MANUSCRIPT TITLE

HerpesFolds: A proteome-wide structural systems approach reveals insights into protein families and activities of all nine human herpesviruses

AUTHORS

Timothy K. Soh[1, 2, 3, 4], Sofia Ognibene[1, 2, 3, 4], Saskia Sanders[1, 2, 3, 4], Benedikt B. Kaufer[5, 6], and Jens B. Bosse[1, 2, 3, 4, *]

[1] Centre for Structural Systems Biology, Hamburg, Germany

[2] Hannover Medical School, Institute of Virology, Hannover, Germany

[3] Cluster of Excellence RESIST (EXC 2155), Hannover Medical School, Hannover, Germany

[4] Leibniz Institute of Virology (LIV), Hamburg, Germany

[5] Institut für Virologie, Freie Universität Berlin, Robert von Ostertag-Straße 7-13, 14163 Berlin, Germany

[6] Veterinary Centre for Resistance Research (TZR), Freie Universität Berlin, 14163 Berlin, Germany

* To whom correspondence should be addressed. Tel: +49 040 8998 87645; Email: jens.bosse@cssb-hamburg.de

ABSTRACT

Structure predictions have become invaluable tools for wet lab science by accelerating the generation of structure-based hypotheses. While AlphaFold-predicted structures are available for most proteins, viral proteins are currently absent from the EMBL/DeepMind AlphaFold database. Here, we provide proteome-wide structure predictions for all nine human herpesviruses and analyze them in depth. All predictions were scored using explicit thresholds. By clustering these predictions into groups with structural similarity, we identified new families, such as the HCMV UL112-113 cluster, that contain a unique beta-barrel conserved in all alpha- and betaherpesviruses. An exhaustive domain-level search found protein families consisting of subgroups with varying numbers of duplicated folds. Using large-scale searches, we identified viral proteins with cellular folds and potential functions, such as the HSV-1 and HSV-2 US2 proteins possessing dihydrofolate reductase folds or the EBV BMRF2 cluster of proteins that might have emerged from cellular equilibrative nucleoside transporters. Our HerpesFolds database can be accessed through an easy-to-use web interface at www.bosse-lab.org/herpesfolds/. It displays 3D models of all protein predictions with their quality scores and structural clusters through an interactive, searchable, open-access web interface.

INTRODUCTION

The *Orthoherpesviridae* family comprises alpha-, beta-, and gammaherpesviruses containing nine human herpesviruses (see Table 1 for virus abbreviations and their respective subfamilies). These are medically important human pathogens, including viruses such as human cytomegalovirus (HCMV), the leading infectious cause of congenital disabilities in the USA. Due to the large number of encoded genes (between 70 and 181, depending on the species), experimental structural information for most viral proteins is not available, severely hampering their functional analysis. Moreover, herpesvirus gene and protein nomenclature are very inconsistent between species. Even worse, some proteins of the same species have multiple names based on their genome position, molecular weight, expression in the infected cell, or packaging into the virion. This situation makes it difficult even for experts to draw parallels between viruses. While some corresponding orthologues between species can be deduced from the UniProtKB annotation, this database is incomplete and sometimes inconsistent. Finally, herpesviruses also have a penchant for duplicating genes, further complicating annotation.

With the development of machine learning-based prediction software[1,2], the accuracy of the structural information has increased so that it can often be directly used for building wet-lab testable hypotheses, making it an invaluable tool for molecular biologists[3]. DeepMind, in cooperation with EMBL-EBI, has published over 200 million protein structures to facilitate this process, covering most of UniProtKB in an open-access database (https://alphafold.ebi.ac.uk). While this is an extraordinary resource for many fields, it currently excludes viral proteins such as the human herpesvirus proteomes.

Herpesvirus phylogeny has classically been sequence-based, such as a recent study that classified the human herpesvirus proteomes into homology families[4]. However, the current revolution in structure prediction quality and rapid structural similarity search algorithms[5] now allow structure prediction-based phylogeny[6]. When applied to large proteins or large phylogenetic distances, *e.g.* across the *Duplodnaviria* realm, this approach clustered species more consistently than the sequence-based ones[7]. Moreover, a recent study using this approach showed that orthopoxviruses acquired cellular proteins, with some losing their original function[8]. Furthermore, structure prediction of the glycoproteins of different flaviviruses identified novel and common mechanisms in the fusion machinery[6].

Here, we predicted the structures of all nine human herpesvirus proteomes, totaling 844 individual proteins. We critically evaluated their accuracy with stringent quality scores and thresholds. We clustered their predicted structures into structurally similar groups and analyzed them using a structural systems virology approach. Many similarities suggest a common ancestor and thus a homologous relationship, whether orthologous, *i.e.* due to speciation, or paralogous, *i.e.* due to gene duplication. We found new members to previous groups identified by sequence, such as the US22 family of HCMV. We could also identify novel groups with novel folds, such as the UL112-113 family harboring a unique beta-barrel domain that is conserved in seven of the nine human herpesviruses. A detailed domain-level similarity search identified several subclusters of proteins harboring different numbers of duplicated domains, deletions, and acquisitions. Using structural similarity searches against cellular proteins, we propose new functions for uncharacterized proteins. Examples include the HSV-1 US2

protein, which is likely a dihydrofolate reductase. We also found cases of exaptation, such as the BMRF2 family of transmembrane proteins that share structural similarity to nucleoside transporters but are likely inactive and now have functions in cell-cell spread. Our structural database is accessible through an open-access and searchable web interface at www.bosse-lab.org/herpesfolds/ that interactively displays the predicted structures and structural clusters. HerpesFolds is a curated database that groups the protein predictions based on structural similarity while linking their established names and groups. It also serves as a reference that can translate the complicated herpesvirus nomenclature for the expert and non-expert alike and sheds light on the relationships of all nine human herpesvirus proteomes. This work will aid in developing wet-lab testable hypotheses and gain new insights into herpesvirus biology and pathology.

MATERIAL AND METHODS

Viruses

The strains and associated reference numbers of the nine human herpesvirus proteomes used for prediction are noted in Table 1. In addition to the full-length proteins, splice variants were also included. A complete list of the predicted proteins can be found in Table S1 Sheet 3.

Software

Protein structure predictions were made with ColabFold[9] and AlphaFold[1] using AlphaFold version 2.3.0. For each protein, 5 models were generated using 3 recycles and a stop-at-score of 100. Structural similarity searches were performed with DaliLite.v5 using local install versions of DALI[10] and Foldseek[5]. Networks were visualized with Gephi v0.10.1. UCSF ChimeraX version 1.7.1 was used for visualization of structures, and PyMOL 2.5.8 was used for RMSD calculations. Data analysis was carried out utilizing custom-written Python scripts, which were partly generated with the assistance of large language models. The complete set of scripts can be accessed through the project's GitHub repository: https://github.com/QuantitativeVirology/Herpesfolds/.

Evaluation of model quality

All proteins were initially predicted using ColabFold[9] as it is more time-efficient. To assess the quality of the resulting models, 3 thresholds were used. Failing any single threshold led to the model being flagged as "fail". We interrogated whether the model optimization had converged by testing the consistency of the predicted local distance difference test (pLDDT) scores associated with every model (Figure S1A and B). To do so, we first calculated the mean and standard deviation of the pLDDT values per model and then the standard deviation of these 5 values, resulting in 'StDev of mean' and 'StDev of

StDev'. To set thresholds, we fit a Gaussian curve to their respective histograms and set the cutoff at the mean of the Gaussian plus 2 times its standard deviation (Figure S1C and D). Therefore, a 'StDev of mean' >3.2 or a 'StDev of StDev' >1.9 was deemed a low-quality model. We also evaluated the predicted template modeling score (pTM) to validate model confidence further. To set a global threshold at which a prediction likely constitutes a folded protein, we randomly chose 106 herpesvirus proteins representing the protein length distribution of the herpesvirus proteomes and scrambled their amino acid sequence. These scrambled sequences were used for model prediction and treated as a negative dataset, assuming that a random residue order should not fold. Receiver Operating Characteristic (ROC) analysis yielded an area under the curve of 91.3%, and the Youden index was used to set a threshold (Figure S1E). Following this analysis, a pTM <0.3150 was deemed a low-quality model. We reanalyzed all models that failed the initial threshold in AlphaFold[1] and rescored them. For example, Figure S1F shows an improved model run in AlphaFold with higher pLDDT scores compared to Figure S1B. The corresponding scripts can be found in our HerpesFolds GitHub repository in the file 'score_colabfold.py'.

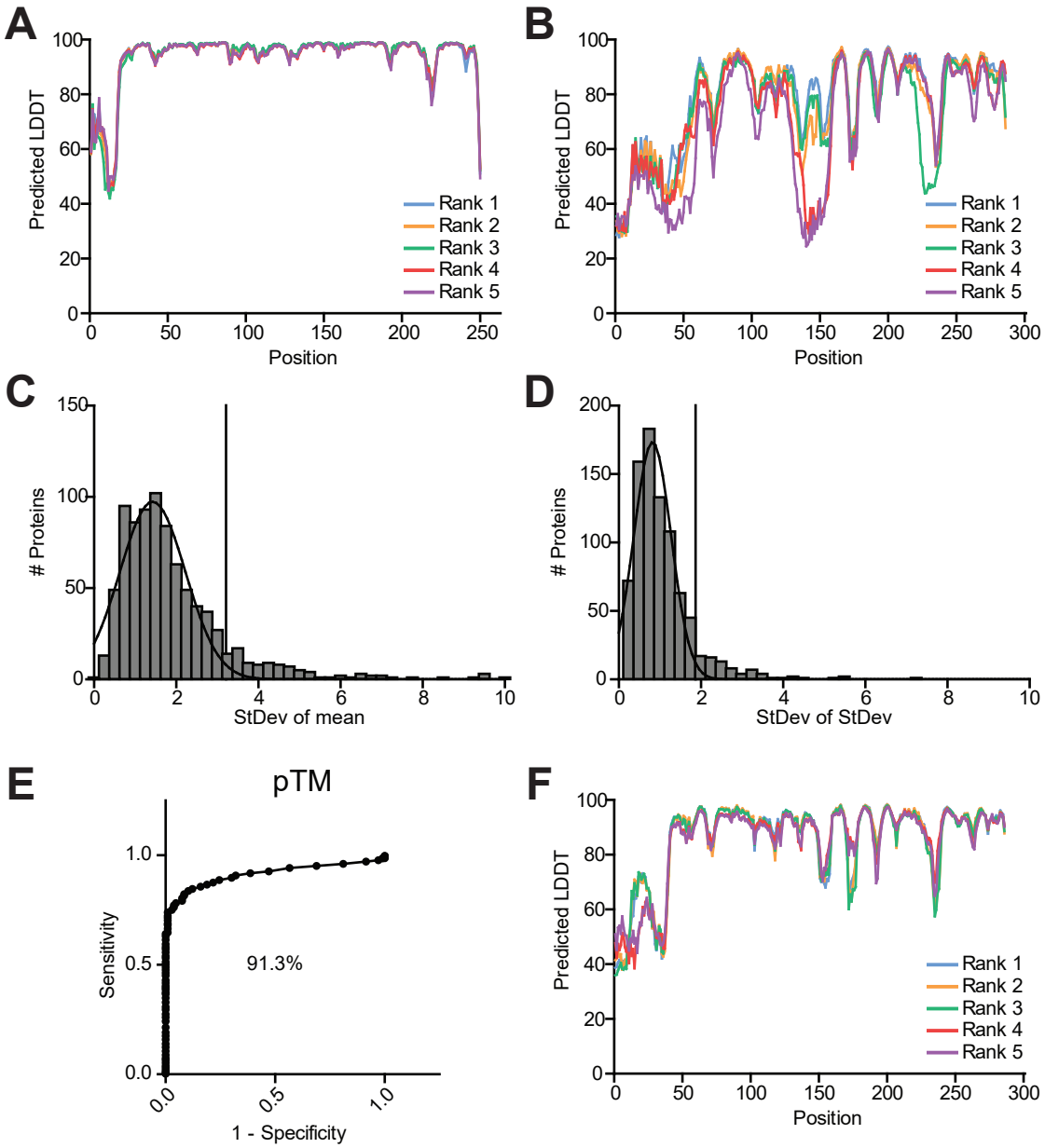

A
Predicted LDDT
Position
Rank 1
Rank 2
Rank 3
Rank 4
Rank 5
B
Predicted LDDT
Position
Rank 1
Rank 2
Rank 3
Rank 4
Rank 5
C
Proteins
StDev of mean
D
Proteins
StDev of StDev
E
pTM
Sensitivity
1 - Specificity
91.3%
F
Predicted LDDT
Position
Rank 1
Rank 2
Rank 3
Rank 4
Rank 5

Figure S1. Developing thresholds for evaluating model quality. A. pLDDT plot of HCMV UL114 as an example of a protein where the pLDDT curves of the 5 models are consistent. B. pLDDT plot of HHV-7 U29 as an example of a protein where the pLDDT curves of the 5 models are inconsistent. C. Histogram of the standard deviation (StDev) of the mean of the pLDDT for each model for all ColabFold models. The threshold (black vertical bar) is set at the mean + 2x StDev, 3.2. D. Histogram of the StDev of the StDev of the pLDDT for each model for all ColabFold models. The threshold (black vertical bar) is set at the mean + 2x StDev, 1.9. E. ROC analysis of the pTM of all ColabFold models that passed the quality scores versus a negative control harboring scrambled amino acid sequences. The area under the curve is 91.3%. F. pLDDT curve of HHV-7 U29 as an example of a protein that passed the quality score thresholds after AlphaFold was used.

