bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SOLeNNoID: A Deep Learning Pipeline For Solenoid Residue Detection in Protein Structures

Georgi Nikov[1], Daniella Pretorius[2], and James W. Murray[3,✉]

[1,2,3]Department of Life Sciences, Imperial College London, Exhibition Road, London SW7 2AZ, United Kingdom

Solenoid proteins are a subset of tandem repeat proteins, which are structurally distinct from globular proteins. Solenoid proteins are defined by their modular, elongated structures, dependent on interactions between adjacent repeats. These proteins are found across all domains of life and have many important functions such as protein binding, enzymatic catalysis, ice binding and nucleic acid binding. Furthermore, engineered variants of solenoid proteins such as DARPins and designed PPR proteins have therapeutic commercial applications. In order to advance the study of natural solenoid proteins and the design of novel solenoid proteins, accurate tools for solenoid detection and annotation are required. As solenoid structures are more conserved than solenoid sequences and owing to recent developments in protein structure prediction, structure-based solenoid detection is preferred. Here we propose SOLeNNoID - a deep learning pipeline for solenoid residue prediction in protein structures. We cover all three solenoid sub-classes: alpha-, alpha/beta- and beta-solenoids. We use a CNN architecture to reason over protein distance matrices and compare our method to existing structure-based methods. Finally, we produce predictions on the entire PDB and demonstrate a 71 percent increase in solenoid-containing entries over the gold-standard RepeatsDB database using our method.

GitHub: `https://github.com/gnik2018/SOLeNNoID`
Zenodo: TBC

Correspondence: *j.w.murray@imperial.ac.uk*

Introduction

Tandem repeat proteins (TRPs) consist of multiple identical or highly similar structural modules and are defined primarily by intra-repeat and inter-repeat interactions between residues adjacent in the primary sequence (1) as opposed to the longer-range interactions between residues distant in the primary sequence observed in globular proteins. It is estimated that about 50 percent of proteins contain at least one tandem repeat, defined as an adjacently repeated amino acid pattern (2). Tandem repeat proteins can be divided into five main classes (3). Class I consists of short (1-to-2-residue) repeats which form crystallites, whereas class II comprises fibrous repeating structures such as collagen and alpha-helical coiled coils. Class III includes solenoid and non-solenoid 'open' structures, characterised by a modular elongated structure which is maintained by inter-repeat interactions. Class IV comprises modular 'closed' structures where the N- and C-terminal repeats of the protein interact. Class V is made up of beads-on-a-string proteins, where each repeat is a small independently folded domain connected to adjacent repeats via linker peptides.

Solenoid proteins (class III) are structurally and functionally diverse. For example, solenoid proteins can bind to nucleic acids (4), catalyse reactions (5), display antifreeze activity (6) and bind to ligands (7). In addition, solenoid proteins are attractive design targets due to their modular architecture and have commercial applications in RNA editing (pentatricopeptide repeat proteins - EditForce) and cancer therapy (DARPins - Molecular Partners), in addition to numerous binding and materials applications developed by academic groups (8). The polypeptide chain in solenoid structures follows a helical path, where individual repeats pack against each other and depend on each other for folding (1). Individual repeats (one turn around the solenoid axis) are made up of secondary structure elements ranging from alpha-helices and beta-strands to 3_{10} and polyproline II helices connected by loops.

Both the design of novel solenoid proteins and the exploration of existing ones rely on the accurate detection of repeating units, repeating regions and entire domains.

Protein sequence information is much more abundant than experimental structural information. For example, the 03 Aug 2022 release of the UniProtKB/TrEMBL protein sequence database (9) contains 226,771,949 entries, while, at the time of writing, the PDB (10, 11) has only 222,624 entries, over a thousand times fewer. Sequence-based tandem repeat detection approaches range from using self-alignment matrices (e.g. TRUST (12)), k-means clustering (T-REKS (13)), Hidden Markov Model (HHRepID (14)), and discrete Fourier transform (REPETITA (15)), to neural-network-based approaches (16–19). In addition, the TRAL library (20, 21) integrates many of the above tandem repeat detectors as well as post-processing, refinement, and annotation modules to evaluate and improve the detected repeats.

Sequence-level tandem repeat detection is more challenging than with structures due to deviations from ideality in tandem repeat sequences such as insertions within and between repeats, and sequence divergence between individual repeats (22). Structural detection of repeats has become more applicable recently, as large databases of accurate structure predictions such as AlphaFold (23) and ESMfold (24) have become available. Examples of recent structural tandem repeat detection methods include TAPO (25), RepeatsDB-lite (26) and PRIGSA2 (22). TAPO comprises an ensemble of 7 sub-

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

methods, which use features such as periodicities of atomic coordinates, conformational alphabet strings, residue contact maps and secondary structure vectors. The outputs of these sub-methods are combined using a support vector machine (SVM) model, to detect tandem repeats. RepeatsDB-lite relies on iteratively matching tandem repeats from a predefined library to the input structure to expand the predicted repeating region. PRIGSA2 is a graph-based algorithm, which analyses the contact network and secondary structure annotation of proteins to identify tandem repeat units.

While many tandem repeat prediction algorithms have been developed, there have been relatively few attempts to leverage neural networks. Convolutional neural networks (CNNs) have been applied to detect tandem repeat proteins using 3D protein structure data (DeepSymmetry (18)) and 2D distance matrix representations of protein structure (Deep-StRIP (19)). Deep-StRIP uses two CNN models - one which classifies a protein as belonging to class III, class IV, or non-tandem-repeat, and one which identifies the repeat region after classification. The predictions of the second model are then further processed using a smoothing algorithm.

Here we focus on the solenoid sub-classes (alpha, alpha/beta, beta) (27) of tandem repeat proteins and propose a structure-based supervised deep learning method, SOLeNNoID, for high-quality annotation of repeating regions. We use a 2D distance matrix representation of 3D protein structure and a semantic segmentation approach, employing a U-Net convolutional neural network architecture (28) to produce a probability that a given amino acid in the protein structure belongs to one of the solenoid classes, or is non-solenoid. We showcase our method and downstream analysis of predictions, compare our approach to state-of-the-art methods such as TAPO, RepeatsDB-Lite and PRIGSA2, and search the PDB for alpha-, alpha/beta- and beta-solenoid proteins to add further entries onto databases such as RepeatsDB (29) and DbStRiPs (30).

Methods

Protein Structure Representation. The structure of a protein can be represented by a matrix of distances between Cα atoms (31). The distance matrix is orientation and translation invariant. Each row n of the distance matrix contains the distances from amino acid x_n to all other amino acids in the protein. Information about the structural environment of that amino acid is contained within the pattern of distances.

Datasets. The training/validation dataset consisted of the dataset produced by Marsella et al (15) and additional beta-solenoid data. Additional beta-solenoid proteins were selected to better represent the diversity of cross-sectional shapes described by Kajava and Steven (32). The dataset comprised a set of protein PDB files and label files. The label files were produced using the annotations in the original dataset, and by annotating structures in PyMOL (33) for the additional beta-solenoid data. Each label file contained a label for each residue in the corresponding PDB file. We assigned an integer value label for each residue, corresponding to either of the four possible classes – non-solenoid (0), beta-solenoid (1), alpha/beta-solenoid (2), or alpha-solenoid (3). We assigned missing data to class 4 to distinguish from non-solenoid data. In total, the dataset consisted of 246 non-solenoid, 52 beta-solenoid, 17 alpha/beta-solenoid and 33 alpha-solenoid proteins. The training and validation datasets were derived by a 80:20 (test:validation) random split of the PDB and corresponding label files.

