Macromolecules

Open Access

pubs.acs.org/Macromolecules

Article

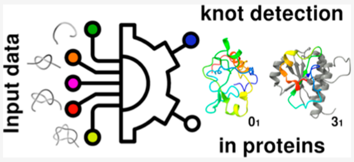
Knots and θ-Curves Identification in Polymeric Chains and Native Proteins Using Neural Networks

Fernando Bruno da Silva, Boštjan Gabrovšek, Marta Korpacz, Kamil Luczkiewicz, Szymon Niewieczerzal, Maciej Sikora, and Joanna I. Sulkowska*

Cite This: *Macromolecules* 2024, 57, 4599−4608

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Entanglement in proteins is a fascinating structural motif that is neither easy to detect via traditional methods nor fully understood. Recent advancements in AI-driven models have predicted that millions of proteins could potentially have a nontrivial topology. Herein, we have shown that long short-term memory (LSTM)-based neural networks (NN) architecture can be applied to detect, classify, and predict entanglement not only in closed polymeric chains but also in polymers and protein-like structures with open knots, actual protein configurations, and also θ-curves motifs. The analysis revealed that the LSTM model can predict classes (up to the 6_1 knot) accurately for closed knots and open polymeric chains, resembling real proteins. In the case of open knots formed by protein-like structures, the model displays robust prediction capabilities with an accuracy of 99%. Moreover, the LSTM model with proper features, tested on hundreds of thousands of knotted and unknotted protein structures with different architectures predicted by AlphaFold 2, can distinguish between the trivial and nontrivial topology of the native state of the protein with an accuracy of 93%.

INTRODUCTION

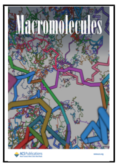
Entanglement is one of the most intriguing topological motifs found in many fields: physics, chemistry, and biology. In each field, it may have a different meaning (e.g., quantum entanglement, mechanical or thermal stability, or biological function), which can be understood after its proper classification. An example of an entanglement is knots, which appear in a daily life, such as those in shoelaces and rope. In mathematics, a knot is a closed, nonself-intersecting curve embedded in a three-dimensional space. A branch of mathematics known as knot theory provides tools to classify and describe properties of knots as well as links or θ-curves. Examples of those tools are so-called knot invariants, i.e., some relatively simple mathematical objects (e.g., numbers, polynomials, etc.), which can be assigned algorithmically to a given knot. Knot invariants remain constant under the ambient isotopy (continuous deformation) of the curve and can thus be used to distinguish or classify knots. Two knots are considered equivalent, i.e., have the same topology, if they can be transformed into each other through an ambient isotopy. More recently, artificial intelligence (AI)-based approaches have also been used to perform topological classifications, e.g., for randomly knotted curves.[1,2] Herein, we test whether AI methods can be used to detect and classify entanglements in polymers and proteins.

Among the types of entanglement found in polymers, in DNAs, and in proteins, knot motifs are the best understood[3,4] and probably the least are θ-curve motifs.[5,6] Mathematically, knots are defined on the closed curves; however, this definition can be extended to proteins which are open chains by properly connecting both ends of the backbone.[7,8] The properties of DNAs, proteins, and polymers may be affected by the complexity of their entanglements.[3,4] Furthermore, for the same number of monomers, knots are common in on-lattice polymers,[9] rare in proteins,[9,10] and only one knotted RNA structure is known.[11]

Knotted motif is present in around 1% of experimentally resolved structures.[12] The most abundant knot type is the simplest trefoil (3_1),[13] but 4_1,[14] 5_2,[15] 6_1,[16] 7_1,[17] and one complex $3_1\#3_1$ knot[18] were also found in proteins. The topological motif of θ-curves is based on the embedding of three-dimensional θ-letter-shaped curves in the protein backbone chains. Such topological motifs arise by taking into account all the covalent connections between nonadjacent amino acids (usually disulfide bonds) and ion-mediated bridges. As reported by Dabrowski-Tumanski et al.,[5] there are 52 nonredundant protein chains, in PDB, embedding nontrivial θ-curves whose architecture can be divided in seven different topologies, $\theta 3_1$, $\theta 0_1\#3_1$, $\theta 4_1$, $\theta 0_1\#4_1$, $\theta 0_1\#5_2$, $\theta 5_4$, and $\theta 8_n$.

However, the number and the topological complexity of the proteins rise if the structures predicted by AI methods are considered.[19] Although AI-type models have been used previously for 3D structure prediction, it is only recently that their results have reached a quality comparable to experimental accuracy. This breakthrough allowed AlphaFold,[20] RoseTTA-

Received: November 30, 2023
Revised: March 28, 2024
Accepted: April 2, 2024
Published: April 15, 2024



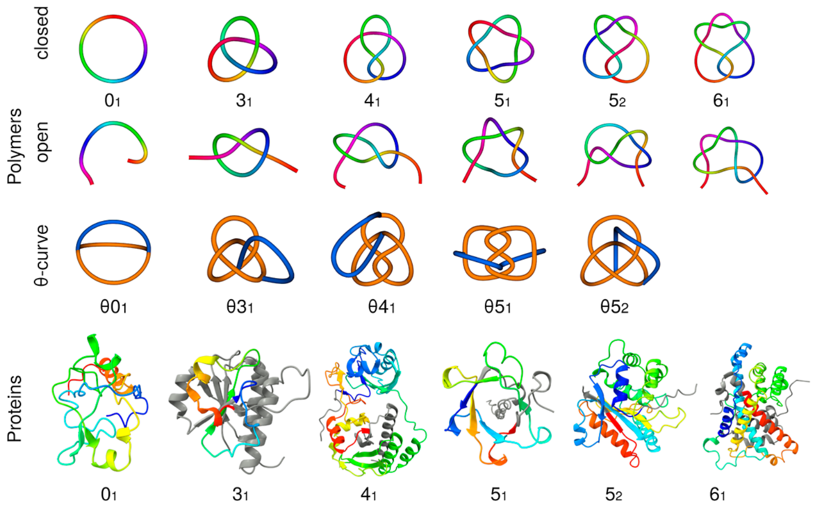


Figure 1. Knot classification of polymeric and protein structures: first and second rows present closed and open polymeric knots (0_1, 3_1, 4_1, 5_1, 5_2, and 6_1). Third row introduces the trivial and the nontrivial θ-curves ($\theta\, 0_1$, $\theta\, 3_1$, $\theta\, 4_1$, $\theta\, 5_1$, and $\theta\, 5_2$). Fourth row highlights the diverse knot classification in native proteins—unknot (0_1, AF ID: A0A0N5CFD3), trefoil knot (3_1, PDB ID: 1J85), figure-eight knot (4_1, PDB ID: 5VIK), cinquefoil knot (5_1, AF ID: P73136), three-twist knot (5_2, AF ID: A0A6P3EV80), and Stevedore knot (6_1, PDB ID: 4N2X).