Criteria for defining protein prediction disorder and structure content

The pLDDT value at a given residue indicates whether a region is predicted to be disordered (pLDDT <0.50) or structured (pLDDT >0.70). For the models that passed all quality scores, we generated histograms of the percentage of the protein below or above these values and fit a Gaussian to both curves. Since the majority of proteins are expected to be structured, a threshold, >47%, for proteins containing a disordered region was set at the mean of the distribution + 2x StDev for the percentage of protein sequence below 0.50, and a threshold, >31%, for proteins with a structured region was set at the mean - 2x StDev for the percentage of protein sequence above a pLDDT of 0.70. The script for extracting pLDDT values from the .json files associated with every prediction can be found in our HerpesFolds GitHub repository in the file 'score_colabfold.py'.

Virus-virus and virus-cell structural similarity searches and establishment of comparable significance scores for DALI and Foldseek results

To cluster viral proteins into groups of structural similarity, we employed DALI as an established algorithm and Foldseek as a recent algorithm designed to work with vast predicted datasets with decreased computation times. Unfortunately, the significance scores are very different. The DALI Z-score measures the number of standard deviations above expected[11], and the Foldseek extreme-value (E-value) reports a score distribution[5]. To find a correlation between both scores, we plotted the DALI Z-scores and Foldseek E-values of the virus-virus outputs common to both searches against each other (Figure S2A and B). We found that the used threshold of <0.001 for Foldseek[8] corresponded with a DALI Z-score of >8.9 (Figure S2B), close to the recommended score of 8 for probable homologs[11]. For this reason, we used the recommended E-value of <0.001 for Foldseek and a DALI Z-score of >8.0 as comparable thresholds for virus-virus similarity searches. Since the results from both algorithms were similar, we only used Foldseek for virus-host and domain-level searches as DALI is orders of magnitude

slower, which would have made these analyses computationally too expensive otherwise. Similarity searches against cellular proteins were done using the PDB database (Table S4 Sheet 1), and the predicted AlphaFold structures of the Swiss-Prot database (Table S4 Sheet 3).

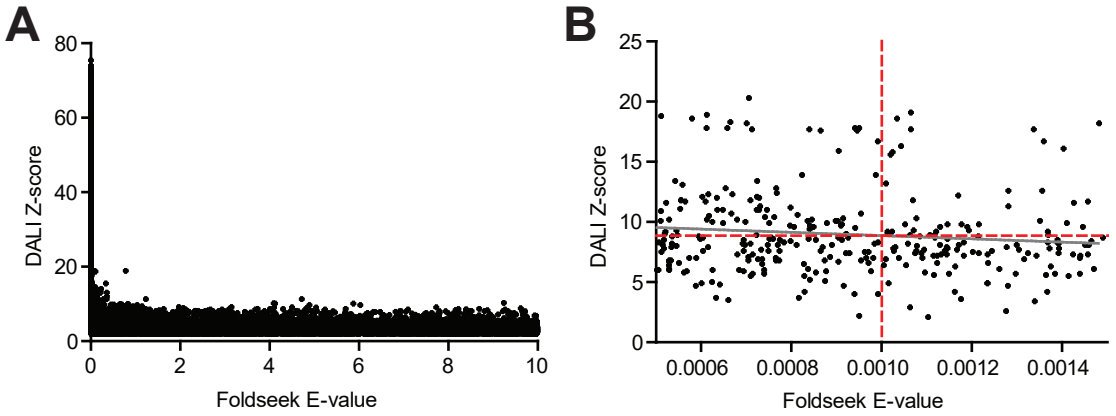

A
80
60
40
20
0
DALI Z-score
0
2
4
6
8
10
Foldseek E-value
B
25
20
15
10
5
0
DALI Z-score
0.0006
0.0008
0.0010
0.0012
0.0014
Foldseek E-value

Figure S2. Comparison of Foldseek and DALI significance scores. A. Scatterplot of all herpesvirus protein pairs scored with both Foldseek and DALI. B. Magnification of plot A for Foldseek E-values around 0.001. Given that an algebraic equation relating the E-value and Z-score is unclear, an approximation over the shown domain was used. Using a line of best fit (grey line), a DALI Z-score of 8.9 (horizontal red line) equates to the recommended significance threshold of a Foldseek E-value of 0.001 (vertical red line).

Clustering of structure predictions by similarity

The Foldseek and DALI outputs are a list of query and target proteins with the associated significance value. We used a threshold of E-value <0.001 for Foldseek and a Z-score >8 for DALI to determine the significant results. Both reciprocal comparisons for each protein pair needed to be significant for inclusion in the clustering. If a protein A was structurally similar to B and B similar to C, then A, B, and C were clustered. Once all clusters were identified, we converted the protein names to their row position in Table S2 Sheet 1. This system allowed us to quickly determine whether a structural cluster contained proteins expected to group, *i.e.* all row numbers are the same, or not, *i.e.* the row numbers are different, based on UniProtKB annotations[12] and previously known homologies[13].

The associated script can be found in our HerpesFolds GitHub repository in the file 'analyze_foldseek_for_structural_similarity_cluster.py'.

Analysis of sliding window structural similarity

To search for alterations at the domain level between protein predictions, we generated "structural snippets" from all predictions using a sliding window approach. We then ran these structural snippets in an all-against-all Foldseek search to identify conserved folds. We empirically tested which window size would result in the highest number of significant hits using Foldseek. Trying a window size of 25, 50, 100, 150, and 200 amino acids, we found the most significant hits at a window size of 150. The sliding window step size was tested at 5, 10, 15, 20, and 25 amino acids. We used a step size of 20 amino acids for the final analysis as it resulted in the maximal number of protein species detected. This approach resulted in over 12000 structural snippets of 150 residues in length, each overlapping by 130 residues with the preceding fragment and covering the whole folded human herpesvirus proteome. All structural fragments were compared pairwise with Foldseek, accounting for about 1.6 million tested combinations. We next assembled 'structural contigs' of overlapping frames that passed the Foldseek threshold of an E-value <0.001 and hit the same target. If a structural snippet matched another structural frame than it was extracted from, it was deemed a "hit". If several structural snippets overlapped, we used them to build larger domains. This analysis resulted in a list of structured areas, which we used to determine if a query protein has a different domain architecture than a structurally similar protein in this

data set. If a snippet of a query protein matched a different part of the query protein, we deemed it to have an "internal duplication". If a target snippet matched the query protein at multiple locations, we considered the query protein to have a "repetitive acquisition". If the query protein was structurally similar to a target protein but the query protein had at least 1 domain that was not found in the target protein, we deemed the protein to have a "domain addition". The associated script can be found in our HerpesFolds GitHub repository in the files 'make_sliding_window_pdb.py', 'Sliding_Window_Analysis.ipynb', 'analyze_foldseek_sliding_window.py' and 'analyze_sliding_window_domains.py'.

Keyword analysis of Foldseek results

Most Foldseek searches against cellular databases resulted in multiple significant matches, each containing protein descriptions and matching protein identifiers. We filtered for the most enriched keywords to extract likely biological functions from these descriptions. The following terms, in addition to single characters, single and double digits, and punctuation, were filtered out as they are too generic:

crystal, a, after, allosteric, allosteric, alpha, an, and, angstrom, angstroms, antibody, apo, as, at, atomic, based, basic, beta, between, bound, by, by, c1, c-c, cell, chain, chi, class, class, coli, complex, complexed, complexes., compound, conserved, cryoem, cryo-em, crystal, c-terminal, c-x-c, de, dehydration, deletion, delta, design, different, discovery, domain, e., e.coli, edition, element, engineered, epsilon, escherichia, eta, fab, for, form, fragment, from, full, full-length, functional, gamma, herpes, herpesvirus, high, holoenzyme, ii, iii, implications, in, inhibitor, inhibitors, ion, iota, its, kappa, lambda, ligand, ligand-binding, loader, loop, low-ph, mode, molecule, monoclonal, motif, mu, mutant, natural, neutralizing, novel, novo, nu, of, omega, omicron, on, one, open, open, opposite, peptide, phi, pi, probable, protein, psi, region, resolution, reveal, rho, sigma, significance, site, specificity, structural, structure, structures, substrate, symmetry, tau, template-primer, the, the, therapeutic, theta, to, type, uncharacterized, unit, upsilon, using, variant, virus, with, xfel, xi, x-ray, zeta

The used script can be found in our HerpesFolds GitHub repository in the files 'Keyword_Frequency_SP_PDB.ipynb'. This script was written with the assistance of a large language model.

RESULTS

ColabFold generates consistent models for the majority of proteins from the 9 human herpesviruses

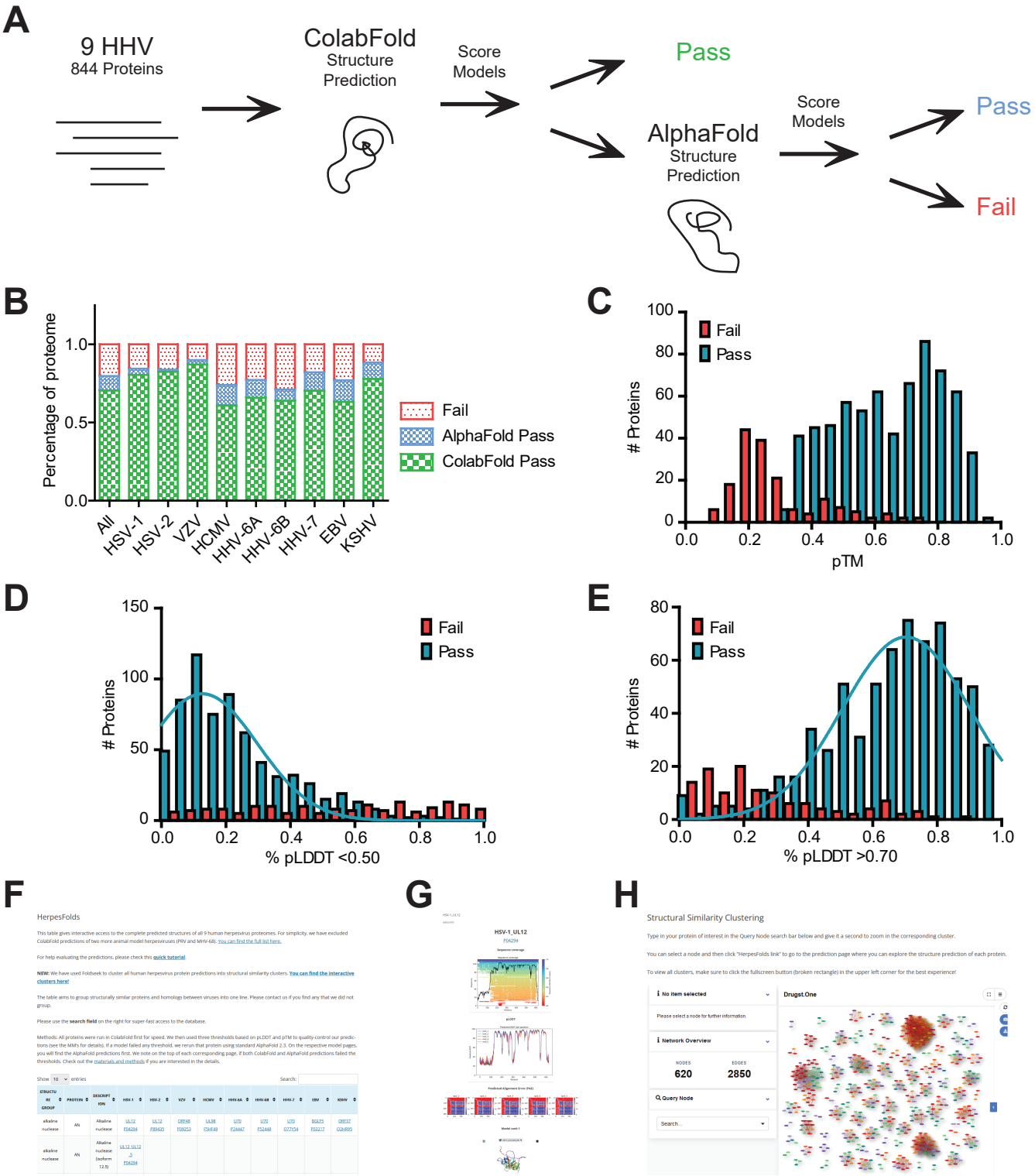
To generate proteome-wide predictions of prototypic strains of all 9 human herpesviruses (Table 1), we initially used ColabFold[9], an implementation of AlphaFold2 that mainly uses a 40-60x faster multiple sequence alignments (MSA) generation (Figure 1A). For each protein, 5 models were generated. For

