Published on Feb 14, 2024 by Arcadia Science **DOI**: 10.57844/arcadia-a7cb-9f5c

Exploring the actin family: A case study for ProteinCartography

We've applied ProteinCartography, a tool for protein family exploration, to the well-studied actin family. We're able to categorize actins and related proteins into distinguishable functional buckets, and we uncovered some surprising hypotheses that could prompt further study.

Contributors (A-Z)

Prachee Avasthi, Brae M. Bigge, Feridun Mert Celebi, Megan L. Hochstrasser, Taylor Reiter, Dennis A. Sun, Ryan York

Version 3 · Mar 31, 2025

Purpose

We recently introduced ProteinCartography, a tool for interactively exploring protein families across species using protein structural comparisons. As an early use case, we chose to investigate a well-known protein family to test the ProteinCartography clustering approach while seeing if we could generate new insights. We selected actin, a cytoskeletal protein present in all eukaryotes and many prokaryotes, which is responsible for many cellular functions, including maintaining cell shape, cell division, cell motility, membrane dynamics, chromatin regulation, and others. Individual species

depend on their actin cytoskeleton to accomplish different cellular functions, so this protein family is a good test case for inferring structure-function relationships.

We found that ProteinCartography clustering separates actin proteins into subfamilies as expected, but we were also able to generate a list of novel observations and hypotheses about how these actin proteins are related structurally and perhaps functionally. We decided to follow up on one observation, that secondary actins from fungi sort into a unique subfamily based on structure **[1]**. We don't plan to follow up on the remaining hypotheses listed here and there may still be more insights hiding within the data. Therefore, if you find a hypothesis or piece of data particularly interesting or compelling, we encourage you to dig deeper.

- This pub is part of the **platform** effort, "Annotation: Mapping the functional landscape of protein families across biology." Visit the platform narrative for more background and context.
- **Data** from this pub, including the full ProteinCartography analysis and results for the actin family including structures, interactive maps, and all associated tables, can be found on Zenodo.
- All associated **code** is available in this GitHub repository.
- The **ProteinCartography pipeline** can be found in this GitHub repository.
- We used **previously generated data** from the 2022-actin-prediction GitHub repository.

Background and goals

We designed the ProteinCartography pipeline to explore protein families based on the structure of each individual protein **[2]**. Briefly, ProteinCartography starts by generating a list of protein structures using sequence- and structure-based searches. It then compares every structure to every other structure in the list, creating a similarity matrix for clustering and mapping. We used it to analyze a few dozen proteins in our initial pub, but we wondered if we could identify shared structural features related to known protein subfunctions and infer new properties of various subfamilies **[2]**.

We previously utilized the actin family for similar protein-based comparative analyses **[3]**. We selected actin because it's important for many distinct cellular functions, including maintaining cell shape, cell motility, cell division, intracellular trafficking, signaling, organellar regulation, membrane remodeling, and many others **[4]**. Additionally, it's been studied enough that we know quite a bit about its structure and function. For example, we know that actin monomers must associate and form long filaments for many of its broader functions in the cell, and we know that actin functions as an ATPase **[4]**. We even know the critical residues for these important functions from experimentally determined structures and biochemical studies **[5]**. Here, we investigated how conservation of these known residues is distributed across the family, and also asked if other unique structural features contribute to the division of the actin family into subfamilies. Our results point to novel structure-function relationships within this well-known family. To see what we found, skip straight to "The results" or continue reading to learn more about our methodology.

> **SHOW ME THE DATA**: Access our clustering data on Zenodo (DOI: 10.5281/zenodo.10641662)

The approach

Preparing the list of proteins

We used actin family proteins previously identified with our actin prediction pipeline **[3]**. In that analysis, we used human β-actin (UniProt ID: P60709) to perform a protein BLAST search against the full NCBI non-redundant (nr) database with no taxonomic restrictions, retrieving 50,000 proteins **[6][7]**, 26,994 of which were available on UniProt. We removed proteins from this list based on their annotation as "fragment" proteins or as "deleted" proteins and downloaded the remaining available proteins from the AlphaFold database **[8][9]**. This resulted in 14,665 actin family protein structures that we analyzed in this study.

You can find a full description of how we prepared this set of protein sequences in this GitHub repository.

Running ProteinCartography Cluster mode

Using the prepared protein structures as input, we ran the ProteinCartography pipeline using “Cluster mode” and the default parameters of the pipeline **[2]**. ProteinCartography’s “Cluster mode” starts with a prepared folder of protein structures. It uses Foldseek to compare proteins and create an all-v-all similarity matrix of TM-scores **[10][11]**. TM-scores are values that tell you how related two protein structures are on a scale of 0–1, where values closer to one are more similar **[12]**. The matrix is then used to cluster the proteins into groups of similar proteins using Leiden clustering **[13]** and to create interactive maps of the comparisons. Various data can be intersected with the maps to help deeply investigate the protein family.

This is outlined in detail in the ProteinCartography GitHub repository.

Combining results of ProteinCartography and the actin prediction pipeline

In addition to the standard overlays in ProteinCartography, users can apply custom overlays to the interactive maps. In this analysis, we used data from our prior work **[3]** to create a custom overlay to add to the plot. In particular, we overlaid the conservation of residues involved in actin polymerization and ATP binding. Additionally, we visualized and statistically analyzed the distributions of some of the metadata. This is outlined in three repositories:

- We describe how we calculated the conservation of important functional residues and built the initial list of proteins in this GitHub repo.
- We created the maps using ProteinCartography, which you can find in this GitHub repo.

- We outline how we prepared files, downloaded structures, and performed additional analyses in this GitHub repo.

Additional methods

We used GitHub Copilot to help write, clean up, and comment our code. We also used ChatGPT to help write some code. We validated all AI-generated code by running known datasets through it.

