PF03977	–	experiment and AlphaFold	3_1
		Na^+ – H^+ antiporter NhaA	CL0064	PF06965	–	experiment and AlphaFold	unknotted
		Auxin efflux carrier	CL0064	PF03547	–	experiment and AlphaFold	unknotted
		Na^+-dependent bicarbonate transporter	CL0064	PF05982	–	experiment and AlphaFold	unknotted
	Na^+ – Ca^{2+}exchanger	Na – Ca^{2+}exchanger	–	PF01699	SLC8 SLC24	experiment and AlphaFold	$K3_1$

The families without SLC ID are not present in mammals. All slipknots and knots possess the same chirality.

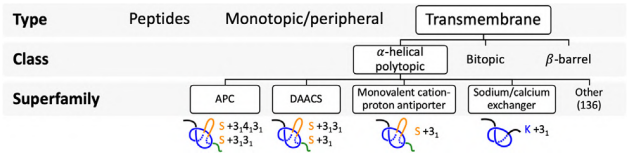


FIGURE 2 Classification and entanglement (type of topology) of membrane proteins based on OPM database. Proteins with nontrivial topology are found only within α-helical polytopic class of transmembrane proteins; they are part of four superfamilies (out of 140). Details about topology conservation in each of the entangled superfamilies are in Table 1. To see this figure in color, go online.

second largest group among all SLCs is the APC superfamily (CL0062), and all its members (with known structures) possess a slipknot topology. This is the largest known group of proteins with slipknot topology based on our results. Additionally, there are two families without clan ID with nontrivial topology: PF01699 (Sodium/calcium exchangers) with knot and PF00375 (Sodium:dicarboxylate symporters) with slipknot topology. Proteins from other smaller clans are either unknotted or their structures remain unknown (potentially possess trivial topology based on AlphaFold (5) and AlphaKnot (6), as we discuss in the next section). Although a lot of information is missing and many structures are yet to be solved, our analysis shows that approximately 25% of all SLC proteins possess nontrivial topology. In contrast, proteins with slipknot and knot topology are in general rare and comprise approximately 1.5% of all proteins (19).

SLC classification contains both facilitated and secondary active transporters. For example, major facilitator superfamily proteins that possess a trivial topology can catalyze transport in three different ways: symport, antiport, and uniport. However, most of them are uniporters, which are responsible for facilitated diffusion or primary active transport (26,28). On the other hand, all reported proteins with slipknot and knot topology are characterized as symporters or antiporters. These proteins catalyze the secondary active transport of various molecules (amino acids, neurotransmitters, cations, anions). A large fraction of SLC proteins with slipknots and knots might indicate that nontrivial topology is advantageous for the function of secondary active transporters.

AlphaFold predictions of transmembrane protein structures

As tertiary structures of many secondary active transporters have yet to be experimentally resolved, the determination of the topology for some of those protein groups remains an open challenge. One of the most important solutions is the in silico modeling (29). Recently, this approach has demonstrated a major rise due to the successful implementation of ML, with over 200 M protein structures predicted by AlphaFold (5,30) and more than 600 M by ESMFold (31). Herein, to identify the possible entanglements for families with no available structures (labeled as "no info" in

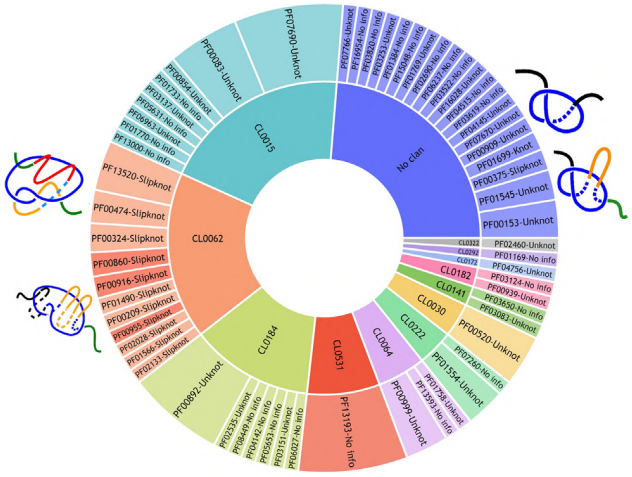


FIGURE 3 Analysis of all secondary active transporters in mammals and humans (SLC tables) shows that approximately 25% of all SLC proteins possess nontrivial topology. The data are grouped according to clans and families from the Pfam database. Additional details are provided for the APC superfamily (CL0062 by Pfam nomenclature). Specifically, NCS2-like proteins are marked with a darker color (the rest of the clan has the LeuT fold). All families without a clan are grouped together (*no clan* label). If there is a protein with resolved structure in the family, the topology (knotted/unknotted) is given. If a structure is not available, no information about the topology is given (*no info* label). To see this figure in color, go online.

Fig. 3), we utilized structures deposited in the AlphaFold database (5). Although predicting entanglement and new knot types might be considered challenging for ML methods due to the lack of a sufficient training set, recent data indicate that AlphaFold was able to accurately predict the first doubly knotted protein (PDB: 8B1N). We focused on SLC proteins and families from OPM superfamilies where the entanglement was detected (Table 1). We determined the knot fingerprint to identify the entire entanglement of the protein backbone.

We found that the majority of predicted protein structures possess the same topology as their homologs with known structures (we have taken into account all organisms and focused on proteins with pLDDT higher than 70). The slipknots were predicted in Pfam families, in which they were found based on experimentally determined structures, namely Sodium:dicarboxylate symporter family (PF00375) (without a clan) and multiple families from clan CL0062 (APC superfamily): Sodium:neurotransmitter symporter family (PF00209), Sodium:solute symporter family (PF00474), Permease family (PF00860), Sulfate permease family (PF00916), HCO3-transporter family (PF00955), Natural resistance-associated macrophage protein (PF01566), and Amino acid permease (PF13520) (Table 1).

Herein, based on the comparative studies of knot matrices, we pose a hypothesis that some of the families, without previous data on their entanglement, could possess various slipknots. This includes a family of K^+ potassium transporters (PF02705) from clan CL0062 (APC

superfamily) with a possible $S3_14_13_1$ fingerprint. If the structures predicted by the AlphaFold method are correct, then all families from the CL0062 clan possess slipknot topology.

Moreover, we identified $S3_1$ fingerprint in PF02690 family and $S4_1$ in PF01384, both Na^+:phosphate symporters with no Pfam clan. To the best of our knowledge, until now, the $S4_1$ fingerprint was not observed in a protein. It could be either a result of incorrect structure prediction by AlphaFold (even though pLDDT is above 70), with $S3_14_13_1$ being a plausible alternative, or perhaps the first encounter of a new slipknot fingerprint.

The only known membrane proteins with knot topology include sodium/calcium exchangers belonging to PF01699 family (3). Using AlphaFold predicted structures, we found two topologies within this family: knotted ($K3_1$) and unknotted. The knot is created by two domains from PF01699 and is located in the center of the protein (Fig. 4). This architecture is typical for this substantial protein family with over 80,000 members. However, a fraction of this family (about 10,000 proteins) possesses a single PF01699 domain. We found such proteins to be unknotted (example protein: C0PNZ9), with a protein chain too short to tie. A more detailed study of this family is necessary since it is only the second time that a family encompassing two topologies has been discovered (the first being ATCase/OTCase family: PF00185) (32).

Overall, we can analyze the topology conservation in two well-represented entangled OPM superfamilies (DAACS and sodium/calcium exchanger have only one family; Table 1). First, the APC superfamily consists of 14 families, and most of them form the same slipknot motif ($S3_14_13_1$ in 11 cases). The remaining three form an $S3_13_1$ slipknot motif. A different situation is observed for the second superfamily: monovalent cation-proton antiporter. Out of seven families, only two form nontrivial topology. Both of them have $S3_1$ slipknot motif and represent strictly bacterial proteins. The remaining families are unknotted and possess proteins from all domains of life. Therefore, this superfamily can serve as an interesting example providing distantly related proteins with different topologies (33).

Knotted NCX transporters

Since sodium/calcium exchangers (NCX; PF01699) are the only membrane transporters that are knotted, we analyzed them to see whether there is a connection between the presence of the knot and the function of a TM protein. Therefore, first, we analyzed the location of the knot in different structures of the same protein. Based on nine structures from PDB, we established that the position of the knot core is consistent: it starts between TM1 and TM2 and ends in the middle of TM8 (Figs. 4 and S1). Note that the borders of the knot are not fixed points but can slightly fluctuate due to differences between crystal structures. Then, we verified the position of the knot in the structures from different organisms (Fig. S2).

The beginning of the knot core lies on the border between the gating domain and the core domain (Fig. 4). Our structural analysis showed that during the transition from

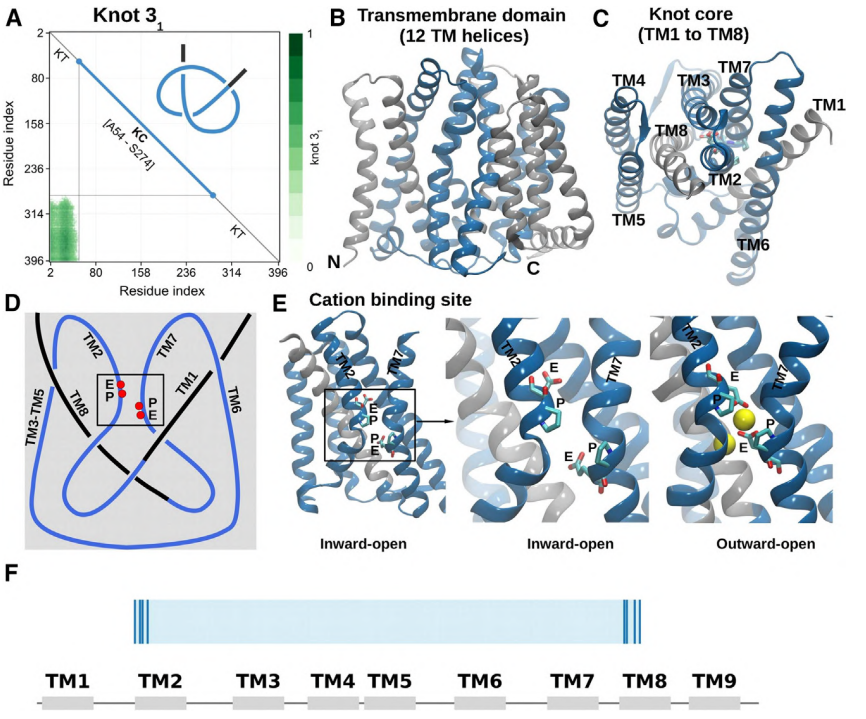


FIGURE 4 Knotted protein backbone in sodium/calcium exchangers (NCX; PF01699). (*A*) The matrix presentation of a protein knotting based on PDB: 4KPP (from *Archaeoglobus fulgidus*), a 3_1 knot (KC, knot core (*blue*); KT, knot tail (*black*)). The color scale shows the dominant knot type formed by a given subchain and the frequency (shown via the color opacity, the green and orange colors are used for knots 3_1 and 4_1, respectively) of its formation. (*B*) Side view of the knotted protein from *Archaeoglobus fulgidus* (PDB: 4KPP). The protein knot core is colored blue and knot tails in gray. (*C*) Top view of the same protein. Only knotted core (TM1–TM8) is shown. (*D*) Schematic presentation of the knotted protein. The binding site is shown in the middle with a black square. Conserved glutamic acid and proline residues are shown as red circles. (*E*) Cation binding site (inward-open structure (PDB: 4KPP from *Archaeoglobus fulgidus*) without the cation bound; outward-open structure (PDB: 5HWY from *Methanococcus jannaschii*) with cation bound). Conserved glutamic acid and proline residues are shown in a stick representation. (*F*) Position of the knot core in different structures of the same protein (UniProtKB: Q57556; based on nine PDB structures). To see this figure in color, go online.

outward-open to inward-open states, the gating domain slides along the core domain, and thus, the turn formed by TM1 and TM2 changes its position relative to the TM7–TM8 of the core domain. In the inward-open structure, TM1–TM2 turn moves away from TM7–TM8, so the inward gate becomes open. In the outward-open structure, TM1–TM2 turn moves toward TM7–TM8 and closes the inward gate. However, more structures and, in particular, structures in inward-open and outward-open states from the same organism are required to analyze whether the flexibility of the knot core position is directly correlated to the conformational motions.

Similarly to the slipknotted TM proteins, knotted sodium/calcium exchangers consist of two homologous repeats (TM1–TM5 and TM6–TM10) that are positioned in the opposite orientation to each other. Each homologous repeat contains a highly conserved region called α-repeat (α-1 (TM2–TM3) and α-2 (TM7–TM8)) important for cation binding and transport (Figs. S2, S3, S4, S5, and S6). Each α-repeat has conserved Glu and Pro residues, and together, they form the cation binding site (Fig. 4 *E* and *F*).

Taking together structural analysis and knotting information, we show that the knot core is conserved in NCX proteins and located within the functional domain (Figs. 4 and S1–S6). Moreover, the conserved α-repeats that form the binding site are located within the knot core in all structures (Figs. 4 and S1–S6). Therefore, knotted topology could be important for the function of these transporters.

Two different slipknot topologies are found in the APC superfamily of transporters

APC superfamily is the largest group of TM proteins with slipknot topology. It contains LeuTs and nucleobase/cation symporters 2 proteins (NCS2-like). Ten families with LeuT fold possess $3_1 4_1 3_1$ slipknot motif (Table 1). The remaining three NCS2-like families have slipknot $3_1 3_1$ (Fig. 2; Table 1).

The most complex slipknot motif $3_1 4_1 3_1$, transporters with LeuT fold

Transporters with LeuT fold possess the most complex slipknot motif $3_1 4_1 3_1$ among TM proteins known to date (8). The matrix representations show that the entire protein backbone forms a slipknot, but continuous subchains form 4_1 knot and two 3_1 knots (Figs. 5, S1, and S10–12). The essential geometric features of these proteins can be characterized as follows.

The most embedded trefoil slipknot 3_1 ($S3_1$) contains two slipknot loops and can be found after removing amino acids from both termini. This trefoil slipknot is formed by the N-terminal half of the structure ($\sim$ 250 aa) as shown in Fig. 5 *A*. The N-terminal slipknot tail (green) and slipknot loop (orange) are located at the TM1 helix, which is further subdivided into TM1a and TM1b by the short unstructured region. This is a hinge region important for the conformational changes of the protein. Interestingly, this is also the border between the slipknot tail and the slipknot loop (Fig. 5 *B*). The transition from the outward- to inward-open state requires movements of several structural elements (34). TM1 helix plays a crucial role in this transition. In the outward-open state, a substrate is bound at the binding site, and TM1a helix (slipknot tail) keeps the inward gate closed: two sodium ions bind to TM1a. Transport of the substrate inside the cell requires hinge-like movement of the TM1b (slipknot loop $S3_1$). The ion-binding site is disturbed, ions are released, and slipknot tail TM1a moves up, opening the inward gate. Interestingly, when the inward gate is closed, residue G24 is located inside the knotted core. Upon motion of the slipknot loop (TM1b), G24 moves outside the knotted core, and the entire slipknot tail moves up (Figs. 5 *B*, *C*, *D,* and *H* and S7). A comparison of 3_1 slipknots in outward- and inward-open states is given in Figs. S7 and S8. Additionally, two conformational states of the same protein are aligned to show precisely the shift of the slipknot elements upon the transition (Fig. S9).

The second embedded knot found in LeuT proteins is a 4_1 slipknot (it can be seen after removing around 350 amino acids from the C-terminus). Fig. 5 *D*, *E*, and *F* shows the exemplary protein structure with its location and geometry. Knotted core in $S4_1$ slipknot is $\sim$ 50 aa longer than in $S3_1$ and slipknot loop is $\sim$ 50 aa. The rest of the structure is a long slipknot tail (green). The $S4_1$ loop is formed by the extracellular loop 4 (EL4). This loop "controls" the opening and closing of the outward gate (Fig. 5 *F* and *H*). In an outward-open state, EL4 (slipknot loop, orange) moves away from the gate, opening it. When the structure transfers into inward-open conformation, the EL4 closes the outward gate (Fig. 5 *G* and *H*). Slipknot 4_1 in outward- and inward-open states is compared in Figs. S7 and S9. Finally, when one includes a longer protein subchain and removes just a few amino acids from the N-terminus, another 3_1 slipknot can be found (Figs. 5 *G*, S7, and S8).

Do uracil transporters possess $3_1 3_1$ or hidden $3_1 4_1 3_1$ topology?

NCS2-like members of the APC superfamily (uracil transporters) possess another type of complex slipknot. Herein, several slipknots are found on the full polypeptide chain when subchains are analyzed. They form a unique knot fingerprint, called $S3_1 3_1$, since two main patches indicating (two) 3_1 slipknots can be seen in the knot fingerprint; see Figs. 6 *A* and S13.

Structures of this protein family consist of two inverted repeats (TM1–TM7 and TM8–TM14). Our topological analysis revealed that more than half of the protein's polypeptide chain (TM1–TM8) is entangled. Slipknot starts at TM1 with a slipknot tail, followed by two slipknot loops

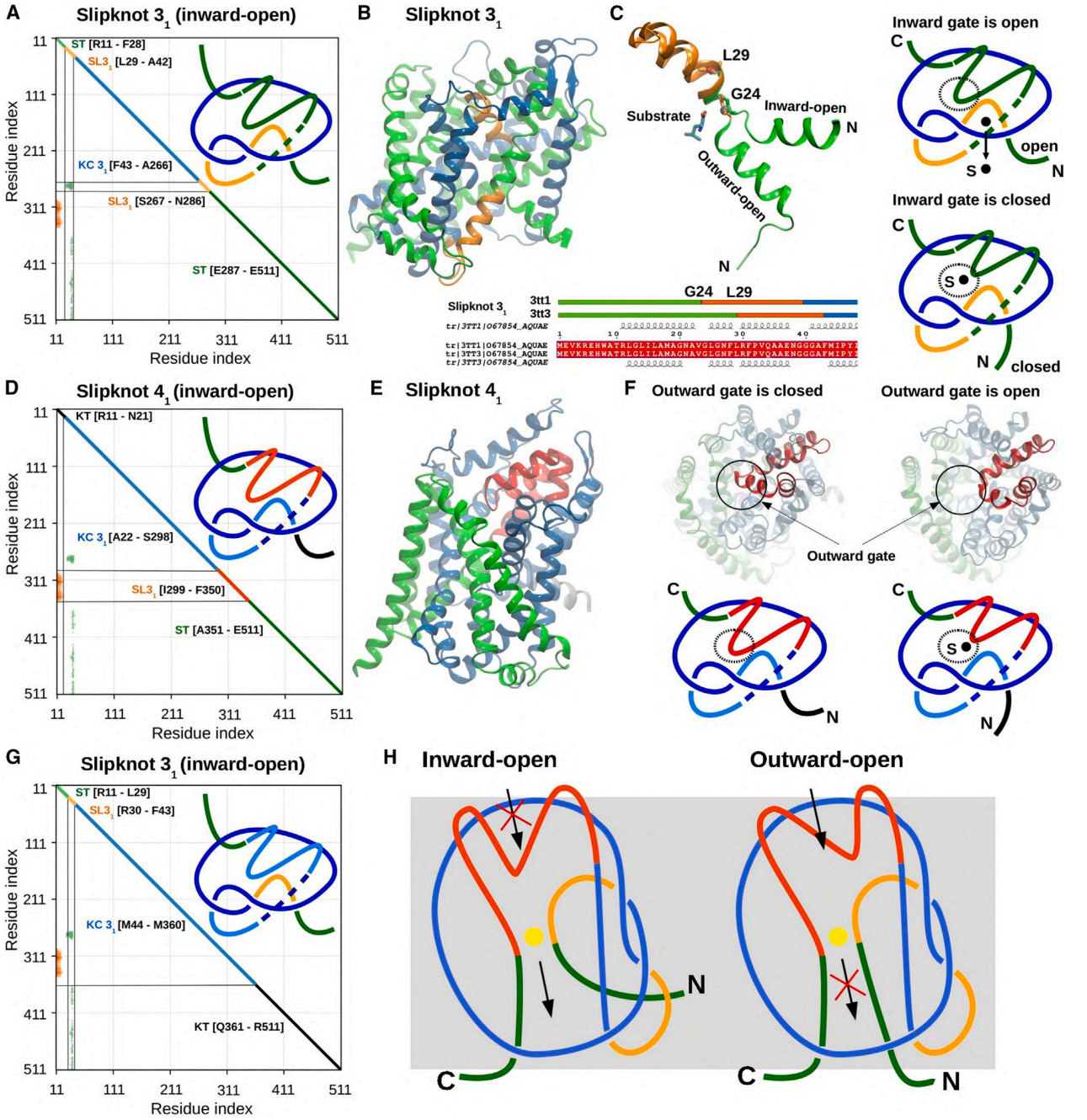


FIGURE 5 Conformation dynamics in LeuT superfamily with $S3_14_13_1$ slipknot motif based on outward (PDB: 3TT1) and inward-open state (PDB: 3TT3). (*A*) Matrix and (*B*) structure with denoted 3_1 slipknot territory determined based on the inward-open state (PDB: 3TT3). ST, slipknot tail (*green*, TM1a); SL, slipknot loop (*orange*, TM1b); KC, knot core (*blue*); KT, knot tail (*gray/black*) with their amino acid ranges written in square brackets. The schematic representation of the slipknot is shown above diagonal. (*C*) superimposed TM1ab of inward-open and outward-open structures. G24 is the residue of the hinge region and the last residue of the 3_1 slipknot loop in the outward-open state. L29 is the last residue of the 3_1 slipknot loop in an inward-open state. In outward-open conformation, the slipknot tail closes the internal gate and blocks the substrate. In an inward-open state, it moves up and allows the substrate to leave. Short alignment compares residues of slipknot 3_1 in inward- to outward-open states. On the right, there is a schematic representation of the slipknot 3_1 (with marked positions of the key residues, G24 and L29, and the substrate) in inward (*upper*) and outward (*lower*) conformations. (*D*) Matrix and (*E*) structure of inward-open structure (PDB: 3TT3) with denoted position of 4_1 slipknot. (*F*) (Upper panel) View of the structures from the extracellular side: in an inward-open state, 4_1 slipknot loop closes the outside gate; in an outward-open state, it moves aside and opens the gate. (Lower panel) Schematic representation of the slipknot 4_1 and position of its loop in inward- (*left*) and outward-open (*right*) states. The dashed circle shows the location of the outward gate. (*G*) Location of the second slipknot 3_1 on the matrix. Note that it almost coincides with the first slipknot 3_1; however, its knotted core is longer. Therefore, the second slipknot 3_1 is found on a different protein subchain. (*H*) Schematic presentation of the position of slipknot loop 4_1 and slipknot tail 3_1 in inward-open and outward-open structures. Both slipknots 4_1 and 3_1 are shown simultaneously. To see this figure in color, go online.

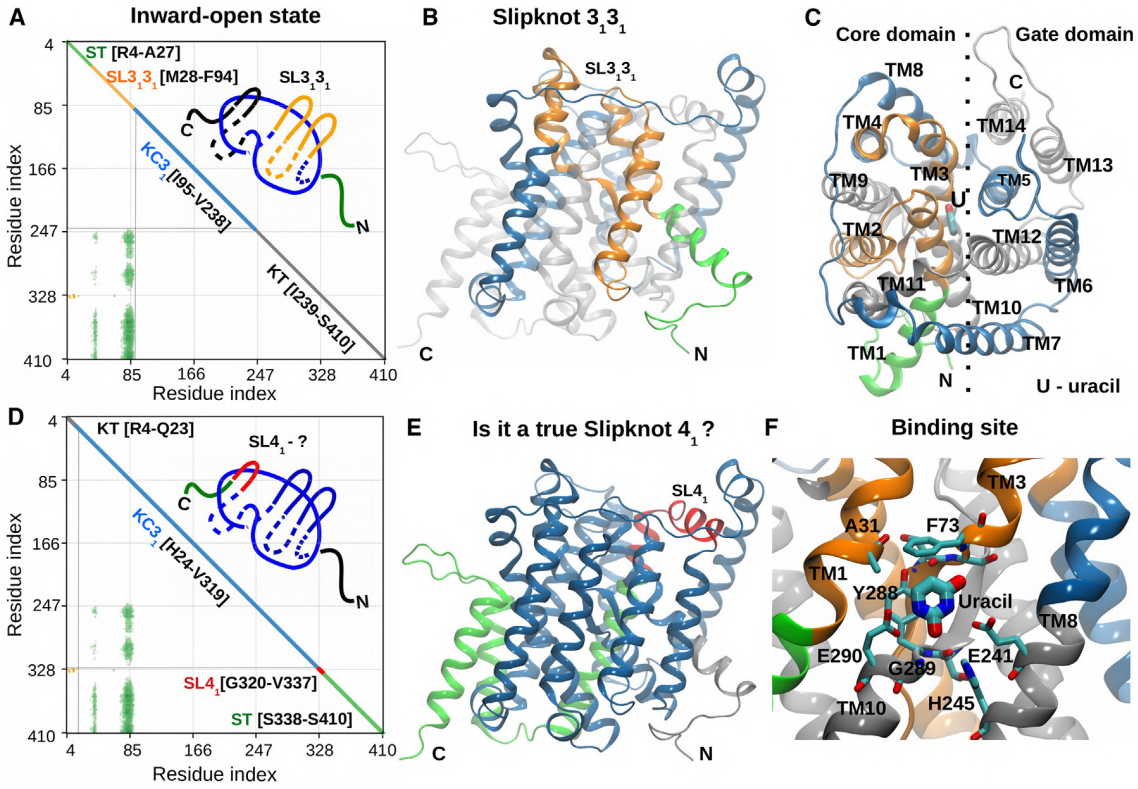


FIGURE 6 Potentially different slipknot motifs (3_13_1 and $3_14_13_1$) are detected in uracil transporters (inward-open state; PDB: 3QE7). (*A*) Matrix presentation of protein topology with S3_13_1 motif marked. (*B*) Structure colored according to S3_13_1 motif. (*C*) Top view of the inward-open structure with colored S3_13_1 motif. The dotted line shows the border between core and gate domains. Uracil molecule located at the binding site is shown by stick presentation. (*D*) Matrix presentation of protein with marked 4_1 slipknot territories ($3_14_13_1$ motif) located at the threshold border. ST, slipknot tail (*green*); SL, slipknot loop (*orange*); KC, knot core (*blue*); KT, knot tail (*gray/black*) with their amino acid ranges written in square brackets. The schematic representation of each slipknot is given in the upper part of the matrix. (*E*) Structure colored according to 4_1 slipknot. (*F*) Magnified view of the binding site. Uracil molecule is shown with thick sticks. Residues within 4 Å from the uracil are shown as sticks. The structure is colored according to the slipknot S3_13_1 motif. To see this figure in color, go online.

(TM1–TM4) (Fig. 6). After the slipknot loops, the knot core starts at TM4 and continues to the middle of TM8. The first 3_1 slipknot loop is a turn formed by the C-terminus of TM1 and the N-terminus of TM2; then it hides in the middle of the structure and then comes up as the second 3_1 slipknot loop (a turn formed by the C-terminus of TM3 and the N-terminus of TM4). Overall, there are two 3_1 slipknot loops (fingerprint 3_13_1; Fig. 6 *B* and *C*).

Similarly to the other transporters, the structure of uracil transporters is made of two functional domains: the core domain and the gate domain (Fig. 6 *B* and *C*). We mapped the slipknot elements on the structure of the transporter to analyze whether slipknot elements are involved in protein function (Fig. 6). Our analysis revealed that slipknot tail TM1, two slipknot loops, and two helices of knot core (TM4 and TM8) are part of the core domain. Although, the three other helices that form the knot core (TM5, TM6, and TM7) are part of the gate domain. The rest of the structure (part of TM8–TM14) is a long knot tail. Although the knot core spans only half of the structure, it binds the core and the gate domain together. Both slipknot loops are located in the middle of the structure, at the interface between the core domain and the gate domain (Fig. 6 *C*). This is the location of the binding site, where the uracil molecule is bound (Fig. 6 *C* and *F*). Together with the residues of the consensus NAT motif (in TM10), residues from the slipknot loop and the knot core take part in uracil coordination (Fig. 6 *F*). In particular, Glu 290, Tyr 288, Gly 289 from TM10 and Glu 241 and His 245 from TM8, Phe 73 from TM3, and Ala 31 from TM1 participate in uracil binding. Notably, some of the key residues involved in uracil binding are located at the borders of the slipknot loop and the knot core (Fig. 6 *F*). In particular, Glu241 is conserved throughout the whole family and required for substrate coordination and co-transporting proton or sodium cation.

However, we were not able to determine whether the slipknot loops and the slipknot tail are involved in the protein transition from inward-open to outward-open state since there is no structure in the outward-open state in the Nucleobase/Cation Symporter family yet. Nevertheless, based on a comparison of the inward-open and occluded structures, we observed the shift of borders of slipknot elements

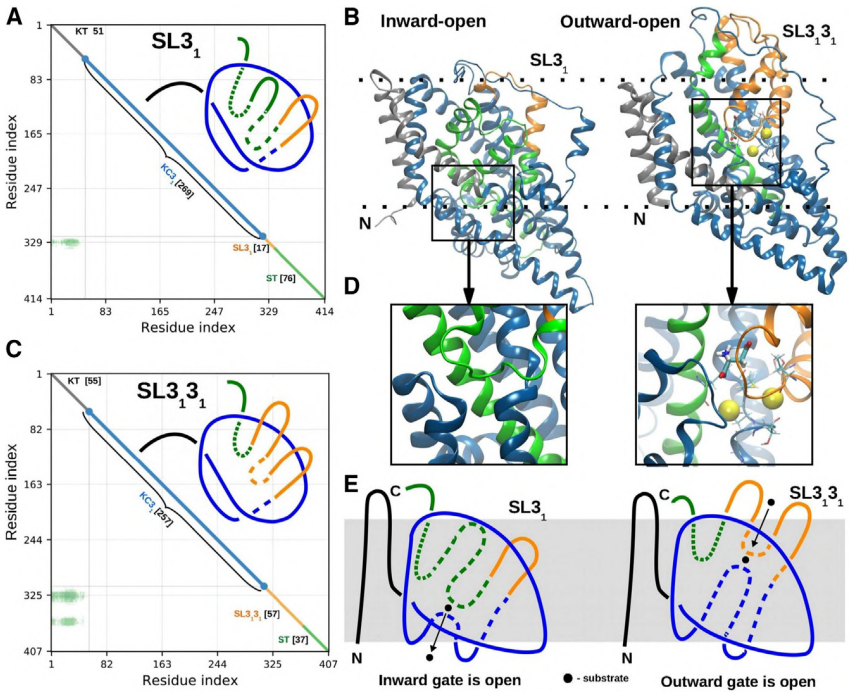


FIGURE 7 Example of different slipknot motifs in SDF transporters, based on inward-open (PDB: 4P19) and outward-open (PDB: 4IZM) state. Matrix presentation of protein in inward-open state (*A*) and in outward-open (*C*) shows respectively one and two slipknot territories, motif $S3_1$ (*A*) and $S3_13_1$ motif (*C*). The schematic representation of each slipknot in the upper part. ST, slipknot tail (*green*); SL, slipknot loop (*orange*); KC, knot core (*blue*); KT, knot tail (*gray/black*). The number of residues is written in square brackets. (*B*) illustrates the difference in protein structures for inward-open state (one slipknot) and outward-open state (two slipknots). Structures are colored according to the slipknot topology. (*D*) Magnified view of the binding site in inward and outward states. (*E*) Schematic drawing of the 3_1 and 3_13_1 slipknots. To see this figure in color, go online.

(Figs. S14 and S16). Therefore, we could conclude that slipknot elements are flexible in this family similar to the other entangled transporters.

Apart from 3_13_1 fingerprint, in some of the structures, 4_1 slipknots are detected with about 40%–55% probability. In this case, the $S3_14_13_1$ motif is detected (Fig. 6 *D* and *E*). However, due to the fact that we use 50% as the probability threshold for the knot's presence in the structure, more structures will be needed to fully characterize this group. In particular, they would provide an answer to the question of whether 4_1 slipknot exists only in some protein folds, or whether it appears only in a particular structural conformation. Another option is that an embedded 4_1 slipknot is a trace of evolution indicating the common origin of LeuT and NCS2-like transporters (Fig. S15).

Uracil transporters are assigned to the same Pfam clan as LeuT, despite significant differences in the fold type. This could indicate that LeuT and NCS2-like proteins are distantly related. However, based on the analysis of the Evolutionary Classification of Protein Domains (ECOD) database (35,36), there is not enough similarity between these families to assign them as possible homologs. Therefore, it is not yet clear whether LeuT and NCS2-like transporters are related or not. Despite the difference in fold types between these transporters, some common features can be observed based on the slipknot fingerprint (Fig. S15). First, in both cases, 3_1 slipknot starts from the tail (shown in green, Figs. 5 *A* and *B* and 6 *A* and *B*) followed by the slipknot loop and then knotted core. However, the slipknot loop is much longer (86 aa) in uracil transporters and forms two slipknot loops instead of one in LeuT proteins. Second, in both structures, the tail of 3_1 slipknot is part of TM1 and the N-terminal region of the protein. The TM1 helix is divided into two topologically different parts: TM1a, slipknot tail, and TM1b, part of the 3_1 slipknot loop (Fig. S15). Third, the knotted core of the 3_1 slipknot occupies approximately half of the structure in uracil and leucine transporters. Therefore, despite the lack of sequence similarity, a fingerprint provides additional information showing the similarity between leucine and uracil transporters.

SDF transporters switch between two fingerprints: 3_1 or 3_13_1 slipknot

Sodium:dicarboxylate symporter family (SDF; classified as proton:glutamate symport protein in DAACS superfamily in the OPM database) possesses slipknot loops located on the C-terminus as opposite to the APC transporters and is embedded in the knot core formed by almost the entire remaining part of the protein chain (Fig. 7).