The test dataset consisted of 49 non-solenoid, 15 beta-solenoid, 7 alpha/beta-solenoid and 11 alpha-solenoid proteins. The non-solenoid proteins were collected from the PDBe (34)while the solenoid proteins were collected from manually reviewed solenoid entries from RepeatsDB. All structures are non-redundant with respect to the training and validation structures to a TM-align TM-score threshold of below 0.6 (35). If an alignment between two proteins produces a TM-score of above 0.5, it can be assumed that the proteins generally have the same fold in SCOP (36) or CATH (37). Thus, as the neural network will be used to identify and annotate proteins broadly in the same folds as the structures in the training dataset, the less stringent threshold of TM-score less than 0.6 was used. The RepeatsDB solenoid labels were amended by excluding loops from repeat definitions and adding parts of partial or imperfect repeats. Table S1 shows a number of statistics relating to these datasets. The datasets contain few solenoid structures but this limited data is sufficient to train a classification network. All datasets are imbalanced: there are more non-solenoid structures and residues compared to either of the solenoid classes, and there are fewer alpha/beta-solenoid structures and residues compared to the alpha- and beta-solenoid classes. For example, in the training dataset, the alpha-solenoid to non-solenoid class imbalance is 1 to 9.6, and the alpha-solenoid to alpha/beta-solenoid class imbalance is 2.2 to 1.

Data Processing. A Cα distance matrix was calculated from each structure file using the BioPython (38) and NumPy (39) libraries in Python3 (40). At points where protein chains were discontinuous, both the labels and the distance matrix entries were replaced with padding values to maintain the register of the amino acid indices. Each distance matrix was pre- and post-padded in both x and y axes to a total length divisible by 128. Each distance matrix was then normalised by the largest distance observed in that matrix.

Per-residue class labels were converted to a 2D label matrix (or 'segmentation mask'). Each value in the label matrix corresponds to a Cα-Cα distance, which for two residues of the same solenoid type corresponds to the solenoid type (1, 2 or 3), and is 0 otherwise. The label matrix gives labelled inter-solenoid distances. For example, consider a 70-residue protein, where residues 1-50 are beta-solenoid, and residues 51-70 are non-solenoid. The label matrix at indices from (1,1) to (50,50) would have a class value of 1, whereas the label matrix at all other indices would have a class value of 0. The label matrix was then padded identically to the distance matrix. Missing data and padding were assigned a separate class (4) to distinguish from the non-solenoid class.

Random cropping was used as a data augmentation technique

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to increase the amount of training data and improve model generalisation. Training distance matrix data consisted of 128x128 distance matrix slices centred on the main diagonal at a random offset. The randomly cropped distance matrix slices had corresponding slices of the label matrix as labels. However, the label matrices were of a size 64x64 and only contained label data for the central 64x64 slice of the corresponding distance matrix. The reasoning here is that the model could learn to better classify the pixels in a 64x64 slice of the distance matrix by 'seeing' a larger 128x128 context. Validation set distance matrices were split into overlapping 128x128 slices with an overlap of 64x64 and without random cropping. Validation set label matrices again corresponded to the central 64x64 slice of the distance matrix. Finally, the label matrices for both training and validation sets were one-hot encoded.

Model Architecture and Training. The model code was written in Python using the Tensorflow (41), Keras (42), Scikit-learn (43) and NumPy (39) libraries. The model has a modified U-Net architecture (28) with a reduced number of filters per layer (Fig. S1). Gaussian noise was applied to the input distance matrix with a standard deviation of 0.0001. The noise was added as a regularisation measure to reduce model overfitting and resulted in modest gains in performance. The convolutional layers within each block had the same number of filters of size 3x3 and stride of 1. All convolutional layers had an exponential linear unit (ELU) activation function (44). The number of filters doubled with each subsequent convolutional block in the encoder arm of the U-Net (Fig. S1, orange) and decreased by a factor of two in the decoder arm of the U-Net (Fig. S1, green). The fraction of activations zeroed by the dropout layer is shown for each convolutional block. Max pooling and transposed convolution with a filter size of 2x2 and a stride of 2 were used. Finally, a label matrix was output by a final convolutional layer with a softmax activation function highlighting regions in the distance matrix which belong to the different classes. The final upsampling block of the U-Net was removed to yield a 64x64 label matrix prediction for each 128x128 input distance matrix slice. Early stopping and model checkpointing callbacks were used with both callbacks tracking loss over the validation dataset. Training was performed for 100 epochs (approx. 1 day) with the categorical cross-entropy loss function and a batch size of 64 using the Adam (45) optimiser with a learning rate of 0.01 on an NVIDIA P1000 GPU. The model weights leading to the lowest loss value on the validation set were saved.

Inference and Further Analysis. For inference, an individual coordinate file and chain ID are specified. The structure (in mmCIF format) is loaded and Cα coordinates and residue indices are used to produce the Cα distance matrix. At inference, the distance matrix is processed analogously to the distance matrices in the validation set. The predicted one-hot label matrices are converted to integer class labels. The class of a residue is the label matrix value at indices (x,y) where x=y=residue index.The non-solenoid and missing data classes are merged.
Predictions are mapped onto the structure and visualised using the nglview library (46), while the sequence mapping is visualised using the Bokeh library (Bokeh Development Team, 2022) in a Jupyter notebook (47). User-defined repeat splitting was achieved by recording the indices of clicked equivalent positions in the nglview viewer. The recorded indices were then used to fill in repeats from the predicted solenoid residues. These repeats were then output in PDB format. The output repeats were processed using mTM-align (48), producing a PDB file with superposed repeats, TM-score and RMSD self-similarity matrices of the repeats, and a fasta file of the aligned repeat sequences. The fasta file was used to produce a logo plot of the repeat sequence alignment with the logomaker library (49).

Benchmark Method Predictions and Processing. For all comparison methods, predictions were collected using the respective servers (TAPO: https://bioinfo.crbm.cnrs.fr/tools/tapo/index_tapo.php; RepeatsDB-Lite: http://old.protein.bio.unipd.it/repeatsdb-lite/; PRIGSA2: https://bioinf.iiit.ac.in/PRIGSA2/) and adapted to the labelling scheme for this method. For TAPO, the top prediction was used. In contrast to SOLeNNoID, the other methods used detect individual repeats instead of individual amino acids. RepeatsDB-Lite and PRIGSA2 provide class labels, whereas TAPO does not. RepeatsDB-Lite and PRIGSA2 repeat predictions were treated as non-solenoid, if predicted to belong to an incorrect class outside the scope of SOLeNNoID (for example, if the ground truth class is alpha-solenoid but instead TIM-barrel is predicted). Precision, Recall, F1-score and multi-class Matthews correlation coefficient (MCC) were used to compare the performance of the solenoid detection algorithms. The scikit-learn classification_report was used to produce Precision, Recall and F1-score metrics and matthews_corrcoef to produce the Matthews correlation coefficient measure. Confusion matrices were produced using scikit-learn confusion_matrix and ConfusionMatrixDisplay.

PDB Solenoid Prediction Processing. mmCIF format files from the PDB as of 30 June 2021 were used. The data collected for each structure included the 4-letter PDB ID, chain ID, the number of residues for the structure and the number, percentage, and residue indices of solenoid predictions for all classes. PDB IDs from the training/validation and test datasets were removed from the final list of processed structures to limit the analysis to new protein chains. A protein chain was considered a member of a given solenoid class, if 50 percent or more of its residues were predicted to belong to this solenoid class. SOLeNNoID predictions for the whole PDB were produced locally over approximately 3.5 days on an M1 MacBook Air with 16 GB RAM.