most proteins, the associated predicted local distance difference test (pLDDT) plots of all five models were very similar. This indicated that the models converged in all five runs to a similar solution, indicating that the best solution was found (Figure S1A). To categorize the models objectively, we established three thresholds based on the pLDDT consistency between models and the predicted template modeling (pTM) scores. See the methods section for more details. Of the 844 proteins, 249 (30%) failed these quality scores (Figure 1B and Table S1 Sheet 1). We used standard AlphaFold[1] (Table S1 Sheet 2) to improve these predictions using the slower but more thorough jackhmmer MSA generation algorithm. This resulted in good predictions for 78 (31%) of the previously rejected models. Therefore, the final dataset consisted of 673 proteins (80% of 844 proteins) with consistent models as indicated by the pLDDT plot (Table S1 Sheet 3). We compared the proteins that failed and passed the quality scores for each threshold criterion (Figure 1C-E). Most proteins that did not pass had low pTM and a small percentage of sequence with a pLDDT >0.70. This correlation suggested that proteins low in folded regions, as indicated by a pLDDT>0.70, had lower prediction confidence. For the proteins not passing the pTM threshold, the pLDDT score can also predict whether a protein region is structured or intrinsically disordered. As shown before, a pLDDT>0.70 indicates a well-modeled region with a Cα root mean square displacement likely less than 1.5 Å[14]. Conversely, pLDDT scores <0.50 likely indicate intrinsically disordered regions (IDR)[1,15,16]. Consequently, the AlphaFold database has been used to predict protein disorder, such as in MobiDB[17]. IDRs have recently been implicated in essential processes such as replication compartment formation of HCMV by liquid-liquid phase separation[18]. Therefore, we determined which well-predicted proteins likely contain disordered regions by quantifying the percentage of pLDDT values each model has below 0.50 or above 0.70. As shown in Figure 1D-E, of the proteins that passed the quality scores, 69 (10%) predicted proteins likely contain a disordered region (cut-off >47% of total sequence), and 624 (93%) of proteins are likely structured (cut-off >31% of total sequence) (Table S1 Sheet 4). We found a fraction of 106 proteins that did not pass the pTM threshold but showed converging model pLDDT plots, indicating that the predictions came to a final solution. These proteins contained extensive sequence stretches with a pLDDT under 0.50, which indicated that these proteins are well-predicted, disordered proteins.

To illustrate the accuracy of the general prediction workflow, we compared our predictions to experimental data that were released after the AlphaFold 2.3 training cut-off of 2021-09-30. Our predictions aligned very well with experimental structures, as depicted for HCMV UL77[19] (Figure S3). The RMSD between the predicted and experimental HCMV UL77 structure (7NXP)[19] was 0.509, indicating a highly consistent model. These observations are in line with a previous community evolution that concluded that AlphaFold models are, on average, as good as experimental structures[3].

Table 1. Viral proteins used for structure prediction.

Subfamily	Virus	Virus abbreviation	Strain	Proteome ID or Reference Sequence	Number of proteins
alpha	herpes simplex virus 1	HSV-1	17	UP000009294	77
alpha	herpes simplex virus 2	HSV-2	HG52	UP000001874	74
alpha	varicella-zoster virus	VZV	Dumas	UP000002602	70
beta	human cytomegalovirus	HCMV	Merlin	UP000000938	181
beta	human herpesvirus 6A	HHV-6A	U1102	NC_001664	90
beta	human herpesvirus 6B	HHV-6B	Z29	NC_000898	97
beta	human herpesvirus 7	HHV-7	RK	NC_001716	84
gamma	Epstein-Barr virus	EBV	B95-8	UP000153037	87
gamma	Kaposi's sarcoma-associated herpesvirus	KSHV	GK18	UP000000942	86



STRUCTURE GROUP	PROTEIN	DESCRIPTION	HSV-1	HSV-2	VZV	HCMV	HHV-6A	HHV-6B	HHV-7	EBV	KSHV
alkaline nuclease	AN	Alkaline nuclease	UL12 P04294	UL12 P89435	ORF48 P09253	UL98 P5HF49	U70 P24447	U70 P52448	U70 Q77154	BGLF5 P03217	ORF37 Q2HR95
alkaline nuclease	AN	Alkaline nuclease (isoform 12.5)	UL12, UL12.5 P04294								

Figure 1. Structure predictions of the proteome of all human herpesviruses. A. Pipeline of model generation and quality scoring. The structures of 844 proteins covering the proteomes of all nine human herpesviruses were initially predicted with ColabFold. After scoring, models that failed any quality score were rerun with AlphaFold. B. Percentage of each viral proteome that passed the quality score. C. Histogram of the pTM scores of the proteins that passed or failed the quality scores. D-E. Histogram of the percentage of each protein with a pLDDT score below 0.50 (D) or above 0.70 (E). In both cases, a Gaussian distribution was fit to identify the distribution for the proteins that passed the quality scores. F. HerpesFolds online database available at www.bosse-lab.org/herpesfolds/. Structure predictions are organized and searchable by homology and structural similarity. G. Example webpage of the displayed information for each predicted model. H. Searchable clustering of all human herpesvirus proteins based on structural similarity.

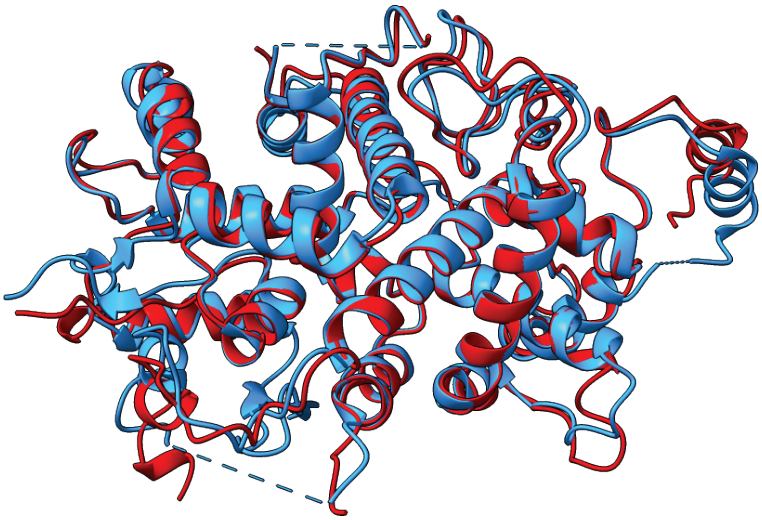


Figure S3. Alignment of predicted HCMV UL77 structure to the experimental structure that was deposited post-AlphaFold training. The deposited structure for HCMV UL77 (PDB 7NXP) (blue) was aligned with the AlphaFold prediction (red), and the comparison has an RMSD of 0.509. Only the residues present in the deposited structure are shown in the predicted model. The corresponding UCSF ChimeraX session can be found at https://github.com/QuantitativeVirology/Herpesfolds/.

HerpesFolds is an easily searchable database of all human herpesvirus protein predictions

To make our proteome-wide predictions accessible and easily searchable, we created a database at www.bosse-lab.org/herpesfolds/ (Figure 1F). The main table contains the individual herpesvirus protein names and their UniProtKB accession numbers. Based on the previous knowledge[12,13], we generated a chart where each row contains proteins with known homology (Table S2 Sheet 1). This way, the database also serves as a reference guide. This feature is important as the nomenclature of human herpesviruses is highly inconsistent and difficult to interpret even for experts. We also integrated the common names based on the UniProtKB annotation. Moreover, each protein entry links to the UniProtKB webpage and an individual webpage containing the 5 animated models and the pLDDT and PAE scores (Figure 1G). If predictions were rerun with AlphaFold, these predictions are included. If the models of both runs failed the thresholds, a warning is displayed at the top of the page. As an important feature, we also integrated a Foldseek[5] search link that directly submits the predicted model to the Foldseek server and displays structurally similar proteins. Finally, we interactively display the structural similarity network described below using Drugst.One[20] (Figure 1H).

Identification of novel clusters, unknown relationships, and novel architectures

We employed structural similarity searches to identify structurally similar proteins that might be homologs between different viral species and determine the potential function of orphan genes. We used both DALI[10] as an established algorithm and the much newer Foldseek[5] algorithm and ran all models against each other. We used comparable thresholds of E-value <0.001 for Foldseek and a DALI Z-score of >8 (see materials and methods) (Table S2 Sheet 2 and 3, respectively). Overall, the resulting structural clusters were similar, with 596 proteins in protein similarity clusters for both search algorithms. 24 proteins were only in clusters generated from significant Foldseek results, while 6 were only included in clusters identified by DALI. One example of a difference is the tegument protein 3 (TEG3) group. Foldseek grouped only the alphaherpesvirus TEG3 proteins with the gammaherpesvirus ORF35 proteins. In contrast, DALI also found a similarity of the roseolovirus (HHV-6A, HHV-6B, and HHV-7) U68 proteins with this group. Since Foldseek performed slightly better overall and is computationally much more efficient, we used it for the remainder of this work. The resulting structural similarity connections were used to cluster the herpesvirus proteins into groups of structural similarity (Figure 2A).

All clusters are available for interactive exploration at www.bosse-lab.org/herpes-network/ and can be viewed in full screen, which aids in exploring the relationships.

Based on the UniProtKB[12] and previous work[4,13], 39 protein groups were expected to be found in all 9 species. A more recent sequence analysis based on domain architecture identified 23 proteins with a consistent domain structure in all 9 species and 169 homology groups[4]. Using structural similarity searches, we identified 23 groups with exactly 1 protein from each of the human herpesviruses and 89 homology groups in total. Many of these clusters matched these previous attempts to categorize herpesvirus proteins in groups based on experimental data and sequence analysis, confirming our approach. An example are the homologs of the alphaherpesvirus UL25 proteins, which are essential capsid-associated proteins and part of the portal cap[19]. We also found clusters of proteins with known and verified enzymatic functions, such as the herpesvirus serine/threonine kinases, named after the prototypic HSV-1 US3 and UL13 proteins (Figure 2B). As only the protein models that passed the quality scores were used, not all groups with exactly one protein from each species were found. For example, HHV-7 U57 is missing from the MCP group because it was deemed a low-confidence model.

Surprisingly, the DNA polymerase catalytic subunit and DNA helicase/primase complex-associated protein exhibit similar folds (Figure 2C). While these proteins have different biological functions, they have some sequence similarity[21]. Furthermore, they are similar to nucleases, with HSV-1 UL8 being homologous to the exonuclease domain of B-family DNA polymerases[22] and HSV-1 UL30 having RNAse H activity[23].

The small capsomere-interacting protein (SCP) is an essential gene in all herpesviruses. However, they clustered as distinct groups based on the subfamily (HSV-2 UL35 is missing as it failed the quality scores) (Figure 2D). Genetic drift may have led to structurally distinct features, as the last common ancestor likely predates the divergence of the subfamilies.

