Protein-protein interfaces in viral capsids are structurally unique

Shanshan Cheng¹ and Charles L. Brooks III^{1,2,3*}

¹ Department of Computational Medicine and Bioinformatics, Medical School,
University of Michigan, Ann Arbor, MI, USA sscheng@umich.edu

² Department of Chemistry, University of Michigan, Ann Arbor, MI, USA
brookscel@umich.edu

³ Biophysics Program, University of Michigan, Ann Arbor, MI, USA

* Corresponding author, 930 North University Avenue, Ann Arbor, Michigan 48109,
U.S.A., (Phone) +1 734-647-6682, (Fax) +1 734-647-1604.

Abstract

Viral capsids exhibit elaborate and symmetrical architectures of defined sizes and remarkable mechanical properties not seen with cellular macromolecular complexes. Given the uniqueness of the higher order organization of viral capsid proteins in the virosphere, we explored the question of whether the patterns of protein-protein interactions within viral capsids are distinct from those in generic protein complexes. Our comparative analysis involving a non-redundant set of 551 inter-subunit interfaces in viral capsids from VIPERdb and 20014 protein-protein interfaces in non-capsid protein complexes from the PDB found 418 generic protein-protein interfaces that share similar physicochemical patterns with some protein-protein interfaces in the capsid set, using the program PCalign we developed for comparing protein-protein interfaces. This overlap in the structural space of protein-protein interfaces is significantly small, with a p-value <

PDB, Protein Data Bank; VIPERdb, VIRus Particle ExploreR; IAU, Icosahedral Asymmetric Unit; PISA, Proteins, Interfaces, Structures and Assemblies

0.0001, based on a permutation test on the total set of protein-protein interfaces. Furthermore, the generic protein-protein interfaces that bear similarity in their spatial and chemical arrangement with capsid ones are mostly small in size with fewer than 20 interfacial residues, which results from the relatively limited choices of natural design for small interfaces, rather than having significant biological implications in terms of functional relationships. We conclude based on this study that protein-protein interfaces in viral capsids are non-representative of patterns in the smaller, more compact cellular protein complexes. Our finding highlights the design principle of building large biological containers from repeated, self-assembling units, and provides insights into specific targets for antiviral drug design for improved efficacy.

Key words: Protein-protein interaction, structural comparison, capsid shells, biological containers, drug specificity

Introduction:

Viral capsids display an elaborate and symmetrical architecture on a scale that is rarely seen in other macromolecular complexes formed by cellular proteins, with a few exceptions such as the bacterial microcompartments¹, the porous clathrin coat² and the chlorophyll protein complex³. Most available structural models of viral capsids observe icosahedral symmetry, where 60 copies of the icosahedral asymmetric unit (IAU) tile the 20 triangular faces. Each IAU contains one or more protein subunits, with the number denoted by the Triangulation number or T-number. Here the term ‘subunit’ is used interchangeably with the term protein relative to the quaternary structure of viral capsids to mean the constituent protomer, rather than referring to part of a protein. Viral capsids can also be viewed as a combination of capsomeres, which are either pentamers (12 of

them in one icosahedral capsid) or hexamers ($10 \times (T-1)$ of them in one icosahedral capsid) of the protein subunits. In $T=1$ viruses, all 60 capsid proteins are placed in an identical environment and form pentamers only. The majority of the $T>1$ viruses with more than 60 capsid proteins in the capsid shell contain a mix of pentamers and hexamers, and obey the quasi-equivalence principle proposed by Caspar and Klug⁴, permitting slightly varied modes of interaction between capsid proteins in different structural environments, with a few exceptions that either still adhere to the overall icosahedral symmetry^{5, 6} or slightly distort such symmetry⁷.

The unique functional role of viral capsid proteins prompts questions regarding their structural characteristics, given that function follows form. Previously, our study on comparing the folded topology of viral capsid proteins and generic cellular proteins revealed the lack of connectivity in the structural space between the two⁸, highlighting the geometric constraints to which the building blocks must conform to form a closed shell⁹. In the current paper, we concern ourselves with whether the protein-protein interfaces in viral capsids are similarly subjected to such selective pressure.

Other than their distinct higher order organization, capsid proteins display an interesting feature of interface plasticity, allowing different intermolecular contacts to be formed by sequence-wise identical proteins in one static quaternary structure¹⁰, as opposed to dynamically controlled variation in intermolecular interactions in some other biological systems¹¹. Understanding how such fine-tuning of protein-protein interactions is achieved in viral capsids will enhance our knowledge of the fundamental principles of protein association, which we can apply to areas such as making better predictions in protein-protein docking studies by improving the scoring functions accordingly. This is

only true, however, if the knowledge we gain from the wealth of interaction data in viral capsids is generalizable to all proteins, which necessitates the assessment of whether the inter-subunit interfaces in viral capsids span the structural interface space of all proteins including cellular ones.