An interesting feature of SDF proteins is the ability to switch between the two types of fingerprints: single 3_1 or double 3_13_1 slipknot (Figs. 7, S17, and S18). Such properties were not observed before for any type of slipknot or knot. This topological feature is a consequence of the transition of the transporter between outward-open and inward-open states. The inward-facing state corresponds to 3_1 slipknot, whereas outward-facing conformation has 3_13_1 slipknot type (Fig. 7). Structural analysis of the protein in inward and outward states shows that two different fingerprints occur due to the change in the position of the 3L4 loop and two helical hairpins (HP2) that are involved in the conformational changes (Figs. 7 *B* and *E* and S17). When

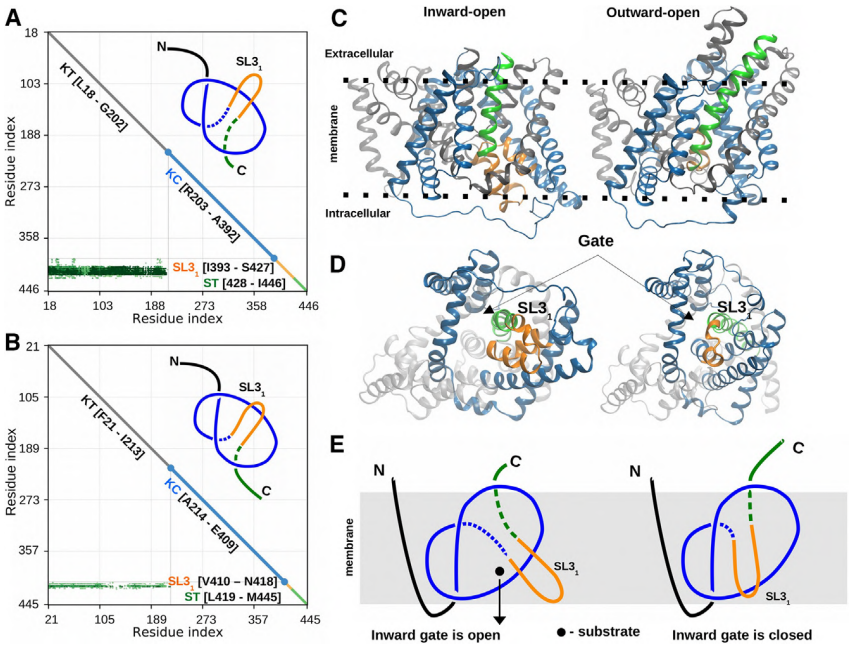


FIGURE 8 Inward- and outward-open states in 2HCT can be detected by different range of 3_1 slipknot territory. (*A* and *B*) Matrix presentation of protein topology (PDB: 5A1S) shows deep and shallow 3_1 slipknots, corresponding to inward- and outward-open states, respectively (notice the difference in the orange loop and its representation on the diagonal). ST, slipknot tail (*green*); SL, slipknot loop (*orange*); KC, knot core (*blue*); KT, knot tail (*gray/black*). The schematic representation of each slipknot is given in the upper part of the matrix. (*C*) Protein structures (colored according to the slipknot topology) showing that in inward-open state, slipknot loop is longer and exits to the extracellular side together with the substrate; in outward-open state, slipknot loop is shorter and moves up toward the extracellular side. (*D*) View from the intracellular side. (*E*) Schematic drawing of slipknot geometry in inward- and outward-open states. To see this figure in color, go online.

a single 3_1 slipknot is observed, the knot core (269 residues long, blue) is followed by a 17-residue-long slipknot loop (orange, Fig. 7 *A*). The rest of the structure (residues 343–400) forms a slipknot tail (green). In outward-open state, the slipknot loop becomes much longer (57 residues) and forms two slipknot loops 3_13_1 (Fig. 7 *C*), whereas the knot core is 12 residues shorter (257 residues). Also, the shift of the slipknot elements upon transition is shown in the sequence alignment of inward- and outward-open states in Fig. S18. Interestingly, proteins mutated to be in outward or inward states also provide similar examples. A mutant of a *Pyrococcus horikoshii* glutamate transporter K55C/A364C is a protein in an inward-facing state with a single 3_1 slipknot (37), whereas the L66C/S300C mutations leading to the outward-facing state (38) result in a 3_13_1 slipknot (Fig. S19). These findings show for the first time to our knowledge that despite being conserved, slipknot topology can change between two fingerprints. Therefore, it has emerged that slipknot topology can have a dynamic nature.

Motion of the slipknot loop correlates with conformational states in 2HCT

2-Hydroxycarboxylate transporter (2HCT) family from the Monovalent cation-proton antiporter superfamily is another example of a unique slipknot fingerprint (Fig. 8). In this family, two helical hairpins, HP1 and HP2, play a crucial role in Na^+/substrate transport. The matrix presentation of the protein shows that the hairpin HP2 is a part of the slipknot loop and knotted core, whereas HP1 is located at the knot tail. These two hairpins are located next to each other but in opposite orientations and form substrate and Na^+ binding sites (Figs. S21 and S22) (39,40). Interestingly, the size of the slipknot loop differs due to the protein's change from the outward- to inward-open conformation. The slipknot loop becomes larger (∼35 residues long) in an inward-facing state and reaches the intracellular side. In this case, the slipknot loop is formed by hairpin HP2 and partially by terminal TM13 helix (Figs. 7, S21, and 22). In the outward-facing state, the slipknot loop moves up and becomes smaller (∼9 residues) (Figs. 7, S21, and 22). In this case, the slipknot loop is formed by residues V410–N418 of hairpin HP2. Changing the size of the slipknot loop shifts the boundaries of the knot core and slipknot tail depending on the protein conformation (Figs. 7, S20, and S23). Also, both hairpin loops contain multiple conserved glycines (GGXG motif, Fig, S24) (39). Although hairpin glycines are highly conserved, only two of them (G187 from hairpin HP1 and G404 from hairpin HP2) are described to be involved in substrate binding (39,40). However, glycine residues are present not only in hairpins, but an unusually high number of glycines is also present throughout the entire structure (Fig. S24). Notably, Gly residues are also highly conserved in other knotted and slipknotted structures. It was shown that the presence of glycine allows for sharp turns in the polypeptide chain that facilitate the threading of a slipknot loop through the knotted loop (8).

Interestingly, the Monovalent cation-proton antiporter superfamily joins together unknotted and slipknotted homologs (Table 1). This indicates that slipknotted and unknotted families share a common ancestor (33). We identified a new protein family with slipknot topology within this superfamily that was not described before. This is Na^+-transporting carboxylic acid decarboxylase family (Fig. S13). Although its

structures differ from the 2-hydroxycarboxylate transporter family, they share the same fold type and the same 3_1 slipknot fingerprint (Figs. 8 and S13).

DISCUSSION

In this work, we reviewed and analyzed in detail all known TM proteins with nontrivial topology. We showed that proteins with nontrivial topology are found only within the α-helical polytopic class, which is represented by 17 families of TM proteins with slipknot and one family with knot topology. All of them are secondary active transporters, which comprise a large group of amino acid transporters, cation antiporters, and symporters.

Although proteins with slipknots and knots constitute only 1% of the total number of proteins deposited in the PDB (19), herein we showed that they form about one-fourth of all known secondary active transporters in mammals and humans. In particular, this implies that the entanglements were successfully adopted in secondary active transporters, which require complex conformational changes during their transport circle. We found that in all analyzed transporters, the location and borders of the slipknot loop, knot core, and tails coincide with the residues and regions required for protein function. For example, residues located at the slipknot and knot crossings are crucial for the substrate binding or hinge points that trigger transitions from inward to outward conformations. Interestingly, structural changes that occur during transitions from inward to outward states are captured by a slipknot fingerprint motif. In particular, this is a feature of families with a slipknot fingerprint: 2HCT, SDF, and uracil transporters. In all cases, the key elements involved in protein motions require sliding of the knotted core and slipknot loop. These findings indicate that slipknot topology was advantageous for complex and coordinated motions of structural elements.

Nontrivial topological motifs are also known to significantly contribute to the structure of the binding sites of other protein types. For example, the substrate binding site of some tRNA methyltransferases is located at the border of 3_1 knot (17). Importantly, there are several analogical pairs of such methyltransferases, with a knotted bacterial protein, and an unknotted human one, e.g., TrmD and Trm5, respectively. In TrmD and other knotted tRNA methyltransferases, the most important part of the S-adenosylmethionine binding site is the adenine-binding loop, located at the knot tail. This structural feature is missing in the unknotted analogs, thereby imposing a diverse substrate binding profile. Those structural differences may be utilized to rationally design selective antimicrobial agents (41). Another example is the human estrone sulfatase, in which the active site is colocalized with the borders of the specific elements of a deep 3_1 slipknot (42).

Also, the topology type correlates with the unique transport mechanisms. LeuTs with $S3_14_13_1$ motif move molecules by rocking bundle mechanism. NCS2-like transporters are described as having mixed rocking bundle and elevator mechanism. Interestingly, these transporters have 3_13_1 slipknot and traces of 4_1 slipknot. Transporters such as SDF with $S3_13_1$ and 2HCT with $S3_1$ motif transport molecules by elevator mechanism. Finally, the only family with a $K3_1$ knot motif has a yet distinct mechanism called two-stroke.

Herein, we show for the first time to our knowledge that TM proteins with slipknot topology can switch between two fingerprint motifs while preserving nontrivial topology. Based on analysis of structures and fingerprints, our study suggests that slipknot and knot topologies are required for maintaining conformational changes during the transport cycle and therefore are directly involved in the function of these transporters. Moreover, knowledge about protein entanglement is helpful in identifying transport mechanisms and key residues involved in the transport cycle (binding site and hinge regions).

We focused on proteins with known structures; however, including structures predicted by the ML approach allowed us to extend the knowledge of the slipknotted and knotted TM proteins. In consequence, we postulate that all Pfam families from APC superfamily (CL0062) are entangled, and their backbones form a slipknot. Moreover, we identified slipknots in two unattributed families: Phosphate transporter family (PF01384) and Na+/Pi-cotransporter (PF02690). Finally, we found unknotted (group of homologs) proteins among PF01699, which is the second observed occurrence of proteins from the same Pfam clan possessing different topologies. This is an exciting result since it is believed that knotted topology is conserved within a single family (8). However, due to the availability of high-quality structure prediction from AlphaFold, it is possible we will see more and more such exceptions. For example, we recently analyzed models from a family that contains proteins with two different knot types: 6_3 and 3_1 (43).

This work provides a comprehensive introduction and characterization of slipknotted and knotted membrane proteins that can facilitate their further research as there are still many interesting aspects that need resolving. One of them is a degradation method of proteins with such complicated topology. Interestingly, an article published in 2020 analyzed the degradation of slipknotted membrane proteins without considering their nontrivial topology, possibly due to lack of this information (44). Four (out of 16) human transporters they studied are slipknotted (SLC1A5, SLC38A1, SLC38A2, and SLC38A9). All of them were susceptible to proteasome degradation that was ligand induced (dTAG system). In this method, the proteins of interest are fused to FKBP12 protein (45). This shows two additional and important points that were overlooked in the original paper: 1) slipknotted proteins can be degraded even with the presence of an additional domain (over 100 amino acid long); 2) the novel dTAG method can also be used to study the degradation of nontrivial proteins. Given the lack of topology data, such important conclusions can easily be missed, emphasizing the need and significance of this work.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2023.10.031.

AUTHOR CONTRIBUTIONS

V.Z. and J.I.S. designed the research. M.S. collected the data and prepared algorithms. M.S., A.S. and B.A.G. ran the calculations and gathered results. V.Z, A.S, A.P.P., and M.S. analyzed results and curated the data. V.Z., J.I.S., A.P.P., and A.S. wrote the paper. V.Z. and A.P.P. prepared figures. J.I.S., V.Z., A.P.P., and M.S. reviewed and edited the paper. J.I.S. supervised the research and administered the project.

ACKNOWLEDGMENTS

This work was supported by the National Science Centre #UMO-2018/31/B/NZ1/04016, #2021/43/I/NZ1/03341, EMBO Installation Grant #2057, and MNiSW grant #0003/ID3/2016/64.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Fagerberg, L., K. Jonasson, L. Berglund…, 2010. Prediction of the human membrane proteome. *Proteomics.* 10:1141–1149.
2. Joseph, J. B., and M. Li. 2014. Deorphanizing the human transmembrane genome: A landscape of uncharacterized membrane proteins. *Acta Pharmacol. Sin.* 35:11–23.
3. Jarmolinska, A. I., A. P. Perlinska, …, J. I. Sulkowska. 2019. Proteins' knotty problems. *J. Mol. Biol.* 431:244–257.
4. Dabrowski-Tumanski, P., P. Rubach, …, J. I. Sulkowska. 2018. KnotProt 2.0: a database of proteins with knots and other entangled structures. *Nucleic Acids Res.* 47:D367–D375.
5. Jumper, J., R. Evans, …, D. Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature.* 596:583–589.
6. Niemyska, W., P. Rubach, …, J. I. Sulkowska. 2022. AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods. *Nucleic Acids Res.* 50:W44–W50.
7. King, N. P., E. O. Yeates, and T. O. Yeates. 2007. Identification of rare slipknots in proteins and their implications for stability and folding. *J. Mol. Biol.* 373:153–166.
8. Sułkowska, J. I., E. J. Rawdon, K. C. Millett, J. N. Onuchic, and A. Stasiak. 2012. Conservation of complex knotting and slipknotting patterns in proteins. *Proc. Natl. Acad. Sci. USA.* 109:E1715–E1723.
9. Ohta, S., M. T. Alam, …, A. Ikai. 2004. Origin of mechanical strength of bovine carbonic anhydrase studied by molecular dynamics simulation. *Biophys. J.* 87:4007–4020.
10. Sułkowska, J. I., P. Sułkowski, P. Szymczak, and M. Cieplak. 2008. Tightening of knots in proteins. *Phys. Rev. Lett.* 100, 058106.
11. Calvaresi, M., A. S. Duwez, …, L. Zhang. 2023. Mechanical tightening of a synthetic molecular knot. *Chem.* 9:65–75.
12. Sułkowska, J. I., P. Sułkowski, P. Szymczak, and M. Cieplak. 2010. Untying knots in proteins. *J. Am. Chem. Soc.* 132:13954–13956.
13. Wang, H., and H. Li. 2020. Mechanically tightening, untying and retying a protein trefoil knot by single-molecule force spectroscopy. *Chem. Sci.* 11:12512–12521.
14. Dabrowski-Tumanski, P., and J. I. Sulkowska. 2017. Topological knots and links in proteins. *Proc. Natl. Acad. Sci. USA.* 114:3415–3420.
15. Niemyska, W., P. Dabrowski-Tumanski, …, J. I. Sulkowska. 2016. Complex lasso: new entangled motifs in proteins. *Sci. Rep.* 6:36895–36910.
16. Dabrowski-Tumanski, Pawel, Dimos Goundaroulis, Andrzej Stasiak, and Joanna I. Sulkowska. 2019. θ-curves in proteins. Preprint at arXiv. https://doi.org/10.48550/arXiv:1908.05919.
17. Christian, T., R. Sakaguchi, …, Y. M. Hou. 2016. Methyl transfer by substrate signaling from a knotted protein fold. *Nat. Struct. Mol. Biol.* 23:941–948.
18. Dabrowski-Tumanski, P., A. Stasiak, and J. I. Sulkowska. 2016. In search of functional advantages of knots in proteins. *PLoS One.* 11, e0165986.
19. Sulkowska, J. I. 2020. On folding of entangled proteins: knots, lassos, links and θ-curves. *Curr. Opin. Struct. Biol.* 60:131–141.
20. Jackson, S. E., A. Suma, and C. Micheletti. 2017. How to fold intricately: using theory and experiments to unravel the properties of knotted proteins. *Curr. Opin. Struct. Biol.* 42:6–14.
21. Colas, C., and E. Laine. 2021. Targeting Solute Carrier Transporters through Functional Mapping. *Trends Pharmacol. Sci.* 42:3–6.
22. Malinauskaite, L., S. Said, …, P. Nissen. 2016. A conserved leucine occupies the empty substrate site of LeuT in the Na+-free return state. *Nat. Commun.* 7, 11673.
23. Rahman, M., F. Ismat, …, S. G. Patching. 2017. Characterisation of the DAACS family Escherichia coli glutamate/aspartate-proton symporter GltP using computational, chemical, biochemical and biophysical methods. *J. Membr. Biol.* 250:145–162.
24. Liao, J., H. Li, …, Y. Jiang. 2012. Structural insight into the ion-exchange mechanism of the sodium/calcium exchanger. *Science.* 335:686–690.
25. Vastermark, A., S. Wollwage, …, M. H. Saier, Jr. 2014. Expansion of the APC superfamily of secondary carriers. *Proteins.* 82:2797–2811.
26. Lomize, M. A., I. D. Pogozheva, …, A. L. Lomize. 2012. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* 40:D370–D376.
27. Bölinger, D., J. I. Sułkowska, H. P. Hsu, L. A. Mirny, M. Kardar, J. N. Onuchic, and P. Virnau. 2010. A Stevedore's protein knot. *PLoS Comput. Biol.* 6, e1000731.
28. Perland, E., and R. Fredriksson. 2017. Classification systems of secondary active transporters. *Trends Pharmacol. Sci.* 38:305–315.
29. Bernsel, A., H. Viklund, A. Hennerdal, and A. Elofsson. 2009. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* 37:W465–W468.
30. Varadi, M., S. Anyango, …, S. Velankar. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50:D439–D444.
31. Lin, Z., H. Akin, …, A. Rives. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science.* 379:1123–1130.
32. Škrbic, T., C. Micheletti, and P. Faccioli. 2012. The role of non-native interactions in the folding of knotted proteins. *PLoS Comput. Biol.* 8, e1002504.
33. Zayats, V., A. P. Perlinska, …, J. I. Sulkowska. 2021. Slipknotted and unknotted monovalent cation-proton antiporters evolved from a common ancestor. *PLoS Comput. Biol.* 17, e1009502.
34. Penmatsa, A., and E. Gouaux. 2014. How LeuT shapes our understanding of the mechanisms of sodium-coupled neurotransmitter transporters. *J. Physiol.* 592:863–869.
35. Cheng, H., R. D. Schaeffer, …, N. V. Grishin. 2014. ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.* 10, e1003926.
36. Cheng, H., Y. Liao, …, N. V. Grishin. 2015. Manual classification strategies in the ECOD database. *Proteins.* 83:1238–1251.
37. Verdon, G., S. Oh, R. N. Serio, and O. Boudker. 2014. Coupled ion binding and structural transitions along the transport cycle of glutamate

transporters. *eLife.* 3. Benoit Roux:e02283. https://doi.org/10.7554/eLife.02283.

38. Reyes, N., S. C. Oh, and O. Boudker. 2013. Binding thermodynamics of a glutamate transporter homolog. *Nat. Struct. Mol. Biol.* 20:634–640. https://doi.org/10.1038/nsmb.2548.
39. Kim, J. W., S. Kim, …, M. S. Jin. 2017. Structural insights into the elevator-like mechanism of the sodium/citrate symporter CitS. *Sci. Rep.* 7:2548–2610.
40. Wöhlert, D., M. J. Grötzinger, …, Ö. Yildiz. 2015. Mechanism of Na+-dependent citrate transport from the structure of an asymmetrical CitS dimer. *Elife.* 4, e09375.
41. Perlinska, A. P., A. Stasiulewicz, …, J. I. Sulkowska. 2020. Restriction of S-adenosylmethionine conformational freedom by knotted protein binding sites. *PLoS Comput. Biol.* 16, e1007904.
42. Hernandez-Guzman, F. G., T. Higashiyama, …, D. Ghosh. 2003. Structure of human estrone sulfatase suggests functional roles of membrane association. *J. Biol. Chem.* 278:22989–22997.
43. Perlinska, A. P., W. H. Niemyska, …, J. I. Sulkowska. 2023. AlphaFold predicts novel human proteins with knots. *Protein Sci.* 32, e4631.
44. Bensimon, A., M. D. Pizzagalli, …, G. Superti-Furga. 2020. Targeted degradation of SLC transporters reveals amenability of multi-pass transmembrane proteins to ligand-induced proteolysis. *Cell Chem. Biol.* 27:728–739.e9.
45. Nabet, B., J. M. Roberts, …, J. E. Bradner. 2018. The dTAG system for immediate and target-specific protein degradation. *Nat. Chem. Biol.* 14:431–441.
46. Freyd, P., D. Yetter, …, A. Ocneanu. 1985. A new polynomial invariant of knots and links. *Bull. Am. Math. Soc.* 12:239–246. https://doi.org/10.1090/S0273-0979-1985-15361-3.
47. Przytycki, J. H., and Pawel Traczyk. 2016. Invariants of links of Conway type. https://arxiv.org/abs/1610.06679.
48. Millett, K. C., E. J. Rawdon, A. Stasiak, and J. I. Sułkowska. 2013. "Identifying knots in proteins. *Biochem. Soc. Trans.* 41:533–537.
49. Christopher J Williams et al. "MolProbity: More and better reference data for improved all-atom structure validation". en. *Protein Sci.* 27:293–315.
50. Dabrowski-Tumanski, P., P. Rubach, …, J. I. Sulkowska. 2021. Topoly: Python package to analyze topology of polymers. *Briefings Bioinf.* 22:bbaa196.
51. El-Gebali, S., J. Mistry, …, R. D. Finn. 2018. The Pfam protein families database in 2019. *Nucleic Acids Res.* 47:D427–D432.
52. Madeira, F., M. Pearce, …, R. Lopez. 2022. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* 50:W276–W279.
53. Robert, X., and P. Gouet. 2014. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* 42:W320–W324.
54. Krieger, E., and G. Vriend. 2014. YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics.* 30:2981–2982.
55. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27-28.

Article 4

AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods

W44–W50 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 24 May 2022*
https://doi.org/10.1093/nar/gkac388

AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods

Wanda Niemyska[1,2], Pawel Rubach[1,3], Bartosz A. Gren[1], Mai Lan Nguyen[1,4], Wojciech Garstka[1,5], Fernando Bruno da Silva[1], Eric J. Rawdon[6] and Joanna I. Sulkowska[1,*]

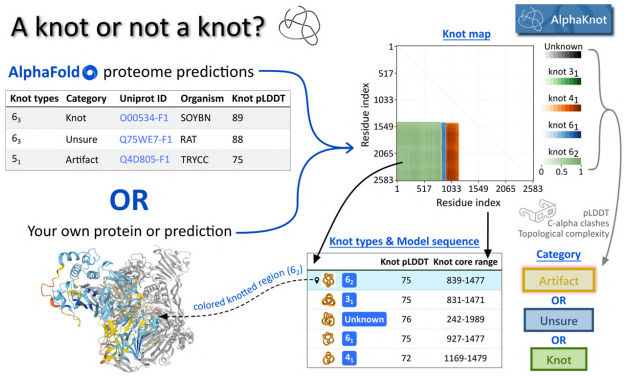
[1]Centre of New Technologies, University of Warsaw, Banacha 2c, Warsaw, Poland, [2]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland, [3]Warsaw School of Economics, Al. Niepodleglosci 162, 02-554, Warsaw, Poland, [4]Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland, [5]Inter-faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland and [6]University of St. Thomas, 2115 Summit Ave, Saint Paul, MN 55105, USA

Received March 29, 2022; Revised April 20, 2022; Editorial Decision April 29, 2022; Accepted May 06, 2022

ABSTRACT

AlphaKnot is a server that measures entanglement in AlphaFold-solved protein models while considering pLDDT confidence values. AlphaKnot has two main functions: (i) providing researchers with a webserver for analyzing knotting in their own AlphaFold predictions and (ii) providing a database of knotting in AlphaFold predictions from the 21 proteomes for which models have been published prior to 2022. The knotting is defined in a probabilistic fashion. The knotting complexity of proteins is presented in the form of a matrix diagram which shows users the knot type for the entire polypeptide chain and for each of its subchains. The dominant knot types as well as the computed locations of the knot cores (i.e. minimal portions of protein backbones that form a given knot type) are shown for each protein structure. Based mainly on the pLDDT confidence values, entanglements are classified as Knots, Unsure, and Artifacts. The database portion of the server can be used, for example, to examine protein geometry and entanglement-function correlations, as a reference set for protein modeling, and for facilitating evolutional studies. The AlphaKnot server can be found at https://alphaknot.cent.uw.edu.pl/.

GRAPHICAL ABSTRACT



INTRODUCTION

The presence of entanglements in proteins is an important phenomenon, inspiring multidisciplinary research involving biology, biophysics, chemistry and mathematics. Proteins with knots and slipknots, intensively studied over last years, provide an important example of this phenomenon. However, the 3D structures of knotted proteins are not easily discovered experimentally.

Currently around 2% (1) of solved protein structures from the PDB are considered to contain non-trivial topologies: knots, slipknots, or links (2). However, the PDB contains only experimentally-solved structures and, therefore, is not a representative set of structures. The set of solved structures is not sufficient to answer the most important questions about the biological role of knots, their evolution, and the identification of sequence motifs responsible for coding non-trivial 3D topologies (3).

However, in the last year the amount of data crucial for structural biology increased significantly as a result of ground-breaking developments based on methods from ar-

*To whom correspondence should be addressed. Tel: +48 22 55 43 675; Fax: +48 22 822 02 11 (Ext 320); Email: j.sulkowska@cent.uw.edu.pl

tificial intelligence: the AlphaFold 2 (AF 2) program by DeepMind (4) predicted structures of >990 000 proteins from 48 complete proteomes, which exceeds the amount of experimental data gathered to date. The quality of some of the data is extraordinary. Furthermore, analysis of AF 2 results for the human genome shows that AF 2 predicts knot types not observed in experimentally-solved structures, as well as new knotted families (5). On the other hand, since there is no sufficient training set, the AlphaFold database includes families whose members (homological proteins from different organisms) possess high pLDDT scores but different topologies, implying that some of the predictions form artificial knots (3) and, thus, are likely not accurate models.

Herein, we present the AlphaKnot server—the first server to identify entanglements in AlphaFold-solved protein models while taking into account the pLDDT confidence values. AlphaKnot acts as a submission-server and as a database for some AlphaFold published predictions. First, the submission-server provides a simple interface that allows researchers to upload AlphaFold predicted models for full topological analysis. Users can upload single chains or multiple models (including multiple models that overlap), in which case the server displays a visual comparison of the knotting in different models. Second, AlphaKnot is a database for the 21 proteomes for which AlphaFold models have been published prior to December 2021 with a browsing feature that allows for filtering by different criteria (such as certain proteomes, knot types, and core lengths).

Thus, in addition to finding entanglements in uploaded or AlphaFold-predicted proteins, AlphaKnot can be used as a tool for improving structure prediction. For example, topological differences between related proteins in combination with pLDDT values could be used to reveal potential areas of improvement (or strengths and weaknesses) for AlphaFold predictions.

MATERIALS AND METHODS

Detection of knots

The algorithm computes knotting in the protein backbone using the coordinates of Cα atoms based on a CIF or PDB file (for details see on-line help). The general idea of knot identification follows earlier works (1).

The full structure (processed either by the server or in the database) is first checked using the HOMFLY-PT polynomial (in the case of probabilistic closing, with a high number 1000 of random closures). If a non-trivial knot is found, then the structure is further processed and the knot core or whole knot map is calculated. By 'finding a knot using probabilistic closure' we mean that the dominant knot type is non-trivial with a probability of at least 48%. When computing a knot map, the faster Alexander polynomial is used. Once we have computed the knot map, we can determine the knot fingerprint and knot cores of all knot types formed by the structure. The whole algorithm is implemented in C/C++.

Graphical content and structure smoothing

The structures are visualized using PDBe Mol* v1.2 which was modified via JavaScript to highlight subchains using different coloring schemes and to hide side chains after highlighting. The PDBe version of Mol* was used to take advantage of implemented confidence coloring and pLDDT information for AlphaFold structures. We provide an additional menu bar to allow users to select the coloring mode, the representation mode (cartoon or backbone) and to highlight interesting subchains in the structure.

The so-called knot map image is generated using the Matplotlib library and manipulated by D3 and JQuery libraries to make an interactive information box of residue index, to highlight a knot in the structure, and to display information corresponding to a given knot cutoff. Users can also display arrows to show interpretations of different knots in the structure, i.e. the knot core, tails, their locations, and their lengths. The web interface is built with the newest versions of popular web libraries, e.g. Bootstrap 5, including the implementation of some simple interface animations and transitions.

Server and database implementation

The frontend of the server is implemented in Python 3 using the Flask framework. The data is stored in a MySQL database and accessed using SQLAlchemy. An asynchronous cluster tasks management solution—the kafka−slurm−agent (https://github.com/prubach/kafka-slurm-agent) was developed to enable seamless large-scale computations necessary to obtain the presented results for all organisms. Thanks to this solution, the knot detection and identification is computed on three independent Linux clusters managed by slurm (https://slurm.schedmd.com/). The kafka-slurm-agent is also used to manage the computing of tasks submitted by online users to our server. The knot detection and identification is computed using the Topoly package (6). Information about the proteins is downloaded from PDBe, RCSB and PFAM SIFTS and RESTful services. The whole service is installed on multicore Linux nodes.

SERVER DESCRIPTION

AlphaKnot consists of two parts: the submission server and the database.

Submission server

The server part of AlphaKnot performs a comprehensive topological analysis of single or multiple models uploaded by users, in the CIF, CIF.gz or PDB formats. The files can be models predicted by AlphaFold, in which case the topology confidence analysis will be most complete. But users can also upload PDB files resulting from other sources, e.g. from RoseTTa (7) predictions or from the RCSB database.

Users can select from several choices which affect the speed and accuracy of the computations. Figure 1 shows the submission page on the left with the various choices on the right. For the lists on the right, the options are shown with the speed of the computation from fastest to slowest. The results page, an example of which is shown in Figure 2, shows the types of knots, pLDDT confidence information for the knotted chain, an image matrix showing the positions of

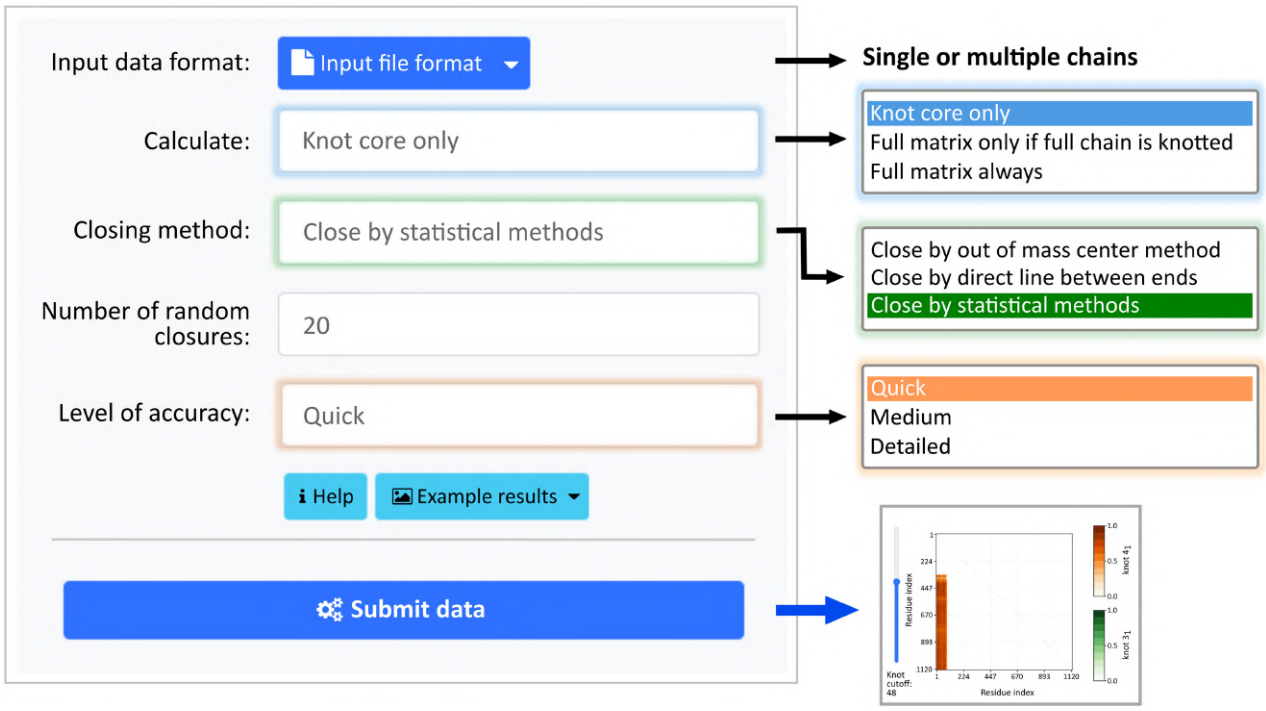


Figure 1. Submission options. The user can choose from several options which determine what methods will be used to describe the entanglement in the structure, and also determine the accuracy of the calculations. The calculation time on the server is directly related to the selected parameters (and to the length and complexity of the structure). For a single structure, computation time varies from a few seconds to several hours.

knotting along the chain, a 3D manipulatable model of the protein with colored knot locations, information about the position of the knot along the amino acid sequence, and other information about the protein.

Database of AlphaFold predictions from 21 proteomes

The database part of AlphaKnot contains knotting information for structures predicted prior to 2022 based on 21 proteomes (with pLDDT >50) published in (4). A browse feature allows users to see and compare all of the knotted proteins. The proteins are categorized manually (by visual inspection) into the categories of Knots (high confidence), Unsure (significant, but not insurmountable, issues exist in the model), and Artifact (very likely the entanglement is due to issues with the model).

This classification does not always agree with pLDDT confidence, e.g. in the Artifacts, users can find proteins with high pLDDT values but very unlikely topologies. Some unusual structures were also checked with the RoseTTa tool (7). For example, proteins in which AlphaFold predicted the new 5_1 knot type, were classified in AlphaKnot as Unsure (e.g. P53336) and Artifact (e.g. Q75JZ0) due to low pLDDT values and the fact that all known models of methyltransferases analyzed to date form 3_1 knots. (8). RoseTTa's models of these structures confirm AlphaKnot's classification, P53336 forms the 5_1 knot, and Q75JZ0 is unknotted. One way to explain the unknotted topology in the case of Q75JZ0 is that it could be due to the wrong template being used for prediction: the protein with PDB code 5ZYO (a protein which is artificially unknotted). We do not see the origin of the 5_1 knot. It would be interesting to conduct an experiment to investigate the existence of the 5_1 knot (potentially the first non-twisted knot type in a protein (9)) in these structures.

A search results in a list which links to an individual page for each protein. The individual pages contain the same information as is shown for submitted jobs, as well as other information about the protein and a list of proteins with similar sequences.