A particularly interesting cluster of so far-unannotated proteins consisted of the likely homologs of HCMV UL112-113. HCMV UL112-113 is essential for viral replication by mediating replication compartment formation through phase separation at viral genomes, thereby recruiting the viral polymerase[18]. The poorly characterized roseolovirus U79 proteins and the alphaherpesvirus UL4/ORF56 proteins clustered with UL112-113. They shared an undescribed and conserved predicted beta-barrel (Figure 3A). These herpesviruses also form replication compartments, and their potential UL112-113 homologs would be prime candidates for further investigation.

Another novel cluster contained the gammaherpesvirus KSHV ORF73/LANA and EBV EBNA1 proteins that harbor a known DNA binding domain[24,25]. Our structure predictions indicate that this DNA binding domain is not only conserved in all betaherpesvirus IE2 homologs (HCMV UL122, roseolovirus U86 proteins) but also in HCMV UL117 and its homologous U84 proteins in the roseoloviruses (Figure 3B). Excitingly, we also found this domain in the alphaherpesvirus UL3/ORF58. While LANA[26], EBNA1[27], the IE2[28,29] homologs, as well as the U84 proteins are involved in transcriptional regulation, the

UL3/ORF58 proteins might not directly regulate transcription[30] but interact with ICP22 to function in nuclear compartments[31].

The HCMV phosphoprotein UL25 formed a group with its roseolovirus U14 homologs and clustered with the HCMV UL35 proteins, which likely arose via gene duplication, constituting the pp85 superfamily[32]. Strikingly, we also found HCMV UL32 and its roseolovirus U11 homologs to be structurally similar. UL32, also known as pp150, is a protein conserved in the betaherpesviruses. It binds directly to capsids as an inner tegument protein while bridging the full virion tegument via its disordered C-terminus[33]. It will be essential to dissect if the structurally similar UL25 and UL35 can also bind capsids and potentially regulate pp150 abundance at pentons, potentially confusing the reconstruction of virus capsid cryo-EM maps. Finally, the poorly studied KSHV ORF48 and likely the EBV BBRF2 protein, excluded from the clustering due to low prediction scores, also clustered with the pp150 homologs. ORF48 is part of virions[34], and it should be assessed if it can bind to the viral capsid as pp150 does.

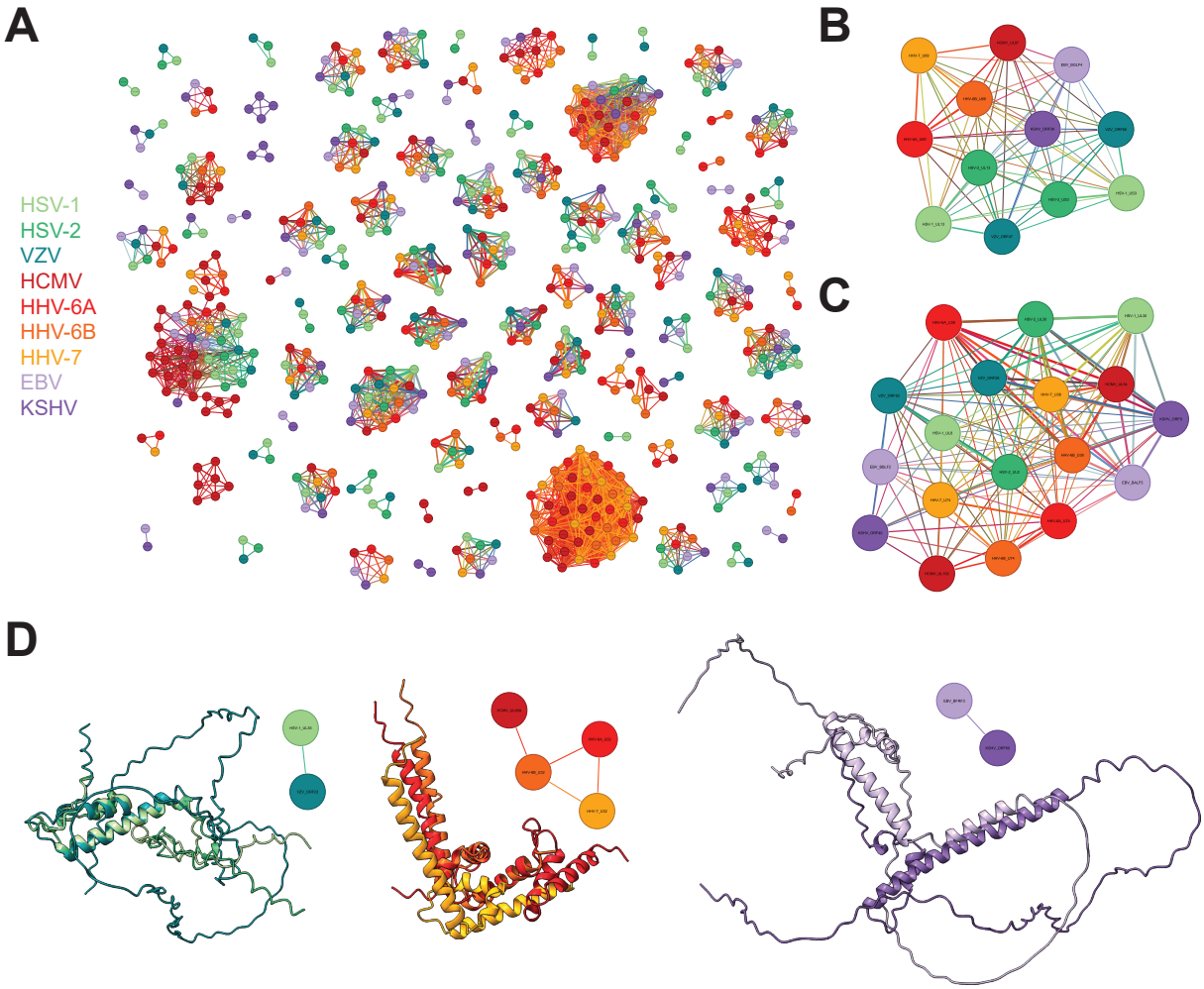

A
B
C
D
HSV-1
HSV-2
VZV
HCMV
HHV-6A
HHV-6B
HHV-7
EBV
KSHV

Figure 2. Structural similarity search identifies protein clusters. A. Graphical display of all structurally similar protein clusters. The 673 herpes proteins that passed the quality score were analyzed pairwise by Foldseek for structural similarity. The color scheme for the viral species is used throughout this work. B. Cluster of viral serine/threonine-protein kinases US3 and UL13. C. Cluster of viral DNA helicase/primase complex-associated protein (HEPA) and DNA polymerase catalytic subunit (DPOL). D. Clusters of small capsomere-interacting protein (SCP). The alpha-, beta-, and gammaherpesviruses clustered separately. The corresponding UCSF ChimeraX session can be found at https://github.com/QuantitativeVirology/Herpesfolds/.

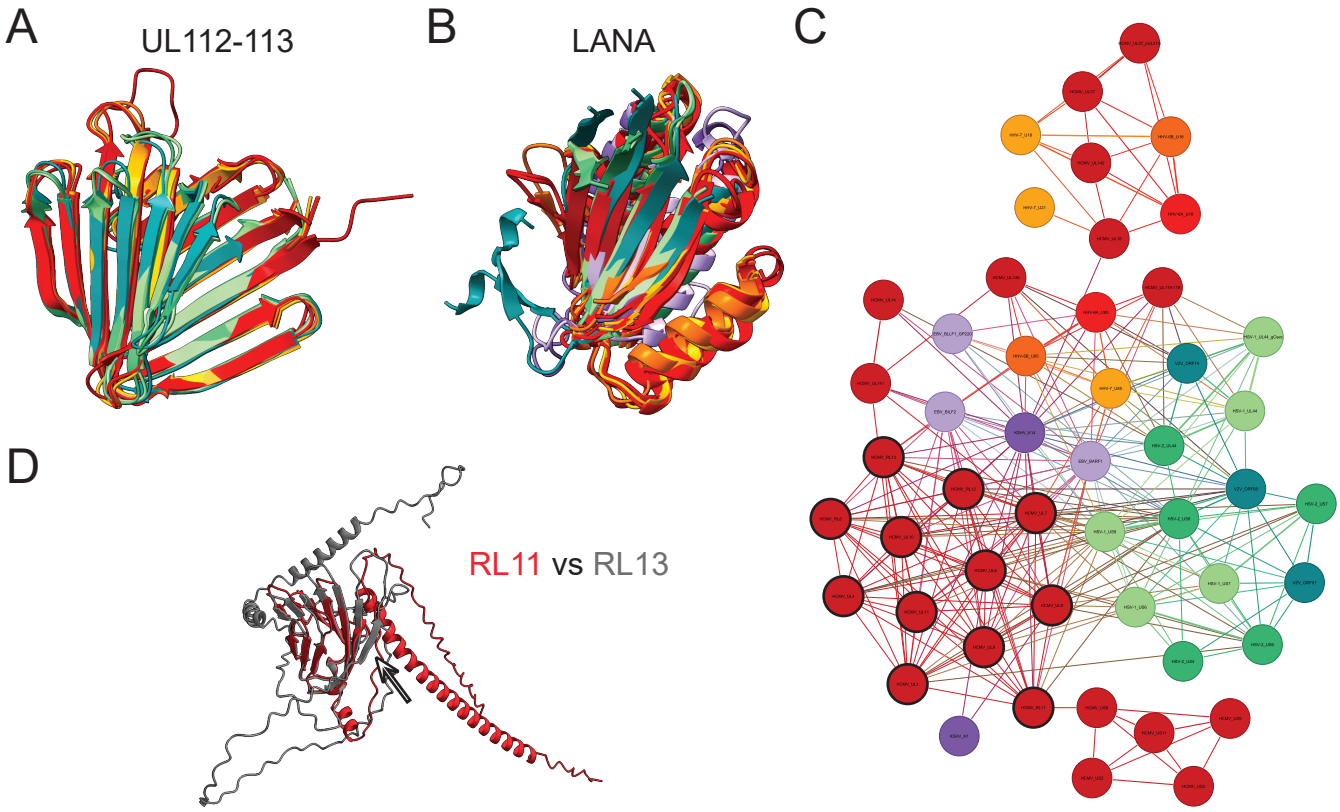

A
UL112-113
B
LANA
C
D
RL11 vs RL13

Figure 3. Structural clustering identifies novel and differently grouped protein groups. A. Structural alignment of the HCMV UL112-113 cluster. Only the conserved beta-barrel-like domain is shown for clarity of HCMV UL112-113, HHV-6A U79_U80, HHV-6B U79, HHV-7 U79, HSV-1 UL4, HSV-2 UL4, and VZV ORF56. B. Structural alignment of the LANA DNA binding domain-containing cluster with only the matching domain shown for clarity of KSHV ORF73/LANA, EBV EBNA1, HHV-6A U84, HCMV UL117, HCMV UL122, HCMV UL123, HHV-6A U86, HHV-6A U90_U87U86, HHV-6B U84, HHV-6B U90_U86, HHV-7 U84, HHV-7 U86, HSV-1 UL3, HSV-2 UL3, and VZV ORF58. C. Ig-like domain-containing cluster harboring the HCMV RL11 protein family. The canonical members are marked with a black outline. D. Structural alignment of RL11 with RL13. The arrow points to the additional beta strand. The corresponding UCSF ChimeraX sessions can be found at https://github.com/QuantitativeVirology/Herpesfolds/.

Immunoglobulin-like domains are common structural elements in herpesvirus proteins

One of the largest clusters we identified consists of viral proteins that contain immunoglobulin (Ig)-like domains. This cluster merged several previously defined protein families. One is the HCMV RL11 family[35]. They share an RL11 domain (RL11D), formed by a conserved tryptophan and two cysteines, and are structurally similar to immunoglobulin domains. They are non-essential in cell culture[36] but have high sequence variability between isolates with some playing a role in modulating immune responses such as UL11[37]. While there are 14 accepted members of this family, they did not form an isolated structural group in our clustering (Figure 3C). Unexpectedly, EBV BILF2 and KSHV K14 but also the well-studied HSV-1 glycoproteins gC/UL44, gD/US6, gI/US7, and gE/US8 were structurally similar to multiple members of this family. EBV BILF2 is encoded in a genome region unique to *lymphocryptoviruses*, and KSHV K14 is also unique to KSHV. Furthermore, HSV-1 and HSV-2 US8 are located in a different part of the genome than HCMV. In addition to the low sequence identity, these data suggest that these Ig-like domain-containing proteins were acquired independently rather than from a common ancestor. Furthermore, gene duplication continued after speciation. While most RL11 members have the same 6-strand antiparallel beta-fold, RL13 has an additional beta-strand (Figure 3D), potentially representing a distinct evolutionary branch and function.