Fold,[21] and Evolutionary Scale Modeling (ESMFold)[22] to collectively release millions of 3D protein structure predictions covering most of the proteomic databases. A survey of AlphaFold database featuring a majority of UniProtKB entries revealed around 700,000 potentially knotted proteins[23] including knot types not seen before in proteins, such as 5_1 knot,[19,24] complex knots such as 6_3,[25] 7_1,[19,24] 8_3,[24] and even a composite knot $3_1\#3_1$[19] (this one is already confirmed experimentally by Bruno da Silva et al.[18]). To our knowledge, no one has checked the AlphaFold Database from θ-curves point of view, but since structures deposited in the PDB represent less than 1% of all known proteins, it is to be expected that more complex θ-curves may exist.

From the other perspective, deep learning models have demonstrated exceptional abilities in recognizing and classifying patterns.[26−28] By training these models on extensive data sets containing random curve[1,2] and protein-like[29] chains with known knot types, it becomes feasible to create reliable algorithms for automating the identification and classification of knots within these intricate structures. Braghetto et al.[30] and Sleiman et al.[31] demonstrate the effectiveness of long short-term memory (LSTM)[32] based on neural network (NN) in accurately discerning knot types in highly geometrically complex entangled structures. Moreover, Sleiman et al.[31] have in fact shown indirectly that these methods can work for open curves since their approach is able to locate the knotted portion.

This raises the question of whether machine learning (ML) models can be trained to recognize knot types in simulated polymers that closely resemble the protein configurations found in nature. More advanced question is whether these techniques can be applied to open knots or other types of topological motifs, such as θ-curves,[5] bonded knots,[33,34] knotoids,[35,36] or lassos,[37] structures also found in proteins and other biopolymers.

To determine the topology of a protein, one can close the open curves by deterministic (e.g., by connecting the two terminals, i.e., the first and the last α-carbons, directly with a straight line)[9] or nondeterministic means,[7] or alternatively, using a knotoid approach.[38] The nondeterministic approach is also called probabilistic or random closure. In such a case, we enclose the entire chain in a sphere centered on the protein structure, and next we connect protein terminals several hundred times to two points randomly chosen on the sphere to enclose the analyzed chain. Subsequently, these two points are connected by an arc lying on the surface of the sphere. After the protein chain has been closed, we can identify the topology of the protein (a knot type) using knot invariants. In the case of random closure, the most frequently observed knot type for a given analyzed chain is then associated with that chain as its dominant knot type.

Most knot invariants, such as the Jones and HOMFLYPT polynomials (or Yamada polynomial in the case of θ-curves),[12,39,40] pose a computational challenge as they are #P-hard to compute. Even the faster Alexander polynomial, which has a time complexity of $O(n^3)$, where n is the number of crossings in a given projection, is still inefficient for proteins that possess complex geometric representations. Employing ML techniques for knot recognition provides other opportunities.[29−31] If it were possible to build such an ML that detects knots on proteins, for example, one could use ML models to filter doubtful cases, e.g., by fixing a very high probability threshold to pick unknots, exclude them, and analyze through polynomials only the presumably knotted configurations.

Herein, we tested the power of ML to distinguish different types of entanglement based on simulated polymers, protein-like structures, and naturally occurring knotting and unknotting in hundreds of thousands of protein structures[24] predicted by AlphaFold 2. We have investigated and demonstrated a highly accurate classification of various types of topologies (Figure 1), such as open and closed trivial and nontrivial entanglement (0_1, 3_1, 4_1, 5_1, 5_2, and 6_1), and θ-curves ($\theta 0_1$, $\theta 3_1$, $\theta 4_1$, $\theta 5_1$, and $\theta 5_2$), utilizing an NN architecture.

■ METHODS

Data Input. Here, we consider three types of structures from the point of view of stiffness: polymers (without local stiffness), protein-like chain (with dihedral and planar constant angles), and proteins. In the case of proteins, the effective bending stiffness of the chain is influenced by secondary structures like α-helices and β-strands, whose presence is dictated by the amino-acid sequence and protein fold. In the case of polymers and protein-like chains, we consider open and closed knots. More details are given below.

Polymers and Protein-like Structures. We considered three types of topological structures: closed knots, open knots, and θ curves, each serving as a model for random polymers. For each of the structure types, we built an initial mathematical model of a structure composed of N beads. For closed knots, the initial model was obtained using KnotPlot software.[41] The initial models of the open knots and θ-curves were obtained by constructing a cubic spline representation of the model. These splines were parametrized by arc length parametrization and sampled at N uniformly distributed points.

From the initial model, the movements were then simulated with the molecular dynamics (MD) simulations in the space enclosed by the cylinder (Figure 2 and Molecular Dynamics Model description) in order to obtain different conformations that could serve further as a training set.

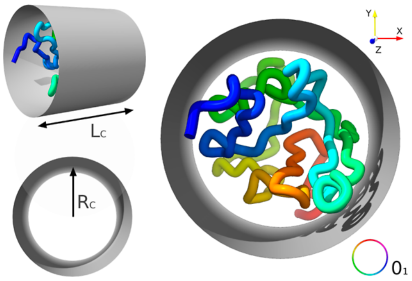


Figure 2. Schematic representation of the protein within the cylinder. L_c and R_c correspond to the cylinder length and radius. The rainbow-colored protein used in the representation corresponds to the AlphaFold ID: A0A0K6IPI7. The AlphaFold structure has an unknotted conformation (0_1). The N and C termini exhibit unrestricted movement along the x and y coordinates, yet are constrained along the z-axis. The cylinder bases were hidden to enhance visualization.

In the case of closed polymers and θ-curves, the action of forces on the polymer cannot change its topology; however, in the case of open knots, it would theoretically be possible for the end of the polymer to slide out/insert through the loops and thus change the topology during the simulation. In order to avoid such situations, the ends of the polymers can slide only on the surface base of the cylinder in which the starting structure was placed first. Furthermore, to model polymers, no constraints were used. On the other hand, to mimic stiffness in the case of protein-like, additional constraints, such as plane angles and dihedrals, were added (see the Molecular Dynamics Model description section).

Proteins. Two sets of proteins with 0_1 and 3_1 knots were investigated based on: (1) the knot and unknotted configuration obtained from MD simulations and (2) the native conformation of unknotted and knotted protein structures predicted by AlphaFold 2, which additionally meets the condition of appropriate length as described below.

In the first case, we selected six protein structures with and without a knot: A0A1H4VHL9 (1), A0A1M4N8C8 (2), and P85286 (3) were used for 0_1 topology and A0A2D6CS53 (4), A0A7L1ILP5 (5), and A0A2E8PTH8 (6) for 3_1 topology. Numbers in parentheses are used for easier identification of proteins. In all cases, the protein length was slightly longer than 128 amino acids, Table 1. Proteins were trimmed to 128 amino acids without affecting the main part of the protein (each amino acid was represented by one bead, Cα atom). To obtain a sufficient number of configurations to train the model, proteins were simulated in the same way as protein-like data.

In the second case, the data were constructed based on all proteins deposited in AlphaFold v2 (over 200 million),[20] with the average quality of the full-chain 3D structures of pLDDT > 70. We used AlphaKnot[11] database to select and download 681,000 potentially knotted proteins. We chose only proteins without clashes in the knot core region (based on the MolProbity tool). Structures that did not pass visual inspection and those with obvious problems were rejected from our analysis. From this data set, we derived two independent data sets that we cut into 128 and 256 beads.