A third aspect that motivates the comparison of interfaces in viral capsids versus those in cellular protein complexes is whether the design of antiviral drugs targeting viral capsid assembly can be rationalized to improve specificity. While focus of current treatment for viral diseases has largely been on viral enzymes¹², viral capsid proteins are emerging as a highly promising yet underexplored therapeutic target, with important theoretical works^{13; 14} laying the foundation for plausible strategies of interfering with the capsid assembly process. Notable studies of antiviral small molecules with well-characterized mechanism of action include the Janssen and the WIN classes of compounds targeting the picornaviruses^{15; 16; 17; 18; 19; 20; 21; 22} and the heteroaryldihydropyrimidines targeting the Hepatitis B viruses^{23; 24; 25; 26}, and more recently small molecule and peptide inhibitors of the HIV virus^{27; 28; 29; 30}. Some of these antiviral agents act as assembly agonists by modifying the assembly kinetics^{28; 31}, some inhibit assembly by altering the inter-subunit interfaces upon binding^{27; 30}, and others misdirect assembly by modulating both the kinetics and the inter-subunit geometry²⁶. As computational efforts towards structure-based drug design continue to propose new candidates for experimental validation³², precaution needs to be taken to evaluate the selectivity of these drug candidates, regardless of the intended mechanism of action. If we find that the inter-subunit interfaces in capsids are indeed uniquely found in viruses,

we can focus on these pathogen-specific sites for therapeutic development with less concern about off-target effects that may disrupt normal cellular activities.

To address the question of whether the modes of protein-protein recognition seen in viral capsids are representative of those found in small oligomeric protein-protein complexes that are typically formed by cellular proteins, we perform a structural comparative analysis across all pairwise-interacting dimers in viral capsids versus those in generic protein complexes for all structural models available to date. The same question was raised by Joel Janin and coworkers³³, but with a different approach that comprehensively surveyed various structural features of capsid protein-protein interfaces, crystal contacts and protein-protein complexes, such as buried surface area, chemical composition and atomic packing. In our work, we make the first attempt to directly test the hypothesis that the inter-subunit protein-protein interfaces found in viral capsids are structurally unique. Figure 1 outlines the design of our analysis. Specifically, we would like to examine if the structural overlap in the protein-protein interface space between the set of all capsid-forming proteins and the set of all non-capsid forming proteins is significantly small. If this is true, we then have statistical evidence supporting the uniqueness of inter-subunit interfaces in viral capsids.

Results:

Similar sizes and intra-dimer sequence divergence, but different oligomerization states and shape complementarity

For the representative 551 capsid protein-protein interfaces and 20014 generic protein-protein interfaces, we first examine the size distribution of each set, specifically

the number of interfacial residues present in a protein dimer. From Figure 3, we see that there is no pronounced difference in the distribution of interface sizes in the two sets, at least not in the range of interface sizes (between 10 and 200) we have considered in our analysis. The capsid set compared to the generic set has a marginally larger number of residues making contacts at an interface. The most frequently seen interface size for a generic protein dimer is 22 interfacial residues, while an inter-subunit interface in viral capsids typically has 24 interfacial residues. Nonetheless, this difference is well within the standard deviation of 30 residues in the distribution and thus negligible. Having comparable interface sizes in the two sets implies that differences identified, if any, should be mainly attributed to the geometry and the chemical properties of the interfacial residues. Next, we analyzed the distribution of intra-dimer sequence identity in the two data sets. Again we found approximately the same proportion of homodimers in the two data sets (67% in the case of the capsid set and 79% in the generic set) (Figure 4). It should be noted that the ratio of homodimers versus heterodimers in our data sets are not reflective of the true distribution in nature, as we applied a structural filtering procedure to only keep representative structures in our data sets. For comparison purposes in this study, the factor of interface size in terms of residue counts and that of intra-dimer sequence divergence may not contribute significantly to differences in the structural property, if any.

The oligomerization states, however, differ significantly between the two sets, as expected. Similar to what was done in³³, we examined the number of pairwise protein-protein interfaces that the protein is simultaneously involved in, i.e. the number of neighbors each protein has. This number of interacting partners gives a crude description

of the higher order organization of proteins in their quaternary states. As shown in Figure 5, a typical generic protein forms a dimer, while capsid proteins are frequently interacting with 5 other proteins within the same shell, consistent with the theory of requiring a 5-edged prototile in order to build a canonical capsid³⁴, highlighting the overall complexity of capsid shell formation. The different oligomerization states may underline potential differences we find in the two sets of proteins.

Another important physical property related to the functional requirement of viral capsid proteins is the level of atom packing at the protein-protein interface, or surface complementarity. It has been reported previously in related studies that weak, reversible interactions between subunits govern the viral capsid assembly^{35; 36; 37}. We computed the shape complementarity of all interfaces in both datasets using the CCP4 program suite³⁸, and their density distribution is shown in Figure 6. Compared to generic protein-protein interfaces, a larger proportion of inter-subunit interfaces in the capsid set are characterized with a poor shape correlation statistic³⁹. This is in line with our expectation that capsid proteins in general do not bind with one another as strongly, which arises from the necessity to both assemble and disassemble at different stages of the life cycle of viruses. Consequently, the similarity in the geometric arrangements of interfacial residues between the two data sets is likely low, as further confirmed in the subsequent result section.

Protein-protein interfaces formed by capsid proteins are distinct from those formed by cellular proteins

The all-against-all comparison between pairwise protein-protein interfaces in viral capsids and cellular protein complexes identified altogether 418 generic interfaces that resemble capsid ones, where the similarity is defined based on a distance cutoff of 0.5. Compared with the distribution of any two sets of the same sizes, estimated from the 10,000 permutation tests, this overlap in structural space falls on the extreme left as shown in Figure 7, which is significantly small with a one-tailed p-value < 0.0001 . We further show in Supplemental Figure S1(a) that such disconnectivity between the capsid set and the generic set in the protein-protein interface space is not the result of differences in the sampling density of the two sets; in fact, the intra-set connectivity is about equal for the two sets. Rather, the capsid interfaces are characterized with a sparse neighborhood that makes them less connected to the rest of protein-protein interfaces (Supplemental Figure S1(b)). We thus have statistical evidence that in terms of interfacial patterns formed between protein dimers, inter-subunit interfaces in viral capsids are not representative of generic protein-protein interfaces found in cellular protein complexes.

