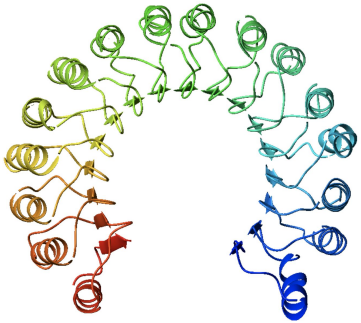
Berkeley
UNIVERSITY OF CALIFORNIA

MATHEMATICS + BERKELEY

CENTER *for* COMPUTATIONAL BIOLOGY

Structure-Aware Annotation of Leucine-Rich Repeat Domains

Boyan Xu



THE CENTRAL DOGMA

DNA

TRANSCRIPTION

mRNA AUGAUCUCGUAA

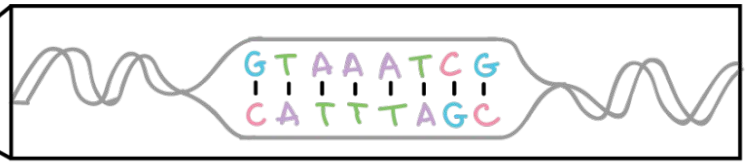
TRANSLATION

Polypeptide Met Ile Ser STOP

protein folding

3D protein structure

flower color gene

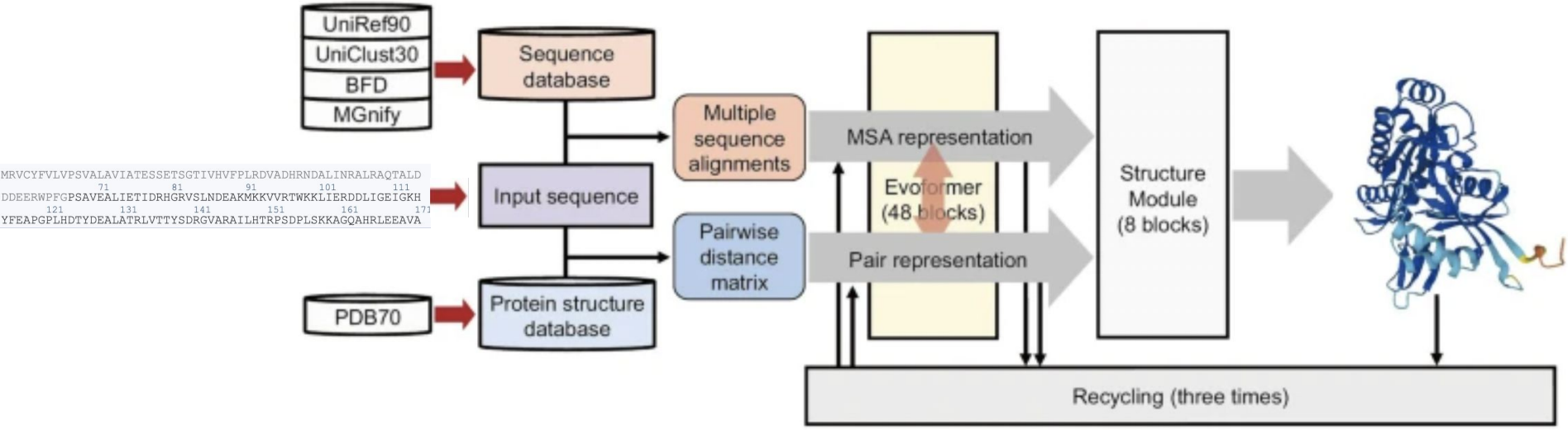


DNA sequence

protein that helps make pigments

purple flowers

HPC-enabled Alphafold 2 produces accurate 3D protein structure prediction



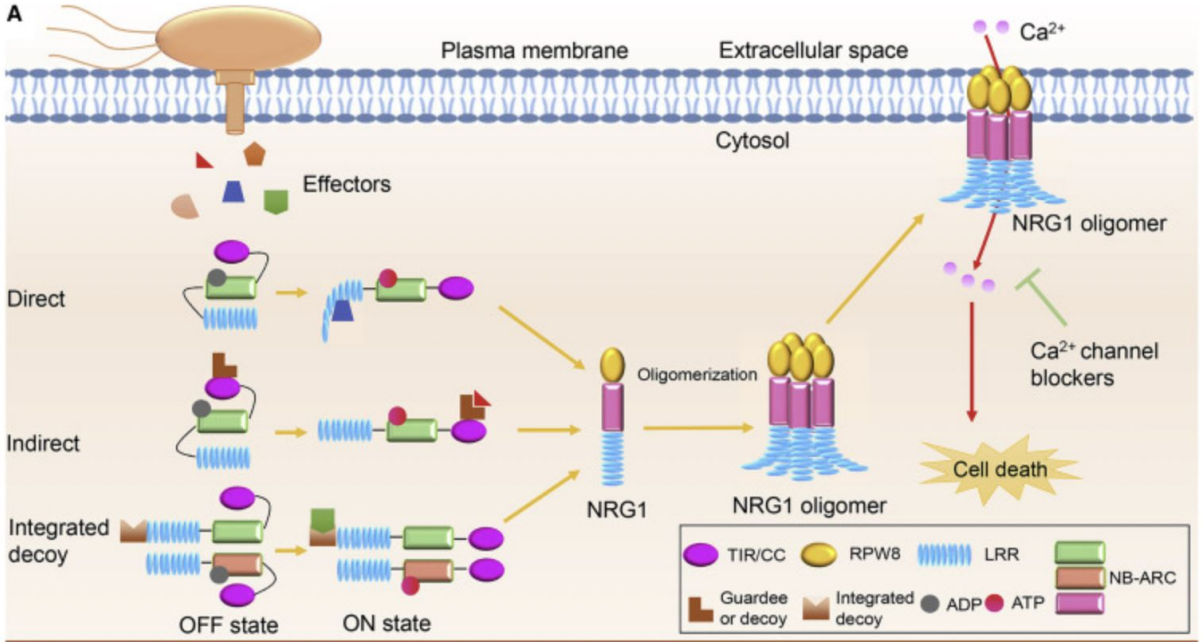