For the data set with 128 beads, we picked proteins that possess knot core not longer than 126 amino acids and trimmed them to 128 amino acids so that the knot was positioned approximately in the middle of the sequence.

For the unknotted data set, we picked high-quality proteins from the SwissProt part of the UniProt Database so that the number matches the knotted set. We chose proteins with sequences longer than 160 amino acids with 128 beads selected from the middle of the protein. Slightly longer proteins were picked to ensure that the cut fragment is not positioned on the flexible ends of the protein as they often present lower pLDDT scores. To address the redundancy, we then applied sequential clustering using the CD-HIT 70% as a threshold to both knotted and unknotted data sets filtering out and removing very similar proteins. Next, we labeled proteins with their InterPro architectures and split them into training and test sets in such a way that proteins with the given architecture can appear only in exactly one of the sets but not in both.

Similarly, we also constructed both knotted and unknotted data sets with the 256 beads. The 254 was the maximum knot core size, and unknotted proteins needed to be of at least length 300. Clustering and architecture separation methods were also applied. For even further testing of the model learning boundaries, we also prepared knotted data sets for 384 and 512 beads (however, finally not used). The number of proteins and architectures are detailed in Table S2 in Supporting Information.

In total, we constructed the following data sets: for 128 beads, the training set consisted of 55,069 unknotted proteins (0_1) and 47,554 knotted proteins (3_1), 102,623 in total; the test set consisted of 23,500 unknotted proteins (0_1) and 20,824 knotted proteins (3_1), 44,324 in total. For 256 beads, the training set consisted of 41,542 unknotted proteins (0_1) and 41,001 knotted proteins (3_1). The test set consisted of 5134 unknotted proteins (0_1) and 5070 knotted proteins (3_1). The complete data sets, based on which clustering and training were made, are available upon request. More information about the details of the selection of data sets is in the following sections.

Machine Learning. *Features.* Each sample is a frame from the simulation, and it is given as a sequence of positions of the N beads of the given polymer, protein-like, or protein structure (called here just a polymer). We preprocessed these input data to a well-defined set of features (input vectors), representing the investigated polymer, and a corresponding set of M labels (output vectors), representing the

Table 1. Proteins Used to Conduct MD Simulations to Train the LSTM Model[a]

knot type	UniProtKB ID	knot core	short name	domains	Sim. in test set
0_1	A0A1H4VHL9		RbsD	IPR007721; IPR023064	no
	A0A1M4N8C8		panD	IPR003190	no
	P85286		NDK B	IPR034907	no
3_1	A0A2D6CS53	1−128	DndE	IPR014969	no
	A0A7L1ILP5	21−67	TRMD/TRM10	IPR007356; IPR016009; IPR028564	no
	A0A2E8PTH8	9−118	Unchar.	no dom.	no

[a] The last column shows that no proteins with similar sequence (proteins at over 80% coverage and 50% sequence identity) were used to test this model. The test set is composed of 23,500 proteins with 0_1 knots and 20,824 proteins with 3_1 knots predicted by AlphaFold 2.

topology type of the polymer. We train two separate ML models using the following two feature sets:

1. The input vector is a sequence of N Cartesian coordinates $\mathbf{x_i} = (x_i, y_i, z_i)$, where $0 \leq i < N$, of each bead in the chain,

2. The input vector is a sequence of $(N - 2)$ relative spherical coordinates, where $(d_i, \theta_i, \varphi_i)$, $0 \leq i < N - 2$, where $d_i(s) = |\mathbf{x}_{i+1} - \mathbf{x}_i|$ is the Euclidean distance between the i-th and $(i + 1)$-th bead, $\theta_i = \angle(\mathbf{x}_i, \mathbf{x}_{i+1}, \mathbf{x}_{i+2})$ is the angle between the i, $(i + 1)$ and $(i + 2)$-th bead, and φ_i is the dihedral angle between beads i, $i + 1$, $i + 2$, $i + 3$, i.e., the angle $\angle(\Sigma_1, \Sigma_2)$, where Σ_1 is the plane on which beads i, $(i + 1)$ and $(i + 2)$ lie and Σ_2 is the plane on which beads $(i + 1)$, $(i + 2)$ and $(i + 3)$ lie. Note that relative spherical coordinates are invariant under rotations and translations of the trajectory obtained from the sequence and thus generate a much smaller configuration space for each protein topology.

Model Architecture. Herein, we investigate the deep learning models, which we built upon a series of LSTM layers (implementation from the *Keras*[42] and *TensorFlow*[43] packages). LSTM models are specialized types of recurrent NNs (RNNs) primarily designed to effectively handle sequences of data, such as natural language processing, time series analysis, and, in our case, spatial coordinate sequences.[44]

LSTM models stand out due to their unique ability to capture long-range dependencies and effectively manage the flow of information through time steps. This is crucial when working with sequences of data that exhibit complex temporal relationships. LSTMs achieve this by incorporating a memory cell and gating mechanisms that allow them to selectively retain and forget information at each time step, making them robust in preserving important context,[32] such as topological information on spatial curves, over extended sequences.

The input to the model is a sequence of features for each sample. In *xyz* coordinates, each sample is a vector of dimension $(N, 3)$, where N denotes the number of beads. In the case of spherical coordinates $(d_i, \theta_i, \varphi_i)$, the input vectors are of dimension $(N - 3, 3)$. The input vector is then passed to three bidirectional LSTM layers with N (or $N - 3$) units each. In addition, we use bidirectional LSTM layers, which allow the model to analyze the structures in both forward and backward directions. We pass the information through a Global Max Pooling Layer to a sequence of three feed-forward (FF) dense layers, with 128, 64, and N_T units, respectively, where N_T is a number of different topologies (labels) included in a training data set. After the first dense layer, a dropout layer with a dropout rate of 0.2 is used to reduce overfitting. A schematic diagram of the model is shown in Figure 3.

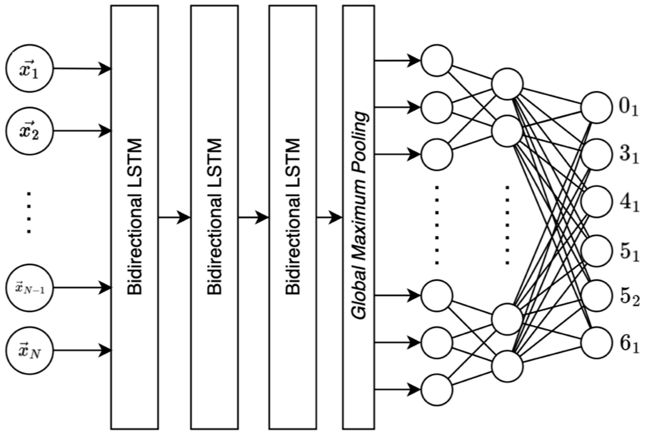


Figure 3. Diagram of the model architecture. A sequence of vectors of dimension $[(N, 3)$ or $(N - 3, 3)$ for spherical coordinates$]$ was used as the sequential input, which is passed to the first of three LSTM bidirectional layers. Subsequently, the global max pooling operation and the FF part, which in the optimal case uses three layers, successively with 128, 64, and N_T neurons.