RepeatsDB and DbStRiPs Solennoid Entry Selection. Data for RepeatsDB solenoid proteins were downloaded from https://www.repeatsdb.org/

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

classification/3. Subsequently, only entries with a 'Classification' tab beginning with '3.1', '3.2', or '3.3' were selected, corresponding to beta-, alpha/beta- and alpha-solenoid entries, respectively. The version of RepeatsDB accessed in this work used data from the PDB as of 14 August 2022. Data for DbStRiPs were downloaded from https://bioinf.iiit.ac.in/dbstrips/data.php, selecting Type 'Structural class' and Sub-type 'Alpha Solenoids', 'Beta Solenoids' and 'Alpha-Beta Solenoids'. The version of DbStRiPs accessed in this work used data from the PDB as of 03 June 2019.

Results

SOLeNNoID. The SOLeNNoID analysis pipeline takes as an input a PDB structure and a chain id. A Cα distance matrix is computed from the coordinates and a label matrix is predicted using the distance patterns in the distance matrix. Finally, per-residue class labels are produced by taking the solenoid class labels from the main diagonal of the label matrix. The class labels are mapped onto the input structure and sequence for visualisation. The algorithm can process arbitrarily large structures and has a quadratic time complexity ($O(n^2)$) (see Fig. S2). In a test on the 6,040 AlphaFoldDB v1 predicted protein structures for *S. cerevisiae*, predictions for structures around 2500 amino acids took less than a minute, whereas predictions for structures up to around 1200 amino acids took less than 10 seconds.

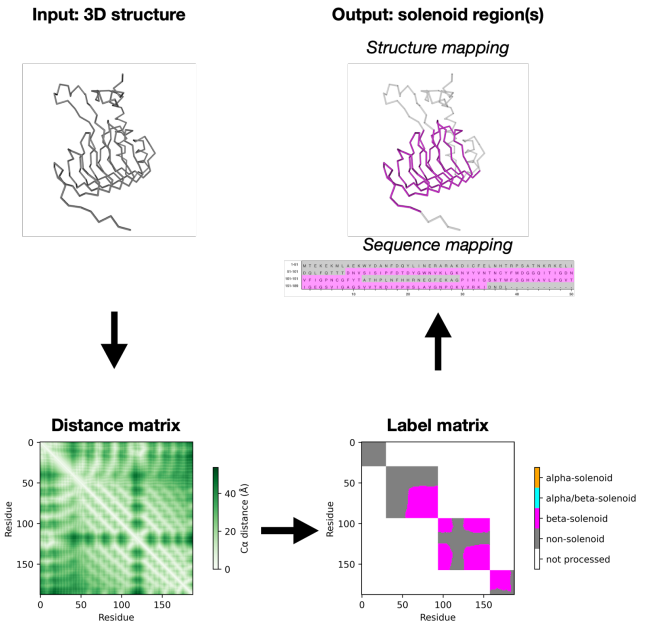

Fig. 1. Illustration of the SOLeNNoID analysis pipeline using PDB ID 3V4E. A distance matrix is calculated from the input protein structure. A label matrix is predicted from the distance matrix using a trained U-Net neural network. Finally, the label matrix classification is converted to a per-residue classification, which is mapped onto the structure and sequence of the input protein.

In addition to the user-friendly visualisation of solenoid predictions, the SOLeNNoID Jupyter notebook also allows interactive repeat definition. Users can quickly select equivalent residues in adjacent repeats in the nglview structure viewer, (Fig. 2 A left, green sticks). These residue selections allow the solenoid region predicted by SOLeNNoID (Fig. 2 A left, magenta ribbon) to be divided into individual repeats. The resulting repeats (Fig. 2 A right) are then output as individual PDB files and processed using mTM-align and logomaker to produce informative visualisations of repeat similarity on the structure (structural similarity matrices with RMSD (Fig. 2 B left) or TM-score (Fig. 2 B right) as the criterion and sequence (logo plot) (Fig. 2 C right) levels. For example, for PDB ID 3V4E, chain A the repeat similarity matrices indicate that repeats 1-4 form a group with high similarity between members. The sequence alignment suggests that the GWG sequence is an insertion found in only one of the repeats. Consensus residues are apparent at some of the remaining positions such as Asn at position 1 and Ile at position 20. The RepeatsDB database provides a similar output for each structure.

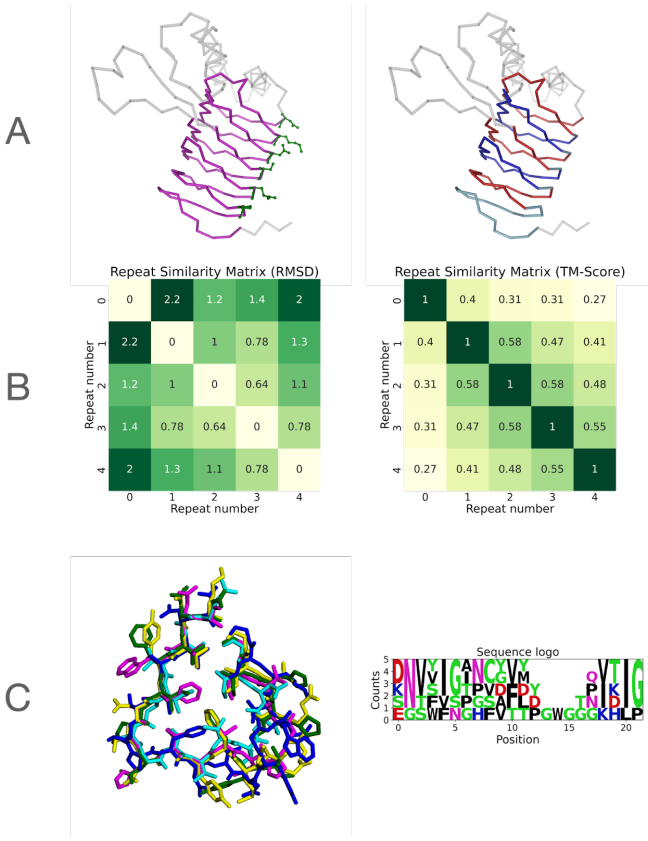

Fig. 2. Repeat splitting and analysis output for PDB ID 3V4E chain A. A) Left: SOLeNNoID beta-solenoid predictions in magenta backbone representation and user-selected equivalent residues in each turn in green ball and stick representation. Right: Automatically defined repeats based on the user input of equivalent residues in the SOLeNNoID predictions. Alternating repeats are shown in blue and red for ease of view. B) mTM-align output repeat similarity matrices using RMSD and TM-score. C) Left: Superposition of all repeats defined from SOLeNNoID predictions and user input in line representation. Each repeat is highlighted in a different colour. Right: Sequence logo (generated using the logomaker library) for the repeats using the fasta file produced by mTM-align. Counts in the Y-axis correspond to the number of repeats of 3V4E chain A in which a residue is observed at a given position. For example, at position 0, Asp is observed in 2 repeats, while Lys, Ser and Glu are observed in 1 repeat each.

Benchmarking against Other Tandem Repeat Detection Methods. SOLeNNoID was benchmarked against

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TAPO, PRIGSA2, and RepeatsDB-Lite on the solenoid structures in the test set. The four metrics used to describe performance on this dataset were precision, recall, F1-score (harmonic mean of precision and recall) and multi-class Matthews correlation coefficient (MCC). SOLeNNoID outputs solenoid residue predictions, rather than repeats. Therefore, repeat predictions of other methods were converted into a per-residue scheme. Unlike RDB-Lite and PRIGSA2, TAPO does not provide class labels. Therefore, there are two ways of comparing performance: one where it is assumed that TAPO would provide correct class labels and one where solenoid classification is treated as a binary classification of residues into a solenoid and a non-solenoid class.

