Published on May 30, 2024 by Arcadia Science **DOI**: 10.57844/arcadia-cae9-96c4

A strategy to validate protein function predictions *in vitro*

We aim to validate ProteinCartography, a tool for structure-based protein clustering, by evaluating two foundational hypotheses: that proteins within a cluster have similar functions and proteins in different clusters have differing functions.

Contributors (A-Z)

Prachee Avasthi, Audrey Bell, Brae M. Bigge, Megan L. Hochstrasser, Atanas Radkov , Dennis A. Sun, Harper Wood, Ryan York

Version 1 · Mar 31, 2025

Purpose

In this pub, we outline a path for validating ProteinCartography, a computational tool for comparative analysis of protein structures across species **[1]**. ProteinCartography produces an interactive map of protein families with individual proteins separated into clusters based on their structural similarity. Our foundational hypotheses are that functionally similar proteins cluster together while proteins with distinct functions cluster separately. We plan to assess this using a couple of test protein families.

We've selected protein families for *in vitro* validation, and that's mostly what we'll focus on in this pub. We started with a list of the most common human proteins in the Protein

Data Bank **[2]** and narrowed it down using criteria outlined below. We selected two candidate protein families, Ras GTPase and deoxycytidine kinase.

Now we face the challenge of selecting individual clusters and proteins to focus on. We go into much more depth on how we're thinking about this for each family in our accompanying dCK and Ras GTPase pubs. Head there for specific information (and to provide family-specific feedback!). We'll update this pub with generalizable takeaways from our studies of each protein family to build a roadmap for validation.

- This pub is part of the **platform effort**, "Functional annotation: mapping the functional landscape of proteins across biology." Visit the platform narrative for more background and context.
- The accompanying pubs, "How can we biochemically validate ProteinCartography with the deoxycytidine kinase family?" and "How can we biochemically validate ProteinCartography with the Ras GTPase family?", present ProteinCartography data and follow-up testing options for our two chosen protein families.
- The **ProteinCartography pipeline** used to run these analyses is available in this GitHub repo.
- The **data** associated with this pub, including ProteinCartography results for the 30 proteins we ran, can be found in this Zenodo repository. An additional four from previous ProteinCartography runs can be found in this Zenodo repo.

Motivation

What is ProteinCartography?

We previously introduced a tool for structural comparison of protein families: ProteinCartography **[2]**. ProteinCartography identifies proteins similar to an input using sequence- and structure-based searches. It aligns the structure of each protein to every other protein to generate a structural similarity score, or TM-score (template modeling score), for each pair of proteins in the analysis **[3]**. It uses these scores to

populate a similarity matrix. It then uses this matrix to cluster proteins into similar groups and to create interactive maps (UMAP or t-SNE) for easy visualization **[4][5][6]**.

The outputs of this analysis can be useful for making predictions about which proteins within families might be structurally similar or identifying which proteins might have novel structural features. Because structure and function are closely related, we hope that this analysis will also let us generate hypotheses about protein function.

Our foundational hypotheses

As we use ProteinCartography's results to infer functional relationships, we want to biochemically validate ProteinCartography to show that the structure-based clustering can really give insights into protein function. To this end, we have two main hypotheses to test (Figure 1):

1 — Proteins within the same cluster have similar biochemical functions.

2 — Proteins in different clusters have functional differences.

We plan to test these hypotheses using candidate protein families that we can assess biochemically. For our first round of validation, we're aiming for a couple protein families that are easy to work with *in vitro* and that produce ProteinCartography results with clearly defined clusters that present many opportunities to test our hypotheses.