The results

Distinct clustering of high-quality protein structures allows for exploration of the actin

family

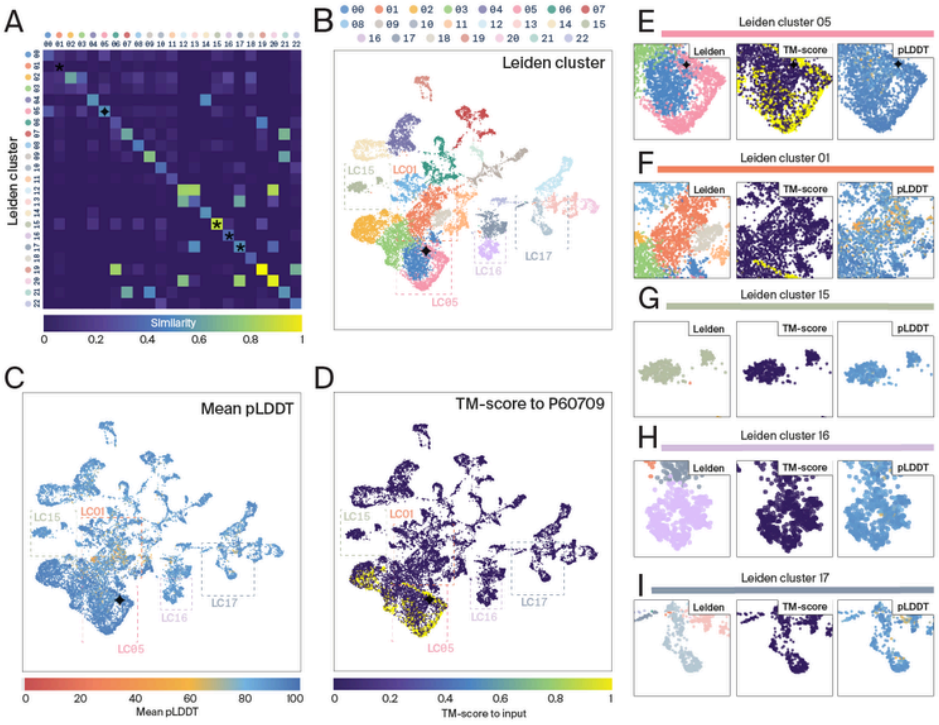

Figure 1

Clustering of the actin family produced well-defined clusters of high-quality proteins.

(A) Cross-cluster similarity matrix. The four-pointed star shows the cluster that contains the input. Other clusters highlighted in this figure are indicated with an asterisk.

(B) The UMAP of actin and related proteins showing Leiden clusters.

(C) The UMAP of actin and related proteins showing mean pLDDT of each protein.

(D) The UMAP of actin and related proteins showing the TM-score of each protein compared to the input.

(E) Zoomed-in version of LC05, where the input protein can be found (represented by the star).

We've also featured LC01 (F), LC15 (G), LC16 (H), and LC17 (I).

After running ProteinCartography, we see proteins sort into 23 (LC00–LC22) distinct clusters. Our input protein, human β-actin, is in cluster LC05 (Figure 1 and Figure 2). A full list of all the proteins in this analysis along with all the aggregated information from the pipeline can be found in the aggregated features file linked below.

tsv `actin_aggregated_features.tsv` Download

Figure 2

Interactive UMAP plot with metadata overlays.

This is the UMAP generated via ProteinCartography “cluster mode” for actin. Our input protein, human β-actin, is in cluster LC05 and is highlighted with a four-pointed star, which can be toggled on and off with the “Input Proteins” button. You can change the overlay using the

"color" drop-down menu. A static version of this graph is available in Figure 1, B.

Before diving into exploring the clusters, we first evaluate cluster quality using the cross-cluster similarity matrix (Figure 1, A and Figure 3). ProteinCartography calculates a mean TM-score for each cluster versus every other cluster and plots these scores in a visual matrix **[2]**. The diagonal of the matrix shows how similar the proteins within each cluster are to each other, a measure we refer to as "cluster compactness." For this analysis, we observe high cluster compactness, as visualized by higher average TM-score values along a clear diagonal, suggesting that the clusters are distinct (Figure 1, A and Figure 3). In particular, Leiden clusters 19 (LC19) and 20 (LC20) seem to be particularly compact (Figure 1, B and Figure 2). Clusters LC12 and LC15 seem to be compact as well, but to a lesser extent (Figure 1, B; Figure 2; and Figure 3). We can also see that cluster pair LC20/LC12 and cluster pair LC06/LC19 show a high mean TM-score, suggesting the clusters are structurally similar, leading us to hypothesize that they could share some functions (Figure 1, A and Figure 3). Of note, our input protein (human ß-actin) is present in LC05, which shares some similarity with LC00, LC01, and LC03 (Figure 1, A; Figure 2; and Figure 3).

