Published on Feb 14, 2025 by Arcadia Science **DOI**: 10.57844/arcadia-a757-3651

Structure-based protein clustering sometimes, but not always, provides insight into protein function

We asked whether ProteinCartography's structure-based protein clustering reflects functional features of proteins. We found that proteins often clustered with proteins that have similar functions, but there were cases when this wasn't the case.

Contributors (A-Z)

Audrey Bell, Brae M. Bigge, Feridun Mert Celebi, Megan L. Hochstrasser, Atanas Radkov, Ryan York

Version 1 · Mar 31, 2025

Purpose

ProteinCartography is a tool for structurally comparing, clustering, and mapping protein families **[1]**. It relies on the idea that structure and function are closely linked, an idea that we tested in this analysis. Our foundational hypotheses are that ProteinCartography will cluster functionally similar proteins together while sorting functionally distinct proteins into different clusters based on structural similarities.

Here, we test these hypotheses using *in vitro* data to help give ProteinCartography users some idea of how well clustering aligns with function and when they should confidently use ProteinCartography results.

Building on our previous work, we investigated this by analyzing biochemically characterized deoxycytidine kinases (dCK), proteins that convert deoxynucleosides to their monophosphate form. We evaluated publicly available biochemical data for 34 dCK homologs, and we biochemically characterized four novel proteins from two specific ProteinCartography clusters **[2]**. We tested the enzymatic activity for each protein and five different substrates. We also noted their general characteristics throughout the purification. We used this data to evaluate ProteinCartography, but we hope this data is also useful to scientists studying dCK and related proteins.

We found that ProteinCartography, which uses global structural alignment, is able to sort proteins into clusters based on their enzymatic function, but it does not always do so. For example, proteins annotated as thymidine kinase that act on deoxythymidine all populate a single cluster. However, while proteins annotated as dCK all cluster together, they don't share all functions. This is likely related to how ProteinCartography compares and clusters proteins, something that we're interested in exploring more, and highlights the importance of combining analyses like these with other analyses to learn more about protein function.

- 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.
- This pub is part of our **validation strategy series** of pubs that starts with "A strategy to validate protein function predictions in vitro" **[3]**. Our original ProteinCartography results for the deoxycytidine kinase family can be found in "How can we biochemically validate protein function predictions with the deoxycytidine kinase family?" **[2]**.
- **Data** from this pub, including ProteinCartography results, expression constructs, purification data and images, and individual protein selection data, is available on Zenodo.
- All associated **code** is available in this GitHub repository.

Background

What is ProteinCartography?

ProteinCartography is a structure-based protein clustering tool designed to compare protein structures from a single family across multiple species **[1]**. It identifies proteins similar to an input and compares the structures to produce an interactive map with clustering information. To see whether the results of ProteinCartography can be used to infer functional relationships, we decided to test the two foundational hypotheses underlying ProteinCartography — that proteins within a cluster have similar functions, and proteins in different clusters have differing functions **[3]**. We chose two model protein families for biochemical testing, one of which is deoxycytidine kinase **[2]**.

What is deoxycytidine kinase (dCK)?

The nucleoside kinase dCK is involved in producing DNA synthesis precursors **[4]**. It phosphorylates deoxycytidine (dC) into deoxycytidine monophosphate (dCMP) and can also convert deoxyadenosine (dA) and deoxyguanosine (dG) into their monophosphate forms **[5]**. Human dCK activates several nucleoside analog prodrugs, including anticancer and antiviral drugs **[4]**. While much is known about the human dCK, non-human homologs present an intriguing area of study due to their potentially distinct properties that could enhance anticancer and antiviral therapies.

Using ProteinCartography to investigate the dCK family

In a previous pub **[2]**, we ran ProteinCartography using the human dCK (UniProt ID: P27707). The analysis produced well-defined clusters, with our input protein in cluster 4 (see Figures 1 and 2 in **[2]**). We identified three additional clusters that we thought were interesting — clusters 2, 8, and 9. We're excited about these clusters because cluster 2 contains almost exclusively plant proteins with long disordered regions, while the proteins in clusters 8 and 9 come from a diverse set of species but show a high structural divergence from the human dCK protein.

In this pub, we'll tell you how and why we selected specific proteins and clusters for *in vitro* testing, we'll correlate existing biochemical data for the dCK family with our ProteinCartography results, and we'll present new activity data from the dCK enzymes in the clusters we selected. Finally, we'll talk through the implications or our results for ProteinCartography predictions.

> **SHOW ME THE DATA**: The data associated with this pub, including ProteinCartography results, expression constructs, purification data and images, and individual protein selection data, are in this Zenodo repository (DOI: 10.5281/zenodo.14517344).

The approach

To evaluate the ProteinCartography results for the dCK family, we compiled data on the activity of various dCK homologs. First, we looked to the literature, where we found activity data for 34 dCK proteins. Additionally, to test the results of ProteinCartography using data we generated, we selected a handful of proteins from a couple specific clusters to evaluate *in vitro.* Using the data from the literature and our generated data, we reviewed our original hypotheses related to ProteinCartography. For more info on each of these steps, keep reading. To jump straight to "The results," click here.

Obtaining data from the literature

Before selecting individual proteins for laboratory study, we conducted a literature review to identify biochemically characterized dCK homologs that could help us evaluate ProteinCartography. We found a review article containing biochemical information for 34 dCK proteins **[6]** (Table 1 and Figure 1). We cross-referenced these proteins with our ProteinCartography clusters by re-running our initial analysis using "Cluster" mode and including the biochemically characterized proteins from the literature as key proteins. We used version 0.5.0 of the pipeline for this analysis. We generated a heatmap and Sankey plot to visualize the data (Figure 1, A-B).

> The **code** for these visualizations is in this GitHub repository.

Selecting proteins for *in vitro* analysis

To select individual proteins to bring into the lab, we first identified representative proteins for each cluster. Using the all-v-all structural similarity matrix generated by ProteinCartography, we selected the protein from each cluster that had the highest similarity to every other protein in its cluster.

Next, to ensure diversity in our representatives, we sub-clustered our clusters of interest. We used the Elbow method to determine the optimal number of clusters **[7]** and Scikit-Learn's k-means clustering algorithm for sub-clustering. To find the representatives for each sub-cluster, we selected the protein with the highest similarity to every other protein within its sub-cluster.

