

METHODS

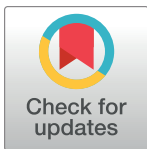
Structure-aware annotation of leucine-rich repeat domains

Boyan Xu^{1,2}, Alois Cerbu², Christopher J. Tralie³, Daven Lim⁴, Ksenia Krasileva^{1,4*}

1 Center for Computational Biology, University of California Berkeley, Berkeley, California, United States of America, **2** Department of Mathematics, University of California Berkeley, Berkeley, California, United States of America, **3** Department of Mathematics and Computer Science, Ursinus College, Collegeville, Pennsylvania, United States of America, **4** Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, California, United States of America

☞ These authors contributed equally to this work.

* kseniak@berkeley.edu



OPEN ACCESS

Citation: Xu B, Cerbu A, Tralie CJ, Lim D, Krasileva K (2024) Structure-aware annotation of leucine-rich repeat domains. PLoS Comput Biol 20(11): e1012526. <https://doi.org/10.1371/journal.pcbi.1012526>

Editor: Ozlem Keskin, Koç University, TÜRKIYE

Received: November 1, 2023

Accepted: September 29, 2024

Published: November 5, 2024

Copyright: © 2024 Xu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: A Jupyter notebook for running the winding number LRR annotator is available at <https://github.com/amcerbu/LRR-Annotation/tree/main>.

Funding: BX has been supported by Department of Energy Computational Sciences Graduate Fellowship grant number DE-SC0020347. KK has been supported by funding from the Innovative Genomics Institute (<https://innovativegenomics.org/>), the Gordon and Betty Moore Foundation (<https://www.moore.org/>), grant number 8802, and the National Institute of Health New Innovator Director's Award, grant number DP2AT011967.

Abstract

Protein domain annotation is typically done by predictive models such as HMMs trained on sequence motifs. However, sequence-based annotation methods are prone to error, particularly in calling domain boundaries and motifs within them. These methods are limited by a lack of structural information accessible to the model. With the advent of deep learning-based protein structure prediction, existing sequenced-based domain annotation methods can be improved by taking into account the geometry of protein structures. We develop dimensionality reduction methods to annotate repeat units of the Leucine Rich Repeat solenoid domain. The methods are able to correct mistakes made by existing machine learning-based annotation tools and enable the automated detection of hairpin loops and structural anomalies in the solenoid. The methods are applied to 127 predicted structures of LRR-containing intracellular innate immune proteins in the model plant *Arabidopsis thaliana* and validated against a benchmark dataset of 172 manually-annotated LRR domains.

Author summary

In immune receptors across various organisms, repeating protein structures play a crucial role in recognizing and responding to pathogen threats. These structures resemble the coils of a slinky toy, allowing these receptors to adapt and change over time. One particularly vital but challenging structure to study is the Leucine Rich Repeat (LRR). Traditional methods that rely just on analyzing the sequence of these proteins can miss subtle changes due to rapid evolution. With the introduction of protein structure prediction tools like AlphaFold 2, annotation methods can study the coarser geometric properties of the structure. In this study, we visualize LRR proteins in three dimensions and use a mathematical approach to 'flatten' them into two dimensions, so that the coils form circles. We then used a mathematical concept called winding number to determine the number of repeats and where they are in a protein sequence. This process helps reveal their repeating patterns with enhanced clarity. When we applied this method to immune receptors from a model plant organism, we found that our approach could accurately identify coiling

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

patterns. Furthermore, we detected errors made by previous methods and highlighted unique structural variations. Our research offers a fresh perspective on understanding immune receptors, potentially influencing studies on their evolution and function.

Introduction

Solenoid domains are a class of protein structures defined by a repeating helical arrangement of their backbone chain. These domains are found in a diverse range of proteins and play important roles in a variety of biological processes, including protein-protein interactions, molecular recognition, and scaffolding [18]. The coil shape of solenoid domains arises from a repeating motif of amino acid residues, known as *tandem repeat units*. The specific amino acid sequence and length of the repeating unit can vary between solenoid domains, resulting in differences in the overall structure and function of the domain. The modular nature of solenoid domains allows for the construction of complex structures by combining different domains in a predictable and controlled manner [6].

Leucine-rich repeat (LRR) domains are a type of curved solenoid domain with repeated units of about 20–30 residues long which contain leucine residues in a beta-strand conformation. These domains are found in a wide range of proteins, including cell surface receptors, enzymes, and structural proteins, and are known to play important roles in protein-protein interactions, signal transduction, and immune recognition [12].

Leucine-rich repeats play a critical role in the function of the NOD-like receptor (NLR) family of proteins in the innate immune system of plants and animals [17]. NLRs are intracellular immune receptors that recognize pathogen-derived molecules and activate downstream signaling pathways to initiate an immune response. NLRs are involved in the recognition of a wide range of pathogens, including bacteria, fungi, and viruses. NLRs typically consist of three domains: an N-terminal domain, a central nucleotide-binding domain, and a C-terminal LRR domain. The LRR domain is responsible for recognizing and binding to pathogen-derived molecules, such as effector proteins or pathogen-associated molecular patterns (PAMPs) [8]. In particular, the LRR domains of plant NLRs are highly diverse and can recognize a wide range of pathogen-derived molecules, allowing plants to mount a robust and specific immune response to a broad range of pathogens. Understanding LRR domains in plant NLRs is important for developing strategies to enhance plant immunity and improve crop resistance to pathogens.

The concave surface of the leucine-rich repeat domain is generally responsible for binding to ligands [11]. The amino acid residues on the concave surface of the LRR domain form a specific pattern of hydrophobic, polar, and charged residues that can interact with specific ligands, such as proteins, peptides, carbohydrates, or nucleic acids. The specificity of ligand binding by LRR domains is determined by the overall shape and chemical properties of the concave surface, which can be highly variable between different LRR-containing proteins [9, 10]. Additionally, LRR domains can contain variable regions and insertions that can modify the binding specificity and affinity of the domain. More recently, studies such as [13] have shown that “post-LRR” domains which lie at the C-terminal end of the LRR are required for successful plant immune response. Accurate annotation of these domains and their constituent repeat units is thus essential to understanding the components which govern protein shape and binding specificity.

Existing methods for annotating LRR domains give unreliable and inconsistent results due to irregularities in sequence motifs. Profile hidden Markov models (HMMs) are widely used,

e.g. by HMMER [4], to annotate protein domains in genomic sequences, but they are sensitive to the size and diversity of the protein family being analyzed and do not perform accurately for rapidly-evolving, highly-divergent families such as LRR [14]. Profile HMMs are also unable to delineate tandem repeat units.

An existing tool, LRRPredictor [7], uses an ensemble of 8 machine learning classifiers to determine the residues which comprise the basic LRR motif of the form “LxxLxL” (where “L” refers to Leucine or other hydrophobic amino acid, and “x” can be any amino acid). We found that LRRPredictor often makes mistakes, particularly in identifying divergent motifs near the C- and N-terminal boundaries of the LRR. Because LRRPredictor, like an HMM, is trained on a specific set of LRR sequences taken from Protein Data Bank [20] (PDB), it incorrectly annotates LRR sequences which diverge from its training set.

With AlphaFold 2 [3], a deep-learning-based model, reliable protein structure prediction has become readily available, enabling domain annotation methods with direct access to geometric data from the protein. We leverage this geometric information to annotate essential features of the LRR domain: start/end position, post-LRR detection, repeat unit delineation, and structural irregularities.

From the perspective of differential geometry, a coiling curve in 3D space is characterized by a linearly increasing winding number around a core curve. We therefore detect the coiling LRR region, as the loci where the winding number is sufficiently close to a line of a fixed slope; the post-LRR domain is then decided as C-terminal sequence downstream from the point at which steady winding terminates. The methods section below describes our procedure for computing the winding number across the length of the protein. In contrast to HMM-based or other data driven techniques, our method is completely unsupervised and driven by simple mathematical methods.