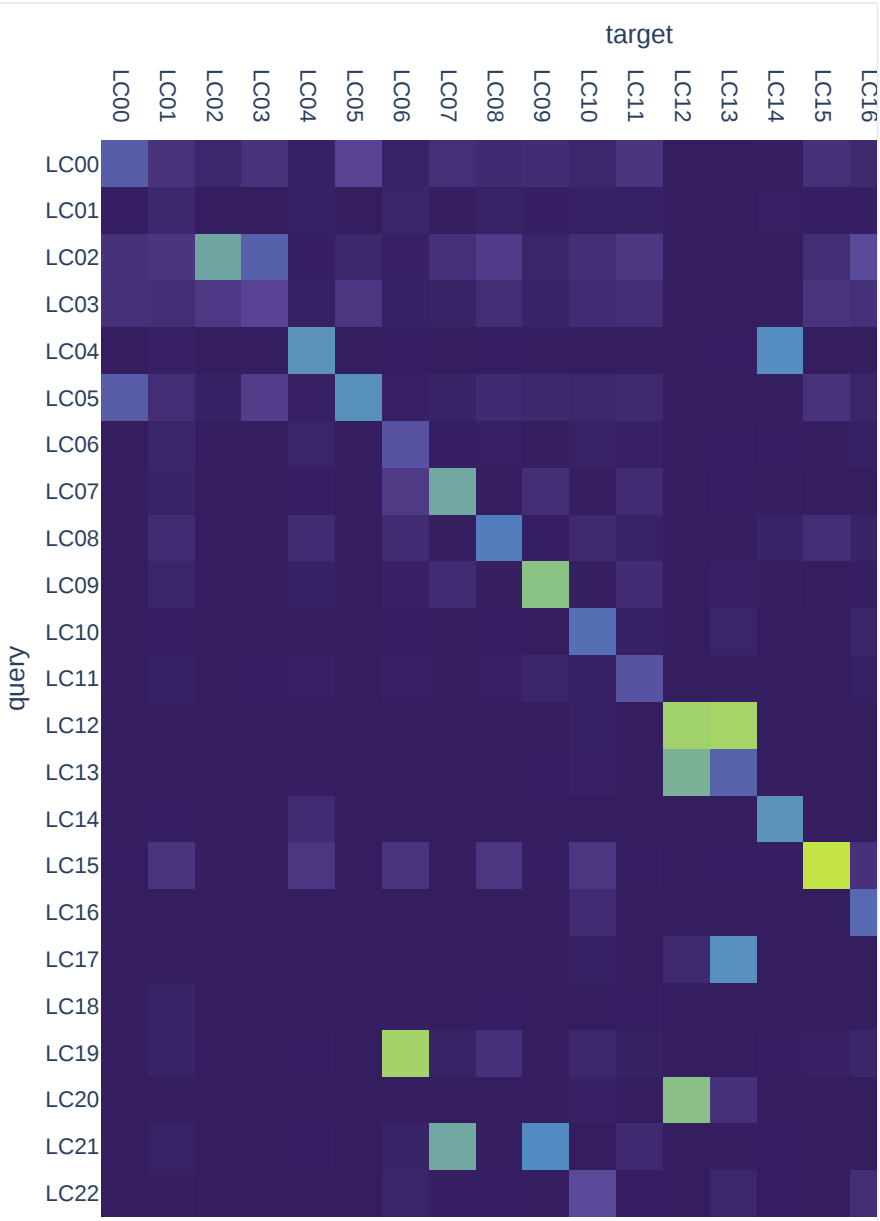

Figure 3

Cross-cluster similarity matrix for the actin family shows distinct and compact clusters.

The similarity matrix shows the mean TM-score of each structure within each cluster versus all other proteins in all other clusters. The diagonal line shows how similar proteins within a cluster are to each other, and thus how compact the clusters are. The input protein is in cluster LC05. You can view a static version of this matrix in Figure 1, A.

Looking at the clusters themselves and applying additional overlays can tell us about the quality of the protein structures within the space (Figure 1, B and Figure 2). The "pLDDT" (predicted local distance difference test) is a per residue score that tells us about the quality of the structure prediction **[14]**. The pipeline averages this value across each predicted structure and displays it as an overlay. Cluster LC05, which contains the input protein, has a high pLDDT (mean is 93.6%), with all other clusters having statistically lower average pLDDT scores (Figure 1, C; Figure 1, E; Figure 2; and Figure 4, A). AlphaFold suggests guidelines for interpreting these scores — pLDDT scores greater than 90 suggest high confidence, and scores between 70 and 90 are modeled well. All the cluster averages fall within this range (Figure 4, A), but there are a handful of proteins that appear to have lower-quality predictions looking at the pLDDT overlay (Figure 1, C; Figure 1, F; Figure 2; and Figure 4, A). Overall, however, most proteins in this analysis are well-structured and well-predicted according to the pLDDT scores.

Knowing that the analysis yielded compact and distinct clusters of high-quality protein structures, we can now dive deeper into what the analysis might tell us about the structure-function relationship throughout this protein family.

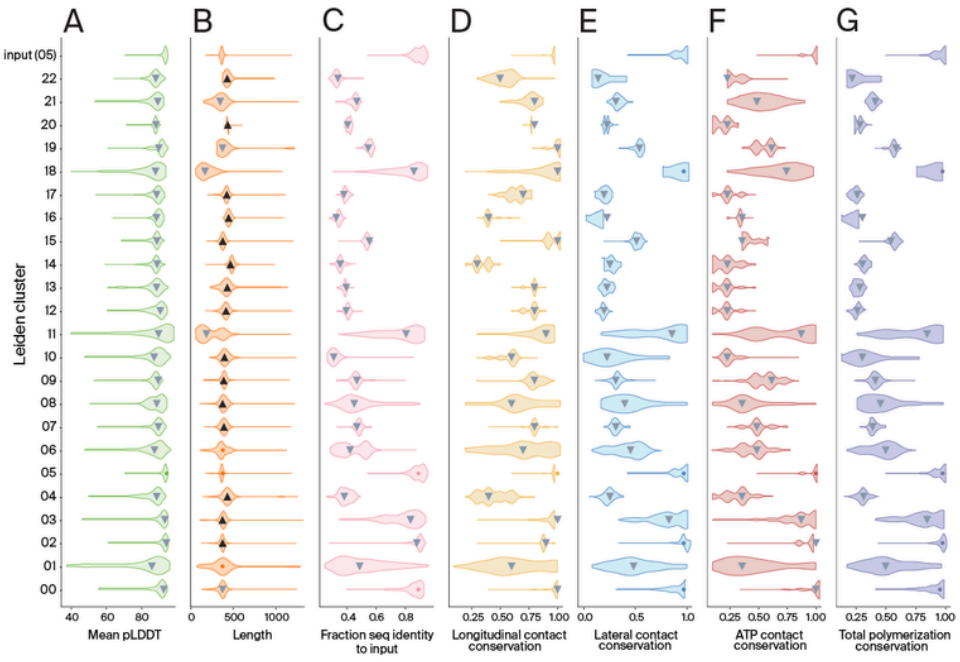

Figure 4

Distributions of metadata.