If TAPO is assumed to provide correct class labels, SOLeNNoID had the best precision on the non-solenoid and beta-solenoid classes, the best recall on the alpha-solenoid class and the best F1-score on the non-solenoid and beta-solenoid classes (see 1). The PRIGSA2 method showed the highest precision on alpha/beta- and alpha-solenoid classes but had lower recall than TAPO and SOLeNNoID. Overall, SOLeNNoID outperformed PRIGSA2 and RDB-lite on F1-score and MCC and performed competitively with TAPO.

	Precision			
Class	SOLeNNoID	TAPO	PRIGSA2	RDB-lite
Non-solenoid	**0.786**	0.780	0.509	0.494
β-solenoid	**0.862**	0.770	0.857	0.502
α/β-solenoid	0.948	0.872	**0.962**	0.801
α-solenoid	0.806	0.912	**0.971**	0.886
	Recall			
Class	SOLeNNoID	TAPO	PRIGSA2	RDB-lite
Non-solenoid	0.760	0.729	**0.958**	0.892
β-solenoid	0.813	**0.836**	0.165	0.553
α/β-solenoid	0.849	**0.964**	0.781	0.106
α-solenoid	**0.914**	0.880	0.562	0.348
	F1-score			
Class	SOLeNNoID	TAPO	PRIGSA2	RDB-lite
Non-solenoid	**0.773**	0.754	0.665	0.636
β-solenoid	**0.837**	0.801	0.276	0.527
α/β-solenoid	0.896	**0.916**	0.862	0.188
α-solenoid	0.857	**0.896**	0.712	0.500
	MCC			
	SOLeNNoID	TAPO	PRIGSA2	RDB-lite
	0.764	**0.768**	0.564	0.384

Table 1. Comparison of method performance on the solenoid test set as a multi-class classification problem.

If all methods are instead used as binary solenoid/non-solenoid classifiers, SOLeNNoID outperformed TAPO in all metrics apart from recall on the solenoid class, where the two methods were equal (see Table S2). PRIGSA2 had the highest precision but very low recall on the solenoid class, ultimately resulting in a low F1-score. Therefore, while most solenoid predictions made by PRIGSA2 were true solenoid residues, many true solenoid residues were not detected. In comparison, TAPO and SOLeNNoID had a more balanced performance.

The full confusion matrices for all methods are shown in Fig. S3. Only SOLeNNoID and TAPO had large values across the entire diagonal (corresponding to residues with agreement between predicted and true label). PRIGSA2 and RDB-Lite incorrectly predicted many alpha- and beta-solenoid residues as non-solenoid. PRIGSA2 showed little to no incorrect predictions between solenoid classes. This matches the high precision but poor recall for PRIGSA2 in Table 1. RepeatsDB-Lite predicted a large number of alpha-solenoid residues as non-solenoid, which is due to some alpha-solenoid residues not being predicted as tandem repeat at all and some being predicted as a non-solenoid repeat type such as TIM-barrel or alpha-barrel. In addition, RepeatsDB-Lite incorrectly predicted many alpha/beta-solenoid residues as beta-solenoid. Both tandem repeat misclassifications by RepeatsDB-Lite are due to an incorrect classification of the master repeat unit used to build up the solenoid region. TAPO and SOLeNNoID had fewer incorrect predictions of solenoid residues as non-solenoid compared to PRIGSA2 and RepeatsDB-Lite. The major type of incorrect predictions for TAPO were non-solenoid residues predicted as beta-solenoid. SOLeNNoID instead incorrectly predicted non-solenoid residues as alpha-solenoid. This is possibly due to similarities between the distance matrices of non-solenoid alpha-helical proteins and alpha-solenoid proteins.

SOLeNNoID also performed well on the full test set, which includes non-solenoid structures (Table S3). On introduction of non-solenoid residues, precision and recall for the non-solenoid class increased, whereas precision for the alpha-solenoid class decreased. Precision, recall and F1-score values for solenoid classes were nearly identical to the values on the solenoid test set only apart from precision on alpha-solenoid residues, which was now lower on introduction of non-solenoid structures. This can also be seen in the bottom right corner of the confusion matrix in Fig. S4. Comparison with the confusion matrix on the solenoid test set indicates that the model incorrectly classifies an additional 256 non-solenoid residues as alpha-solenoid.

SOLeNNoID PDB Predictions. 599,443 chains (175,084 unique PDB IDs) from the PDB were processed using SOLeNNoID. Exploration of the list of alpha-solenoid predictions revealed 6 PDB IDs with over 100 chains each (3J3Q, 3J3Y, 6X63, 6QVK, 6QZ0, 6QYD). Manual inspection identified that these structures are not solenoid-containing but instead contain elements which resemble alpha-solenoid repeats. These structures were removed, leading to the final predictions outlined in Table S4. The total number of unique chains identified as containing solenoid regions by SOLeNNoID was 9,326, or about 1.56 percent of all processed PDB entries. Considering only unique PDB IDs, 4,148 PDB IDs were identified as majority solenoid, representing 2.37 percent of processed PDB IDs. This result agrees with the work of Chakrabarty and Parekh (30), who also found that approximately 2 percent of PDB entries contain solenoid regions. The total number of unique PDB IDs is lower than the sum of the numbers of unique PDB IDs of individual solenoid classes, as a single PDB ID can contain multiple solenoid chains of different classes. Therefore, using unique PDB IDs underestimates the number of solenoid structures. The mean number of solenoid residues

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for alpha-solenoid proteins was the highest at 348, followed by alpha/beta-solenoid proteins at 295, and beta-solenoids at 198 residues. The maximum number of solenoid residues detected in a single chain was highest for alpha-solenoid proteins (3,340: PDB ID 6TAX chain A), followed by alpha-beta-solenoid proteins (732: PDB ID 6S6Q chain B), and finally beta-solenoid proteins (631: PDB ID 6Z7P chain A). SOLeNNoID predictions also suggested that alpha-solenoid entries were by far the most abundant in the PDB. For example, there were 7.5 times more predicted alpha-solenoid chains than alpha/beta-solenoid chains.

Comparison of PDB Predictions with Other Tandem Repeat Databases. Solenoid entries in the RepeatsDB and DbStRiPs databases had significant overlap with SOLeNNoID predictions on the PDB (see Fig. 3). There were 1,139 PDB IDs in common between RepeatsDB, DbStRiPs, and SOLeNNoID predictions, and a further 1,005 PDB IDs, which were shared between either RepeatsDB and SOLeNNoID predictions, or DbStRiPs and SOLeNNoID predictions. Therefore, SOLeNNoID predictions were supported by agreement with databases populated by distinct tandem repeat detection methods. However, SOLeNNoID also identified many PDB entries not covered by these databases. Due to detecting multiple chains from protein complexes, the number of solenoid-containing chains predicted by SOLeNNoID was about 2.6 times larger than the number of solenoid-containing PDB IDs. SOLeNNoID predictions had the second largest number of total solenoid chains (9,326) and method/database-specific solenoid chains (5,321).

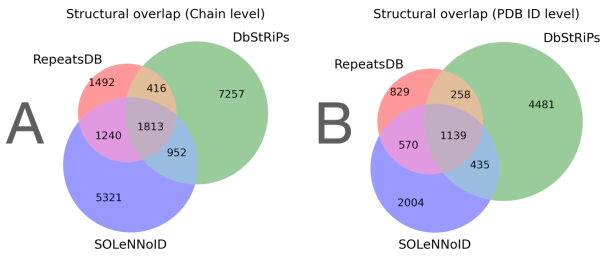

Fig. 3. Venn diagrams showing overlap between the structures detected as solenoid by SOLeNNoID and solenoid structures in the RepeatsDB and DbStRiPs databases. Red - solenoid structures found only in the RepeatsDB database, green - solenoid structures found only in the DbStRiPs database, blue - solenoid structures found only using SOLeNNoID. Grey - solenoid structures found in all databases, pink - solenoid structures found both in RepeatsDB and by SOLeNNoID, cyan - solenoid structures found both in DbStRiPs and by SOLeNNoID, yellow - solenoid structures found both in RepeatsDB and DbStRiPs. A) Overlap between databases and SOLeNNoID predictions when considering unique individual chains. B) Overlap between databases and SOLeNNoID predictions when considering unique PDB IDs.