The same conclusion can be arrived at even when a different distance cutoff is chosen. Figure 8 plots the cumulative fraction of 20014 interfaces that are within a certain structural distance of their nearest neighbor in the complementary set of 551 interfaces. The blue curve corresponding to the comparison between the capsid set and the generic set is slightly shifted to the right of the grey curves representing the 10,000 permutations, which suggests that inter-subunit interfaces in viral capsids are more different from generic protein-protein interfaces compared to what happens as a result of random chance.

This holds true across the entire spectrum of structural distances, including the cutoff of 0.5 chosen previously.

To assess if the same property of uniqueness holds true for any subset of related protein-protein interfaces versus their complement set, we performed the same pipeline of analysis on six subsets of protein-protein interfaces grouped by their function, their structural homology and their origin. For proteins with the same function, we examined interfaces formed by histone proteins and muscle proteins as two examples, the latter including actin, myosin, titin and nebulin. For proteins sharing structural homology, we looked at the superfamily level of proteins and analyzed globins, which are alpha-helical, and TIM barrels, which are alternating alpha-helices and beta-strands. For proteins grouped by organism, we studied bacteria and cattles as samples from different domains. In all six subsets of protein-protein interfaces, we did not establish the distinction between the subset and their complement set (Table 1), stressing that finding viral capsid interfaces to be structurally unique is not trivial.

Overlap in structural space of protein-protein interfaces

As a general trend, the larger a generic interface is, the more different it likely is from any inter-subunit interface in viral capsids, as evident from the connectivity profile shown in Figure 9. As we examine the 418 generic interfaces that resemble at least one inter-subunit interface in some capsid, we see their structural analogues in the capsid set are generally small in size, with more than 70% of them having fewer than 20 interfacial residues (Figure 10). Agreeably, for protein dimers that form a small area of contact, there are most likely only limited ways of arranging a few points (i.e. interfacial residues)

spatially. Thus for the smaller interfaces in viral capsids, it is intuitively easier to find similar patterns among generic interfaces. Furthermore, interactions rendered by a few interfacial residues are unlikely to contribute to a great amount of binding energy compared with larger interfaces (which should not be confused with the idea of a few hot spot residues anchoring protein-protein association in general). In fact, most of these protein-protein interfaces in viral capsids that overlap with generic ones are found in between capsomeres, which were shown to be stable assembly intermediates in some (but not all) viral capsids⁴⁰, rather than within a capsomere. This overlap in structural space of protein-protein interfaces in the two sets of data thus may not have significant biological implications.

Unlike the case with comparing the folded topology between capsid proteins and generic proteins, where we found a few representative classes of cellular proteins that resemble capsid ones which also form symmetric oligomers, no similar trend can be easily identified with the capsid-like generic interfaces in terms of the functional annotation of the constituent protein monomers. A complete list of the diverse range of GO terms associated with these protein interfaces is summarized in Supplemental Table 2. Among these, the most abundant hits are associated with proteins in the immune system. This is not surprising, given the frequent encounters and close interaction between the host immune system and the viral coat proteins before the virus gains access to the cell interior. What appear most frequently in these small, capsid-like generic interfaces are clusters of discontinuous fragments from loop regions, rather than well-defined secondary structural elements, and the constituent proteins forming these small contacts are typically part of larger oligomeric complexes as opposed to homodimers. Figure 11

illustrates a few examples of these cases, with the constituent proteins covering functional classes such as transferases (Figure 11 (A)), cell adhesion (Figure 11 (B)) and ligases (Figure 11 (C)).

Discussion:

Result of comparative study is not sensitive to quality of structural data

Structural comparison between biological molecules relies on having reasonable structural models solved by experimental techniques. The large, elaborate nature of macromolecular complexes such as viral capsids often creates limits on the sample availability and regularity, resulting in lower resolution of the structures determined. Consequently, the atomic coordinates of these structural models are less accurate and sometimes with details missing, compared to those of small protein complexes involving cellular proteins. Furthermore, in the case of viral capsids, the symmetry of the assembled architecture is utilized for averaging over the experimental density of all asymmetric units for coordinate derivation, leading to additional sources of imprecision. Nonetheless, the structural comparison tool that we used in this study scores the similarity between a pair of protein-protein interfaces with reduced representation of their structural information, both geometrically and chemically, and we have demonstrated previously⁸ that our comparison metric can recognize significant similarity between related protein-protein interfaces in artificially created data that are highly corrupted. Given the robustness of our method against noisy data, it is unlikely that the conclusion drawn based on our comparative study will be affected by the inaccuracy of the atomic positions in these structural models.

