The language of proteins and cells

Claire McWhite

University of Arizona

- **B.S. Biochemistry and Cell Biology –** Rice University
- **Ph.D. Cell and Molecular Biology –** The University of Texas at Austin
  - Advisor: Edward Marcotte
- **Postdoc. Lewis-Sigler Scholar –** Princeton University
  - Mentors:
    - Joshua Akey: Human evolutionary genomics
    - Mona Singh: Computer science, cancer driver mutations
    - Martin Jonikas: Algal CO2-fixation
- **Currently Assistant Professor in Molecular Biology at the University of Arizona**

NIH: Lung cancer molecular biology
Rice: Yeast synthetic biology

Trained as an experimental biologist.

UT: Proteomics, interactomics

Princeton and Arizona: Biological language models

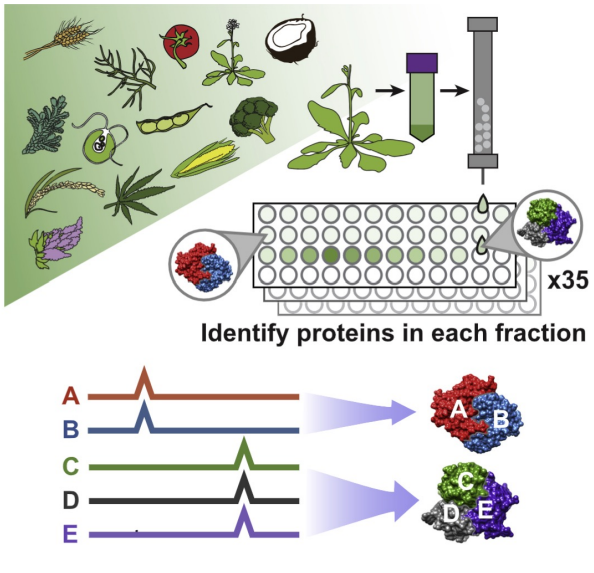
Cell

Resource

A Pan-plant Protein Complex Map Reveals Deep Conservation and Novel Assemblies

Claire D. McWhite,[1,6] Ophelia Papoulas,[1,6] Kevin Drew,[1] Rachael M. Cox,[1] Viviana June,[1] Oliver Xiaoou Dong,[2,3] Taejoon Kwon,[4] Cuihong Wan,[1,5] Mari L. Salmi,[1] Stanley J. Roux,[1] Karen S. Browning,[1] Z. Jeffrey Chen,[1] Pamela C. Ronald,[2,3] and Edward M. Marcotte[1,7,*]

Derived conserved, stable protein complexes from biochemical fractionation and proteomics of 13 plant species



Multiple novel stable protein complexes conserved across plants

New methods for:

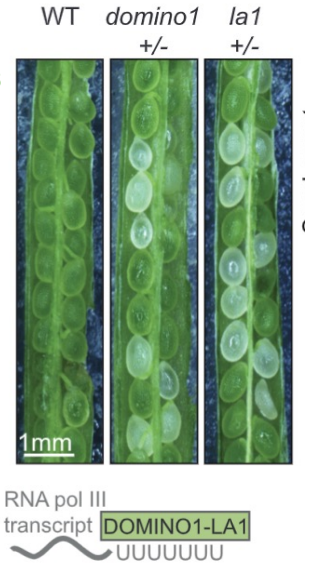
- Interpretation of multispecies data
- Protein identification from polyploid species

Examples of new protein complexes

Novel DOMINO1-LA1 complex

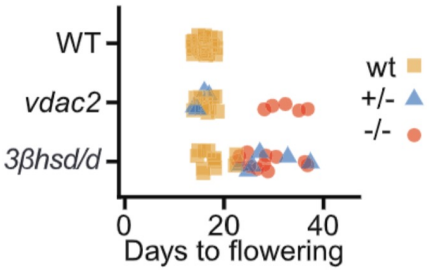
Failed embryogenesis phenotype

Arabidopsis seed pod



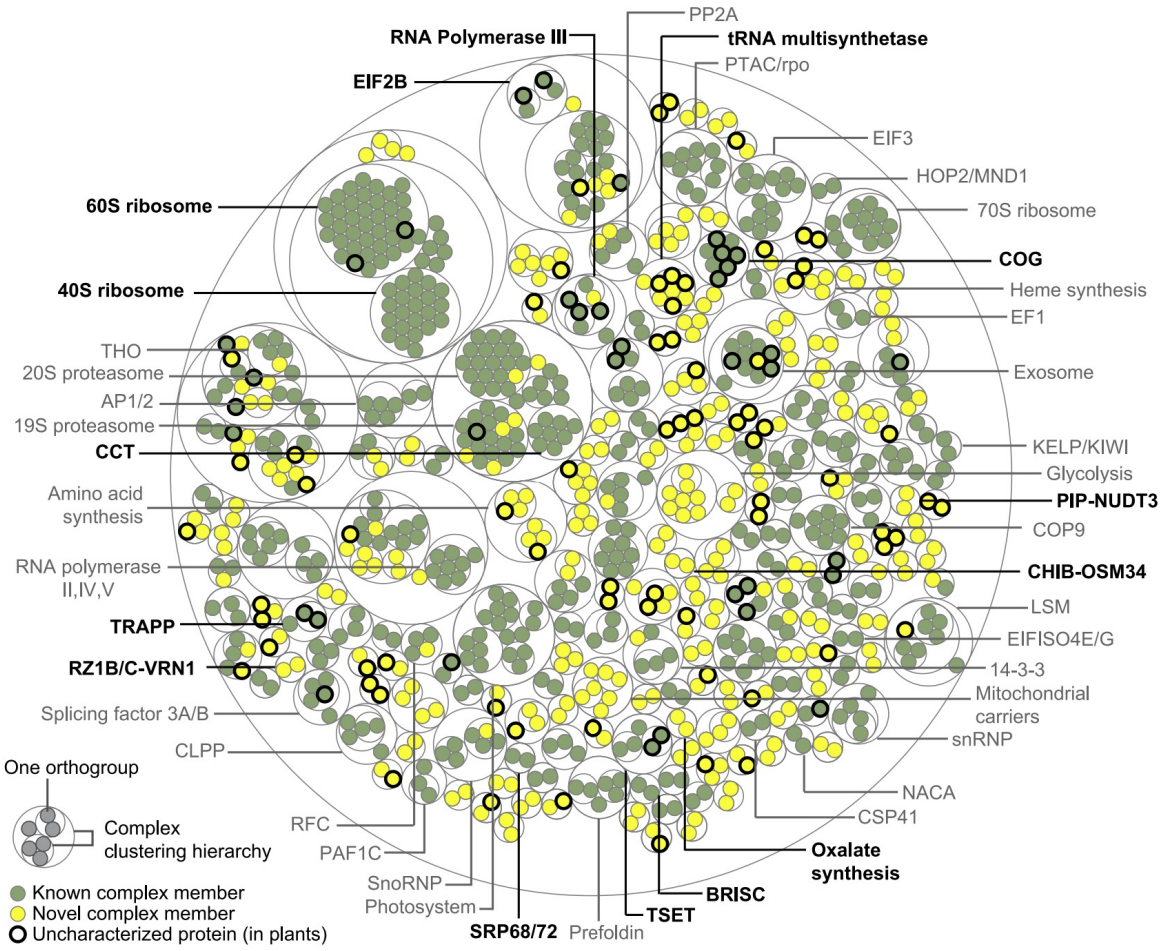
Novel VDAC2-3βHDD complex

Late flowering and infertility phenotype



Arabidopsis main stem

- LA1 found broadly across eukaryotes, DOMINO1 plant-specific


PP2A
RNA Polymerase III
tRNA multisynthetase
PTAC/rpo
EIF2B
EIF3
HOP2/MND1
60S ribosome
70S ribosome
COG
40S ribosome
Heme synthesis
EF1
THO
Exosome
20S proteasome
AP1/2
19S proteasome
KELP/KIWI
CCT
Glycolysis
Amino acid synthesis
PIP-NUDT3
COP9
RNA polymerase II,IV,V
CHIB-OSM34
LSM
TRAPP
EIFISO4E/G
RZ1B/C-VRN1
14-3-3
Mitochondrial carriers
Splicing factor 3A/B
snRNP
CLPP
NACA
CSP41
RFC
Oxalate synthesis
PAF1C
SnoRNP
BRISC
Photosystem
TSET
SRP68/72
Prefoldin
One orthogroup
Complex clustering hierarchy
Known complex member
Novel complex member
Uncharacterized protein (in plants)

Article

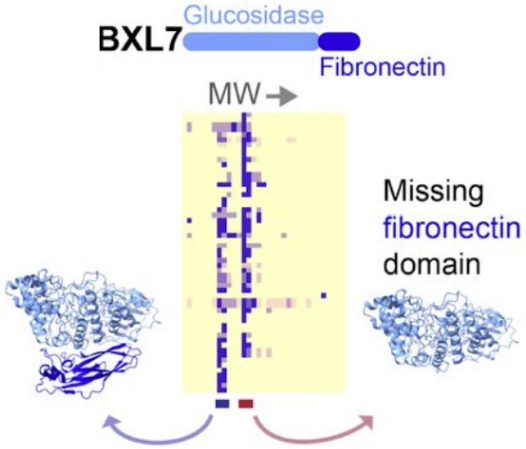
TRANSPARENT PROCESS OPEN ACCESS CHECK FOR UPDATES

molecular systems biology

Alternative proteoforms and proteoform-dependent assemblies in humans and plants

Claire D McWhite[1✉], Wisath Sae-Lee[2], Yaning Yuan[3], Anna L Mallam[2], Nicolas A Gort-Freitas[4], Silvia Ramundo[5], Masayuki Onishi[3] & Edward M Marcotte[2]

Observed that, in rare cases, a protein seen in proteomics would be missing peptides from one or more domains



Surveyed human, plant, and algae proteomics data for these short proteoforms

- ~1% of proteins appear to occur in a form missing one or more domains
- Found conserved truncation proteoforms
- Found proteoform-specific protein complexes

Ongoing frustrations interpretating protein data

- Can measure abundance, detect protein variants, find protein interactions
- However**, frustrating** lack of annotations for many proteins
- Started my postdoc around the time of the first next-gen protein language models
  - The same type of deep learning models used to capture and translate human languages, but created from protein sequences
- **Can these models help us uncover biological properties that aren't obvious from sequence alone?**

Protein sequences are like a language

- Sequences of tokens, where token meaning depends on context

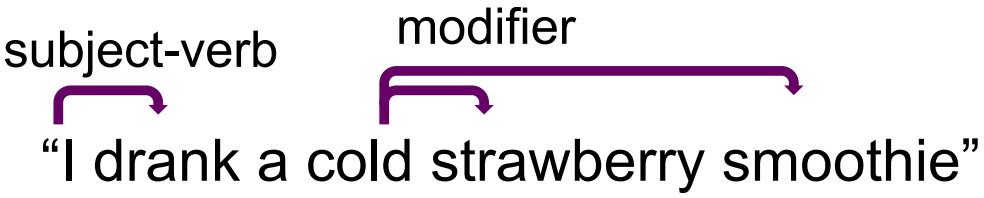
- Language: “I threw a coin in the wishing **well**” 