Table 2 shows the number of putative solenoid chains and PDB IDs found by SOLeNNoID with PDB IDs distinct from the entries in RepeatsDB and DbStRiPs. A total of 4,638 new chains and 2,004 unique PDB IDs were found with most belonging to the alpha-solenoid class. This means that 683 additional chains were discovered by SOLeNNoID for PDB IDs represented in either/both of the comparison databases. The discovery of additional solenoid structures by SOLeNNoID is due to two factors: 1) the ability of SOLeNNoID to detect examples which are not accessible to other methods, such as RepeatsDB-Lite and PRIGSA2, owing to, for example, the inability of these methods to process very large structures and 2) using a more recent version of the PDB with a larger number of chains and IDs. While the latter point applies when comparing SOLeNNoID predictions with DbStRiPs, it does not apply to RepeatsDB, as the most recent version of the database (v3.2) is based on a version of the PDB from 2022, whereas SOLeNNoID was used to process a version of the PDB from 2021. The scale of our contribution to solenoid protein annotation is therefore demonstrated as a 71.6 percent increase over the "gold standard" RepeatsDB database. These results further highlight the need for multiple approaches to solenoid detection. Different methods show significant overlap but also complementarity and, when used together, enable a better coverage of solenoid protein structures and regions.

	α-solenoid	α/β-solenoid	β-solenoid	Total
Number of chains	3964	329	345	4638
Number of PDB IDs	1746	163	120	2004
% increase over RepeatsDB (PDB IDs)	86.9	48.2	26	71.6

Table 2. Number of structures detected by SOLeNNoID and not present in other databases and percentage increase over entries in RepeatsDB.

Discussion

SOLeNNoID uses a CNN trained on alpha carbon distance matrices to rapidly process arbitrarily large structures and provide user-friendly output of solenoid region predictions mapped to both structure and sequence. We also present downstream interactive repeat analysis to extract structural and sequence similarities between repeats.

SOLeNNoID differs from other solenoid detection methods in that it highlights individual solenoid residues rather than repeats. Test set benchmark data shows that SOLeNNoID is competitive with TAPO and outperforms RepeatsDB-Lite and PRIGSA2. Furthermore, SOLeNNoID has balanced precision and recall. Therefore, it can be used to both annotate solenoid regions and detect new solenoid proteins, which may be overlooked if either RepeatsDB-Lite or PRIGSA2 are used. RepeatsDB-Lite depends heavily on the quality of the repeating unit library it uses to structurally align against a query protein, as well as the 'Master' unit, which defines the class of the repeating region. Some incorrect repeat assignments by RepeatsDB-Lite are due to incorrect 'Master' unit assignments - such as alpha-solenoid proteins detected as alpha-barrels, and alpha/beta-solenoid proteins detected as beta-solenoids. TAPO, in contrast, has a balanced performance as it uses several structural features in a SVM model. However, TAPO does not provide class labels. A limitation of SOLeNNoID is that it is dependent on the quality of the training dataset and solenoid proteins far outside the structural space covered by this dataset may be incorrectly predicted or omitted. All methods show certain biases on the test dataset presented here, whether predicting solenoid residues as non-solenoid, or vice versa, or confusing solenoid classes. These biases highlight the need for a consensus approach to

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

solenoid region prediction by using the outputs of several different repeat detection methods.

Using SOLeNNoID, we have detected hundreds to thousands of new proteins with the different solenoid classes from the largest database of experimentally determined structures, the PDB. In our analysis, 50 percent of residues in a chain must be detected as solenoid - this is a stringent threshold used to remove false positive solenoid protein hits. However, this will also have the effect of under-estimating true solenoid protein numbers as solenoids with partial coverage are omitted. The PDB search for solenoid proteins we conducted revealed substantial overlap with existing databases of solenoid proteins but also complementarity of the different methods. We present over 5000 new protein chains and over 2000 new PDB IDs not covered by existing databases - a 71 percent increase over the current version of RepeatsDB. Our findings suggest that the best way to address the challenge of finding new solenoid proteins is to combine multiple approaches.

With the rise of predicted structural databases with hundreds of millions of structures such as the AlphaFold and ESMFold databases, efficient analysis of structures is a key bioinformatic task. Our solenoid detection system is a contribution to analysing this flood of structural information.

ACKNOWLEDGEMENTS

We acknowledge computational resources and support provided by the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232). Georgi Nikov was supported by a EPSRC DTP training grant (EP/R513052/1).