Google DeepMind

Highly accurate protein structure prediction with AlphaFold. 2021

NERSC Perlmutter

Innate immune receptors in plants

NOD-like receptors (NLRs) bind sense pathogen effectors and trigger immune response.
Plant NLRs typically contain Leucine-Rich Repeats (LRR)



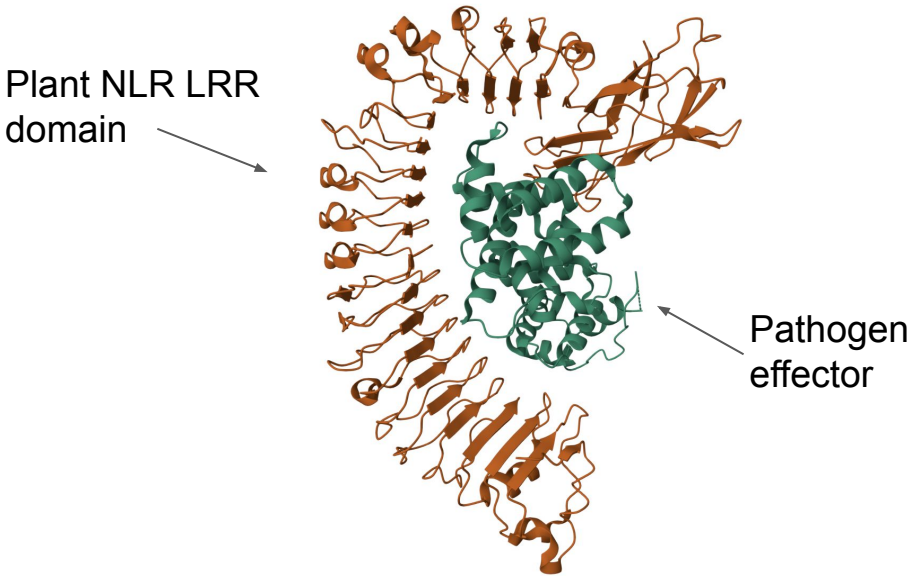
ZAR1 resistosome and helper NLRs: Bringing in calcium and inducing cell death. Chen et al. 2021

Hypersensitive response on leaf

3D structure of LRR domain influences binding specificity of innate immune receptors

Model plant *Arabidopsis thaliana* infected with downy mildew pathogen.