Domain-swapped interfaces are not treated differently

Domain swapping can be found in homomeric complexes, and it refers to two identical proteins exchanging the same structural elements to form dimer interfaces that replace the original intramolecular contacts in the monomeric states. Since its first recovery in diphtheria toxin⁴¹, many domain-swapped structures of diverse origin have been identified⁴². These interfaces are best characterized by their highly intertwined nature, blurring the boundary between intramolecular contacts and intermolecular ones. In our work, we have not delineated this class of protein-protein interfaces from the rest of generic interfaces for special consideration, for the main reason that despite the possible advantage of the domain swapping mechanism in formation of large, stable protein complexes, domain swapping is still rare (on the order of 10) among currently solved structures⁴³ and remains less well understood for unambiguous annotation. This being the case, we do not expect the negligible fraction contributed by this class of protein-protein interfaces to have significant bearing on the overall conclusion of our study.

Implications of protein-protein interfaces in capsids being unique

Viral capsid shells represent large macromolecular assemblies that exist on a scale not seen in most of cellular macromolecular complexes. Given the unique function of capsid proteins in making shells, we are interested in whether the sticky patches on these building blocks that piece them together are also structurally unique, in order to pinpoint features critical for the design principle of large biological containers. Our hypothesis testing has provided strong statistical evidence of the distinctiveness of the interfacial

patterns found between capsid proteins from those between cellular proteins, which has the important implication that how capsid proteins recognize and associate with one another is retained specifically by evolution to favor the higher order organization. This conclusion that capsid interfaces are subjected to evolutionary constraints is in agreement with the previous findings that changes in inter-subunit geometry have direct impact on the oligomerization states of protein subunits⁴⁴, which in turn should confer selective pressure on protein-protein interactions. Combined with our earlier observation that the folded topology of viral capsid proteins is also segregated from those of cellular proteins, we arrive at the conclusion that the basic shape of these Lego pieces and the molecular recognition sites on their edges act concertedly to create the sophisticated shell architecture as designed.

In terms of rationalizing therapeutic design of antiviral drugs, the finding of this work favors the view that pathogen specificity can be achieved, given that protein-protein interfaces in viral capsids are significantly different from those involved in cellular activities. While this statement holds true, we would like to caution against the danger of introducing inhibitors targeting the smaller interfaces in viral capsids, especially those not found within the same capsomere, which can possibly fall into the structural overlap between capsid protein-protein interfaces and generic ones. As shown in Supplemental Table 2, these potential off-targets in the cellular domain cover a wide range of biological functions, including various enzymatic activities and gene regulation that are crucial for life. A more conservative approach would be to focus on those larger interfaces that provide greater stabilization energy for the capsid shell, in order to minimize undesirable cross-reaction with cellular proteins. Remarkably, crystal structures of known antiviral

agents that directly interfere with protein-protein interfaces (as opposed to targeting individual proteins as in the case of the picornavirus inhibitors or affecting interface geometry allosterically as in the case of the CAI inhibitor for HIV) show that they bind to interfaces within a capsomere rather than between capsomeres; the HAP compound targeting the Hepatitis B viruses inserts into the C1D7 and D1C9 interfaces²³, which are part of a hexamer (the numbering here follows the convention in the VIPERdb database⁴⁵), and the class of HIV inhibitors reported in ²⁸ also sit between a hexameric interface formed by the N-terminal domain of one capsid protein and the C-terminal domain of another monomer. These binding sites of the known drugs coincide with what we predict as favorable, given they are most likely virus-specific.

Lastly, our study has confirmed that interfacial patterns in viral capsids are not representative of those in cellular protein complexes. While most cellular proteins function by forming binary interactions, either transiently or permanently, viral capsid proteins are characterized by interacting with multiple partners simultaneously in the multiple-component assemblies. What we are most interested in with regards to properties of inter-subunit interfaces in viral capsids is the control of quasi-equivalence, which provides a great example of exquisite fine-tuning of modes of interaction via conformational switching of identical gene products^{10; 46}. However, as we established in this study, the larger interfaces within capsomeres do not share much structural similarity with generic protein-protein interfaces, and it is precisely the alternation between the concave pentamers and the flat hexamers that best manifests the quasi-equivalent property. Therefore, what we learn about physical principles governing protein-protein

recognition from capsid proteins may not easily extend to protein-protein interactions at large.

The above aspects cover the implications of our findings in this work. What our conclusion does not imply, however, is the general uniqueness of structural characteristics of all viral proteins. The exclusive sets of folded topology and distinct patterns of association with one another found in viral capsid proteins are correlated with their specific function that is missing for cellular life forms. Unlike capsid proteins, other viral proteins may assume roles that partially overlap with cellular proteins. Examples include viral proteases that process polypeptides for maturation, which contain motifs for hydrolysis of peptide bonds similar to cellular proteases that lyse misfolded peptides for recycling, and reverse transcriptases in retroviruses necessarily share similar nucleotide binding sites with DNA polymerases in cellular machinery. Related studies on the structural relationship between general classes of viral proteins and cellular proteins have been carried out, with focuses on virus-host interaction in humans⁴⁷ and thermodynamic stabilities⁴⁸. The first study established extensive overlap between virus-host interactions and endogenous interactions within the host for competitive binding, suggesting that the uniqueness of capsid protein-protein interfaces is a result of functional requirement rather than their viral origin. The second study concluded on the high adaptability of viral proteins for effective interaction with proteins in the host, as opposed to thermostable proteins, which are less tolerant of the deleterious effects of mutations. This second study presents an additional aspect of viral proteins, specifically their chemical composition and disorder propensity. One needs to be cautious in extrapolating this unique

biophysical property of viral proteins to all protein-protein interfaces that are viral derived without further investigation.