Features unique to AlphaKnot

The functions available on AlphaKnot are not available on any other server/website/software. AlphaFold's pLDDT measures the confidence of a predicted model, computed per amino acid. This is critical information which is needed in order to interpret the reliability of the knotting within a model. While there are very few resources for researchers trying to analyze entanglement in chains in general, AlphaKnot is the only platform for computing knotting of protein chains which includes pLDDT values. For user-submitted files, or within the included database, AlphaKnot identifies the knot cores, i.e. shortest chains realizing a given knot, within the chain and uses the pLDDT values to characterize to what extent the results should be trusted. Furthermore, in the 3D model, the full chain (or a given subchain) can be colored relative to the pLDDT values to allow the user to visualize the confidence one might have in the model. The other coloring modes, default (which is monochrome) and

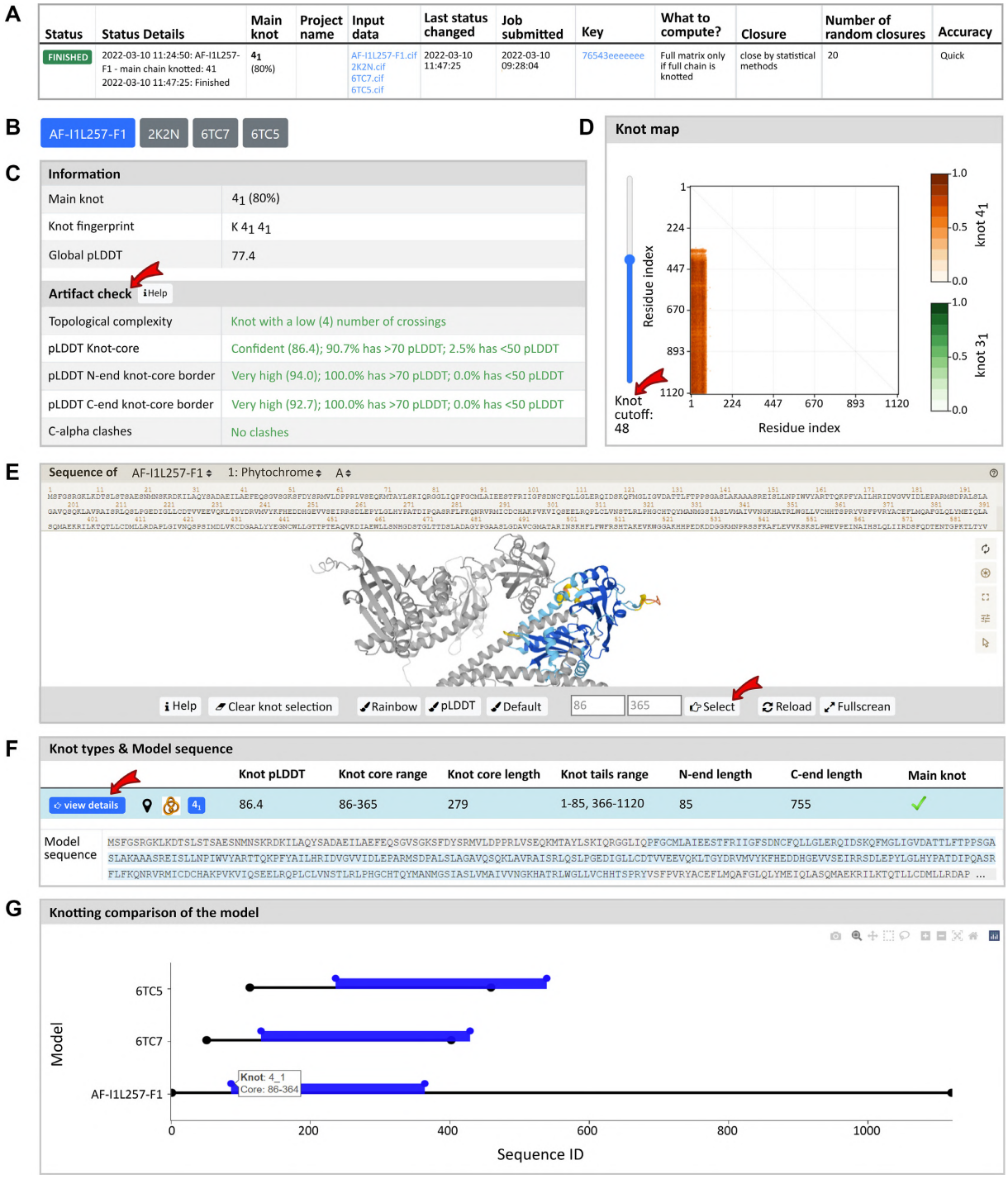


Figure 2. Example page with the output from a submission to AlphaKnot. Here the multiple chains option was chosen and four proteins with at least 40% sequence similarity were analyzed: I1L257, Q2JIZ5, P93527, B4YB07 (UniProtKB ID). (**A**) Job status table. The table shows details of the job, information about the methods used, the chosen parameters for the job, and the main results. (**B**) Structure choice. For multiple chain uploads, the user can choose the structure to view. Here I1L257 is selected, which forms the knot 4_1. (**C**) Information and Artifact check. The tables present basic information on the topology of the structure and results of some automatic tests on the quality of the structure and its knotted region. These results can help the user to recognize if the knotted topology may be an artifact—any information in red suggests that the user should be wary of the results. (**D**) The knot map (if full matrix was computed). The user can choose the cutoff—knot frequency required to show it on the matrix. (**E**) View of the structure. The user can choose the coloring method and select the fragment of the chain which is colored. In the figure the knot core region is colored by the pLDDT values, which is helpful in assessing whether the knot is an artifact. (**F**) Knot types and model sequence. The table contains detailed information about all knots formed by fragments of the chain (if full matrix was computed). By selecting a knot, the corresponding sequence is highlighted. (**G**) Comparison of the location and types of knots in each (knotted) structure. The color of the bar corresponds to the type of knot. The user can see on the chart that three of the four structures (besides Q2JIZ5) form the knot 4_1.

rainbow (where the color changes as one passes through the protein) provide modes that are more useful for visualizing the 3D nature of the protein.

Sometimes AlphaFold models are created by predicting smaller segments of the protein which overlap. As is shown in Figure 2, users can upload portions of AlphaFold predictions that overlap. In such a case, the models can disagree on the location or type of knot that is formed over a subchain. When multiple portions of a common protein are uploaded for analysis, AlphaKnot provides a tool for visualizing the similarity and differences in the knotting between the models, as is seen in Figure 2G. In that figure, we see that three of the four predictions predict a 4_1 knot, but with slightly different minimal cores. From these results, one can assess how stable the knotting is within the set of predictions. Since the knot does not appear in one of the models, the predictions are not consistent and this suggests areas within the structure that garner further attention.

Improvements over our KnotProt server

AlphaKnot also has a number of improvements over our server KnotProt (1,10). Many of these were inspired by the unique opportunities (such as the pLDDT values mentioned above) and challenges due to the size and complexity of knots observed in the AlphaFold models. First, the Topoly package (freely available, see (6)), is fully integrated into AlphaKnot and is able to recognize more complex knot types than KnotProt. More precisely on the KnotProt, knot types through eight crossings are recognized, while on the AlphaKnot, knot types through 12 crossings are recognized. Second, AlphaKnot allows users to specify a cut-off percentage in the interactive knot map. At the cut-off value of 0.3, for example, one sees all subchains whose predominate non-trivial knot type occurs with probability at least 0.3. As in KnotProt, the opacity of the color in each cell corresponds to this probability value. Increasing the cut-off value results in fewer, but more robustly knotted chains. This feature allows users to visualize how and where the knotting is created within the protein. Third, AlphaKnot computes knot cores for uploaded models. This functionality does exist in some other software for proteins, like KymoKnot (11), however without the critical pLDDT analysis specific for AlphaFold predicted structures. Fourth, AlphaKnot implements a number of speed improvements to make the calculations more efficient. The chain lengths of the protein models from AlphaFold are much larger than what is seen in the PDB. Furthermore, the complexity of knotting observed is much higher. These improvements were needed so that the submission server would be feasible.

User experience of the AlphaKnot

The AlphaKnot server and database has a number of features focused on providing an intuitive and robust user experience. First, the advanced search allows users to search by knot type, core length, tail length, or fingerprint filtered by proteome (or proteomes) and within the classes of Knots, Unsure or Artifacts. The filtering choices are shown in Figure 3. Second, by hovering over an amino acid in the 3D model or hovering over an amino acid in the sequence, one

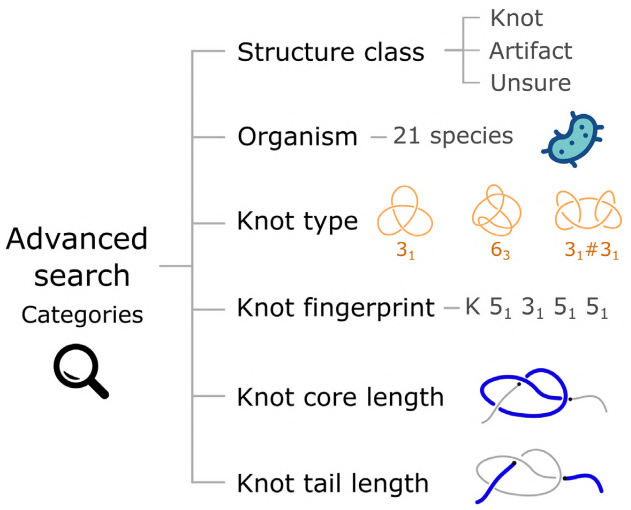


Figure 3. Features that can be searched for knotted proteins in Advanced Search in the AlphaKnot database.

sees the protein information, and the amino acid's type, number and pLDDT score for the amino acid. Third, the knot core search algorithm is improved and provides a more accurate calculation of the core and N- and C-termini tails. Fourth, each database protein (or uploaded model) provides output from a number of automated checks on the integrity of the model. For example, if the distance (in space) between amino acids is too small the user receives a warning. Also, low pLDDT scores for the full chain, and within knotted subchains, are highlighted for the user. Lastly, for every structure in the database, there are links which direct the user to the corresponding entries in UniProt and AlphaFold. When a model overlaps with an experimentally-solved structure from the PDB, the corresponding PDB and KnotProt links are also given.

ONLINE DOCUMENTATION

AlphaKnot provides a rich set of online documentation with sections (i) about—providing an overview of the website, (ii) knot detection—detailing the methods used to compute the knotting, (iii) how to interpret the knotting data—explaining how one can interpret the entanglement based on the knot map, (iv) how to use the server—showing how users can upload proteins, the choices available, the speed-effect on the different choices, and how to interpret the results from single or multiple chain output, (v) how to search and browse the database—guiding the user on how to search within the database of AlphaFold predictions, (vi) database statistics—presenting information about the protein models in the database as well as downloadable text files which list information about these models, (vii) API—documenting an API for downloading sets of data according to different search criteria.

APPLICATIONS

Artificial intelligence (AI) methods are used commonly to determine protein structures from less investigated genomes

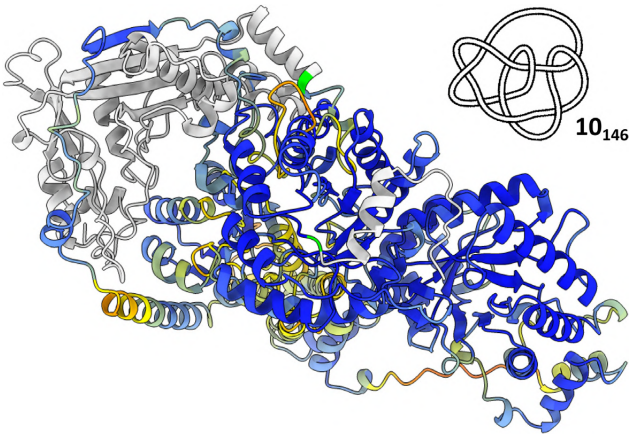


Figure 4. Protein (UniProtKB ID: Q54P92) with very high pLDDT value, but whose topology (knot with 10 crossings $K10_{146}$, shown with schematic) suggests that the prediction is not reliable.

or with unknown homological models. AI methods and the constantly growing AlphaFold database facilitate a wide range of research based on 3D protein structures. However, a careful analysis of knotting within these models can provide unique insights into the the predicted structures which can be used for many different applications. Below we provide some examples.

The analysis of all proteins from the PDB has shown that the topology is conserved between proteins with the same function but low sequence similarity (3) with one exception (12). Computations by the AlphaKnot server on proteins in AlphaFold database with the same biological function but from different organisms show that in some cases the topology indeed is strictly conserved, implying that a detected topology has a high probability of being correctly predicted. As a result, new knotted protein families have been discovered, e.g. tRNA-uridine aminocarboxypropyltransferase 1 (PF03942) with a 3_1 knot located in the active site. On the other hand, in some cases the topologies are seemingly random (5), strongly suggesting that the modeled structures are not correct, e.g. as is shown in Figure 4. There are also cases when the topology is conserved only between some homological proteins, e.g. integrin as noted in (5). Such families open the door for an evolutional analysis (determining the origin of a non-trivial topology) or even the detection of an amino acid motif responsible for creating a non-trivial topology. Also, some potential new types of knots have been found, e.g. in Putative methyltransferase (UniProtKB ID: YGR283C). This knot type, denoted 5_1 in mathematical notation, would necessitate a more complex folding path than any other knotted protein characterized to date. This knot is called potential since not all homological structures deposited in the AlpahFold possess it.

Other potential applications include in silico drug design or structural biology studies. For example proteins from the SPOUT family, such as the TrmD protein are on the list of high importance proteins due to 12 drug-resistant bacteria (based on World Health Organization) that should be treated as high-priority pathogens. This makes TrmD, a vital bacterial protein present in all of these strains, an excellent antimicrobial target for drug design. On the other hand, in all experimentally-solved structures from the SPOUT clan, the active site is embedded in a tightly packed knot even though they share low sequence similarity. Therefore, a correct prediction of topology is necessary for a successful search for a potent drug. Moreover, the majority of knotted proteins are enzymes which have an active site embedded in the knot. Accurate predictions of knots are crucial for further structural biology investigations.

We present one further example which highlights a possible issue with relying on pLDDT values and how AlphaKnot can provide critical insights in such cases. Figure 4 shows a protein structure (UniProtKB ID: Q54P92) with a high pLDDT reliability which is certainly not a realistic model from the topological perspective. The knot formed is 10_{146}, which is much more complicated than any protein deposited in the PDB. The high pLDDT values suggest a good quality fold prediction for the structure. The pLDDT values along the knot core are also high. However, the AlphaKnot topological analysis reveals that the topological arrangement is highly unlikely. We used the AlphaKnot server to analyze 27 homological structures deposited in the AlphaFold database and found that the majority have the trivial topology (although two form the 3_1 knot and some form the 10_{146} knot). Furthermore, we used RoseTTa to predict the structure for Q54P92 and found that the model is unknotted. Together, these results suggest that all members are probably unknotted. Without the use of AlphaKnot, this prediction could be interpreted as being reliable based on the pLDDT values. However, AlphaKnot has exposed that researchers should be wary of this structure.

We would like to point out that some of the potentially wrongly predicted topologies may be due to the fact that some structures in the PDB database contain gaps. Modeling long gaps as straight segments can easily change the types of knots observed. Since the structures predicted by AlphaFold do not have gaps, one may not be aware that AlphaFold used structures with gaps in the learning process. One example is the family including the sodium-driven chloride bicarbonate exchanger. The original predicted structure, based on the human genome UniProtKB ID: Q6U841, has good quality (high pLDDT score over 70) and forms a 3_1 knot. Currently 86 homologues can be found in AlphaFold. We found that 85 out of its 86 homologs from other organisms are unknotted. That is, although AF 2 achieves incredible effectiveness in protein structure prediction, there is clearly still room for improvement in specific cases.

The AlphaKnot server is based on protein analysis, therefore it has many biological applications. Nevertheless, its versatility makes it useful in other areas of research as well.

SUMMARY

AlphaKnot provides a user-friendly submission server combined with a rich database of knotting analysis for AlphaKnot predicted structures. The use of AlphaFold's pLDDT values is critical for interpreting the reliability of the knotting within models and is only found on this server. The 3D models colored by pLDDT and multi-chain analysis for overlapping protein subchains allows users to efficiently vi-

sualize the quality of the models. In addition, AlphaKnot provides unique insights into the reliability of predictions which are not available by using the pLDDT alone. Furthermore, the AlphaKnot server is a tool to conduct fast topological analysis of sets of related structures (e.g. from different organisms) to assess the quality of the predictions and suggest possible improvements for the AlphaFold algorithm.

DATA AVAILABILITY

The AlphaKnot web server, with the online documentation, is publicly available at https://alphaknot.cent.uw.edu.pl.

ACKNOWLEDGEMENTS

We thank Center for Machine Learning from University of Warsaw (Center4ML) for help with modeling protein structures using AlphaFold v2.0 pipeline.

FUNDING

National Science Centre [UMO-2018/31/B/NZ1/04016 to J.I.S.]; European Biology Organization [EMBO Installation Grant 2057 to J.I.S.]; National Science Foundation [1720342 to E.J.R.]. Funding for open access charge: National Science Centre [UMO-2018/31/B/NZ1/04016 to J.I.S.]

Conflict of interest statement. None declared.

REFERENCES

1. Jamroz,M., Niemyska,W., Rawdon,E.J., Stasiak,A., Millett,K.C., Sułkowski,P. and Sulkowska,J.I. (2015) KnotProt: a database of proteins with knots and slipknots. *Nucleic Acids Res.*, **43**, D306–D314.
2. Sulkowska,J.I. (2020) On folding of entangled proteins: knots, lassos, links and θ-curves. *Curr. Opin. Struct. Biol.*, **60**, 131–141.
3. Sułkowska,J.I., Rawdon,E.J., Millett,K.C., Onuchic,J.N. and Stasiak,A. (2012) Conservation of complex knotting and slipknotting patterns in proteins. *Proc. Nat. Acad. Sci. U.S.A.*, **109**, E1715–E1723.
4. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
5. Perlinska,A.P., Niemyska,W.H., Gren,B.A., Rubach,P. and Sulkowska,J.I. (2022) New 63 knot and other knots in human proteome from AlphaFold predictions. bioRxiv doi: https://doi.org/10.1101/2021.12.30.474018, 01 January 2022, preprint: not peer reviewed.
6. Dabrowski-Tumanski,P., Rubach,P., Niemyska,W., Gren,B.A. and Sulkowska,J.I. (2020) Topoly: Python package to analyze topology of polymers. *Brief. Bioinformatics*, **22**, bbaa196.
7. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D. *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
8. Jarmolinska,A.I., Perlinska,A.P., Runkel,R., Trefz,B., Ginn,H.M., Virnau,P. and Sulkowska,J.I. (2019) Proteins' knotty problems. *J. Mol. Biol.*, **431**, 244–257.
9. Bölinger,D., Sułkowska,J.I., Hsu,H.-P., Mirny,L.A., Kardar,M., Onuchic,J.N. and Virnau,P. (2010) A Stevedore's protein knot. *PLoS Comput. Biol.*, **6**, e1000731.
10. Dabrowski-Tumanski,P., Rubach,P., Goundaroulis,D., Dorier,J., Sułkowski,P., Millett,K.C., Rawdon,E.J., Stasiak,A. and Sułkowska,J.I. (2018) KnotProt 2.0: a database of proteins with knots and other entangled structures. *Nucleic Acids Res.*, **47**, D367–D375.
11. Tubiana,L., Polles,G., Orlandini,E. and Micheletti,C. (2018) KymoKnot: a web server and software package to identify and locate knots in trajectories of linear or circular polymers. *Eur. Phys. J. E*, **41**, 72.
12. Virnau,P., Mirny,L.A. and Kardar,M. (2006) Intricate knots in proteins: function and evolution. *PLoS Comput. Biol.*, **2**, e122.

Article 5

AlphaFold predicts novel human proteins with knots

Received: 16 September 2022 | Revised: 16 March 2023 | Accepted: 20 March 2023

DOI: 10.1002/pro.4631

FULL-LENGTH PAPER

THE PROTEIN SOCIETY WILEY

AlphaFold predicts novel human proteins with knots

Agata P. Perlinska[1] | **Wanda H. Niemyska**[1,2] | **Bartosz A. Gren**[1,3] | **Marek Bukowicki**[3] | **Szymon Nowakowski**[2,3] | **Pawel Rubach**[1,4] | **Joanna I. Sulkowska**[1]

[1]Centre of New Technologies, University of Warsaw, Warsaw, Poland

[2]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland

[3]Faculty of Physics, University of Warsaw, Warsaw, Poland

[4]Warsaw School of Economics, Warsaw, Poland

Correspondence
Joanna I. Sulkowska, Centre of New Technologies, University of Warsaw, Banacha 2c 02-097, Warsaw, Poland.
Email: j.sulkowska@cent.uw.edu.pl

Funding information
Ministerstwo Nauki i Szkolnictwa Wyższego, Grant/Award Number: 0003/ID3/2016/64; Narodowe Centrum Nauki, Grant/Award Number: 2021/43/I/NZ1/03341

Review Editor: Nir Ben-Tal

Abstract

The fact that proteins can have their chain formed in a knot is known for almost 30 years. However, as they are not common, only a fraction of such proteins is available in the Protein Data Bank. It was not possible to assess their importance and versatility up until now because we did not have access to the whole proteome of an organism, let alone a human one. The arrival of efficient machine learning methods for protein structure prediction, such as AlphaFold and RoseTTaFold, changed that. We analyzed all proteins from the human proteome (over 20,000) determined with AlphaFold in search for knots and found them in less than 2% of the structures. Using a variety of methods, including homolog search, clustering, quality assessment, and visual inspection, we determined the nature of each of the knotted structures and classified it as either knotted, potentially knotted, or an artifact, and deposited all of them in a database available at: https://knotprot.cent.uw.edu.pl/alphafold. Overall, we found 51 credible knotted proteins (0.2% of human proteome). The set of potentially knotted structures includes a new complex type of a knot not reported in proteins yet. That knot type, denoted 6_3 in mathematical notation, would necessitate a more complex folding path than any knotted protein characterized to date.

KEYWORDS

biological function, evolution, folding, new knotted folds

1 | INTRODUCTION

The AlphaFold method (Jumper et al., 2021) has already led to high-quality predictions for thousands of protein structures from different genomes. AlphaFold's approach bypasses the need to understand the complex rules that determine the physical protein-folding process. An in silico approach can lead to the prediction of new folds and structures which are hard to determine using approaches such as crystallization NMR or cryo-EM. AlphaFold uses a per-residue confidence score (pLDDT) to estimate the quality of prediction. However, is the pLDDT score sufficient for quality assessment, especially in the case of long multi-domain proteins? Since it is known that the probability of a linear chain being knotted increases with its length, can folding rules be simply bypassed? Are these longer protein structures more complicated than we can currently model accurately?

The fact that proteins can have their chain formed into a knot has been known for almost 30 years (Mansfield, n.d.; Mansfield, 1997; Takusagawa & Kamitori, 1996). Topology, and more specifically knot theory, provides tools to recognize different types of knots (e.g., Alexander and HOMFLY-PT polynomials [Alexander, 1928; Freyd et al., 1985; Przytycki & Traczyk, 2016]). Although knot theory defines a knot as a "closed curve", we can still study knots in proteins by connecting their N and C ends to use knot theory methods (Millett et al., 2013). So far, five types of knots have been found in experimentally solved protein structures (besides unknot these are 3_1, 4_1, 5_2, and 6_1 prime knots [Jamroz et al., 2015; Dabrowski-Tumanski et al., 2019) and $3_1\#3_1$ complex knot (Bruno da Silva et al., 2023]). All of these prime knots are twist knots which means that they can be easily tied by twisting a loop of an open chain, pulling one end of the chain through the loop, and connecting the ends. Does AlphaFold predict protein structures with other types of knots?

Herein, we show the advantages of considering the topology of a model generated by a protein structure prediction method. In most cases, the presence of knots in proteins is conserved within a family, thus the topology of the predicted model should be the same as its homologs. Otherwise, there is a strong probability that the model was incorrectly predicted.

Furthermore, the topological analysis of proteins, especially human ones, can lead to the understanding of how evolution coded such high levels of protein organization (Sulkowska, 2020). Recent studies have shown that some proteins form open knots in their native folded structure (Mallam & Jackson, 2005; Mansfield, 1994; Sulkowska, 2020). These protein knots resemble well-known rope knots (Sułkowska et al., 2012). In general, the percentage of known knotted proteins is much lower than would be expected in random polymers with a similar length, compactness, and flexibility (Virnau et al., 2006). Among known protein structures only the simplest of knot types have been observed (Jamroz et al., 2015) and, furthermore, only ones which can be constructed by single threading (Taylor, 2000). Numerical simulation has shown that many proteins can self-tie (a Beccara et al., 2013; Li et al., 2012; Noel et al., 2013; Sułkowska et al., 2009) (even a protein with the 6_1 knot (Bölinger et al., 2010)), and knotting can be assisted by a chaperone (Soler et al., 2016) or ribosome (Baiesi et al., 2019; Chwastyk & Cieplak, 2015; Dabrowski-Tumanski et al., 2018), especially in the case of deeply knotted proteins (Jamroz et al., 2015). An in vitro investigation additionally has shown that some proteins can self-tie (Wang, Chen, & Hsu, 2015), although chaperones facilitate their folding (Andrews et al., 2013; Jackson et al., 2017; King et al., 2010; Mallam & Jackson, 2012; Wang, Liu, et al., 2015; Ziegler et al., 2016). Interestingly, although its biological role is not clear (Jackson et al., 2017; Sulkowska, 2020), knotting has been found in proteins from all branches of the tree of life and is conserved even among the sequences with low similarity (Sułkowska et al., 2012). The presence of knots in proteins raises many fundamental questions, from which we list here just a few: (1) Is the small percentage of knotted proteins deposited in the PDB observed due to the specificity of the data? (2) Are there more complex knots which have yet to appear in the PDB? (3) What is the origin of knotted proteins? (4) Is topology strictly conserved? (5) What role do knots play in proteins?

Based on human protein structures predicted by AlphaFold, we have found answers to some of these questions. We have conducted a comprehensive review of all 23,391 structures predicted by AlphaFold for the human proteome. For each protein structure, we determined the dominant knot type (Jamroz et al., 2015) and found the location of the knot cores (i.e., minimal portions of the protein backbone that form a given knot type). As a result, we found 340 knotted structures that we then further analyzed. All these proteins are deposited in an online database to enable further in vivo and in silico investigation, available at: https://knotprot.cent.uw.edu.pl/alphafold. They are also available in AlphaKnot database alongside with knotted proteins of 20 other proteomes at https://alphaknot.cent.uw.edu.pl. (Niemyska et al., 2022)

After careful evaluation, we concluded that over 75% of the models containing knots are artifacts. However, most of the previously known knotted human proteins were correctly predicted, which shows that AlphaFold is capable of modeling structures with complicated topology (see Table S1 for details about all proteins we classified as knotted). Compelling evidence for this is shown by a bacterial protein from *Aquifex aeolicus*, which AlphaFold models with correct knotted topology even though the single template available (TrmD protein, PDB ID: 1oy5, UniProtKB ID: O67463) is unknotted. On the other hand, it is not always the case since there are examples of errors in AlphaFold models that are due to problems in the templates from the PDB (Tata et al., 2022).

All the structures that were not classified as artifacts were put through a more detailed evaluation that included a topology analysis of homologous proteins found in the AlphaFold database. Using sequence clustering we grouped the proteins by similarity and checked if the topology is conserved both within each cluster and within the whole family of homologs. As a result, we obtained two additional pieces of information: (1) further verification of a given knot's presence in the human

protein; and (2) whether the knot is robustly present (conserved) in each family. Interestingly, these results also emphasized the models that were of good quality but probably of the wrong topology. Since we obtained the models from an early version of AlphaFold Protein Structure Database we recalculated problematic structures by our locally installed AlphaFold 2 version. For example, the model of Testis anion transporter 1 protein (UniProtKB ID: Q96RN1) available in the AlphaFold Protein Structure Database v.1 has a 3_1 knot (with high average pLDDT for the knot core being 86.1), whereas 96% of its homologs are unknotted. After recalculation with AlphaFold 2, we obtained a more credible unknotted model of this protein. Finally, for additional verification of the knotted models, we used RoseTTaFold to predict their structures and compare topologies.

Here, we established how many knotted proteins are predicted to be in a single proteome and found new families of knotted proteins. Recently, an article searching for the most complex knot types was published (Brems et al., 2022). It also analyzes AlphaFold predicted structures (including human proteins) and the topology results agree with ours. However, this extensive survey focuses on the most complex knot types and does not report any new knotted proteins with less than five crossings nor with chains longer than 600 aa nor with shallow knots (with tails shorter than 5 aa) nor with low quality (pLDDT lower than 80) (Brems et al., 2022). In our study, we find that all newly identified knotted human proteins are 3_1 knotted.

2 | TYPES OF PROTEIN KNOTS FOUND

Thus far four main types of knots have been observed in experimentally solved proteins—3_1, 4_1, 5_2, and 6_1 (Jamroz et al., 2015). Here, we have found three types—3_1, 5_2, and 6_3 (Table 1). The 6_3 knot has not been seen before in a protein. Note that the 4_1 and 6_1 knots that have not been detected in our study were previously observed mostly in bacterial and plant proteomes (16 proteins with 4_1 and two with 6_1).

2.1 | New 3_1 knotted proteins

Overall, we found 51 credible (robustly) knotted structures. Within this group, there are proteins already known as knotted because either their structure was resolved experimentally, or they are part of an established knotted family (e.g., SPOUT clan of knotted methyltransferases). Interestingly, in some cases, we were able to find knots in proteins considered unknotted before due to their incomplete crystal structure (as in the Ion channel regulator).

The remaining proteins make a group of newly identified knotted proteins (Table 2). All of them contain a subchain forming the 3_1 knot which is overall the most common knot type found in proteins.

2.2 | Ecdysoneless protein

We found the human Ecd (ecdysoneless) protein (UniProtKB: O95905) to have been modeled by AlphaFold with a deep 3_1 knot. (The same topology is formed in the models from RoseTTaFold). This 644 amino acid long structure has the knot positioned between 208 and 314 amino acid with average pLDDT for this region being 93/100. The knot is conserved in the homologs of the protein, including the original *Dropsophila melanogaster* Ecdysoneless protein (UniProtKB ID: Q9W032) from which the name was derived. None of the homologous proteins have structures resolved experimentally, including the whole SGT1 family (Pfam ID: PF07093) that they are part of. This makes the knot identification

TABLE 1 Number of knotted structures with different knot types found within the human proteome based on AlphaFold prediction.

Knot types	No. knotted structures	No. potentially knotted	No. artifact knots
31	49	16	189
41	0	1	12
51	0	0	10
52	2	1	17
62	0	0	3
63	0	1	0
Other	0	0	41
All	51	18	271

TABLE 2 Proteins newly identified as knotted. Upper part: proteins with conserved knots found in models predicted by AlphaFold. Note that the topology matches between structures modeled by different softwares. Lower part: potentially knotted proteins with good quality models. Close homologs are represented by proteins with at least 60% sequence identity. See Methods section for details regarding structure classification.

	UniprotKB ID	Knot type AlphaFold/ RoseTTaFold	Protein name	Homologs with conserved topology (%)	No. homologs	Close homologs with conserved topology (%)	No. close homologs
CONSERVED KNOTTED	O95905	$3_1/3_1$	Protein ecdysoneless homolog	100	19	100	2
	Q14CN2	$3_1/3_1$	Calcium-activated chloride channel regulator 4	100	19	100	3
	Q5W0V3	$3_1/3_1$	FHF complex subunit HOOK interacting protein 2A	100	10	100	3
	Q86V87	$3_1/3_1$	FHF complex subunit HOOK interacting protein 2B	100	11	100	2
	Q8N612	$3_1/3_1$	FHF complex subunit HOOK interacting protein 1B	87	31	100	4
	A8K7I4	$3_1/3_1$	Calcium-activated chloride channel regulator 1	100	19	100	5
	Q9UQC9	$3_1/3_1$	Calcium-activated chloride channel regulator 2 exchanger 4	100	19	100	2
	Q96GM8	$3_1/3_1$	Target of EGR1 protein 1	90	10	100	4
	Q8N5C7	$3_1/3_1$	TRNA-uridine aminocarboxypropyltransferase 1	100	14	100	3
	Q8NBA8	$3_1/3_1$	TRNA-uridine aminocarboxypropyltransferase 2	100	9	100	3
POTENTIALLY KNOTTED	O00534	$6_3/3_1$	von Willebrand factor A domain-containing protein 5A	32	19	75	4
	Q9Y4D8	$4_1/0_1$	E3 ubiquitin-protein ligase HECTD4	0	0	0	0
	Q13683	$3_1/3_1$	Integrin alpha-7	39	90	100	2
	P23229	$3_1/3_1$	Integrin alpha-6	38	95	100	2
	P53708	$3_1/3_1$	Integrin alpha-8	37	95	100	4
	P06756	$3_1/3_1$	Integrin alpha-V	39	92	100	4
	P08648	$3_1/0_1$	Integrin alpha-5	39	94	67	3
	P26006	$3_1/0_1$	Integrin alpha-3	38	95	100	4
	P08514	$3_1/0_1$	Integrin alpha-IIb	39	95	100	2
	Q13075	$3_1/0_1$	Baculoviral IAP	100	8	100	6
	Q8TE82	$3_1/0_1$	SH3 domain and tetratricopeptide repeat-containing protein 1	100	7	100	2
	Q8TF17	$3_1/0_1$	SH3 domain and tetratricopeptide repeat-containing protein 2 subfamily U member 1	100	7	100	2
	Q9NYU2	$3_1/0_1$	UDP-glucose: glycoprotein glucosyltransferase 1	53	15	100	2
	Q7RTX0	$3_1/0_1$	Taste receptor type 1 member 3	4	325	100	6
	Q9NQV8	$3_1/0_1$	PR domain zinc finger protein 8	100	3	100	2
	Q6UXX5	$3_1/0_1$	Inter-α-trypsin inhibitor H6	0	0	0	0
	Q5CZC0-F17	$3_1/0_1$	Fibrous sheath-interacting protein 2	0	0	0	0
	Q709C8-F1	$3_1/0_1$	Vacuolar protein sorting-associated protein 13C	0	0	0	0

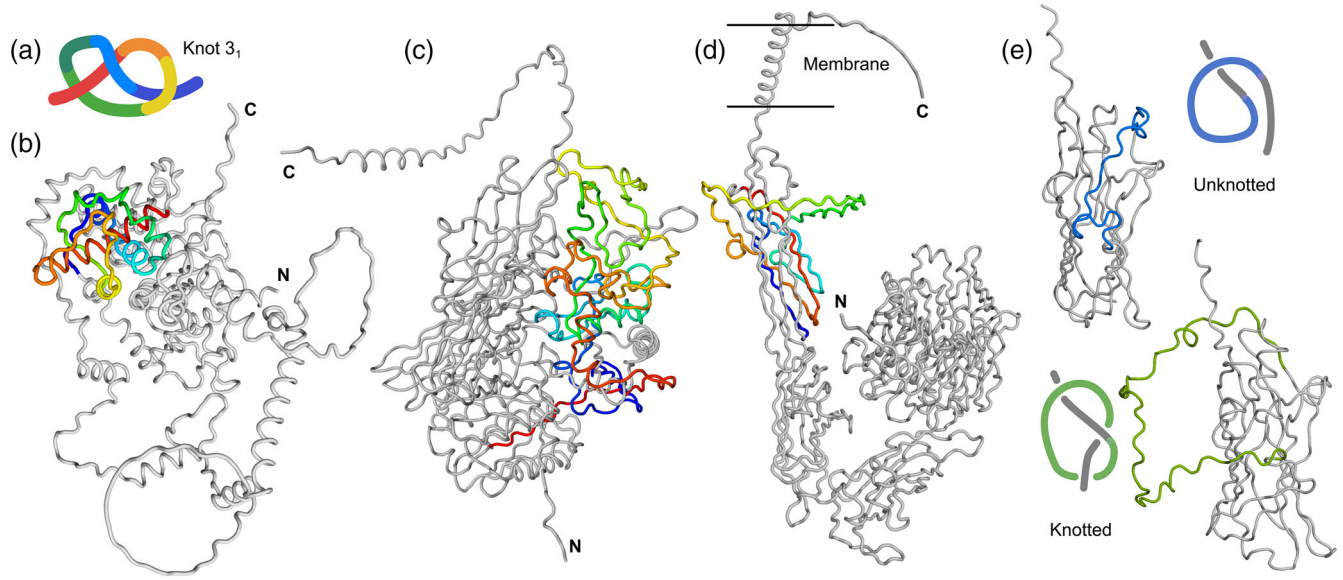


FIGURE 1 Selected proteins with a 3_1 knot. (a) Simplified representation of a 3_1 knot found in proteins. (b) Ecdysoneless protein (UniProtKB: O95905). (c) Calcium-activated chloride channel regulator 1 (UniProtKB ID: A8K7I4). (d) Integrin alpha IIb (UniProtKB ID: P08514). All knotted cores are shown in a rainbow color scheme to guide the eye. (e) Upper panel: unknotted integrin alpha 1 (UniProtKB ID: P56199). Lower panel: 3_1 knotted integrin alpha 3 (UniProtKB ID: P26006).

in this protein even more important, since SGT1 is a large family that contains over 2000 sequences. In human cells, Ecd was found to function as a regulator of the tumor suppressor p53 protein (Zhang et al., 2006). Also, the deficiency of Ecd lowers the fidelity of mRNA splicing (Erkelenz et al., 2021).