Cryo-electron microscopy of LRR-effector binding

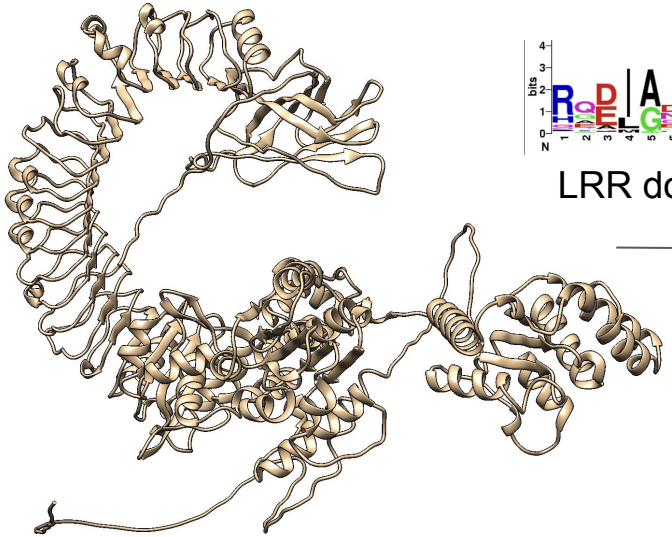


Cryo-EM structure of plant NLR RPP1 LRR-ID domain in complex with ATR1 2020

Protein domain annotation

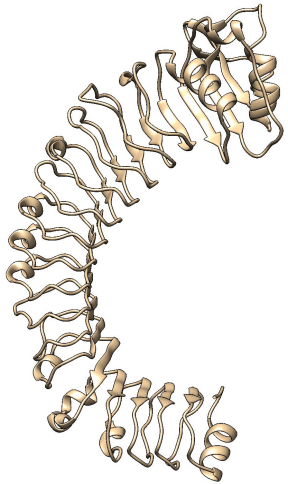
Domains are functional subregions of proteins which fold independently. Domain annotation identifies where these functional units are located within the protein sequence.

Whole protein



LRR domain annotation

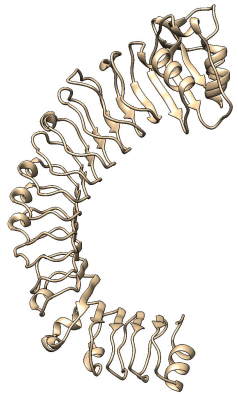
LRR domain only



Standard method of domain annotation: Hidden Markov Models

Annotation is typically done using Hidden Markov Models (HMMs) trained on sequence motifs of protein domains, but these are often inaccurate for highly divergent sequences such as LRR.

HMM-based annotation – incomplete



Actual LRR domain (nearly 100 residues longer)