Methods

Datasets used in this study

161 NLR protein sequences, i.e. *NLRome*, were obtained from the reference proteome of *A. thaliana* Col-0 TAIR10 as described previously using hmmsearch [32] and the extended NB-ARC Hidden Markov Model [2]. Of these 161 NLRs, 127 had AlphaFold-predicted structures available on AlphaFoldDB [3, 30]. The training dataset used for LRRpredictor, which contained manual annotations of LRR motif positions, was downloaded from supplemental data of [7]. We ran AlphaFold 2 prediction on a supercomputer cluster with default parameters and selected the best-scoring model for further analysis. We have included the protein amino acid sequences and corresponding pdb files in the [GitHub repository](#) where we host all the code used in this study.

Outline of methods

Our treatment of protein structures follows the outline below. Fig 1 shows the results of steps 1–4, while Fig 2 shows the results of steps 5–6.

1. **Obtaining the backbone.** Given the space curve $\gamma(t)$ representing the positions of the α -carbons, obtain a smoothed backbone curve $\gamma_\sigma(t)$ by convolving γ with a Gaussian.
2. **Parallel transport & framing.** Parallel-transport a frame along the backbone to produce, at each position t , an orthonormal basis for the plane normal to $\gamma'_\sigma(t)$. This yields a two-dimensional coordinate system $A(t)$ for each t .

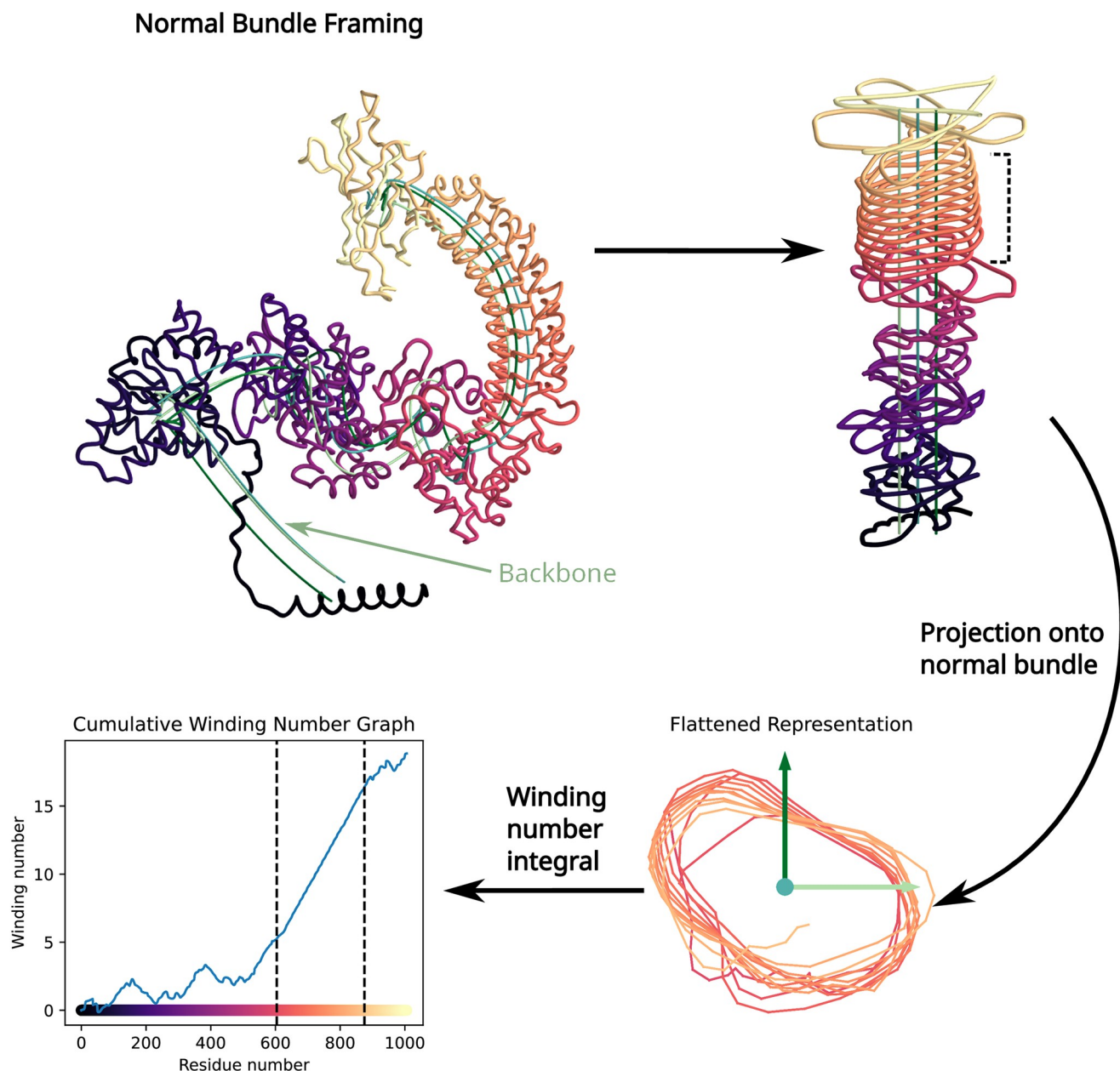


Fig 1. Embedding of protein backbone curve into normal bundle followed by projection onto an orthonormal frame yields a 2D curve containing a flattened slinky shown in lower right. The cumulative winding number, computed using the classic formula from calculus, is computed from the projection. Sloped linear segments of the winding number curve indicate coiling. Protein shown is *A. thaliana* NLR with TAIR [1] ID AT3G44400.2.

<https://doi.org/10.1371/journal.pcbi.1012526.g001>

3. **The flattened representation.** For each t , compute the coordinates of $\gamma(t) - \gamma_o(t)$ according to $A(t)$. This produces a two-dimensional “flattened” curve $\phi(t)$ representing the position of γ relative to its backbone.
4. **Cumulative winding number.** Compute the cumulative winding number $W_\phi(t)$ of ϕ about the origin.
5. **Secant line statistics; median slope.** Compute the median slope of secant lines to W_ϕ to infer the number m of residue positions per helical repeat unit in the LRR domain.

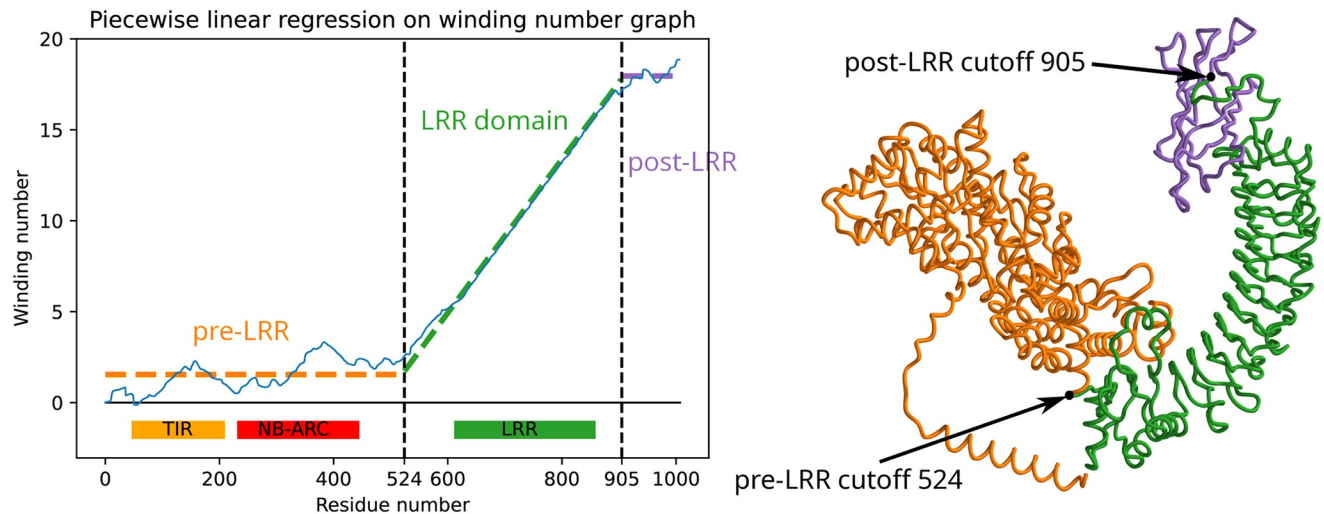


Fig 2. A discontinuous clipped ReLU function is regressed on the graph of the winding number function for *A. thaliana* NLR with TAIR [1] ID AT3G44400.2. The breakpoints of the regression yields the start and end positions of the LRR domain, highlighted in green. InterPro [19] domain annotations are shown below regression plot.