2.3 | Ion channel regulator

Calcium-activated chloride channel regulator 1 (CLCA1; UniProtKB ID: A8K7I4) is a 914 amino acid long protein with a 3_1 knot in its structure modeled by AlphaFold. Interestingly, this protein was resolved experimentally, although only one domain was crystallized (domain VWA, between residues 303–459). Also, this domain by itself is unknotted—the knot is present only in a full-length protein since it is located at residues 76–311 (Figure 1b). (Note, that when all subchains are considered the internal knot [slipknot] can be detected. This motif will be called $K3_13_1$). The knotted core is marked by AlphaFold as a high-confidence region (average pLDDT 92.4/100). Moreover, we found that paralogs of this protein (also predicted by AlphaFold: CLCA2 and CLCA4) and other homologs all have the same knot type. Similarly, the models calculated by RoseTTaFold also have a 3_1 knot. This shows that the prediction of a knot in this group of proteins is credible and the knot itself is a conserved feature that is likely to be advantageous for the protein.

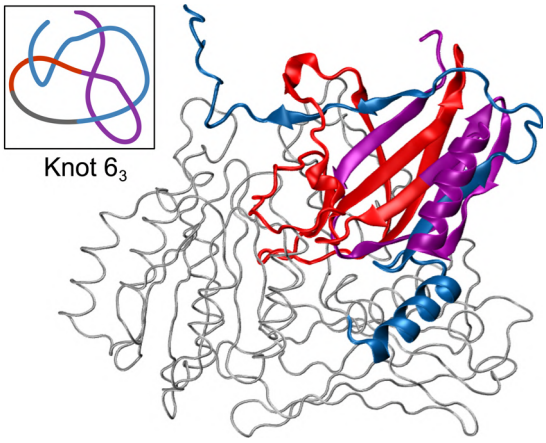


FIGURE 2 Von Willebrand factor A domain-containing protein 5A (UniProtKB ID: O00534) with a potential new type of protein knot—6_3 shown with essential strands colored in red, blue, and purple. Upper left panel shows a simplified knotted core of the protein.

CLCA proteins undergo self-cleavage (CLCA1 at position 695) to form the N- and C-terminal parts of the proteins. The N-terminal portion is responsible for the function of the protein, which is activating ion channels (Yurtsever et al., 2012). Based on our analysis we now know that this part is knotted, which was not previously observed. This information will shed new light onto the studies of CLCA proteins. Moreover, this is a great example that emphasizes the importance of having full structure information available for any protein studies.

2.4 | Knot or not?

We placed many constraints on when a model should be described as knotted (see Methods section). The structures that only met some were labeled as *potentially knotted*—usually due to either a low level of confidence in crucial parts of the knotted core or the topology not being conserved within the protein family. Those structures require further validation. In fact, this group forms a ready-to-use list of the most important protein targets for finding new knotted motifs, and thus can be used to significantly broaden the knowledge of entanglement in proteins (Table 2).

2.5 | Potentially knotted—6_3 knot in BCSC-1

We found a new complex type of knot in the von Willebrand factor A domain-containing protein 5A (BCSC-1, breast cancer suppressor candidate 1, Figure 2). The 6_3 knot is located in a high confidence region (average per-residue confidence score [pLDDT] is 88.6) between amino acids 45 and 625. It covers most of the protein, which is 786 residues long. However, because it has quite long tails (which consist of 45 and 161 amino acids, respectively), it is considered to be located rather deep within the model. This suggests that the knot will be stable in this structure and will not be untied spontaneously by thermal fluctuations of the protein.

Even though the model presents itself as accurate based on pLDDT, its 6_3 knot type is not conserved in other species. AlphaFold predicts both 6_3 and 3_1 knots within the BCSC-1 group of homologs (Table 2). Moreover, the model generated with RoseTTaFold also has a 3_1 knot. Finally, the models generated by ESMfold also have both 6_3 and 3_1 knots. Therefore, further experimental examination is needed to verify the existence of this complex knot, especially since this is not a twist knot (i.e., is not the result of twisting of a loop followed by a single threading [Taylor, 2000]). To tie such a knot, the protein chain must cross the energy barrier at least twice during folding as it is pulled through twisted loops. None of the experimentally resolved structures in PDB possess this type of knot. Right now, they are only found in predicted structures—here we report a 6_3 and recently 5_1 and 7_1 were found (Brems et al., 2022).

The BCSC-1 protein is interesting not only due to its complex topology, but also because of its crucial function—there are studies showing that it is involved in cancer development and can act as its suppressor (Di et al., 2018).

2.6 | Potentially knotted—integrins alpha

Often experimentally resolved structures have missing fragments that are not structurally organized. Such regions can be crucial for forming a knot since the way they will be modeled into the structure can either make a knot or not. The integrins alpha are a great example that shows the importance of this aspect, since within their models predicted by AlphaFold we found both knotted and unknotted structures.

We detected the 3_1 knot in seven out of 18 human integrins alpha (Table 2). In each of them, the knot is located in the C-terminal part of the protein (in the domain called Calf2; Figure 1C). Based on the AlphaFold predicted structures, it appears that whether an integrin will be knotted is determined by a single loop. Only integrins with longer loops are knotted (green region in Figure 1D). Unfortunately, this loop in most of the structures has a low confidence (pLDDT) score, making its location uncertain. Similarly, the topology of integrins' homologs is also mixed, making topology difficult to assess.

Therefore, we cannot verify whether the knot is present in these proteins—experimental evidence is needed. However, if only some of the integrins are indeed knotted, integrins would provide a second example of a family containing topologically distinct proteins—the first being aspartate/ornithine carbamoyltransferases (Sułkowska et al., 2008).

3 | CONSERVED KNOTTED PROTEINS

Table 3 shows proteins for which an entangled structure of a distinct homolog is known, and thus they are expected to be knotted. Most of them belong to four families (UCH, SPOUT, sodium/calcium exchanger, carbonic anhydrase-related). First, let us discuss transmembrane proteins, for which the knot conservation we found most interesting. Herein, all eight transmembrane proteins belong to only two families SLC24 and SLC8 (sodium/calcium exchanger). They all share a 9-helix inverted repeat motif, which forms a transmembrane channel, its active part 6-helix responsible for ion transport forms a 3_1 knot. SLC24 and SLC8 represent distant homologs of only one known family of knotted membrane proteins (represented by CAX, NCX) (Jarmolinska et al., 2019). These proteins belong to a huge membrane family PF01699, represented by more than 47,000 sequences. In this context, one is tempted to speculate whether all members possess nontrivial topology. Nevertheless, the 3_1 knot is conserved between SLC24, SLC8, and CAX,

TABLE 3 Proteins expected to be entangled. Here are the proteins that do not have a resolved structure but are part of a known family of knotted proteins, thus have a high probability of being correctly predicted as knotted. Close homologs are represented by proteins with at least 60% sequence identity.

UniprotKB ID	Knot type (AlphaFold)	Protein name	Homologs with conserved topology (%)	No. homologs	Close homologs with conserved topology (%)	No. close homologs
Q92560	5_2	Ubiquitin carboxyl-terminal hydrolase BAP1	88	8	75	4
Q6PF06	3_1	tRNA methyltransferase 10 homolog B	100	41	100	3
Q6IN84	3_1	rRNA methyltransferase 1, mitochondrial	96	28	100	2
Q9UJK0	3_1	18S rRNA aminocarboxypropyltransferase	81	37	100	3
P32418	3_1	Sodium/calcium exchanger 1	100	36	100	7
Q9UPR5	3_1	Sodium/calcium exchanger 2	100	36	100	9
P57103	3_1	Sodium/calcium exchanger 3	100	36	100	9
Q9UI40	3_1	Sodium/potassium/calcium exchanger 2	87	46	100	1
Q9HC58	3_1	Sodium/potassium/calcium exchanger 3	91	67	100	2
Q8NFF2	3_1	Sodium/potassium/calcium exchanger 4	88	34	50	2
Q71RS6	3_1	Sodium/potassium/calcium exchanger 5	88	42	100	3
Q6J4K2	3_1	Mitochondrial sodium/calcium exchanger protein	91	69	100	3
O75493	3_1	Carbonic anhydrase-related protein 11	99	156	100	6
Q9NS85	3_1	Carbonic anhydrase-related protein 10	99	159	100	7
Q9Y2D0	3_1	Carbonic anhydrase 5B, mitochondrial	98	172	100	3
P35218	3_1	Carbonic anhydrase 5A, mitochondrial	95	181	100	3

NCX proteins which share less than 30% similarity. Evolutionary conservation of a knotted topology, in this case, could imply its beneficial role, as it has been suggested for slipknotted membrane proteins (which also belong to Alpha-helical polytopic Transmembrane classes (based on OPM classification; Zayats et al., 2021). Therein, it was suggested that the channel may be held together more tightly when its α-helices are strapped together by a slipknot loop embracing several of the helices (Sułkowska et al., 2012).

Second, ubiquitin carboxyl-terminal hydrolase BAP1 (UCH-L2) with a 5_2 knot is a distant homolog of other UCH proteins. All homologs (UCHL1, UCHL3, UCHL5, UBL1-YEAST) with known structures possess the same type of topology (including internal 3_1 knots), even though their sequence similarity is very low, around 23% (Sułkowska et al., 2012). BAP1 is known to play a role in cancer, functioning both as a tumor suppressor and as a metastasis suppressor.

Third, there are plenty of carbonic anhydrase-related proteins already deposited in the PDB, they all are found to possess a 3_1 knot when a full sequence is determined. Finally, proteins with the UniProtKB ID: Q6PF06, Q6IN84, Q9UJK0 are members of SPOUT family (Tkaczuk et al., 2007), which is known to contain proteins with a conserved 3_1 knot embedded in the active site (Christian et al., 2016).

3.1 | Artifacts

More than 75% of knotted human structures from AlphaFold we found to be artifacts. In such structures, the knot is often predicted in low confidence regions (pLDDT < 50), which means its position and presence is not reliable. Moreover, about 1/3 of the structures with artifact knots represent large proteins (>2700 aa) that were predicted using multiple overlapping models. In fact, more than 50% of the artifact structures, versus only 17% of the credibly knotted structures, have more than 1000 residues. This behavior is expected, since the difficulty of predicting a structure (including its topology) increases with the length of the protein chain. Thus, topology can be used as a tool to identify the quality of experimental or simulated data. Some examples of interesting artifacts are described on the website.

4 | CONCLUSION

While knots are relatively rare in proteins, their conservation may suggest a functional utility. Many of them are enzymes with a knot located in the active site. However, as they are not common, only a fraction of such proteins is available in the PDB. It was not possible to assess their importance and versatility up until now because we did not have access to the whole proteome of an organism, let alone a human one. The arrival of efficient machine learning methods for protein structure prediction, such as AlphaFold, RoseTTaFold, and ESMFold changed that.

Herein, we analyzed all proteins from human proteome (over 20,000) determined with AlphaFold in search for knots and found them in less than 2% of the structures. Information about their topology, knot position, and quality (pLDDT score) is available in a database at: https://knotprot.cent.uw.edu.pl/alphafold. Importantly, after detailed analysis we found that the majority of the knotted proteins are artifacts and only 51 structures (0.2% of the proteome) are credibly knotted.

The set of credibly knotted proteins includes new knotted families and new types of knots (Table 2). The set of potentially knotted structures includes a new complex knot type in proteins. That knot type, denoted 6_3 in mathematical notation, would necessitate a more complex folding path than any knotted protein characterized to date.

To sum up, we select proteins with the best chance of being entangled, and which can be further analyzed without immediate structure determination. However, it would be beneficial to verify experimentally the structures of several proteins, including those of uncertain topology (Table 2), like integrins alpha. Also, our work emphasizes the importance of validation of the predicted models, as pLDDT in some cases is not sufficient to distinguish between correct and wrong models, in particular in terms of their topology.

5 | MATERIALS AND METHODS

The human proteome was downloaded from the AlphaFold Protein Structure Database (Jumper et al., 2021). It consisted of 23,391 structures (20,504 proteins). All the models were analyzed by computing the HOMFLY-PT polynomial for 200 random closures. The details of the method are explained in (Jamroz et al., 2015). We determined the position of the knot core in the structure based on the so-called matrix model (Jamroz et al., 2015). Homologs for each protein were obtained with GGSearch from EMBL-EBI (Madeira et al., 2022) using E-value cutoff 10^{-10} and searching within the AlphaFold Protein Structure Database. For the analysis of the conservation of topology in protein families, we clustered the proteins using CD-hit with a 60% sequence identity cutoff (Huang et al., 2010). Proteins with very low pLDDT were remodeled by our group using AlphaFold (version 2.1.1) (Jumper et al., 2021), with the program installed locally. For the RoseTTaFold prediction, we used software available on-line (Baek et al., 2021).

A structure with a knot is classified as credible when random closures form a nontrivial knot type more frequently than a trivial knot, the pLDDT score for the knot core is above 50, there are no clashes in the knot core region (checked with MolProbity tool [Williams et al., 2018]), the same knot type is found in more than 80% of the protein's homologs and in a model generated by RoseTTaFold, and no suspicious geometry was found after visual inspection. The structures which did not pass the visual inspection were classified as potentially knotted. The structures with obvious problems are called artifacts.

ACKNOWLEDGMENTS

We thank Eric Rawdon and Andrzej Stasiak for proof reading this article and the Center for Machine Learning, UW for computing support. This work was supported by the Polish Ministry for Science and Higher Education 0003/ID3/2016/64 (to JIS), the National Science Centre #UMO-2021/43/I/NZ1/03341 (to JIS) and COST EUTOPIA action. This research was carried out with the support of the Interdisciplinary Centre for Mathematical and Computational Modelling ICM) University of Warsaw under computational allocation no GS82-12.

AUTHOR CONTRIBUTIONS

Agata Paulina Perlińska: Formal analysis (equal); investigation (equal); methodology (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Wanda Niemyska:** Data curation (equal); validation (equal); investigation (equal); methodology (equal);writing – review and editing (equal). **Bartosz Greń:** Data curation (equal), investigation (equal); methodology (equal); software (equal); visualization (equal). **Marek Bukowicki:** Software (equal). **Szymon Nowakowski:** Software (equal). **Paweł Rubach:** Data curation (equal); investigation (equal); methodology (equal); software (equal); writing – review and editing (equal). **Joanna Sulkowska:** Formal analysis (equal); investigation (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Agata P. Perlinska https://orcid.org/0000-0003-2806-1190
Wanda H. Niemyska https://orcid.org/0000-0003-4167-1807
Bartosz A. Gren https://orcid.org/0000-0001-9971-6807
Marek Bukowicki https://orcid.org/0009-0009-2638-3437
Szymon Nowakowski https://orcid.org/0000-0002-1939-9512
Pawel Rubach https://orcid.org/0000-0001-5487-609X
Joanna I. Sulkowska https://orcid.org/0000-0003-2452-0724

REFERENCES

a Beccara S, Škrbić T, Covino R, Micheletti C, Faccioli P. Folding pathways of a knotted protein with a realistic atomistic force field. PLoS Comput Biol. 2013;9(3):e1003002.

Alexander JW. Topological invariants of knots and links. Trans Am Math So. 1928;30(2):275–306.

Andrews BT, Capraro DT, Sulkowska JI, Onuchic JN, Jennings PA. Hysteresis as a marker for complex, overlapping landscapes in proteins. J Phys Chem Lett. 2013;4(1):180–8.

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373(6557):871–6.

Baiesi M, Orlandini E, Seno F, Trovato A. Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding. Sci Rep. 2019;9(1):1–2.

Bölinger D, Sułkowska JI, Hsu HP, Mirny LA, Kardar M, Onuchic JN, et al. A Stevedore's protein knot. PLoS Comput Biol. 2010;6(4):e1000731.

Brems MA, Runkel R, Yeates TO, Virnau P. AlphaFold predicts the most complex protein knot and composite protein knots. Protein Sci. 2022;31(8):e4380.

Bruno da Silva F, Lewandowska I, Kluza A, Niewieczerzal S, Augustyniak R, Sulkowska JI. First crystal structure of double knotted protein TrmD-Tm1570—inside from degradation perspective. bioRxiv. 2023; 532328.

Christian T, Sakaguchi R, Perlinska AP, Lahoud G, Ito T, Taylor EA, et al. Methyl transfer by substrate signaling from a knotted protein fold. Nat Struct Mol Biol. 2016;23(10):941–8.

Chwastyk M, Cieplak M. Cotranslational folding of deeply knotted proteins. J Phys Condens Matter. 2015;27(35):354105.

Dabrowski-Tumanski P, Piejko M, Niewieczerzal S, Stasiak A, Sulkowska JI. Protein knotting by active threading of nascent polypeptide chain exiting from the ribosome exit channel. J Phys Chem B. 2018;122(49):11616–25.

Dabrowski-Tumanski P, Rubach P, Goundaroulis D, Dorier J, Sułkowski P, Millett KC, et al. KnotProt 2.0: a database of proteins with knots and other entangled structures. Nucleic Acids Res. 2019;47(D1):D367–75.

Di D, Chen L, Guo Y, Wang L, Wang H, Ju J. Association of BCSC-1 and MMP-14 with human breast cancer. Oncol Lett. 2018;15(4):5020–6.

Erkelenz S, Stanković D, Mundorf J, Bresser T, Claudius AK, Boehm V, et al. Ecd promotes U5 snRNP maturation and Prp8 stability. Nucleic Acids Res. 2021;49(3):1688–707.

Freyd P, Yetter D, Hoste J, Lickorish WR, Millett K, Ocneanu A. A new polynomial invariant of knots and links. Bulletin (new series) of the American mathematical. Society. 1985;12(2):239–46.

Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26(5):680–2.

Jackson SE, Suma A, Micheletti C. How to fold intricately: using theory and experiments to unravel the properties of knotted proteins. Curr Opin Struct Biol. 2017;42:6–14.

Jamroz M, Niemyska W, Rawdon EJ, Stasiak A, Millett KC, Sułkowski P, et al. KnotProt: a database of proteins with knots and slipknots. Nucleic Acids Res. 2015;43(D1):D306–14.

Jarmolinska AI, Perlinska AP, Runkel R, Trefz B, Ginn HM, Virnau P, et al. Proteins' knotty problems. J Mol Biol. 2019;431(2):244–57.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.

King NP, Jacobitz AW, Sawaya MR, Goldschmidt L, Yeates TO. Structure and folding of a designed knotted protein. Proc Natl Acad Sci. 2010;107(48):20732–7.

Li W, Terakawa T, Wang W, Takada S. Energy landscape and multi-route folding of topologically complex proteins adenylate kinase and 2ouf-knot. Proc Natl Acad Sci. 2012;109(44):17789–94.

Madeira F, Pearce M, Tivey AR, Basutkar P, Lee J, Edbali O, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. 2022;50(W1):W276–9.

Mallam AL, Jackson SE. Folding studies on a knotted protein. J Mol Biol. 2005;346(5):1409–21.

Mallam AL, Jackson SE. Knot formation in newly translated proteins is spontaneous and accelerated by chaperonins. Nat Chem Biol. 2012;8(2):147–53.

Mansfield ML. Are there knots in proteins? Nat Struct Biol. 1994;1(4):213–4.

Mansfield ML. Fit to be tied. Nat Struct Biol. 1997;4(3):166–7.

Millett KC, Rawdon EJ, Stasiak A, Sułkowska JI. Identifying knots in proteins. Biochem Soc Trans. 2013;41(2):533–7.

Niemyska W, Rubach P, Gren BA, Nguyen ML, Garstka W, Bruno da Silva F, et al. AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods. Nucleic Acids Res. 2022;50(W1):W44–50.

Noel JK, Onuchic JN, Sulkowska JI. Knotting a protein in explicit solvent. J Phys Chem Lett. 2013;4(21):3570–3.

Przytycki JH, Traczyk P. Invariants of links of Conway type. arXiv preprint *arXiv*:1610.06679 2016.

Soler MA, Rey A, Faísca PF. Steric confinement and enhanced local flexibility assist knotting in simple models of protein folding. Phys Chem Phys. 2016;18(38):26391–403.

Sulkowska JI. On folding of entangled proteins: knots, lassos, links and θ-curves. Curr Opin Struct Biol. 2020;60:131–41.

Sułkowska JI, Rawdon EJ, Millett KC, Onuchic JN, Stasiak A. Conservation of complex knotting and slipknotting patterns in proteins. Proc Natl Acad Sci. 2012;109(26):E1715–23.

Sułkowska JI, Sułkowski P, Onuchic J. Dodging the crisis of folding proteins with knots. Proc Natl Acad Sci. 2009;106(9):3119–24.

Sułkowska JI, Sułkowski P, Szymczak P, Cieplak M. Stabilizing effect of knots on proteins. Proc Natl Acad Sci. 2008;105(50): 19714–9.

Takusagawa F, Kamitori S. A real knot in protein. J Am Chem Soc. 1996;118(37):8945–6.

Tata RB, Alsulami AF, Sheik Amamuddy O, Blundell TL, Tastan BÖ. Slipknot or crystallographic error: a computational analysis of the plasmodium falciparum DHFR structural folds. Int J Mol Sci. 2022;23(3):1514.

Taylor WR. A deeply knotted protein structure and how it might fold. Nature. 2000;406(6798):916–9.

Tkaczuk KL, Dunin-Horkawicz S, Purta E, Bujnicki JM. Structural and evolutionary bioinformatics of the SPOUT superfamily of methyltransferases. BMC Bioinform. 2007;8(1):1–31.

Virnau P, Mirny LA, Kardar M. Intricate knots in proteins: function and evolution. PLoS Comput Biol. 2006;2(9):e122.

Wang I, Chen SY, Hsu ST. Unraveling the folding mechanism of the smallest knotted protein, MJ0366. J Phys Chem B. 2015; 119(12):4359–70.

Wang LW, Liu YN, Lyu PC, Jackson SE, Hsu ST. Comparative analysis of the folding dynamics and kinetics of an engineered knotted protein and its variants derived from HP0242 of helicobacter pylori. J Phys Condens Matter. 2015;27(35): 354106.

Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al. MolProbity: more and better reference data for improved all-atom structure validation. Protein Sci. 2018;27(1): 293–315.

Yurtsever Z, Sala-Rabanal M, Randolph DT, Scheaffer SM, Roswit WT, Alevy YG, et al. Self-cleavage of human CLCA1 protein by a novel internal metalloprotease domain controls calcium-activated Chloride Channel activation*◆. J Biol Chem. 2012;287(50):42138–49.

Zayats V, Perlinska AP, Jarmolinska AI, Jastrzebski B, Dunin-Horkawicz S, Sulkowska JI. Slipknotted and unknotted monovalent cation-proton antiporters evolved from a common ancestor. PLoS Comput Biol. 2021;17(10):e1009502.

Zhang Y, Gurumurthy CB, Kim J, Bhat I, Gao Q, Dimri G, et al. The human orthologue of drosophila ecdysoneless protein interacts with p53 and regulates its function. Cancer Res. 2006;66(14):7167–75.

Ziegler F, Lim NC, Mandal SS, Pelz B, Ng WP, Schlierf M, et al. Knotting and unknotting of a protein in single molecule experiments. Proc Natl Acad Sci. 2016;113(27):7533–8.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Perlinska AP, Niemyska WH, Gren BA, Bukowicki M, Nowakowski S, Rubach P, et al. AlphaFold predicts novel human proteins with knots. Protein Science. 2023;32(5):e4631. https://doi.org/10.1002/pro.4631

Article 6

Discovery of a trefoil knot in the RydC RNA: challenging previous notions of RNA topology

Research Article

Discovery of a trefoil knot in the RydC RNA: Challenging previous notions of RNA topology

Wanda Niemyska [1,2,*], Sunandan Mukherjee [3], Bartosz A. Gren [1], Szymon Niewieczerzal [1], Janusz M. Bujnicki [3,*] and Joanna I. Sulkowska [1,*]

*1 - **Centre of New Technologies,** University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland*
*2 - **Faculty of Mathematics,** Informatics and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland*
*3 - **Laboratory of Bioinformatics and Protein Engineering,** International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland*

***Correspondence to Wanda Niemyska, Janusz M. Bujnicki and Joanna I. Sulkowska:** wanda@mimuw.edu.pl (W. Niemyska), janusz@iimcb.gov.pl (J.M. Bujnicki), jsulkowska@cent.uw.edu.pl (J.I. Sulkowska)*
https://doi.org/10.1016/j.jmb.2024.168455
Edited by: David Mathews

Abstract

Knots are very common in polymers, including DNA and protein molecules. Yet, no genuine knot has been identified in natural RNA molecules to date. Upon re-examining experimentally determined RNA 3D structures, we discovered a trefoil knot 3_1, the most basic non-trivial knot, in the RydC RNA. This knotted RNA is a member of a small family of short bacterial RNAs, whose secondary structure is characterized by an H-type pseudoknot. Molecular dynamics simulations suggest a folding pathway of the RydC RNA that starts with a native twisted loop. Based on sequence analyses and computational RNA 3D structure predictions, we postulate that this trefoil knot is a conserved feature of all RydC-related RNAs. The first discovery of a knot in a natural RNA molecule introduces a novel perspective on RNA 3D structure formation and on fundamental research on the relationship between function and spatial structure of biopolymers.

Introduction

The spatial structure of biopolymers is closely related to their function [1–5]. Therefore, a lot of effort is put into understanding the spatial structure of these molecules. Among the various shapes, there are special – knotted configurations. We have personal experience with knots for example from sailing or we still remember the headphones with a cable pulled out of the pocket, almost every time maliciously tied in a knot. Fig. 1 shows the simplest knots, i.e. those with the smallest number of crossings when projected onto a plane: the version with closed curves (these are studied by the mathematical knot theory) and open ones (physical knots). Knotting is statistically almost certain in long polymer chains, especially under spatial confinement. Polymers form knots with increasing probability with increasing chain length and compactness [6,7]. A different situation is observed in the case of biopolymers such as RNA, DNA and proteins.

The most well known and historically first discovered are knots in circular DNA molecules [8]. Particularly, knots in DNA are most common in plasmids and viral DNA, and in prokaryotic circular chromosomes, where they are formed and resolved by the action of topoisomerases. Such knots in DNA can significantly influence processes such as replication, transcription, and packing into viruses [9,10]. We refer readers to the latest reviews [11].

For a long time, it was thought that the backbone of a protein formed only a trivial topology, but it is now known that at least 2% of proteins form

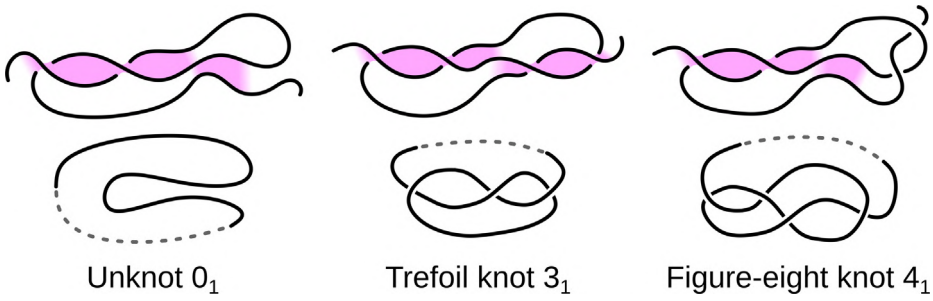


Fig. 1. Three simplest knot types, each type in different column. At the top of the figure are examples of differently knotted RNA pseudoknot structures. Hydrogen bonds are represented by a pink colour. Below are physical knots drawn with solid black lines – created by open curves, such as RNA chains. Together with the dashed gray lines, they form closed curves – mathematical knots. The big number in the knot name indicates minimal possible number of crossings of any knot's projections.

knots, links, lassos and theta-curves [12–16]. Experimentally solved protein structures mainly form 3_1 and 4_1 knots, however proteins with knots $5_2, 6_1$ and $3_1\#3_1$ [17] have also been identified. The universe of topology can be even more complicated when proteins with predicted structures are included [18].

In the case of proteins, knots occur in all branches of the tree of life, including proteins with different biological functions [19]. The general role of the knot is unknown, however often the knotted region coincides with the active site of proteins [20]. In some cases, knot constrains have been shown to play a role in binding ligands [21,22], or increase mechanical stability [23,24].

Ribonucleic acid (RNA) molecules play an informative, structural, and metabolic roles in all living cells [25]. RNA molecules are capable of folding into complex secondary and tertiary structures through base pairing and stacking interactions. Typically, the base pairing leads to the formation of double-stranded helices. A structure known as a pseudoknot occurs when at least four distinct regions of an RNA chain participate in a non-nested base pairing arrangement, e.g., the first region pairs with the third, while the second region pairs with the fourth. It is important to emphasize that the term 'pseudoknot' in the context of RNA's secondary structure does not imply that the RNA's main chain forms an actual knot in its tertiary structure (see Fig. 1). Although RNA knots were debated in [26] as early as 1998, to date no authentic RNA knots were found in natural RNA molecules. Several computational explorations of experimentally determined RNA structures have only been able to identify a handful of knotted RNA molecules, all of which were ribosomal RNAs (rRNAs) with length of around 3000 nucleotides [27]. These potential knots, found in structures with RCSB PDB identifiers **3J5L** [28], **4BW0** [29], and **4V9D** [30], have been argued to be artifacts of the modeling procedure, owing to the lack of knots in closely related high-resolution structures. Possible explanations for why knots are not observed in RNA have been postulated in [27,31].

Artificial knotted RNAs can be made in the laboratory. For example, a synthetic RNA has been designed so that it can adopt two different topological states (a circle and a trefoil knot) when ligated into a cyclic molecule and it was determined that the interconversion of these two species can be catalyzed by Escherichia coli DNA topoisomerase III, indicating that this enzyme can act as an RNA topoisomerase [32]. This suggests that there is no fundamental reason why relatively short knotted RNAs could not exist in nature. Micheletti and colleagues further listed potential candidates naturally existing RNAs, opening a path toward the systematic search for RNA knots [31]. In 2021, another type of entanglement was observed in RNA, in the genome of Zika virus [33]. By *knot* the authors meant here a *knot-like* structure formed by a pseudoknot. Such structure is also called a *lasso* which was identified firstly in proteins [15]. Lasso is a loop closed by a bond connecting distant residues and crossed by another part of the chain. Lassos can also be knots, but most often they are not - and in the case of the Zika virus, the knot-like structure is not a true knot. Recently, it was reported that among the RNA chains present in the PDB, as many as 18% form lassos and 0.5% form interlaces [34,35]. Authors suggested that most of these entanglements may be artifacts, but each one should be looked at separately. Lassos are similarly frequent in proteins [15,36]. The function of this geometrical shape in proteins has already been studied and suggested important to perform a biological function [2,3], or lead to protein misfolding [3].

In this paper, we present a discovery of the first true knotted RNA structure. It is a trefoil knot 3_1 in the small RNA (sRNA) RydC (chain **4V2S Q**), whose 3D structure was determined in complex with the Hfq protein, solved at the resolution of 3.48 Å (see Fig. 2) [37]. The RydC RNA was known to form an H-type pseudoknot at the level of the

secondary structure, but the knot in terms of the topology of the RNA chain remained unnoticed. Based on extensive molecular dynamics simulations with structure based model, we present a possible RydC knotting pathway. Like many other bacterial sRNAs, RydC modulates gene expression. It was suggested that its stability in vivo was due to the pseudoknot structure and interaction with the RNA chaperone Hfq [37]. The discovery of a knotted RNA allows us to raise the question of whether and how this type of topology affects physical properties (e.g. stability) and biological function. We discuss the possibility of existence of knots in the structures of other members of the RydC family and in other families with similar size and secondary structures. So far, no structures were determined for other RNAs related to RydC, which we postulate as an interesting and important goal for the near future.

Results

Knotted RNA chain found in the sRNA RydC

We considered all currently available high- and medium-resolution RCSB PDB entries that include RNA chains (see Materials and Methods). The RNA backbone was closed using the stochastic approach [13] and we applied topological invariants to recognize the knot types (see Materials and Methods).

We only found one plausible knot in the RNA, which is most likely a real feature of the RNA structure, and not an artifact of a modeling process. A left-handed trefoil knot -3_1 (see Fig. 2) is present in the small regulatory RNA RydC from *Salmonella enterica* subsp. *enterica* serovar Typhimurium, whose structure was determined by X-ray crystallography in complex with the Hfq protein hexamer (RCSB PDB ID **4V2S**) at a resolution of 3.48 Å [37]. A search of the RCSB database with RydC as a query revealed no overall similarities to other experimentally determined RNA 3D structures.

RydC is a bacterial non-coding RNA. RydC is thought to regulate an mRNA, yejABEF, which encodes an ABC transporter protein. RydC binds the Hfq protein, which causes a conformational change in the RNA molecule. The Hfq/RydC complex is then thought to bind to the target mRNA and induce its degradation.

The RydC sequence (chain Q in the crystal structure **4V2S**) consists of 65 nucleotide residues (nts) and its five 5′-terminal nts are not modeled. Nts 20–21 and a part of nt 19 are disordered in the crystal structure, and their atoms are not included in the model, which creates an apparent gap. That unmodeled part of the chain cannot wrap around a neighboring chain part to remove the knot, even if the missing part of the chain is built in a stretched conformation that maintains physically realistic bond lengths. Therefore, the unmodeled short chain segment in the crystal structure does not affect the RydC RNA chain topology (see Fig. 2). We rebuilt the missing three nts 19–21 (CUC) using Maestro program (a part of the Schrodinger package) [38]. The structure with such a reconstructed gap was used in all further analyzes and simulations.

In the crystal structure, the RNA bridges two Hfq hexamers, with the 3′-end of the RNA (nts U61, U62, C63, and U64) interacting with the proximal

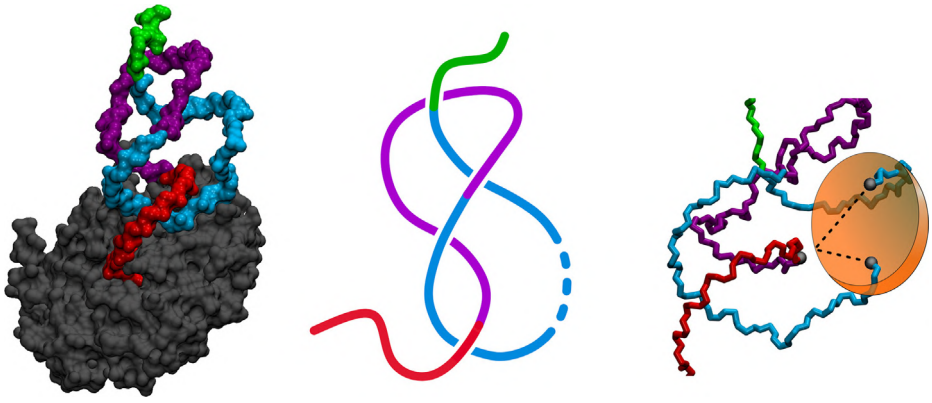


Fig. 2. Schematic representation of the knot in RydC RNA. Left panel: Crystal structure (RCSB PDB ID: **4V2S**) of RydC RNA (chain Q, only the backbone shown, in colors) in complex with the Hfq protein (chains A, B, C, D, E, F, CPK representation, in gray). Middle panel: The knot core of 3_1 knot is formed by nts 11–50 colored in purple and blue: nts 33–50 form a loop (in purple) around which the 5′ end (in blue) wraps and pierces. The tails (in green and red) are of a length 6 and 15 nts, respectively. Right panel: Missing atomic coordinates between atoms O5′19 and P22. The missing atoms could not be modeled in a way that unties the knot, because the rebuilt fragment would have to go around the loop. An orange ellipsoid marks the area beyond which the reconstructed RNA chain fragment cannot go, even assuming that the straightened chain from atom O5′19 to atom P22 (of length 21.1 Å) can be freely arranged in space.

face of the principal hexamer, and the 5′-end (nts G8, U9, A10, and G11) interacting with the lateral surface of another hexamer in the asymmetric unit. The knot core (nts 11–50 that form a knot) is located in between these regions of RNA–protein interactions. Biophysical analyses suggested that the 1:2 complex of RydC with two Hfq hexamers is unlikely to represent the biologically relevant species and that contacts made in the crystal by RydC with the neighbouring Hfq hexamers can form on the proximal face of the same, single hexamer in the 1:1 complex. In particular, the U-rich 3′-end of RydC was proposed to bind in a recessed channel on the proximal face of Hfq, with that part of the RNA chain being flexible, possibly with intrinsic disorder or multiple conformations of the polyU tail [37].