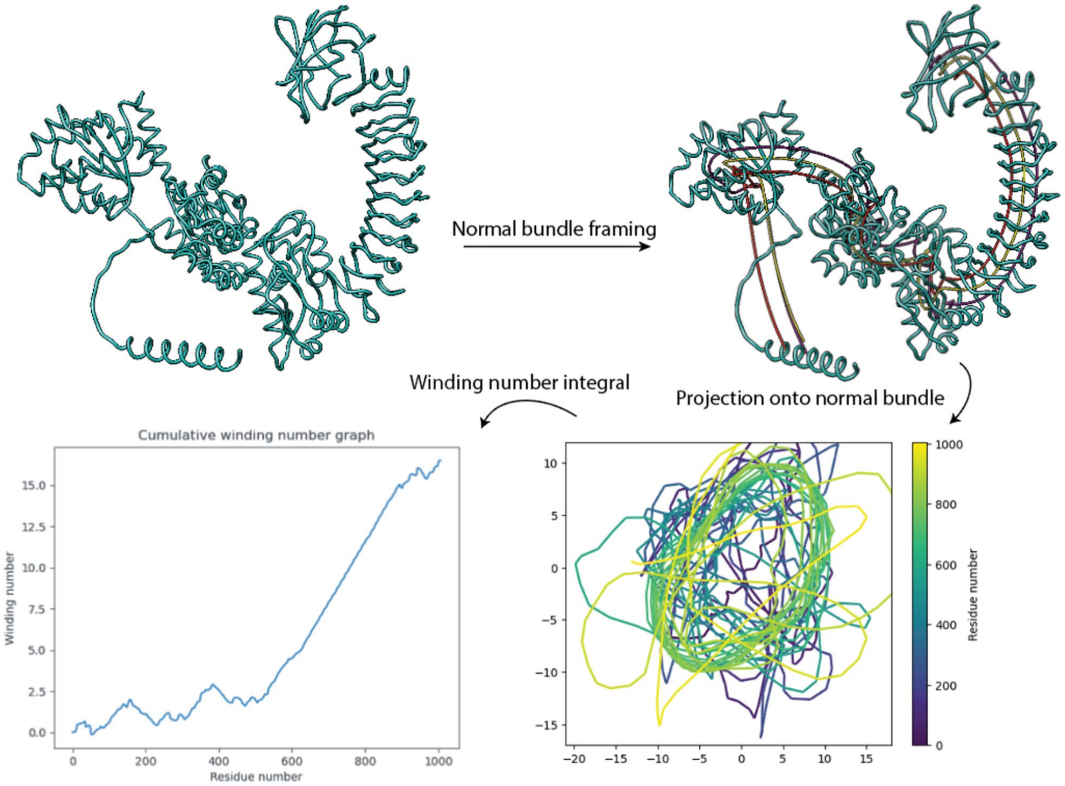
Parallel transport algorithm on orthonormal frame enables winding number computation on 2D projection

Normal bundle framing algorithm:

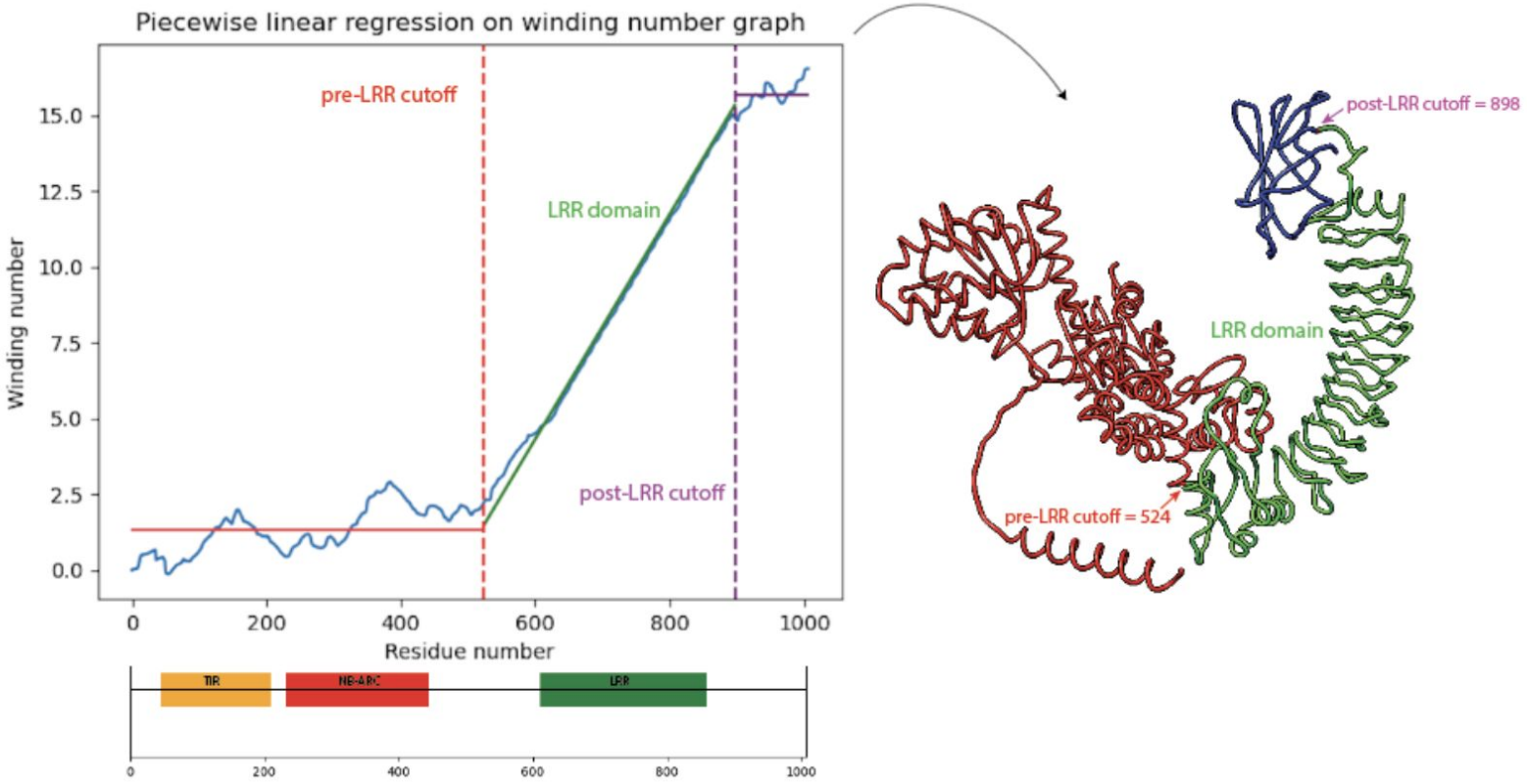
1. Randomly initialize first pair of orthonormal vectors.
2. Project current orthonormal pair onto next normal plane. Compute closest pair of orthonormal vectors using SVD.
3. Iterate step 2 across length of protein.