Molecular Dynamics Model Description. *Model Construction.* Systems of closed knots and open knots consisted of 64, 128, and 256 beads; θ-curves and composite θ-curves were made up of 92, 188, and 380 beads. We introduced bond, angle, and dihedral angle potentials for beads adjacent to the chain. Beads that did not interact via backbone potentials interacted with the repulsive part of the Lennard-Jones potential. The total potential has the following form

$$V = \sum_{\text{bonds}} \varepsilon_b (r - r_0)^2 + \sum_{\text{angles}} \varepsilon_a (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} (\varepsilon_{d1}[1 + \cos(\phi - \phi_0)] + \varepsilon_{d2} [1 + \cos(3(\phi - \phi_0))]) + \sum_{\text{non-bonded}} \varepsilon \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12},$$

where $\epsilon_b = 100\ \epsilon/\text{Å}^2$, $\epsilon_a = 2\ \epsilon/\text{rad}^2$, $\epsilon_{d1} = 0.1\ \epsilon$, and $\epsilon_{d2} = 0.01\ \epsilon$. The hard-sphere diameter of the beads was taken to be equal to 2.0 Å, and the distance between neighboring beads along the chain was around 3.8 Å. We also considered a simplified model of the polymer with switched-off angles and dihedral potentials.

In the case of the triple-bonded beads which form the θ-curve topology, at each junction we added dihedral and planar angles as if they were added at any place along the chain. We used the same force constants, and the θ_0 and ϕ_0 were taken from the generated starting conformation. The method in which we generated these starting structures led in general to different values for these angles at every junction.

Cylindrical Box. We used the same confinement for systems with the given type of entanglements in order to preserve similar packing/compactness of systems with different topology, since, e.g., chain with a 0_1 knot has a tendency to more spread than a chain with a 5_2 knot. Generally, we considered several sizes of the confinement; however, in each case, height of the cylinder was equal to the diameter of its base ($L_\mathrm{c} = 2R_c$, see Figure 2). The walls were simulated with the Lennard-Jones potential

$$V_{\text{cage}} = \epsilon_c \left[\left(\frac{2}{d_i} \right)^4 - 2 \left(\frac{2}{d_i} \right)^2 + 1 \right]$$

where d_i is the distance between the i-th bead and the wall, and $\epsilon_c = 10.0\epsilon$. The cutoff distance is 2 Å, and within that distance from the wall, interactions between beads and the cylinder were present.

In the case of open knots, to prevent changes in topology during the time evolution of the system, positional restraints were imposed on the extreme beads. Each of them could move only in a plane parallel to the bases of the cylindrical cage at a distance of around 2.0 Å of the base. Each of the extreme beads was restrained in the vicinity of the opposite base.

MD Simulation and Sampling. All simulations were conducted using Gromacs v4.5.4[45] with introduced potential for the cylindrical cage.[46] A leapfrog stochastic dynamics integrator with an inverse friction constant of 1.0 was used. The time step was equal to 0.0005 τ. For temperature, we use Gromacs units (for reduced units, one has to multiply it by 1/0.0083145).

Simulations were run at temperature set to 50 ϵ/k_B, and conformations were saved every 10^2 steps for knots and every 10^4 steps for θ-curves. In the case of proteins, conformations were generated by means of MD simulations performed at seven different temperatures in a range between 20 ϵ/k_B and 100 ϵ/k_B, and the optimal results were obtained at a temperature equal to 20 ϵ/k_B. Conformations were saved every 10^2 steps.

The topology of each model was additionally checked after the simulation to ensure that it had not changed.

Topology Determination. All of the models 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.[47] The fingerprint method was used to determine the position of the knot core.[7] A structure is classified as knotted when random closures form a nontrivial knot more frequently than a trivial knot.

All knotted proteins were downloaded from AlphaKnot 2.0 using its API (https://alphaknot.cent.uw.edu.pl/api). We selected only the proteins with 3_1 topology and the knot core shorter or equal to 126 or 254 amino acids (depending on the data set, with 128 or 256 total beads). The AlphaKnot Database also provided the knot core positions for those proteins.

RESULTS

This study investigated the optimal approach and the best performance of the NNs model to identify and classify knots and θ-curves based on their 3D structure. The data used can be divided into three main types: polymers, protein-like, and proteins. For each polymer and protein-like with closed and open knots, there are six types of topologies (0_1, 3_1, 4_1, 5_1, 5_2, and 6_1). In the case of θ-curves, there are five classes ($\theta 0_1$, $\theta 3_1$, $\theta 4_1$, $\theta 5_1$, and $\theta 5_2$), as shown in Figure 1. Lastly, for proteins, we consider two classes (0_1 and 3_1), based on the topology that is observed in the native state of proteins.

Training and Testing. The data set was split in a typical sequential fashion: the first 70% of samples (frames) from the simulation were taken as the training set, the next 10% to the validation set, and the last 20% to the testing set. This approach was introduced to avoid leakage of data from the training data set since frames close together in the simulation show small variation from each other, which means that they may not be independent of each other.

For proteins, the approach was different. In this case, we tested two different solutions to the problem. In the first one, training was performed on simulated configuration for six proteins with 0_1 and 3_1 type of knots, see Figure 4 and Table 1. In the second one, training was performed on a set composed of native protein conformations: 55,069 with 0_1 knot and 47,554 with 3_1 knot, which we identified and preprocessed based on AlphaFold prediction. However, tests for both approaches (simulation-based and native conformation-based) were conducted on a data set also consisting of native protein conformations derived from AlphaFold predictions of: 23,500 with 0_1 knot and 20,824 with 3_1 knot. In this way, additional difficulty was introduced to the simulation-based approach due to the far greater differences between the different native structures than it is in the case with many different conformations but derived from the same simulation as it takes place for polymers and protein-like. Moreover, proteins from the training and test sets possess different architectures. All training, validation, and test sets were normalized using Standard Scaling (removing the mean and scaling to the unit variance).

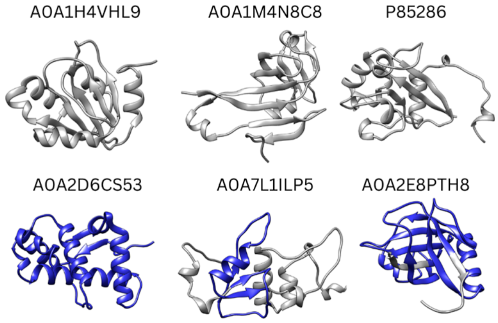


Figure 4. Cartoon representation of the 3D structure of the proteins used in the simulation to train the LSTM model. The blue color indicates the position of the knot core. Each protein has a different protein architecture.

The models are optimized using the RMSprop optimizer[48] for open knots (due to better convergence behavior) and the ADAM optimizer[49] for the other cases and the categorical cross-entropy loss function. The batch size is set to 32. The number of epochs varies, depending on the specific experiment. Due to the rapid stabilization of both loss and accuracy for the validation and training set (2−3 epochs), no difference in the model's performance was noticed for values between 3 and 50 epochs. In general, we tested many different hyperparameters, and most parameters are listed in Table S1. In the case of proteins of length 256 beads, a grid search (training the model for all combinations of given hyperparameter values) was performed; the range of parameters taken for optimization and the top 10 results obtained can be found in Table S3. The selected parameters give the best performances of the models for a given type of input data. The accuracy metric is used to evaluate the performance of the model.