<https://doi.org/10.1371/journal.pcbi.1012526.g002>

6. **Piecewise-linear regression & gradient descent.** By gradient descent on an appropriate loss function, find a piecewise-linear regression of W_φ with slopes alternating between zero and m . Regions of the regression with slope m correspond to solenoidal regions of the protein structure.

Obtaining the backbone

Let $\gamma(t)$, $t \in \{0, \dots, n\}$ be a discrete space curve representing the positions of the α -carbons in a protein structure. This curve can be represented as three scalar functions of t : $\gamma(t) = (\gamma_x(t), \gamma_y(t), \gamma_z(t))$. Let g_σ be the mean-zero Gaussian distribution with standard deviation σ :

$$g_\sigma(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2/(2\sigma^2)}. \quad (1)$$

We define the “backbone” to the structure

$$\gamma_\sigma(t) := ((g_\sigma \star \gamma_x)(t), (g_\sigma \star \gamma_y)(t), (g_\sigma \star \gamma_z)(t)), \quad (2)$$

where $\star$ is the convolution, defined $(p \star q)(t) := \sum_s p(t-s)q(s)$, where the sum is over all sensible indices s . Throughout in our computations, we set $\sigma = 20$.

Parallel transport & framing

First, we compute the tangent vector $\gamma'_\sigma(t)$ to the backbone by convolving γ with the derivative of a Gaussian, i.e. with

$$g'_\sigma(t) = -\frac{t}{\sigma^3\sqrt{2\pi}} e^{-t^2/(2\sigma^2)}. \quad (3)$$

This is a standard technique [31] for defining derivatives of discrete data, since convolution associates with differentiation as $(d/dt)(p \star q) = ((dp/dt) \star q) = (p \star (dq/dt))$. In order to

measure the winding of γ around its backbone γ_σ , we need a consistent representation of the position of γ relative to γ_σ ; in effect, we need to “straighten” the backbone and carry γ along for the ride.

Now that we have $\gamma'_\sigma(t) = (g'_\sigma \star \gamma)(t)$, we will produce a sequence of orthonormal bases for the planes orthogonal to $\gamma'_\sigma(t)$ at each residue t . Our method starts with a frame at $t = 0$ and parallel-transport it along the backbone as follows:

1. Given $\gamma'_\sigma(0)$, the initial tangent to the backbone, let $A(0)$ be any 3×2 real matrix with orthonormal columns such that $A(0)^T \gamma'_\sigma(0) = \mathbf{0}$ (i.e., the columns of $A(0)$ complete $\gamma'_\sigma(0)$ to an orthonormal basis for $\mathbb{R}^3$).
2. Given $A(t-1)$, let $B(t)$ be the matrix whose columns are orthogonal projections of the columns of $A(t-1)$ onto the complement of $\gamma'_\sigma(t)$. Symbolically,

$$B(t) = A(t-1) - \frac{1}{\|\gamma'_\sigma(t)\|^2} \gamma'_\sigma(t)^T A(t) \gamma'_\sigma(t). \quad (4)$$

The columns of $B(t)$ are likely not orthonormal.

3. Let $A(t)$ be the 3×2 matrix with orthonormal columns that is closest (in the Frobenius norm) to $B(t)$. Numerically, $A(t)$ is found by computing the SVD of $B(t)$ and replacing its singular values with 1's (the standard solution to the “Orthogonal Procrustes Problem” [15, 16]). Note that the columns of $A(t)$ span the same subspace as those of $B(t)$, so $A(t)$ has columns guaranteed orthogonal to $\gamma'_\sigma(t)$.
4. Repeat steps 2 and 3 for $t = 1, \dots, n$.

The flattened representation

The flattened representation is now a plane curve $\varphi(t) = A(t)^T(\gamma(t) - \gamma_\sigma(t))$. It can be thought of as γ from the perspective of an observer traveling along the backbone γ_σ and oriented according to the frames $A(t)$.

Cumulative winding number

For a continuous-time plane curve $z(t) = (x(t), y(t))$ with polar representation $(r(t) \cos(\theta(t)), r(t) \sin(\theta(t)))$, the winding number is defined

$$W_z(t) = \frac{1}{2\pi} \int_0^t \theta'(s) ds = \frac{1}{2\pi} \int_0^t \frac{x(s)y'(s) - y(s)x'(s)}{x(s)^2 + y(s)^2} ds. \quad (5)$$

This quantity tracks the total number of rotations accumulated by a ray pointing at $z(s)$, as s moves in the interval $[0, t]$.

In our case, given the discrete plane curve $\varphi(t) = (x(t), y(t))$, we define a discrete version of the cumulative winding number by

$$W_\varphi(t) = \frac{1}{2\pi} \sum_{s=1}^t \arctan \left(\frac{y(s)x(s-1) - x(s)y(s-1)}{x(s)x(s-1) + y(s)y(s-1)} \right). \quad (6)$$

The summand accumulates the angle between rays to consecutive points $\varphi(s-1)$ and $\varphi(s)$ along the discrete curve. Fig 1 provides a graphical example of the backbone, parallel-transported normal bundle, flattened representation, and cumulative winding number plot.

Secant line statistics; median slope

To make piecewise-linear regression tractable, we remove slope as an optimization parameter, and instead infer it from the statistics of secant lines to W_φ . First we choose parameters $0 < d < D$ (in the `median_slope` method, these are `small = 100` and `big = 250`, respectively). We will consider only secant lines with endpoints a, b where $d \leq b - a \leq D$. Associated to such a secant line is a slope $m_{a,b} = (W_\varphi(b) - W_\varphi(a))/(b - a)$, and a score $S_{a,b}$ computed as follows. First define

$$R_{a,b} = \sum_{t=a}^b \left[\left(W_\varphi(t) - m_{a,b}t \right) - \frac{1}{b-a+1} \sum_{s=a}^b \left(W_\varphi(s) - m_{a,b}s \right) \right]^2. \quad (7)$$

In other words, $R_{a,b}$ measures the total squared deviation of $(W_\varphi(t) - m_{a,b}t)$ away from its mean; we have $R_{a,b} = 0$ if and only if W_φ coincides with its secant line on $t \in [a, b]$. Now let the score $S_{a,b} = (b - a)/(1 + R_{a,b})$, rewarding long secant lines and penalizing deviations from linear behavior.

The “median slope” is chosen by a voting process. First we determine the minimum and maximum slopes, call them m and M . We create $\sqrt{N}$ score bins, where N is the number of secant lines, i.e. the number of pairs (a, b) with $0 \leq a, b \leq n$ and $d \leq b - a \leq D$. For each secant line with endpoints a, b , its score $S_{a,b}$ accumulates in the bin with index $\lfloor (m_{a,b} - m)/(M - m) \rfloor$. After this procedure, the slope returned is $m + (i/\sqrt{N})(M - m)$, where i is the index of the bin with largest score. We use this slope in subsequent regression tasks.

Our “median slope” computation is conceptually similar to the Hough transform [33], a computer vision method for detecting segments in images via a voting process across a parameterized space of lines in the plane.

Piecewise-linear regression & gradient descent

The “median slope” m associated to the winding W_φ approximates the reciprocal of residues per repeat unit in the LRR domain—as residue position t changes by m , winding number increases by 1, i.e. φ completes one revolution around the origin. To annotate the domain in which W_φ exhibits this linear, slope- m behavior, we fit a piecewise-linear, discontinuous function which is constant in the pre-LRR region, slope- m in the LRR domain, and constant in the post-LRR region. More precisely, associated to a choice of breakpoints ($0 = a_0 < a_1 < \dots < a_k = n$) is a regression function that is constant on $[a_0, a_1)$, slope- m on $[a_1, a_2)$, constant on $[a_2, a_3)$, and so on. Most of the cumulative winding number plots were well-approximated with $k = 2$ (two breakpoints); we discuss larger k below.