KNOT or NOT? Homologs of 4V2S_Q

The number of non-coding RNA sequences is approximately 34 million based on the RNAcentral database [39], while there are only about seven thousand structures with polymer entity type RNA in RCSB PDB database [40], indicating that the 3D structure space of RNA molecules is largely unexplored [41]. This may suggest that on the one hand, there may be more knotted RNA molecules among those with unknown 3D structures, but on the other hand, if only one knotted RNA is found among those with known structures, it could be a sole exception or an experimental artifact.

We postulated that closely related protein molecules retain the knot in the evolution [24], and the same principle could apply to RNA molecules as well. According to the Rfam database, RydC is a member of a relatively small family of RNAs (**RF00505**), which comprises 21 members exclusively from *Gammaproteobacteria*, out of which there are only 10 non-identical sequences. Thus far, RydC from *Salmonella* is the only family member characterized experimentally. We searched the RNAcentral database, and were able to find only eight further sRNAs with non-identical sequences that exhibit sequence and structure conservation with RydC. This gives us 18 different sequences of RydC homologs in total.

The search also identified other RNA sequences with some similarity to the knotted core of RydC (sequence identity down to $\geqslant 60\%$), but further analyses of these sequences revealed that they all lack the typical H-type pseudoknotted secondary structure characteristic for the RydC family, and they are sequence fragments found in the middle of eukaryotic long non-coding RNAs, hence they are not homologous to RydC. The sequences are summarized in the Table 1 (for more details see Supplementary Table 2). We predicted (without known sRNA) 3D models for all of the identified sequences similar to RydC using the VFold [42] and checked if they form knots. In order to make the predictions unbiased, we excluded the RydC structure (PDB ID 4v2s) while using VFold to predict the 3D structures. VFold first predicts 2D structures using Vfold2D, and based on this prediction generates 3D structures using Vfold3D/VfoldLA methods based on the fragment assembly, using A-form helices with loop and motif templates, extracted from the known RNA 3D structures. It proposes 10 spatial models for a given sequence.

We found that all non-redundant 18 sequences of RydC homologs (Fig. 3) had at least 40% Vfold-predicted 3D structures with a knot (Supplementary Fig. 4). On the other hand, fragments of RNAs from various eukaryotes, with relatively low sequence similarity to RydC and with no similarity of predicted secondary structure, were either never folded by Vfold into knotted structures, or knotted models appeared very rarely (e.g., one model out of 10). This suggests that the knotted topology observed in the RydC RNA from *Salmonella enterica* subsp. enterica serovar Typhimurium is a conserved feature of all sequences in the RydC family, while remote sequence similarity to RydC is an insufficient feature to enable the knot formation. We carried out additional simulations with SimRNA, which agreed with VFold predictions (see Supplementary Information).

Burton et al. [31] analyzed RNA molecules from the Pseudoknot database and proposed that 12 of them could be forming actual knots. Since their analysis in 2015, for three of these RNAs (*H. sapiens* hTER, *S. cerevisiae* Sc_18S-PKE21-7/8, and bacteriophage Q-β PKIdX) 3D structures were determined experimentally, demonstrating the absence of knots. We analyzed the remaining nine candidates for knotted RNAs with so far unknown 3D structures using the VFold, and found presence of knotted conformation in six candidates. Out of the six candidates, only the plasmid ColIb-P9 (ColIb-P9_repZ) formed knot in all the predicted models. For the rest of the cases, less than 50% of the predicted models have shown knotted conformation, of which VFold was unable to predict any 3D model for two of the candidates (Supplementary Table 1).

Protein-style RNA knot tying

The problem of RNA folding can be understood in two ways, one approach focuses on predicting one or more secondary RNA structures from a given base sequence bypassing the kinetic pathway (gluing different pieces of RNA together). From a topological point of view, there are no constraints in predicting a knot here. In the second, when we talk about folding, the exact movement an RNA chain must make to form its native conformation is studied. In this case, the RNA chain that is folding starts with a random configuration and remains cohesive all the time. From this perspective, an intriguing question arises as to how this sRNA chain will tie itself into a knot.

Table 1 Frequency of knots detected in 3D models generated by the Vfold [42] for sequences of RydC homologs (in bacteria and in *Trichuris trichiura*) and for remotely similar (most likely non–homologous) fragments of longer RNA sequences in eukaryotes. The table presents results for 30 non-identical sequences or fragments of longer RNA sequences that exhibited >= 60% identity to RydC. These fragments are grouped in rows according to the frequency of knotted structures among their predicted models (stated in the first column). The second column indicates the number of fragments in the group. The third column presents a range of identities between the sequence of **4V2S_Q** and sequences in the group. The last column lists the organisms in which full RNA sequences containing fragments from that group were found, along with the number of different sequences found (one sequence if no number is given in the brackets). There can be many RNA sequences containing the same short fragment similar to **4V2S_Q** and therefore the sequence numbers in the 4th column do not add up to the number in the 2nd column.

Knotted models	# Similar fragments	Identity [%]	Organisms (# sequences)
90% – 100%	3	70.97 – 90.48	*Citrobacter koseri* (2), *Enterobacter aerogenes, Klebsiella aerogenes, Klebsiella oxytoca*
50% – 80%	4	67.74 – 86.67	*Escherichia coli* (129), *Escherichia fergusonii, Klebsiella pneumoniae* (62), *Shigella boydii, Shigella dysenteriae, Shigella sonnei, Synthetic Escherichia, Trichuris trichiura*
20% – 40%	11	69.35 – 98.41	*Citrobacter youngae, Danio rerio, Klebsiella sp.* (3), *Klebsiella oxytoca* (2), *Salmonella enterica* (4), *Shigella sonnei*
10%	2	66.1 – 71.74	*Homo sapiens* (2), *Nothobranchius furzeri*
0%	10	62.5 – 71.43	*Chlamydomonas reinhardtii, Chrysemys picta, Gallus gallus* (4), *Homo sapiens* (4), *Hordeum vulgare, Pongo abelii, Taeniopygia guttata, Triticum aestivum* (4)
Total	30	62.5–98.41	

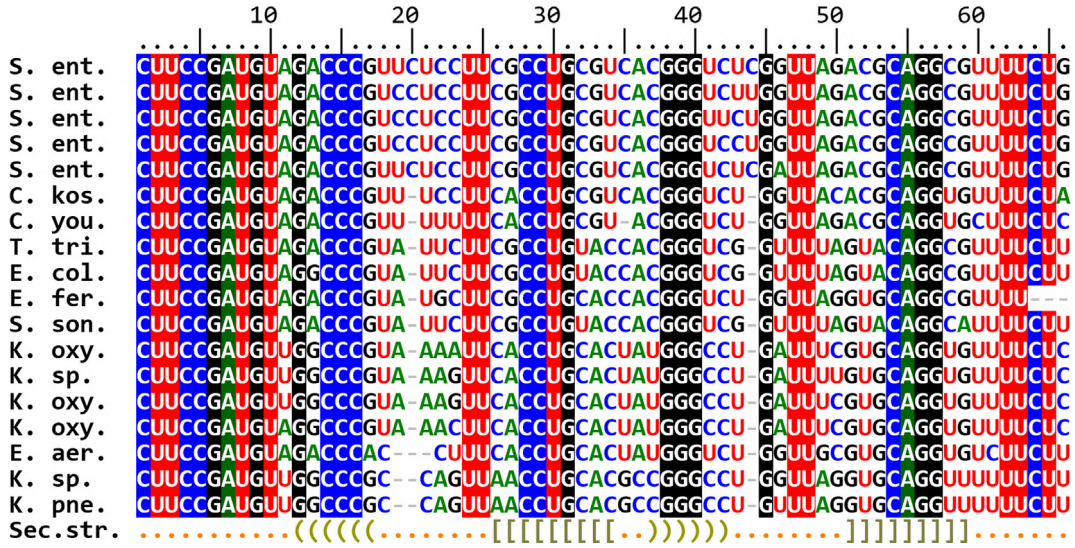


Fig. 3. RydC family alignment. Residues invariant in all family members are highlighted. The bottom row indicates the consensus secondary structure in the dot and bracket format.

To understand potential folding and knotting pathways we conducted extensive molecular dynamics simulations based on a coarse grained structure based model (SBM) in the temperature just below estimated T_f (see Materials and Methods).

We found that it is possible to observe self-tying of investigated sRNA, more specifically in 73% of trajectories the structure folded to its native form, with a knot 3_1. In all trajectories, we observed that the structure instantly formed a loop in its native position (see Materials and Methods) and the loop remained formed during the entire folding pathway. Then in most cases the knot 3_1 was tied and native form achieved. However, during the folding route incorrect steps may occur and in several percent of the trajectories the structure tied a knot 4_1 (see Fig. 4), which was a kinetic trap. Escaping from such kinetic traps requires a backtracking mechanism, which was observed in the case of knotted proteins [43–45]. The average times that structure remained in different intermediate states are shown in Table 2. The structure exits the kinetic trap with the knot 4_1 six times faster than

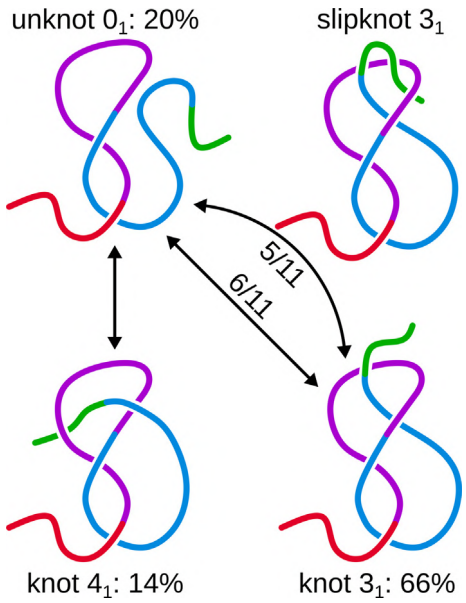


Fig. 4. Schematic representation of different intermediate states of **4V2S_Q** folding. Top-left: loop is formed in all simulations and in 20% of them structure remains in this state. Bottom-left: kinetic trap with a knot 4_1. Bottom-right: native state with a knot 3_1, which can be reached directly from an unknotted state (in 36% simulations) or via slipknot (top-right; in 30% simulations). Configurations with knots 3_1 and 4_1 differ in whether the tail (in green) goes through the loop (in purple) from below or from above.

it unties the knot 3_1 that is present in its native form. To describe kinetics of folding routs in details we use number of created native contacts (Q), which are usually used in the case of proteins. The average value of Q in topological states of $0_1, 3_1$ and 4_1 was 42%, 76% and 66%, respectively (see Fig. 5).

The fact that the loop is formed immediately in its native position, long before the 3′ end of the chain is pulled through it, suggests that the RydC RNA chain folds into a knot in *protein-style* [45]. In contrast, in random polymers, the tail often passes through a large-sized loop, and only later does the loop tighten. We found that in 97% of *knotting events* (see Materials and Methods), when structure forms a knot 3_1, it does it in protein-style.

Moreover, in about half of the folding pathways that lead to the knotted conformation in our simulations we observed an intermediate configuration containing a slipknot (like in the case of proteins [46,47]). This means that when tying the knot first the tail arm was pulled through the loop (see Fig. 4), not just the end of the tail. In the case of RydC RNA, the part of the tail passing through the loop is not very long, consisting of 6 nts. Thus both folding pathways, with or without the slipknot, are similar to each other. This is probably why they are similarly frequent (45% and 55% knotting events, respectively).

Table 2 The average times along all trajectories that structure remained continuously in different intermediate states and in native state.

	Average (continuous) time [τ]
till loop is formed	362
to form a knot after a loop is formed *(if at all)*	17,988
with a knot 4_1	3041
with a knot 3_1	18,806

Discussion and conclusions

So far, there has been no consensus in the scientific literature regarding the existence of true topological knots in natural RNAs. Previous studies suggested that there are no knotted RNAs among those with experimentally determined structures in the RCSB PDB database. The only cases reported in the literature are likely artifacts of modeling. While candidates for knotted RNAs have been proposed, these RNAs have yet to be experimentally confirmed to exhibit knots. Meanwhile, artificially knotted RNAs have been successfully designed and synthesized. Based on the current evidence, it appears that true knots are exceedingly rare in natural RNAs.

In our study, we discovered a genuine knot in the crystal structure of RydC RNA bound to the Hfq protein. To the best of our knowledge, this is the first instance of a true knot found in an experimentally determined structure involving an RNA molecule.

Although the resolution of the crystal structure is moderate and the quality of the RNA model is not perfect (based on [48]), the chain topology is resolved and exhibits a knot beyond any reasonable doubt. However, it will be of great value to perform further experiments to solve structures for other sequences, for which VFold [42] predicts knotted structures, as well as for the pseudoknotted RNAs hypothesized to form a knot [31]. These experiments may change our thinking about the possible geometry and topology of the 3D structures of millions of RNA molecules with unknown structures. One may ask why the crystal structure has adopted the typical knot type? The specific type of the knot is mostly dictated by the short sequence and the relatively simple secondary structure of RydC, which provides just two loops, through which the ends of the chain can thread. According to our computational analyses, the RydC sequence and the sequences of its homologs have an intrinsic tendency to hold into a true knot, so we believe it is not a direct effect of the crystal structure or the formation of the complex with the Hfq protein. The

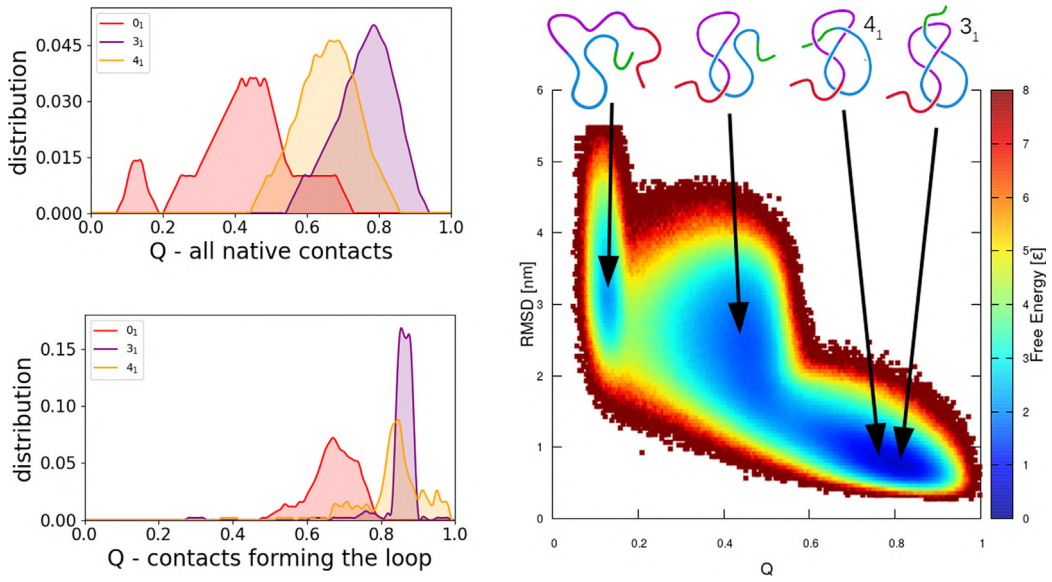


Fig. 5. Intermediate states in folding. Left panel: distribution of Q values in three topological states in all trajectories; at the top all native contacts, at the bottom only those forming the loop. Right panel: plot of free energy vs. Q and RMSD, where you can see three intermediate states in the folding process: I. unfolded, II. unknotted with the loop at its position, III. knotted. The plot does not show separate states for 3_1 and 4_1 - in both states a lot of native contacts are created and these states merge into one on the plot. To go from knot 4_1 to knot 3_1, structure has to go through state II.

ability of the RydC sequence alone (without Hfq) to form a knot in solution will need to be validated experimentally, which is beyond the scope of this study.

The rarity of knots in RNA 3D structures may indicate evolutionary or functional constraints. It is possible that the main driving force of RNA structure formation, namely base-pairing, can be the main obstacle for the process of knot formation, for instance threading of a long single-stranded (unpaired) stretch or RNA through a loop that needs to remain unpaired as well. Unraveling the knots in RNAs and the mechanism of their formation will necessitate both advances in experimental techniques to resolve complex RNA structures and theoretical breakthroughs to predict and understand the potential functional implications of knots.

Our computational analyses suggest that RydC folds in a protein-style, not like a homopolymer. During folding into a native form the chain first creates a stable loop required for a knot formation.

Following AlphaFold's success in protein structure prediction [49], tools for 3D RNA structure prediction are also being developed rapidly, with around 18 tools available, 8 have been published after 2020 [41]. There are already tools that take into account the protein or RNA topology when analyzing modeled 3D structures and some modeling strategies include rejection of models that exhibit knots [18,50,35]. The discovery of a true knot in the RydC RNA is an important result that should be taken into account in the development of RNA modeling tools.

Materials and Methods

The data set. The list of RNA representative chains was obtained from the 3.269 version of the BGSU RNA database (released on 2023–0208). Only structures with resolution of 4 Å or better were chosen – that gave 2580 equivalence classes and 3202 representative chains. We added to this set 13 chains formed by joining pairs of chains in one molecule, the ends of which are no more than 4 Å apart. This gives a total dataset size of 3,215 RNA chains. Three RNA chains discussed to be knotted or unknotted in [27], i.e. **2GYA_0**, **1C2W_B**, **3JYX_5**, were not included in our dataset due to their low resolution (15 Å, 7.5 Å and 8.9 Å, respectively). Everywhere in the text of this article and in the Supplement, the symbol after the underscore (e.g. 0 in **2GYA_0**) denotes the name of the chain, i.e., the value from *auth_asym_id* column of the CIF file.

Identification of knots. To detect a knot in an RNA chain we analyze its sugar-phosphate backbone (specifically the atoms P, O5′, C5′, C4′, C3′, O3′). The gaps are connected by a straight line. We use Alexander and Homfly-PT polynomials to identify the exact knot types, the last one to recognize the chirality of a knot. Both polynomials can be applied only to closed chains (rings) and have high computational complexity, thus we first bridge the termini of the RNA chain and simplify the ring geometry at fixed topology [51]. We use *probabilistic closure*, randomly closing the chain multiple times (e.g. 100) and pulling out the statistics [13]. We call a chain *knotted* if a

specific non-trivial knot occurs most frequently, more than 48% of the time. We applied this algorithm, implemented in [52], and among 3,215 RNA chains we found just one knotted structure, **4V2S_Q**. Left-handed trefoil knot -3_1 was formed 65% times among 1000 random closures, while trivial knot 28% times. The knot core (minimal portion of a chain that is knotted into 3_1) was located with a bottom-up approach [53]. In the **4V2S_Q** chain the 3_1 knot core is formed by nts 11–50. The remaining parts of the chain, so-called *tails*, are of length 5 nts and 15 nts. In general we use Topoly [52] package to detect knots.

Structure Based Model for simulations. We constructed and used a coarse grained structure based model (SBM) in which each nucleotide is represented by two interaction sites (beads): (i) one mimicking ribose-phosphate group (these beads act as a backbone), and (ii) another bead representing the base. We introduce native interactions between all beads which do not interact via local interactions, by considering all pairs of beads which are within the distance of 11 Å in the native (crystal) structure. These native non-bonded interactions are mimicked by the Gaussian potential implemented in [54].

Along the backbone we introduce bond-length, bond-angle, and dihedral angle potentials. Beads which do not interact via backbone potentials nor they form native contacts, interact with the repulsive part of the Lennard-Jones.

We validated the model by comparing it with all-atom representation approach [55] on three RNA structures [56–58]. Our model captures key features of the studied systems observed with all-atom representation approach (see Supplement).

All simulations were conducted using Gromacs v4.5.4 with the Gaussian potential. A leap-frog stochastic dynamics integrator with an inverse friction constant of 1.0 was used. The time step was equal to 0.0005τ.

We run 515 simulations of $50,000\,\tau$ (i.e. 100 million steps, 40,000 frames), at reduced temperature $T^* = 1.14$ (137 in Gromacs units), which corresponds to the estimated folding temperature T_f.

Identification and analysis of knotting events in trajectories. While searching for the knots along trajectories we were closing chains in a deterministic way, using method called hereafter *out of the mass center* [59]. By this we mean extending the termini in the direction from the center of mass and closing on a large sphere. This was due to computational costs. However, both closing methods (out of the mass center and probabilistic approach) agree in the vast majority of cases [60].

To identify moments in the trajectory where the chain *stably* forms a knot, we looked for series of 50 consecutive frames with a knotted chain following series of 50 frames with an unknotted chain. We checked the knot every 5th frame to speed up the calculations. In total, we found 639 knotting events in 515 trajectories. Among them 528 correspond to knot 3_1, 109 to knot 4_1 and two to knot 5_2.

Next, to characterize geometrically how a knot is tied, we counted the knot core for 10 frames from the moment of knotting, and looked at how it changes. In its native form, the knot core consists residues 11–50. The tail at the loop (at 3′ end) is of 15 nts, and the tail passing through the loop (at 5′ end) is of 5 nts (see Fig. 2). If the tail at the loop was at least 10 nts long already in the first knotted frames (so that the loop did not move significantly after the knot was formed), we considered the knot to be tied in protein-style [45]. We found that this was the case in 97% of knotting events when structure folds into 3_1 knot. Finally, if the tail passing through the loop was at least 3 nts long already in the first knotted frames, we considered the knot to be tied via slipknot. This was the case in 45% of knotting events.

Analysis of loop formation in trajectories. We considered a loop to be created when at least 90% of native contacts between nts 30–34 and 49–53 were formed. There are 33 contacts between those nucleotides (17% of all native contacts), where by contact we mean a pair of nucleotides with a distance of less than 11 Å (see Supplement).

Note on choosing sugar-phosphate backbone. It is worth noting that the choice of atoms on which to base the simplified representation of the RNA chain may affect the type of knot that we find in that RNA chain. For example, in the **7NAF_1** chain, the backbone based only on phosphates forms a knot 4_1 (when closed out of mass center). However, the backbone based on C3' atoms forms a knot 3_1, and backbones based separately on C4′, C5′, O3′, or O5′, or on all the atoms P, O5′, C5′, C4′, C3′, O3′ - are unknotted. A similar situation occurs for 17 out of 3,215 RNA chains, which is about 0.5% of cases. Therefore, we decided on the least coarse-grained model, i.e. on a backbone composed of all the atoms P, O5′, C5′, C4′, C3′, O3′, when searching the set of RNA chains for knots.

Accession numbers. RCSB PDB ID: 1C2W, 2GYA, 2K95, 3J5L, 3JYX, 4BW0, 4U6F, 4V2S, 4V9D, 6WLN, 6WLO, 6ZQG, 7LHD, 7NAF, 7TRF, 7ZUX, 8CBJ. RFAM ID: RF00140, RF00165, RF00505, RF01087, RF01089, RF01090.

Funding

This work was supported by Warsaw University [BOB-IDUB-622–90/2021 to W.N]; COST EUTOPIA action to W.N., B.A.G., J.I.S.; the National Science Centre [#UMO-2018/31/B/NZ1/04016, #UMO-2021/43/I/NZ1/03341, to J.I.S., 2017/26/A/NZ1/01083 to J.M.B., and 2021/43/D/NZ1/03360 to S.M.]. We gratefully acknowledge

Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH, CI TASK, WCSS) for providing computer facilities and support within computational grant no. PLG/2022/016043.

Declaration of Generative AI and AI-assisted technologies in the writing process

Disclaimer: The authors of this paper have utilized the ChatGPT language model solely for the purpose of improving the quality of the English language. At no point was ChatGPT used to modify, alter, or interpret the data presented. All scientific conclusions, data analyses, and interpretations remain the sole responsibility of the authors.

DATA AVAILABILITY

Data will be made available on request.

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.

Acknowledgments

We thank Maciej Antczak for fruitful discussions. We thank Roksana Malinowska and Mai Lan Nguyen for help with running Maestro.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jmb.2024.168455.

Received 17 October 2023;
Accepted 18 January 2024;
Available online 23 January 2024

Keywords:
Entanglement of biomolecules;
Topology in soft matter;
RNA 3D structure;
RNA folding;
Molecular dynamics

References

[1]. Jobe, A., Liu, Z., Gutierrez-Vargas, C., Frank, J., (2019). New insights into ribosome structure and function. *Cold Spring Harb. Perspect. Biol.* **11** (1), a032615.

[2]. Dabrowski-Tumanski, P., Gren, B., Sulkowska, J.I., (2019). Statistical properties of lasso-shape polymers and their implications for complex lasso proteins function. *Polymers* **11** (4), 707.

[3]. Q.V. Vu, Y. Jiang, I. Sitarik, M.S. Li, E.P. O'Brien, A new class of protein misfolding is observed in all-atom folding simulations, bioRxiv, 2022.

[4]. Tezuka, Y., Oike, H., (2001). Topological polymer chemistry: systematic classification of nonlinear polymer topologies. *J. Am. Chem. Soc.* **123** (47), 11570–11576.

[5]. Niewieczerzał, S., Sulkowska, J.I., (2019). Supercoiling in a Protein Increases its Stability. *Phys. Rev. Lett.* **123** (13), 138102.

[6]. Sumners, D.W., Whittington, S., (1988). Knots in self-avoiding walks. *J. Phys. A: Math. Gen.* **21** (7), 1689–1694.

[7]. Micheletti, C., Marenduzzo, D., Orlandini, E., (2011). Polymers with spatial or topological constraints: Theoretical and computational results. *Phys. Rep.* **504** (1), 1–73.

[8]. Sumners, D.W., (1995). Lifting the curtain: Using topology to probe the hidden action of enzymes. *Not. Am. Math. Soc.* **42** (5), 528–537.

[9]. Wasserman, S.A., Dungan, J.M., Cozzarelli, N.R., (1985). Discovery of a predicted DNA knot substantiates a model for site-specific recombination. *Science* **229** (4709), 171–174.

[10]. Bates, A., Maxwell, A., (2005). DNA Topology. Oxford University Press, Oxford bioscience.

[11]. C. Micheletti, "DNA Knots," Tezuka, Y., Deguchi, T. (eds) Topological Polymer Chemistry, 2022.

[12]. Taylor, W.R., (2000). A deeply knotted protein structure and how it might fold. *Nature* **406** (6798), 916–919.

[13]. Jamroz, M., Niemyska, W., Rawdon, E.J., Stasiak, A., Millett, K.C., Sułkowski, P., Sulkowska, J.I., (2015). KnotProt: a database of proteins with knots and slipknots. *Nucleic Acids Res.* **43** (D1), D306–D314.

[14]. P. Dabrowski-Tumanski, A.I. Jarmolinska, W. Niemyska, E. J. Rawdon, K.C. Millett and J.I. Sulkowska (2017). LinkProt: A database collecting information about biological links, Nucleic Acids Res., 45 (D1), 2017, D243–D249.

[15]. Niemyska, W., Dabrowski-Tumanski, P., Kadlof, M., Haglund, E., Sułkowski, P., Sulkowska, J.I., (2016). Complex lasso: new entangled motifs in proteins. *Sci. Rep.* **6**, 36895.

[16]. Sulkowska, J.I., (2020). On folding of entangled proteins: knots, lassos, links and theta-curves. *Curr. Opin. Struct. Biol.* **60**, 131–141.

[17]. F. Bruno da Silva, I. Lewandowska, A. Kluza, S. Niewieczerzal, R. Augustyniak, and J. Sulkowska, "First crystal structure of double knotted protein TrmD-Tm1570 - inside from degradation perspective," BioRxiv, 2023.

[18]. Niemyska, W., Rubach, P., Gren, B.A., Nguyen, M.L., Garstka, W., Bruno da Silva, F., Rawdon, E.J., Sulkowska, J.I., (2022). AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods. *Nucleic Acids Res.* **50** (W1), W44–W50.

[19]. Jarmolinska, A.I., Perlinska, A.P., Runkel, R., Trefz, B., Ginn, H.M., Virnau, P., Sulkowska, J.I., (2019). Proteins' knotty problems. *J. Mol. Biol.* **431** (2), 244–257.

[20]. Perlinska, A.P., Stasiulewicz, A., Nawrocka, E.K., Kazimierczuk, K., Setny, P., Sulkowska, J.I., (2020). Restriction of S-adenosylmethionine conformational freedom by knotted protein binding sites. *PLoS Comput. Biol.* **16** (5), e1007904.

[21]. Perlinska, A.P., Kalek, M., Christian, T., Hou, Y.-M., Sulkowska, J.I., (2020). Mg2+-dependent methyl transfer by a knotted protein: A molecular dynamics simulation and quantum mechanics study. *ACS Catal.* **10** (15), 8058–8068.
[22]. Christian, T., Sakaguchi, R., Perlinska, A.P., Lahoud, G., Ito, T., Taylor, E.A., Yokoyama, S., Sulkowska, J.I., Hou, Y.-M., (2016). Methyl transfer by substrate signaling from a knotted protein fold. *Nat. Struct. Mol. Biol.* **23** (10), 941–948.
[23]. Virnau, P., Mirny, L.A., Kardar, M., (2006). Intricate knots in proteins: Function and evolution. *PLoS Comput. Biol.* **2** (9), e122.
[24]. Sułkowska, J.I., Rawdon, E.J., Millett, K.C., Onuchic, J. N., Stasiak, A., (2012). Conservation of complex knotting and slipknotting patterns in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **109** (26), E1715–E1723.
[25]. Lewis, M.W., Verma, A., Eckdahl, T.T., (2021). A new modeling paradigm for the RNA folding problem. *J. Heuristics* **27**, 695–717.
[26]. VanLoock, M., Harris, B., Harvey, S., (1998). To knot or not to knot? Examination of 16S ribosomal RNA models. *J. Biomol. Struct. Dyn.* **16** (3), 709–713.
[27]. Micheletti, C., Di Stefano, M., Orland, H., (2015). Absence of knots in known RNA structures. *Proc. Natl. Acad. Sci. U.S.A.* **112** (7), 2052–2057.
[28]. Arenz, S., Ramu, H., Gupta, P., Berninghausen, O., Beckmann, R., Vázquez-Laslop, N., Mankin, A.S., Wilson, D.N., (2014). Molecular basis for erythromycin-dependent ribosome stalling during translation of the ErmBL leader peptide. *Nat. Commun.* **5** (1), 3501.
[29]. Huang, L., Lilley, D.M., (2013). The molecular recognition of kink-turn structure by the L7Ae class of proteins. *RNA* **19** (12), 1703–1710.
[30]. Dunkle, J.A., Wang, L., Feldman, M.B., Pulk, A., Chen, V. B., Kapral, G.J., Noeske, J., Richardson, J.S., Blanchard, S.C., Cate, J.H.D., (2011). Structures of the Bacterial Ribosome in Classical and Hybrid States of tRNA Binding. *Science* **332** (6032), 981–984.
[31]. Burton, A.S., Di Stefano, M., Lehman, N., Orland, H., Micheletti, C., (2016). The elusive quest for RNA knots. *RNA Biol.* **13** (2), 134–139.
[32]. Wang, H., Di Gate, R.J., Seeman, N.C., (1996). An RNA topoisomerase. *Proc. Natl. Acad. Sci. U.S.A.* **93** (18), 9477–9482.
[33]. Zhao, M., Woodside, M., (2021). Mechanical strength of RNA knot in Zika virus protects against cellular defenses. *Nat. Chem. Biol.* **17**, 975–981.
[34]. Popenda, M., Zok, T., Sarzynska, J., Korpeta, A., Adamiak, R.W., Antczak, M., Szachniuk, M., (2021). Entanglements of structure elements revealed in RNA 3D models. *Nucleic Acids Res.* **49** (17), 9625–9632.
[35]. Luwanski, K., Hlushchenko, V., Popenda, M., Zok, T., Sarzynska, J., Martsich, D., Szachniuk, M., Antczak, M., (2022). RNAspider: a webserver to analyze entanglements in RNA 3D structures. *Nucleic Acids Res.* **50** (W1), W663–W669.
[36]. Dabrowski-Tumanski, P., Niemyska, W., Pasznik, P., Sulkowska, J.I., (2016). LassoProt: server to analyze biopolymers with lassos. *Nucleic Acids Res.* **44** (W1), W383–W389.
[37]. Dimastrogiovanni, D., Fröhlich, K.S., Bandyra, K.J., Bruce, H.A., Hohensee, S., Vogel, J., Luisi, B.F., (2014). "*Recognition of the small regulatory RNA RydC by the bacterial Hfq protein,*" *eLife* **vol. 3**, e05375.
[38]. S. Maestro, "Maestro," Schrödinger, LLC, New York, NY, vol. 3, 2023.
[39]. The RNAcentral Consortium, "RNAcentral: a hub of information for non-coding RNA sequences," Nucleic Acids Res., vol. 47, no. D1, pp. D221–D229, 2018.
[40]. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, B.P.E., "The Protein Data Bank," Nucleic Acids Res., vol. 28, pp. 235–242, 2020.
[41]. Schneider, B., Sweeney, B.A., Bateman, A., Cerny, J., Zok, T., Szachniuk, M., (2023). When will RNA get its AlphaFold moment? *Nucleic Acids Res.* **51** (18), 9522–9532.
[42]. Li, J., Zhang, S., Zhang, D., Chen, S.-J., (2022). Vfold-Pipeline: a web server for RNA 3D structure prediction from sequences. *Bioinformatics* **38** (16), 4042–4043.
[43]. Norcross, T.S., Yeates, T., (2006). A framework for describing topological frustration in models of protein folding. *J. Mol. Biol.* **362**, 605–621.
[44]. Gosavi, S., Chavez, L.L., Jennings, P.A., Onuchic, J.N., (2006). Topological frustration and the folding of interleukin-1beta. *J. Mol. Biol.* **357**, 986–996.
[45]. Sułkowska, J.I., Sułkowski, P., Onuchic, J., (2009). Dodging the crisis of folding proteins with knots. *Proc. Natl. Acad. Sci. U.S.A.* **106** (9), 3119–3124.
[46]. Wallin, S., Zeldovich, K.B., Shakhnovich, E.I., (2007). The folding mechanics of a knotted protein. *J. Mol. Biol.* **368** (3), 884–893.
[47]. Noel, J.K., Sułkowska, J.I., Onuchic, J.N., (2010). Slipknotting upon native-like loop formation in a trefoil knot protein. *Proc. Natl. Acad. Sci. U.S.A.* **107** (35), 15403–15408.
[48]. Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall III, W.B., Snoeyink, J., Richardson, J.S., Richardson, D.C., (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383.
[49]. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.
[50]. Biesiada, M., Pachulska-Wieczorek, K., Adamiak, R.W., Purzycka, K.J., (2016). RNAComposer and RNA 3D structure prediction for nanotechnology. *Methods* **103**, 120–127.
[51]. Koniaris, K., M, M., (1991). Knottedness in ring polymers. *Phys. Rev. Lett.* **66** (17), 2211–2214.
[52]. Dabrowski-Tumanski, P., Rubach, P., Niemyska, W., Gren, B.A., Sulkowska, J.I., (2020). Topoly: Python package to analyze topology of polymers. *Briefings Bioinf.* **vol. 22, no. 3**, bbaa196.