Domain-level structural similarity searches reveal domain duplications and deletions

We performed the initial structural similarity search with full-length proteins. While this will identify commonalities between proteins, it will, by definition, not identify differences. To address how the domain architecture of a protein can differ between species, we performed a structural similarity search with overlapping structural snippets using a sliding window approach (see materials and methods) (Table S2 Sheet 4). We identified 2758 structurally similar connections between proteins, which

includes new connections for 107 proteins relative to the full-length analysis (Table S3 Sheet 1 and 2). This analysis identified 23 internal duplications, where domains were likely duplicated in the same protein, repetitive acquisitions, where domains of a different protein matched the query protein multiple times at different positions, and domain additions, where multiple domains of a protein matched different proteins (Figure 4A and Table S3 Sheet 3).

We found that HSV-1 UL15 was structurally similar to the previously identified HSV-2 UL15 but also HSV-1 UL9 (Figure 4B). UL15 is part of the tripartite terminase, essential for packaging viral genomes into newly formed capsids, and UL9 is the replication origin-binding protein. The fragment that aligns is oriented in the opposite direction relative to the core in UL15 versus UL9, which suggests that this region is not involved in an external interaction and has a conserved structural integrity function.

An important example of domain duplications is in the large betaherpesvirus US22 family cluster. No experimental structure of this large family of proteins is available, and sequence-based methods indicate that the family consists of at least 12 members in HCMV[38,39]. Using full-length structural similarity searches, we identified three HCMV proteins, UL26, its isoform UL26.21, and UL27, and UL38, which share the same conserved fold (Figure 4C). Importantly, this group consists of four different domain architectures depending on whether the protein contains 0, 1, 2, or 4 copies of the US22 domain (Figure 4D). The majority of proteins are similar to US22 with two domains. The list of members with only one domain, which includes the newly identified proteins: HCMV UL26, HCMV UL26_UL26.21, HCMV UL27, HCMV UL38, HHV-6A U19, HHV-6B U19, and HHV 7 U19. The proteins without a US22 domain are HHV-6A U4, HHV-6B U4, and HHV-7 U4. They are connected to this group through the HHV-6A U7 and HHV-7 U7 proteins, which appear to have an additional domain relative to US22, which the U4 proteins consist of exclusively. The proteins with four copies of the US22 domain are HCMV UL29, HHV-6A U7, and HHV-7 U7. Oddly, HHV-6B U7, unlike the other roseoloviruses, only has two domains, suggesting a divergence in this species. Furthermore, many group members were identified to contain an internal duplication and/or repetitive acquisition. The low RMSDs between the two domains of US22 and between the domains from different proteins are consistent with a domain duplication (Figure 4D). These data suggest that the US22 family is bigger than previously thought and that domain additions and deletions are particularly common in this family.

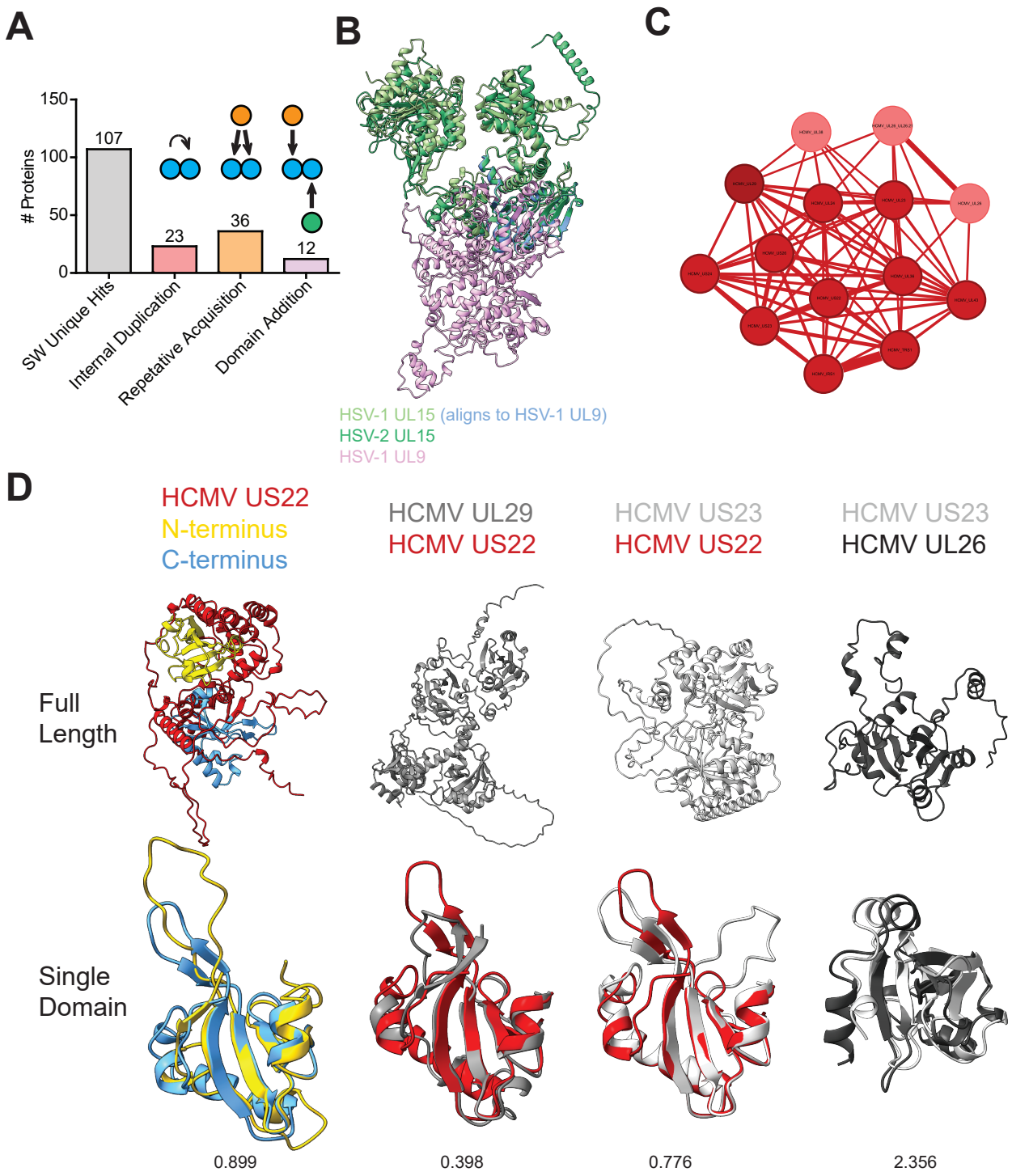

A
150
100
50
0
Proteins
107
23
36
12
SW Unique Hits
Internal Duplication
Repetative Acquisition
Domain Addition
B
HSV-1 UL15 (aligns to HSV-1 UL9)
HSV-2 UL15
HSV-1 UL9
C
D
HCMV US22
N-terminus
C-terminus
HCMV UL29
HCMV US22
HCMV US23
HCMV US22
HCMV US23
HCMV UL26
Full Length
Single Domain
0.899
0.398
0.776
2.356

Figure 4. Domain-level similarity search identifies conserved domain duplications. A. Domain-level similarities identified using a sliding window approach. The "SW Unique Hits" are the number of proteins that had a structurally similar query-target pair in the sliding window analysis that was absent from the full-length analysis. "Internal Duplications" are proteins where parts of the protein are similar to another part of itself. "Repetitive Acquisition" are proteins where a piece of a different protein matches the query protein multiple times at different positions. "Domain Addition" refers to proteins with multiple domains matching different proteins. Illustrative cartoons are shown above the corresponding bar. B. HSV-1 UL15 is aligned to the known homolog HSV-2 UL15 and the structurally similar target HSV-1 UL9, which was only identified in the domain-level search. The part of HSV-1 UL15 that aligns with HSV-1 UL9 is colored blue. C. The cluster of HCMV proteins that contain HCMV US22. The proteins that contain 2 domains, like the canonical US22, are shown in the HCMV color. The proteins with 1 domain are shown in light red, and the protein that has 4 domains is shown in dark red. D. The tertiary structure of the core domain is similar between the proteins in this group. The top row is the full-length protein, the bottom row is a structural alignment of the core domain, and the RMSD of the core domain alignment is written below. HCMV US22 has 2 domains, shown in yellow and blue, which are aligned. HCMV UL29 has 4 domains and was aligned to US22. HCMV US22 was aligned to HCMV US23, which has 2 domains, and HCMV US23 was aligned to HCMV UL26, which has 1 domain. In the case of multiple domains, the best-fitting domain combination is shown. The corresponding UCSF ChimeraX sessions can be found at https://github.com/QuantitativeVirology/Herpesfolds/.

Structural similarity search of human herpesvirus protein models against cellular proteins predicts viral functions

Many of the identified clusters contained understudied proteins. We expanded the structural similarity search to cellular proteins using Foldseek to determine their potential biological activity from the PDB database (Table S4 Sheet 1), and the predicted AlphaFold structures of the Swiss-Prot database (Table S4 Sheet 3). The PDB contains experimentally verified structures, but it is limited in diversity. In contrast, the AlphaFold predictions of the Swiss-Prot database enable comparisons to a much more extensive predicted database less biased by experimental limitations. We combined the significant hits from the two databases and identified structurally similar cellular proteins. To find putative functions for the viral protein groups, the functional description of the homologous proteins was mined for keywords (Table S4 Sheets 2 and 4). Unfortunately, the depositors' descriptions of structures in the PDB database are not systematic. To handle these variations, we calculated the frequencies of all words in the descriptions, and the words with the highest frequencies that likely represent the protein function were manually chosen as the annotation. Each viral homology group identified above was then given a HerpesFolds annotation. Since the PDB, but not the AlphaFold database, contains structures of viral proteins, it is important to note that some hits in the PDB database are because they were aligning with themselves.

Some predicted viral protein structures did not map to any other protein fold in the searched databases, indicating that these are unique folds (Table S4 Sheet 5). We found that the viral alkaline nucleases had a significantly different fold to cellular nucleases (Figure 5A). A striking example was the cytoplasmic envelopment protein 2, which is conserved in all herpesviruses and for which no structure exists (Figure 5A). These proteins are crucially involved in virion envelopment[40], and it will be important to analyze their role in modulating the cellular secretory system through a unique fold.

We could annotate several poorly studied viral proteins with potential functions using this workflow. One example is HSV-1 US2, structurally similar to dihydrofolate reductases (DHFRs) (Figure 5B-D). The most significant cellular target was the DHFR from *Mesocricetus auratus* (commonly known as the golden or Syrian hamster) (PDB 3EIG). Furthermore, HSV-1 US2 was structurally similar to KSHV ORF2, a known DHFR homolog with enzymatic activity[41,42]. While HSV-1 and HSV-2 US2 have low sequence identity to the *M. auratus* DHFR, they have a fold similar to viral and cellular DHFRs. Alignment to the *Escherichia coli* DHFR identified potential residues in the active site (Figure S4)[43]. No DHFR activity has been reported for the US2 proteins, and thus, functional assays will be needed to determine if the US2s still have DHFR activity or if this fold was exaptated.