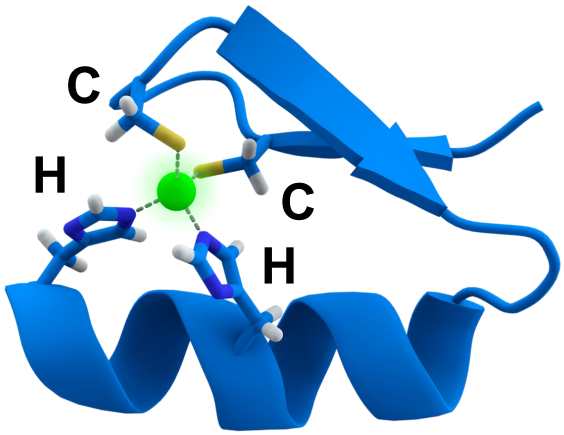
“I did **well** on the exam”

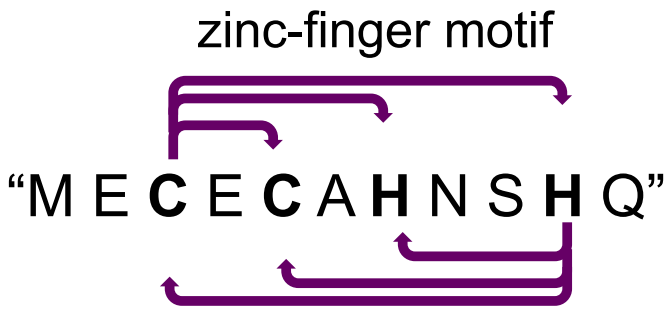
- Protein: “M E C E **R** A H N S H Q **R** S K **R** Q N T F A”

Language models capture syntax of human language, and patterns sequences

- There are specific relationships between tokens

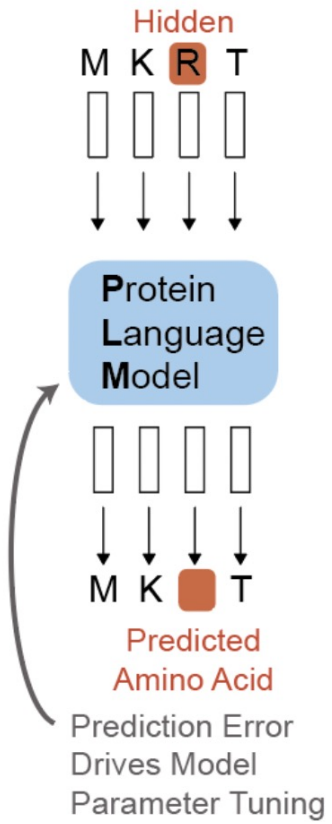






How are language models made?

- One method: Predict hidden amino acids
- The model learns to use amino acid context to predict identity of hidden amino acids
  - Inaccuracies in the prediction are how model weights are learned, over millions of sequences



Input sentence:	The cat meows
The model sees:	The cat [hidden word]
First guesses are random:	The cat [philosophizes, lamp, ..]
Parameters are updated until:	The cat [meows, purrs, sleeps..]

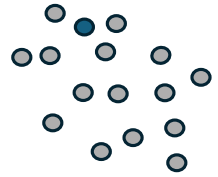
Models capture what is **permitted** in a language

Language models produce numeric representations of a sequence

A protein sequence

M E T A G H A I L L I A I → Protein Language Model → 1x1152 Numeric vector → Captures protein meaning

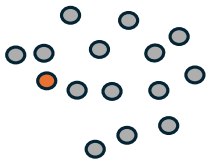
Embeddings for many proteins



Expressed genes in a cell (ordered by abundance)

CD3D | CD3E | TRAC | LCK … → Single-cell Foundation Model → 1x1152 Numeric vector → Captures cell identity

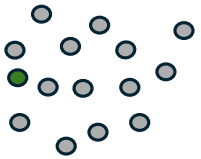
Embeddings for many cells



Peptide mass spectra

m/z: [101.07, 129.10, 175.12, 244.17]
intensity: [0.12, 0.35, 0.08, 0.91] → Spectral Foundation Model → 1x1152 Numeric vector → Captures spectral properties

Embeddings for many spectra



Lab Goals

- Discover functions for uncharacterized proteins
- Identify new features of disease-relevant proteins
- Understand what determines molecular interactions

Project 1: Finding signatures of protein function and cell identity

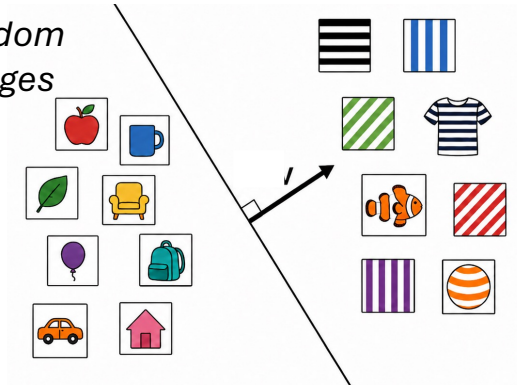
A new framework for working with language models

We identify numeric signatures of different features

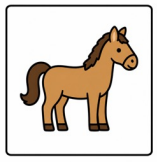
Image -> Model -> Image embedding

Identify shift in embeddings associated with “stripes”

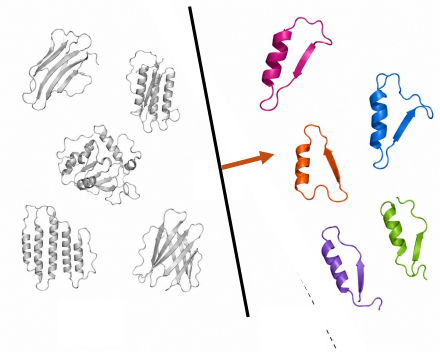
Random images



Stripe-related shift/concept aligns more closely with zebra embedding than horse.



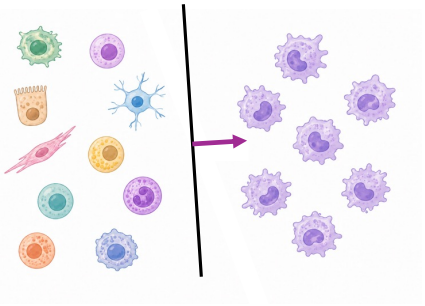
Protein sequence -> Protein embedding



Random proteins *Zinc fingers*

Identifiable concept of “Zinc-Finger”(1)

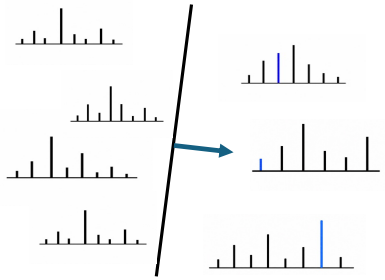
Single-cell expression -> Cell embedding



Random cells *Macrophages*

Identifiable concept of “Macrophage”

Peptide spectra-> Spectra embedding

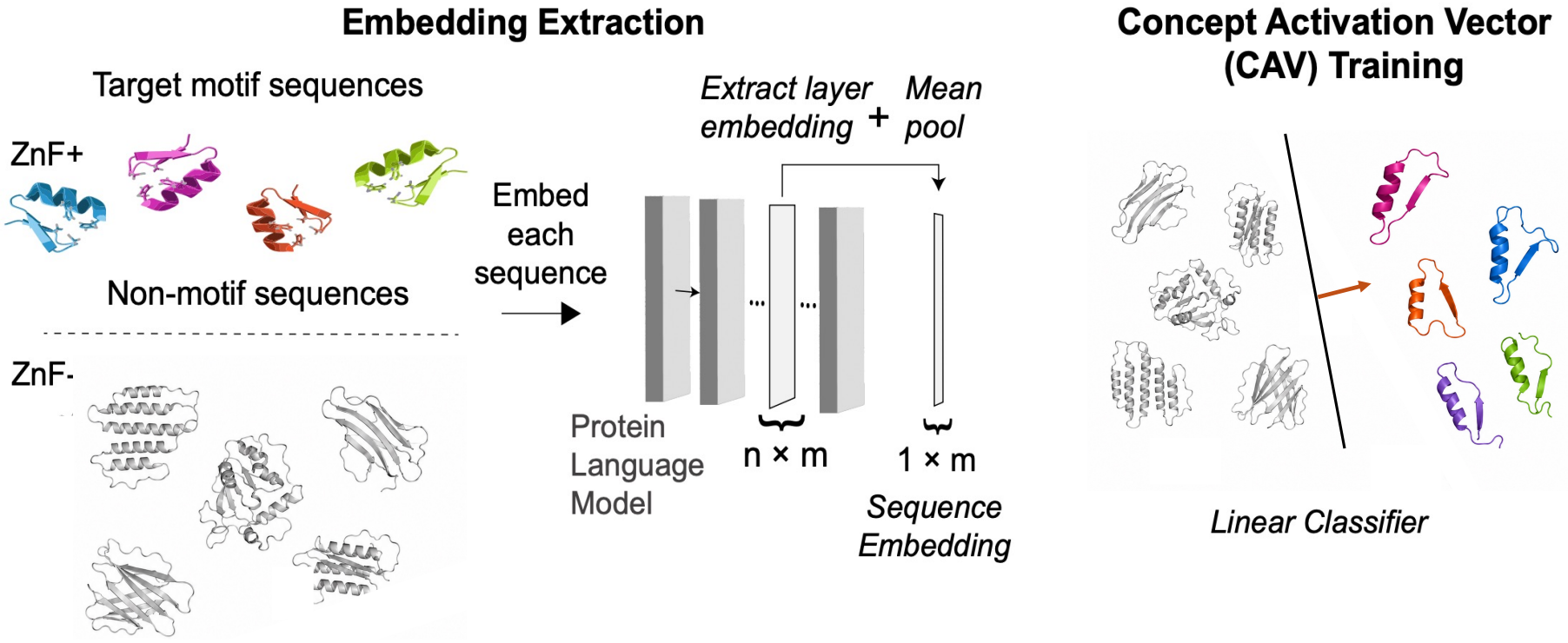


Random spectra *w/* SUMOylation

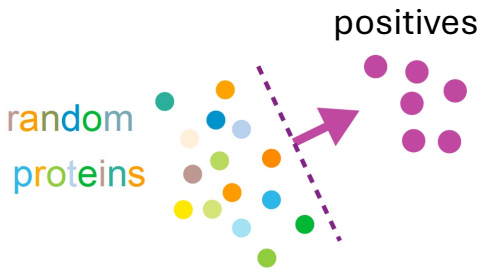
Identifiable concept of “SUMOylation”

(1) Shamail & McWhite, Automated Protein Motif Localization using Concept Activation Vectors in Protein Language Model Embedding Space, ArXiv 2025