Finally, we confirmed that the proteins we selected would be soluble using the web server for Protein-Sol **[8]** (Table 2).

> All of the **scripts** are available in this GitHub repository and the **resulting data** are available in this Zenodo repository.

Purifying selected dCK proteins

We based our expression and purification protocol on a previously successful protocol for human dCK purification **[9]**.

Cloning

We synthesized and cloned the codon-optimized sequences for our proteins of interest into the pET28a(+) vector for *E. coli* expression using Twist Biosciences. The constructs include a 6× N-terminal His tag, as well as a Human Rhinovirus (HRV) 3C

cut site. We ultimately didn't use the cut site, as the purified proteins were active with the tag. We've included our protein expression constructs in this Zenodo repository.

Induction

We transformed the constructs into *E. coli* BL21 (DE3) cells (NEB - C2527H) and incubated overnight at 37˚C in 9 mL YT media with 50 µg/mL kanamycin. The following day, we added 10 mL of the overnight culture to 1 L (4 L for the bc-dNK) of fresh YT media with 50 µg/mL kanamycin. We incubated the cells with shaking until the OD600 reached about 0.6, after which we induced expression with 0.1 mM IPTG. After 4 hours of shaking at 37 °C, we collected the cells via centrifugation at 4,000 × g for 15 minutes and snap-froze the pellets in liquid nitrogen before storing them at −80 °C. All proteins besides human dCK had lower yields. Yield could be increased by optimizing purification buffer conditions, for example, adding a protease inhibitor or altering the pH to account for the pI. The almond protein, pd-dNK, showed signs of degradation during purification, but we were able to get enough active protein to analyze.

Lysis

We thawed the cell pellets and resuspended them in a resuspension buffer containing 50 mM HEPES, pH 7.5, and 500 mM NaCl (pH adjusted with NaOH). We sonicated (VWR 76193-590) the cells, keeping them on ice, using a ⅜-inch horn at 50% amplitude for 2 minutes, with 20-second on and 20-second off intervals, for three total cycles. We clarified the lysate by centrifuging for 15 minutes (or longer if necessary) at 19,000 × g at 10 °C.

Purification

For affinity chromatography, we used a 1 mL HisTrapFF column (Cytiva - 17-5319-01), an AKTA system, and the resuspension buffer described above as our wash buffer and as our elution buffer (with 200 mM imidazole). We combined the elution fractions from the affinity run and concentrated them to 1 mL using a 10 kDa MWCO Pierce concentrator (ThermoFisher - 88528) before injecting the sample onto the size exclusion chromatography column (HiPrep 16/60 Sephacryl S-200 HR; Cytiva - 17116601). The buffer we used for size exclusion chromatography contained 20 mM HEPES, pH 7.5,

200 mM sodium citrate, 2 mM EDTA (pH adjusted with 10 M NaOH). Chromatograms (affinity and size exclusion) are in the Zenodo repository.

Checking concentration and purity

We evaluated protein concentration with a Bradford assay reagent kit (Thermofisher - 23236) and a SpectraMax iD3 plate reader at 595 nm. We prepared a standard curve of bovine serum albumin (Thermofisher - 23209) with eight different standard protein concentrations ranging from 0 µg/mL to 2,000 µg/mL.

We confirmed protein identity and purity with gel electrophoresis. We used Any kD™ Mini-PROTEAN® TGX™ precast protein Gels, 15 wells (Bio-Rad - 4569036) and the Precision Plus Protein™ Dual Color Standards as the molecular weight ladder (Bio-Rad - 1610394). We prepared the 1× running buffer from a commercial stock (10× Tris/Glycine/SDS buffer; Bio-Rad - 1610732). We ran the gels for 30 minutes at a constant voltage of 200 V in a tetra electrophoresis chamber (Bio-Rad - 1658004). We stained the gels using commercial Coomassie solutions (Bio-Rad - 1610436 and 1610438). We imaged the final destained gel using an Azure 600 gel imaging system.

For the western blot, we transferred protein to nitrocellulose membranes (Bio-Rad - 1620112) using a Trans-Blot Turbo transfer system (Bio-Rad - 1704150) and the built-in StandardSD method. We prepared the 1× transfer buffer from a commercial stock (10× Tris/Glycine buffer; Bio-Rad - 1610734), with 20% (vol/vol) final concentration of methanol (VWR - BDH1135-4LG). We blocked the membrane with a commercial casein solution (1× Tris Buffered Saline with 1% Casein; Bio-Rad - 1610782), containing 0.1% (vol/vol) Tween-20 (Bio-Rad - 1610781), for 30 minutes at room temperature with shaking. We incubated the blot at room temperature with shaking, first with a primary anti-His antibody at 1:2,000 dilution (Histidine Tag Antibody | AD1.1.10; Bio-Rad - MCA1396) and then with a secondary antibody at 1:5,000 dilution (Goat-anti-mouse IgG (H+L), HRP conjugate; Advansta - R-05071-500). After the transfer and the blocking step, between the two antibody incubation steps, and after the secondary antibody incubation, we rinsed the membrane several times with a 1× buffer, prepared from a commercial stock (10× Tris Buffered Saline; Bio-Rad - 1706435) containing 0.1% (vol/vol) final concentration Tween-20 (Bio-Rad - 1610781). We visualized the protein using the WesternBright ECL-HRP Substrate (Advansta - K-12045-D20) with the Azure 600 gel imaging system.

Assessing biochemical activity of dCK proteins

We assessed activity of the protein with the Kinase-Glo® luminescent kinase assay kit from Promega (V6071). We prepared the luminescence reagent that we added to each assay according to the manufacturer's instructions. We did each assay in a 50 µl total volume, containing 40 µl of enzyme and 5 µl of dN substrate at 500 µM final concentration [Cayman Chemical; dC - 34708; dG - 9002864; dA - 27315; thymidine (dT) - 20519; deoxyuridine (dU) - 27803], and 5 µl of ATP, also at 500 µM final concentration (Cayman Chemical - 14498). For each assay, we used 0.4 mg/mL final protein concentration. We incubated the reactions at room temperature without shaking for 60 minutes, after which we added 50 µl of the luminescence reagent and incubated for 10 minutes at room temperature. We measured the outputs with a SpectraMax iD3 plate reader (integration: 1,000 ms and read height: 1 mm). We performed the assays for each protein and each deoxynucleoside in triplicate. We calculated enzyme activity as the luminescence signal per minute per mg protein and normalized the activity values so that the sample with the highest enzyme activity was set to 100%.