We've plotted the distribution of mean pLDDT (A), length (B), fraction sequence identity compared to input (C), and important functional residue conservation for each cluster (D–G). We performed a Mann–Whitney U test to statistically compare each cluster to the cluster containing the input protein (LC05). Gray arrows pointing down represent clusters where the distribution is statistically lower than the input cluster, and black arrows pointing up represent clusters where the distribution is statistically higher than the input cluster. The placement of either an arrow or a dot shows the mean of each distribution.

Taxonomy, length, and annotation score overlays provide functional insights

To help us understand more about the proteins in this analysis, we applied an annotation score overlay. The annotation score overlay tells us how confident the existing UniProt annotations are. UniProt annotations with a score of five mean the

protein annotations are experimentally determined, while a score of one is a lower-confidence annotation. In this analysis, 13,668 proteins out of the total 14,608 (about 94%) have annotation scores of one, while only 64 proteins have annotation scores of five. 22 of those proteins (34%) were grouped in LC05, where our input protein is found, while only 6% of all proteins were in LC05 and no other cluster had as many proteins with annotation scores of five (Figure 2; Figure 5 B; and Figure 5 E). Therefore, despite this being a well-studied family, the majority of protein annotations are low-confidence predictions, leaving plenty of room for discovery.

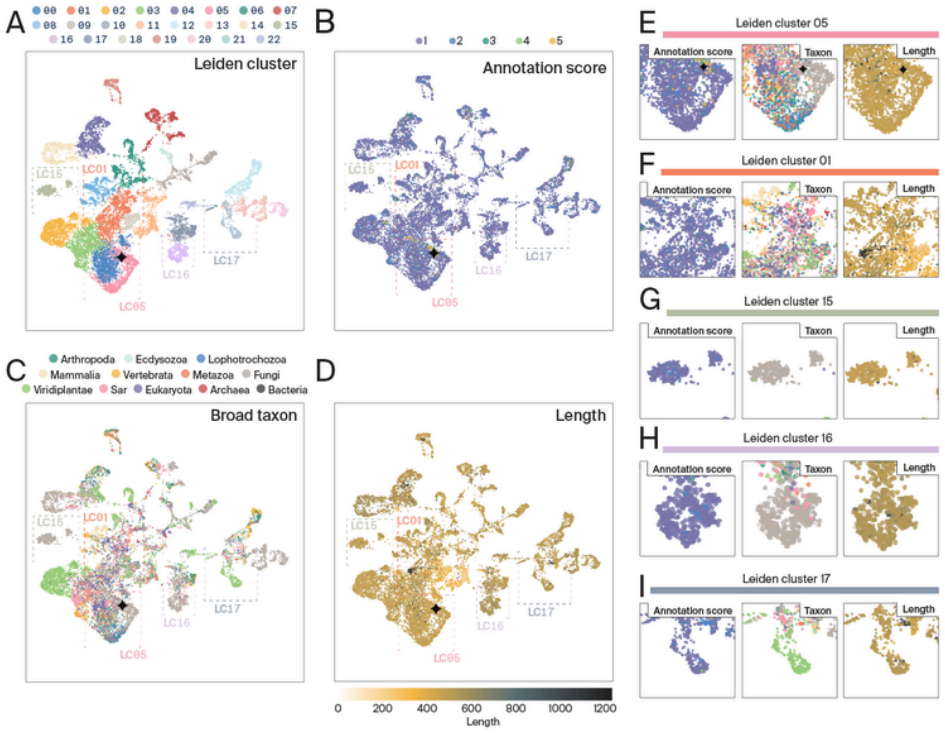

Figure 5

Overlays indicate patterns that could drive clustering.

(A) The UMAP of actin and related proteins showing Leiden clusters.

(B) The UMAP of actin and related proteins showing the UniProt annotation score of each protein.

(C) The UMAP of actin and related proteins showing the broad taxon into which each protein was sorted.

(D) The UMAP of actin and related proteins showing the length.

(E) Zoomed-in version of LC05 where the input protein can be found (represented by the star).

We also featured LC01 (F), LC15 (G), LC16 (H), and LC17 (I).

We can apply additional overlays, like the broad taxonomy of the proteins (Figure 2 and Figure 5, B). We assigned the broad taxa into which ProteinCartography sorts proteins to make interpreting the plots easier, but it's important to note that the taxonomic

depth isn't uniform. We know that actin is widely expressed, particularly in eukaryotes, and we see that the taxonomic origins of actin proteins in the map are generally quite mixed. We mostly see eukaryotes, but there is a smattering of bacterial and archaeal proteins throughout the space.

While the overall map is taxonomically mixed, some regions are primarily composed of organisms from a single broad taxonomic group. For example, clusters LC13, LC14, LC15, LC16 and LC20 contain mostly fungal species (Figure 2; Figure 5, C; and Figure 5, G). Meanwhile, clusters LC02, LC17, and LC22 contain mostly proteins from plants (Figure 2; Figure 5, C; and Figure 5, I). We expect structures to vary with phylogeny, but we're interested in better understanding the clustering differences we see between taxonomic groups and how these might relate to protein structure and function. For example, the clustering we see in Figure 5, C suggests there may be specific structural/functional features that separate fungal and plant proteins from other similar proteins, so this could be a useful place for us to start.

In addition to looking at the taxonomy of proteins throughout the space, we can also look at features of the proteins, like length. We filtered out proteins annotated as fragments on UniProt in our analysis, but because we know the length of most actins — 375 amino acids — we can use the length to tell us when proteins are surprisingly long or short compared to what we consider "normal" **[4][7]**. Some proteins are much longer than the conserved 375 amino acids (up to 1279 residues), and clusters LC11 and LC18 contain many proteins that are shorter than 375 amino acids (with the shortest just 60 amino acids long) (Figure 2; Figure 4, B; Figure 5, D; and Figure 5, F). To perform cellular functions, individual actin molecules bind together to form filaments with structures that are generally quite well-conserved across isoforms and species. Differences in monomer length could affect the structure and dynamics of those filaments in "true" actins. Additionally, because this analysis contains actin-related proteins in addition to true actins, these long or short proteins could have totally distinct functions, which we'll discuss more in the next section.