Extracting the numeric signature of zinc-finger from protein language models

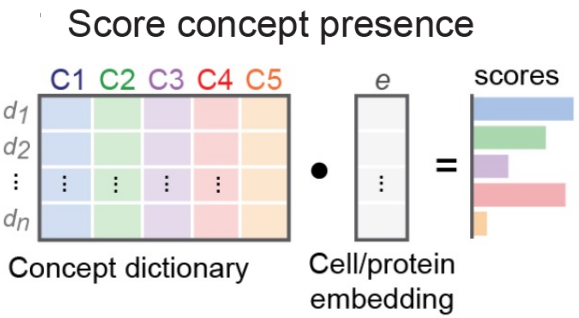


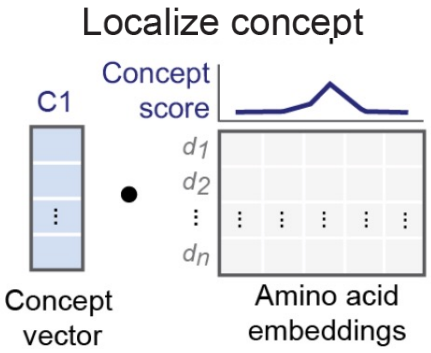
Useful properties of concept signatures

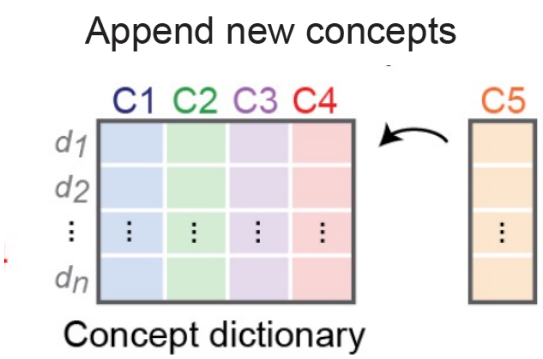


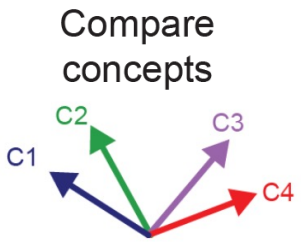
Training a concept just needs simple linear classifier between embeddings

The direction (↗) is the **concept** vector







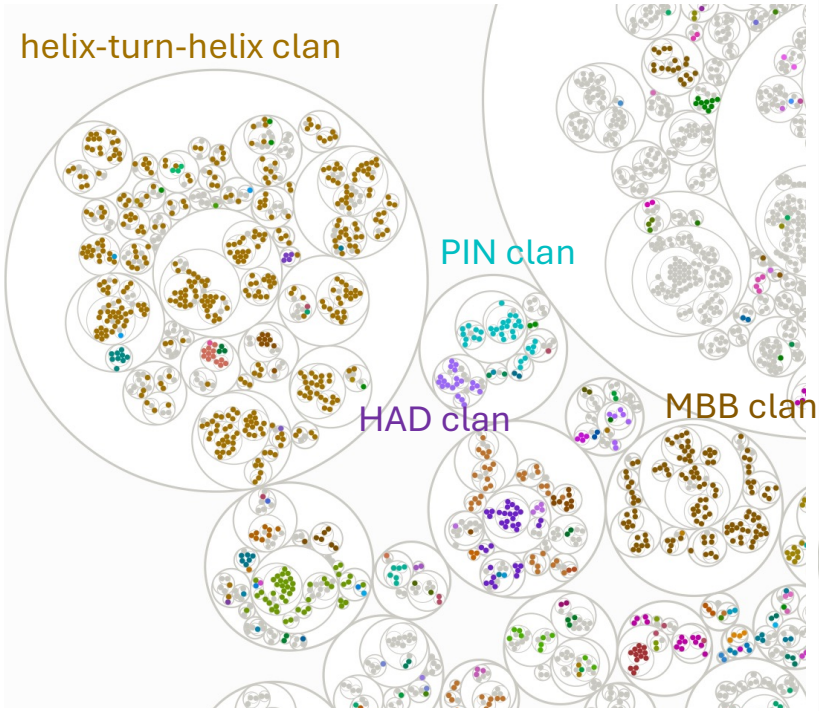


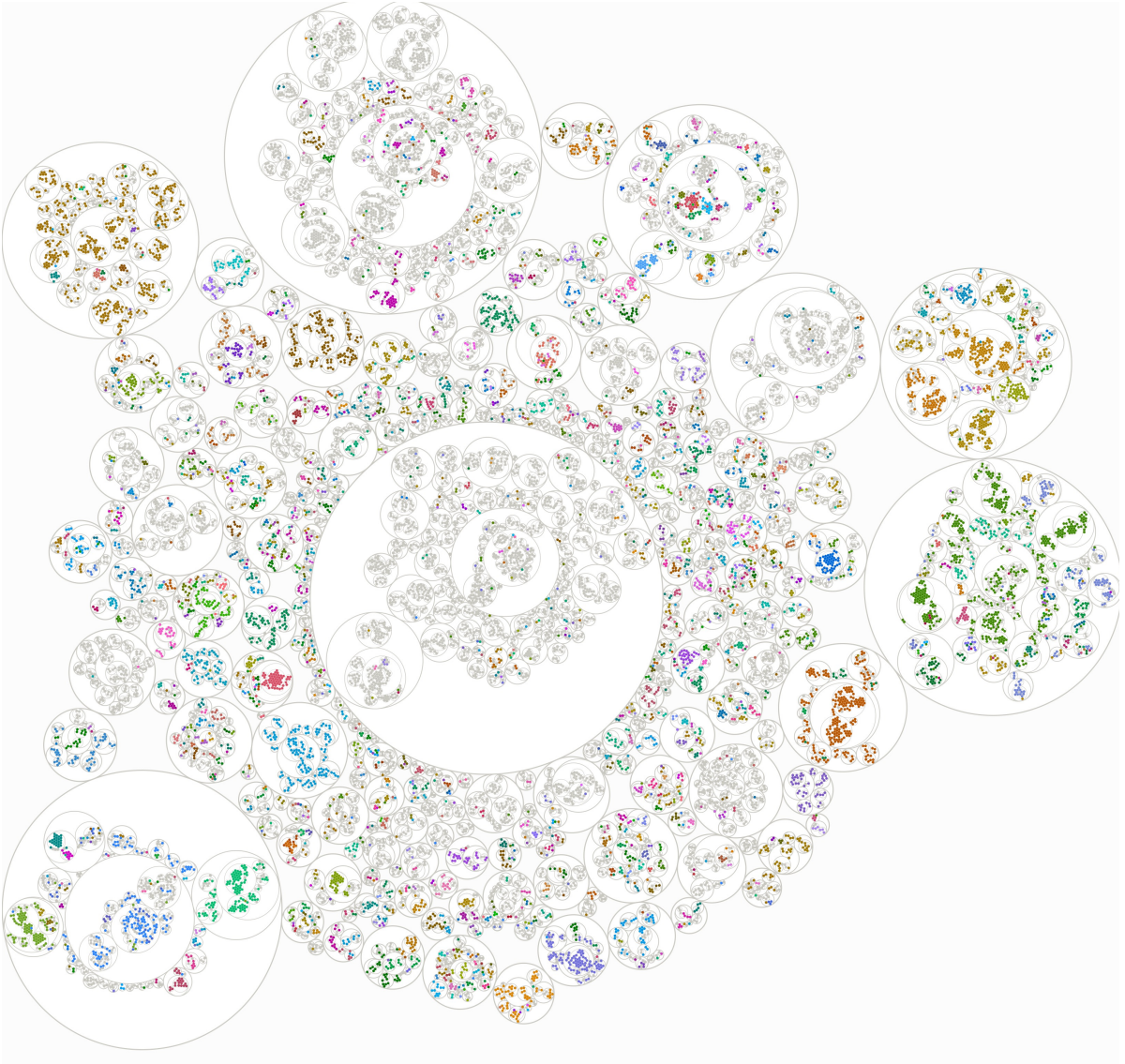
What can we do with concept signatures:

vignette 1: Hidden enzymes and new binding sites in uncharacterized domains

Hierarchical clustering of 27000 PFAM motifs

Colored by Pfam Clan





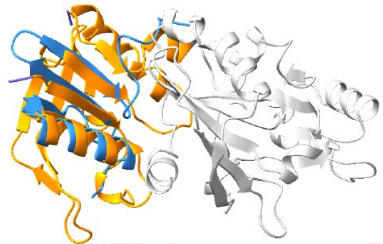
We find many unannotated homologies in PFAM domains

Simple homology

PF25029
MOM1-like domain
vs.
Cyclin_N

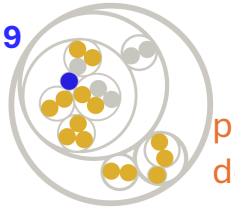
Misannotation:
ATP-grasp-like is actually pre-ATP-Grasp-like

PF18419
ATP-grasp-like domain vs.
Pre-ATP-Grasp domain(GshB)



ATP-Grasp domain (GshB)

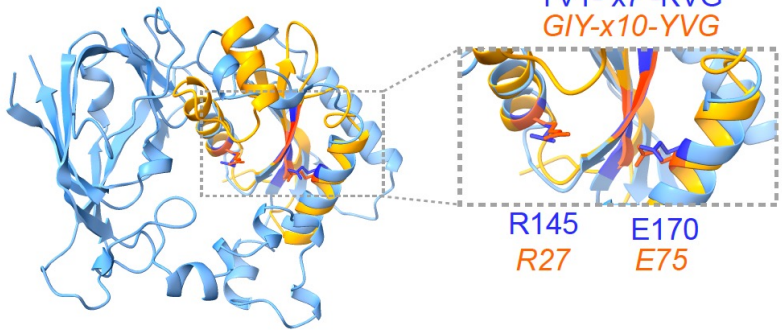
PF18419



pre-ATP-Grasp domains

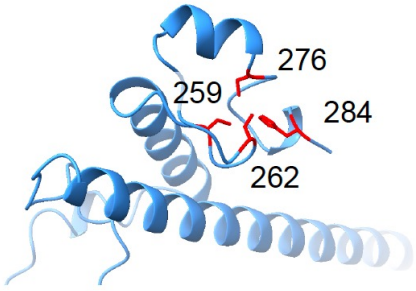
Hidden homology: An endonuclease domain hidden in uncharacterized DUF2797

DUF2797 vs. *GIY-YIG_endonuc*



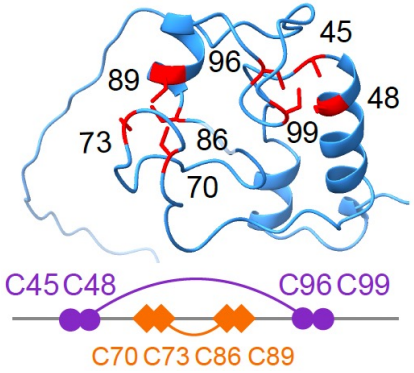
Multiple completely novel zinc sites

DUF4187 / GPATCH11 C-terminus



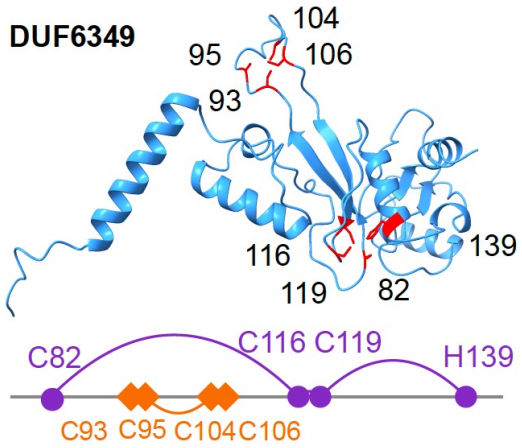
GPATCH11 causes retinal/neurodevelopmental disease in humans