Bibliography

1. Fabio Parmeggiani and Po Ssu Huang. Designing repeat proteins: a modular approach to protein design, 8 2017. ISSN 1879033X.
2. Matteo Delucchi, Elke Schaper, Oxana Sachenkova, Arne Elofsson, and Maria Anisimova. A new census of protein tandem repeats and their relationship with intrinsic disorder. *Genes*, 11:407, 4 2020. ISSN 20734425. doi: 10.3390/genes11040407.
3. Andrey V. Kajava. Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology*, 179:279–288, 9 2012. ISSN 10478477. doi: 10.1016/j.jsb.2011.08.009.
4. Sam Manna. An overview of pentatricopeptide repeat proteins and their applications. *Biochimie*, 113:93–99, 6 2015. ISSN 0300-9084. doi: 10.1016/J.BIOCHI.2015.04.004.
5. James G. Ferry. The γ class of carbonic anhydrases. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1804:374–381, 2 2010. ISSN 1570-9639. doi: 10.1016/J.BBAPAP.2009.08.026.
6. Abirami Baskaran, Manigundan Kaari, Gopikrishnan Venugopal, Radhakrishnan Manikkam, Jerrine Joseph, and Parli V. Bhaskar. Anti freeze proteins (afp): Properties, sources and applications - a review. *International journal of biological macromolecules*, 189:292–305, 10 2021. ISSN 1879-0003. doi: 10.1016/J.IJBIOMAC.2021.08.105.
7. Jessica K. Bell, Gregory E.D. Mullen, Cynthia A. Leifer, Alessandra Mazzoni, David R. Davies, and David M. Segal. Leucine-rich repeats and pathogen recognition in toll-like receptors. *Trends in Immunology*, 24:528–533, 10 2003. ISSN 14714906. doi: 10.1016/S1471-4906(03)00242-4.
8. Frances Gidley and Fabio Parmeggiani. Repeat proteins: designing new shapes and functions for solenoid folds. *Current Opinion in Structural Biology*, 68:208–214, 6 2021. ISSN 0959-440X. doi: 10.1016/J.SBI.2021.02.002.
9. Alex Bateman, Maria Jesus Martin, Sandra Orchard, Michele Magrane, Rahat Agivetova, Shadab Ahmad, Emanuele Alpi, Emily H. Bowler-Barnett, Ramona Britto, Borisas Bursteinas, Hema Bye-A-Jee, Ray Coetzee, Austra Cukura, Alan Da Silva, Paul Denny, Tunca Dogan, Thank God Ebenezer, Jun Fan, Leyla Garcia Castro, Penelope Garmiri, George Georghiou, Leonardo Gonzales, Emma Hatton-Ellis, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Petteri Jokinen, Vishal Joshi, Dushyanth Jyothi, Antonia Lock, Rodrigo Lopez, Aurelien Luciani, Jie Luo, Yvonne Lussi, Alistair MacDougall, Fabio Madeira, Mahdi Mahmoudy, Manuela Menchi, Alok Mishra, Katie Moulang, Andrew Nightingale, Carla Susana Oliveira, Sangya Pundir, Guoying Qi, Shriya Raj, Daniel Rice, Milagros Rodriguez Lopez, Rabie Saidi, Joseph Sampson, Tony Sawford, Elena Speretta, Edward Turner, Nidhi Tyagi, Preethi Vasudev, Vladimir Volynkin, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie Claude Blatter, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casals-Casas, Edouard de Castro, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Guillaume Keller, Arnaud Kerhornou, Vicente Lara, Philippe Le Mercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Batista Neto, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Christian Sigrist, Karin Sonesson, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S. Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai Su Yeh, and Jian Zhang. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49:D480–D489, 1 2021. ISSN 0305-1048. doi: 10.1093/NAR/GKAA1100.
10. Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank, 1 2000. ISSN 03051048.
11. Stephen K. Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V. Crichlow, Cole H. Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M. Duarte, Shuchismita Dutta, Zukang Feng, Sai Ganesan, David S. Goodsell, Sutapa Ghosh, Rachel Kramer Green, Vladimir Guranovic, Dmytro Guzenko, Brian P. Hudson, Catherine L. Lawson, Yuhe Liang, Robert Lowe, Harry Namkoong, Ezra Peisach, Irina Persikova, Chris Randle, Alexander Rose, Yana Rose, Andrej Sali, Joan Segura, Monica Sekharan, Chenghua Shao, Yi Ping Tao, Maria Voigt, John D. Westbrook, Jasmine Y. Young, Christine Zardecki, and Marina Zhuravleva. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 49:D437–D451, 1 2021. ISSN 0305-1048. doi: 10.1093/NAR/GKAA1038.
12. Radek Szklarczyk and Jaap Heringa. Tracking repeats using significance and transitivity. *Bioinformatics*, 20:i311–i317, 8 2004. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTH911.
13. Julien Jorda and Andrey V. Kajava. T-reks: identification of tandem repeats in sequences with a k-means based algorithm. *Bioinformatics*, 25:2632–2638, 10 2009. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTP482.
14. A. Biegert and J. Söding. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, 24:807–814, 3 2008. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTN039.
15. Luca Marsella, Francesco Sirocco, Antonio Trovato, Flavio Seno, and Silvio C.E. Tosatto. Repetita: Detection and discrimination of the periodicity of protein solenoid repeats by discrete fourier transform. volume 25, pages 289–295. Oxford Academic, 6 2009. doi: 10.1093/bioinformatics/btp232.
16. David Fournier, Gareth A. Palidwor, Sergey Shcherbinin, Angelika Szengel, Martin H. Schaefer, Carol Perez-Iratxeta, and Miguel A. Andrade-Navarro. Functional and genomic analyses of alpha-solenoid proteins. *PLoS ONE*, 8:79894, 11 2013. ISSN 19326203. doi: 10.1371/journal.pone.0079894.
17. Gareth A. Palidwor, Sergey Shcherbinin, Matthew R. Huska, Tamas Rasko, Ulrich Stelzl, Anup Arumughan, Raphaele Foulle, Pablo Porras, Luis Sanchez-Pulido, Erich E. Wanker, and Miguel A. Andrade-Navarro. Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Computational Biology*, 5:1000304, 3 2009. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000304.
18. Guillaume Pagès and Sergei Grudinin. Deepsymmetry: using 3d convolutional networks for identification of tandem repeats and internal symmetries in protein structures. *Bioinformatics*, 35:5113–5120, 12 2019. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTZ454.
19. Kanak Garg and Saksham Gupta. Deep-strip: Deep learning approach for structural repeat identification in proteins. *ACM International Conference Proceeding Series*, pages 48–54, 5 2022. doi: 10.1145/3543377.3543385.
20. Matteo Delucchi, Paulina Näf, Spencer Bliven, and Maria Anisimova. Tral 2.0: Tandem repeat detection with circular profile hidden markov models and evolutionary aligner. *Frontiers in Bioinformatics*, 0:20, 6 2021. ISSN 2673-7647. doi: 10.3389/FBINF.2021.691865.
21. Elke Schaper, Alexander Korsunsky, Julija Pečerska, Antonio Messina, Riccardo Murri, Heinz Stockinger, Stefan Zoller, Ioannis Xenarios, and Maria Anisimova. Tral: tandem repeat annotation library. *Bioinformatics*, 31:3051–3053, 9 2015. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTV306.
22. Broto Chakrabarty and Nita Parekh. Prigsa2: Improved version of protein repeat identification by graph spectral analysis. *Journal of Biosciences*, 45:1–16, 12 2020. ISSN 09737138. doi: 10.1007/S12038-020-00058-X/FIGURES/8.
23. Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50:D439–D444, 1 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1061.
24. Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 3 2023. ISSN 10959203. doi: 10.1126/SCIENCE.ADE2574/SUPPL_FILE/SCIENCE.ADE2574_SM.PDF.
25. Phuong Do Viet, Daniel B. Roche, and Andrey V. Kajava. Tapo: A combined method for the identification of tandem repeats in protein structures, 9 2015. ISSN 18733468.
26. Layla Hirsh, Lisanna Paladin, Damiano Piovesan, and Silvio C.E. Tosatto. Repeatsdb-lite: A web server for unit annotation of tandem repeat proteins. *Nucleic Acids Research*, 46:W402–W407, 7 2018. ISSN 13624962. doi: 10.1093/nar/gky360.
27. Bostjan Kobe and Andrey V. Kajava. When protein folding is simplified to protein coiling: The continuum of solenoid protein structures, 10 2000. ISSN 09680004.
28. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:234–241, 5 2015. ISSN 16113349.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29. Tomás Di Domenico, Emilio Potenza, Ian Walsh, R. Gonzalo Parra, Manuel Giollo, Giovanni Minervini, Damiano Piovesan, Awais Ihsan, Carlo Ferrari, Andrey V. Kajava, and Silvio C.E. Tosatto. Repeatsdb: A database of tandem repeat protein structures. *Nucleic Acids Research*, 42:1–6, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1175.
30. Broto Chakrabarty and Nita Parekh. Dbstrips: Database of structural repeats in proteins. *Protein Science*, 2021. ISSN 1469896X. doi: 10.1002/PRO.4052.
31. Liisa Holm. Dali and the persistence of protein shape. *Protein Science*, 29:128–140, 1 2020. ISSN 0961-8368. doi: 10.1002/pro.3749.
32. Andrey V. Kajava and Alasdair C. Steven. β-rolls, β-helices, and other β-solenoid proteins, 2006. ISSN 00653233.
33. LLC Schrödinger and Warren DeLano. Pymol.
34. Sameer Velankar, Glen Van Ginkel, Younes Alhroub, Gary M. Battle, John M. Berrisford, Matthew J. Conroy, Jose M. Dana, Swanand P. Gore, Aleksandras Gutmanas, Pauline Haslam, Pieter M.S. Hendrickx, Ingvar Lagerstedt, Saqib Mir, Manuel A.Fernandez Montecelo, Abhik Mukhopadhyay, Thomas J. Oldfield, Ardan Patwardhan, Eduardo Sanz-García, Sanchayita Sen, Robert A. Slowley, Michael E. Wainwright, Mandar S. Deshpande, Andrii Iudin, Gaurav Sahni, Jose Salavert Torres, Miriam Hirshberg, Lora Mak, Nurul Nadzirin, David R. Armstrong, Alice R. Clark, Oliver S. Smart, Paul K. Korir, and Gerard J. Kleywegt. Pdbe: Improved accessibility of macromolecular structure data from pdb and emdb. *Nucleic Acids Research*, 44:D385–D395, 2016. ISSN 13624962. doi: 10.1093/nar/gkv1047.
35. Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Research*, 33:2302–2309, 4 2005. ISSN 0305-1048. doi: 10.1093/NAR/GKI524.
36. Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. Scop: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 4 1995. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80134-2.
37. Michael Knudsen and Carsten Wiuf. The cath database. *Human Genomics*, 4:207, 2 2010. ISSN 14797364. doi: 10.1186/1479-7364-4-3-207.
38. Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25:1422–1423, 3 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163.
39. Stéfan Van Der Walt, S. Chris Colbert, and Gaël Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13:22–30, 3 2011. ISSN 15219615. doi: 10.1109/MCSE.2011.37.
40. G Van Rossum and F L Drake. Python 3 reference manual; createspace. *Scotts Valley, CA*, page 242, 2009.
41. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. 3 2016.
42. François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.
43. Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 10 2011. ISSN 15324435.
44. Djork Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 11 2015.
45. Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. International Conference on Learning Representations, ICLR, 12 2015.
46. Hai Nguyen, David A. Case, and Alexander S. Rose. Nglview–interactive molecular graphics for jupyter notebooks. *Bioinformatics*, 34:1241–1242, 4 2018. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTX789.
47. Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter development team. Jupyter notebooks - a publishing format for reproducible computational workflows. In Fernando Loizides and Birgit Scmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90, Netherlands, 2016. IOS Press.
48. Runze Dong, Zhenling Peng, Yang Zhang, and Jianyi Yang. mtm-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics*, 34:1719, 5 2018. ISSN 14602059. doi: 10.1093/BIOINFORMATICS/BTX828.
49. Ammar Tareen and Justin B. Kinney. Logomaker: beautiful sequence logos in python. *Bioinformatics*, 36:2272–2274, 4 2020. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTZ921.
50. Wes Mckinney. Data structures for statistical computing in python. 2010.
51. Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6:3021, 4 2021. ISSN 2475-9066. doi: 10.21105/JOSS.03021.
52. John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science and Engineering*, 9:99–104, 5 2007. ISSN 15219615. doi: 10.1109/MCSE.2007.55.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Methods