Polymers and Protein-like Chains. *Closed and Open Knots.* First, we investigate the LSTM on polymer chains with closed and open knots; the procedure for open knots simulations is described in Data input in Methods section. In both cases, the chains had 128 beads and were divided into six classes, as described in Figure 3 and Table 2. As shown in Figure 5, the confusion matrices provide a comprehensive evaluation and detail the ML model's performance. These findings provide vital insights into the task of more precise topology classification. Such results display classification among the actual and predicted classes. Figure 5A, normalized confusion matrix for closed knots, shows a satisfactory prediction for the first five classes (0_1, 3_1, 4_1, 5_1, and 5_2); however, a slight difference was observed for the 6_1 knot. Due to the complexity of 6_1 knot, the ML may often confuse it with 4_1 and 5_2, also observed in

Table 2. Best Results for Each Type of Structure Tested

chain type	topology type	no. beads	no. classes	features	optimizer	accuracy [%]
polymers	closed knots	128	6	(x_i, y_i, z_i)	ADAM	90
	open knots	128	6	(x_i, y_i, z_i)	ADAM	92
	θ-curves	92	5	(x_i, y_i, z_i)	RMSprop	95
protein-like	closed knots	128	6	$(d_i, \theta_i, \varphi_i)$	ADAM	84
	open knots	128	6	$(d_i, \theta_i, \varphi_i)$	RMSProp	99
	θ-curves	92	5	$(d_i, \theta_i, \varphi_i)$	ADAM	79
proteins (sim)	open knots	128	2	$(d_i, \theta_i, \varphi_i)$	ADAM	67
proteins (native)	open knots	128	2	$(d_i, \theta_i, \varphi_i)$	ADAM	93
proteins (native)	open knots	256	2	$(d_i, \theta_i, \varphi_i)$	ADAM	86

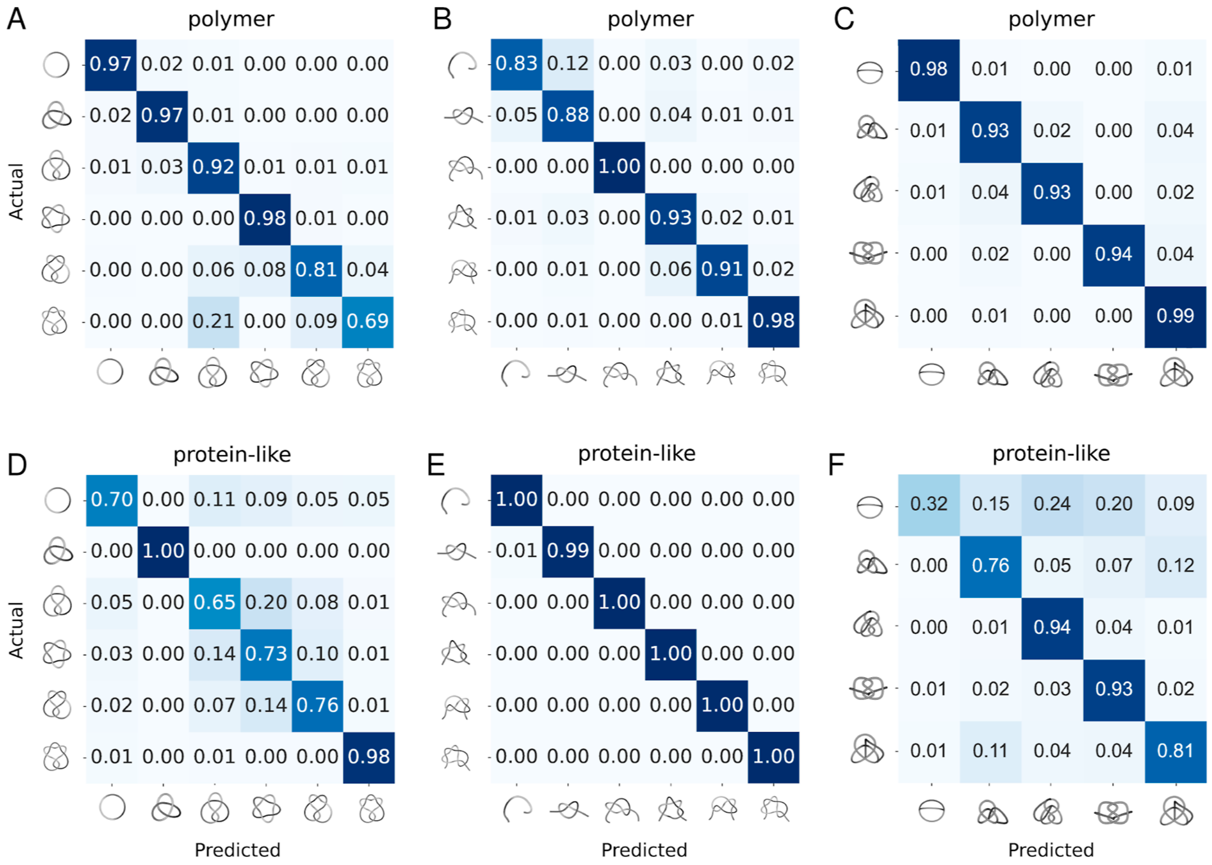


Figure 5. Normalized confusion matrices for polymer, protein-like, and θ-curves. (A−C) Panels correspond to the polymers with closed, open, and θ-curve knots. (D−F) Panels correspond to the protein-like with closed, open, θ-curve knots. Open and closed knots are represented by six classes (0_1, 3_1, 4_1, 5_1, 5_2, and 6_1), while θ-curves with five classes ($\theta 0_1$, $\theta 3_1$, $\theta 4_1$, $\theta 5_1$, and $\theta 5_2$). The blue scale color corresponds to the probability, dark blue for high probability, and light blue for low probability.

reference.[30] Interestingly, all three exhibit a high degree of topological similarity as they are both twist knots with an even number of twists.

Despite the slight difference observed, the confusion matrix specifically between the labels corresponding to the 4_1 and 6_1 knots suggests that the model learned to recognize associate geometrical patterns and transformations to a specific label, that is, the topological class. Since proteins are not closed chains, we constructed open polymeric chains with the same knot types as closed knots. The goal is to assess and analyze the efficacy of NN in classifying knot types that closely approximate to the real system, proteins (native conformation). In Figure 5B, the confusion matrix presents a satisfactory predictor for all knot types investigated in this study. In both cases, LSTM was trained with the same features, and the results display a good accuracy, 90, and 92% for closed and open knots, respectively. The results in Figure 5A,B, as well as in Table 2, indicate that LSTM can identify and distinguish different topological configurations.