Additional methods

We used ChatGPT to write some of the text, as well as to suggest wording ideas and then chose which small phrases or sentence structure ideas to use. We also used ChatGPT to help critique, clarify, and streamline text that we wrote.

We generated figures in this pub using code in the arcadia-pycolor GitHub repo **[10]**.

> All **code** we generated and used for this pub is available in this GitHub repository (DOI: 10.5281/zenodo.14814709), including scripts used to evaluate the literature and identify representative proteins. Additionally, all **data** we generated for this pub is available in this Zenodo repository.

The results

We evaluated our ProteinCartography results with data from the literature and data we generated in-house. To jump to our analysis of all this data as a whole, click here.

> **SHOW ME THE DATA**: The data associated with this pub, including ProteinCartography results, expression constructs, purification data and images, and individual protein selection data can be found in this Zenodo repository.

Biochemically characterized dCK proteins from the literature demonstrate the utility of ProteinCartography clustering

Before selecting dCK proteins to characterize *in vitro*, we searched the literature for pre-existing enzymatic data. The review article “Non-Viral Deoxyribonucleoside Kinases – Diversity and Practical Use” **[6]** includes biochemical data for 34 dCK enzymes, listed in Table 1, that we used to evaluate ProteinCartography (Figure 1).

Figure 1

ProteinCartography results for dCK family.

The interactive UMAP for the dCK family produced using ProteinCartography with the 34 additional proteins from the literature added. The human protein is in LC04.

Members of the dCK family are known to act on multiple deoxynucleosides (dNs), a distinguishing feature of proteins in this family (Figure 2). Sixteen of the 34

characterized enzymes are annotated as thymidine kinase (TK, TK1, TK1a, TK1b, TK2) and show high activity towards dT (Figure 2). The one exception is the *Xenopus laevis* enzyme, annotated as TK2, which shows the highest activity towards dC. Of the five biochemically characterized dCK proteins, three are annotated as dCK and have a dominant activity towards dC, while two are annotated as dCK2 and have a dominant activity towards dG (Figure 2). The dAKs and dGKs generally have matching annotations and biochemical activity (Figure 2). The dNKs show broad specificity towards two or more deoxynucleosides.

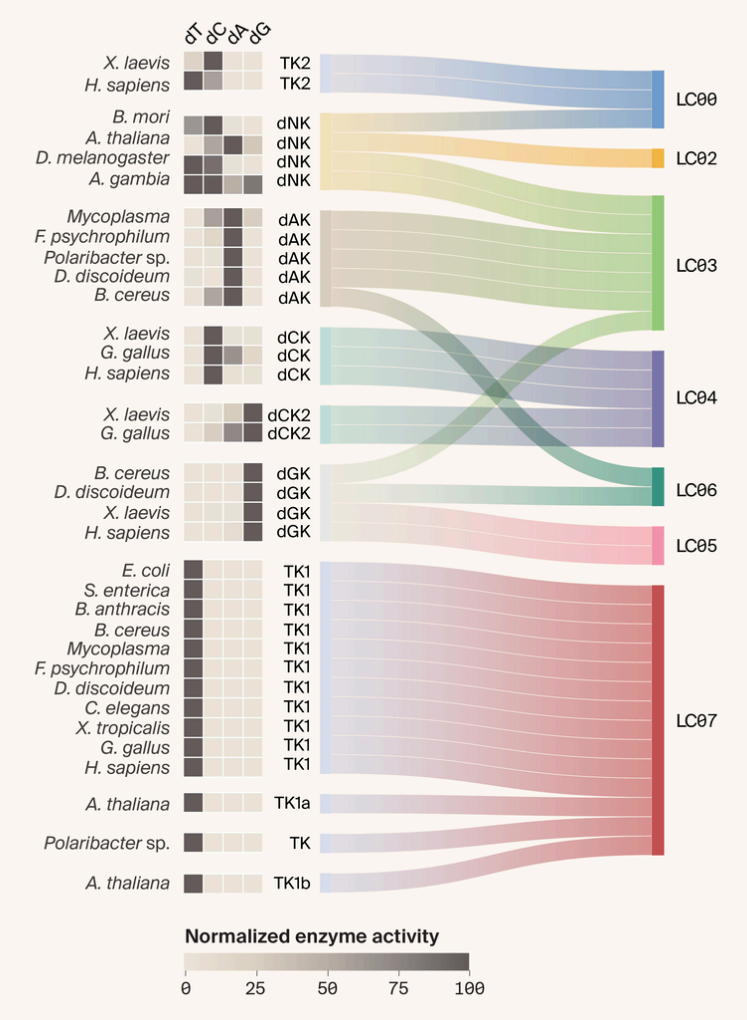

Figure 2

ProteinCartography's structure based clustering largely reflects the function of characterized dCK proteins.

Normalized enzymatic activity of characterized dCK proteins compiled in this review article are shown in the heatmap on the left. Enzymatic activity refers to catalytic efficiency (k_{cat}/K_m). The data are normalized so that the highest activity for each protein is set at 100%. These proteins were sorted into ProteinCartography clusters on the right. A Sankey plot connects the enzymatic activity to the clustering. Each cluster is represented by a box, the color of which matches the cluster colors in Figure 1.