[53]. Rawdon, E., Millett, K.C., Sulkowska, J.I., Stasiak, A., (2013). Knot localization in proteins. *Biochem. Soc. Trans.* **41**, 538–541.
[54]. Noel, J.K., Whitford, P.C., Sanbonmatsu, K.Y., Onuchic, J.N., (2010). SMOG@ ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res.* **38** (W1), W657–W661.
[55]. G. Lazzeri, C. Micheletti, S. Pasquali, and P. Faccioli, "RNA folding landscapes from explicit solvent all-atom simulations," arXiv preprint arXiv:2205.12603, 2022.
[56]. Dufour, D., de la Pena, M., Gago, S., Flores, R., Gallego, J., (2009). Structure-function analysis of the ribozymes of chrysanthemum chlorotic mottle viroid: a loop-loop interaction motif conserved in most natural hammerheads. *Nucleic Acids Res.* **37** (2), 368–381.
[57]. Nonin-Lecomte, S., Felden, B., Dardel, F., (2006). NMR structure of the aquifex aeolicus tmRNA pseudoknot pk1: new insights into the recoding event of the ribosomal trans-translation. *Nucleic Acids Res.* **34** (6), 1847–1853.
[58]. Kim, N.-K., Zhang, Q., Zhou, J., Theimer, C.A., Peterson, R.D., Feigon, J., (2008). Solution structure and dynamics of the wild-type pseudoknot of human telomerase RNA. *J. Mol. Biol.* **384** (5), 1249–1261.
[59]. Caraglio, M., Micheletti, C., Orlandini, E., (2017). Physical Links: Defining and detecting inter-chain entanglement. *Sci. Rep.* **7**, 1156.
[60]. Virnau, P., (2010). Detection and visualization of physical knots in macromolecules. *Phys. Procedia* **6**, 117–125.

Article 7

Knotted artifacts in predicted 3D RNA structures

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Knotted artifacts in predicted 3D RNA structures

Bartosz A. Gren[1☯], **Maciej Antczak**[2,3☯], **Tomasz Zok**[2], **Joanna I. Sulkowska**[1*], **Marta Szachniuk**[2,3*]

1 Centre of New Technologies, University of Warsaw, Warsaw, Poland, **2** Institute of Computing Science, Poznan University of Technology, Poznan, Poland, **3** Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

☯ These authors contributed equally to this work.
* j.sulkowska@cent.uw.edu.pl (JIS); mszachniuk@cs.put.poznan.pl (MS)

Citation: Gren BA, Antczak M, Zok T, Sulkowska JI, Szachniuk M (2024) Knotted artifacts in predicted 3D RNA structures. PLoS Comput Biol 20(6): e1011959. https://doi.org/10.1371/journal.pcbi.1011959

Editor: Shi-Jie Chen, University of Missouri, UNITED STATES

Received: March 8, 2024

Accepted: June 1, 2024

Published: June 20, 2024

Data Availability Statement: The RNA models predicted within CASP15 are available at https://predictioncenter.org/download_area/CASP15/predictions/RNA/. RNAspider is accessible at https://rnaspider.cs.put.poznan.pl/ Other scripts used to identify and analyse knots and entanglements are available in the GitHub repository https://github.com/ilbsm/CASP15_knotted_artifacts All other data are within the manuscript and its Supporting information files.

Funding: This work was supported by the National Science Centre, Poland (https://www.ncn.gov.pl/

Abstract

Unlike proteins, RNAs deposited in the Protein Data Bank do not contain topological knots. Recently, admittedly, the first trefoil knot and some lasso-type conformations have been found in experimental RNA structures, but these are still exceptional cases. Meanwhile, algorithms predicting 3D RNA models have happened to form knotted structures not so rarely. Interestingly, machine learning-based predictors seem to be more prone to generate knotted RNA folds than traditional methods. A similar situation is observed for the entanglements of structural elements. In this paper, we analyze all models submitted to the CASP15 competition in the 3D RNA structure prediction category. We show what types of topological knots and structure element entanglements appear in the submitted models and highlight what methods are behind the generation of such conformations. We also study the structural aspect of susceptibility to entanglement. We suggest that predictors take care of an evaluation of RNA models to avoid publishing structures with artifacts, such as unusual entanglements, that result from hallucinations of predictive algorithms.

Author summary

- 3D RNA structure prediction contests such as CASP and RNA-Puzzles lack measures for topology-wise evaluation of predicted models. Thus, predictors happen to submit potentially inappropriate conformations, for example, containing entanglements that are prediction artifacts.
- Automated identification of entanglements in 3D RNA structures is computationally hard. Distinguishing correct from incorrectly entangled conformations is not trivial and often requires expert knowledge.
- We analyzed 3D RNA models submitted to CASP15 and found that all entanglements in these models are artifacts.

en) - grants 2018/31/B/NZ1/04016 and 2021/43/I/NZ1/03341 to JIS. MA, TZ, and MS acknowledge the support of the Poznan University of Technology (https://www.put.poznan.pl/en) and the Institute of Bioorganic Chemistry, Polish Academy of Sciences (https://www.ibch.poznan.pl/en.html) - statutory funds. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

- Compared to non-ML, machine learning-based methods are more prone to generating entanglements that are not present in natural RNAs.
- To increase the reliability of 3D RNA structure prediction, it is necessary to reject abnormally entangled structures in the modeling stage.

Introduction

The birth of the third decade of this century brought a sudden surge of interest in modeling 3D RNA structures. The latter was, among other things, a by-product of the COVID-19 pandemic, whose main actor was the RNA virus, and intensive research on the development of RNA-based vaccines against COVID-19. The second major factor was the spectacular success brought about by the application of deep neural networks to model protein structures [1]. As a result, new methods have emerged to predict RNA structures, most of which use machine learning models in an end-to-end approach or at selected stages of the 3D folding process [2]. Many of them have undergone virgin benchmarking while competing in recent RNA-Puzzles and CASP15 initiatives, which aim to blindly evaluate predicted 3D RNA models and identify the best predictive tools. The results of both competitions show that none of the new methods has made a breakthrough in the quality and accuracy of the prediction of the 3D RNA structure so far. The latter are evaluated using various measures of distance (DI, GDT-TS, lDDT, MCQ, RMSD), similarity (INF, TM-score, LCS), and quality (Clash score) [3–8]. No measure can directly assess the accuracy of the 3D model topology and its compatibility with the topology of the target. Consequently, awareness of topological irregularities in 3D RNA predictions is negligible in the RNA community, and the predicted models happen to contain them. These anomalies include entanglements that are absent from known experimental structures. We can consider them locally, taking into account the secondary and tertiary structure of the molecule (we then speak of entanglements of structural elements), and globally, studying the spatial arrangement of the RNA backbone (we then analyze topological knots).

The entanglement of structural elements occurs when two elements of the RNA structure are in spatial conflict, that is, one of them punctures the other to form a lasso, interlace (known as a link in knot theory) [9] (Fig 1A) or genus type (the genus trace represents how interconnected and densely packed the structure is in three dimensions) [10] (Fig 1E). Elements such as loops, stems (consisting of dinucleotide steps), and single-stranded fragments can be involved in entanglements of structure elements (Fig 1B). From the viewpoint of knot theory, lassos are not knots. Interlaces can be interpreted as Hopf links (the simplest type of two interlaced loops) defined on loops traced by the phosphodiester and hydrogen bonds; the first contribute to the formation of the nucleotide chain and the latter to base pairs. The formation of many types of entanglements of structure elements, like some interlaces (loop & dinucleotide step, dinucleotide step & dinucleotide step) and deep lassos, contradicts the RNA folding hierarchy and is therefore hardly possible. We do not find them in high-resolution experimentally determined RNAs [11]. They occur in structures modeled *in silico*, in which case they should be regarded as artifacts of computational procedures.

The formation of topological knots in RNA structures has been studied by a few research teams so far. The authors of [12], although they came across knotted 16S rRNA domains, argued about the low likelihood of topological knots in large native RNAs and suggested examining experimental RNA models for knotting before their publication. In this spirit, authors of [13, 14] created a method to assemble viral RNA genomes that avoided the creation of

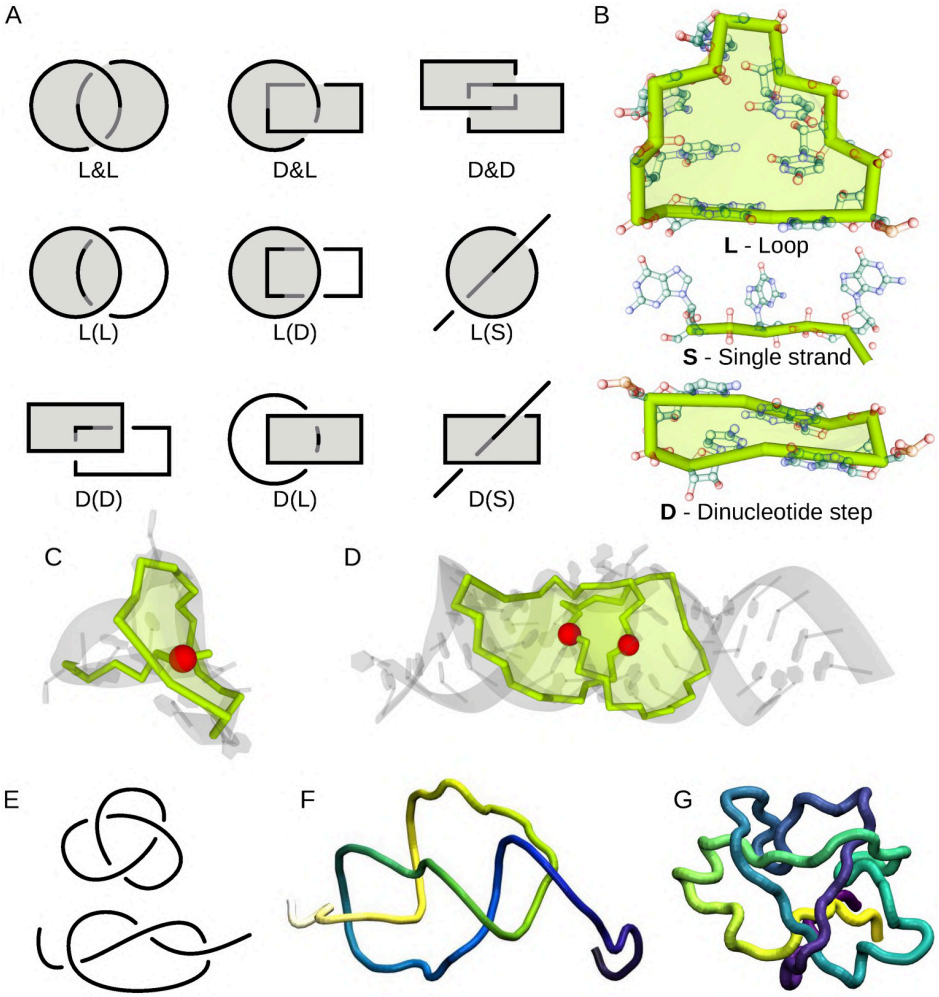


Fig 1. Types of entanglements. (A) Schematic drawings of interlaces (D&D, D&L, L&L) and lassos (D(D), D(L), D(S), L(D), L(L), L(S)). Loop (L) is represented by a circle, dinucleotide step (D) by a rectangle, and single strand (S) by a segment. (B) Example structural elements L, D, and S, and entanglements of structure elements—(C) L(S)-type lasso formed by a 6-nt apical loop and the 5'-end threaded through it, (D) misfolded conformation of two loops forming L&L-type interlace instead of kissing loops. (E) Diagrams of closed and opened trefoil knot, and two molecules with trefoils—(F) sRNA RydC (PDB ID: 4V2S:G [18]) and (G) PHD finger-like domain-containing protein 5A (PDB ID: 5ZYA:C [19]).

https://doi.org/10.1371/journal.pcbi.1011959.g001

topological defects. Micheletti *et al.* [15] found three knotted rRNAs solved by cryo-EM, but due to the absence of knots in high-resolution structures, concluded on some thermodynamic or kinetic mechanisms that minimize the entanglement of biologically viable structural RNAs. Yet a little later, the same group suggested that the properties of some of the predicted RNA secondary structures indicate the potential to form knots [16]. Only recently, a trefoil (3_1), the first non-trivial topological knot, was identified in high-resolution experimental RNA structure [17]. However, this is not sufficient to infer the formation of knots in RNA structures and their possible influence on the function of the molecule, as we can for proteins [20–24]. In addition, in the case of proteins, 6 different types of knots (3_1, 4_1, 5_2, 6_1) have already been identified [25], including those predicted by AI methods such as $3_1\#3_1$ [26] and 7_1 [27], whose

structure has been experimentally confirmed while other more complex knots [28] await confirmation.

Theoretical studies on ideal (ghost) polymers predict that probability of polymer being unknotted P_{unknot} falls exponentially with increasing length L and the inverse of the Kuhn length b of the chain: $P_{\text{unknot}}(L;b) = \exp(-\alpha L/b)$ [29–31], where the Kuhn length b is a quantity proportional to stiffness, and α is a constant that depends on a theoretical model used. This behavior is not observed for RNA, as the only known knotted RNA molecule is only 65 nt long [17]. In addition, no correlation was found between the size of the RNA molecule and its susceptibility to form entanglements of structural elements [9]. RNAs are brush-like real biopolymers living in a polar, polyelectrolyte environment, and their folding is driven by the minimization of free energy (theirs and the solvents). Research on the folding of knotted proteins reveals that although the knotted state may enhance protein stability, a topological energy barrier must first be overcome to reach the native knotted state [32–38]. This topological energy barrier hinders the formation of knotted proteins. Since methods that predict the native state of biopolymers do not fold molecules *de novo* but directly guess these native states, they have a harder job avoiding states unavailable due to topological energy barriers [39, 40].

In this work, we have analyzed all 3D RNA models predicted in the recent CASP competition. Using existing computational tools, RNAspider [11] and Topoly [41], we have scanned these predictions for entanglements of structural elements and topological knots. We have studied the structural and methodological aspects of susceptibility to entanglement generation in RNA models. All entanglements found are artifacts of the modeling procedures. Methods using deep learning entangle RNA chains more often than non-ML algorithms and generate quite complex topological knots. We believe that predictive methods should automatically reject models with invalid chain entanglements. This would improve the reliability and quality of the prediction of the 3D RNA structure.

Materials and methods

Benchmark data

To identify, count, and classify knots and entanglements of structural elements in the 3D RNA structure models predicted in CASP15, we downloaded the data from the competition website in September 2023. The CASP15 resources are available at https://predictioncenter.org/download_area/CASP15/predictions/RNA/ [42]). Data were cleaned of irrelevant metadata using the clean-casp-headers.awk script and grouped by 12 RNA targets: R1107 (69 nt), R1108 (69 nt), R1116 (157 nt), R1117 (30 nt), R1126 (363 nt), R1128 (238 nt), R1136 (374 nt), R1138 (720 nt), R1149 (124 nt), R1156 (135 nt), R1189 (118 nt) and R1190 (118 nt). The data set contained a total of 62 reference structures (for some targets, there was more than one structure) and 1,660 models computationally predicted by 41 modeling groups.

Methods used for entanglements of structure elements

The RNAspider web server was applied to identify and classify entanglements of structure elements in 3D RNA structures [11]. The system detects lassos and interlaces and assigns them to nine subclasses (Fig 1A and 1B)—L&L, L&D, D&D, L(S), L(D), L(L), D(S), D(D), and D(L)—based on the types of elements involved, i.e., loops, dinucleotide steps (both are closed structural elements), and single-stranded fragments (open element). It was run with the default settings of the advanced parameters. We then calculated the depth of each identified lasso formed from the loop L(*), where * stands for S, D, or L. This was done using an additional script (2_analyze_depth.py [43]), as RNAspider does not cover such functionality. The script was fed with information on the intersection points that RNAspider provided for every closed

structure element. The intersection point is determined by 3D coordinates, where the backbone or hydrogen bond in a canonical base pair punctures a surface that covers a closed element. We considered two cases. In the case of L(S), when a single strand is taken to lasso by the loop, we computed the depth as the minimum number of its nucleotides towards the 5' and 3' ends from the intersection point. In the second case, when a closed element is lassoed (L (L), L(D)) or when a single strand punctures the loop twice (L(S.)), the lasso is characterized by two intersection points. Therefore, we computed the depth as the number of nucleotides between the intersection points. Knowing the depth, we divided the lassos created by the loops into shallow (depth $\leq$ 5 nts) and deep (depth $>$ 5 nts). Following our experience with molecular dynamics, we hypothesize that shallow lassos L(*) can spontaneously disentangle during structure folding and therefore may not be treated as structure anomalies even though no lassos occur in the reference structures. In contrast, we consider interlaces, D(*) lassos, and deep L(*) lassos as modeling artifacts.

Methods used for topological knots

Topological knots were detected and identified using the Topoly Python package [41]. This collection of scripts offers features to study the topology of polymers and a generation of artificial, random polymer chains of a given topology. Here, the coordinates of the sugar-phosphate backbone atoms (P, O5′, C5′, C4′, C3′, O3′) were extracted from the 3D structure data. To identify knots, the Alexander polynomial [44] was calculated using the `topoly.alexander()` function, which before calculations closes the chain randomly by projecting RNA endpoints on the big sphere around the molecule and connecting them (two-point probabilistic closure, 200 closures, explained more in [45, 46]). We treated the structure as knotted if $<$50% of the closures were unknots. At first, Topoly identified 84 knotted RNAs. They were visually inspected to confirm entanglements. For some of them, a direct closure seemed to make more sense. We discarded 7 of these models since they were unknots after direct closure. 67 predictions failed to process by Topoly due to non-unique atom coordinates (47 models) or too tangled structure (20 models). Structures in the latter subset were visually inspected and confirmed to be too densely packed to be correct. We assigned them to a category named TTC (too tangled to check) and did not include them in further analysis, unlike trefoils (Fig 1C) and complex knots with known classification.

Results and discussion

In this work, we analyzed entanglements in 3D RNA models submitted to CASP15 following the pipeline presented in S1 Fig. We took into account the entanglements of structural elements and topological knots. It is important to emphasize that none of the reference structures, of which there were 62 models, was entangled from the point of view of the structure elements or topological knots. In contrast, of the 1,660 predicted models, 160 models have either entanglements of structure elements or topological knots, 83 models have only entanglements of structural elements, 34 models have only topological knots, and 43 models have both (S1 Table). Fig 2A shows a Venn diagram detailing the number of predicted models that contain given types of entanglements. Note that the existence of topological knots in the 3D RNA model does not equate to entanglement of the structural elements of the model and vice versa. In 126 entangled models, we encountered 173 interlaces and 300 lassos (Fig 2B). The latter group included 67 deep and 33 shallow conformations of the L(*) type. Note that among the lassos, the L(S) and L(D) types are most abundantly represented in the RNA predictions. Intuition dictates that this is entirely reasonable since forming a lasso around a single or double helix should be relatively easy. In contrast, pushing a non-single-stranded fragment of a

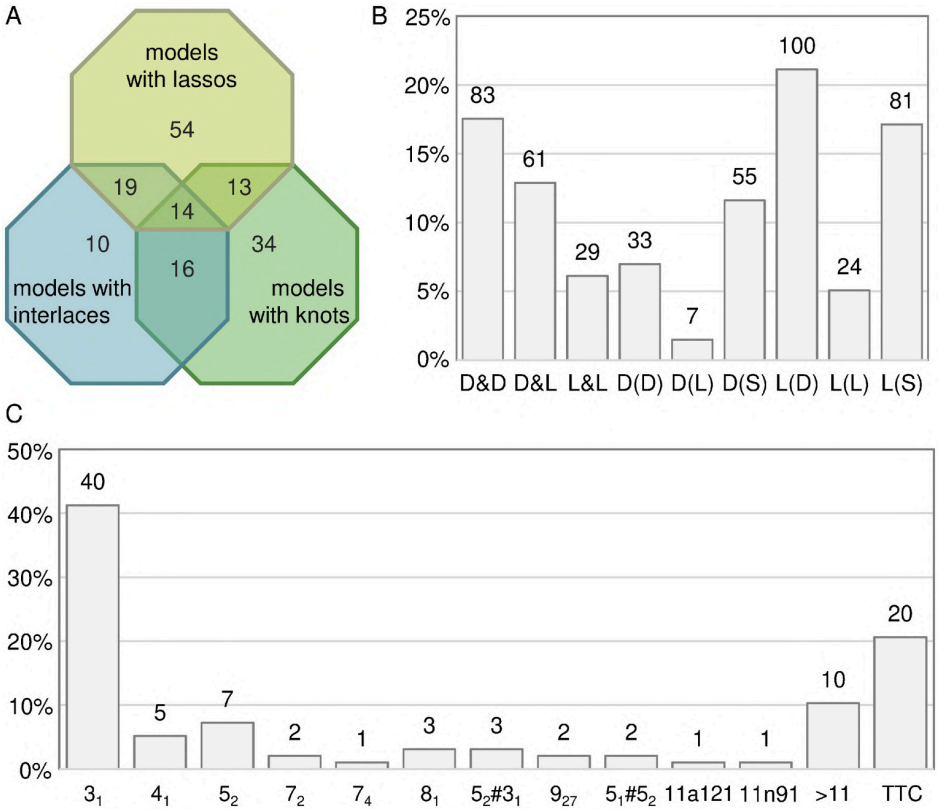


Fig 2. (A) Models with various entanglement types in numbers. (B) Entanglements of structure elements and (C) topological knots in RNA predictions by type. Column labels in (B) and (C) show the total number of entanglements of a given type across all predictions.

https://doi.org/10.1371/journal.pcbi.1011959.g002

structure through a dinucleotide step or a loop through a loop seems quite difficult. Therefore, these types of entanglements are relatively rare in the models. In 77 RNA models with topological knots, we found a predominance of trefoils—they account for more than half of all knots with assigned classes excluding the TTC subset (Fig 2C). TTC, the second large subgroup, is made up of 20 models whose knotting is too complex to classify correctly.

Target-focused analysis

In this part of our experiment, we focused on the structural aspect of the structure entanglement problem. We asked whether, among the RNA targets in CASP15, we could distinguish structures that show a greater /lower susceptibility to entanglement during computer modeling. Answering this question required analyzing the predicted models clustered by target (see Fig 3 and Table 1).

The target structures were divided by the CASP assessors into natural (8 targets) and synthetic (4 targets). The former cluster distinguished between easy (R1117), medium (R1107, R1108), and difficult (R1116, R1149, R1156, R1189, R1190) targets. Non-natural ones included R1126, R1128, R1136, and R1138. Assignment of a target to particular group indicates the ease/difficulty of its structure prediction due to the similarity to known experimental

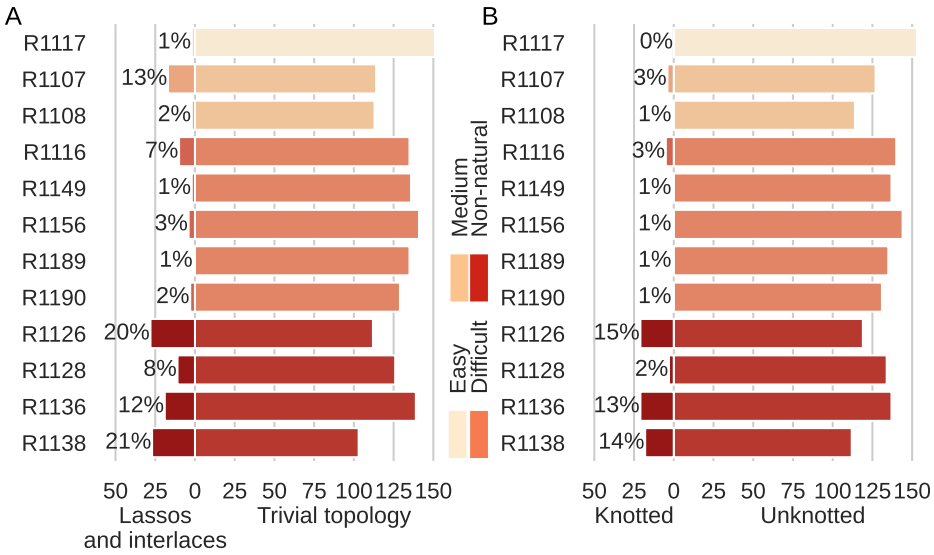


Fig 3. Distribution of entangled 3D RNA structure predictions by target. Target structures are grouped by difficulty [8]. Results are displayed separately for (A) entanglements of structure elements (lassos and interlaces) and (B) topological knots.

https://doi.org/10.1371/journal.pcbi.1011959.g003

Table 1. Entanglements in RNA predictions by target.

Predictions for	#Models	#Interlaces	#Lassos	#Trefoils	#Other knots	#TTC
natural targets	1,095	17	62	14	0	0
synthetic targets	565	156	232	26	37	20
total number	1,660	173	294	40	37	20

https://doi.org/10.1371/journal.pcbi.1011959.t001

structures [8]. In general, 33% (41) of the entangled structures are predictions of natural RNA targets, and 67% (85) are models from the non-natural cluster.

Based on the analysis of entanglements of structural elements (Fig 3A), we can say that the probability of entangled predictions for natural RNAs equals 0.03, while for non-natural targets it is 0.15, which is 5x higher. If we look at the sets of predictions per target, we can see that the entangled structures represent 8–20% for the synthetic targets. In contrast, in the sets of models for natural targets, the percentage of entangled predictions is 0.74–2.76%. The exceptions here are R1107 (12.98%) and R1116 (6.90%), the former of which is a moderately difficult structure with a pseudoknot, and the latter is classified as difficult.

The diagram prepared for the topological knots (Fig 3B) has similar characteristics as in the case of entangled structure elements. The highest number of knotted models is found for the targets for which we observe the most entanglements of structural elements. The study reveals that the knotting probability is 0.01 for natural structures and 0.12 (10x higher) for non-natural ones. All knotted predictions of natural structures are trefoils (3_1, simplest non-trivial knot). For non-natural structures, only more complex knots appear, moreover, they make up the majority of knotted structures.

Some predictions contain more than one entanglement. Table 1 presents a distribution of various types of entanglements across predictions for natural and synthetic targets. For

simplicity, topological knots were divided into trefoils (the simplest non-trivial knot) and other knots (more complicated ones). The rightmost column represents unclassified knots from non-physically dense structures (TTC—too tangled to check knot type).

Finally, let us add that the largest number of entangled 3D models (both from the point of view of topological knots and entangled structure elements) was identified in the predictions for the three largest targets, R1138 (720 nts), R1136 (374 nts), and R1126 (363 nts), all of which are synthetic. However, note that, as shown in [9], no simple relationship has yet been observed between the entanglements of structural elements and the size of the structure. Given that topological knots more complex than trefoil are found only in the predictions of synthetic targets, it seems that the large number of entanglements and their complexity are a result of the specificity of the non-natural target rather than the structure size.

Method-focused analysis

Next, we analyzed 3D RNA models by prediction group and examined the problem of structure entanglement from the point of view of the method. We checked how many nontrivial topologies, distinguishing between topological knots, lassos, and interlaces, appear in the models generated by a given prediction method (Fig 4). Of the 41 groups, 16 submitted only unentangled models (a total of 510 predictions, which makes 30% of all predictions). The remaining 25 groups predicted 1,150 models, of which 126 (10%) include entanglements of structure elements, and 78 (6%) are knotted from the topological point of view.

We divided the predictive methods into traditional and ML-based according to the information provided in the CASP15 book of abstracts [47]. Groups that applied machine learning predicted more models on average than groups using traditional methods—17/41 groups (41%) used ML and predicted 814/1,660 models (49%); 24/41 groups (59%) used traditional approaches and submitted 846/1,660 models (51%). Among 814 RNA models generated by machine learning-based methods, 132 included entangled structure elements, and 67 had a topological knot (34 were trefoils). For comparison, in the set of 846 RNA predictions by traditional algorithms, we found only 33 with entangled structure elements, and 10 with a knot (6 were trefoils). It follows that ML methods are 4x more prone to generate models containing entanglements of structural elements and 7x more prone to predict a knotted model than traditional algorithms. In particular, ML methods generated 6x more trefoils and 8x more complex knots than non-ML approaches. Moreover, all 20 structures, which were too tangled to check their knot type (TTC), were also predicted by ML methods. The distribution of the entanglement types is presented in Table 2.

Finally, recall that CASP accepts submissions from two categories of participants, web servers, and human groups. Predictions from the former category are fully automatic and must be submitted within 72 hours of publishing the target sequence. Human groups have 3 weeks to make predictions and can utilize any method to support the modeling process, including laboratory experiments to refine their models. With this in mind, we checked whether topological knots and entanglements of structure elements are more often in the web server than in human predictions. Among the 41 participants, there were 9 web servers. They submitted a total of 423 models, including 29 (0.07%) with entanglements. 32 human groups predicted 1,237 models, of which 97 (0.08%) were entangled. The number of entanglements of each type in the models submitted by both categories of participants is shown in Table 2. Based on these data, the thesis that automated predictions are more likely to get entangled than those submitted by experts cannot be confirmed.

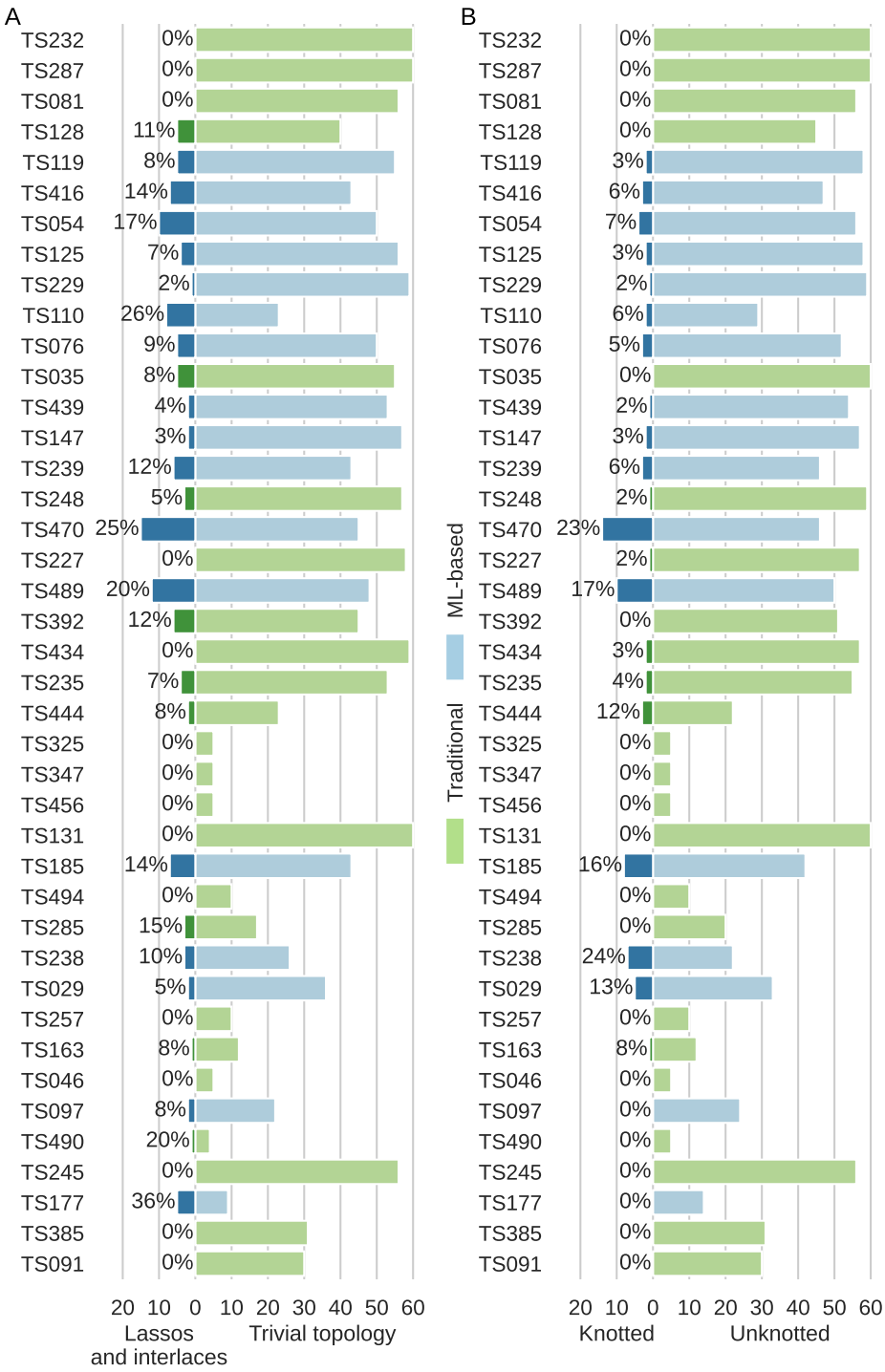


Fig 4. Distribution of entangled 3D RNA structure predictions by method. Modeling groups (classified as traditional or ML-based) are listed due to their ranking in CASP15 with the best one at the top. Results are shown separately for (A) entanglements of structure elements (lassos and interlaces) and (B) topological knots.

https://doi.org/10.1371/journal.pcbi.1011959.g004

Table 2. Entanglements in RNA predictions by method.

Predicted by	#Models	#Interlaces	#Lassos	#Trefoils	#Other knots	#TTC
traditional methods	846	62	103	6	4	0
ML-based methods	814	111	191	34	33	20
servers	423	47	66	7	10	8
human predictors	1,237	126	228	33	27	12
total number	1,660	173	294	40	37	20

https://doi.org/10.1371/journal.pcbi.1011959.t002

Example predictions with artifacts

First, let us present the 3D RNA prediction that contains a lasso and was generated using the machine learning-based method. The R1107TS416_2 model was submitted by the TS416 group (AIchemy_RNA). It targeted the natural 69-nt-long RNA structure of the human CPEB3 HDV-like ribozyme (PDB ID: 7QR4) [48] (target ID: R1107). The native structure contains a pseudoknot formed between the dangling 5'-end (residues 1–6) and the three-way junction (residues 31–36). The R1107TS416_2 model is not an ideal reconstruction of the target either from the point of view of secondary structure (S2 Fig) (INF_{all} = 0.77) or the 3D topology (RMSD = 7.81Å, TM-score = 0.392). The model contains the L(S)-type lasso formed between the 26-nt-long hairpin closed by a pseudoknotted base pair 0:6–0:31 and a 34-nt-long single strand (0:36–0:69) (Fig 5A). Both of these structure elements also exist in the native structure, although in the latter a single strand bypasses the loop. In the predicted model, the intersection of the area inside the loop is between residues 0:A62 and 0:U63 of the puncture strand. The single strand passes quite a distance from the chain that forms the loop. Therefore, we did not observe clashed atoms in this part of the structure; in general, the Clash score is low and equals 14.02. According to the threshold adopted for shallow lassos, the entanglement found belongs to the deep category; its depth equals 6 nucleotides.