DUF2039 / C9orf85



C9orf85 recently discovered member of the human BRISC complex

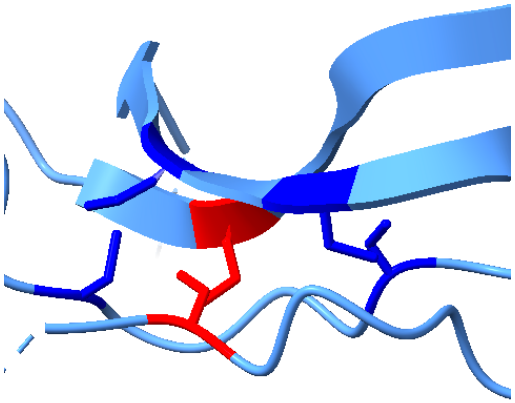
DUF6349



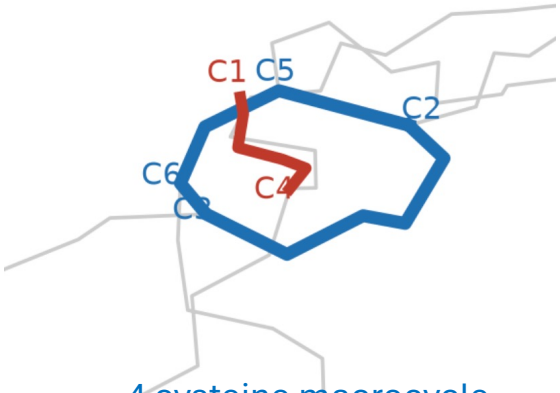
Very rare/interesting tandem zinc-site topology

52 new candidates zinc sites structurally confirmed, still more to screen

A novel cystine knot in human CILP



Cartilage intermediate layer protein 1

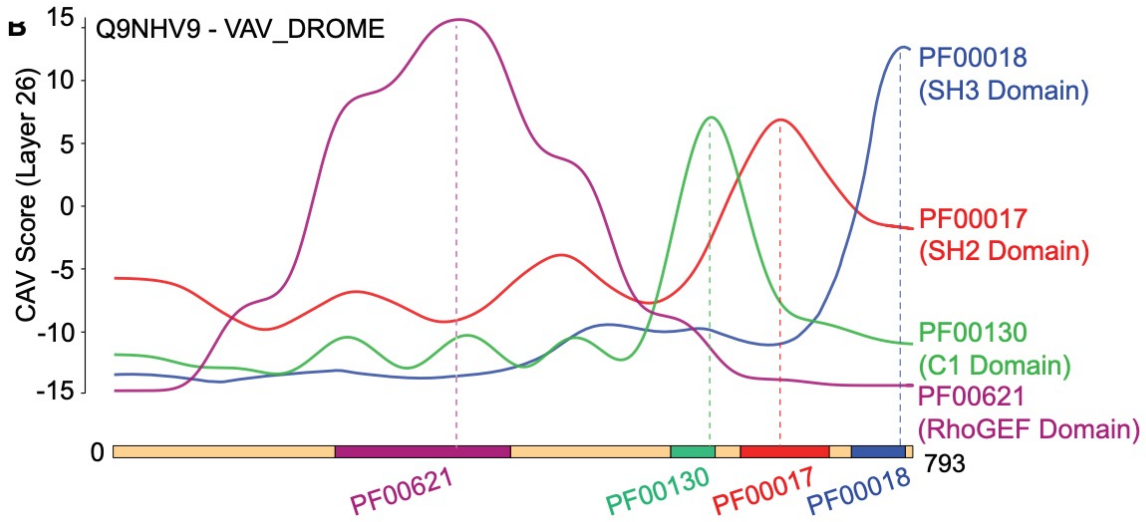


4 cysteine macrocycle

Threading disulfide

- CLIP linked to osteoarthristis and cartilage degradation
- Known to bind TGF-β1, but binding site unknown
- However, other growth factor-binding domains use a *cystine knot motif* to bind TGF-β1

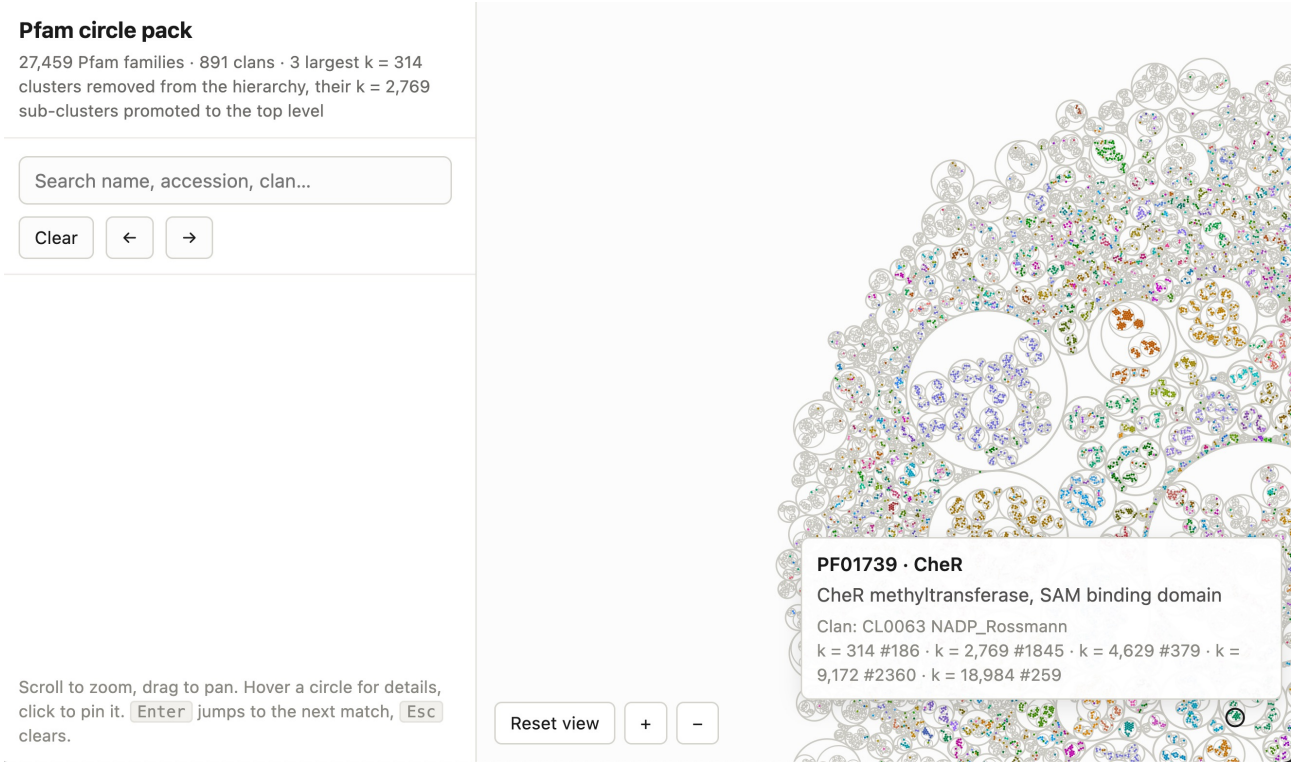
We can also use PFAM domain concept signatures to detect and locate domains in proteins



(1) Shamail & McWhite, **Automated Protein Motif Localization using Concept Activation Vectors in Protein Language Model Embedding Space**, BiorXiv, 2025

PFAM clustering can be explored here

- https://mcwhitelab.org/pfam-circlepack/

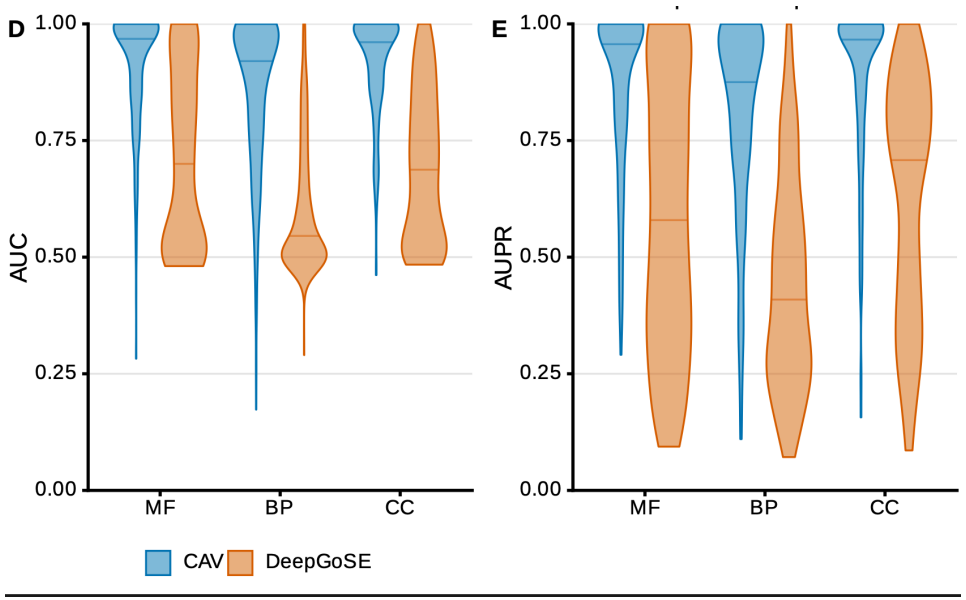


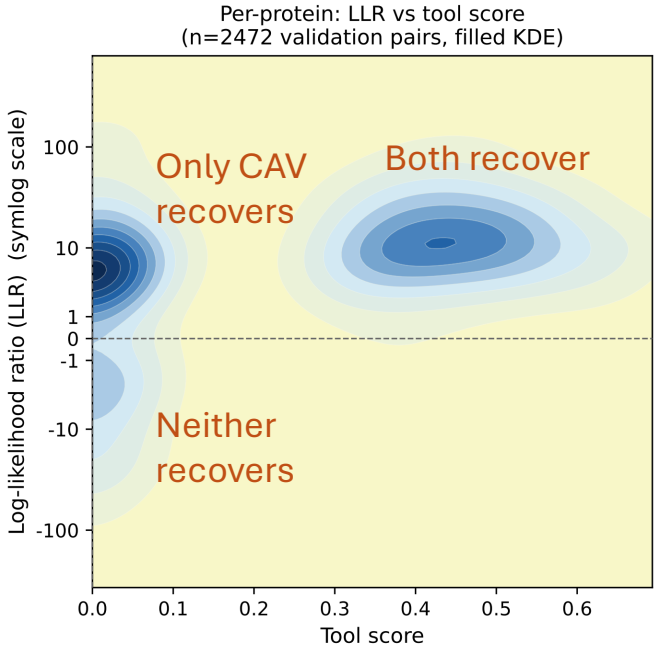
What can we do with concept signatures: **vignette 2**: Predicting what uncharacterized proteins do