Organism	Annotation	UniProt ID	Cluster
Homo sapiens (human)	TK2	O00142	0
Xenopus laevis (African clawed frog)	TK2	Q8UVZ9	0
Bombyx mori (silk moth)	dNK	Q9BKL3	0
Arabidopsis thaliana (thale cress)	dNK	A0A654ENJ6	2
Anopheles gambiae (African malaria mosquito)	dNK	Q86LB8	3
Drosophila melanogaster (fruit fly)	dNK	Q9XZT6	3
Dictyostelium discoideum (social amoeba)	dAK	Q54YL2	3
Bacillus cereus	dGK	Q81JC3	3
Mycoplasma mycoides subsp. *mycoides* SC (mycoplasma)	dAK	Q93IG4	3
Flavobacterium psychrophilum	dAK	A6GWA3	3
Polaribacter sp. MED 152	dAK	A2U3R9	3
Homo sapiens (human)	dCK	P27707	4
Gallus gallus (chicken)	dCK	Q5ZMF3	4
Gallus gallus (chicken)	dCK2	Q5ZJM7	4
Xenopus laevis (African clawed frog)	dCK	A0A1L8HV70	4
Xenopus laevis (African clawed frog)	dCK2	Q6DD33	4
Homo sapiens (human)	dGK	Q16854	5
Xenopus laevis (African clawed frog)	dGK	Q6GPW6	5
Dictyostelium discoideum (social amoeba)	dGK	Q54UT2	6
Bacillus cereus	dAK	Q0H0H5	6
Homo sapiens (human)	TK1	P04183	7
Gallus gallus (chicken)	TK1	P04047	7
Xenopus tropicalis (Western clawed frog)	TK1	Q5I0A2	7
Caenorhabditis elegans (roundworm)	TK1	F3Y5P8	7

Organism	Annotation	UniProt ID	Cluster
Arabidopsis thaliana (thale cress)	TK1a	Q9S750	7
Arabidopsis thaliana (thale cress)	TK1b	F4KBF5	7
Dictyostelium discoideuml (social amoeba)	TK1	Q27564	7
Escherichia coli	TK1	P23331	7
Salmonella enterica	TK1	Q7CQF3	7
Bacillus anthracis	TK1	Q81JX0	7
Bacillus cereus	TK1	Q0H0H6	7
Ureaplasma parvum serovar 3 (mycoplasma)	TK1	Q9PPP5	7
Flavobacterium psychrophilum	TK1	A6GYI4	7
Polaribacter sp. MED 152	TK	A2TYX7	7

Table 1

dCK proteins in the established dataset we further analyzed in this pub.

A similar table first appeared in this review article.

Next, we investigated where these proteins fall within our ProteinCartography map. Cluster 4, which contains the human dCK protein, also contains the other characterized dCK proteins (Figure 1, Figure 2, and Table 1). The two dCK proteins, dCK and dCK2, do act on different dNs according to the biochemical data, but perhaps share enough of their structure that they still cluster together. There's a tight cluster of TK1 proteins in cluster 7, but proteins annotated as TK2 fall into a separate cluster (Figure 1, Figure 2, and Table 1). This aligns well with the function of these proteins as TK1s act almost exclusively on dTs, while TK2s act on dTs and dGs (Figure 1 and Figure 2). This suggests that in some cases ProteinCartography is able to sort proteins into structure-based clusters that do reflect some function.

The results are less straightforward for the dAKs and dGKs. Four of the five dAK proteins are in cluster 3, with the other dAK falling into another cluster (Figure 1, Figure 2, and Table 1). There are small structural differences between this dAK and the others in the cluster, mostly around the N and C termini, but we found no clear functional

reason for them to cluster separately. Proteins annotated as dGK are distributed between three clusters with two proteins landing in cluster 5 (Figure 1, Figure 2, and Table 1). This is interesting as dGKs do seem to exclusively act on dG (Figure 1 and Figure 2). Perhaps there are structural differences in these proteins that don't affect substrate specificity. The dNKs, whose activity varies, are distributed into three clusters as well, with two landing together in cluster 3 with the dAKs (Figure 1, Figure 2, and Table 1). Interestingly, although the protein from *Bombyx mori* is annotated as a dNK, it's sorted into cluster 0 with the TK2 proteins, which reflects its function (Figure 1 and Figure 2). Together, this provides evidence that while ProteinCartography can separate proteins based on function, it doesn't always do so. However, it's possible that there are other functional differences between these proteins beyond enzymatic activity and are therefore not reflected in this analysis.

In vitro analysis of dCK proteins further highlights ProteinCartography's utility

The previously characterized proteins show that ProteinCartography can sort proteins into clusters based on function, but it also showed that there are cases when it doesn't. We wondered if there were additional functions beyond enzymatic activity that might better align with clustering. For example, for the 34 proteins in the study, we only had enzymatic data. We hoped working with the proteins ourselves might allow us to look at other functions that could lead to differences in structures and, therefore, clustering. Additionally, purifying and analyzing proteins in the lab allowed us to more directly compare proteins purified in the same lab, using the same purification strategy and the exact same assay conditions. To learn how we chose which dCK enzymes to test ourselves, read on. To skip straight to what we found, click here.

Selecting clusters and individual proteins for biochemical characterization

We planned to directly compare proteins within the same cluster and proteins in different clusters. Which clusters we chose for this comparison wasn't necessarily important, so we selected clusters that were interesting for reasons beyond ProteinCartography validation. We previously identified **[3]** three interesting clusters in addition to the cluster containing our input protein (cluster 4) — clusters 2, 8, and 9 (see

Figure 2 in "How can we biochemically validate protein function predictions with the deoxycytidine kinase family?"). We polled the Twitter/X community to help us select a single cluster. We decided to focus on LC02, the top choice in the Twitter/X poll, and the input-containing cluster, LC04, for our subsequent protein selection and validation studies of the ProteinCartography results. Thanks to all who voted!

Cluster 2 almost exclusively contains plant proteins that are longer than the human protein with disordered regions at the N- or the C-terminus. However, the part of the plant proteins that align well with the human dCK protein is well-structured and quite conserved. Only one of these proteins was included in the list of previously biochemically characterized proteins in the previous section. We also selected proteins from cluster 4 because this cluster contains our input protein. We can test the hypothesis that proteins within a cluster function similarly by comparing the input human dCK to another protein in this cluster and by comparing the proteins in cluster 2 to each other. Comparing proteins from both of these two clusters should let us test our hypothesis that proteins from different clusters have distinct activities.