Figure 5D,E presents the results for protein-like chains whose structure is characterized by constraints such as internal forces, like: bonds (two consecutive atoms) (d), angles (θ), and dihedral angles (ϕ). In the case of proteins, their rigidity is due to secondary structure motifs: α helices, β strands, and turns. Thus, in the case of polymers depending on the given persistence length, their range is from highly flexible to relatively rigid. Protein-like chain studied here encompasses the spectrum between polymers and proteins. The performance of the LSTM trained based on the features d, θ, and ϕ for closed knots (Figure 5D) was moderate, which is evident by the confusion matrix result. The model could not predict the knot types accurately except for 3_1 and 6_1 knots. On the other hand, the LTSM trained using identical features, as for protein-like closed knot structures, showed the highest performance when trained and tested with open knots. All knots were well predicted, and the accuracy for closed and open knots is 84 and 99%, respectively.

As observed, knot-type identification using LSTM may present low accuracy depending on the features of the model on which it was trained on. Polymers and protein-like with open and closed knots were set with different parameters (see Table S1 in Supporting Information) to capture the best LSTM NN performance.

θ-Curves. In the case of the θ-curves simulated without constraints based on the same feature shape as for the knots, i.e. $(N, 3)$, we obtained rather good results (95% accuracy) as one could expect. However, in the case of θ-curves obtained in constrained simulations, it is only 66% accurate. However, while a closed knot is a circle embedded in $\mathbb{R}^3$, a θ-curve is a spatial graph that consists of three edges with two common vertices.[50] For this reason, we proposed changing the shape of the features to $\left(\frac{N+4}{3}, 9\right)$. Each of the $\frac{N+4}{3}$ rows contains the (x, y, z) coordinates of three beads (one from each strand). This change brought an increase in the accuracy to 79%. However, the efficiency of trivial θ-curve recognition drastically dropped. This is probably due to the high flexibility of its structure. A knot with many crossings restricts the movements of the polymer, while a trivial knot does not, which means that it can obtain many different conformations, making it difficult for the NN to capture the dependencies.

Proteins. In the context of protein structure classification, a new challenge arises with model performance. Two approaches were tried in this case. The first was based on generating a conformation with MD simulation for randomly selected proteins and training a model on them. The second approach was based on training the model on part of a data set composed of native conformation of proteins. For both approaches, the trained models were tested on the same test data set consisting of 44,324 native protein structures. In both cases, due to the complexity of the protein structures, $(d_i, \theta_i, \varphi_i)$ variables were used.

Simulation-Based Training Set. A natural step in the transition from polymers to proteins is to consider conformations of proteins generated by MD simulations. To prepare the training set, three proteins without a knot and three with the 3_1 knot were selected, see Table 1 and Figure 4. Each protein has a different protein architecture, Figure 4. To obtain a sufficient number of configurations to train the model, proteins were simulated in the same way as protein-like data. We analyzed combinations of all possible knotted−unknotted pairs. We found that the best performance—of 67% accuracy—was obtained on the data from simulations of proteins (2) and (5). It is the simplest example of employing individual proteins in generating a training set for the ML model, and the results are based on individual trajectories. However, such an approach can be a starting point for studying data from the dynamic processes like protein folding or other large rearrangements of protein structure over the course of MD simulations, and it can be used, e.g., in recognition of knots in folding kinetics.

Training Set based on Native Conformation of Proteins. The second approach was to train the model on native protein structures. Two data sets were prepared for training and testing purposes, consisting of knotted and unknotted proteins. Knotted proteins were selected from AlphaKnot 2.0.[11] To ensure the diversity of the knotted data set (different protein architecture), proteins were grouped by the InterPro domains appearing in the knot core and clustered using CD-HIT[51] at the 70% identity level. We concentrated on the 3_1 topology as it is the most common topology and appears in the widest range of protein architectures, thus ensuring that enough data can be provided for model training. The unknotted data set was prepared, as described in the Data Input section.

Trivial and knotted proteins were randomly distributed, and we ensured that each architecture could be present only in either the training or test set. In addition, we calculated the rmsd between pairs of proteins with different architectures for 1000 of the 17,000 possible pairs to ensure that there was no data leakage (one architecture can be composed of several different domains). We found that rmsd is above 4 and 6 Å for 73 and 98% of pairs, respectively (details are given in the Supporting Information). Thus, there should be enough structural diversity between architectures for the model to train properly.

In total, the training set was composed of 55,069 unknotted proteins (11,426 architectures) and 47,554 knotted proteins (602 architectures). The test set had 23,500 unknotted proteins (4843 architectures) and 20,824 knotted proteins (454 architectures). The full list of architectures and protein IDs is available upon request.

Following the results for the case with 128 beads, we tested the accuracy of our model on the larger sequences of 256 beads. The training set consisted of 41,542 unknotted proteins (6680 architectures) and 41,001 knotted proteins (2296 architectures). The test set consisted of 5134 unknotted proteins (1089 architectures) and 5070 knotted proteins (932 architectures).

Results Comparison. The results obtained for knotted protein structures are shown in Table 3 and as a confusion matrix in Figure 6. The simulation-based model is denoted as proteins (sim), and the native-structure-based model with 128 and 256 beads is denoted as proteins (native). Due to the slightly unbalanced test set for the proteins, other metrics such as precision (positive predictive value, $\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$), recall (true positive rate, $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$), and F1-score ($\mathrm{F1} = \frac{2\cdot\mathrm{PPV}\cdot\mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}}$) were also presented. Precision measures the accuracy of positive predictions by calculating the ratio of true positives to the sum of the true positives and false positives. Recall shows the ability to classify relevant instances within the all positive class by calculating the ratio of true positives to the sum of true positives and false negatives. The F1-score, which is the harmonic mean of precision and recall, provides a more comprehensive metric of the model with an unbalanced test set.

Table 3. Selected Metrics for Protein Models

type	no. beads	acc. [%]	topol.	precision	recall	F1-score
sim	128	67	0_1	0.75	0.55	0.64
			3_1	0.61	0.79	0.69
native	128	93	0_1	0.89	0.99	0.94
			3_1	0.99	0.86	0.92
native	256	86	0_1	0.80	0.96	0.88
			3_1	0.95	0.74	0.83

A model trained on native protein structures with 128 beads outperforms the one trained on a simulated data set due to the definitely greater diversity of the first one. The native protein-based training set gives a diverse pool of examples, which provides a greater coverage of the configuration space, allowing the variability and complexity characteristic of different proteins to be captured. In contrast, a model trained on a limited set of proteins lacks this expansive diversity, limiting its ability to generalize across various proteins. By learning from a broader spectrum of structures, the model gains a deeper understanding of the complex relationships among smaller sequences, a whole structure, and topology, resulting in more accurate classification.

The results for the model trained on the 256 bead window show that the accuracy dropped as compared to the 128 bead model. One of the possible explanations could be the limited memory capability for LSTM models. The recommended method of applying the model for larger proteins might be thus to use the sliding window technique: cutting 128 beads from overlapping parts of the protein.