We define a loss function, similar in spirit to 7, as follows. First, for a function $f(t)$ and endpoints $a < b$, define $V_{a,b}(f)$ to be the total squared deviation of f from its mean on $[a, b]$:

$$V_{a,b}(f(t)) = \sum_{t=a}^{b-1} \left(f(t) - \frac{1}{b-a} \sum_{s=a}^{b-1} f(s) \right)^2. \quad (8)$$

Choose constants C, D , and define the loss associated to the partition $(a_0, \dots, a_k)$:

$$L(a_0, \dots, a_k) = \sum_{j=0}^{k/2} (C V_{a_{2j}, a_{2j+1}}(W_\varphi(t)) + D V_{a_{2j+1}, a_{2j+2}}(W_\varphi(t) - mt)). \quad (9)$$

The loss is a weighted measurement of the total squared deviation between W_φ and the regression function we are fitting, with the weights C and D determining how harshly we penalize

deviations from linearity (slope- m behavior) in the LRR region. In our code, we found that $C = 1$ and $D = 1.5$ worked well. Our optimization problem now becomes: find $(a_0, \dots, a_k)$ minimizing $L(a_0, \dots, a_k)$.

We solve the optimization problem by gradient descent on L : we form a finite-difference gradient ∇L whose j^{th} entry is

$$(\nabla L)_j = (L(a_0, \dots, a_j + 1, \dots, a_k) - L(a_0, \dots, a_j, \dots, a_k)), \quad (10)$$

choose a learning rate $\epsilon > 0$, increment the vector of breakpoints by $-\epsilon \nabla L$, and iterate.

Refinements and alternatives

Loss histograms & four-breakpoint regressions. A small number (ten out of 127) of proteins in our dataset contained hairpin loops or other localized deviations from solenoidal geometry in the LRR region, and regressions with $k = 2$ breakpoints were not satisfactory. We found the standard deviation of the difference between W_φ and the regressing function inside the LRR region, i.e. $V_{a_1, a_2} / (a_2 - a_1)$, is high in such cases. Fig 3 shows the distribution of these values. We repeat the regression with four breakpoints, instead of two, to deal with these edge cases.

Fig 4 shows the result of fitting a regressing function with four, instead of two breakpoints.

Laplacian circular coordinates. In the previous sections, we used piecewise linear regression on the cumulative winding number to isolate the LRR domain. In the process, we estimated the winding number, which can also give us *instantaneous phase*, or the angle along each loop, on the LRR domain sequence. In this section, we briefly describe another technique

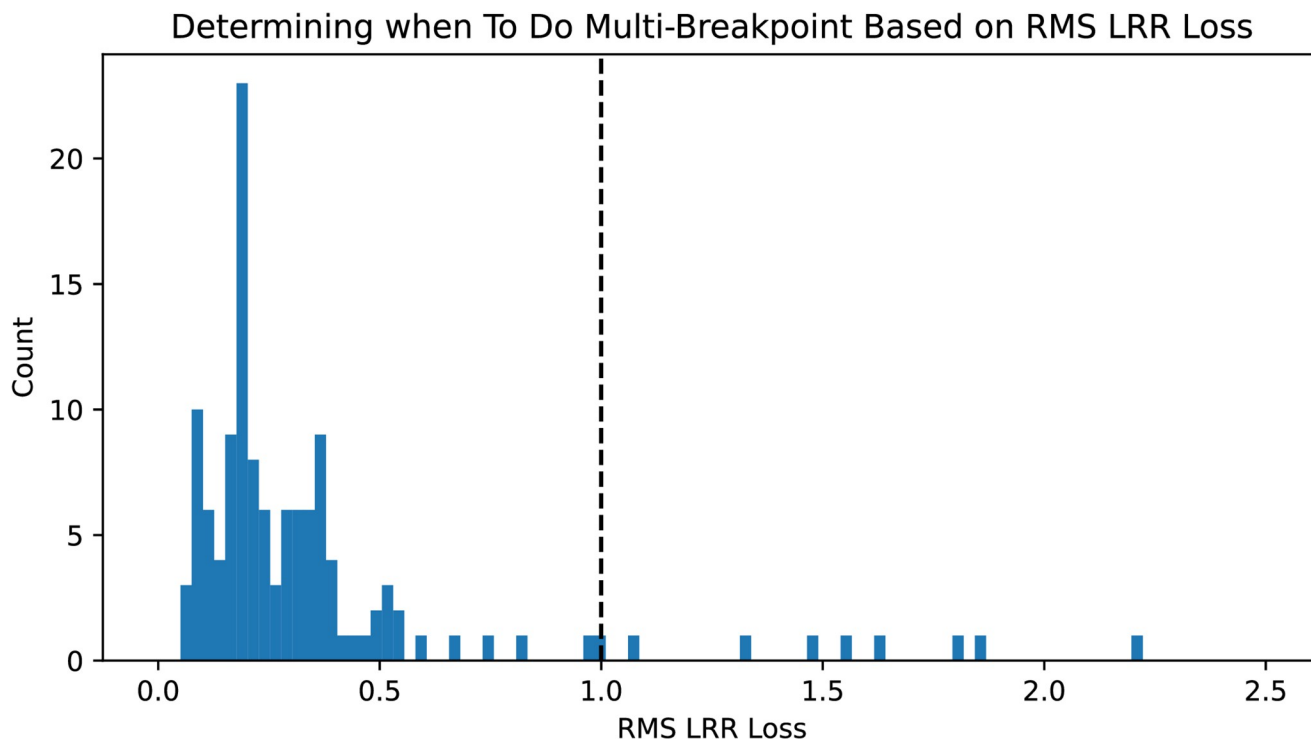


Fig 3. We determine when to redo the regression using 4 breakpoints by examining the RMS of the LRR component of the loss. This term is above our threshold of 1 for 9/127 of the proteins in *A. thaliana*.