Existing annotations align with ProteinCartography clustering

Because the actin family is extensively studied, most proteins are assigned some annotation in UniProt. The actin family is composed of several smaller subfamilies,

including “actins,” but also several distinct “actin-related proteins” and some “actin-like proteins.” Together, this means that we can use these annotations to determine how well our clustering aligns with existing information and to generate hypotheses about why certain proteins are clustered together. In general, we found that the existing annotations do align well with the clustering results. However, we also found evidence that existing annotations are not always reliable.

We see several clusters, including LC00, LC01, LC02, LC03, and LC05, that contain mostly proteins annotated as “actin” or a similar variation (Figure 6). This group of clusters contains almost all primary eukaryotic actins, including our input protein, human β-actin, which is in LC05. Surprisingly, *Giardia* actin, the most divergent actin currently studied with a sequence identity of less than 60% compared to human actin, can be found in LC03 along with many conventional actins (Figure 6) **[15][16]**. Despite the sequence divergence, this unusual actin shares enough of its structure with other actins to still be considered an actin, and therefore, likely has similar functions to “normal” actins. However, it is interesting that the “true” actins are sorted into separate clusters, as opposed to a single, large cluster. Evaluating some of the underlying structural and functional differences that result in this separation could lead to a better understanding of actin evolution across species.

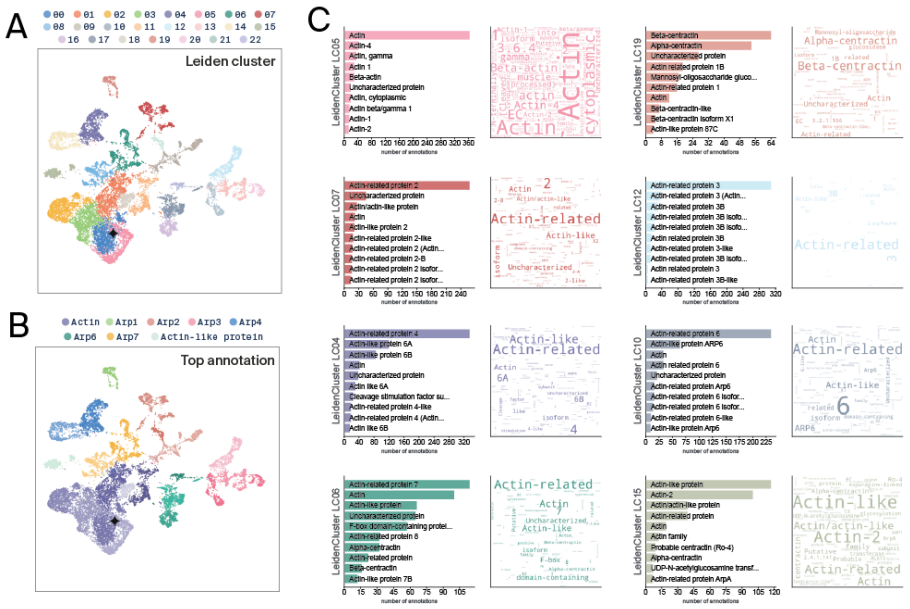

Figure 6

Semantic analysis of the actin family shows that protein clustering generally agrees with existing annotations but that existing annotations are not always reliable.

(A) The map of the actin family showing Leiden clusters.

(B) The map of actin proteins with the top annotation from each cluster represented in the color overlay. Different hues of each color are to help discriminate between individual clusters.

C) Colors correspond to Leiden cluster color. Each cluster has a ranked bar chart that shows its top ten full annotation strings, as well as a word cloud that shows top annotation words.

LC19 is composed of “centractins,” “actin-related protein 1,” or “Arp1” (Figure 6). Arp1 is a subunit of the dynactin complex, which binds microtubules and dynein **[17]**. Dynactin, and thus Arp1, is involved in intracellular transport, nuclear positioning, and chromosome movement. In the dynactin complex, Arp1 forms a short filament that never achieves the length of conventional actin filaments **[18]**. Within this single Arp1 cluster, there are many proteins annotated as “hypothetical proteins” or “proteins of unknown function.” We hypothesize that these unannotated proteins could be Arp1s.

There are also Arp1 proteins in LC06 and LC14 that could have differing functions from those in LC19.

We also see a few distinct clusters — LC07, LC09, and LC21 — annotated as “actin-related protein 2” or “Arp2” (Figure 6). Similarly, we see distinct clusters, LC12, LC13, LC17, and LC20, annotated as “actin-related protein 3” or “Arp3” (Figure 6). Arp2 and Arp3 are members of the Arp2/3 complex, which binds to primary actin filaments and nucleates new filaments in a characteristic branch **[19]**. Arp2 and Arp3 serve as the first monomers in the new actin filaments, but are unable to form stable filaments on their own **[20][21]**. Interestingly, both Arp2 and Arp3 subfamilies are broken up into clusters with a cluster or two that are primarily composed of proteins from a single taxonomic group, while other clusters are more homogenous. In the case of Arp2, LC09 and LC21 contain more fungal proteins, while the majority of LC07 consists of metazoan proteins (Figure 5, C and Figure 6). For Arp3, this distinction is even more apparent — LC20 contains mostly fungal proteins, LC13 contains mostly metazoan proteins, and LC17 contains primarily plant proteins (Figure 5, C; Figure 5, I; and Figure 6). One could investigate why these subfamilies are broken up into smaller clusters. Could there be structural or functional differences between the Arp2s or Arp3s in specific clusters? One could also use this dataset to investigate co-evolution of proteins that function together. Specifically, Arp2 and Arp3 form a complex that binds actin. Using the information from this analysis, one could determine whether these two proteins evolved together by checking whether Arp2 and Arp3 from individual species cluster together.