■ CONCLUSIONS

In summary, our investigation of LSTM NN models on the polymer, protein-like, and protein chains, with both closed and open knots as well as θ-curves, has provided valuable insights into their classification and identification. The analysis revealed that the LSTM model can predict classes (up to the 6_1 knot) accurately not only for closed knots but also for open polymeric chains, resembling native proteins. Application of LSTM with proper features has the ability to classify knot types in protein-like chains, displaying robust prediction capabilities with an

https://doi.org/10.1021/acs.macromol.3c02479

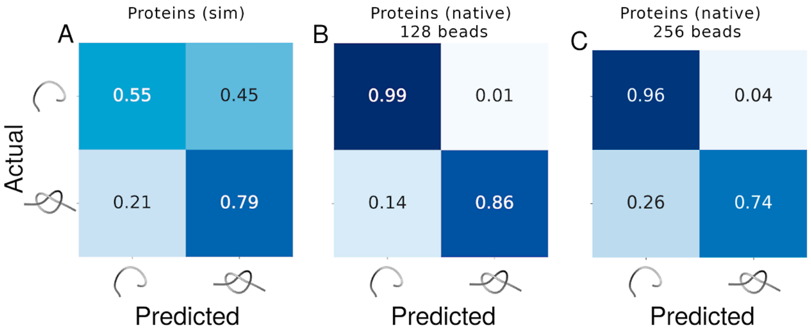


Figure 6. Confusion matrices for proteins with two classes of topologies (0_1 and 3_1) for proteins model trained on (A) simulated data [proteins (sim)] and model trained on (B) native structures [protein (native)] with 128 beads and (C) native structures [protein (native)] with 256 beads.

accuracy of 99%. This suggests the model's efficacy in handling systems that closely approximate real-world scenarios.

The aim to precisely identify and classify knotted protein architectures stands as a considerable challenge in scientific research. Two different methodologies were employed: the first approach involved many conformations of a single protein obtained from MD simulations, while the second utilized a single conformation from many different proteins predicted by AlphaFold 2.0. In both cases, the model was tested on native conformation of 23,500 unknotted and 20,824 knotted proteins sampled from a wide range of domain architectures. Despite different techniques, both approaches take into account the same features to address the intricate nature of the protein structures. The model trained on simulations displayed promising performance with an accuracy of 67%. On the other hand, the second approach considers a training set composed only of native conformation of knotted (47,554) and unknotted (55,069) 128 bead proteins, leading to an accuracy of 93%. For the analyzed longer proteins, i.e., 256 beads in length, the results obtained are slightly worse, and the accuracy is 86%. Such results are promising since we ensured that there was no overlap in the architecture of proteins between the training and testing sets, and one of the reasons for the underperformance for longer sequences could be the limited memory capacity of the LSTM networks. One considered solution to this problem might be to use a model trained on shorter sequences that tests individual overlapping fragments in a longer sequence in a sliding-window approach. This approach can also be applied to much longer proteins, e.g., human proteins. Only note that it will allow knots smaller than 126 or 254 amino acids to be found.

Currently, there are around 600,000 potentially knotted proteins.[11] Based on AlphaFold-Multimer-predicted protein complex structures, it has been estimated that approximately 1.72% of the predicted structures contain topological links.[52] The complexity and number of other types of nontrivial topologies such as lasso and links in a single proteins chain are not known since it was not checked in AlphaFold or EMSFold. Creating an LSTM model to review such data (as a first filter supported later with classic methods) would provide knowledge of the number of entangled proteins in a given genome and type of entanglements and thus provide data that can be further used to analyze, e.g., potential correlation between topology and biological function of a given protein. An LSTM model could be also used to scan for the possession of nontrivial topologies in *de novo* prediction or kinetics simulations of biopolymers. In conclusion, the classification of knots remains a challenge, especially concerning entanglement motif in biopolymers. Within this study, we have shown that the LSTM model based on NN architecture can achieve very good accuracy in recognizing some knots, open knots, and θ-curves within a polymer, protein-like, and protein chains. Consequently, one could expect that the LSTM approach can be further developed toward accurately predicting, distinguishing, and classifying more complex nontrivial topologies in highly complex and disorderly structures.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.macromol.3c02479.

> Description of protein architectures; type of structure tested for different features and optimizers; summary of structural similarity between domains; best protein models (256 beads) from grid search; and number of knotted proteins clustered for the data sets of 128, 256, 384, and 512 beads (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Joanna I. Sulkowska − *Centre of New Technologies, University of Warsaw, Warsaw 02-097, Poland;* orcid.org/0000-0003-2452-0724; Email: j.sulkowska@cent.uw.edu.pl

Authors

Fernando Bruno da Silva − *Centre of New Technologies, University of Warsaw, Warsaw 02-097, Poland;* orcid.org/0000-0002-0285-8700

Boštjan Gabrovšek − *Faculty of Mechanical Engineering, University of Ljubljana, Ljubljana 1000, Slovenia; Institute of Mathematics, Physics and Mechanics, Ljubljana 1000, Slovenia*

Marta Korpacz − *Centre of New Technologies, University of Warsaw, Warsaw 02-097, Poland; Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw 02-097, Poland*

Kamil Luczkiewicz − *Centre of New Technologies, University of Warsaw, Warsaw 02-097, Poland; Faculty of Physics, University of Warsaw, Warsaw 02-097, Poland*

Szymon Niewieczerzal − *Centre of New Technologies, University of Warsaw, Warsaw 02-097, Poland*

Maciej Sikora − *Centre of New Technologies, University of Warsaw, Warsaw 02-097, Poland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.macromol.3c02479

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Science Centre (2022/47/B/NZ1/03480 and 2021/43/I/NZ1/03341 to J.I.S.). B.G. was supported by the Slovenian Research and Innovation Agency grant N1-0278. We would like to thank B. Gren and W. Garstka for the assistance in protein structure preparation. We would like to thank P. Rubach for IT support and S. Nowakowski for all discussion.

■ REFERENCES

(1) Gukov, S.; Halverson, J.; Ruehle, F.; Sułkowski, P. Learning to unknot. *Mach. Learn.: Sci. Technol.* **2021**, *2* (2), 025035.

(2) Hughes, M. C. A neural network approach to predicting and computing knot invariants. *J. Knot Theory Ramif.* **2020**, *29* (03), 2050005.

(3) Sulkowska, J. I. On folding of entangled proteins: knots, lassos, links and θ-curves. *Curr. Opin. Struct. Biol.* **2020**, *60*, 131−141.

(4) Hsu, S. T. D. Folding and functions of knotted proteins. *Curr. Opin. Struct. Biol.* **2023**, *83*, 102709.

(5) Dabrowski-Tumanski, P.; Goundaroulis, D.; Stasiak, A.; Sulkowska, J. I. θ-curves in proteins. *arXiv* **2019**, arXiv:1908.05919.

(6) O'Donnol, 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), 9181−9188.

(7) Millett, K. C.; Rawdon, E. J.; Stasiak, A.; Sułkowska, J. I. Identifying knots in proteins. *Biochem. Soc. Trans.* **2013**, *41* (2), 533−537.

(8) Tibor, E.; Annoni, E. M.; Brine-Doyle, E.; Kumerow, N.; Shogren, M.; Cantarella, J.; Shonkwiler, C.; Rawdon, E. J. Performance of the uniform closure method for open knotting as a bayes-type classifier. *arXiv: Geometric Topology* **2020**, arXiv:2011.08984.

(9) Lua, R. C.; Grosberg, A. Y. Statistics of knots, geometry of conformations, and evolution of proteins. *PLoS Comput. Biol.* **2006**, *2* (5), No. e45.