<https://doi.org/10.1371/journal.pcbi.1012526.g003>

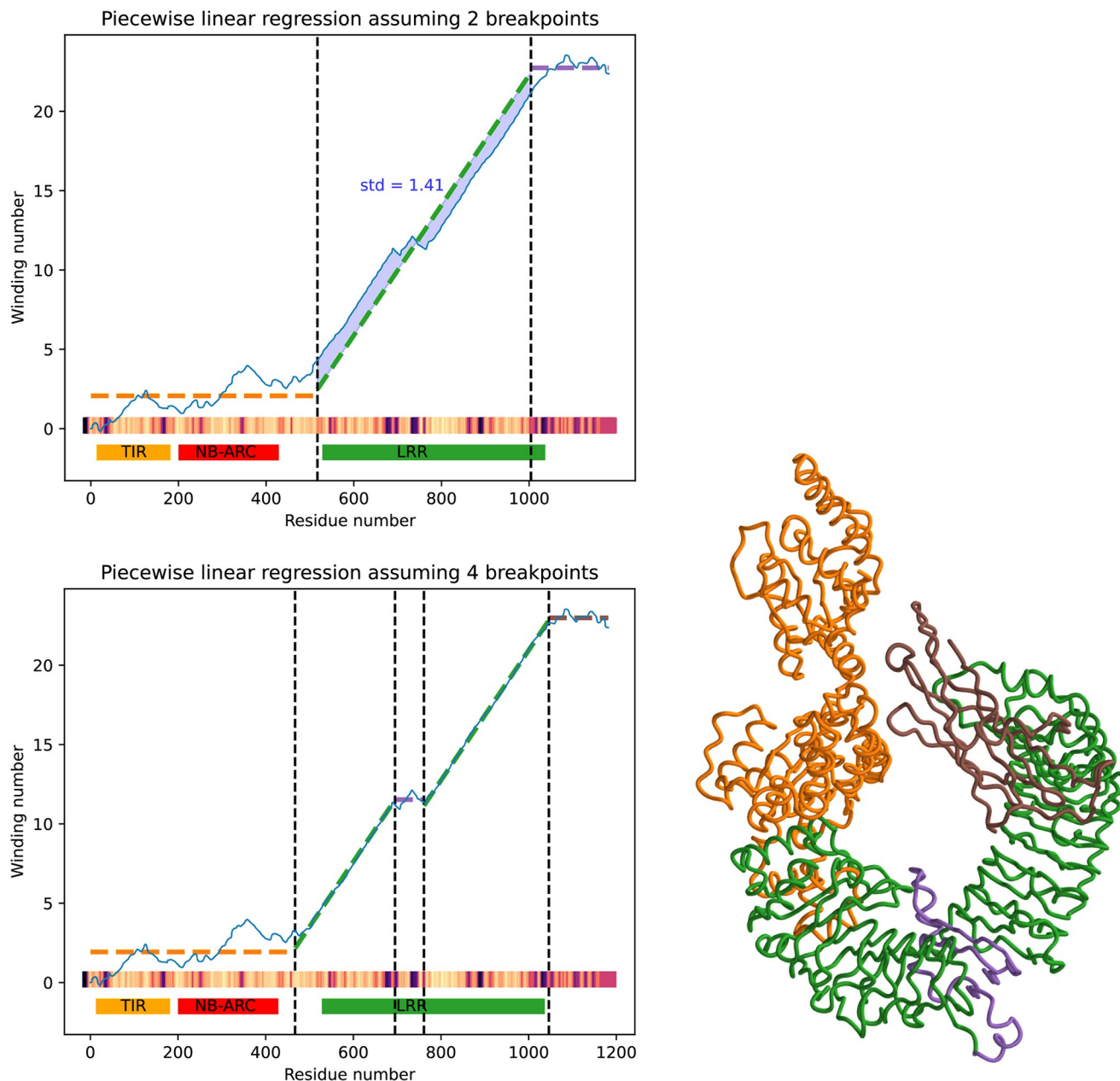


Fig 4. Four breakpoint piecewise linear regression enables detection of a non-coiling structure (highlighted in purple at right) which deviates from the usual coiling in the LRR domain. Below regression plot, a heat map shows the pLDDT (predicted local distance difference test), a per-residue confidence measure given by AlphaFold 2 which is elevated in the non-coiling region. Bottom of plot shows HMM-based InterPro domain annotations which fail to detect non-coiling region within LRR domain. TAIR ID is AT1G72840.2.

<https://doi.org/10.1371/journal.pcbi.1012526.g004>

based on graph theory for estimating instantaneous phase of LRR regions, which we evaluate alongside the parallel transport method in Section 7.2.

Before setting up the graph, we perform some preprocessing to make LRR solenoid region as circular as possible. First, we nullify some of the torsion by once again computing the tangent vectors on the LRR solenoid. This time, however, we set $\sigma = 1$, and convolve $\gamma(t)$ with $g'_1(t)$ instead of $g'_{20}(t)$ to preserve the loop structures. To further accentuate periodic features,

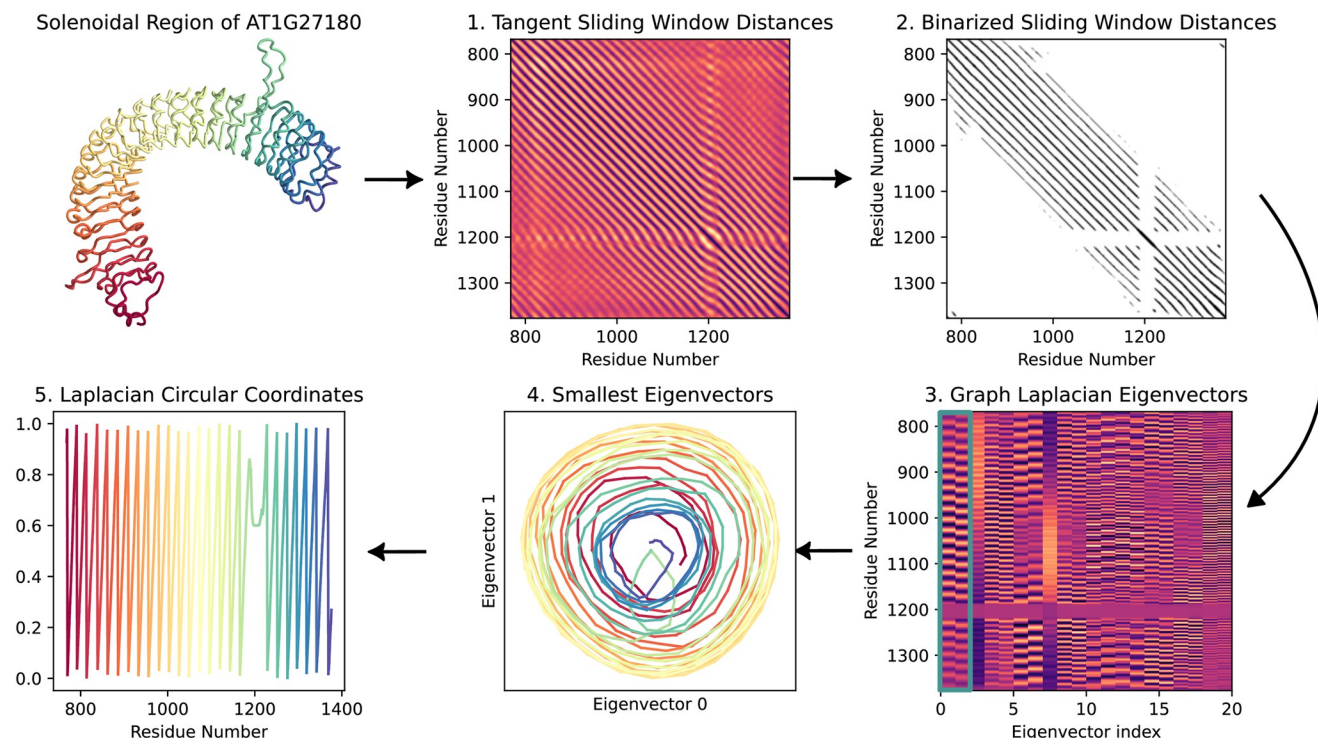


Fig 5. Graph Laplacian eigenvectors of mutual nearest neighbor graph on LRR solenoid curve tangent vectors. LRRPredictor residues are shown as blue horizontal lines on eigenmatrix plot. The 0th and 1st eigenvectors have period matching the expected period of the solenoid as determined by LRRPredictor. Leading eigenvectors of graph Laplacian are periodic and are $\pi/4$ -phase shifted, thereby yielding projections of LRR coil onto a winding around a circle in a 2D-plane. Phase estimation using the formula $\theta = \tan^{-1}(\frac{y}{x})$ of LRR coil at bottom taking values between $-\pi$ and π .

<https://doi.org/10.1371/journal.pcbi.1012526.g005>

we perform a multivariate sliding window embedding [5] of window size 24 (roughly the length of the LRR period) with delay time 1 on each component of the tangent vector field. The formula for such a sliding window embedding of some sequence $f[t]$ is

$$SW_1^{24} f[t] := \begin{bmatrix} \gamma[t] \\ \gamma[t+1] \\ \gamma[t+2] \\ \vdots \\ \gamma[t+24] \end{bmatrix} \in \mathbb{R}^{24+1} \quad (11)$$

We concatenate together $SW_1^{24} \gamma'_{i\sigma}$ for each of the three components of the tangent vector $\gamma'_{i\sigma}$, resulting in a sequence in 75-dimensional Euclidean space. We then construct a 50-mutual-nearest-neighbors graph on the sliding window embedding.