In the earlier study by Bahadur *et al.*³³, various chemical and physical features of protein-protein interfaces in 49 icosahedral viral capsids were analyzed and compared with those in oligomeric protein complexes, homodimers and crystal contacts. They reported that the average atom packing in capsid interfaces is lower than that in generic protein complexes based on the packing index⁴⁹, which agrees with our calculation using the shape complementarity metric. This geometric property reflects the need of capsid protein subunits to assemble and disassemble at different stages, which is very likely subjected to selective pressure. In the same study, Bahadur *et al.* also found a big difference in the chemical composition of interfacial residues between the two groups, with the capsid set having more non-polar residues and fewer charged residues. Our analysis shows otherwise, with an enrichment of polar residues and a depletion of charged residues, although the difference is largely within the variations among individual interfaces (Supplementary Figure S2(a)). The apparent disparate findings could be attributed to the different data sets included for analysis, which suggests the highly diverse nature of the chemical properties of protein-protein interfaces. Furthermore, both the previous study and our current work established that capsid protein-protein interactions are multi-component in nature, involving far more neighbors per protein than those in generic protein complexes, highlighting the difference in the resulting oligomerization states. Additionally, we found that the average disorder propensity of interfacial residues in the capsid interface set is marginally greater than that in the generic protein interface set (Supplementary Figure S2(b)), which may correlate

with the greater flexibility in capsid interfaces, manifested as quasi-equivalence. Taken together, capsid protein-protein interfaces display an array of characteristics that are not representative of generic protein-protein interfaces at large, most of which likely arise from the necessity of forming the capsid shell.

Concluding remarks

To conclude, we have shown in this study via rigorous hypothesis testing that inter-subunit interfaces in the large, elaborate viral capsid assemblies are significantly different from protein-protein interfaces found in the smaller and simpler cellular protein complexes. This difference in capsid interfaces is most likely a consequence of the functional requirement for making the specific architecture of shells for encapsulation of the viral genome, which on average involves more neighbors per protein subunit as well as assumes a more expanded organization compared with cellular complexes that are often times binary and more compact and collapsed in nature. Our results should provide deeper understanding of the nature of self-association in large biological containers for creating predictable designs in a variety of therapeutic and materials science applications.

Materials and Methods:

Data collection

Each viral capsid shell contains 60 copies of the IAU, which consists of T-number of protein subunits. Therefore all pairwise dimer interfaces involving at least one protein subunit from the first IAU sufficiently represent all unique protein-protein interfacial patterns found in that particular virus. For the 421 entries in the database Virus Particle ExploreR (VIPERdb)⁴⁵, which include procapsids (capsids in the premature form) as well

as subviral particles that form as a result of modified conditions, yet all obeying icosahedral symmetry, we collect 1930 dimers involving subunits from the first asymmetric unit, with at least 10 interfacial residues and not exceeding 200 interfacial residues in those dimers. Interfacial residues are defined as the set of all residues that are in contact with at least one residue in the binding partner, where a contact between two residues is defined by having at least two heavy atoms, one from each residue, within 4.5 Å. When structural models have reduced information, such as backbone models, a C α -C α distance cutoff specific to amino acid pair type is used instead⁵⁰. The cutoffs of the minimal and maximal number of interfacial residues were chosen based on the rationale that pairwise interacting protein dimers with too small an interface are insufficiently stabilized in the complexed state, and too large a protein-protein interface typically involves multiple domains. Moreover, 50 additional entries of capsid protein dimers that we identified in the generic interface set (described in the following section) were removed from that set and added to the capsid interface set (the complete list of these 50 entries is found in Supplemental Table 1). These 1980 capsid dimers were then clustered using complete-linkage hierarchical agglomerative clustering cut at a PC-score (explained in the Comparison metric section) of 0.5 into 551 groups, from each of which a median structure is chosen to be included in the final representative set of capsid protein-protein interfaces. It is worth noting that the final clustering procedure has led to only one or few representatives of quasi-equivalent interfaces being retained in the data set, because quasi-equivalent interfaces are usually similar enough to be filtered off as structurally redundant. However, we believe that the representative set captures the

coarse overall geometric and chemical moieties found in capsid shells and thus suffices for our comparative study.

For the generic interface set, 75694 structural models were collected from the Protein Data Bank (PDB)⁵¹, which were screened against Proteins, Interfaces, Structures and Assemblies (PISA)⁵² to extract 165257 protein dimer interfaces assigned to be biologically significant (as opposed to due to crystal packing). 123395 remained after pruning those with fewer than 10 interfacial residues or more than 200 interfacial residues. This set was then grouped into dimers that have pairwise sequence identity of at least 50% for both chains. Within each group, a clustering was performed based on the structural similarity of protein-protein interfaces and the representatives chosen using a PC-score cutoff of 0.5, given similar monomers can interact in different orientations relative to each other. The entries corresponding to inter-subunit interfaces in viral capsids are removed from this reduced set and appended to the viral capsid set as described earlier. Finally, the same structural clustering procedure is performed to retain representative dimers only to constitute our generic dimer set consisting of 20014 pairwise generic interfaces.