LC04 and LC14 are annotated as “actin-related protein 4” or “Arp4” and LC10, LC16, and LC22 are annotated as “actin-related protein 6” or “Arp6.” Arp4 and Arp6 are most known for their nuclear roles. Arp4 prevents polymerization of actin within the nucleus and regulates gene expression, among other things **[22][23][24]**). Arp6 is involved in maintenance of the nucleolus and regulation of transcription **[25]**. As with Arp2 and Arp3, we also see these subfamilies broken up into clusters where LC04 and LC10 are quite mixed but LC14 and LC16 are composed mostly of proteins from fungi and LC22 is almost completely proteins from plants (Figure 5, C and Figure 6). The pipeline could be separating proteins based on structural differences that one might expect based on phylogeny, but further investigation could determine if these differences are explained by phylogenetic differences alone or if Arp4 and Arp6 proteins from different taxa have more extreme structural differences that lead to differences in overall function.

Finally, there are “actin-like proteins” in many clusters, but we found one cluster composed primarily of actin-like proteins or secondary actins (“Actin-2,” “Actin II,” etc.) — LC15 (Figure 6). We found this particularly interesting because of LC15's very clean separation from clusters of actins, with an average within-cluster TM-score of about 0.86, strongly suggesting the structures in this cluster are unique, which could indicate a functional distinction between these actin-like proteins and true actins (Figure 6). Looking more closely, LC15 is nearly 100% fungal proteins (Figure 5, C and Figure 5, G). To our knowledge, these secondary actins or actin-like proteins from fungi have not previously been explored and further analysis could provide insight into actin biology and fungal biology.

In summary, we found that all clusters contained a highly represented annotation that corresponds to a specific actin subfamily. Additionally, we found that most actin subfamilies were represented in only a handful of clusters. This suggests that ProteinCartography clustering can distinguish specific subfamilies, which in this protein family have differing functions.

However, despite the usefulness of the existing annotations for checking the reliability of our clustering approach, we also found that existing annotations are themselves not always reliable. Clusters have distinct annotations that are more highly represented than other annotations, but nearly every cluster still contains proteins annotated as simply “actin” or some variation. This suggests that “actin” is often applied as an annotation for proteins across this large, functionally diverse family. Additionally, we saw proteins annotated as “actin family” or “actin-like” across many clusters, and even in this well-known family, there are many proteins annotated as “uncharacterized protein” or “hypothetical protein.” This suggests that this type of analysis could be broadly useful for generating hypotheses about protein families, even those that are already very well-studied.

Mapping conservation of important functional residues helps identify novel drivers of clustering

To gain more information about the space, ProteinCartography lets users create custom overlays using metadata gathered from elsewhere or analyses performed outside the pipeline. For actin, we previously ran an analysis to determine the

conservation of residues involved in actin's key biochemical functions, which are essential for the protein to perform its many roles in the cell **[4][3][5]**.

Actin can exist in two primary states: single actin monomers (monomeric actin) or actin monomers strung together into filaments (filamentous actin). The process whereby actin switches between these states is termed polymerization or depolymerization, and is essential for actin's many functions. Actin polymerizes when monomers bind to each other through lateral contacts and longitudinal contacts **[5]**. This process also requires that actin bind and hydrolyze ATP. Because actin is so well-studied, the specific regions of the protein involved in lateral contacts, longitudinal contacts, and ATP binding are known. Therefore, we looked for the conservation of these residues in each of our proteins, as we previously detailed **[3]**.

We used the information generated from that analysis to create a custom overlay so we could observe how the conservation of those important residues is distributed across the actin maps generated by ProteinCartography (Figure 7).

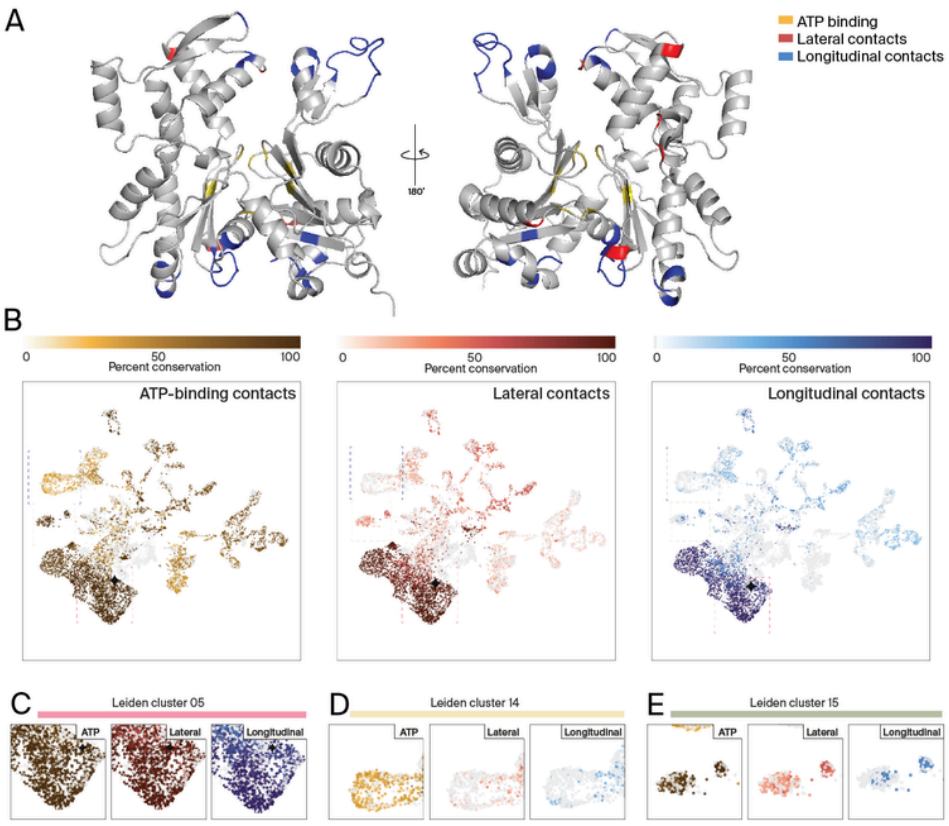

Figure 7

Overlaying additional information on the map can help inform functional predictions.