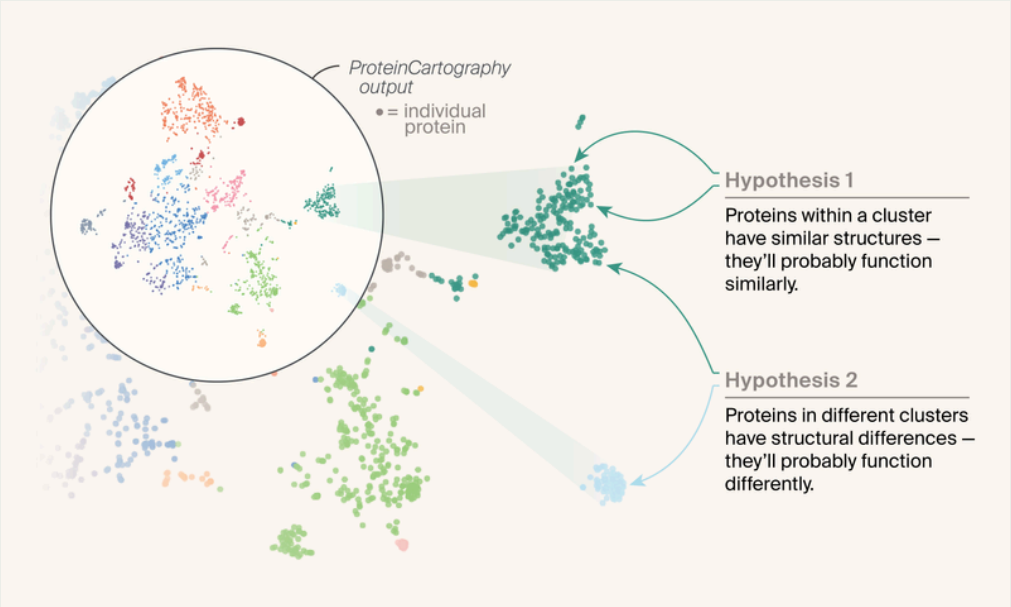

Figure 1

Foundational hypotheses we intend to test via biochemical validation.

The ProteinCartography generated t-SNE for MAPK10 (UniProt ID: P53779) with examples of our hypotheses indicated. This ProteinCartography analysis was originally done in our initial ProteinCartography pub and full data for this analysis can be found there and in our Zenodo repo.

The plan

As we work toward validating ProteinCartography, we'll go through the following steps. We'll update this pub so that it can serve as a roadmap for future validation plans and for how one might follow up on ProteinCartography results.

So far, we've selected two protein families for initial validation using a strategy discussed below. If you'd like to read about these protein families and see some practical examples of this process, visit the pubs for our candidate protein families: deoxycytidine kinase or Ras GTPase.

Step 1: Decide which protein families to focus on

To test these hypotheses, we first had to identify protein families to work with. For our initial analyses, we aimed for families that are easy to work with *in vitro* and that had ProteinCartography outputs with defined clusters and functions that we can realistically assay in the lab. We came up with a list of criteria that we thought were essential (Table 1, column 1).

Rather than considering the entire protein universe, we decided to start somewhere with a more tractable number of protein families to choose from. We turned to the list of the 200 most-studied human proteins in the Protein Data Bank (PDB) **[7]**. These proteins have many experimentally determined protein structures, which means the proteins have likely been purified. A note that because these proteins have been deeply studied and because they're easy to work with, they may represent a class of proteins that's potentially more likely to validate ProteinCartography. However, for this first round of validation, we wanted to aim for lower-hanging fruit. For future validations, we may use protein families that more thoroughly stress-test the tool to find the edges of its functionality.

To narrow this down further, we went through our criteria from Table 1 and eliminated proteins in a stepwise manner. From our list of 200 purifiable proteins, we eliminated any proteins that didn't have commercially available assay kits and that hadn't been previously purified from a bacterial host. We also eliminated proteins that were outside our standard length and structural confidence (mean pLDDT) criteria for the ProteinCartography pipeline **[1]**. For example, because the AlphaFold database uses a length cutoff of 1,280 amino acids, we eliminated any proteins over this length, and we eliminated any proteins with a significant amount of disorder (mean pLDDT < 80) **[8][9]** as they're not well-suited for structural comparisons **[1][9]** (at the time of writing, the AlphaFold database FAQ lists the length cutoff and significant disorder limitations described). This left us with 34 proteins, listed in Table 2.