We selected representatives for each cluster by identifying the protein with the highest similarity to every other protein in the cluster. We also sub-clustered the clusters of interest to select additional proteins that are more representative of the diversity of the larger clusters. We evaluated the solubility and the predicted isoelectric point (pI) of each representative protein. The proteins in Table 2 are the representatives identified. One of the cluster 2 sub-clusters presented a representative that was predicted to be insoluble. Therefore, we substituted in the Rickettsiales protein for this sub-cluster, which is unique in that it's one of the only proteins in this cluster not from plants.

UniProt ID	Cluster	Annotation	Organism	Predicted molecular weight	Pred s solu
A0A4Y1QVV5	LC02	P-loop containing nucleoside triphosphate hydrolases superfamily protein (pd-dNK)	*Prunus dulcis* (almond)	55.0 kDa	
A0A3P6ASY1	LC02	Deoxynucleoside kinase domain-containing protein (bc-dNK)	*Brassica campestris* (field mustard)	27.5 kDa	
A0A2A5BCG8	LC02	Deoxynucleoside kinase domain-containing protein (rb-dNK)	Unidentified Rickettsiales bacterium	23.2 kDa	
A0A7J5YK87	LC04	Deoxynucleoside kinase domain-containing protein (dm-dCK)	*Dissostichus mawsoni* (Antarctic toothfish)	29.2 kDa	
P27707	LC04	Deoxynucleoside kinase	*Homo sapiens* (human)	33.0 kDa	

Table 2

Selected proteins for in-lab analysis.

We determined the predicted scaled solubility on the Protein-Sol website, where higher values indicate higher predicted solubility. The average protein in *E. coli* has a predicted scaled solubility of 0.45.

The human dCK shares some, but not all, functions with a protein from its cluster

To start, we purified the human dCK protein using a published expression and purification protocol **[9]** and confirmed that it exists as a dimer in its native state (Figure 2, A and C). The kinase activity of our purified human dCK closely matched its reported activity, acting primarily on dC and less so dA and dG **[11]** (Figure 2, B).

To determine if proteins within a structure-based cluster share biochemical functions, we purified and analyzed the Antarctic toothfish (*Dissostichus mawsoni*) dCK (dm-dCK), which resides in the same cluster as the human dCK (Figure 3, A). We found that dm-dCK, like human dCK, behaved as a dimer (Figure 3, C). If we think of assembly of monomers into an oligomeric form as another function of this protein, this can be counted as another instance where proteins within a ProteinCartography cluster share functions. The Antarctic toothfish protein, dm-dCK, showed similar activity to our input protein against a comparable selection of deoxynucleoside substrates, with the exception of the activity towards dT (Figure 3, B). While the human dCK enzyme didn't show any activity towards dT, the dm-dCK did (Figure 3, B).

These results lend support to our hypothesis while also generating some questions. Functions conserved between the two proteins from cluster 4, human dCK and dm-dCK, include behavior as a dimer, enriched activity towards dC, and lesser activity towards dA and dG. A function not conserved between the two proteins is the activity towards dT. This suggests that while ProteinCartography can separate proteins based on function, it doesn’t separate on every function. This is expected, as proteins are complex and perform many different functions.

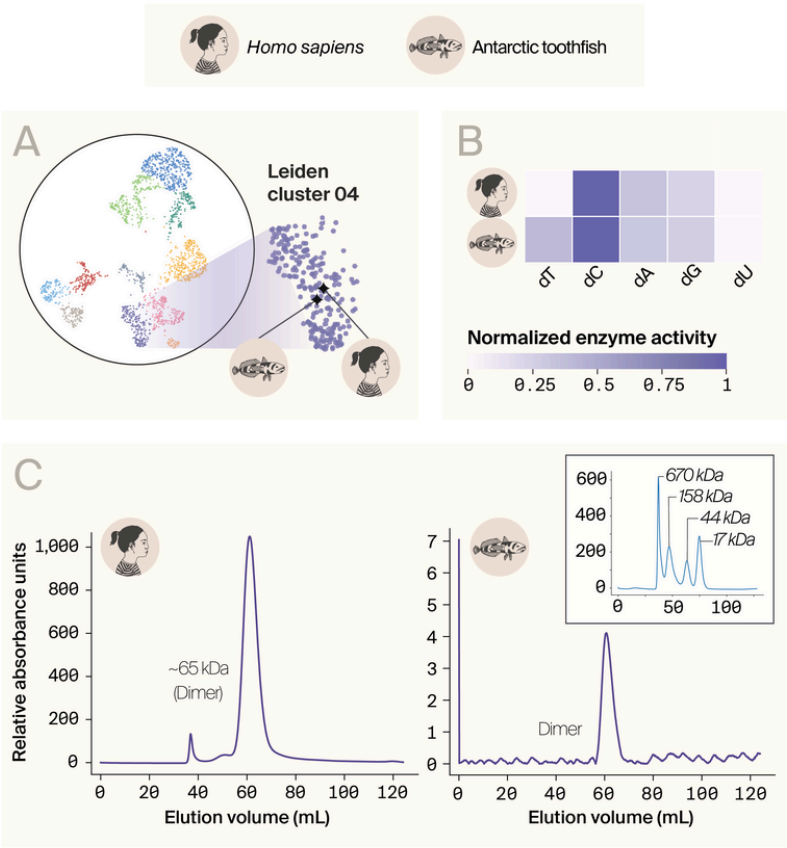

Figure 3

The human dCK and the antarctic toothfish protein from cluster 4 share similar functions.

(A) We first analyzed the cluster containing our input protein the human dCK (P27077). This cluster also contains the Antarctic toothfish protein (A0A7J5YK87).

(B) We measure kinase activity for the human dCK and the antarctic toothfish protein using five substrates. We calculated enzyme activity as the luminescence signal per minute per mg protein. We set the highest activity to 100% and normalized the data accordingly. We also show that our measured human data matches that of the literature.

(C) Size exclusion chromatography results show that both the human and antarctic toothfish dCK proteins form dimers. The graph on the left shows the analyzed

commercial standards that we used to estimate the weight of the purified human dCK protein. The size exclusion data and all accompanying gels and western blots from the purification are on Zenodo. Additionally, gels can be found in Supplementary Figure 1.