(A) Structure of actin with the regions required for ATP binding and polymerization (lateral and longitudinal) contacts highlighted visualized using open-source PyMOL.

(B) Maps of the actin family showing where ATP-binding residues and polymerization contacts are more or less conserved.

(C) LC05, which contains the input protein, has high ATP-binding residue and polymerization contact conservation.

(D) LC14 has low ATP-binding residue and polymerization contact conservation.

(E) LC15 has high ATP-binding residue conservation and low polymerization contact conservation.

We found that overall, the clusters that contain “true” actins have higher conservation of ATP-binding residues, lateral contacts, and longitudinal contacts than the rest of the space (Figure 7, B). Additionally, ATP-binding residues seem to be much more conserved throughout than lateral or longitudinal contact sites (Figure 7, B). This is expected based on the biology, as even actin-related proteins can bind ATP. However, clusters LC04 and LC14, which primarily contain Arp4s, and LC10, LC16, and LC22, which primarily contain Arp6s appear to have a less conserved ATP-binding site (Figure 7, B and Figure 7, D). This could suggest that their ATP-binding kinetics or specific functions differ from “true actins.”

When we look closer at particular clusters — for example, LC15, which contains primarily secondary actins from fungi — we see that while ATP-binding contacts seem to be fairly well-conserved, both lateral and longitudinal contacts are not (Figure 7, B and Figure 7, E). This could further suggest functional differences between these secondary actins and the conventional actins in LC05. Specifically, we could hypothesize that proteins in this cluster are able to bind ATP but have distinct polymerization kinetics.

> All **code** generated and used for the pub is available in this GitHub repository (DOI: 10.5281/zenodo.10642492), including notebooks used for preparing data, obtaining structures, and plotting the data.

ProteinCartography analysis of the actin family generates novel hypotheses

While using actin to learn about ProteinCartography, we generated a number of interesting hypotheses about the actin family itself. For example, we were particularly interested in LC15, the cluster of secondary actins or actin-like proteins that contains almost exclusively proteins from fungi. Because this group of proteins had not been previously characterized, we decided to dive deeper into this analysis in a separate pilot study. In that pilot, we tried to elucidate the functions of these secondary fungal actins using additional ProteinCartography analysis along with phylogenetic analysis and trait mapping **[1]**.

In addition, we have several hypotheses that we don't plan to investigate. We'd love for readers to dive deeper into these hypotheses and any others that they find in this data if they're interested:

- First, there are many uncharacterized proteins throughout the clusters that researchers interested in the cytoskeletons of their organisms of interest might want to study. This type of analysis could even potentially be used for annotation transfer via structural comparison. For example, LC19, which contains primarily Arp1s or centractins, contains 27 uncharacterized proteins (about 10% of the total cluster). Perhaps some of these uncharacterized proteins have Arp1-related functions. Similar hypotheses could be drawn for each cluster.
- How does the overall variation in length throughout the proteins in the map correspond to functional differences?
- What are the factors in our "true" actin clusters that are causing them to sort into a handful of different clusters? The actins in LC00, LC01, LC02, LC03, and LC05 are primarily annotated as "actin," and we found that most actins we looked for, including the very divergent *Giardia* actin, sorted into one of these four clusters. It could be interesting to see if there are more subtle differences between actins that are causing the further division of true actins into these clusters. Additionally, it could be interesting to look at the consequences of those more subtle differences — are there differences in actin dynamics or actin-binding proteins in these species?
- The division of Arp2s and Arp3s into distinct clusters could suggest functional differences between the branched actin nucleating Arp2/3 complexes of certain species. Specifically, we see that fungal proteins seem to cluster separately from others — could fungi have Arp2/3 complexes that perform functions different from other Arp2/3 complexes? One way to approach this could involve investigating the structural relationships of the other members of the Arp2/3 complex. As mentioned above, this could also provide an opportunity to study co-evolution — do Arp2s and Arp3s similarly cluster or not?
- Similarly, we see division of Arp4s and Arp6s into distinct clusters within each subfamily. Could there be functional differences between these clusters that cause them to fall into different clusters based on their structures? Additionally, these clusters have lower conservation of the residues involved in ATP binding. Could Arp4s and Arp6s have different specific functions or dynamics of ATP binding?

- LC06 and LC08 contain proteins annotated as “actin-related protein 7” or “Arp7” and “actin-related protein 8” or “Arp8.” This cluster also contains many proteins annotated as “F-box domain-containing protein.” This is a unique domain that has mostly only been investigated in plants **[26]**. The top taxon in this cluster is plants, but there are other taxonomic groups represented as well. It could be interesting to investigate whether this domain is more widespread and what function it serves outside of plants.
- Finally, while we’ve thoroughly investigated the conservation of residues that are known to be important for actin function, there is still room to determine if there are other important conserved residues that help make an actin an actin.

If you use these data, please let us know in a comment on this pub! We’d love to know how you’re using hypotheses generated via the ProteinCartography pipeline.

Key takeaways

In addition to the generation of hypotheses related to the actin family, this analysis also helped us better understand the pipeline’s performance. Based on these results, we plan to continue using this family as a standard dataset for future development and testing of the pipeline. We found:

- ProteinCartography was able to separate primary actins from actin-related and actin-like proteins. Specifically, we found clusters for actin, ARP1, ARP2, ARP3, ARP4, and ARP6.
- Existing annotations are not always reliable, as nearly every cluster contained some variation of “actin” as a common annotation.
- ProteinCartography is useful for generating new hypotheses even for this well-known protein family.