We ran ProteinCartography **[1]** using the standard parameters (searching for 5,000 hits total) on those 34 proteins. We looked for maps with well-defined clusters that appeared to contain representatives from multiple broad taxonomic groups. Using those criteria, we narrowed those 34 protein families down to 14. We dug deep on our top five, including HRas/KRas GTPases (UniProt IDs: P01112 and P01116), glycogen

synthase kinase 3 beta (GSK3ß) (P49841), lysozyme C (P61626), a tyrosine kinase (P43405), and deoxycytidine kinase (P27707). For these five protein families, we scaled up our ProteinCartography runs, asking it to fetch 10,000 total similar proteins from each family to capture additional protein diversity. We found that lysozyme C lacked taxonomic diversity, GSK3ß returned many hits with low-confidence predicted structures, and the tyrosine kinase lacked annotation diversity in existing annotations (all proteins had similar annotations). Other families had similar issues. While these are ProteinCartography outputs that we would eventually like to dive deeper into, for this round of validation, we chose protein families that would help us test the clustering outputs in the simplest possible manner. We chose two protein families so we can test our hypotheses through orthogonal experiments and rely on just one of the families if in-lab analyses prove challenging for the other.

> **SHOW ME THE DATA**: The data associated with this pub, including ProteinCartography results for 30 proteins we ran, can be found in this Zenodo repository (DOI: 10.5281/zenodo.11264123). An additional four from previous ProteinCartography runs can be found in this Zenodo repo (DOI: 10.5281/zenodo.8377393)

The families we settled on are deoxycytidine kinases and Ras GTPases. For both families, we have open questions for which we're seeking feedback. Visit the pubs to learn more and provide your feedback!

Criteria	How we met this criterion	Number of proteins after filtering
Protein must be purifiable	Started with a list of previously purified proteins	200
Protein must meet standard length and pLDDT criteria for ProteinCartography	Eliminated any proteins over 1,280 amino acids or with an average pLDDT under 80	34
Protein activity must be assayable	Eliminated any proteins that didn't have a commercially available kit	34
Standard ProteinCartography outputs must present testable hypotheses	Eliminated any proteins that didn't have well-defined clusters representing a broad taxonomic range	14
Scaled-up ProteinCartography outputs must present testable hypotheses	Chose the top two most interesting	2

Table 1

Criteria for protein family selection.

Protein	UniProt ID	Data source	Length	Average pLDDT
Superoxide dismutase [Cu-Zn]	P00441	**[1]**	154	98
Peptidyl-prolyl cis-trans isomerase A	P62937	**[1]**	165	98
Glutathione S-transferase P	P09211	This study	210	98
Carbonic anhydrase 2	P00918	This study	260	97
Pancreatic alpha-amylase	P04746	This study	511	97
Dihydrofolate reductase	P00374	**[1]**	187	96
Histone deacetylase 8	Q9BY41	This study	377	95
Lysozyme C	P61626	This study	148	94
Transforming protein RhoA	P61586	This study	193	94
DNA polymerase beta	P06746	This study	335	94
Nicotinamide phosphoribosyltransferase	P43490	This study	491	94
⭑ GTPase HRas	P01112	This study	189	93
Hypoxia-inducible factor 1-alpha inhibitor	Q9NWT6	This study	349	93
⭑ GTPase KRas	P01116	This study	189	92
Fibroblast growth factor 1	P05230	This study	155	91
Interstitial collagenase	P03956	This study	469	91
Serine/threonine-protein kinase pim-1	P11309	This study	313	90
⭑ Deoxycytidine kinase	P27707	This study	260	89

Protein	UniProt ID	Data source	Length	Average pLDDT
Glycogen synthase kinase-3 beta	P49841	[1]	420	89
Cyclin-dependent kinase 2	P24941	This study	298	88
Beta-secretase 1	P56817	This study	501	88
Caspase-3	P42574	This study	277	86
Vitamin D3 receptor	P11473	This study	427	85
Serine/threonine-protein kinase PLK1	P53350	This study	603	85
Tyrosine-protein kinase Lck	P06239	This study	509	84
Tyrosine-protein kinase SYK	P43405	This study	635	84
Urokinase-type plasminogen activator	P00749	This study	431	82
Aldo-keto reductase family 1 member B1	P15121	This study	316	98
Casein kinase II subunit alpha	P68400	This study	391	91
Mitogen-activated kinase 1	P28482	This study	360	91
Mitogen-activated kinase 14	Q16539	This study	360	89
Macrophage metalloprotease	P39900	This study	470	88
Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1	Q13526	This study	163	93
Renin	P00797	This study	406	86

Table 2

Proteins we analyzed with ProteinCartography.