CAVs trained for 2000 GO terms, and 1000 Enzyme Commission terms

Concept vectors are excellent Gene Ontology-term predictors

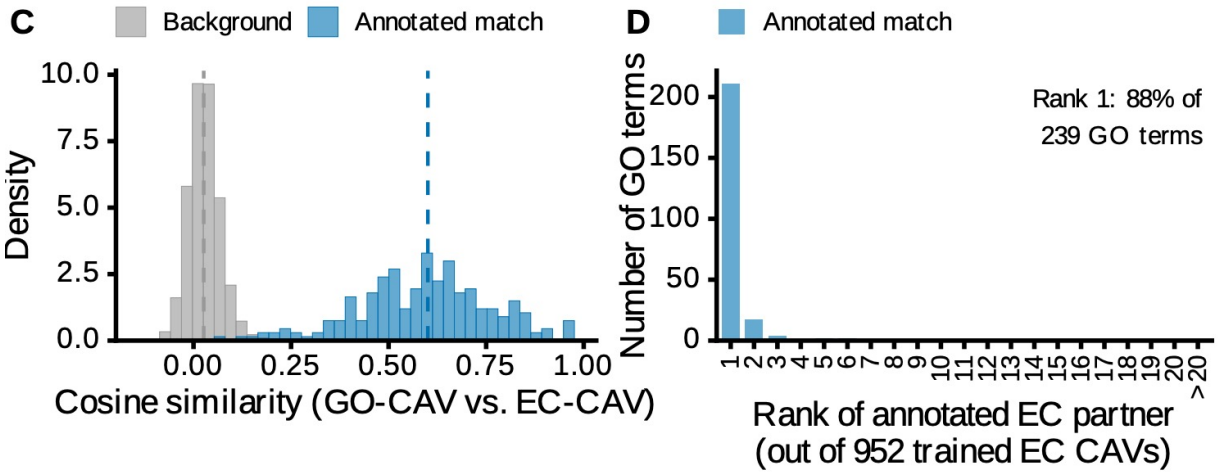
- Validation set: 2500 proteins newly annotated with molecular function (post 2021)
- Compared to DeepGoSE predictions (sequence-based multiclassifier)

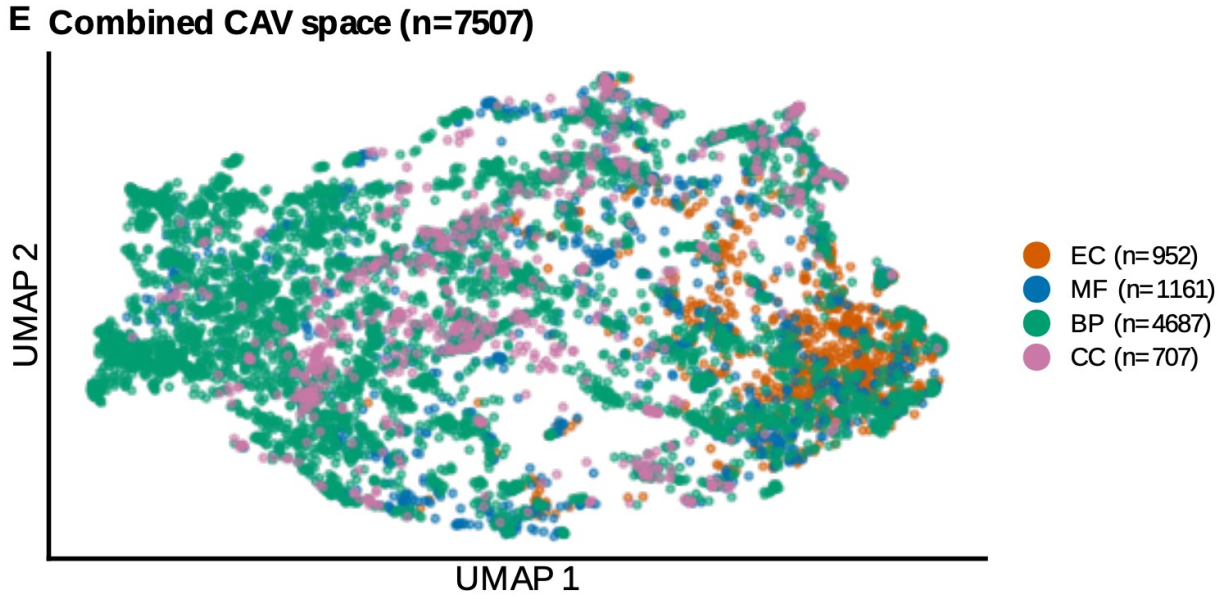




Comparing vectors similarity of GO-CAVs and EC-CAVS recovers the hand-annotated GO-EC mapping

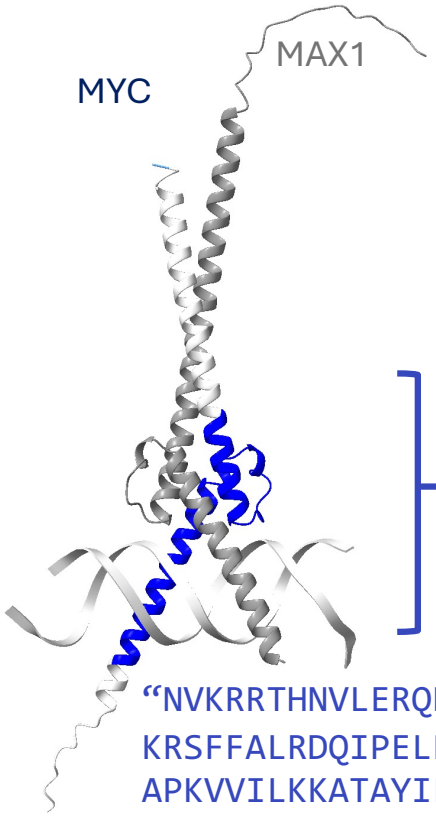
Natural combined embedding space for two different sources of annotations





What can we do with concept signatures:
vignette 3: Studying DNA-binding and co-receptor specificity

Surprisingly, a *single example sequence* can sometimes define a searchable biological concept



"NVKRRTHNVLERQRRNEL
KRSFFALRDQIPELENNEK
APKVVILKKATAYILS"

E-box-binding region of MYC

Learn **concept signature** for CACGTG-binding region of MYC

Rank proteins by how much their embedding contains this **signature**

Top 20 **E-box-binding** aligned proteins

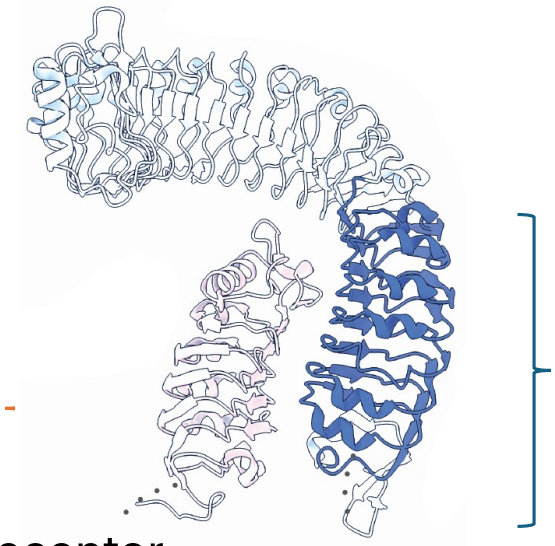
1	MYCN
2	MYCL
3	MYC
4	MYCLP1
5	USF1
6	USF2
7	MAX
8	TFAP4
9	MNT
10	MLXIPL
11	MLXIP
12	MAD4
13	MLX
14	MAD1
15	MXI1
16	MAD3
17	TCFL5
18	STRA8
19	TFEC
20	MITF

All proteins known to bind **CACGTG E-box**

Similarity finds proteins that look like MYC; the CAV finds proteins that share what is distinctive about MYC DNA-binding region.

I’m particularly interested in Leucine-Rich-Repeat Receptor Kinases

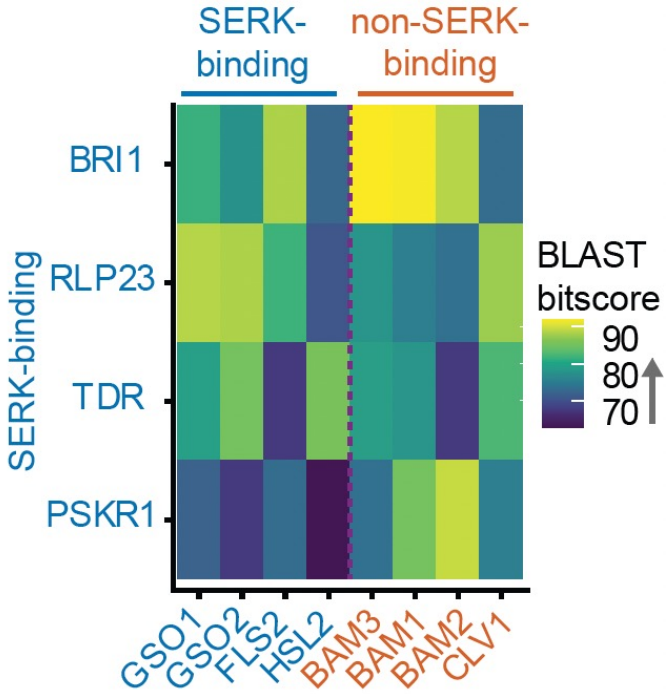
LRR-RK’s bind co-receptors



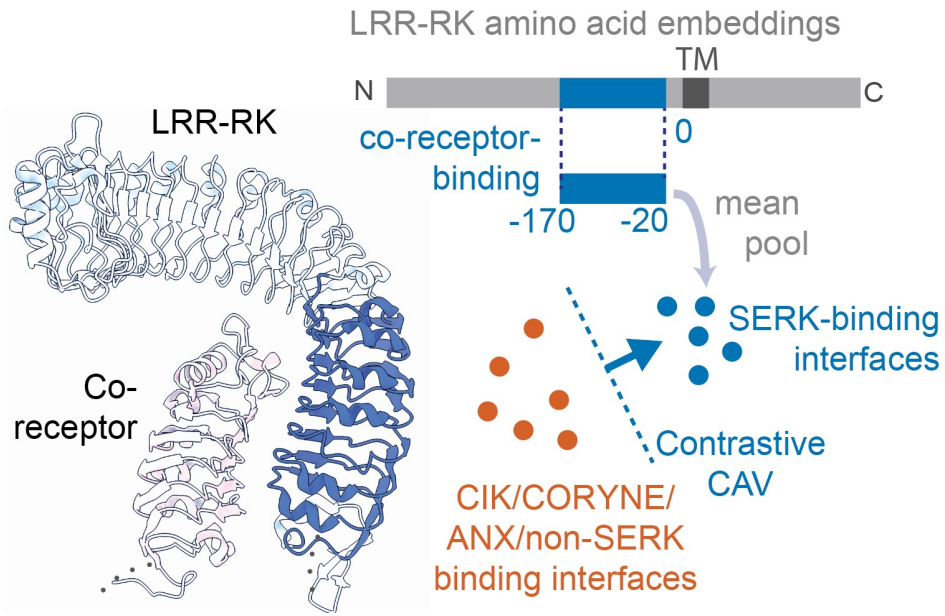
Co-receptor (SERK, CIK, CORYNE, ANX)