Training Set				
	α-solenoid	α/β -solenoid	β-solenoid	Non-solenoid
Number of structures	26	13	41	196
Avg. number of residues	292	263	310	255
Avg. % of solenoid residues	78	78	60	0
Number of residues	6031	2735	7201	57988
Validation Set				
	α-solenoid	α/β-solenoid	β-solenoid	Non-solenoid
Number of structures	7	4	11	50
Avg. number of residues	346	246	327	269
Avg. % of solenoid residues	75	75	53	0
Number of residues	1880	756	1813	16016
Test Set				
	α-solenoid	α/β-solenoid	β-solenoid	Non-solenoid
Number of structures	11	6	15	49
Avg. number of residues	581	513	450	263
Avg. % of solenoid residues	74	78	55	0
Number of residues	4023	2428	3265	18833

Table S1. Training, validation and test set statistics.

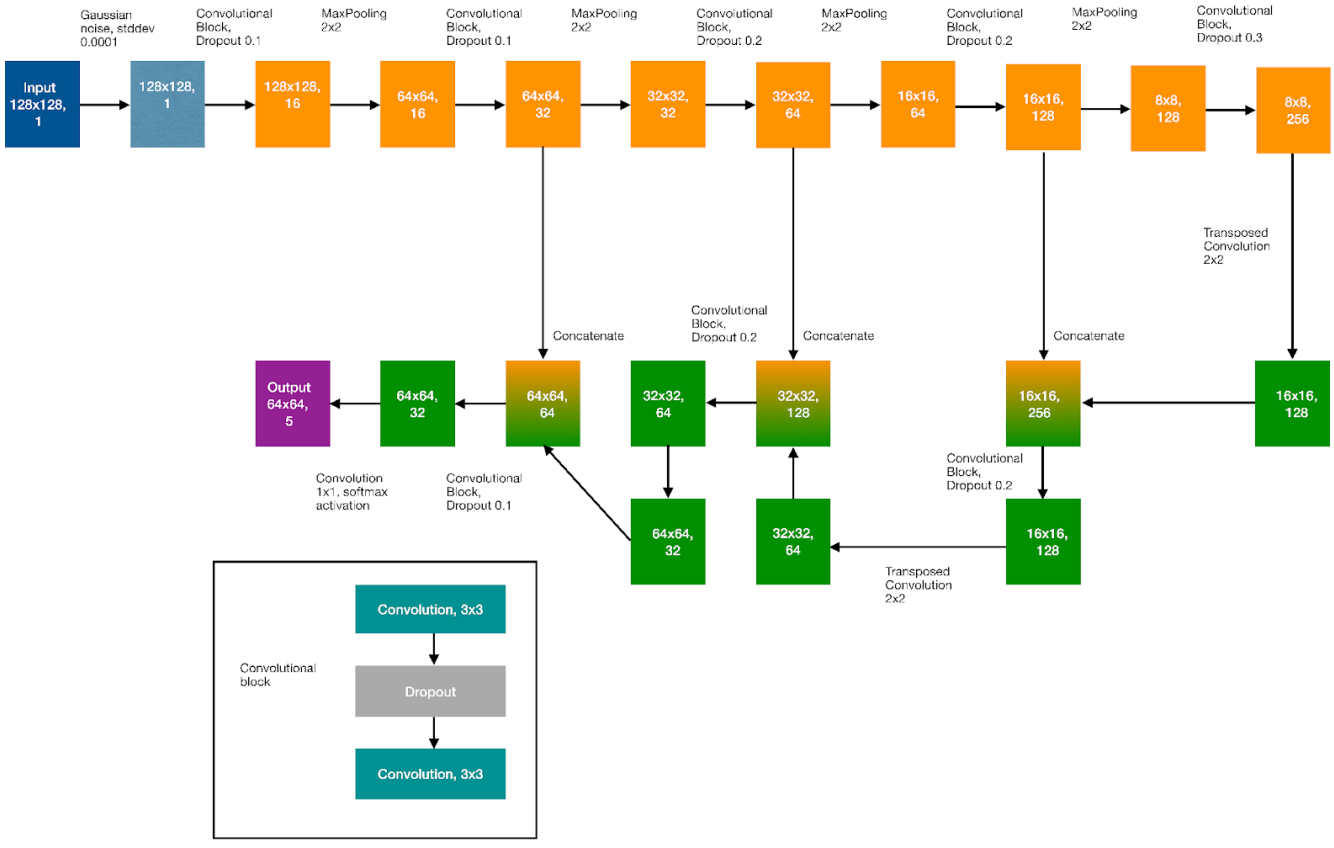

Fig. S1. Illustration of the U-Net architecture used in this work. Each convolutional block consists of the three layers shown in the box. Each square denotes the dimensions of the input, intermediate tensor or output.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Prediction Time Test. The prediction time test was conducted using the 6,040 AlphaFold2-predicted S. cerevisiae structures (v1 database) in the EBI AlphaFold database. The script was executed on the Imperial College HPC with 16 CPU cores, 96 GB RAM and 4 RTX6000 GPUs. The prediction time was defined from before loading a structure to after obtaining solenoid predictions for the structure.

Data Processing and Visualisation. Data analysis was carried out using the Pandas library (50) in Python3 (40). Data visualisation was carried out using the Seaborn (51), matplotlib (52), matplotlib-venn (https://github.com/konstantint/matplotlib-venn) and logomaker (49) libraries. Protein structure visualisation was carried out using the nglview library (46) in Python3.