From the mutual-NN graph we compute leading eigenvectors of the unweighted graph Laplacian [23]. An example is shown in Fig 5. Intuitively, the graph Laplacian is a generalization of a discrete second derivative operator to graphs. For the same reason that sines and cosines are eigenfunctions of the second derivative operator with associated eigenvalues proportional to the frequency, eigenvectors of the graph Laplacian on a graph of a circle are sine-cosine pairs, up to a phase, that go through an integer number of cycles over one revolution of

the circle, and lower frequency pairs have smaller eigenvalues [24]. We expect a near circular graph in the mutual-NN graph in the periodic LRR region, and the Laplacian eigenvectors are known to degrade gracefully in the presence of imperfections. Therefore, we expect the two eigenvectors with the smallest eigenvalue to be approximately periodic and $\pi/4$ -phase shifted. If we use the two entries of these eigenvectors as x - and y -coordinates, respectively, we obtain a projection of the LRR coil onto a circle winding in the plane. Our phase estimation θ along the LRR coil is simply obtained as $\theta = \tan^{-1}(\frac{y}{x})$, as shown in Fig 5 below.

We note that a similar phase-estimation scheme with the graph Laplacian of mutual nearest neighbors has been used to order photographs along a loop [25] and to parameterize periodic videos [5]. Furthermore, a spiritually similar but more computationally intensive topological phase estimation based on cohomology [28, 29] has been used to recognize patterns in motion capture data [26] and to detect head orientation from neural data [27].

Results

Cumulative winding number reveals errors made by ML-based LRR repeat unit delineator

We ran the LRR annotation tool LRRPredictor [7] on the 127 NLRs from *A. thaliana* to obtain predicted locations of the LRR motif “LxxLxL.” Let $R_1, \dots, R_\ell$ denote the starting residues for the LRR motifs predicted by LRRPredictor. The analogous measurement in our model is to record the residues at which our cumulative winding number W_φ crosses integers.

To compare the two prediction schemes, we evaluate our cumulative winding number at the residues returned by LRRPredictor. That is, we form the list of numbers $(W_\varphi(R_1), \dots, W_\varphi(R_\ell))$. If the models are in agreement, the running difference $(W_\varphi(R_2) - W_\varphi(R_1), \dots, W_\varphi(R_\ell) - W_\varphi(R_{\ell-1}))$ should equal the all-ones vector $(1, \dots, 1)$ (that is, the structure should wind exactly once around the core between residues R_{j-1} and R_j). The “discrepancy”

$$D(R_1, \dots, R_\ell) := \sqrt{\sum_{j=2}^{\ell} (W_\varphi(R_j) - W_\varphi(R_{j-1}) - 1)^2} \quad (12)$$

quantifies the extent to which this is not the case. A number of LRRPredictor outputs contained false predictions in which consecutive motif start sites R_j and R_{j-1} appear close together—often only a couple residues apart. Such duplicate predictions result in a high discrepancy $D(R_1, \dots, R_\ell)$ because the difference $W_\varphi(R_j) - W_\varphi(R_{j-1})$ as computed in formula (12) above is close to 0.

To test the validity of our winding number computation, we ran the discrepancy computation on the LRRPredictor outputs on the 127 *A. thaliana* reference proteome NLRs as well as AlphaFold 2 structures for the training dataset for LRRpredictor, a manually-annotated “ground truth” dataset of LRR motifs on 172 experimentally-derived LRR structures taken from Protein Data Bank. These PDB protein structures were derived from a diverse set of organisms comprising bacteria, fungi, plants, and animals.

We found consistently low discrepancy values for the ground truth set with mean 0.127. By comparison, *A. thaliana* NLRome discrepancy values were generally low with mean 0.373, but exhibited higher values in cases where LRRpredictor made mistakes. Fig 6 below shows a pair of overlaid histograms comparing discrepancy values for both the NLRome dataset and validation dataset (S1 and S2 Tables). The discrepancy values are much lower on the LRRPredictor ground truth dataset compared to the NLRome dataset, implying that our technique makes fewer mistakes than LRRPredictor does on new data. Fig 7 demonstrates how the discrepancy is able to catch duplicate motif predictions made by LRRPredictor. These results demonstrate

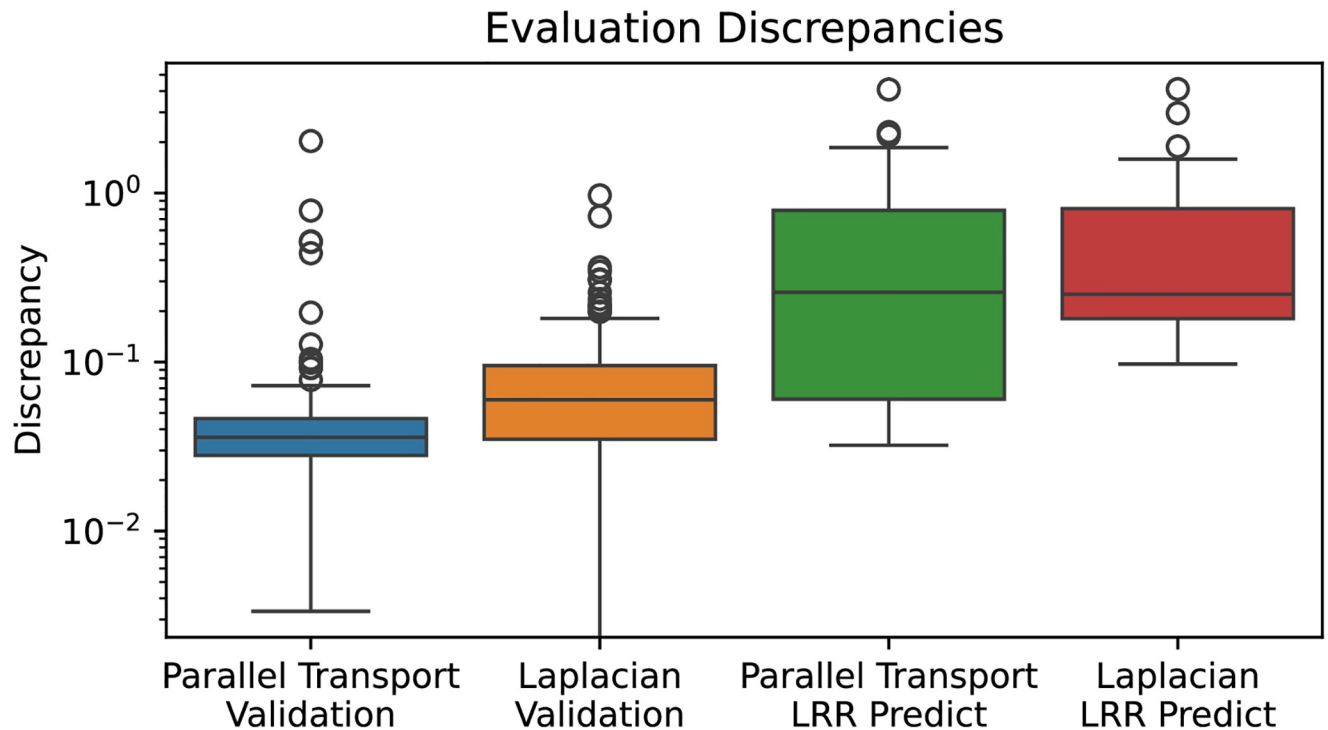


Fig 6. Discrepancies for LRRpredictor outputs on 127 *A. thaliana* (green and red) NLRs are higher than those for manually-annotated LRR repeat units used as the training set for the LRRpredictor model (blue, orange). Thus, the cumulative winding number computation faithfully recapitulates the periodicity of the LRR coil.

<https://doi.org/10.1371/journal.pcbi.1012526.g006>

not only the winding number's ability to accurately model the LRR coil, but also its generalizability to non-NLR LRR's derived from species other than *A. thaliana*.

Structural anomaly detection by sliding window L^2 distance from Laplacian eigenvector winding number to line

Many LRR coils have hairpin loops and other structural anomalies which deviate from coiling. In these anomalous regions, the leading eigenvectors deviate from their usual periodic behavior. Applying the winding number formula (Eq 6) above to the pair of leading graph Laplacian eigenvectors leads to a cumulative winding number within the LRR domain which is better able to discern small hairpins compared to the previous winding number computation based on normal bundle projection. As shown in Fig 8 below, we detect a small hairpin as a spike in L^2 distance between the winding number and its median slope.