Co-receptor-binding region

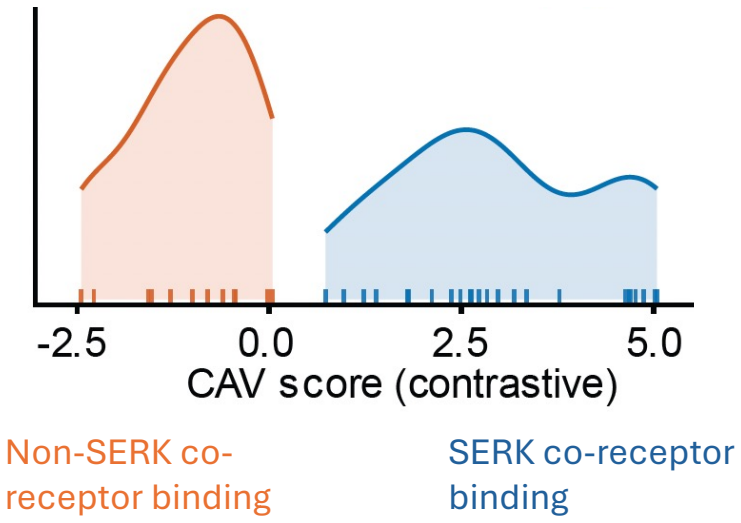
Which class of co-receptor an LRR-RK will bind is not always predictable by sequence



Simple classifier distinguishes co-receptor preference



Trained on Arabidopsis LRR-RKs only



Prediction on non-Arabidopsis LRR-RKs only

What can we do with concept signatures: **vignette 4**: Navigate single-cell datasets

+ alternate route to differential expression

Language models produce numeric representations of a sequence

A protein sequence

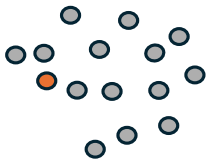
M E T A G H A I L L I A I → Protein Language Model → 1x1152 Numeric vector → Captures protein meaning

Embeddings for many proteins

Expressed genes in a cell (ordered by abundance)

CD3D | CD3E | TRAC | LCK … → Single-cell Foundation Model → 1x1152 Numeric vector → Captures cell identity

Embeddings for many cells

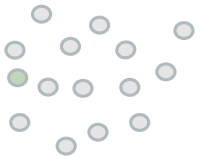


Peptide mass spectra

m/z: [101.07, 129.10, 175.12, 244.17]
intensity: [0.12, 0.35, 0.08, 0.91]

→ Spectral Foundation Model → 1x1152 Numeric vector → Captures spectral properties

Embeddings for many spectra

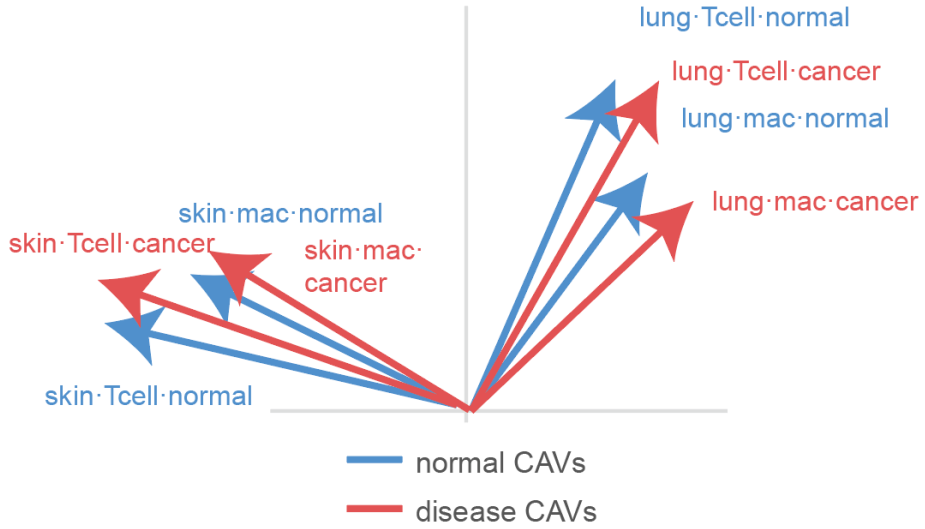


Concepts vectors are a new framework to dissect transcriptomic data

- Connect gene expression to concept signatures
- Connect gene expression to *transitions* between concepts (i.e. normal vs. cancer)
- Easily project away confounding variables (batch, etc.)

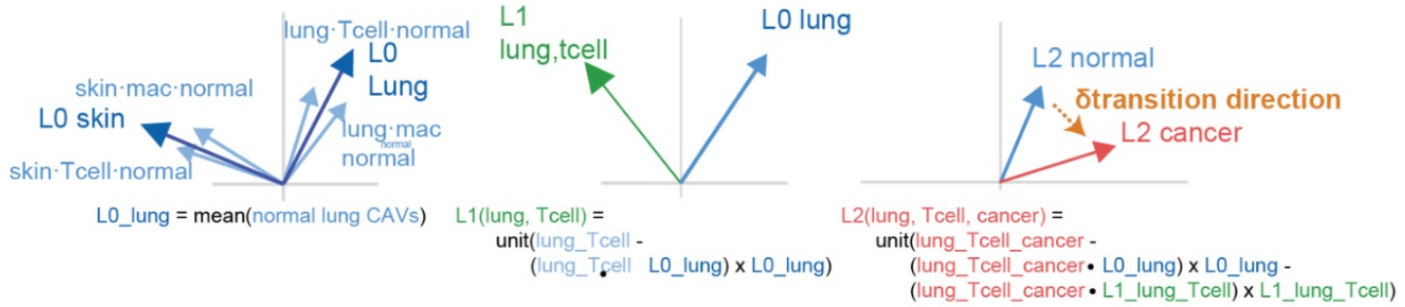
Extract Concepts for:

- Tissues
- Cell types
- Disease States
- Pathways
- Patient Age
- etc.



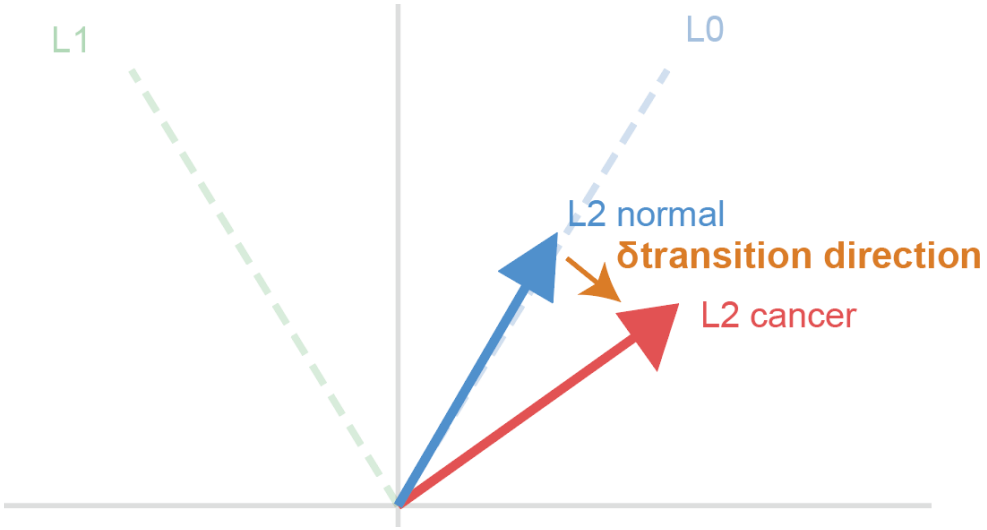
Vector operations can isolate specific signals

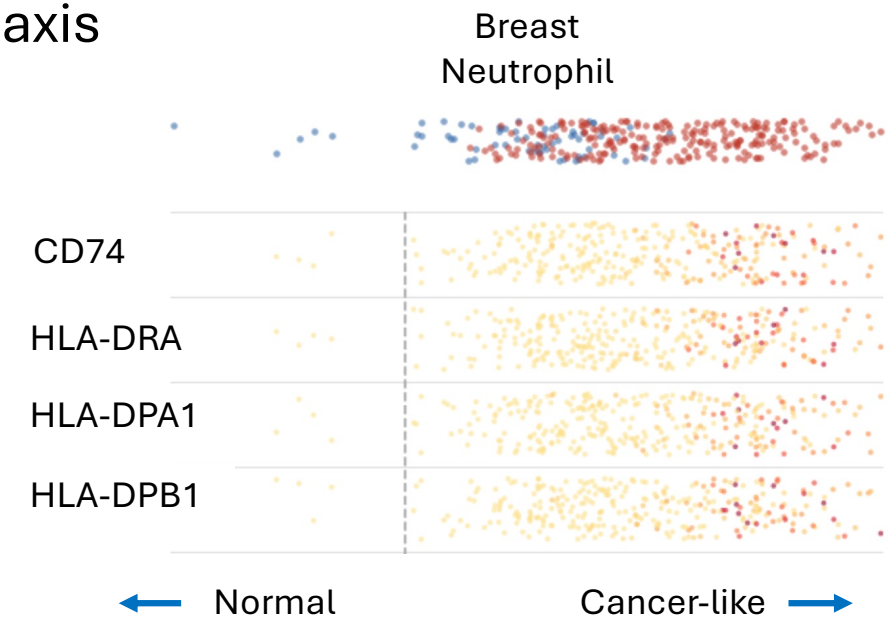
- 1. Remove the shared “lung” direction
- 2. Remove the shared “T-cell identity” direction
- 3. What remains is the normal tissue → lung cancer **transition direction**



We can then connect the transition direction to gene expression!