(10) Jarmolinska, A. I.; Perlinska, A. P.; Runkel, R.; Trefz, B.; Ginn, H. M.; Virnau, P.; Sulkowska, J. I. Proteins' knotty problems. *J. Mol. Biol.* **2019**, *431* (2), 244−257.

(11) Niemyska, W.; Mukherjee, S.; Gren, B.; Niewieczerzal, S.; Bujnicki, J. M.; Sulkowska, J. I. Discovery of a trefoil knot in the rydc rna: Challenging previous notions of rna topology. *J. Mol. Biol.* **2024**, *436*, 168455.

(12) Jamroz, M.; Niemyska, W.; Rawdon, E. J.; Stasiak, A.; Millett, K. C.; Sułkowski, P.; Sulkowska, J. I. KnotProt: a database of proteins with knots and slipknots. *Nucleic Acids Res.* **2015**, *43* (D1), D306−D314.

(13) Kamitori, S. A real knot in protein. *J. Am. Chem. Soc.* **1996**, *118* (37), 8945−8946.

(14) Biou, V.; Dumas, R.; Cohen-Addad, C.; Douce, R.; Job, D.; Pebay-Peyroula, E. The crystal structure of plant acetohydroxy acid isomeroreductase complexed with nadph, two magnesium ions and a herbicidal transition state analog determined at 1.65 å resolution. *EMBO J.* **1997**, *16* (12), 3405−3415.

(15) Taylor, W. R. A deeply knotted protein structure and how it might fold. *Nature* **2000**, *406* (6798), 916−919.

(16) 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* (4), No. e1000731.

(17) Hsu, M.-F.; Sriramoju, M. K.; Lai, C.-H.; Chen, Y.-Ru.; Huang, J.-S.; Ko, T.-P.; Huang, K.-Fa.; Hsu, S.-T. D. Structure, dynamics, and stability of the smallest and most complex 71 protein knot. *J. Biol. Chem.* **2024**, *300* (1), 105553.

(18) Bruno da Silva, F.; Lewandowska, I.; Kluza, A.; Niewieczerzal, S.; Augustyniak, R.; Sulkowska, J. I. First crystal structure of double knotted protein trmd-tm1570 − inside from degradation perspective. *bioRxiv* **2023**.

(19) Brems, M. A.; Runkel, R.; Yeates, T. O.; Virnau, P. Alphafold predicts the most complex protein knot and composite protein knots. *Protein Sci.* **2022**, *31* (8), No. e4380.

(20) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50* (D1), D439−D444.

(21) 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. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373* (6557), 871−876.

(22) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379* (6637), 1123−1130.

(23) Sramkova, D.; Sikora, M.; Uchal, D.; Klimentova, E.; Perlinska, A. P.; Nguyen, M. L.; Korpacz, M.; Malinowska, R.; Rubach, P.; Simecek, P.; et al. Knot or not? sequence-based identification of knotted proteins with machine learning. *Protein Sci.* **2023**.

(24) Niemyska, W.; Rubach, P.; Gren, B.; Nguyen, M. L.; Garstka, W.; Bruno da Silva, F.; Rawdon, E.; Sulkowska, J. Alphaknot: server to analyze entanglement in structures predicted by alphafold methods. *Nucleic Acids Res.* **2022**, *50*, W44−W50.

(25) Perlinska, A. P.; Niemyska, W. H.; Gren, B. A.; Bukowicki, M.; Nowakowski, S.; Rubach, P.; Sulkowska, J. I. Alphafold predicts novel human proteins with knots. *Protein Sci.* **2023**, *32* (5), No. e4631.

(26) Graves, A.; Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 855−868.

(27) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436−444.

(28) Yun, K.; Huyen, A.; Lu, T. Deep neural networks for pattern recognition. *arXiv* **2018**, arXiv:1809.09645.

(29) Vandans, O.; Yang, K.; Wu, Z.; Dai, L. Identifying knot types of polymer conformations by machine learning. *Phys. Rev. E* **2020**, *101* (2), 022502.

(30) Braghetto, A.; Kundu, S.; Baiesi, M.; Orlandini, E. Machine learning understands knotted polymers. *Macromolecules* **2023**, *56* (7), 2899−2909.

(31) Sleiman, J. L.; Conforto, F.; Fosado, Y. A. G.; Michieletto, D. Geometric learning of knot topology. *Soft Matter* **2024**, *20* (1), 71−78.

(32) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735−1780.

(33) Gabrovšek, B. An invariant for colored bonded knots. *Stud. Appl. Math.* **2021**, *146* (3), 586−604.

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

(35) Goundaroulis, D.; Dorier, J.; Benedetti, F.; Stasiak, A. Studies of global and local entanglements of individual protein chains using the concept of knotoids. *Sci. Rep.* **2017**, *7* (1), 6309.

(36) Gabrovšek, B.; Gügümcü, N. Invariants of multi-linkoids. *Mediterr. J. Math.* **2023**, *20* (3), 165.

(37) Dabrowski-Tumanski, P.; Niemyska, W.; Pasznik, P.; Sulkowska, J. I. LassoProt: server to analyze biopolymers with lassos. *Nucleic Acids Res.* **2016**, *44* (W1), W383−W389.

(38) Vladimir, T. Knotoids. *Osaka J. Math* **2012**, *49* (1), 195−223.

(39) Dabrowski-Tumanski, P.; Rubach, P.; Niemyska, W.; Gren, B. A.; Sulkowska, J. I. Topoly: Python package to analyze topology of polymers. *Briefings Bioinf.* **2020**, *22* (3), bbaa196.

(40) Grünbaum, D. Narrowing the gap between combinatorial and hyperbolic knot invariants via deep learning. *J. Knot Theory Ramif.* **2022**, *31* (01), 2250003.

(41) Scharein, R. G. Interactive Topological Drawing. Ph.D. Thesis, Department of Computer Science; The University of British Columbia, 1998.

(42) Chollet, F.; et al. *Keras*, 2015. https://keras.io.

(43) Abadi, M.n.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Craig, C.; Corrado, G.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.

(44) Van Houdt, G.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929−5955.

(45) Hess.; et al. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *3*, 435−447.

(46) Niewieczerzal, S.; Sulkowska, J. I. Knotting and unknotting proteins in the chaperonin cage: Effects of the excluded volume. *PLoS One* **2017**, *12* (5), No. e0176744.

(47) Freyd, P.; Yetter, D.; Hoste, J.; Lickorish, W. B. R.; Millett, K.; Ocneanu, A. A new polynomial invariant of knots and links. *Bull. Am. Math. Soc.* **1985**, *12*, 239−246.

(48) Tieleman, T.; Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning* **2012**, *4*, 26−31.

(49) Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

(50) No, S.; Oh, S.; Yoo, H. Topological aspects of theta-curves in cubic lattice. *J. Phys. A Math. Theor.* **2021**, *54* (45), 455204.

(51) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22* (13), 1658−1659.

(52) Hou, Y.; Xie, T.; He, L.; Tao, L.; Huang, J. Topological links in predicted protein complex structures reveal limitations of alphafold. *Commun. Biol.* **2023**, *6* (1), 1098.