We also identified a novel cluster containing EBV BMRF2, KSHV ORF58, and HSV-1 UL43 (with HSV-2 UL43 being absent due to low-quality scores). Importantly, EBV BMRF2, KSHV ORF58, and VZV ORF15 predictions were significantly similar to the human equilibrative nucleoside transporter (ENT) 4 (Figure 5E-G). This cellular protein is involved in transporting metabolites and regulating adenosine concentrations[44]. However, no experimental data currently supports a role for any of these viral proteins in transmembrane transport. Instead, EBV BMRF2 can change cellular morphology by modulating the actin cytoskeleton[45] and is essential for efficient cell-cell spread[46]. UL43 colocalizes with proteins involved in membrane fusion and potentially interacts with the gN/gM complex[47], which is engaged in syncytia formation[48]. These commonalities suggest a function in cell-cell spread for this protein cluster, which does not align with the function of BMRF2 and UL43 as a transmembrane transporter. Moreover, comparative cross-sections of the viral proteins and ENT4 illustrate that they do not have a channel necessary for metabolite transport[49] (Figure S5A). Several transporters depend on accessory proteins for correct trafficking to the plasma membrane, such as the lysosomal transporter MFSD1, which depends on GLMP[50]. This is also true for EBV BMRF2, which needs BDLF2[45] for correct trafficking to the plasma membrane, and KSHV ORF58, which needs ORF27[34]. Interestingly, AlphaFold predictions of the viral heterodimers suggest that even if channels were present in BMRF2 and ORF58, they would be blocked by the partner protein (Figure S5B). These data indicate that the similarity to ENT4 may stem from "radical" exaptation, where the transporter activity was lost and replaced by a role in viral cell-to-cell spread.

All human herpesviruses encode for deoxyuridine triphosphatases (dUTPases)[51] called HSV-1 UL50, HSV-2 UL50, VZV ORF8, HCMV UL72, U45 of roseoloviruses, KSHV ORF54 and EBV BLLF3. Cellular dUTPases are homotrimers that form three active centers at their interfaces. In contrast, the human herpesvirus dUTPases are monomers consisting of a fusion of two dUTPase domains, as shown for EBV BLLF3[52]. While the respective alpha- and gammaherpesvirus proteins still have enzymatic activity,

the betaherpesvirus homologs have lost them and are involved in immune modulation[53]. Human beta- and gammaherpesviruses code for 17 more dUTPase-related proteins (DURPs), likely due to gene duplications[51]. Using domain-level search, we found that all DURPs were predicted to harbor a three-domain architecture in contrast to the two domains of the viral dUTPases. However, the separate domains diverged quite significantly from the canonical dUTPase fold as indicated by their RMSD (Figure 5H), which is in line with a recent report on the crystal structure of the HCMV DURP UL82 (pp71), which describes that it has lost its dUTPase function[54].

Herpesviruses have been suggested to induce autoimmunity through molecular mimicry[55,56]. Several studies have attempted to identify viral proteins that might induce autoimmunity due to a similarity to a cellular protein[57,58]. Protein sequence-based analyses could not identify viral proteins that would induce autoimmunity to known cellular proteins[57], such as myelin, in multiple sclerosis[58]. Our structural similarity search identified no direct structural match of myelin to a viral protein. Interestingly, multiple members of the immunoglobulin domain-containing group, such as HSV-1 US8, mapped to myelin protein zero-like protein 1, and others, such as EBV BARF1, mapped to myelin-associated glycoprotein. These myelin-related proteins could be involved and warrant further investigation.

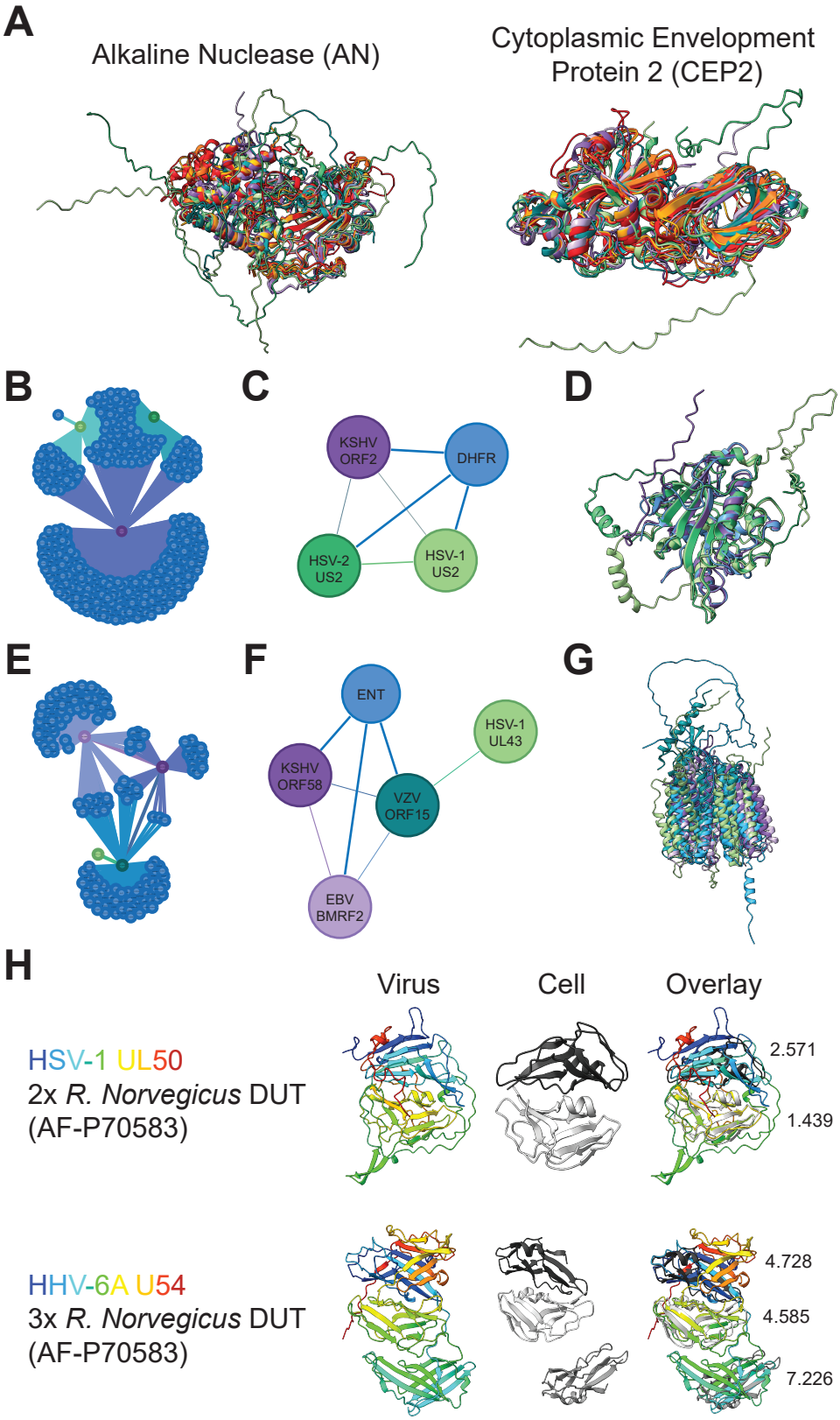

A
Alkaline Nuclease (AN)
Cytoplasmic Envelopment Protein 2 (CEP2)
B
C
KSHV ORF2
DHFR
HSV-2 US2
HSV-1 US2
D
E
F
ENT
HSV-1 UL43
KSHV ORF58
VZV ORF15
EBV BMRF2
G
H
Virus
Cell
Overlay
HSV-1 UL50
2x R. Norvegicus DUT
(AF-P70583)
2.571
1.439
HHV-6A U54
3x R. Norvegicus DUT
(AF-P70583)
4.728
4.585
7.226

Figure 5. Annotation of viral protein functions from structural similarity search. A. Structural alignment of the alkaline nuclease (AN) and cytoplasmic envelopment protein 2 (CEP2) from all human herpesviruses. CEP2 was found in all human herpesviruses but had no significant cellular hit. B. Cluster of KSHV ORF2, HSV-1 US2, and HSV-2 US2 containing the cellular hits. Each cellular node (blue) represents a specific protein. C. Summary of B for clarity. D. Structural alignment of the proteins in C. P04753 dihydrofolate reductase (DHFR) was the most significant cellular hit. E. Cluster of EBV BMRF2, KSHV ORF58, HSV-1 UL43, and VZV ORF15 containing the cellular hits. Each cellular node (blue) represents a specific protein. F. Summary of E for clarity. G. Structural alignment of the proteins in F. The most significant cellular hit was Q8R139 equilibrative nucleoside transporter 4 (ENT4). H. Structural architecture of the classical DUT HSV-1 UL50 and the DURP HHV-6A U54. The rainbow-colored viral protein illustrates how the polypeptide folds back on itself. The most significant cellular hit, AF-P70583 DUT from *Rattus norvegicus*, is shown multiple times so that it can be aligned to each domain in the viral protein. The RMSD for each alignment is shown on the right of the overlay. The disordered termini were hidden for clarity. The corresponding UCSF ChimeraX sessions can be found at https://github.com/QuantitativeVirology/Herpesfolds/.

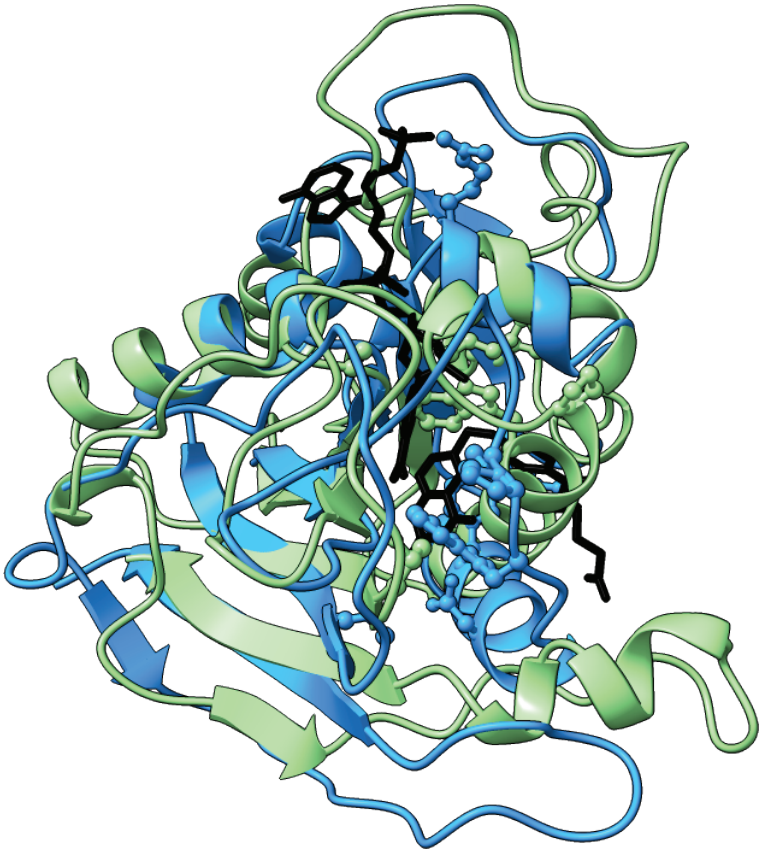


Figure S4. Alignment of HSV-1 US2 to the E. coli DHFR identifies potential residues involved in the active site. The E. coli DHFR (PDB 1RX2) (blue) was aligned to HSV-1 US2 (green), and the cofactors folic acid and nicotinamide-adenine-dinucleotide phosphate are shown in black. The atomic structure of the catalytic residues of the E. coli DHFR (M20, P21, W22, D27, F31, R44, R57, G95, G96, T113) and the potential catalytic residues that are in similar positions in US2 (P27, W30, R52, R71, G128, A129, T154) are shown. The corresponding UCSF ChimeraX session can be found at https://github.com/QuantitativeVirology/Herpesfolds/.