Supplementary Results

Prediction Time Test. To evaluate the time taken by SOLeNNoID to process a structure and make predictions, a test was conducted on the 6,040 v1 AlphaFoldDB predicted protein structures for S. cerevisiae. The result in Fig. S3 below shows that prediction time had a second order polynomial relationship with respect to the length of a protein. However, predictions for structures around 2500 amino acids took less than a minute under the conditions used, whereas predictions for structures up to around 1200 amino acids took less than 10 seconds. This demonstrates that SOLeNNoID can rapidly produce predictions even for large structures. In comparison, the TAPO, PRIGSA2 and RepeatsDB-Lite servers can take on the order of minutes to produce a prediction for a structure over 1000 amino acids.

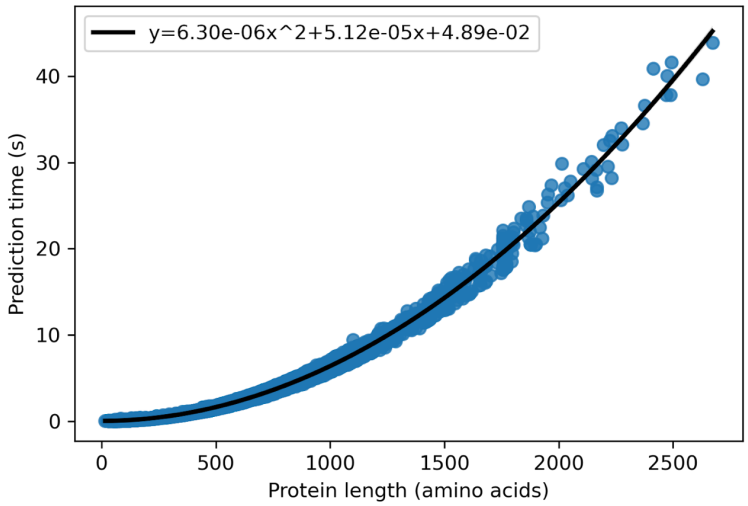

Fig. S2. Relationship between protein length and prediction time for SOLeNNoID method on the YEAST v1 AlphaFoldDB dataset. Top left - second order polynomial equation fit to the data using numpy.polyfit.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Precision			
Class	SOLeNNoID	TAPO	PRIGSA2	RDB-lite
Non-solenoid	**0.786**	0.780	0.509	0.494
Solenoid	0.870	0.855	**0.954**	0.892
	Recall			
Class	SOLeNNoID	TAPO	PRIGSA2	RDB-lite
Non-solenoid	0.760	0.729	**0.958**	0.892
Solenoid	**0.886**	**0.886**	0.488	0.494
	F1-score			
Class	SOLeNNoID	TAPO	PRIGSA2	RDB-lite
Non-solenoid	**0.773**	0.754	0.665	0.636
Solenoid	**0.877**	0.870	0.646	0.636
	MCC			
	SOLeNNoID	TAPO	PRIGSA2	RDB-lite
	0.651	0.625	0.455	0.387

Table S2. Comparison of method performance on the solenoid test set as a binary (solenoid/non-solenoid) classification problem. The metrics used are precision, recall and F1-score for single classes and multi-class MCC as a global metric. Values in bold represent the best performance across methods.

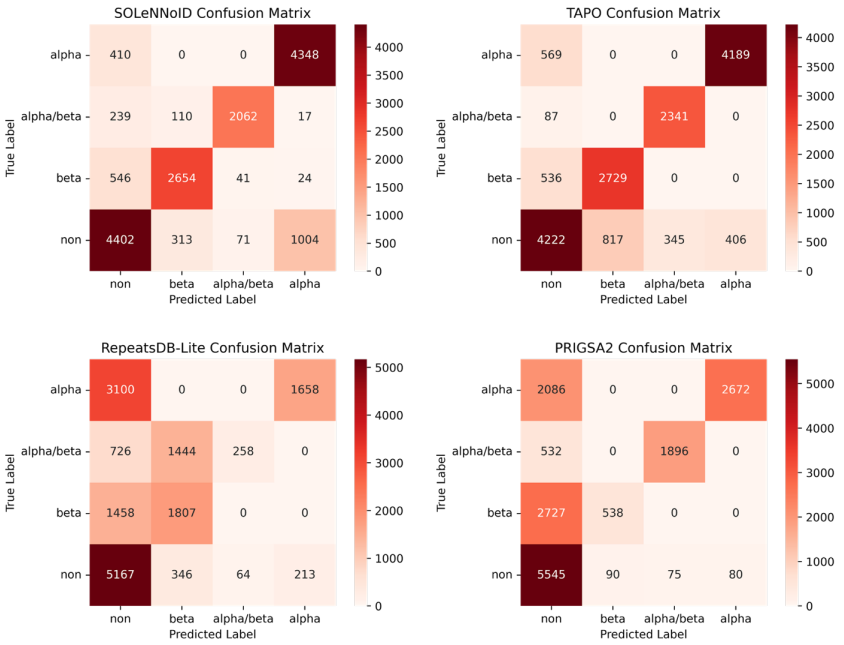

Fig. S3. Multi-class confusion matrices for SOLeNNoID, TAPO, PRIGSA2 and RepeatsDB-Lite on the solenoid test set. True labels are shown on the y-axis and predicted labels are shown on the x-axis. The numbers represent the number of protein residues placed within each category of the confusion matrix. A darker red colour indicates a larger number of residues.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.22.604558; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Class	Precision	Recall	F1-score
Non-solenoid	0.935	0.912	0.924
β-solenoid	0.860	0.813	0.836
α/β-solenoid	0.948	0.849	0.896
α-solenoid	0.770	0.914	0.835
MCC			
0.811			

Table S3. Performance of SOLeNNoID on the whole test set as a multi-class classification problem.

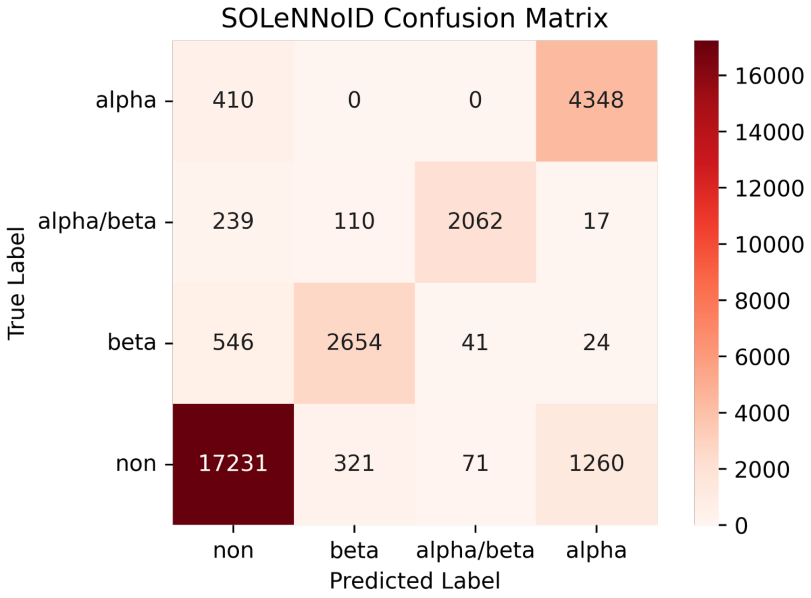

Fig. S4. Multi-class confusion matrix for the SOLeNNoID model on the full test dataset. True labels are shown on the y-axis and predicted labels are shown on the x-axis. The numbers represent the number of protein residues placed within each category of the confusion matrix. A darker red colour indicates a larger number of residues.

	α-solenoid	α/β-solenoid	β-solenoid
Number of chains	7371	981	974
Number of PDB IDs	3326	490	366
Min. number of solenoid residues	36	49	28
Mean number of solenoid residues	348	273	198
Max. number of solenoid residues	3340	732	631

Table S4. Statistics for PDB entries detected as solenoids by SOLeNNoID.