The other example, R1136TS110_4 model, was submitted by the TS110 group (DF_RNA). It is an ML-driven prediction of the synthetic construct, which is the 3D structure of a broccolli-pepper aptamer FRET tile in the ligand-bound state (PDB ID: 7ZJ4) [49] (target ID: R1136). The reference structure consists of 374 nucleotides and has a non-trivial topology with kissing-loop interactions between two hairpins (0:50-0:60+0:142–0:152; 0:75–0:85+0:121-0:131). The secondary structure of the predicted model (S2 Fig) is quite well reconstructed (INF_{all} = 0.82), while the 3D fold clearly deviates from the native (RMSD = 44.35Å, TM-score = 0.304). Clash score = 49.78. The model contains six entanglements of structural elements—two D(D), two D&D, one D&L, and one D(L) (Fig 5B), formed between two 10-nt-long hairpin loops (0:142-0:152; 0:50-0:60) and seven dinucleotide steps (0:119-0:120+0:132–0:133; 0:159–0:160+0:216–0:217; 0:160–0:161+0:215–0:216; 0:335–0:336+0:345–0:346; 0:49–0:50+0:60–0:61; 0:336–0:337+0:344–0:345; 0:293–0:294+0:299-0:300). They are not only artifacts of modeling but also incorrect conformations hardly possible to form while RNA folding.

The next example presents different topological knots in the 3D RNA models predicted for the largest RNA target of CASP15, that is, R1138 (720 nts). The native structure (synthetic construct) is a young conformer of a 6-helix bundle of RNA with clasp (PDB IDs: 7PTK, 7PTL) [50]. As shown in Fig 6A, it does not contain topological knots. Model R1138TS227_2 (Fig 6B) was generated by a traditional (non-ML) approach used by the TS227 group (GinobiFold). RMSD = 48.23Å and TM-score = 0.204 indicate a significant deviation of its 3D topology from the target structure, while the secondary structure is quite well-reproduced (INF_{all} = 0.82). Interestingly, clashes are almost non-existent (Clash score = 0.30). This model is knotted and

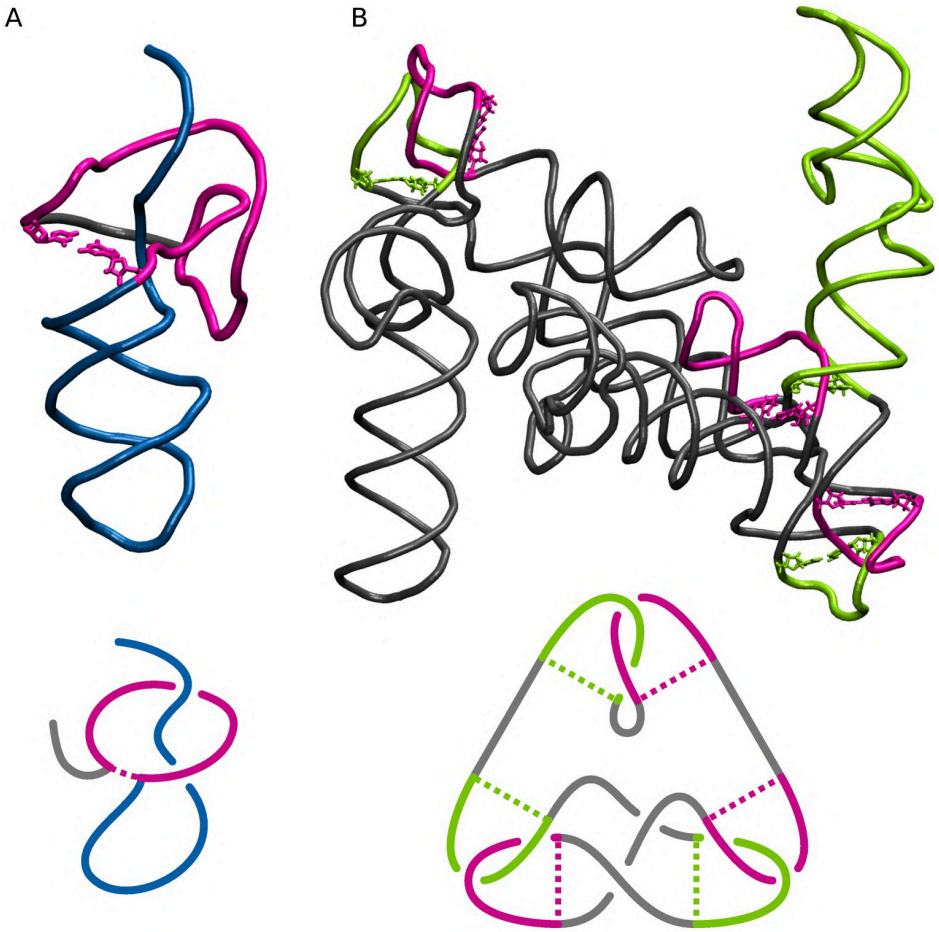


Fig 5. Two models with entangled structural elements predicted by ML-based methods and diagrams showing included entanglements. (A) The R1107TS416_2 model (69 nt) with a lasso. (B) The R1136TS110_4 model (375 nt) with highlighted three interlaces forming the 3_1 knot. Hydrogen bonds closing an entangled loop are marked with dotted lines. Secondary structures of both models are shown in S2 Fig.

https://doi.org/10.1371/journal.pcbi.1011959.g005

forms a trefoil (3_1), the simplest non-trivial knot. On the other hand, the R1138TS054_3 model was predicted using the machine learning-based method by the TS054 group (Ultra-Fold). Its ratings are similar to the previous example (RMSD = 39.73Å, TM-score = 0.186, INF_{all} = 0.85, and Clash score = 4.81), however, this structure forms a more complicated 7_2 knot. This clearly shows that existing evaluation measures are not correlated with the complexity of the structure entanglement.

Conclusion

An analysis of the 1,660 3D RNA models predicted within CASP15 showed that predictive methods using machine learning are four times more likely than traditional tools to generate structures with entanglements, which are artifacts of the computational process. We hypothesize that the generation of entanglements may be due to the algorithm prioritization of more compact structures. The predictions submitted by web servers and human groups contain the same percentages of entangled models. The implication is that predictors do not use any, either

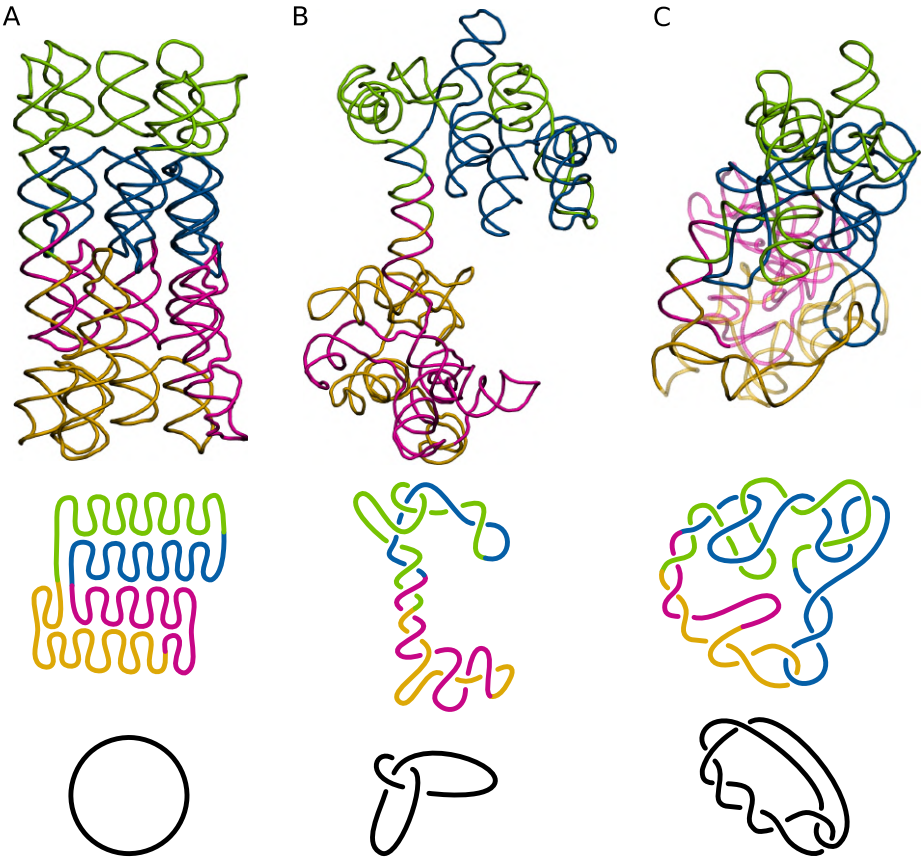


Fig 6. 3D models predicted for the same target containing various topological knots and schematics of the latter. (A) Target structure R1138 (mature, 720 nt). (B) The R1138TS227_2 model generated by traditional method. (C) The R1138TS054_3 model predicted by ML-based method. Secondary structures of all three models are shown in S3 Fig.

https://doi.org/10.1371/journal.pcbi.1011959.g006

automatic or non-automatic, verification of their models for topological anomalies. The characteristics of the modeled structure also appear to affect the probability of entanglement. In the set of all predictions, the largest number of entangled models was generated for large targets of synthetic RNA molecules. The probability of entanglement in this subset was significantly higher than for natural structures. We believe that enriching prediction methods with procedures to validate the topology would increase the accuracy of 3D RNA structure prediction. Currently, such validation can be easily done with the RNAspider and Topoly packages applied in the presented work. However, in the future, other ways of measuring topological defects could be designed, such as a loss function that would identify incorrect entanglements and fix them. This is not a straightforward task. Well-established methods for the recognition of knotted closed chain topologies are knot invariants such as Alexander, Jones, and HOMFLY-PT polynomials [44, 51–53]. All of them are computed using skein relations, which have exponential time complexity dependent on the number of crossings on a polymer projection. The exception is the in-point calculation of the Alexander polynomial that uses a determinant of the Alexander matrix and runs with a polynomial-time complexity. However, RNA structures are mainly open chains. For such cases, the Jones polynomial for knotoids [54–56] can be used, the value of which continuously changes with the coordinates of the chain. The

alternative is to close the chain and use methods designed for closed chains [57]. Chain closure allows safe use of the KMT chain reduction algorithm [58] and greatly reduces the complexity of the chain to analyze. However, note that different chain closures can result in different knots. Due to the time of gradient computation, none of the above methods seems suitable to be applied as a loss function for predictive algorithms. More suitable candidates for a loss function might be machine learning models trained specifically for knot recognition [59–61].

Supporting information

S1 Table. Entangled RNA 3D models predicted in CASP15.
(ODS)

S1 Fig. Computational pipeline used to identify and analyze entangled 3D RNA models in CASP15 submissions.
(PDF)

S2 Fig. Secondary structures of two models with entangled structure elements predicted by ML-based methods. (A) The R1107TS416_2 model (69 nt) with a lasso. (B) The R1136TS110_4 model (375 nt) with two D&D entanglements (D1 interlaced with D1′ and D2 interlaced with D2′) and one D&L entanglement (D3 interlaced with L3). Colors match those applied in Fig 5.
(EPS)

S3 Fig. Secondary structures of models predicted for the same target containing various topological knots. (A) Target structure R1138 (mature, 720 nt). (B) The R1138TS227_2 model generated by a traditional method. (C) The R1138TS054_3 model predicted by an ML-based method. Colors match those applied in Fig 6.
(EPS)

Author Contributions

Conceptualization: Joanna I. Sulkowska, Marta Szachniuk.

Formal analysis: Maciej Antczak, Tomasz Zok, Marta Szachniuk.

Funding acquisition: Joanna I. Sulkowska.

Investigation: Bartosz A. Gren, Joanna I. Sulkowska, Marta Szachniuk.

Methodology: Bartosz A. Gren, Maciej Antczak, Tomasz Zok, Marta Szachniuk.

Project administration: Marta Szachniuk.

Software: Bartosz A. Gren, Maciej Antczak, Tomasz Zok.

Supervision: Marta Szachniuk.

Validation: Maciej Antczak, Joanna I. Sulkowska, Marta Szachniuk.

Visualization: Bartosz A. Gren, Tomasz Zok.

Writing – original draft: Bartosz A. Gren, Maciej Antczak, Tomasz Zok, Marta Szachniuk.

Writing – review & editing: Joanna I. Sulkowska, Marta Szachniuk.

References

1. Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. Proteins. 2021; 89(12):1687–1699. https://doi.org/10.1002/prot.26171 PMID: 34218458
2. Schneider B, Sweeney B, Bateman A, Cerny J, Zok T, Szachniuk M. When will RNA get its AlphaFold moment? Nucleic Acids Res. 2023; 51(18):9522–9532. https://doi.org/10.1093/nar/gkad726 PMID: 37702120
3. Parisien M, Cruz JA, Westhof E, Major F. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. RNA. 2009; 15(10):1875–1885. https://doi.org/10.1261/rna.1700409 PMID: 19710185
4. Szachniuk M. RNApolis: computational platform for RNA structure analysis. Foundations of Computing and Decision Sciences. 2019; 44(2):241–257. https://doi.org/10.2478/fcds-2019-0012
5. Wiedemann J, Zok T, Milostan M, Szachniuk M. LCS-TA to identify similar fragments in RNA 3D structures. BMC Bioinform. 2017; 18:456. https://doi.org/10.1186/s12859-017-1867-6 PMID: 29058576
6. Magnus M, Antczak M, Zok T, Wiedemann J, Lukasiak P, Cao Y, et al. RNA-Puzzles toolkit: A computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. Nucleic Acids Res. 2020; 48(2):576–588. https://doi.org/10.1093/nar/gkz1108 PMID: 31799609
7. Carrascoza F, Antczak M, Miao Z, Westhof E, Szachniuk M. Evaluation of the stereochemical quality of predicted RNA 3D models in the RNA-Puzzles submissions. RNA. 2021; 28(2):250–262. https://doi.org/10.1261/rna.078685.121 PMID: 34819324
8. Das R, Kretsch RC, Simpkin AJ, Mulvaney T, Pham P, Rangan R, et al. Assessment of three-dimensional RNA structure prediction in CASP15. Proteins: Struct, Funct, Bioinf. 2023; 91(12):1747–1770. https://doi.org/10.1002/prot.26602
9. Popenda M, Zok T, Sarzynska J, Korpeta A, Adamiak RW, Antczak M, et al. Entanglements of structure elements revealed in RNA 3D models. Nucleic Acids Res. 2021; 49(17):9625–9632. https://doi.org/10.1093/nar/gkab716 PMID: 34432024
10. Zając S, Geary C, Andersen ES, Dabrowski-Tumanski P, Sulkowska JI, Sułkowski P. Genus trace reveals the topological complexity and domain structure of biomolecules. Sci Rep. 2018; 8(1):17537. https://doi.org/10.1038/s41598-018-35557-3 PMID: 30510290
11. Luwanski K, Hlushchenko V, Popenda M, Zok T, Sarzynska J, Martsich D, et al. RNAspider: a webserver to analyze entanglements in RNA 3D structures. Nucleic Acids Res. 2022; 50(W1):W663–W669. https://doi.org/10.1093/nar/gkac218 PMID: 35349710
12. VanLoock M, Harris B, Harvey S. To knot or not to knot? Examination of 16s ribosomal RNA models. J Biomol Struct Dyn. 1998; 16(3):709–713. https://doi.org/10.1080/07391102.1998.10508282 PMID: 10052626
13. Poblete S, Guzman HV. Structural 3D domain reconstruction of the RNA genome from viruses with secondary structure models. Viruses. 2021; 13(8):1555. https://doi.org/10.3390/v13081555 PMID: 34452420
14. Cruz-León S, Assenza S, Poblete S, Guzman HV. In: Comas-Garcia M, Rosales-Mendoza S, editors. RNA multiscale simulations as an interplay of electrostatic, mechanical properties, and structures inside viruses. Springer International Publishing; 2023. p. 27–56.
15. Micheletti C, Di Stefano M, Orland H. Absence of knots in known RNA structures. PNAS. 2015; 112(7):2052–2057. https://doi.org/10.1073/pnas.1418445112 PMID: 25646433
16. Burton AS, Di Stefano M, Lehman N, Orland H, Micheletti C. The elusive quest for RNA knots. RNA Biol. 2016; 13(2):134–139. https://doi.org/10.1080/15476286.2015.1132069 PMID: 26828280
17. Niemyska W, Mukherjee S, Gren B, Niewieczerzal S, Bujnicki JM, Sulkowska JI. Discovery of a trefoil knot in the RydC RNA: challenging previous notions of RNA topology. J Mol Biol. 2024; 436(6):168455. https://doi.org/10.1016/j.jmb.2024.168455 PMID: 38272438
18. Dimastrogiovanni D, Frohlich KS, Bandyra KJ, Bruce HA, Hohensee S, Vogel J, et al. Recognition of the small regulatory RNA RydC by the bacterial Hfq protein. eLife. 2014; 3:e05375. https://doi.org/10.7554/eLife.05375 PMID: 25551292
19. Finci LI, Zhang X, Huang X, Zhou Q, Tsai J, Teng T, et al. The cryo-EM structure of the SF3b spliceosome complex bound to a splicing modulator reveals a pre-mRNA substrate competitive mechanism of action. Genes Dev. 2018; 32:309–320. https://doi.org/10.1101/gad.311043.117 PMID: 29491137
20. King NP, Yeates EO, Yeates TO. Identification of rare slipknots in proteins and their implications for stability and folding. J Mol Biol. 2007; 373(1):153–166. https://doi.org/10.1016/j.jmb.2007.07.042 PMID: 17764691
21. Sulkowska JI, Rawdon EJ, Millet KC, Onuchic JN, Stasiak A. Conservation of complex knotting and slipknotting patterns in proteins. Biophys J. 2012; 102(3):253a. https://doi.org/10.1073/pnas.1205918109 PMID: 22685208

22. Hou YM, Matsubara R, Takase R, Masuda I, Sulkowska JI. TrmD: a methyl transferase for tRNA methylation with m1G37. The Enzymes. 2017; 41:89–115. https://doi.org/10.1016/bs.enz.2017.03.003 PMID: 28601227
23. Sulkowska JI. On folding of entangled proteins: knots, lassos, links and θ-curves. Curr Opin Struct Biol. 2020; 60:131–141. https://doi.org/10.1016/j.sbi.2020.01.007 PMID: 32062143
24. Hsu STD. Folding and functions of knotted proteins. Curr Opin Struct Biol. 2023; 83:102709. https://doi.org/10.1016/j.sbi.2023.102709 PMID: 37778185
25. Jamroz M, Niemyska W, Rawdon EJ, Stasiak A, Millett KC, Sułkowski P, et al. KnotProt: a database of proteins with knots and slipknots. Nucleic Acids Res. 2015; 43(D1):D306–D314. https://doi.org/10.1093/nar/gku1059 PMID: 25361973
26. da Silva FB, Lewandowska I, Kluza A, Niewieczerzal S, Augustyniak R, Sulkowska JI. First crystal structure of double knotted protein TrmD-Tm1570–inside from degradation perspective. bioRxiv [pre-print]. 2023; p. 2023–03.
27. Hsu MF, Sriramoju MK, Lai CH, Chen YR, Huang JS, Ko TP, et al. Structure, dynamics, and stability of the smallest and most complex 71 protein knot. J Biol Chem. 2024; 300(1). https://doi.org/10.1016/j.jbc.2023.105553 PMID: 38072060
28. Niemyska W, Rubach P, Gren BA, Nguyen ML, Garstka W, Bruno da Silva F, et al. AlphaKnot: server to analyze entanglement in structures predicted by AlphaFold methods. Nucleic Acids Res. 2022; 50(W1): W44–W50. https://doi.org/10.1093/nar/gkac388 PMID: 35609987
29. Sumners D, Whittington S. Knots in self-avoiding walks. J Phys A Math Gen. 1988; 21(7):1689. https://doi.org/10.1088/0305-4470/21/7/030
30. Pippenger N. Knots in random walks. Discrete Appl Math. 1989; 25(3):273–278. https://doi.org/10.1016/0166-218X(89)90005-X
31. Deguchi T, Tsurusaki K. Universality of random knotting. Phys Rev E. 1997; 55(5):6245. https://doi.org/10.1103/PhysRevE.55.6245
32. Sulkowska JI, Sulkowski P, Onuchic J. Dodging the crisis of folding proteins with knots. Biophys J. 2009; 96(3):81a. https://doi.org/10.1073/pnas.0811147106 PMID: 19211785
33. Li W, Terakawa T, Wang W, Takada S. Energy landscape and multiroute folding of topologically complex proteins adenylate kinase and 2ouf-knot. PNAS. 2012; 109(44):17789–17794. https://doi.org/10.1073/pnas.1201807109 PMID: 22753508
34. Sułkowska JI, Noel JK, Onuchic JN. Energy landscape of knotted protein folding. PNAS. 2012; 109 (44):17783–17788. https://doi.org/10.1073/pnas.1201804109 PMID: 22891304
35. a Beccara S, Škrbić T, Covino R, Micheletti C, Faccioli P. Folding pathways of a knotted protein with a realistic atomistic force field. PLoS Comput Biol. 2013; 9(3):e1003002. https://doi.org/10.1371/journal.pcbi.1003002 PMID: 23555232
36. Soler MA, Rey A, Faísca PF. Steric confinement and enhanced local flexibility assist knotting in simple models of protein folding. Phys Chem Chem Phys. 2016; 18(38):26391–26403. https://doi.org/10.1039/C6CP05086G PMID: 27722468
37. Jackson SE, Suma A, Micheletti C. How to fold intricately: using theory and experiments to unravel the properties of knotted proteins. Curr Opin Struct Biol. 2017; 42:6–14. https://doi.org/10.1016/j.sbi.2016.10.002 PMID: 27794211
38. Dabrowski-Tumanski P, Sulkowska JI. Topological knots and links in proteins. PNAS. 2017; 114 (13):3415–3420. https://doi.org/10.1073/pnas.1615862114 PMID: 28280100
39. Dabrowski-Tumanski P, Stasiak A. AlphaFold Blindness to Topological Barriers Affects Its Ability to Correctly Predict Proteins' Topology. Molecules. 2023; 28(22):7462. https://doi.org/10.3390/molecules28227462 PMID: 38005184
40. Sramkova D, Sikora M, Uchal D, Klimentova E, Perlinska AP, Nguyen ML, et al. Knot or Not? Sequence-Based Identification of Knotted Proteins With Machine Learning. bioRxiv [pre-print]. 2023; p. 2023–09.
41. Dabrowski-Tumanski P, Rubach P, Niemyska W, Gren BA, Sulkowska JI. Topoly: Python package to analyze topology of polymers. Brief Bioinform. 2021; 22(3):bbaa196. https://doi.org/10.1093/bib/bbaa196 PMID: 32935829
42. Kryshtafovych A, Antczak M, Szachniuk M, Zok T, Kretsch RC, Rangan R, et al. New prediction categories in CASP15. Proteins: Struct, Funct, Bioinf. 2023; 91(12):1550–1557. https://doi.org/10.1002/prot.26515
43. Zok T. BioCommons: A Robust Java Library for RNA Structural Bioinformatics. Bioinformatics. 2021; 37 (17):2766–2767. https://doi.org/10.1093/bioinformatics/btab069 PMID: 33532837
44. Alexander JW. Topological invariants of knots and links. Trans Am Math Soc. 1928; 30(2):275–306. https://doi.org/10.1090/S0002-9947-1928-1501429-1

45. Millett KC, Rawdon EJ, Stasiak A, Sulkowska JI. Identifying knots in proteins. Biochem Soc Trans. 2013; 41(2):533–537. https://doi.org/10.1042/BST20120339 PMID: 23514149

46. Jamroz M, Niemyska W, Rawdon EJ, Stasiak A, Millett KC, Sułkowski P, et al. KnotProt: a database of proteins with knots and slipknots. Nucleic Acids Res. 2015; 43(D1):D306–D314. https://doi.org/10.1093/nar/gku1059 PMID: 25361973

47. CASP15 Abstract book; 2022. Available from: https://predictioncenter.org/casp15/doc/CASP15_Abstracts.pdf.

48. Przytula-Mally AI, Engilberge S, Johannsen S, Olieric V, Masquida B, Sigel RK. Anticodon-like loop-mediated dimerization in the crystal structures of HdV-like CPEB3 ribozymes. bioRxiv [pre-print]. 2022; p. 2022–2209.

49. Kretsch RC, Andersen ES, Bujnicki JM, Chiu W, Das R, Luo B, et al. RNA target highlights in CASP15: Evaluation of predicted models by structure providers. Proteins: Struct, Funct, Bioinf. 2023; 91 (12):1600–1615. https://doi.org/10.1002/prot.26550 PMID: 37466021

50. McRae EKS, Rasmussen HO, Liu J, Boggild A, Nguyen MTA, Sampedro Vallina N, et al. Structure, folding and flexibility of co-transcriptional RNA origami. Nat Nanotechnol. 2023; 18:808–817. https://doi.org/10.1038/s41565-023-01321-6 PMID: 36849548

51. Jones VF. Hecke algebra representations of braid groups and link polynomials. In: New Developments In The Theory Of Knots. World Scientific; 1987. p. 20–73.

52. Freyd P, Yetter D, Hoste J, Lickorish WR, Millett K, Ocneanu A. A new polynomial invariant of knots and links. Bull Amer Math Soc. 1985; 12(2):239–246. https://doi.org/10.1090/S0273-0979-1985-15361-3

53. Przytycki JH, Traczyk P. Invariants of links of Conway type. arXiv [pre-print] arXiv:161006679. 2016;.

54. Gügümcü N, Kauffman LH. New invariants of knotoids. Eur J Comb. 2017; 65:186–229. https://doi.org/10.1016/j.ejc.2017.06.004

55. Dorier J, Goundaroulis D, Benedetti F, Stasiak A. Knoto-ID: a tool to study the entanglement of open protein chains using the concept of knotoids. Bioinformatics. 2018; 34(19):3402–3404. https://doi.org/10.1093/bioinformatics/bty365 PMID: 29722808

56. Panagiotou E, Kauffman LH. Knot polynomials of open and closed curves. P Roy Soc A. 2020; 476 (2240):20200124. https://doi.org/10.1098/rspa.2020.0124 PMID: 32922152

57. Millett KC, Rawdon EJ, Stasiak A, Sułkowska JI. Identifying knots in proteins. Biochem Soc Trans. 2013; 41(2):533–537. https://doi.org/10.1042/BST20120339 PMID: 23514149

58. Koniaris K, Muthukumar M. Self-entanglement in ring polymers. J Chem Phys. 1991; 95(4):2873–2881. https://doi.org/10.1063/1.460889

59. Vandans O, Yang K, Wu Z, Dai L. Identifying knot types of polymer conformations by machine learning. Phys Rev E. 2020; 101(2):022502. https://doi.org/10.1103/PhysRevE.101.022502 PMID: 32168694

60. Gukov S, Halverson J, Ruehle F, Sułkowski P. Learning to unknot. Mach Learn Sci Technol. 2021; 2 (2):025035. https://doi.org/10.1088/2632-2153/abe91f

61. Bruno da Silva F, Gabrovsek B, Korpacz M, Luczkiewicz K, Niewieczerzal S, Sikora M, et al. Knots and θ-Curves Identification in Polymeric Chains and Native Proteins Using Neural Networks. Macromolecules. 2024;. https://doi.org/10.1021/acs.macromol.3c02479

Article 8

Topoly: Python package to analyze topology of polymers

Briefings in Bioinformatics, 22(3), 2021, 1–8

doi: 10.1093/bib/bbaa196
Problem Solving Protocol

Topoly: Python package to analyze topology of polymers

Pawel Dabrowski-Tumanski, Pawel Rubach, Wanda Niemyska, Bartosz Ambrozy Gren and Joanna Ida Sulkowska

Corresponding author: Joanna Ida Sulkowska, Centre of New Technologies, University of Warsaw, Warsaw, 02-097, Poland; Faculty of Chemistry, University of Warsaw, 02-093, Warsaw, Poland. Tel.: +48-22-55-43678 E-mail: jsulkowska@cent.uw.edu.pl;

Abstract

The increasing role of topology in (bio)physical properties of matter creates a need for an efficient method of detecting the topology of a (bio)polymer. However, the existing tools allow one to classify only the simplest knots and cannot be used in automated sample analysis. To answer this need, we created the Topoly Python package. This package enables the distinguishing of knots, slipknots, links and spatial graphs through the calculation of different topological polynomial invariants. It also enables one to create the minimal spanning surface on a given loop, e.g. to detect a lasso motif or to generate random closed polymers. It is capable of reading various file formats, including PDB. The extensive documentation along with test cases and the simplicity of the Python programming language make it a very simple to use yet powerful tool, suitable even for inexperienced users. Topoly can be obtained from https://topoly.cent.uw.edu.pl.

Key words: knot; lasso; graph; cyclotide; protein; DNA

Introduction

The history of science provides dozens of cases when the properties of matter were dictated not by its composition, but rather by its spatial arrangements. In the 20th century, great attention was paid to monitor the chirality of the compounds. Recently, researchers moved one step further and focused on the analysis of the influence of the topology on the properties of polymers. In particular, the effect of knots, slipknots and links in linear polymers was studied in designed compounds [3, 5, 6, 55, 62, 66], DNA [4, 57, 59] and proteins [10, 13, 19, 37, 63, 67]. The inclusion of branching points showed the existence of the θ-curve topology, lasso motif or cystine knots [14, 18, 21, 49], where the depth of the piercing affects the properties of the polymer [3, 51].

However, for complex structures, the identification of the topologically nontrivial motif becomes a time-consuming and hard task even for specialists. Therefore, a tool automatically identifying the nontrivial topology is required. For some topologies (usually knots or links), one may use some dedicated web servers [16–18, 32, 39, 41, 42, 64], plugins [25, 26, 43] or standalone packages [23, 31, 33]. However, such an approach is not sufficient when one needs to perform a meticulous analysis

Pawel Dabrowski-Tumanski, PhD, is post-doc at the Centre of New Technologies, University of Warsaw, Warsaw, Poland.
Pawel Rubach, PhD, is an assistant professor at the Warsaw School of Economics, Warsaw, Poland.
Wanda Niemyska, PhD, is an assistant professor at the Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland.
Bartosz Gren, is a PhD candidate at the Faculty of Physics, University of Warsaw, Warsaw, Poland.
Joanna Sulkowska, PhD DSc, is the head of the Interdisciplinary Laboratory of Biological Systems Modelling at the Centre of New Technologies, University of Warsaw, Warsaw, Poland.
Submitted: 14 May 2020; **Received (in revised form):** 15 July 2020

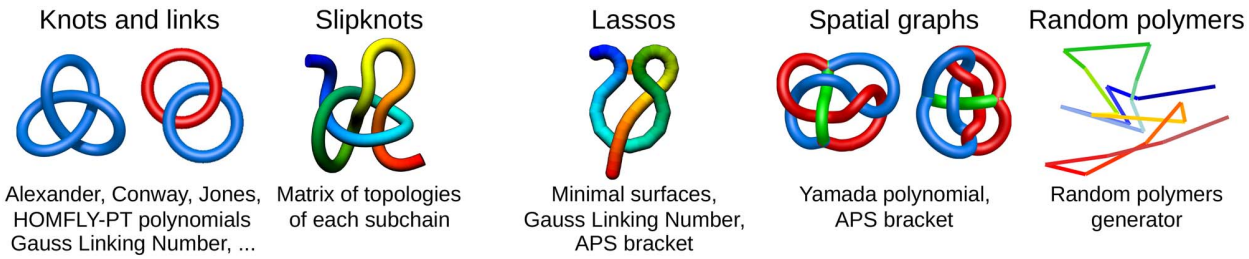


Figure 1. The exemplary structures, which can be identified using the Topoly package, and the methods suitable for their detection (below the schemes). The schemes present (from left to right) 3_1 knot, Hopf link, 3_1 slipknot, L_1 lasso, $\theta 5_1 \theta$-curve, $H6_1$ handcuff graph and a random polymer with 20 steps.

of the whole library of polymer structures, use non-standard methods or analyze new topologies (such as branched polymers). To address this need, we have developed a flexible and powerful tool—the Topoly package (https://topoly.cent.uw.edu.pl).

Topoly is a Python3 package, allowing one to identify and analyze any polymers' topology studied so far. Apart from standard knot-determining techniques (Alexander, Conway, Jones and HOMFLY-PT polynomials [2, 34]) it includes less-known methods (Kauffman and APS Brackets, Kauffman and BLM/Ho polynomials [11, 20, 30, 35, 36]), a polynomial for the analysis of spatial graphs (Yamada polynomial [65]), minimal surface analysis to identify the lasso topology [49] or the Gauss linking number (GLN) [46, 47, 56]. Exemplary structures, which can be analyzed with Topoly, are presented in Figure 1. To study statistics, or simply to test the behavior of the functions provided, the user may also generate random loops and lassos (tadpoles), θ-curves, two-component links and handcuff graphs (dumbbells).

Apart from self-generated structures, Topoly as a versatile tool accepts structures provided as a set of coordinates in various formats (XYZ, PDB, mmCIF, Mathematica and other) or as an abstract code (Planar Diagram or Ewing–Millett [24] code). For open structures (such as protein chains), the package provides the option of chain closure with various methods [48, 53]. As usual in case of a Python package, Topoly's results can be easily captured and parsed further according to the needs of the user. Topoly comes with a thorough manual (including the description of all functions) and a tutorial project (https://github.com/ilbsm/topoly_tutorial) that includes real-life examples of its usage.

Getting started

Topoly is available for Python3 running on 64-bit Linux or Mac OS X. It is distributed using the standard Python package manager—PyPI—and can be easily installed by invoking *pip install topoly*, provided a recent version of the *pip* installer is used.

The easiest way to get started with Topoly is to download or clone the tutorial project that we provide at https://github.com/ilbsm/topoly_tutorial and follow the three-step instructions in the README.

Topoly features

Input files and format translation

The main function of Topoly is to analyze the topology of the structure given either as 3D coordinates (PDB, mmCIF, XYZ, Mathematica formats) or as an abstract code (Planar Diagram or Ewing–Millett [24] code). In case the structure is given as PDB or mmCIF file (e.g. directly from the RCSB database), the coordinates of C_α atoms are extracted. The user may also supply the

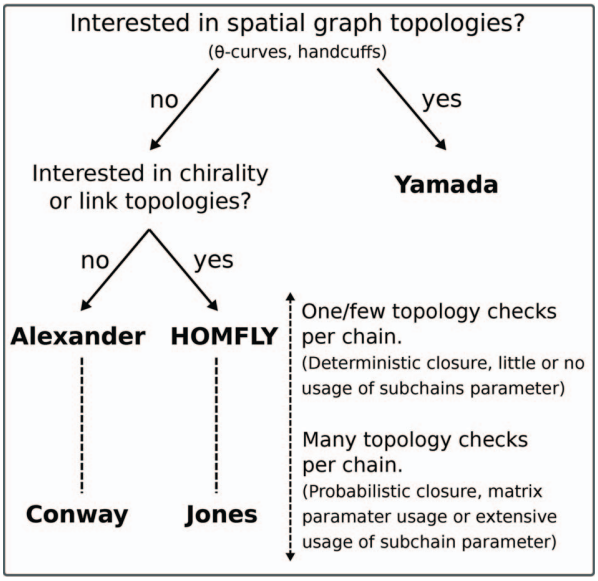


Figure 2. Decision tree with our proposition of choosing optimal invariant for topology check. We recommend usage of other invariants only for more specific goals, e.g. identification of topologies with more than eight crossings and theoretical research of topological invariants.

spatial graphs (with branching points) with each arc separated by a line without coordinates.

The coordinates and the codes for the structures are stored in the memory and can be written out in the desired format. In particular, this option can be used to translate between the formats, e.g. PDB → Mathematica, or to obtain the abstract representation, e.g. Mathematica → PD code. Currently, the recreation of the coordinates (PD → Mathematica) is not possible, if the coordinates were not given.