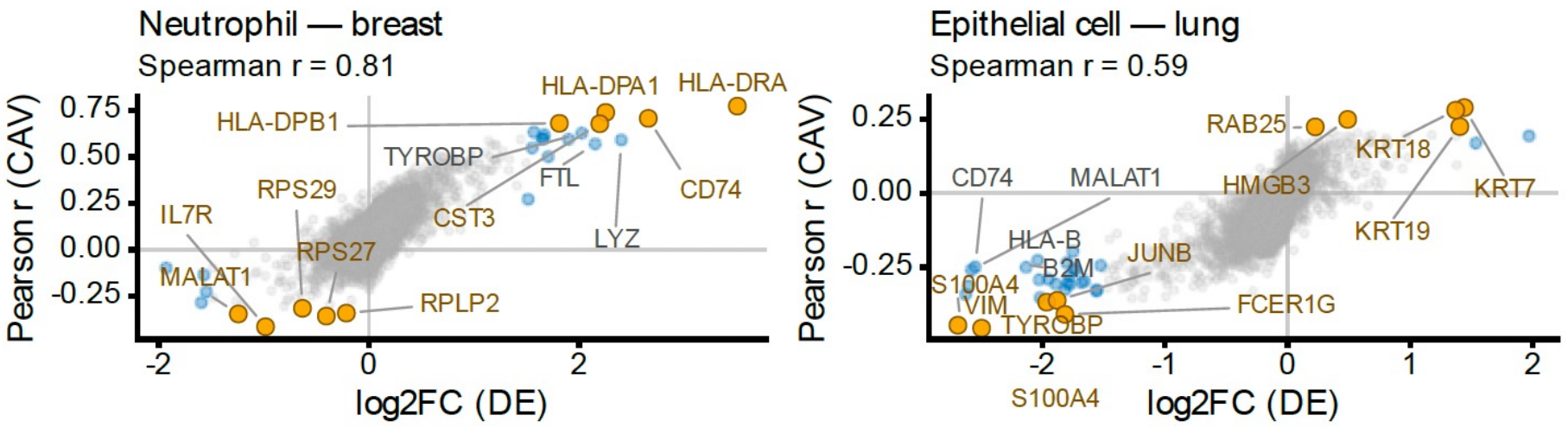
By placing cells along the normal → cancer transition, we can find genes whose expression varies along the axis





Captures the enrichment of tumor-associated $HLA{-}DR^{+}CD74^{+}$ antigen presenting neutrophils!

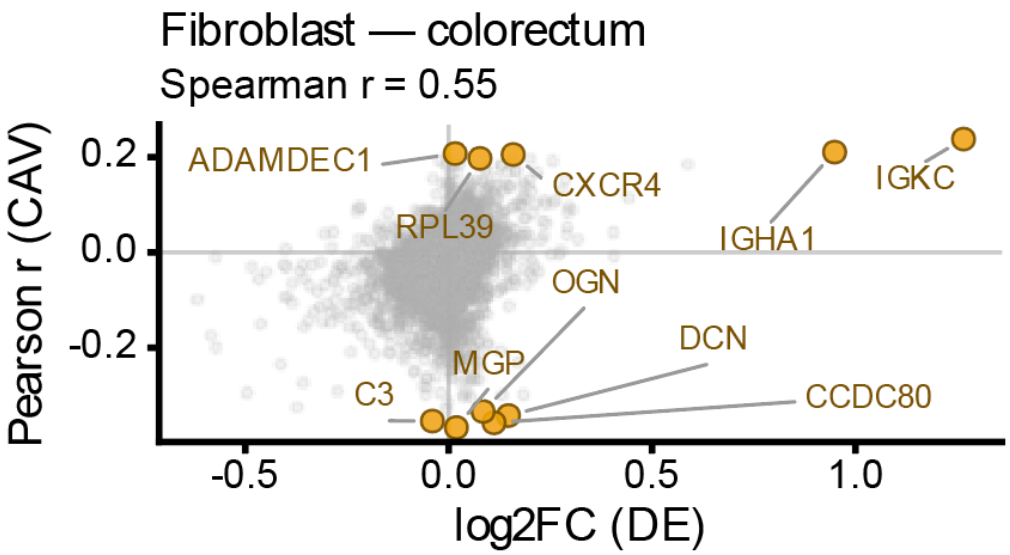
Cancer-axis correlated genes is quite similar to differential expression



Log2FC calculated with mixed-effect model

- No common algorithmic steps with DiffE
- Very fast
- Also captures low expression genes

Captures cell state differences that do not show up in differential expression



- Captures increase in immunogenic-state cells
  - ADAMDEC1, CXCR1

2. Discovering novel repeat motifs

There is a strong six-mer repeat in this sequence, that repeats six times,

It’s quite hard to see even
if you know the pattern

[G/P]-R-X-**S**-X-**[D/E]**

GNLGAGSSALPVEKAIGROLRDTQSEIAECY

GYGRRSCDTDPRFSIDAGRFSLDAGRVSVD

DPRYSFEEPRASWDGYLIGIRAAAAPMRMPSM

LSVVEDSRPVRNHVHRSDTHIPVEKSPQVSE

Using the repeat motif to focus structural modelling

[G/P]-R-X-S-X-[D/E]

GNLGAGSSALPVEKAIGROLRDTQSEIAECY

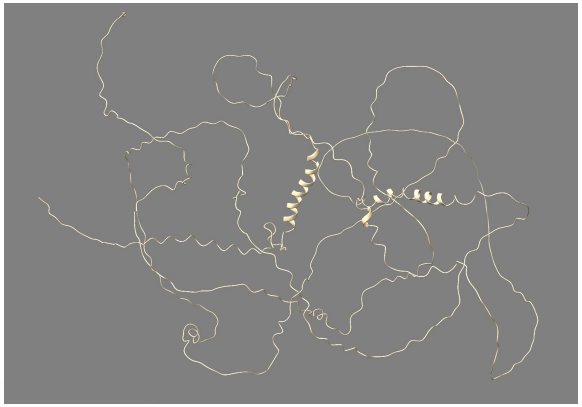
GYGRRSCDTDPRFSIDAGRFSLDAGRVSVD

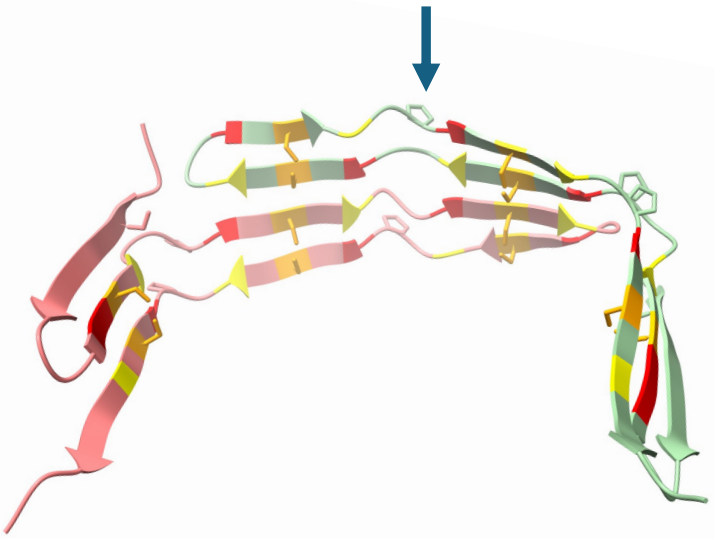
DPRYSFEEPRASWDGYLIGIRAAAAPMRMPSM

LSVVEDSRPVRNHVHRSDTHIPVEKSPQVSE

*folds as a dimer
*little serine bridges between repeats

Opl2 Arabidopsis

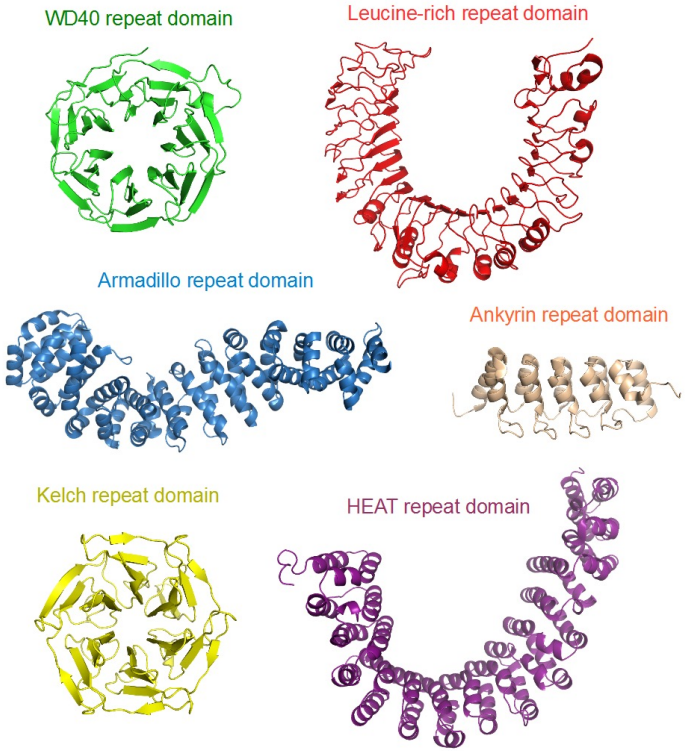




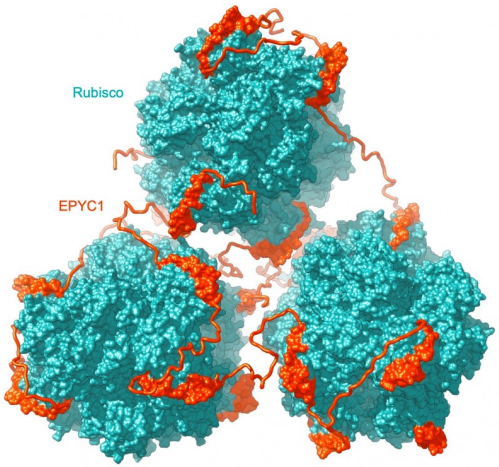
Protein repeats

Any motif ~3-30 amino acids long that occurs multiple times in a protein, more than expected by chance