Cumulative winding number formula

$$w(s) := \frac{1}{2\pi}\int_0^s d\theta = \frac{1}{2\pi}\int_0^s \frac{1}{x^2+y^2}\left(x\frac{dy}{dt} - y\frac{dx}{dt}\right)dt$$

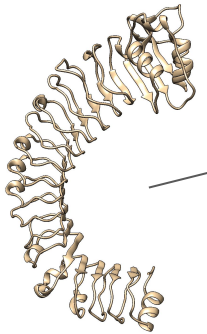


Unsupervised LRR domain annotation via piecewise linear regression on cumulative winding number

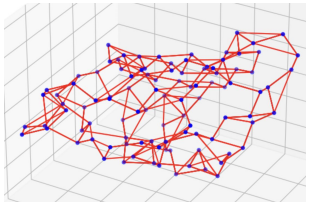


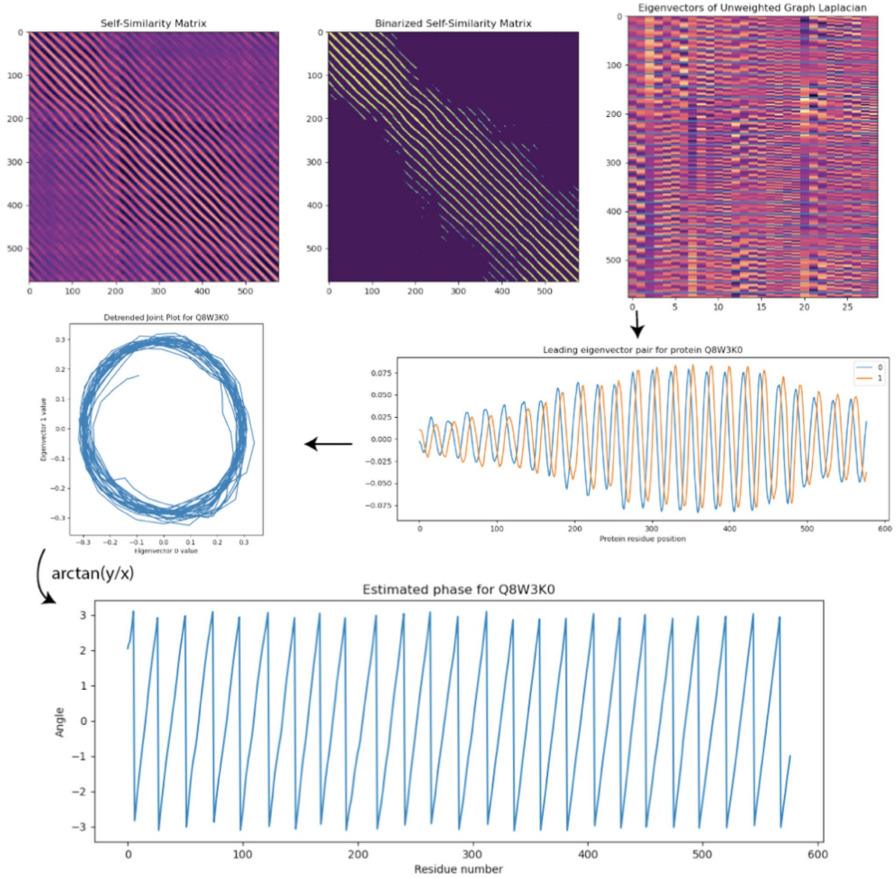
Eigenvectors of graph Laplacian on nearest neighbors graph yield solenoid phase estimation