Discussion

The emergence of AlphaFold 2 has catalyzed a paradigm shift in protein structure prediction, facilitating access to genome-wide high-quality structural models. Traditional sequence homology-based domain annotation techniques, like LRRPredictor, often face challenges with LRRs, especially in proteins with high sequence divergence. While evolutionary divergence might veil the sequence homology of LRR units, their core structural topology, characterized by 20–30 amino acid stretches typically involved in protein-protein interactions, often remains conserved, acting as a distinct structural signature.

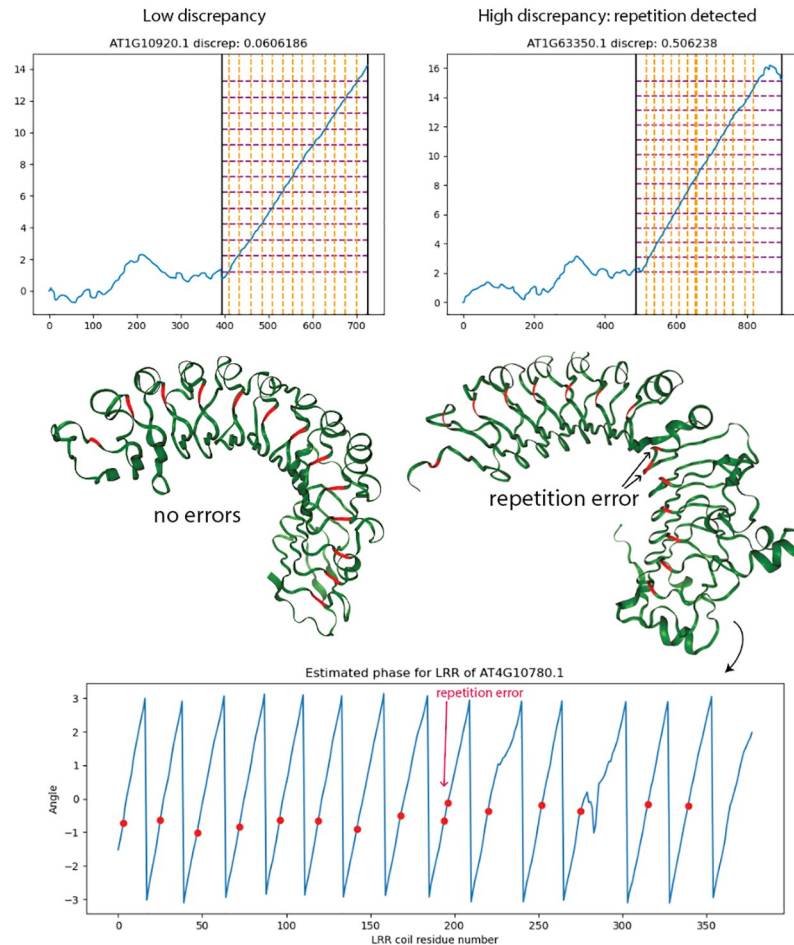


Fig 7. LRRPredictor discrepancy computation reveals proteins with erroneously repeated predictions. NLRs with high-discrepancy LRRPredictor outputs tend to carry repetition errors or missing motif annotations. Orange vertical lines overlaid on winding number plot depict LRRPredictor residues, while purple horizontal lines depict the integer-spaced grid which best approximates the winding number graph evaluated at LRRPredictor residues. A repetition error can be seen in the grid representation as a doubled orange line around residue 685. At bottom, LRRPredictor residues are mapped onto graph Laplacian eigenvector phase estimation, revealing an pair of duplicates with adjacent phase.

<https://doi.org/10.1371/journal.pcbi.1012526.g007>

This study uses AlphaFold 2 to generate a 3D space curve from a protein sequence, which subsequently is projected into the 2D plane by identifying a series of “slinky” cross-sections. Through computing the cumulative winding number on the resultant 2D curve and employing piecewise linear regression, the linearly sloped region, identified as the LRR domain, is discerned. Our method pivots on the application of geometric data analysis to illuminate structural motifs that remain elusive to sequence analysis alone.

The use of geometric and topological concepts in our method aligns with previous studies that have explored Topological Data Analysis (TDA) in protein structure and dynamics [21, 22]. For instance, SINATRA Pro has been used to identify biophysical signatures in protein dynamics by detecting topological differences between protein structures [21]. Similarly, TopologyNet integrates TDA with deep learning for biomolecular property predictions [22]. Our approach builds on these foundational ideas by leveraging large-scale AI/ML-derived databases like AlphaFoldDB, showcasing the potential of combining AI-based structural

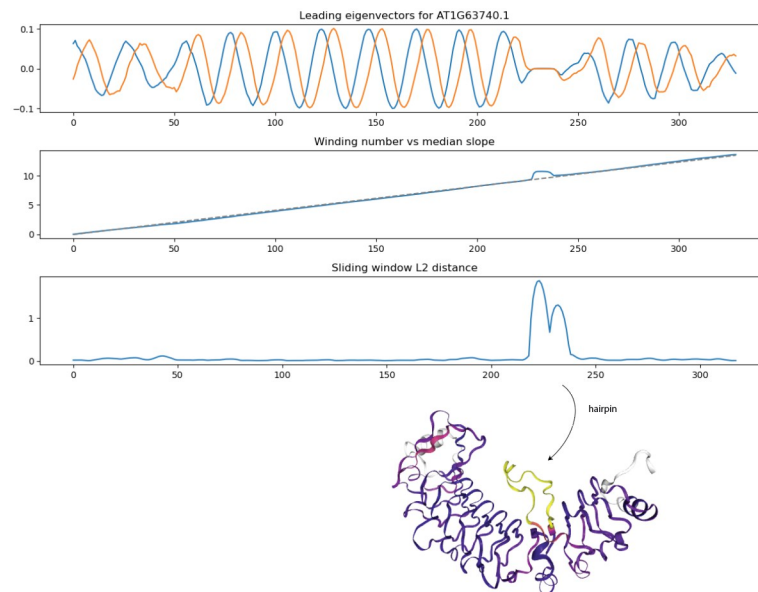


Fig 8. Sliding window L^2 distance (SWL2D) from winding number to median secant line detects small hairpins/insertions in LRR coil domain. Structure at bottom is colored according to SWL2D where yellow values are higher.

<https://doi.org/10.1371/journal.pcbi.1012526.g008>

predictions with geometric and topological analyses for advanced domain annotation. The amalgamation of advanced protein structure prediction technologies and mathematical models, as demonstrated in our approach, underscores the potential for widening our understanding of protein function across varied biological systems.

Our method yields several kinds of precise results: (a) it identifies the start and end sites of the LRR domain with greater accuracy than HMM-based methods, (b) it annotates repeat units more reliably than the existing LRRPredictor, (c) it identifies misannotations by other annotation/prediction tools, and (d) it reveals structural anomalies within the LRR domain that deviate from conventional coiling behaviors. These findings not only underscore the utility of our approach but also present a robust framework for delving into the intricate structural patterns intrinsic to LRR domains.

While we benchmarked our work on LRR domains in NLR proteins, the intrinsic methodology has the capacity for broader applications, likely extending to other linear solenoid protein domains like armadillo (ARM), tetratricopeptide (TPR), and ankyrin (ANK) repeats, all of which feature distinctive repeat sequences and structural configurations. However, the method is unlikely to work well on circular solenoid domains such as beta propellers (e.g. WD40) because, unlike linear solenoids, those structures do not consistently wind around a core curve.

Our method does come with limitations. For instance, while it can detect non-coiling structural anomalies within the LRR domain, the origin, authenticity, and potential functionality of these regions remain ambiguous. Moreover, our structure-based annotation method, albeit effective for domains with a straightforward geometric description like LRRs, might not be universally applicable to other protein domains without developing a new geometric model tailored to them. This underscores a potential limitation when juxtaposing sequence-based versus structure-based domain annotation, highlighting a future avenue warranting exploration: developing geometric models for other protein domains.