Comparison metric

For quantifying the similarity between two interfaces of a given pair of protein dimers, we use the program PCalign⁵⁰, which returns a normalized score, PC-score, based on a structural alignment of the two interfaces. PC-score takes the following form:

$$\text{PC-score} = \text{PC-score}_{\text{raw}}^{\ln 0.3 / \ln(0.14 + 0.29 \times 0.97^{L_{ave}})}, \quad (1)$$

where L_{ave} is the average of the number of interfacial residues for the pair of protein dimers compared, and $PC\text{-score}_{raw}$ is computed by

$$PC\text{-score}_{raw} = \frac{f_c}{L_{ave}} \sum_{i=1}^{L_{ali}} \frac{1}{1 + 0.25 \times (1 - I_{ii}(\text{same chem type})) + \frac{d_{ii}^2}{16}}, \quad (2)$$

based on the structural alignment of the two interfaces identified by PCalign. Here f_c is the ratio of common contacts between the two interfaces aligned given by

$$f_c = \frac{\overrightarrow{N^1} \cdot \overrightarrow{N^2}}{(\overrightarrow{N^1} \cdot \overrightarrow{N^1} + \overrightarrow{N^2} \cdot \overrightarrow{N^2})/2}, \quad (3)$$

where $\overrightarrow{N^1}$ and $\overrightarrow{N^2}$ are matrices representing the contact maps of the aligned interfacial residues in the pair of interfaces being compared. The dot operation represents the inner product of two matrices. $I_{ii}(\text{same chem type})$ is the indicator function that takes the value of 1 if the i^{th} pair of aligned residues share the same chemical functional group and 0 if they do not (see⁵⁰). d_{ii} is the spatial distance in Å between the Ca atoms of the i^{th} pair of aligned residues. This scoring function ranges from 0 to 1, with 1 resulting from comparison of identical interfaces (self-comparison).

Quantify overlap in structural space

To have a quantitative measure of the extent to which generic protein-protein interfaces represent patterns found at capsid inter-subunit interfaces, we perform the following analysis. We carry out $M \times N$ comparisons of protein-protein interfaces using PCalign between all M members in the capsid set and all N members in the generic set. For each of the N generic protein-protein interfaces, we find its nearest neighbor in the capsid set, where the structural distance is defined by $(1 - PC\text{-score})$. For a given generic

protein-protein interface, the distance between itself and its nearest neighbor in the capsid set reflects how connected it is to the capsid set in the interface space. We then select the generic protein-protein interfaces whose distances with their respective nearest neighbor are less than 0.5, and this set of generic interfaces is considered to share significant structural similarity with some inter-subunit interface in viral capsids, representing the overlapping region in the Venn diagram shown in Figure 1. The count of these capsid-like generic interfaces is used as the test statistic for our hypothesis testing.

Statistical significance of the test statistic

To assess whether the shared structural space between generic protein-protein interfaces and inter-subunit interfaces found in viral capsids is significantly small, we estimate the p-value of our test statistic by a permutation test, as summarized in Figure 2. Similar to what was done previously⁸, the total set of protein-protein interfaces is first randomly partitioned into two sets, A and B, that have the same number of interfaces as the capsid set and the generic set respectively. We then count the number of interfaces in the larger set B that highly resemble at least one interface in the smaller set A, which is our variable of interest. By repeating this experiment 10,000 times, we obtain the distribution of structural overlap between any two mutually exclusive sets of the given sizes, and can therefore estimate the probability of obtaining a value smaller than or equal to our test statistic by random chance.

Acknowledgements:

We thank Dr. Janet Smith (University of Michigan, Ann Arbor) for her helpful suggestions in both deriving our data sets and adding the discussion about domain

swapping. We acknowledge the funding support for this work from NSF grant MCB1121575.

References:

1. Tanaka, S., Kerfeld, C. A., Sawaya, M. R., Cai, F., Heinhorst, S., Cannon, G. C. & Yeates, T. O. (2008). Atomic-level models of the bacterial carboxysome shell. *Science* **319**, 1083-6.
2. Fotin, A., Cheng, Y., Sliz, P., Grigorieff, N., Harrison, S. C., Kirchhausen, T. & Walz, T. (2004). Molecular model for a complete clathrin lattice from electron cryomicroscopy. *Nature* **432**, 573-9.
3. Boekema, E. J., Hifney, A., Yakushevskaya, A. E., Piotrowski, M., Keegstra, W., Berry, S., Michel, K. P., Pistorius, E. K. & Kruip, J. (2001). A giant chlorophyll-protein complex induced by iron deficiency in cyanobacteria. *Nature* **412**, 745-8.
4. Caspar, D. L. & Klug, A. (1962). Physical principles in the construction of regular viruses. *Cold Spring Harb Symp Quant Biol* **27**, 1-24.
5. Liddington, R. C., Yan, Y., Moulai, J., Sahli, R., Benjamin, T. L. & Harrison, S. C. (1991). Structure of simian virus 40 at 3.8-A resolution. *Nature* **354**, 278-84.
6. Stehle, T., Yan, Y., Benjamin, T. L. & Harrison, S. C. (1994). Structure of murine polyomavirus complexed with an oligosaccharide receptor fragment. *Nature* **369**, 160-3.
7. Gipson, P., Baker, M. L., Raytcheva, D., Haase-Pettingell, C., Piret, J., King, J. A. & Chiu, W. (2014). Protruding knob-like proteins violate local symmetries in an icosahedral marine virus. *Nat Commun* **5**, 4278.
8. Cheng, S. & Brooks, C. L., III. (2013). Viral capsid proteins are segregated in structural fold space. *PLoS Comput Biol* **9**, e1002905.
9. Mannige, R. V. & Brooks, C. L., III. (2008). Tiling nature of virus capsids and the role of topological constraints in natural capsid design. *Phys Rev E Stat Nonlin Soft Matter Phys* **77**, 051902.
10. Johnson, J. E. (1996). Functional implications of protein-protein interactions in icosahedral viruses. *Proc Natl Acad Sci U S A* **93**, 27-33.
11. Milligan, R. A. (1996). Protein-protein interactions in the rigor actomyosin complex. *Proc Natl Acad Sci U S A* **93**, 21-6.
12. Bocanegra, R., Rodriguez-Huete, A., Fuertes, M. A., Del Alamo, M. & Mateu, M. G. (2012). Molecular recognition in the human immunodeficiency virus capsid and antiviral design. *Virus Res* **169**, 388-410.
13. Caspar, D. L. (1980). Movement and self-control in protein assemblies. Quasi-equivalence revisited. *Biophys J* **32**, 103-38.
14. Prevelige, P. E., Jr. (1998). Inhibiting virus-capsid assembly by altering the polymerisation pathway. *Trends Biotechnol* **16**, 61-5.
15. Hiremath, C. N., Grant, R. A., Filman, D. J. & Hogle, J. M. (1995). Binding of the antiviral drug WIN51711 to the sabin strain of type 3 poliovirus: structural comparison with drug binding in rhinovirus 14. *Acta Crystallogr D Biol Crystallogr* **51**, 473-89.
16. Grant, R. A., Hiremath, C. N., Filman, D. J., Syed, R., Andries, K. & Hogle, J. M. (1994). Structures of poliovirus complexes with anti-viral drugs: implications for viral stability and drug design. *Curr Biol* **4**, 784-97.
17. Hiremath, C. N., Filman, D. J., Grant, R. A. & Hogle, J. M. (1997). Ligand-induced conformational changes in poliovirus-antiviral drug complexes. *Acta Crystallogr D Biol Crystallogr* **53**, 558-70.
18. Rossmann, M. G. (1989). The structure of antiviral agents that inhibit uncoating when complexed with viral capsids. *Antiviral Res* **11**, 3-13.