Proteins we moved forward with for validation are indicated with stars (⭐).

Future directions

Now that we've selected which protein families to focus on for our initial validation, we're seeking feedback on how we decide which protein clusters to focus on and how to select individual proteins from within clusters. Additionally, we're beginning to plan how we'll actually assay biochemical functions for our protein families.

Step 2: Select clusters to focus on

Our ProteinCartography runs for Ras GTPase and dCK generated 12 clusters for each protein family. For our first round of validation, we want to test our foundational hypotheses on only a handful of clusters. We can identify appropriate clusters based on additional information from ProteinCartography. In addition to the Leiden cluster overlay shown in Figure 1, we also get metadata overlays, including overlays that can tell us about the broad taxonomy of the proteins, characteristics like length, and how similar the proteins are to our input proteins. Additionally, we get an analysis that tells us more about the UniProt annotations for proteins in our space, called a semantic analysis.

In accompanying pubs, we outline all of this data for both deoxycytidine kinases and Ras GTPases. We've selected clusters that we find interesting based on these analyses and request your feedback on deciding which ones to use for our initial validation.

Step 3: Pick individual proteins to bring into the lab

Once we select which clusters to focus on, we'll need a plan for selecting individual proteins to bring into the lab. A typical cluster can contain hundreds of individual

proteins. Our goal for this first round of validation is to keep the number of proteins we analyze relatively low, so we want to be thoughtful about picking proteins. We'd love your input on ways that we might tackle this challenge.

Step 4: Biochemically analyze function across proteins

We have plans in place for purification and simple activity assays, but we'd love to know if there are additional ways to evaluate biochemical or protein-level function that might be useful for validating ProteinCartography.

Summary

We're working toward validating our ProteinCartography tool by testing two foundational hypotheses:

1. Proteins within the same cluster have similar biochemical functions.
2. Proteins in different clusters have functional differences.

We're sharing our strategy for validation as we generate it to gather feedback from the community, but also to provide a roadmap for future validation and for how one might use ProteinCartography results.

So far, we've addressed our first open question — how to select protein families for validation. Further analyses of these protein families can be found in the accompanying pubs:

> **How can we biochemically validate protein function predictions with the...**
>
> deoxycytidine kinase family? **[10]**
>
> Ras GTPase family? **[11]**

Next, we'll work to answer the remaining questions, including how we select clusters to test, how we select individual proteins, and how we go about biochemical analyses of

these proteins.

References

1

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

2

Berman H, Henrick K, Nakamura H. (2003). Announcing the worldwide Protein Data Bank. https://doi.org/10.1038/nsb1203-980

3

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

4

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

5

Belkina AC, Ciccolella CO, Anno R, Halpert R, Spidlen J, Snyder-Cappione JE. (2019). Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. https://doi.org/10.1038/s41467-019-13055-y

6

McInnes L, Healy J, Saul N, Großberger L. (2018). UMAP: Uniform Manifold Approximation and Projection. https://doi.org/10.21105/joss.00861

7

Li Z, Buck M. (2021). Beyond history and “on a roll”: The list of the most well-studied human protein structures and overall trends in the protein data bank. https://doi.org/10.1002/pro.4038

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

Avasthi P, Bigge BM, Radkov A, Wood H, York R. (2024). How can we biochemically validate protein function predictions with the deoxycytidine kinase family? https://doi.org/10.57844/ARCADIA-1E5D-E272

11

Avasthi P, Bigge BM, Radkov A, Wood H, York R. (2024). How can we biochemically validate protein function predictions with the Ras GTPase family? https://doi.org/10.57844/ARCADIA-74AD-345F