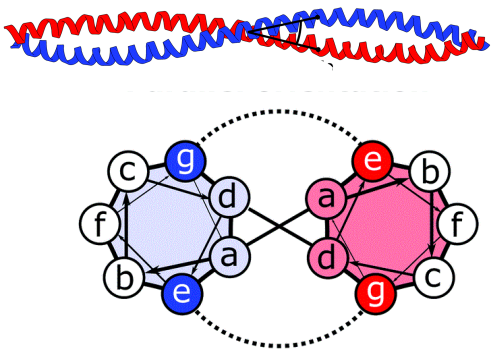
Structural tandem repeats



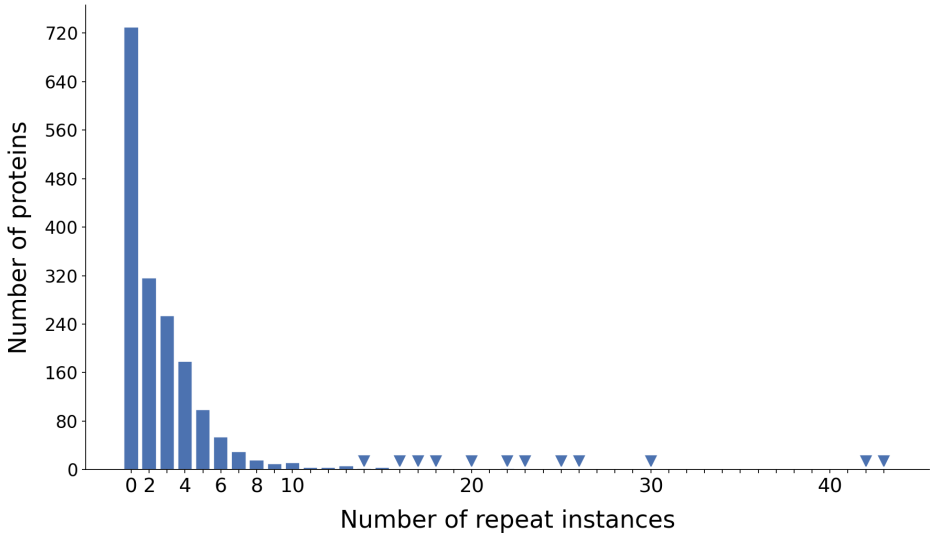
SLiMs

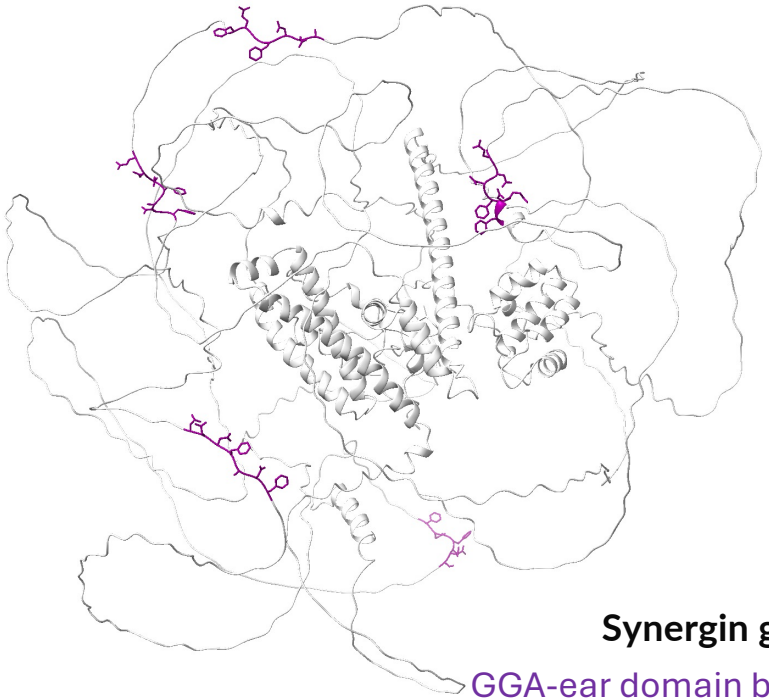


Coiled-coil heptads



- About 50% of human proteins have some repeat structure

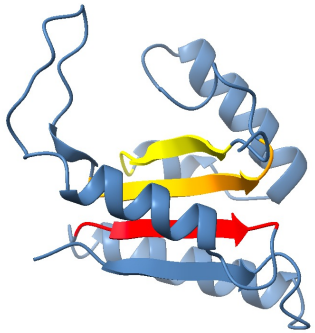




Synergin gamma

GGA-ear domain binding motif

[ADEQS]-D-D-F-x-[DE]-F



Amphipathic repeat motif

[PST]-[AVR]-Ψ-[LF]-[FL]-Ψ-[SP]-x

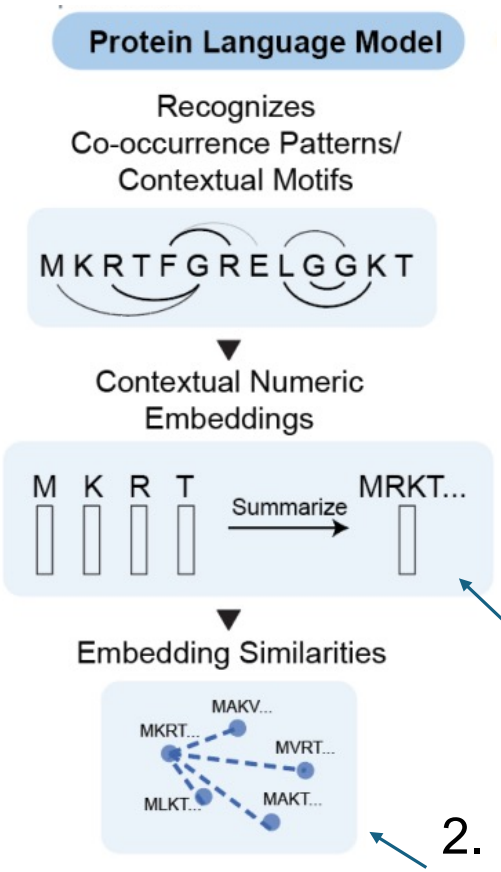
Thank you for listening!

mcwhitelab.org/tv

Project 3: Evaluating Pretrained Protein Language Model embeddings as proxies for functional similarity

Shaw, Love, and McWhite, 2025, Journal of Molecular Evolution

Do pretrained embeddings capture differences in molecular function?



- Many powerful PLM applications rely on additional training with labeled data
  - Structure prediction, protein design, drug–target prediction, antibody engineering, etc.

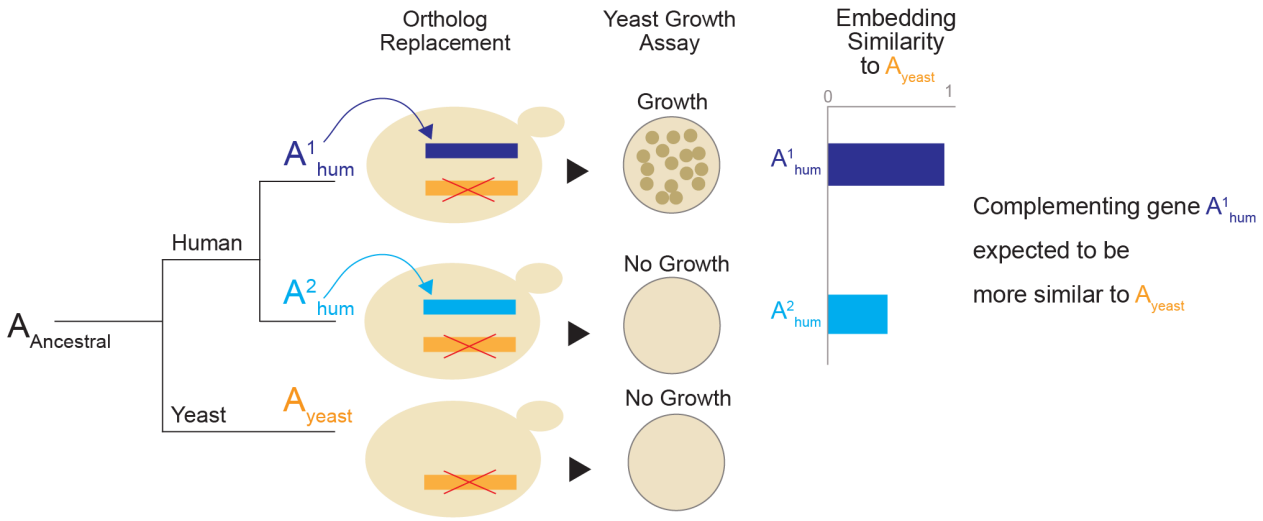
- Mismatch:
  - biological questions are often specific and focused, while supervised learning requires large labeled datasets

1. Can pretrained PLM embeddings represent protein molecular function directly, without task-specific training?

2. Does embedding distance track subtle functional divergence?

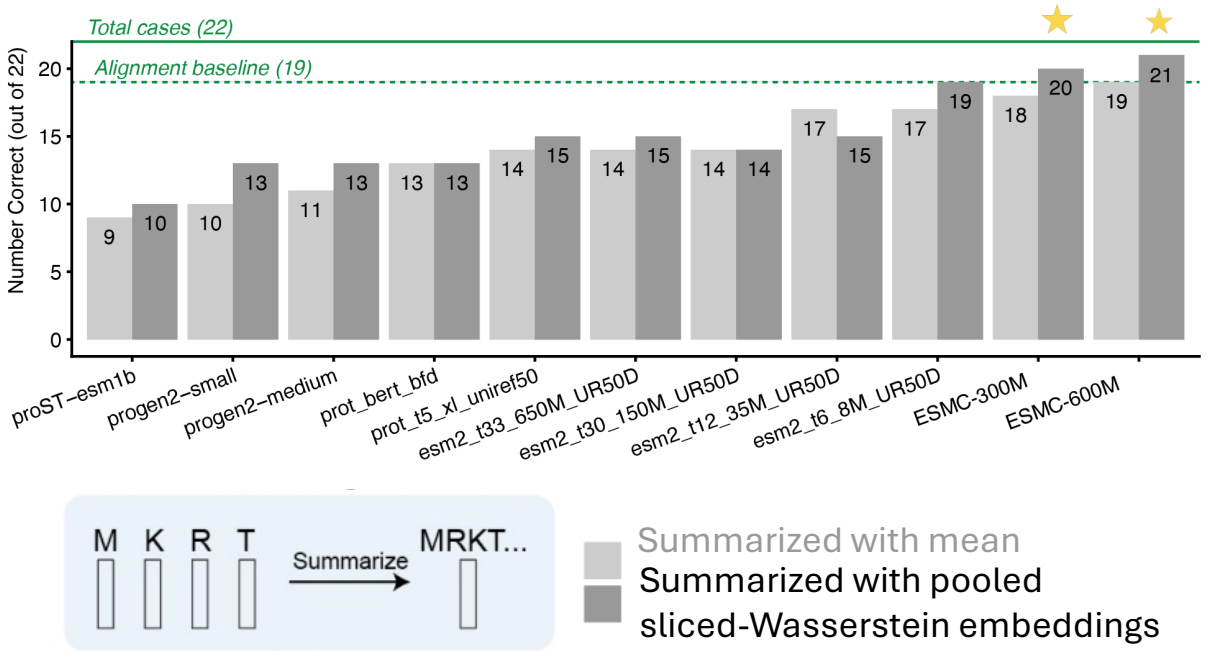
Cross-species complementation assays are a strict test of protein functional similarity

- “Remove an essential yeast gene, and see if a corresponding human gene allows the yeast to grow”



Kachroo .. McWhite.. et al. (2017), Laurent et al. (2020), and Garge et al. (2020),

Embeddings can be better than sequence similarity at prioritizing the complementing protein



Summarized with mean

Summarized with pooled sliced-Wasserstein embeddings

Shaw, Love, and McWhite, 2025, Journal of Molecular Evolution

Key ideas from *Shaw, Love & McWhite, 2025*

- Embedding similarity reflects structure/function more than ancestry
- Clearly laid out the model reasons why PLMs capture evolutionary outcomes, not evolutionary history
- Embedding distance is expected to diverge from phylogenetic distance at fine scales

Impact:
A recent Trends in Genetics perspective heavily cited and recapitulated these ideas as a framework for thinking about PLM embeddings in molecular evolution.

Rethinking molecular evolution through protein language model embeddings

Rosa Fernández [1] · Sergi Valverde [1] · Aureliano Bombarely [2] · Ildefonso Cases [3] · Scott A. Handley [4] · Ana M. Rojas [3]