19. Smith, T. J., Kremer, M. J., Luo, M., Vriend, G., Arnold, E., Kamer, G., Rossmann, M. G., McKinlay, M. A., Diana, G. D. & Otto, M. J. (1986). The site of attachment in human rhinovirus 14 for antiviral agents that inhibit uncoating. *Science* **233**, 1286-93.
20. Diana, G. D., Pevear, D. C., Otto, M. J., McKinlay, M. A., Rossmann, M. G., Smith, T. & Badger, J. (1989). Inhibitors of viral uncoating. *Pharmacol Ther* **42**, 289-305.
21. McKinlay, M. A., Pevear, D. C. & Rossmann, M. G. (1992). Treatment of the picornavirus common cold by inhibitors of viral uncoating and attachment. *Annu Rev Microbiol* **46**, 635-54.
22. Rossmann, M. G. (1988). Antiviral agents targeted to interact with viral capsid proteins and a possible application to human immunodeficiency virus. *Proc Natl Acad Sci U S A* **85**, 4625-7.
23. Bourne, C. R., Finn, M. G. & Zlotnick, A. (2006). Global structural changes in hepatitis B virus capsids induced by the assembly effector HAP1. *J Virol* **80**, 11055-61.
24. Stray, S. J. & Zlotnick, A. (2006). BAY 41-4109 has multiple effects on Hepatitis B virus capsid assembly. *J Mol Recognit* **19**, 542-8.
25. Stray, S. J., Bourne, C. R., Punna, S., Lewis, W. G., Finn, M. G. & Zlotnick, A. (2005). A heteroaryldihydropyrimidine activates and can misdirect hepatitis B virus capsid assembly. *Proc Natl Acad Sci U S A* **102**, 8138-43.
26. Bourne, C., Lee, S., Venkataiah, B., Lee, A., Korba, B., Finn, M. G. & Zlotnick, A. (2008). Small-molecule effectors of hepatitis B virus capsid assembly give insight into virus life cycle. *J Virol* **82**, 10262-70.
27. Tang, C., Loeliger, E., Kinde, I., Kyere, S., Mayo, K., Barklis, E., Sun, Y., Huang, M. & Summers, M. F. (2003). Antiviral inhibition of the HIV-1 capsid protein. *J Mol Biol* **327**, 1013-20.
28. Blair, W. S., Pickford, C., Irving, S. L., Brown, D. G., Anderson, M., Bazin, R., Cao, J., Ciaramella, G., Isaacson, J., Jackson, L., Hunt, R., Kjerrstrom, A., Nieman, J. A., Patick, A. K., Perros, M., Scott, A. D., Whitby, K., Wu, H. & Butler, S. L. (2010). HIV capsid is a tractable target for small molecule therapeutic intervention. *PLoS Pathog* **6**, e1001220.
29. Sticht, J., Humbert, M., Findlow, S., Bodem, J., Muller, B., Dietrich, U., Werner, J. & Krausslich, H. G. (2005). A peptide inhibitor of HIV-1 assembly in vitro. *Nat Struct Mol Biol* **12**, 671-7.
30. Ternois, F., Sticht, J., Duquerroy, S., Krausslich, H. G. & Rey, F. A. (2005). The HIV-1 capsid protein C-terminal domain in complex with a virus assembly inhibitor. *Nat Struct Mol Biol* **12**, 678-82.
31. Katen, S. P., Chirapu, S. R., Finn, M. G. & Zlotnick, A. (2010). Trapping of hepatitis B virus capsid assembly intermediates by phenylpropenamide assembly accelerators. *ACS Chem Biol* **5**, 1125-36.
32. ElSawy, K. M., Twarock, R., Verma, C. S. & Caves, L. S. (2012). Peptide inhibitors of viral assembly: a novel route to broad-spectrum antivirals. *J Chem Inf Model* **52**, 770-6.
33. Bahadur, R. P., Rodier, F. & Janin, J. (2007). A dissection of the protein-protein interfaces in icosahedral virus capsids. *J Mol Biol* **367**, 574-90.
34. Mannige, R. V. & Brooks, C. L., III. (2009). Geometric considerations in virus capsid size specificity, auxiliary requirements, and buckling. *Proc Natl Acad Sci U S A* **106**, 8531-6.
35. Zlotnick, A. (2003). Are weak protein-protein interactions the general rule in capsid assembly? *Virology* **315**, 269-74.
36. Rapaport, D. C. (2008). Role of reversibility in viral capsid growth: a paradigm for self-assembly. *Phys Rev Lett* **101**, 186101.