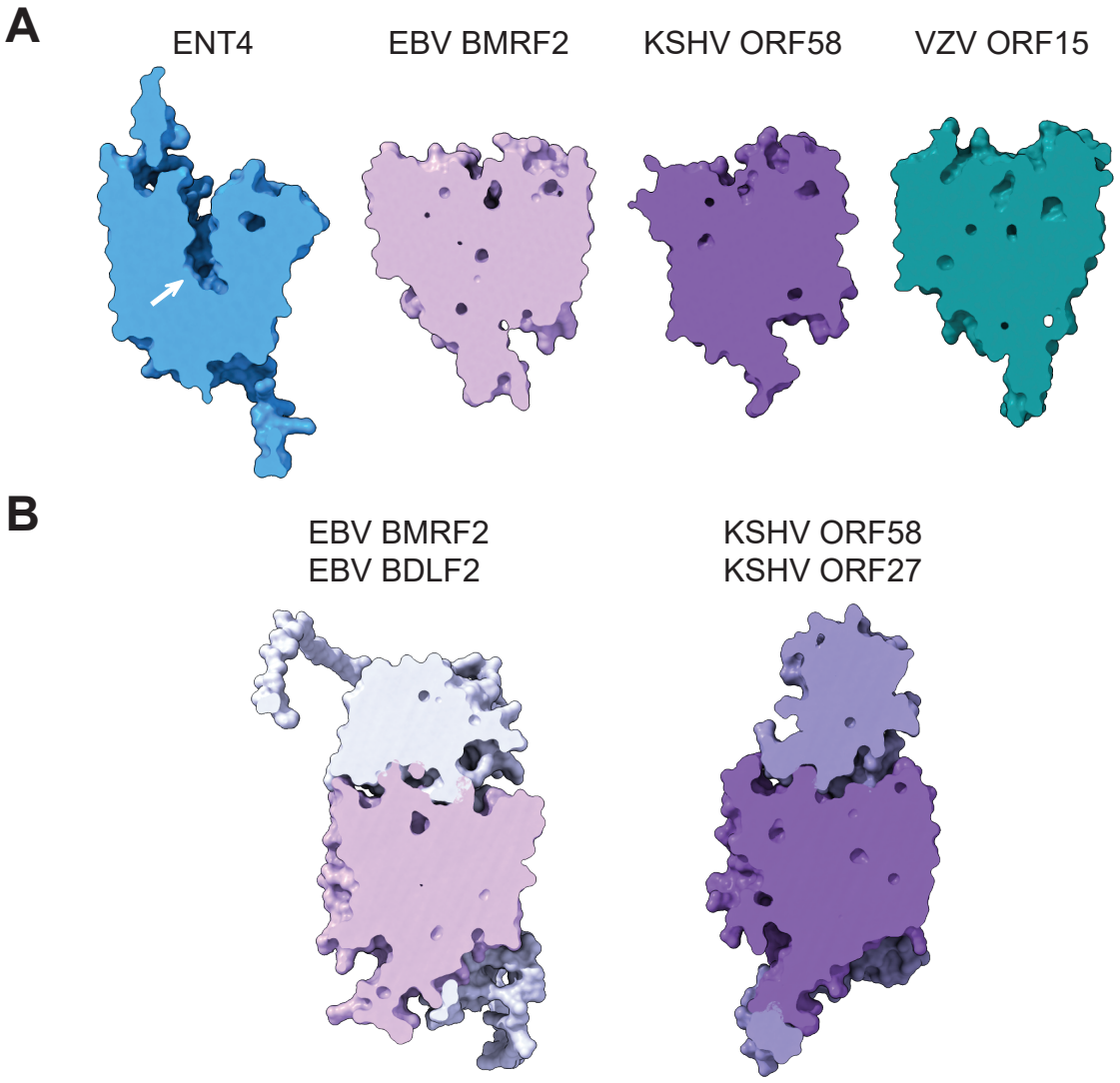

A
ENT4
EBV BMRF2
KSHV ORF58
VZV ORF15
B
EBV BMRF2
EBV BDLF2
KSHV ORF58
KSHV ORF27

Figure S5. The herpesvirus ENTs are unlikely to have transporter activity. A. Cross-sections of surface plots of the monomer predictions of ENT4 and the structurally similar viral proteins. The white arrow points to the metabolite channel in ENT4. B. Structural predictions of the known EBV (BMRF2 pink and BDLF2 white) and KSHV (ORF58 purple and ORF27 light purple) heterodimers. The corresponding UCSF ChimeraX sessions can be found at https://github.com/QuantitativeVirology/Herpesfolds/.

DISCUSSION

Understanding protein structures is critical to comprehending pathogen-host interactions. Unfortunately, experimental structural data is lacking for most human herpesvirus proteins. Machine learning-based structure prediction tools have made huge strides in accuracy and provide stringent quality scores for their predictions[1,2,9]. Accordingly, recent community evaluations have found "that AF2 [AlphaFold2]-predicted structures, on average, tend to give results that are as good as those derived from experimental structures"[3] while also cautioning that "AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination" [14].

The human herpesviruses encode for at least 844 proteins. For most, no experimental structure has been solved. Here, we applied proteome-wide predictions to provide a structural systems view of all annotated human herpesvirus proteins and developed stringent scoring workflows to validate our predictions. We generated clusters of proteins with similar folds using all-against-all full-length structural similarity searches. This approach confirmed previously known clusters, such as the viral G-protein coupled receptors and kinases, and even special cases, such as the gLs clustering within their corresponding sub-family[4], underlining its validity. Importantly, we identified previously unknown protein clusters sharing conserved folds, such as the UL112-113 family that carries a novel beta-barrel-like fold and likely plays an essential role in genome replication.

Still, structural clustering might not always indicate direct evolutionary relatedness. The large Ig-like domain cluster appears to be a group of proteins with a common fold, likely the result of independent acquisitions rather than originating from a common ancestor or gene duplication. Using domain-level searches, we identified differences between structurally related proteins such as viral dUTPases and the US22 family. Both consist of subgroups with different numbers of duplicated domains that correlate with their described activity. Finally, we searched for cellular proteins with matching folds. We identified cases of potential exaptation, such as the membrane transporter ENT-4 by human alpha-and gammaherpesviruses and unique folds that do not match any cellular protein, such as the CEP2 proteins.

This body of data provides many valuable hypotheses for experimental validation. To accelerate the use of this data and research on human herpesvirus, we have generated an easy-to-use, searchable

web interface for the community, accessible at www.bosse-lab.org/herpesfolds/. The database groups proteins by structural relationship and allows interactive evaluation of all structures. PDB files of each prediction can be downloaded or directly submitted to the Foldseek structural similarity search server to find similar cellular proteins in up-to-date databases.

This work is a first step based on high-confidence monomeric predictions of viral proteins. It neglects that some proteins may fold differently in complex with others, adopt multiple conformations, or need to form oligomers to function, such as gB[59]. Future work will need to take into account protein complex composition and stoichiometry. However, accomplishing this task is challenging as this information is unavailable for most viral proteins, and generating all likely and possible combinations is computationally highly demanding. Furthermore, posttranslational modifications will likely play a role in many such complexes. Still, protein complex prediction software is advancing, and the next logical step will be predicting viral protein-protein interaction networks. Moreover, predictions of protein-ligand complexes[60,61] or nucleic acids[62] are now within reach.

The methodology presented here applies to any virus or virus family. It can easily be extended to analyze all reading frames of a given organism, not only the annotated ORFs. As the databases of predicted and experimental viral protein structures grow, more connections between different viral pathogens and their hosts will be made. The speed at which these *in silico* models can be generated will drive rapid hypothesis design and subsequent experimental confirmation.

SUPPLEMENTARY TABLES

Table S1. Consistency of pLDDT scores of AlphaFold models. The quality scores for ColabFold are in Sheet 1, and AlphaFold are in Sheet 2. Sheet 3 lists all proteins and details whether they passed the quality scores. Sheet 4 lists the percentages of each protein sequence with a pLDDT of <0.50 (likely disordered) and >0.70 (likely structured).

Table S2. Identified herpesvirus structural similarity clusters. Sheet 1 lists the currently known homologous proteins between herpesvirus species and the Foldseek outputs. Sheet 2 lists the newly identified structural similarity groups based on Foldseek analysis with reference to the previously known homolog groups shown in Sheet 1. The identified function based on structural similarity to cellular structures is annotated. Sheet 3 is the same as Sheet 2, except that DALI outputs were used for clustering by structural similarity. Sheet 4 lists the structural clusters based on domain-level search instead of full-length protein search. The difference to the full length analysis is annotated.

Table S3. Results of the domain-level structural similarity search. Sheet 1 lists all pairwise similarities found using Foldseek for full-length similarity searches (columns A-B) and domain-level searches using structural snippets generated by the sliding window approach (columns C-D) for comparison. Pairs only identified in the domain-level search are listed again (columns E-F). Sheet 2 is a list of the sequence ranges of query proteins that match the sequence ranges of a target protein in the domain-level search. Sheet 3 lists all domain-level search results as well as a categorization of whether the hits can be categorized into three categories: "Internal Duplication" (the query protein matched itself), "Repetitive Acquisition" (the query protein matched a target sequence multiple times), or "Domain Addition" (the query protein has a domain that is not found in a related protein).

Table S4. Structural similarity search against cellular proteins. Sheet 1 lists the Foldseek search results of the viral protein structural predictions against the PDB database structural similarity search. Sheet 2 lists the frequency of each word in the description of the hits from sheet 1 for each protein domain (column B). Viral keywords were subtracted to remove known functions and self-hits from experimental viral structures in the PDB (column C). Sheet 3 lists the Foldseek structural similarity search results of the viral protein structural predictions against the AlphaFold-Swissprot. Sheet 4 lists the frequency of each word in the description of the hits from sheet 3 for each protein domain (column B). Viral keywords were subtracted to remove known functions and self-hits to experimental viral structures in the PDB (column C).

AUTHOR CONTRIBUTIONS

Timothy K. Soh: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

Sofia Ognibene: Data curation, Software

Saskia Sanders: Data curation, Software

Benedikt Kaufer: Data curation, Writing – original draft, Writing – review & editing, Funding acquisition, Project administration,

Jens B. Bosse: Conceptualization, Data curation, Funding acquisition, Methodology Project administration, Supervision, Writing – original draft, Writing – review & editing

ACKNOWLEDGEMENTS

We would like to thank the Topf lab at CSSB for help in the initial phases of this project, Milot Midarta and Martin Steinegger for help with Foldseek and for providing links to Foldseek in HerpesFolds and Lisa Holm for advice in using DALI as a local installation. We would also like to thank Michael Hartung and Olga Zolotareva for allowing us to use DrugSt.One for cluster representation. Finally, we thank Christian Löw for his help in evaluating if the BMRF2 protein cluster codes for functional transporters. This research was partly supported by the Maxwell computational resources operated at Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany.

FUNDING

Work in the Bosse lab was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2155—project no. 390874280, the DFG-funded RTG 2771 Humans and Microbes, project no. 453548970RTG, the DFG-funded RTG 2887, project number 49735088, by the Wellcome Trust through a Collaborative Award (209250/Z/17/Z) and the Leibniz ScienceCampus InterACt, funded by the BWFGB Hamburg and the Leibniz Association (W75/2022 InterACt and "Hamburg- X Infektionsforschung". Moreover, the Kaufer and Bosse labs are both funded through the DFG Research Unit FOR5200 DEEP-DV (443644894) project BO 4158/5-1 and KA 3492/12-1 awarded to JBB and BBK respectively. BKK is further by the ERC consolidator grant (ERC-CoG ENDo-HERPES, 101087480).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).

2 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876, doi:10.1126/science.abj8754 (2021).

3 Akdel, M. *et al.* A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* **29**, 1056-1067, doi:10.1038/s41594-022-00849-w (2022).

4 Zmasek, C. M., Knipe, D. M., Pellett, P. E. & Scheuermann, R. H. Classification of human Herpesviridae proteins using Domain-architecture Aware Inference of Orthologs (DAIO). *Virology* **529**, 29-42, doi:10.1016/j.virol.2019.01.005 (2019).

5 van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243-246, doi:10.1038/s41587-023-01773-0 (2024).

6 Oliver, M. R. *et al.* Structures of the Hepaci-, Pegi-, and Pestiviruses envelope proteins suggest a novel membrane fusion mechanism. *PLoS Biol* **21**, e3002174, doi:10.1371/journal.pbio.3002174 (2023).

7 Evseev, P., Gutnik, D., Shneider, M. & Miroshnikov, K. Use of an Integrated Approach Involving AlphaFold Predictions for the Evolutionary Taxonomy of Duplodnaviria Viruses. *Biomolecules* **13**, doi:10.3390/biom13010110 (2023).

8 Mutz, P. *et al.* Exaptation of Inactivated Host Enzymes for Structural Roles in Orthopoxviruses and Novel Folds of Virus Proteins Revealed by Protein Structure Modeling. *mBio* **14**, e0040823, doi:10.1128/mbio.00408-23 (2023).

9 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682, doi:10.1038/s41592-022-01488-1 (2022).

10 Holm, L. & Sander, C. Dali: a network tool for protein structure comparison. *Trends Biochem Sci* **20**, 478-480, doi:10.1016/s0968-0004(00)89105-7 (1995).