Supporting information

S1 Table. Discrepancy values for *A. thaliana* NLROME dataset.
(CSV)

S2 Table. Discrepancy values for LRRPredictor training dataset.
(CSV)

Acknowledgments

We thank Daniil Prigozhin for running LRRPredictor on *A. thaliana* NLROME, Chandler Sutherland and the Krasileva Lab for providing feedback and suggestions on this project and resulting manuscript.

Author Contributions

Conceptualization: Boyan Xu, Ksenia Krasileva.

Data curation: Boyan Xu, Daven Lim.

Formal analysis: Boyan Xu, Alois Cerbu, Christopher J. Tralie, Daven Lim.

Funding acquisition: Ksenia Krasileva.

Investigation: Boyan Xu, Daven Lim.

Methodology: Boyan Xu, Alois Cerbu, Christopher J. Tralie, Daven Lim.

Software: Boyan Xu, Alois Cerbu, Christopher J. Tralie.

Supervision: Ksenia Krasileva.

Validation: Boyan Xu, Christopher J. Tralie, Daven Lim.

Visualization: Boyan Xu, Alois Cerbu, Christopher J. Tralie.

Writing – original draft: Boyan Xu, Alois Cerbu.

Writing – review & editing: Boyan Xu, Alois Cerbu, Christopher J. Tralie, Ksenia Krasileva.

References

- Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al. The Arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome. *Genome*. 2015; 53(8):474–485. <https://doi.org/10.1002/dvg.22877> PMID: 26201819
- Bailey PC, Schudoma C, Jackson W, Baggs E, Dagdas G, Haerty W, et al. Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. *Genome Biol*. 2018; 19(1):1–18. <https://doi.org/10.1186/s13059-018-1392-6> PMID: 29458393
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021; 596(7873):583–589. <https://doi.org/10.1038/s41586-021-03819-2> PMID: 34265844
- Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011; 39(suppl_2):W29–W37. <https://doi.org/10.1093/nar/gkr367> PMID: 21593126
- Tralie CJ, Berger M. Topological eulerian synthesis of slow motion periodic videos. In: 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE; 2018. p. 3573–3577.
- Park K, Shen BW, Parmeggiani F, Huang PS, Stoddard BL, Baker D. Control of repeat-protein curvature by computational protein design. *Nat Struct Mol Biol*. 2015; 22(2):167–174. <https://doi.org/10.1038/nsmb.2938> PMID: 25580576
- Martin EC, Sukarta OCA, Spiridon L, Grigore LG, Constantinescu V, Tacutu R, et al. LRRpredictor—a new LRR motif detection method for irregular motifs of plant NLR proteins using an ensemble of classifiers. *Genes*. 2020; 11(3):286. <https://doi.org/10.3390/genes11030286> PMID: 32182725

8. Tamborski J, Krasileva KV. Evolution of plant NLRs: from natural history to precise modifications. *Annu Rev Plant Biol.* 2020; 71:355–378. <https://doi.org/10.1146/annurev-arplant-081519-035901> PMID: 32092278
9. Prigozhin DM, Krasileva KV. Analysis of intraspecies diversity reveals a subset of highly variable plant immune receptors and predicts their binding sites. *Plant Cell.* 2021; 33(4):998–1015. <https://doi.org/10.1093/plcell/koab013> PMID: 33561286
10. Barragan AC, Weigel D. Plant NLR diversity: the known unknowns of pan-NLRomes. *Plant Cell.* 2021; 33(4):814–831. <https://doi.org/10.1093/plcell/koaa002> PMID: 33793812
11. Padmanabhan M, Cournoyer P, Dinesh-Kumar SP. The leucine-rich repeat domain in plant innate immunity: a wealth of possibilities. *Cell Microbiol.* 2009; 11(2):191–198. <https://doi.org/10.1111/j.1462-5822.2008.01260.x> PMID: 19016785
12. Ng A, Xavier RJ. Leucine-rich repeat (LRR) proteins: integrators of pattern recognition and signaling in immunity. *Autophagy.* 2011; 7(9):1082–1084. <https://doi.org/10.4161/auto.7.9.16464> PMID: 21606681
13. Saucet SB, Esmenjaud D, Van Ghelder C. Integrity of the post-LRR domain is required for TIR-NB-LRR function. *Mol Plant Microbe Interact.* 2021; 34(3):286–296. <https://doi.org/10.1094/MPMI-06-20-0156-R> PMID: 33197377
14. Bateman A, Coghill P, Finn RD. DUFs: families in search of function. *Acta Crystallogr F.* 2010; 66(10):1148–1152. <https://doi.org/10.1107/S1744309110001685> PMID: 20944204
15. Wahba G. A least squares estimate of satellite attitude. *SIAM Rev.* 1965; 7(3):409–409. <https://doi.org/10.1137/1007077>
16. Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A.* 1976; 32(5):922–923. <https://doi.org/10.1107/S0567739476001873>
17. Jones JDG, Vance RE, Dangl JL. Intracellular innate immune surveillance devices in plants and animals. *Science.* 2016; 354(6316):aaf6395. <https://doi.org/10.1126/science.aaf6395> PMID: 27934708
18. Jang H, Stevens P, Gao T, Galperin E. The leucine-rich repeat signaling scaffolds Shoc2 and Erbin: cellular mechanism and role in disease. *FEBS J.* 2021; 288(3):721–739. <https://doi.org/10.1111/febs.15450> PMID: 32558243
19. Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 2021; 49(D1):D344–D354. <https://doi.org/10.1093/nar/gkaa977> PMID: 33156333
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. *Nucleic Acids Res.* 2000; 28(1):235–242. <https://doi.org/10.1093/nar/28.1.235> PMID: 10592235
21. Tang WS, da Silva GM, Kirveslahti H, Skeens E, Feng B, Sudijono T, et al. A topological data analytic approach for discovering biophysical signatures in protein dynamics. *PLoS Comput Biol.* 2022; 18(5):e1010045. <https://doi.org/10.1371/journal.pcbi.1010045> PMID: 35500014
22. Cang Z, Wei GW. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput Biol.* 2017; 13(7):e1005690. <https://doi.org/10.1371/journal.pcbi.1005690> PMID: 28749969
23. Chung FRK. Spectral graph theory. American Mathematical Soc.; 1997.
24. Godsil C, Royle GF. Algebraic graph theory. Springer Science & Business Media; 2001.
25. Averbuch-Elor H, Cohen-Or D. Ringit: Ring-ordering casual photos of a temporal event. *ACM Trans Graph.* 2015; 34(3):1–11. <https://doi.org/10.1145/2735628>
26. Vejdemo-Johansson M, Pokorny FT, Skrabla P, Kragic D. Cohomological learning of periodic motion. *Appl Algebra Eng Commun Comput.* 2015; 26(1):5–26. <https://doi.org/10.1007/s00200-015-0251-x>
27. Rybakken E, Baas N, Dunn B. Decoding of Neural Data Using Cohomological Feature Extraction. *Neural Comput.* 2019; 31:68–93. https://doi.org/10.1162/neco_a_01150 PMID: 30462582
28. De Silva V, Vejdemo-Johansson M. Persistent cohomology and circular coordinates. In: Proceedings of the twenty-fifth annual symposium on Computational geometry. 2009. p. 227–236.
29. Perea JA. Sparse circular coordinates via principal $\mathbb{Z}$ -bundles. In: Topological Data Analysis: The Abel Symposium 2018. Springer; 2020. p. 435–458.
30. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* 2024; 52(D1):D368–D375. <https://doi.org/10.1093/nar/gkad1011> PMID: 37933859
31. Mokhtarian F, Mackworth AK. A theory of multiscale, curvature-based shape representation for planar curves. *IEEE Trans Pattern Anal Mach Intell.* 1992; 14(8):789–805. <https://doi.org/10.1109/34.149591>

32. Misty J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 2013; 41(12):e121–e121. <https://doi.org/10.1093/nar/gkt263> PMID: 23598997
33. Duda RO, Hart PE. Use of the Hough transformation to detect lines and curves in pictures. *Commun ACM.* 1972; 15(1):11–15. <https://doi.org/10.1145/361237.361242>