Three proteins in another cluster share some, but not all, functions

We also selected three proteins from cluster 2 to purify and analyze, including the almond (*Prunis dulcis)* dNK (pd-dNK), the field mustard (*Brassica campestris)* dNK (bc-dNK), and a dNK from a Rickettsiales bacterium (rb-dNK) (Table 2). As we did with human dCK and dm-dCK, we compared the functions of these three proteins to test whether proteins in the same cluster share functions (Figure 4, A).

All three proteins from this cluster eluted from the size exclusion run at very high molecular weights, indicating that they either form a multimer or aggregate, but the protein is active after purification (Figure 4, C). This oligomerization could be considered a function that's shared by proteins within a cluster. Two of the proteins in cluster 2, rb-dNK and pd-dNK, had high activities towards all of the tested deoxynucleosides, including dU (Figure 4, B). The final protein from this cluster, bc-dNK, has the highest activity towards dG but also acted dT, dC, and dU (Figure 4, B).

Similar to the comparison we made with the protein in cluster 4, some functions are conserved between all three proteins, while some are not. All three proteins form some higher-order multimer and act on multiple deoxynucleosides. However, rb-dNK and pd-DNK seem to act less specifically than bc-dNK. These results support the idea that ProteinCartography can separate proteins based on some, but not all, of their functions.

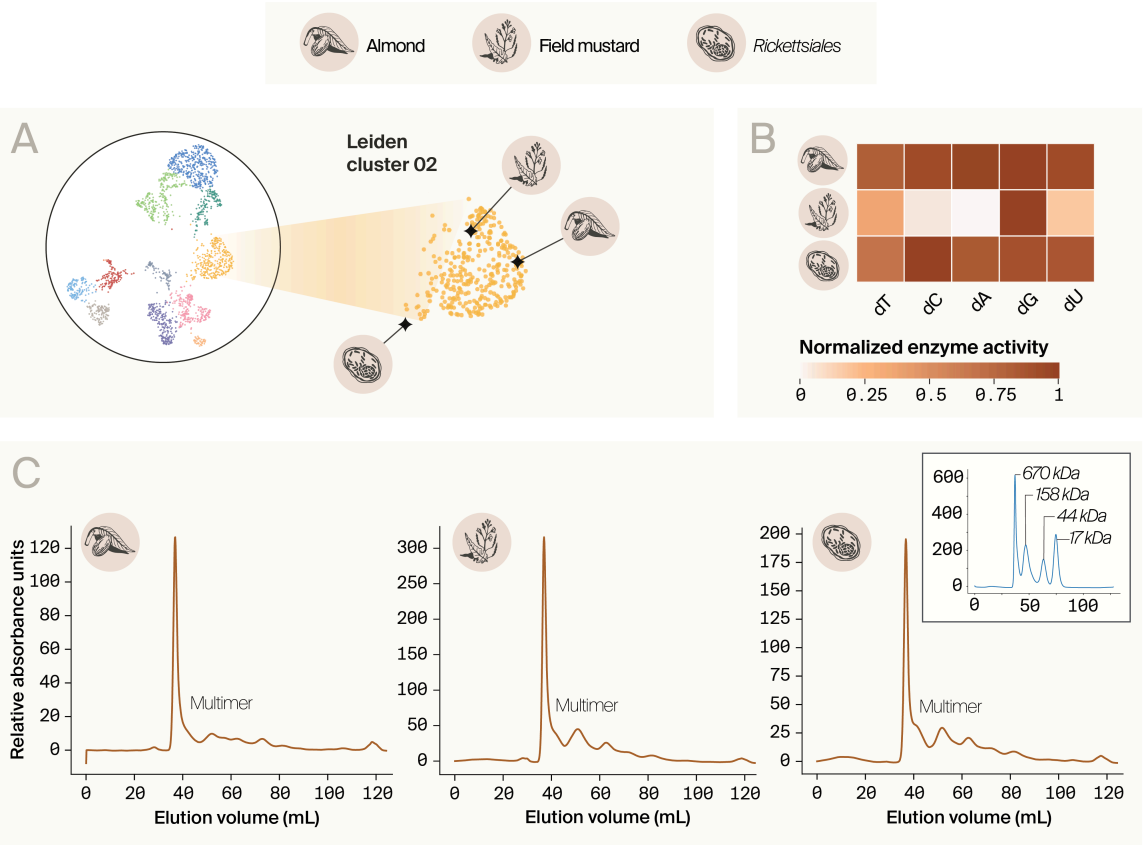

Figure 4

Three proteins from cluster 2 share similar functions.

(A) We next analyzed cluster 2 which contains primarily plant proteins. We specifically looked at the almond protein (A0A4Y1QVV5), the field mustard protein (A0A3P6ASY1), and the Ricketsiales protein (A0A2A5BCG8).

(B) We measure kinase activity for each enzyme using five substrates. We calculated enzyme activity as the luminescence signal per minute per mg protein. We set the highest activity to 100% and normalized the data accordingly.

(C) Size exclusion chromatography results show that all three proteins elute at a higher than expected molecular weight. The graph on the left shows the analyzed commercial standards. The size exclusion data and all accompanying gels and western blots from the purification are on Zenodo. Additionally, gels can be found in Supplementary Figure 1.

Proteins in different clusters have some distinct biochemical features

To test if proteins in different clusters have different functions, we compared the proteins from cluster 4 to the proteins from cluster 2. First, the proteins in cluster 4, which contains the human dCK and dm-dCK, eluted as a dimer from our size exchange column, while the proteins in cluster 2 eluted as multimers larger than dimers. In every case that we've tested, this oligomerization "function" aligns with ProteinCartography clustering.

We previously established that the functions aren't totally conserved within the clusters. However, the activity profiles of proteins within a cluster are much more similar to each other than to the activity profiles of proteins from the other cluster (Figure 5). All proteins act on dC and dG to some degree, meaning that the two clusters do share functions, which isn't totally unexpected as they're all from the same family of proteins (Figure 5). We're also able to identify functional differences between the proteins in the two clusters. The proteins in cluster 4 act primarily on dCK, while the proteins in cluster 2 are generally less specific, acting on the deoxynucleosides that aren't substrates for the proteins in cluster 4 (Figure 5).