In the future, we want to add tools to the ProteinCartography pipeline that automate some of the analyses used in this pub, such as highlighting important functional residue conservation across the space and providing distributions of protein features across clusters, as in Figure 4.

Please let us know if you apply the ProteinCartography pipeline for other uses. The pipeline is still in development, and we’re actively looking for ways to improve, so any

feedback is welcome and appreciated.

References

1. Avasthi P, Bigge BM, Kolb I, Mets DG, Morin M, Patton AH, York R. (2024). A structurally divergent actin conserved in fungi has no association with specific traits. https://doi.org/10.57844/ARCADIA-9768-F6C5

2. Avasthi P, Bigge BM, Celebi FM, Cheveralls K, Gehring J, McGeever E, Mishne G, Radkov A, Sun DA. (2024). ProteinCartography: Comparing proteins with structure-based maps for interactive exploration. https://doi.org/10.57844/ARCADIA-A5A6-1068

3. Avasthi P, Bigge BM, Reiter T. (2024). Defining actin: Combining sequence, structure, and functional analysis to propose useful boundaries. https://doi.org/10.57844/ARCADIA-YNTH-KH70

4. Dominguez R, Holmes KC. (2011). Actin Structure and Function. https://doi.org/10.1146/annurev-biophys-042910-155359

5. Chou SZ, Pollard TD. (2019). Mechanism of actin polymerization revealed by cryo-EM structures of actin filaments with three different bound nucleotides. https://doi.org/10.1073/pnas.1807028115

6. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. (2009). BLAST+: architecture and applications. https://doi.org/10.1186/1471-2105-10-421

7. Consortium TU. (n.d.). UniProt: the Universal Protein Knowledgebase in 2023. https://doi.org/10.1093/nar/gkac1052

8. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. (2021). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. https://doi.org/10.1093/nar/gkab1061

9 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. https://doi.org/10.1038/s41586-021-03819-2

10 Barrio-Hernandez I, Yeo J, Jänes J, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M. (2023). Clustering predicted structures at the scale of the known protein universe. https://doi.org/10.1101/2023.03.09.531927

11 van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. (2022). Fast and accurate protein structure search with Foldseek. https://doi.org/10.1101/2022.02.07.479398

12 Zhang Y, Skolnick J. (2004). Scoring function for automated assessment of protein structure template quality. https://doi.org/10.1002/prot.20264

13 Traag VA, Waltman L, van Eck NJ. (2019). From Louvain to Leiden: guaranteeing well-connected communities. https://doi.org/10.1038/s41598-019-41695-z

14 Mariani V, Biasini M, Barbato A, Schwede T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. https://doi.org/10.1093/bioinformatics/btt473

15 Steele-Ogus MC, Johnson RS, MacCoss MJ, Paredez AR. (2021). Identification of Actin Filament-Associated Proteins in Giardia lamblia. https://doi.org/10.1128/spectrum.00558-21

16 Paredez AR, Nayeri A, Xu JW, Krtková J, Cande WZ. (2014). Identification of Obscure yet Conserved Actin-Associated Proteins in Giardia lamblia. https://doi.org/10.1128/ec.00041-14

17 Urnavicius L, Zhang K, Diamant AG, Motz C, Schlager MA, Yu M, Patel NA, Robinson CV, Carter AP. (2015). The structure of the dynactin complex and its interaction with dynein. https://doi.org/10.1126/science.aaa4080

18 Bingham JB, Schroer TA. (1999). Self-regulated polymerization of the actin-related protein Arp1. https://doi.org/10.1016/s0960-9822(99)80095-5

19 Goley ED, Welch MD. (2006). The ARP2/3 complex: an actin nucleator comes of age. https://doi.org/10.1038/nrm2026

20 Robinson RC, Turbedsky K, Kaiser DA, Marchand J-B, Higgs HN, Choe S, Pollard TD. (2001). Crystal Structure of Arp2/3 Complex.

https://doi.org/10.1126/science.1066333

21 Gournier H, Goley ED, Niederstrasser H, Trinh T, Welch MD. (2001). Reconstitution of Human Arp2/3 Complex Reveals Critical Roles of Individual Subunits in Complex Structure and Activity. https://doi.org/10.1016/s1097-2765(01)00393-8

22 Nie W-F, Wang J. (2021). Actin-Related Protein 4 Interacts with PIE1 and Regulates Gene Expression in Arabidopsis. https://doi.org/10.3390/genes12040520

23 Mołoń M, Stępień K, Kielar P, Vasileva B, Lozanska B, Staneva D, Ivanov P, Kula-Maximenko M, Molestak E, Tchórzewski M, Miloshev G, Georgieva M. (2022). Actin-Related Protein 4 and Linker Histone Sustain Yeast Replicative Ageing. https://doi.org/10.3390/cells11172754

24 Yamazaki S, Gerhold C, Yamamoto K, Ueno Y, Grosse R, Miyamoto K, Harata M. (2020). The Actin-Family Protein Arp4 Is a Novel Suppressor for the Formation and Functions of Nuclear F-Actin. https://doi.org/10.3390/cells9030758

25 Kitamura H, Matsumori H, Kalendova A, Hozak P, Goldberg IG, Nakao M, Saitoh N, Harata M. (2015). The actin family protein ARP6 contributes to the structure and the function of the nucleolus. https://doi.org/10.1016/j.bbrc.2015.07.005

26 Kandasamy MK, McKinney EC, Meagher RB. (2008). ACTIN-RELATED PROTEIN8 Encodes an F-Box Protein Localized to the Nucleolus in Arabidopsis. https://doi.org/10.1093/pcp/pcn053