LRR domain



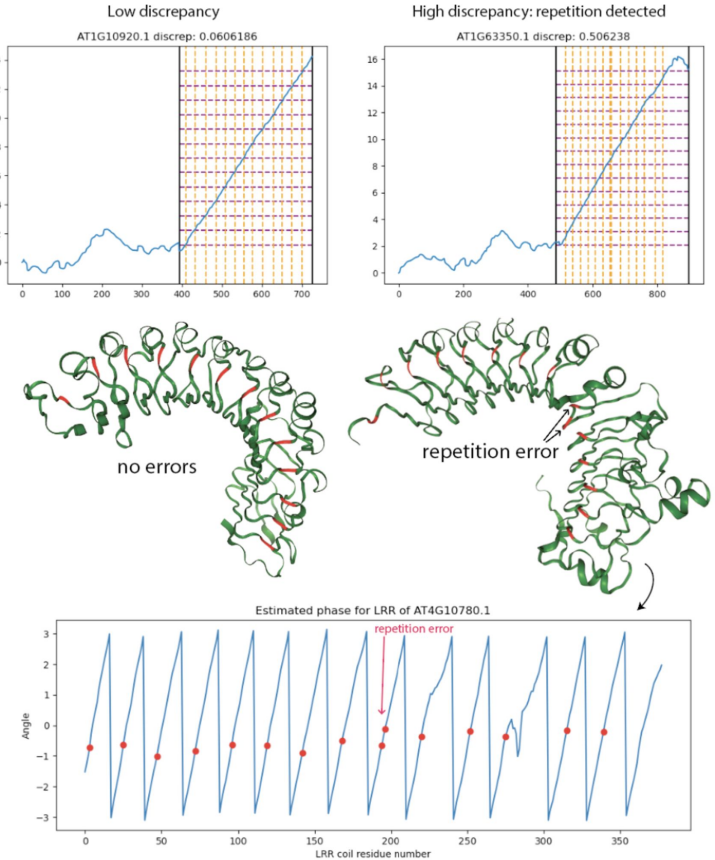
NN-graph





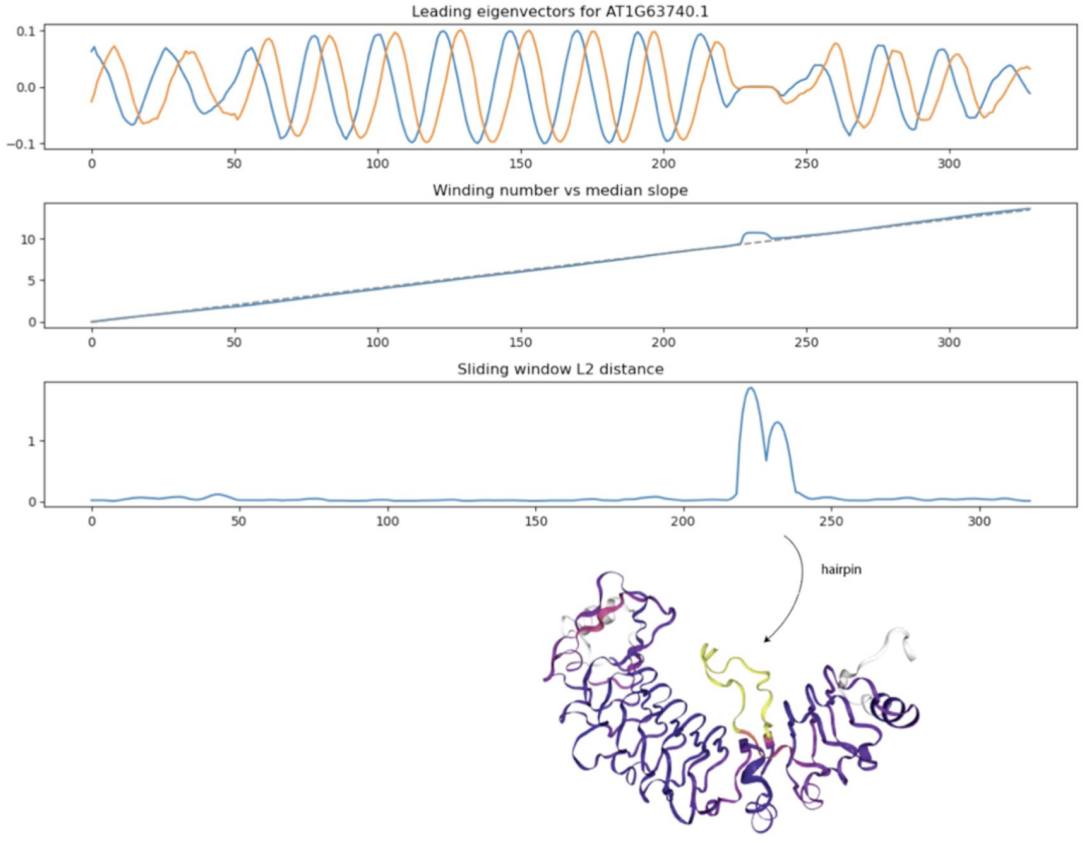
Geometric methods detect errors made by trained models

Our methods can correct for mistakes LRRPredictor, a machine learning model trained on sequence motifs to annotate LRR's.



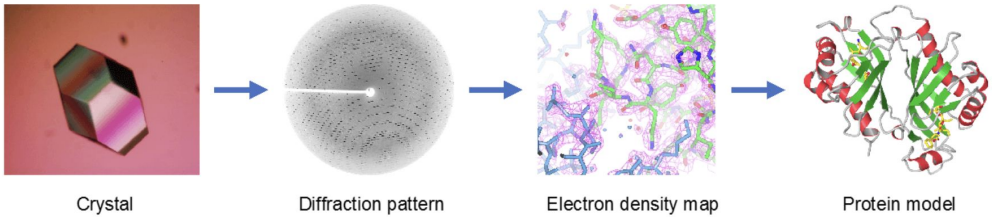
LRRpredictor—A New LRR Motif Detection Method for Irregular Motifs of Plant NLR Proteins Using an Ensemble of Classifiers. 2020

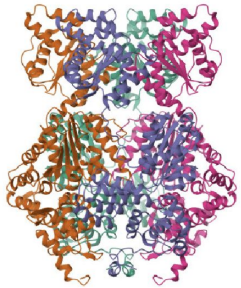
Winding number on graph laplacian eigenvector reveals structural anomalies in LRR coil



Summary

- Protein structure helps us gain insight into their function, but there are far more protein sequences available than we can possibly run Cryo-EM or X-ray crystallography on.
- Accurate in-silico protein structure prediction enables us to systematically characterize large datasets of proteins.
- Future improvements in structure prediction will enable us to better analyze further aspects of protein function, particularly in multi-protein interactions, enzyme function, and more.





Acknowledgements

Krasileva Lab @ UC Berkeley

Alois Cerbu
Graduate student, UC Berkeley

Daven Lim
Undergraduate, UC Berkeley

Chris Tralie
Assistant Professor, Ursinus College

Daniil Prigozhin
Project Scientist, LBNL

Ksenia Krasileva
Assistant Professor, UC Berkeley

berkeley urap undergraduate research apprentice program

NIH

DOE CSGF