11 Holm, L. Using Dali for Protein Structure Comparison. *Methods Mol Biol* **2112**, 29-42, doi:10.1007/978-1-0716-0270-6_3 (2020).

12 The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

13 in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis* (eds A. Arvin *et al.*) (2007).

14 Terwilliger, T. C. *et al.* AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat Methods* **21**, 110-116, doi:10.1038/s41592-023-02087-4 (2024).

15 Necci, M., Piovesan, D., Predictors, C., DisProt, C. & Tosatto, S. C. E. Critical assessment of protein intrinsic disorder prediction. *Nat Methods* **18**, 472-481, doi:10.1038/s41592-021-01117-3 (2021).

16 Piovesan, D., Monzon, A. M. & Tosatto, S. C. E. Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci* **31**, e4466, doi:10.1002/pro.4466 (2022).

17 Piovesan, D. *et al.* MobiDB: 10 years of intrinsically disordered proteins. *Nucleic Acids Res* **51**, D438-D444, doi:10.1093/nar/gkac1065 (2023).

18 Caragliano, E. *et al.* Human cytomegalovirus forms phase-separated compartments at viral genomes to facilitate viral replication. *Cell Rep* **38**, 110469, doi:10.1016/j.celrep.2022.110469 (2022).

19 Naniima, P. *et al.* Assembly of infectious Kaposi's sarcoma-associated herpesvirus progeny requires formation of a pORF19 pentamer. *PLoS Biol* **19**, e3001423, doi:10.1371/journal.pbio.3001423 (2021).

20 Maier, A. *et al.* Drugst.One - A plug-and-play solution for online systems medicine and network-based drug repurposing. *ArXiv* (2023).

21 Bermek, O. & Williams, R. S. The three-component helicase/primase complex of herpes simplex virus-1. *Open Biol* **11**, 210011, doi:10.1098/rsob.210011 (2021).

759 22 Kazlauskas, D. & Venclovas, C. Herpesviral helicase-primase subunit UL8 is inactivated B-
760 family polymerase. *Bioinformatics* **30**, 2093-2097, doi:10.1093/bioinformatics/btu204
761 (2014).
762 23 Weisshart, K., Kuo, A. A., Hwang, C. B., Kumura, K. & Coen, D. M. Structural and functional
763 organization of herpes simplex virus DNA polymerase investigated by limited proteolysis. *J*
764 *Biol Chem* **269**, 22788-22796 (1994).
765 24 Bochkarev, A. *et al.* Crystal structure of the DNA-binding domain of the Epstein-Barr virus
766 origin-binding protein, EBNA1, bound to DNA. *Cell* **84**, 791-800, doi:10.1016/s0092-
767 8674(00)81056-9 (1996).
768 25 Stevenson, J. P. Correspondence: Fish diseases research. *Vet Rec* **95**, 325-326,
769 doi:10.1136/vr.95.14.325 (1974).
770 26 Renne, R. *et al.* Modulation of cellular and viral gene expression by the latency-associated
771 nuclear antigen of Kaposi's sarcoma-associated herpesvirus. *J Virol* **75**, 458-468,
772 doi:10.1128/JVI.75.1.458-468.2001 (2001).
773 27 Altmann, M. *et al.* Transcriptional activation by EBV nuclear antigen 1 is essential for the
774 expression of EBV's transforming genes. *Proc Natl Acad Sci U S A* **103**, 14188-14193,
775 doi:10.1073/pnas.0605985103 (2006).
776 28 Ball, C. B. *et al.* Human Cytomegalovirus IE2 Both Activates and Represses Initiation and
777 Modulates Elongation in a Context-Dependent Manner. *mBio* **13**, e0033722,
778 doi:10.1128/mbio.00337-22 (2022).
779 29 Gravel, A., Tomoiu, A., Cloutier, N., Gosselin, J. & Flamand, L. Characterization of the
780 immediate-early 2 protein of human herpesvirus 6, a promiscuous transcriptional activator.
781 *Virology* **308**, 340-353, doi:10.1016/s0042-6822(03)00007-2 (2003).
782 30 Ludwig, C. H. *et al.* High-throughput discovery and characterization of viral transcriptional
783 effectors in human cells. *Cell Syst* **14**, 482-500 e488, doi:10.1016/j.cels.2023.05.008 (2023).
784 31 Xing, J. *et al.* Comprehensive characterization of interaction complexes of herpes simplex
785 virus type 1 ICP22, UL3, UL4, and UL20.5. *J Virol* **85**, 1881-1886, doi:10.1128/JVI.01730-10
786 (2011).
787 32 Szymanska-de Wijs, K., Dezeljin, M., Bogdanow, B. & Messerle, M. Viral determinants
788 influencing intra- and intercellular communication in cytomegalovirus infection. *Curr Opin*
789 *Virol* **60**, 101328, doi:10.1016/j.coviro.2023.101328 (2023).
790 33 Bogdanow, B. *et al.* Spatially resolved protein map of intact human cytomegalovirus virions.
791 *Nat Microbiol* **8**, 1732-1747, doi:10.1038/s41564-023-01433-8 (2023).
792 34 Nabiee, R., Syed, B., Ramirez Castano, J., Lalani, R. & Totonchy, J. E. An Update of the Virion
793 Proteome of Kaposi Sarcoma-Associated Herpesvirus. *Viruses* **12**, doi:10.3390/v12121382
794 (2020).
795 35 Davison, A. J. *et al.* Homology between the human cytomegalovirus RL11 gene family and
796 human adenovirus E3 genes. *J Gen Virol* **84**, 657-663, doi:10.1099/vir.0.18856-0 (2003).
797 36 Dunn, W. *et al.* Functional profiling of a human cytomegalovirus genome. *Proc Natl Acad Sci*
798 *U S A* **100**, 14223-14228, doi:10.1073/pnas.2334032100 (2003).
799 37 Osanyinlusi, S. A. *et al.* Human Cytomegalovirus pUL11, a CD45 Ligand, Disrupts CD4 T Cell
800 Control of Viral Spread in Epithelial Cells. *mBio* **13**, e0294622, doi:10.1128/mbio.02946-22
801 (2022).
802 38 Chee, M. S. *et al.* Analysis of the protein-coding content of the sequence of human
803 cytomegalovirus strain AD169. *Curr Top Microbiol Immunol* **154**, 125-169, doi:10.1007/978-
804 3-642-74980-3_6 (1990).
805 39 Menard, C. *et al.* Role of murine cytomegalovirus US22 gene family members in replication
806 in macrophages. *J Virol* **77**, 5557-5570, doi:10.1128/jvi.77.10.5557-5570.2003 (2003).
807 40 Maninger, S. *et al.* M94 is essential for the secondary envelopment of murine
808 cytomegalovirus. *J Virol* **85**, 9254-9267, doi:10.1128/JVI.00443-11 (2011).

809 41 Sarid, R., Flore, O., Bohenzky, R. A., Chang, Y. & Moore, P. S. Transcription mapping of the
810 Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) genome in a body cavity-
811 based lymphoma cell line (BC-1). *J Virol* **72**, 1005-1012, doi:10.1128/JVI.72.2.1005-
812 1012.1998 (1998).
813 42 Nicholas, J. *et al.* Novel organizational features, captured cellular genes, and strain variability
814 within the genome of KSHV/HHV8. *J Natl Cancer Inst Monogr*, 79-88,
815 doi:10.1093/oxfordjournals.jncimonographs.a024179 (1998).
816 43 Howell, E. E. Searching sequence space: two different approaches to dihydrofolate reductase
817 catalysis. *Chembiochem* **6**, 590-600, doi:10.1002/cbic.200400237 (2005).
818 44 Barnes, K. *et al.* Distribution and functional characterization of equilibrative nucleoside
819 transporter-4, a novel cardiac adenosine transporter activated at acidic pH. *Circ Res* **99**, 510-
820 519, doi:10.1161/01.RES.0000238359.18495.42 (2006).
821 45 Loesing, J. B., Di Fiore, S., Ritter, K., Fischer, R. & Kleines, M. Epstein-Barr virus BDLF2-BMRF2
822 complex affects cellular morphology. *J Gen Virol* **90**, 1440-1449, doi:10.1099/vir.0.009571-0
823 (2009).
824 46 Walston, J. J. *et al.* The Epstein-Barr Virus Glycoprotein BDLF2 Is Essential for Efficient Viral
825 Spread in Stratified Epithelium. *J Virol* **97**, e0152822, doi:10.1128/jvi.01528-22 (2023).
826 47 Klupp, B. G., Altenschmidt, J., Granzow, H., Fuchs, W. & Mettenleiter, T. C. Identification and
827 characterization of the pseudorabies virus UL43 protein. *Virology* **334**, 224-233,
828 doi:10.1016/j.virol.2005.01.032 (2005).
829 48 El Kasmi, I. & Lippe, R. Herpes simplex virus 1 gN partners with gM to modulate the viral
830 fusion machinery. *J Virol* **89**, 2313-2323, doi:10.1128/JVI.03041-14 (2015).
831 49 Wright, N. J. & Lee, S. Y. Structures of human ENT1 in complex with adenosine reuptake
832 inhibitors. *Nat Struct Mol Biol* **26**, 599-606, doi:10.1038/s41594-019-0245-7 (2019).
833 50 Massa Lopez, D. *et al.* The lysosomal transporter MFSD1 is essential for liver homeostasis
834 and critically depends on its accessory subunit GLMP. *Elife* **8**, doi:10.7554/eLife.50025
835 (2019).
836 51 Davison, A. J. & Stow, N. D. New genes from old: redeployment of dUTPase by herpesviruses.
837 *J Virol* **79**, 12880-12892, doi:10.1128/JVI.79.20.12880-12892.2005 (2005).
838 52 Tarbouriech, N., Buisson, M., Seigneurin, J. M., Cusack, S. & Burmeister, W. P. The
839 monomeric dUTPase from Epstein-Barr virus mimics trimeric dUTPases. *Structure* **13**, 1299-
840 1310, doi:10.1016/j.str.2005.06.009 (2005).
841 53 Williams, M. V., Cox, B. & Ariza, M. E. Herpesviruses dUTPases: A New Family of Pathogen-
842 Associated Molecular Pattern (PAMP) Proteins with Implications for Human Disease.
843 *Pathogens* **6**, doi:10.3390/pathogens6010002 (2016).
844 54 Eberhage, J. *et al.* Crystal structure of the tegument protein UL82 (pp71) from human
845 cytomegalovirus. *Protein Sci* **33**, e4915, doi:10.1002/pro.4915 (2024).
846 55 Zhao, Z. S., Granucci, F., Yeh, L., Schaffer, P. A. & Cantor, H. Molecular mimicry by herpes
847 simplex virus-type 1: autoimmune disease after viral infection. *Science* **279**, 1344-1347,
848 doi:10.1126/science.279.5355.1344 (1998).
849 56 Soldan, S. S. & Lieberman, P. M. Epstein-Barr virus and multiple sclerosis. *Nat Rev Microbiol*
850 **21**, 51-64, doi:10.1038/s41579-022-00770-5 (2023).
851 57 Begum, S. *et al.* Molecular Mimicry Analyses Unveiled the Human Herpes Simplex and
852 Poxvirus Epitopes as Possible Candidates to Incite Autoimmunity. *Pathogens* **11**,
853 doi:10.3390/pathogens11111362 (2022).
854 58 Lunemann, J. D. *et al.* EBNA1-specific T cells from patients with multiple sclerosis cross react
855 with myelin antigens and co-produce IFN-gamma and IL-2. *J Exp Med* **205**, 1763-1773,
856 doi:10.1084/jem.20072397 (2008).
857 59 Heldwein, E. E. *et al.* Crystal structure of glycoprotein B from herpes simplex virus 1. *Science*
858 **313**, 217-220, doi:10.1126/science.1126548 (2006).

60 Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat Methods* **20**, 205-213, doi:10.1038/s41592-022-01685-y (2023).

61 Pang, M. *et al.* CoDock-Ligand: combined template-based docking and CNN-based scoring in ligand binding prediction. *BMC Bioinformatics* **24**, 444, doi:10.1186/s12859-023-05571-y (2023).

62 Baek, M. *et al.* Accurate prediction of protein-nucleic acid complexes using RoseTTAFoldNA. *Nat Methods* **21**, 117-121, doi:10.1038/s41592-023-02086-5 (2024).