Overall, the comparison between clusters 2 and 4 supports the idea that proteins from different structure-based clusters show at least some distinct functions. They form different higher-order structures and have differing substrate specificity.

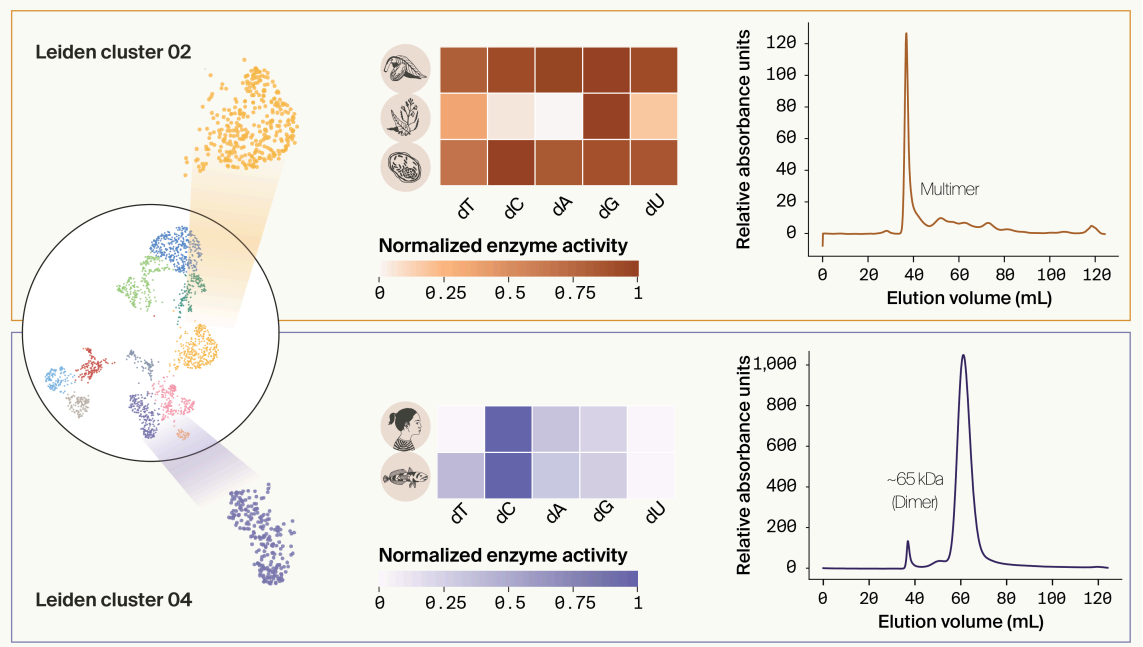

Figure 5

Proteins in cluster 2 and cluster 4 have different functions.

We compare the activity and tertiary structure of proteins in cluster 2 and cluster 4. We see that proteins in cluster 2 tend to act on multiple substrates while proteins in cluster 4 tend to act primarily on dC. We also find that proteins in cluster two form multimers as demonstrated on the right, while proteins in cluster 4 form dimers.

Bringing it all together

In this pub, we presented data from the literature for 34 proteins related to dCK and generated our own data to add to that list. We found instances where function clearly aligned with cluster separation and instances where it was less clear. For example, the 14 TK1 proteins that act exclusively on dT all landed in cluster 7, supporting the idea that ProteinCartography can sort proteins into structural clusters based on their function (Figure 6). Similarly, the proteins in cluster 0 all act on both dT and dC, but not dA and dG, while all the proteins in cluster 5 act primarily on dG (Figure 6). However, the activities of dCK family proteins towards different substrates are more mixed for clusters like 3 and 6, suggesting that ProteinCartography doesn't always separate proteins based on function. This is also a trend we see for the proteins we selected for

our in-lab analysis. Most proteins in cluster 4, which contains the human dCK protein, act most strongly on dC, with the exception of the dCK2s (Figure 6). The proteins in cluster 2 seem to act on all substrates to some degree (Figure 6).

There are many possibilities for why ProteinCartography sometimes, but not always, sorts proteins based on function. First, ProteinCartography performs a global structural alignment, so perhaps in these cases there are subtle local structural differences between proteins that a global alignment doesn't pick up. For example, we know that proteins in cluster 2 are generally longer with large disordered termini. ProteinCartography is much more likely to pick up these larger differences than subtle differences that might account for differences in enzymatic activity.

Looking at a ProteinCartography map is like taking a bird's eye view of the similarities between proteins. ProteinCartography creates a continuous distribution that the clustering tries to discretize. On average, proteins in cluster 2 are likely more similar to other proteins in cluster 2 than in other clusters. However, upon closer evaluation, the reality is more nuanced. For example, two of the characterized proteins in cluster 3, Q86LB8 and Q9XZT6 from *Drosophila* and *Anopheles*, are actually more closely related to the characterized proteins in cluster 0 that have similar functions than to the characterized proteins in cluster 3. These two proteins have an average TM-score of 0.80 compared to the rest of the characterized proteins in cluster 3, while they have an average TM-score of 0.90 compared to the characterized proteins in cluster 0. The functions of these proteins align with these findings, but this isn't always the case. The proteins in cluster 4, which have some functional diversity, have an average structural similarity of 0.94, while the proteins in the very tight cluster of TK1 proteins in cluster 7 that exclusively actin on dT have an average TM-score of only 0.84.

Because clustering tries to sort a continuous distribution into discrete groups of proteins, there's no "correct" clustering, only clustering that's more or less reflective of the properties we care about. ProteinCartography uses the baseline Foldseek settings to create the all-vs-all similarity matrix, meaning that it only calculates structural similarity scores for the top 1,000 proteins for each protein based on an initial alignment of structure-representing 3Di sequences **[12]**. Changing the parameters does alter clustering, so tuning these parameters could help us get closer to clusters that reflect function. However, the truly optimal parameters for each protein family and use case are likely different, so perhaps this is something users should experiment with for their own use cases.