Knot and link analysis

The knots and links are in general well defined for closed components. For such structures, the user may use the Alexander, Conway, Jones, HOMFLY-PT polynomial, and Kauffman Bracket but also less-known Yamada, Kauffman and BLM/Ho polynomials, and APS Bracket, which, to our best knowledge, are not implemented anywhere else (Table 1). In Figure 2 is our proposition when to use each invariant. By default, the result of each function is the knot type (e.g. 3_1) corresponding to the structure analyzed. The knot type is obtained by comparing the polynomial value with the local polynomial library, containing all knots and links (prime and composite) with up to eight crossings. The user may also request the chirality of the knot and link using the

Table 1. Table of available invariants and their capabilities of distinguishing different chiralities, links, and spatial graphs. They are sorted by relative speed when running with deterministic closure and no subchain analysis. Note that Conway and Jones polynomials are relatively faster, when using probabilistic closure or during extensive subchain analysis

Invariant	Speed rank	Chirality	Links	Spatial graphs
Alexander polynomial	1st	no	no	no
HOMFLY polynomial	2st	yes	yes	no
Conway polynomial	3rd	no	no	no
Jones polynomial	3rd	yes	yes	no
Yamada polynomial	4th	yes	yes	yes
BLMHo polynomial	4th	no	yes	no
Kauffman polynomial	4th	yes	yes	no
Kauffman bracket	5th	yes	yes	no
APS bracket	5th	yes	yes	yes

keyword argument *chiral=True*. In such case, the structure type is returned with a sign (e.g. $+3_1$) if possible, where the Doll–Hoste [22] convention is used for links.

Instead of the link type, the user may request the original, untranslated polynomial value (e.g. $q + q^3 - q^4$) with the keyword argument *translate=False*. This option may be used, e.g. to calculate the polynomial for more complicated structures (e.g. with nine crossings). Such a result can be then stored in a user-created (external) dictionary of polynomials and included in further calculations.

In the case of open-chain structures (such as proteins), the chain usually needs to be closed first in order to analyze its topology. The Topoly package provides several closure methods, including stochastic closure, in which the topology is calculated for many random closures (more examples are discussed below). In such cases, the result of the calculation is a dictionary with the obtained types of topologies and their probabilities (e.g. {'3_1': 0.6, '4_1': 0.3, '5_2': 0.1}).

Spatial graphs

One of the key features of the Topoly package is the possibility to analyze the topology of spatial graphs, i.e. polymers with branching points. Such structures appear naturally in biology (e.g. glycogen, cystine knots and cyclotides, lasso proteins and peptides [29, 44, 49] or replicating cyclic DNA [1, 52]). The analysis of such structures may be performed using, e.g. the Yamada polynomial or APS bracket, to the best of our knowledge implemented solely in the Topoly package. In particular, one can analyze the topology of θ-curves or handcuff graphs (Figure 1). The Topoly package contains a local library of Yamada polynomials for all prime and composite θ-curves as well as handcuff graphs with up to seven crossings.

Spanning minimal surface and detection of surface piercings

Another feature of the Topoly package is the ability to construct a triangulated minimal surface spanning a closed loop. The coordinates of the surface-forming triangles may be exported or used to analyze piercings through the loop (minimal surface method). This technique was used to analyze polymer rings in melt [58] and to analyze the organization of chromatin domains [45] or to identify complex lasso proteins [49], where the tail pierces the loop closed by a disulfide bridge (Figure 1). To perform such analysis, the user has to specify the indices of the loop closing bridge. In the case of PDB or mmCIF files, the bridges are recognized automatically. By default, only the disulfide bridges are taken into account but the user may also include other covalent bridges (e.g. amide) or the bridges mediated by ions.

Alternatively, the piercings through the loop may be analyzed using the GLN. Originally, the linking number counts how many times one loop pierces the other. It was initially defined only for a closed loop; however, its integral definition may be also extended to open structures [9]. In such a case, fractional values are obtained and the significantly nonzero value of the GLN should indicate a piercing [50]. The user may analyze either two parts of the same chain (specified by the indices) or two separate chains. The GLN was already used to study the conformation and entanglement of proteins [7, 8, 13, 27, 46, 47] and chromosomes [60].

Analysis of subchains

Apart from the analysis of the whole structure, in case of the knotted chain and GLN approach, one can also analyze the entanglement of each subchain. As the subchain is specified by two indices—begin and end—the information about the topology of each subchain can be encoded in a two-dimensional matrix, indexed by the beginning and the end of the subchain. Such an approach in the case of knots is called the knot fingerprint matrix (Figure 3) [37, 61]. The representation of proteins' chains in the form of a matrix leads to the discovery of slipknots (the overall unknotted structures which have a nontrivial subchain) and a strict conservation of complex knotting patterns (compositions of slipknots) within and between several protein families, despite their large sequence divergence [61]. The GLN matrix was used, e.g. to spot high winding of the chain around closed loops [50].

For the GLN and each knot polynomial function, the matrix analysis may be invoked by using the keyword *matrix=True*. By default, the matrix is presented as a dictionary indexed by each subchain. The user may, however, plot the matrix and save in a desired format (PNG, SVG, PDF).

Structure importing, generating and finding

Apart from the library of polynomial values, the Topoly package includes also a library of planar diagram codes for knots/links with up to eight crossings and θ-curves and handcuff graphs for up to seven crossings. Therefore, the user has a vast library for his/her own analysis.

Alternatively, Topoly allows sampling of random structures. This option may be useful to study the statistics of random chains. Topoly includes the possibility to sample random walks, loops, lassos, θ-curves, two-component links and handcuff graphs (Figure 1), generated using Cantarella *et al.*'s [12]

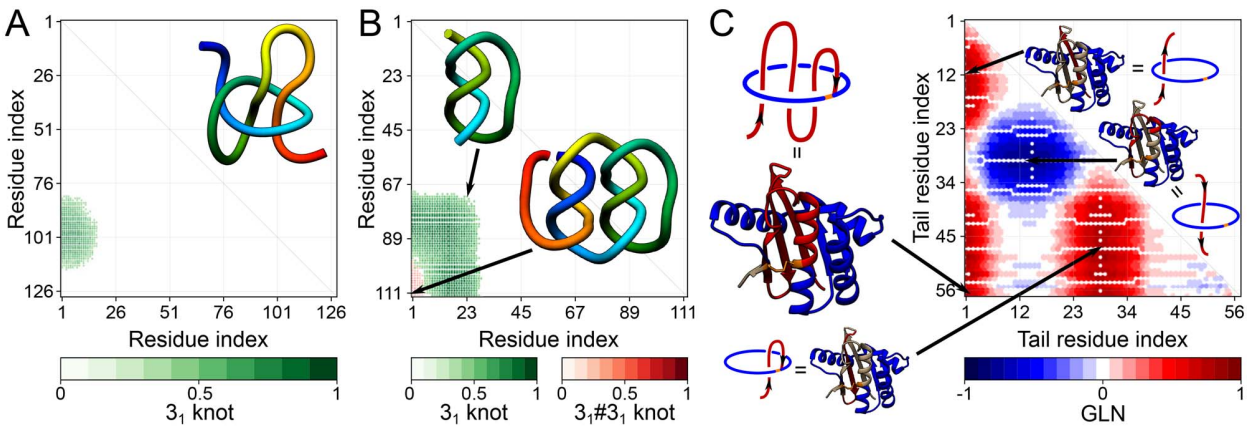


Figure 3. The exemplary matrices obtained with Alexander polynomial (left and middle) and GLN from Topoly package. Structures analyzed from left to right: 3_1 slipknot, composite knot $3_1\#3_1$ formed by artificial structures and L_3 lasso formed by hyperthermophilic archaeal RNase HI (PDB code 2ehg). The visualization of the analyzed structures was placed in the top-right corners of the matrices. In case of the 2egh structure (right panel) the loop (in blue) is closed by a cysteine bridge (orange) between Cys58-Cys145 and pierced by the N-terminal tail (red).

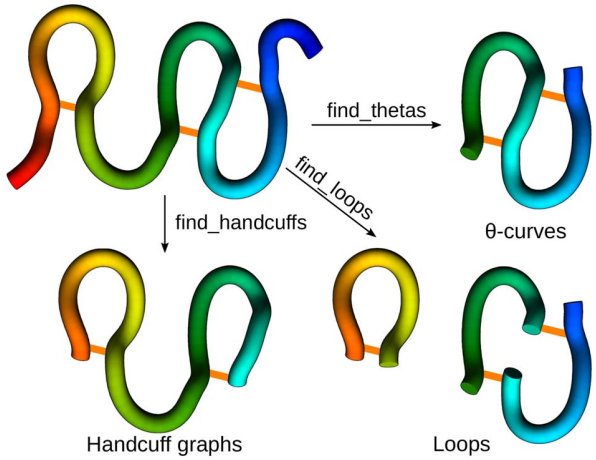


Figure 4. The exemplary results of finding functions implemented in Topoly. For the exemplary structure (top-left) with three disulfide bonds (marked as orange stripes) exemplary results of *find_loops*, *find_thetas* and *find_handcuffs* are presented. In total, four loops, one θ-curve and eight handcuffs were found in this case.

algorithm. The structures may be presented as a Python list or the coordinates may be saved to separate files.

On the other hand, the user may analyze complex, multi-branched polymers (like proteins with disulfide bridges), extracting loops, θ-curves, links and handcuff graphs using the appropriate *find_* function (Figure 4). Such an approach was already used to find θ-curves in proteins [14].

Finally, for a given matrix fingerprint, the user may also identify the representative subchains, which may differ in topology. Such analysis is currently used in the KnotProt database to prescribe the knot fingerprint, as a concatenation of topological types of each representative subchain (e.g.$+6_1 + 6_14_1 + 3_1$ in the case of an α-haloacid dehalogenase protein). In Topoly, the representative subchains with their topologies are obtained using the *find_spots* function.

Programmed speedup

Due to the run-time complexity of some algorithms (e.g. calculating the topology of every possible subchain), the Topoly calculation is accelerated on various levels of the implementation. On the algorithmic level, the sampling of subchains is optimized. The knot fingerprint matrix calculation is divided into two parts. In the first one, only a subset of points is calculated. The size of the subset is controlled by the *density* parameter (the higher the value, the less points are calculated in the first step). The surroundings of these points are calculated in the second step only if a nontrivial topology was identified in the first step. Moreover, the planar diagram codes of the structures are compared with the local library; therefore, only the yet unknown structures are calculated. Furthermore, the calculation of the matrix is done in parallel (governed by the keyword *run_parallel*) using all the computational power available (unless the user limits the number of cores by setting the value of *parallel_workers*). Finally, for the Alexander polynomial (dedicated to matrix calculation) the analysis may be also done using the GPU units that support the CUDA technology.

Other options

Apart from the topological analysis, the Topoly package provides some auxiliary functions helpful in the analysis. In particular, for an open structure, the user may close it (with *close_curve* function) using several methods, including closure on a big sphere surrounding the structure, extending rays in one direction or extending the termini from the center of the mass (Figure 5). The closed structure may be then reduced (*reduce_structure* function) with the implementation of the Koniaris–Muthukumar–Taylor algorithm [40, 63] or by a chain of Reidemeister moves. To visualize the effect of chain closure or reduction, one may use the function *plot_graph* which opens the Matplotlib window in which the structure is shown (Figure 5).

For a structure in a given format (e.g. Mathematica-generated file), Topoly allows one to translate it into different supported formats (e.g. C_α PDB file) using the *translate_code* function. In particular, this way one can obtain, e.g. the planar diagram code of a 3D structure.

By default, the obatined matrix fingerprint is presented as a Python dictionary. As the user may want to plot the result on their own, it may be translated to a string (KnotProt format) or a list applicable to Matplotlib or Gnuplot using the *translate_matrix* function.

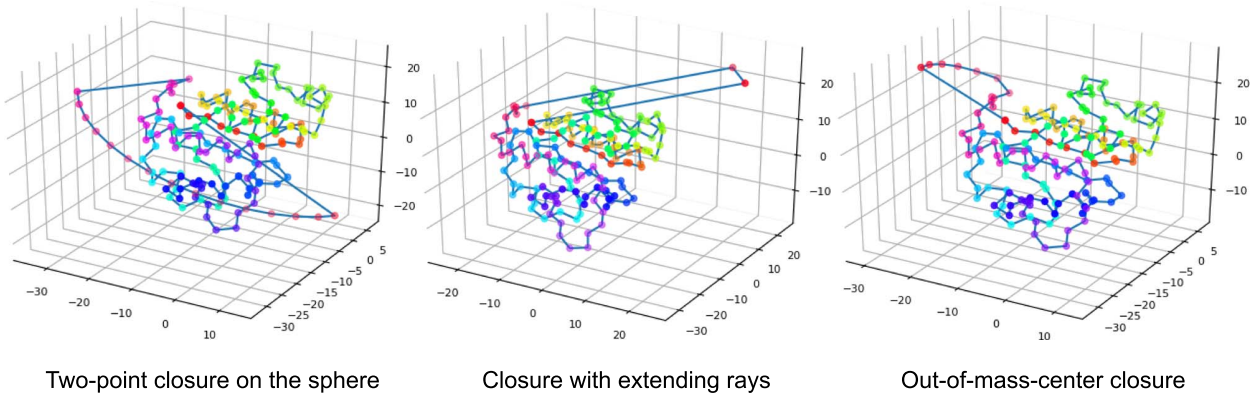


Figure 5. The exemplary closure methods. From left to right—two-point closure on the sphere (two points are randomly chosen on a large sphere and connected with an arc visible in the front), extending rays in one direction, extending the termini from the center of the mass, and connected with an arc visible in the front. In all three panels, the same structure (with PDB code 1j85) was closed. The structures were drawn with the Topoly implemented function *plot_graph*.

Finally, the user may also use Topoly to find a match for a given polynomial in the local library using the *find_matching* function. This may be used to find the link chirality if the appropriate polynomial was calculated with some external software.

Exemplary cases

We will present four exemplary cases.

(i) Calculation of knot topology for UCH protein (PDB code 2len, chain A, model 2) with the Conway polynomial.

```
>>>from topoly import conway, Closure
>>># statistics of knots for different closures
>>> print(conway('2len.pdb', pdb_chain='A',pdb_model=2))
{'5_2': 0.61, '0_1': 0.22, '3_1': 0.035, '4_1': 0.035, 'Unknown': 0.030, '5_1': 0.025, '7_6': 0.015, '6_1': 0.01, '7_3': 0.01, '7_4': 0.005, '8_4': 0.005}
>>> # the explicit Conway polynomial for closed structure
>>> print(conway('2len.pdb', translate=False, poly_reduce=False, closure=Closure.MASS_CENTER))
    1 + 2z^2
```

z We assume the file '2len.pdb' is stored locally during the analysis after having been downloaded from the RCSB server. We explicitly write the parameters. First, the statistics of knots are obtained. In particular, the 5_2 knot is obtained with a probability of 0.61 (in 61% of closures, Figure 6A). In the last step, we calculate the polynomial value $(1 + 2z^2)$ for the structure closed with arms extended from the center of mass. We do not translate the polynomial value to a matching knot type (5_2). This analysis shows the diversity of the detected topology which is a consequence of using a statistical approach to close the chain as well as of the internal complexity of the protein backbone.

(ii) Plotting and analyzing the knot fingerprint matrix for UCH37 (PDB code 4i6n).

```
>>> from topoly import alexander, homfly, find_spots
>>> # calculating and plotting the matrix
matrix_result = alexander('4i6n.pdb', matrix_map=True, map_arrows=False)
# finding the spot centers
spots = find_spots(matrix_result, map_cutoff=0.36)
>>> print(spots)
'5_2': [(2, 228)], '3_1': [(5, 173), (6, 211)]
>>> # establishing chirality
>>> for knot in spots.keys():
... for spot in spots.get(knot):
... print(homfly('4i6n.pdb', chain_boundary=[spot], chiral=True, closure=Closure.MASS_CENTER))
(2, 228) -5_2
(5, 173) -3_1
(6, 211) -3_1
```

We assume the file '4i6n.pdb' is stored locally during analysis and the indices are renumbered to start from 1. The matrix is calculated with the Alexander polynomial (the fastest, devoted to matrix calculations). Next, the spots (representative subchains) are identified (with indices 2–228 for 5_2 knot, 6–211 for 3_1 slipknot and 5–173 for the most deeply embeded 3_1 slipkot; Figure 6B). In the last step, we analyze the chirality of each subchain found in *find_spots* step. It turns out that the representative subchain with the 3_1 topology has the same chirality. As a consequence, the composition of these commands allows one to determine the fundamental topological properties of proteins, such as depth of knots and slipknots, their locations, as well as to compare the entanglement between different proteins. Similar properties in the case of lassos can be determined based on the GLN matrix.

(iii) Calculating the lasso type for all the loops in cerato-platanin (PDB code 3sum).

```
>>> from topoly import lasso_type
>>># gathering information about lassos
>>> lassos = lasso_type('3sum.pdb', more_info=True)
>>> for lasso in lassos.keys():
```

```
... loop = lasso
... motif = lassos[lasso]['class']
... piercing_N = lassos[lasso]['crossingsN']
... piercing_C = lassos[lasso]['crossingsC']
... print(loop, motif, piercing_N + piercing_C)
(43, 80) L+1C ['+97']
(83, 145) L+1N ['+72']
(144, 153) L0 []
```

We assume the file '3sum.pdb' is stored locally during the analysis. We calculate the lasso types and in the last step, we print the result, loop by loop, including the loop-delimiting indices, the topology of the lasso and the indices of piercing residues. The Mathematica visualization of surface files obtained for nontrivial lassos are shown in Figure 6C. The composition of such commands allows for a comprehensive analysis of the lasso type topology—location of the loop and piercings and depth of piercings. Both parameters are important to study kinetics pathways [28, 54], stability [38] and statistical and biological properties [15].

(iv) Generation of loops for statistics.

```
>>> from topoly import generate_loop, jones
>>> # for statistic analysis
>>> from collections import Counter
>>> # generating loops
>>> loops = generate_loop(100, 1000, output='list')
>>> # establishing the topology of the loops
>>> result = [jones(loop, closure=Closure.CLOSED) for
loop in loops]
>>> # printing the statistics
>>> print(Counter(result)) Counter({'0_1': 700, '3_1':
170, '4_1': 27, '3_1#3_1': 23,...
```

We generate 1000 loops with 100 nodes each and calculate the corresponding knot statistics. For a better visualization, the results summarized using extarnal tool—Counter. The results are presented in Figure 6D. Such analysis can be useful in the comparison of statistical properties between, e.g. proteins and random polymers.

Test project and package manual

The further growing, base of examples can be found in the tutorial project, available at https://github.com/ilbsm/topoly_tutorial. All the functions are described extensively in the manual, present at https://topoly.cent.uw.edu.pl

Tool architecture

The Topoly package contains a set of libraries, executables and Python3 modules. A significant part of the available algorithms is implemented in C or C++ and wrapped using Cython with Python3 code. Some algorithms, however, are directly implemented in Python.

Requirements

The Topoly PyPI package contains both Python code as well as executable binaries and compiled shared libraries written in C and C++. Some algorithms have two implementations; one of which is faster but requires a CUDA compatible GPU and the CUDA framework version 8.0 or newer. The Topoly PyPI package is compatible with Linux systems starting with CentOS 6 (following the ManyLinux2010 specification) and with Mac OS starting with 10.9 Mavericks, making most Linux and Mac OS X supported.

Both Linux and Mac OS X packages are built for Python3 (3.5, 3.6, 3.7 and 3.8) and require the following dependent packages: biopython, cycler, kiwisolver, matplotlib, numpy, pyparsing, python-dateutil, scipy, six. In the case Topoly is installed using pip, these dependencies should be installed automatically.

Package structure

The Topoly package contains

- Python modules that should be installed to the relevant Python3 modules location: /usr/local/lib/python3.x/site-packages (administrator installation) or $USER/.local/lib-/python3.x/site-packages (user installation),
- executables and shared libraries that should be installed in /usr/local/bin and /usr/local/lib or *$USER/.local/bin* and $USER/local/lib,
- documentation and test examples available in /usr/local/share/doc/topoly or $USER/.local/share/doc/topoly.

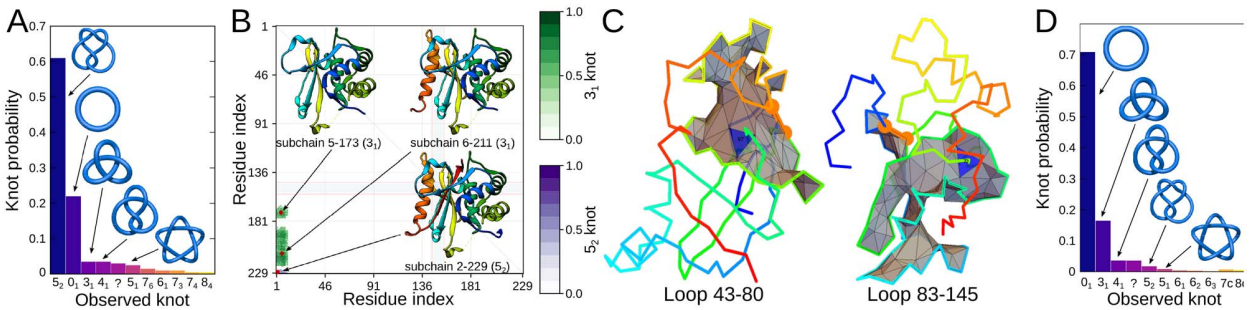


Figure 6. The graphical result of the exemplary cases. (A) The histogram of knots obtained for different closures of the UCH protein (PDB code 2len, test case 1). (B) The knot fingerprint matrix obtained for UCH37 protein (PDB code 4i6n, test case 2). (C) The surfaces spanned on the main chain of cerato-platanin (PDB code 3sum) closed by bridges 43–80 (left panel) and 83–145 (right panel) with the piercings indicated by the blue triangle (test case 3). The third bridge results in a trivial lasso. The surfaces are visualized with Mathematica. (D) The histogram of different knot types obtained for 1000 random loops with the length 100 (test case 4).

PyPI packages can also be installed in Python virtual environments such as venv or virtualenv. In that case, Python modules, executables and libraries will be found in folders relative to the main directory of the virtual environment.

Conclusions

In this work, we presented the Topoly Python3 package—a tool which allows for calculating the topology of linear and branched polymers. To make it most simple for even inexperienced users we enhanced it with the possibility to read common file formats (including PDB and mmCIF). The package includes all the tools used so far to analyze the topology of proteins and many more additions. According to our knowledge, it is the only tool allowing for the calculation of the Yamada polynomial for the analysis of spatial graphs. To facilitate the statistical analysis, we also enhanced the package with methods for the generation of random polymers and the identification of such structures in complex spatial graphs. The tools are optimized to enable the screening of many structures (e.g. conformations of a protein during folding) or very long biopolymers (e.g. chromatin), as well as the conduction of a detailed analysis of the knot fingerprint.

Therefore, we hope that the versatility of the package will be appreciated by many users with different background, allowing for a fundamental topological analysis, as well as the determination of parameters important to understand the biological function or stability (e.g. depth of a knot, lasso, kinetics pathways). Topoly should also help in other applications including the designing of new entanglements in investigated biopolymers (new branch points, location of closed loops, etc.).

Key Points

- Comprehensive and easy to use Python3 package enables the distinguishing of knots, slipknots, links and spatial graphs through the calculation of different topological polynomial invariants.
- It enables the creation of the minimal spanning surface on a given loop, e.g. to detect a lasso motif or to generate random closed polymers.
- Analyzes XYZ, PDB, mmCIF and Mathematica file formats and Python lists, PD codes, and EM codes.

Package manual

Package manual is present at https://topoly.cent.uw.edu.pl

Acknowledgments

We would like to thank Kenneth C. Millett, Eric J. Rawdon and Andrzej Stasiak for their help in developing some of the methods and all remarks.

Funding

National Science Centre (2017/24/T/NZ1/00490 to P.D.T., 2012/07/E/NZ1/01900 to J.I.S.), European Molecular Biology Organization Installation (2057 to J.I.S.) and the Polish Ministry for Science and Higher Education (0003/ID3/2016/64 to J.I.S.).

References

1. Adams DE, Shekhtman EM, Zechiedrich EL, *et al*. The role of topoisomerase IV in partitioning bacterial replicons and the structure of catenated intermediates in DNA replication. *Cell* 1992;**71**(2):277–88.
2. Alexander JW. Topological invariants of knots and links. *Trans Amer Math Soc* 1928;**30**(2):275–306.
3. Aoki D, Takata T. Mechanically linked supramolecular polymer architectures derived from macromolecular [2]rotaxanes: synthesis and topology transformation. *Polymer* 2017;**128**:276–96.
4. Arsuaga J, Vazquez M, McGuirk P, *et al*. DNA knots reveal a chiral organization of DNA in phage capsids. *Proc Natl Acad Sci* 2005;**102**(26):9165–9.
5. Ayme J-F, Beves JE, Campbell CJ, *et al*. Template synthesis of molecular knots. *Chem Soc Rev* 2013;**42**(4):1700–12.
6. Ayme J-F, Beves JE, Leigh DA, *et al*. A synthetic molecular pentafoil knot. *Nat Chem* 2012;**4**(1):15.
7. Baiesi M, Orlandini E, Seno F, *et al*. Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding. *Sci Rep* 2019;**9**(1):1–12.
8. Baiesi M, Orlandini E, Trovato A, *et al*. Linking in domain-swapped protein dimers. *Sci Rep* 2016;**6**(1):1–11.
9. Banchoff T. Self linking numbers of space polygons. *Indiana Univ Math J* 1976;**25**(12):1171–88.
10. Bölinger D, Sułkowska JI, Hsu H-P, *et al*. A Stevedore's protein knot. *PLoS Comput Biol* 2010;**6**(4):e1000731.
11. Brandt RD, Lickorish WBR, Millett KC. A polynomial invariant for unoriented knots and links. *Invent Math* 1986;**84**(3):563–73.
12. Cantarella J, Duplantier B, Shonkwiler C, *et al*. A fast direct sampling algorithm for equilateral closed polygons. *J Phys A* 2016;**49**(27):275202.
13. Caraglio M, Micheletti C, Orlandini E. Physical links: defining and detecting inter-chain entanglement. *Sci Rep* 2017;**7**(1):1156.
14. Dabrowski-Tumanski P, Goundaroulis D, Stasiak A, *et al*. θ-curves in proteins. arXiv preprint arXiv:1908.05919, 2019.
15. Dabrowski-Tumanski P, Gren B, Sulkowska JI. Statistical properties of lasso-shape polymers and their implications for complex lasso proteins function. *Polymers* 2019;**11**(4):707.
16. Dabrowski-Tumanski P, Jarmolinska AI, Niemyska W, *et al*. LinkProt: a database collecting information about biological links. *Nucleic Acids Res* 2017;**45**(D1):D243–D249.
17. Dabrowski-Tumanski P, Niemyska W, Pasznik P, *et al*. LassoProt: server to analyze biopolymers with lassos. *Nucleic Acids Res* 2016;**44**(W1):W383–9.
18. Dabrowski-Tumanski P, Rubach P, Goundaroulis D, *et al*. KnotProt 2.0: a database of proteins with knots and other entangled structures. *Nucleic Acids Res* 2018;**47**(D1):D367–75.
19. Dabrowski-Tumanski P, Sulkowska JI. Topological knots and links in proteins. *Proc Natl Acad Sci USA* 2017;**114**(13):3415–20.
20. Dabrowski-Tumanski P, Sulkowska JI. The aps-bracket—a topological tool to classify lasso proteins, RNAs and other tadpole-like structures. *React Funct Polym* 2018;**132**:19–25.
21. Daly NL, Craik DJ. Bioactive cystine knot proteins. *Curr Opin Chem Biol* 2011;**15**(3):362–8.
22. Doll H, Hoste J. A tabulation of oriented links. *Math Comp* 1991;**57**(196):747–61.
23. Dorier J, Goundaroulis D, Benedetti F, *et al*. Knoto-ID: a tool to study the entanglement of open protein chains using the concept of knotoids. *Bioinformatics* 2018;**34**(19):3402–4.

24. Ewing B, Millett KC. A load balanced algorithm for the calculation of the polynomial knot and link invariants. In: *The Mathematical Heritage of CF Gauss*. Singapore: World Scientific, 1991, 225–66.
25. Gierut AM, Dabrowski-Tumanski P, Niemyska W, *et al.* PyLink: a PyMOL plugin to identify links. *Bioinformatics* 2019;**35**(17):3166–8.
26. Gierut AM, Niemyska W, Dabrowski-Tumanski P, *et al.* PyLasso: a PyMOL plugin to identify lassos. *Bioinformatics* 2017;**33**(23):3819–21.
27. Grønbæk C, Hamelryck T, Røgen P. GISA: using Gauss Integrals to identify rare conformations in protein structures. *PeerJ* 2020;**8**:e9159.
28. Haglund E, Sułkowska JI, He Z, *et al.* The unique cysteine knot regulates the pleotropic hormone leptin. *PLoS One* 2012;**7**(9):e45654.
29. Hegemann JD, Zimmermann M, Xie X, *et al.* Lasso peptides: an intriguing class of bacterial natural products. *Acc Chem Res* 2015;**48**(7):1909–19.
30. Ho CF. A polynomial invariant for knots and links—preliminary report. *Abstracts Amer Math Soc* 1985;**6**:300.
31. Jablan SV, Radmila S. *LinKnot: Knot Theory by Computer*, Vol. **21**. Singapore: World Scientific, 2007.
32. Jamroz M, Niemyska W, Rawdon EJ, *et al.* KnotProt: a database of proteins with knots and slipknots. *Nucleic Acids Res* 2014;**43**(D1):D306–14.
33. Jarmolinska AI, Gambin A, Sulkowska JI. Knot_pull—python package for biopolymer smoothing and knot detection. *Bioinformatics* 2020;**36**(3):953–55.
34. Jones VFR. A polynomial invariant for knots via von Neumann algebras. In: *Fields Medallists' Lectures*. Singapore: World Scientific, 1997, 448–58.
35. Kauffman LH. State models and the Jones polynomial. *Topology* 1987;**26**(3):395–407.
36. Kauffman LH. An invariant of regular isotopy. *Trans Amer Math Soc* 1990;**318**(2):417–71.
37. King NP, Yeates EO, Yeates TO. Identification of rare slipknots in proteins and their implications for stability and folding. *J Mol Biol* 2007;**373**(1):153–66.
38. Knappe TA, Linne U, Robbel L, *et al.* Insights into the biosynthesis and stability of the lasso peptide capistruin. *Chem Biol* 2009;**16**(12):1290–8.
39. Kolesov G, Virnau P, Kardar M, *et al.* Protein knot server: detection of knots in protein structures. *Nucleic Acids Res* 2007;**35**(Suppl_2):W425–8.
40. Koniaris K, Muthukumar M. Knottedness in ring polymers. *Phys Rev Lett* 1991;**66**(17):2211.
41. Lai Y-L, Chen C-C, Hwang J-K. pknot v.2: the protein knot web server. *Nucleic Acids Res* 2012;**40**(W1):W228–31.
42. Lai Y-L, Yen S-C, Yu S-H, *et al.* pknot: the protein knot web server. *Nucleic Acids Res* 2007;**35**(Suppl_2):W420.
43. Lua RC. PyKnot: a PyMOL tool for the discovery and analysis of knots in proteins. *Bioinformatics* 2012;**28**(15):2069–71.
44. Maksimov MO, Pan SJ, Link AJ. Lasso peptides: structure, function, biosynthesis, and engineering. *Nat Prod Rep* 2012;**29**(9):996–1006.
45. Michieletto D, Orlandini E, Marenduzzo D. Polymer model with epigenetic recoloring reveals a pathway for the de novo establishment and 3D organization of chromatin domains. *Phys Rev X* 2016;**6**(4):041047.
46. Millett KC. Knotting and linking in macromolecules. *React Funct Polym* 2018;**131**:181–90.
47. Millett KC. Topological linking and entanglement in proteins. In: *Topology and Geometry of Biopolymers*, Vol. 746. Northeast University, Boston, Massachusetts: American Mathematical Society, 2020, 201.
48. Millett KC, Rawdon EJ, Stasiak A, *et al.* Identifying knots in proteins. *Biochem Soc Trans* 2013;**41**(2):533–7.
49. Niemyska W, Dabrowski-Tumanski P, Kadlof M, *et al.* Complex lasso: new entangled motifs in proteins. *Sci Rep* 2016;**6**:36895.
50. Niemyska W, Millett KC, Sulkowska JI. GLN—a method to reveal unique properties of lasso type topology in proteins. ArXiv preprint, 2020.
51. Niewieczerzał S, Sulkowska JI. Supercoiling in a protein increases its stability. *Phys Rev Lett* 2019;**123**(13):138102.
52. ODonnol D, Stasiak A, Buck D. Two convergent pathways of DNA knotting in replicating DNA molecules as revealed by θ-curve analysis. *Nucleic Acids Res* 2018;**46**(17):9181–8.
53. Perego C, Potestio R. Computational methods in the study of self-entangled proteins: a critical appraisal. *J Phys Condens Matter* 2019;**31**(44):443001.
54. Perego C, Potestio R. Searching the optimal folding routes of a complex lasso protein. *Biophys J* 2019;**117**(2):214–28.
55. Perret-Aebi L-E, von Zelewsky A, Dietrich-Buchecker C, *et al.* Stereoselective synthesis of a topologically chiral molecule: the trefoil knot. *Angew Chem Int Ed* 2004;**43**(34):4482–5.
56. Ricca RL, Nipoti B. Gauss' linking number revisited. *J Knot Theory Ramifications* 2011;**20**(10):1325–43.
57. Siebert J, Kivel A, Atkinson L, *et al.* Are there knots in chromosomes? *Polymers* 2017;**9**(8):317.
58. Smrek J, Grosberg AY. Minimal surfaces on unconcatenated polymer rings in melt. *ACS Macro Lett* 2016;**5**(6):750–4.
59. Sogo JM, Stasiak A, Martinez-Robles ML, *et al.* Formation of knots in partially replicated DNA molecules. *J Mol Biol* 1999;**286**(3):637–43.
60. Sulkowska JI, Niewieczerzal S, Jarmolinska AI, *et al.* KnotGenome: a server to analyze entanglements of chromosomes. *Nucleic Acids Res* 2018;**46**(W1):W17–24.
61. Sułkowska JI, Rawdon EJ, Millett KC, *et al.* Conservation of complex knotting and slipknotting patterns in proteins. *Proc Natl Acad Sci* 2012;**109**(26):E1715–23.
62. Takata T, Aoki D. Topology-transformable polymers: linear-branched polymer structural transformation via the mechanical linking of polymer chains. *Polym J* 2018;**50**(1):127–47.
63. Taylor WR. A deeply knotted protein structure and how it might fold. *Nature* 2000;**406**(6798):916.
64. Tubiana L, Polles G, Orlandini E, *et al.* KymoKnot: a web server and software package to identify and locate knots in trajectories of linear or circular polymers. *Eur Phys J E* 2018;**41**(6):72.
65. Yamada S. An invariant of spatial graphs. *J Graph Theory* 1989;**13**(5):537–51.
66. Yamamoto T, Tezuka Y. Topological polymer chemistry: a cyclic approach toward novel polymer properties and functions. *Polym Chem* 2011;**2**(9):1930–41.
67. Zhao Y, Chwastyk M, Cieplak M. Structural entanglements in protein complexes. *J Chem Phys* 2017;**146**(22):225102.