One of the novelties of ProteinCartography is that it uses structural comparisons to identify matches and for clustering. This has some benefits. For example, because we use structure-based searches (in addition to sequence-based searches) to identify similar proteins we're able to cast a larger net. For example, in this analysis, we have proteins that have as little as 8.5% identity compared to our input protein that we identified using protein structure. The Antarctic toothfish and the human protein share 70% of their sequences, so it's not surprising to find them in the same cluster and with similar functions. However, the three proteins in cluster 2 have less than 40% sequence identity. Despite this, they share structural similarity and some functional similarity. On the flip side, because ProteinCartography is based on rigid, global structural alignment, it might not pick up on small changes at the active site for example.

Finally, it's typical to focus primarily on enzymatic activity when comparing enzymes, but we use the term "function" broadly. Other "functions" or functional properties we could look at include things like stability, tertiary structure, and other functions in the cell. These auxiliary functions should be considered. In our in-lab analysis we found differences in oligomerization states between proteins in cluster 2 and cluster 4, suggesting this could be another "function" that's picked up by ProteinCartography for this family (Figure 6). It would be useful to apply or develop assays that can be generalized and used on multiple protein families quickly to gather more multi-dimensional data about proteins.

Overall, based on our results for the dCk family, ProteinCartography can be a useful tool for investigating protein function, but it should be used alongside other tools and analyses.

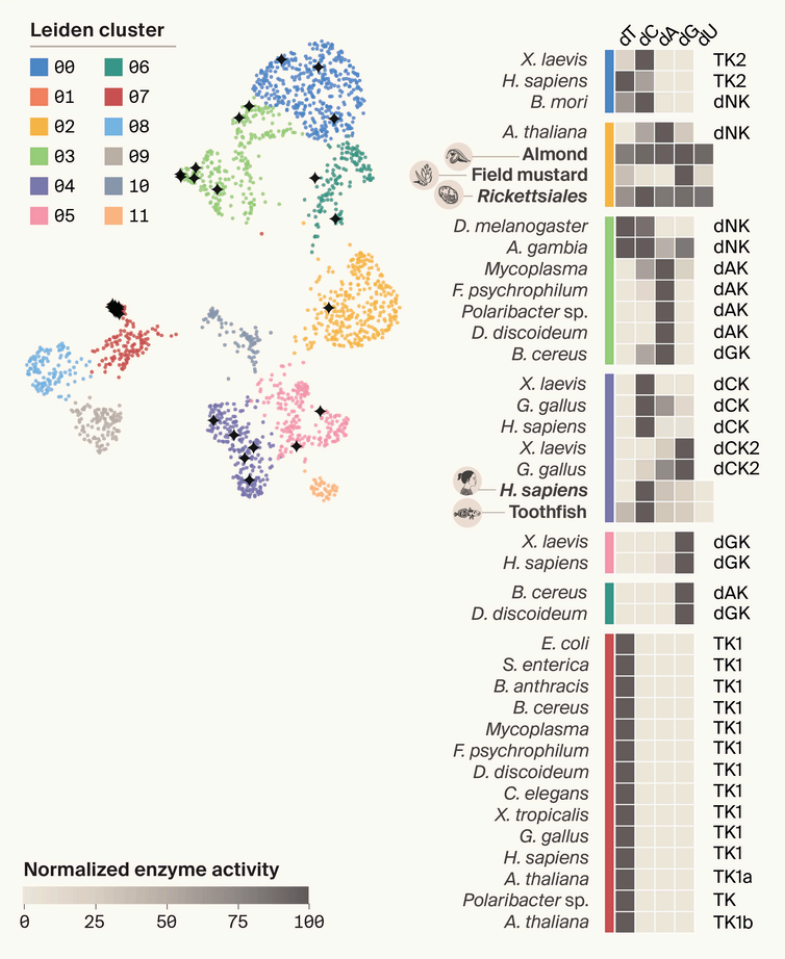

Figure 6

ProteinCartography sometimes but not always sorts proteins into structure-based clusters that reflect function.

We bring together the existing literature data and our experimental data to look at how protein functions are distributed across the ProteinCartography map. Proteins analyzed in this study are represented as four-point stars in the map in the upper left.

Key takeaways

- ProteinCartography separates proteins based on their global protein structures. We asked if these global protein structure relationships could be used to learn anything

about the function of the proteins.

- ProteinCartography can sort proteins based on their functions. However, it doesn't always do so.
- ProteinCartography can be used to learn more about protein function and to form hypotheses but should be used alongside other tools and analyses designed to study protein function.

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. 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

3. Avasthi P, Bigge BM, Radkov A, Wood H, York R. (2024). A strategy to validate protein function predictions in vitro. https://doi.org/10.57844/ARCADIA-CAE9-96C4

4. Sabini E, Hazra S, Ort S, Konrad M, Lavie A. (2008). Structural Basis for Substrate Promiscuity of dCK. https://doi.org/10.1016/j.jmb.2008.02.061

5. Shewach DS, Reynolds KK, Hertel L. (1992). Nucleotide specificity of human deoxycytidine kinase. https://pubmed.ncbi.nlm.nih.gov/1406603/

6. Slot Christiansen L, Munch-Petersen B, Knecht W. (2015). Non-Viral Deoxyribonucleoside Kinases – Diversity and Practical Use. https://doi.org/10.1016/j.jgg.2015.01.003

7. Thorndike RL. (1953). Who Belongs in the Family? https://doi.org/10.1007/bf02289263

8. Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. (2017). Protein–Sol: a web tool for predicting protein solubility from sequence.

https://doi.org/10.1093/bioinformatics/btx345

9 Sabini E, Hazra S, Konrad M, Lavie A. (2007). Nonenantioselectivity Property of Human Deoxycytidine Kinase Explained by Structures of the Enzyme in Complex with <scp>l</scp>- and <scp>d</scp>-Nucleosides. https://doi.org/10.1021/jm0700215

10 Arcadia Science. (2024). arcadia-pycolor. https://github.com/Arcadia-Science/arcadia-pycolor

11 Chottiner EG, Shewach DS, Datta NS, Ashcraft E, Gribbin D, Ginsburg D, Fox IH, Mitchell BS. (1991). Cloning and expression of human deoxycytidine kinase cDNA. https://doi.org/10.1073/pnas.88.4.1531

12 van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. (2023). Fast and accurate protein structure search with Foldseek. https://doi.org/10.1038/s41587-023-01773-0