37. Zlotnick, A. (2007). Distinguishing reversible from irreversible virus capsid assembly. *J Mol Biol* **366**, 14-8.
38. Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* **67**, 235-42.
39. Lawrence, M. C. & Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *J Mol Biol* **234**, 946-50.
40. Xie, Z. & Hendrix, R. W. (1995). Assembly in vitro of bacteriophage HK97 proheads. *J Mol Biol* **253**, 74-85.
41. Bennett, M. J., Choe, S. & Eisenberg, D. (1994). Domain swapping: entangling alliances between proteins. *Proc Natl Acad Sci U S A* **91**, 3127-31.
42. Liu, Y. & Eisenberg, D. (2002). 3D domain swapping: as domains continue to swap. *Protein Sci* **11**, 1285-99.
43. Rousseau, F., Schymkowitz, J. & Itzhaki, L. S. (2012). Implications of 3D domain swapping for protein folding, misfolding and function. *Adv Exp Med Biol* **747**, 137-52.
44. Perica, T., Chothia, C. & Teichmann, S. A. (2012). Evolution of oligomeric state through geometric coupling of protein interfaces. *Proc Natl Acad Sci U S A* **109**, 8127-32.
45. Carrillo-Tripp, M., Shepherd, C. M., Borelli, I. A., Venkataraman, S., Lander, G., Natarajan, P., Johnson, J. E., Brooks, C. L., III & Reddy, V. S. (2009). VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Res* **37**, D436-42.
46. Prasad, B. V. & Schmid, M. F. (2012). Principles of virus structural organization. *Adv Exp Med Biol* **726**, 17-47.
47. Franzosa, E. A. & Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. *Proc Natl Acad Sci U S A* **108**, 10538-43.
48. Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezhovsky, I. N. & Tawfik, D. S. (2009). Do viral proteins possess unique biophysical features? *Trends Biochem Sci* **34**, 53-9.
49. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* **336**, 943-55.
50. Cheng, S., Zhang, Y. & Brooks, C. L., III. (2015). PCalign: a method to quantify physicochemical similarity of protein-protein interfaces. *BMC Bioinformatics* **16**, 33.
51. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.
52. Krissinel, E. & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774-97.

Table 1. Interfaces formed by six subsets of related proteins are not statistically distinct from their complement set. Two examples from each of the grouping methods, by function, by structural homology and by organism, are shown. With the exception of protein-protein interfaces grouped by function being marginally distinguishable with a p-value bordering on statistical significance (0.05), the rest are all highly connected to their complement set.

Protein-protein interface subsets		Subset size	p-value of distinctiveness
Group by function	Histone	135	0.055
	Muscle	401	0.086
Group by structural homology	Globin	109	0.668
	TIM barrel	17	0.410
Group by organism	<i>Thermus thermophilus</i>	488	0.120
	<i>Bos taurus</i>	390	0.136

Figure 1. Comparison between inter-subunit interfaces in viral capsids and protein-protein interfaces in generic protein complexes. The protein-protein interfaces formed by capsid proteins, illustrated on the left (PDB: 3KIC), constitutes the structural space drawn in blue, while the generic protein-protein interfaces, exemplified by those found in an RNA polymerase elongation complex shown on the right (PDB: 2O5I), constitutes the structural space drawn in grey. Overlap in the Venn diagram refers to the subset of generic protein-protein interfaces that sufficiently resemble some capsid inter-subunit interfaces, based on certain quantitative criterion for structural similarity, and hence signifies the extent to which interfacial patterns found in cellular protein complexes can be represented by those in viral capsids. This hypothesis testing design echoes that in the previous work (Figure 2 of ⁸).

Figure 2. Permutation test for estimating the statistical significance. Under our null hypothesis that inter-subunit interfaces in viral capsids are no different from generic protein-protein interfaces, we can exchange labels between the capsid set and the non-capsid set (i.e. partition the total set) in a random fashion to obtain a set A that mirrors the capsid set, and their complement set B that is the equivalent of the generic set. All possible values of the structural overlap between set A and set B under the rearrangement of the labels give us the distribution of structural overlap, from which we can obtain the statistical significance of our test statistic.

Figure 3. Density distribution of interface sizes for the generic set and the capsid set. The two largely overlap, with the generic set having a marginally larger proportion of interfaces in the range of 15 to 40 interfacial residues.

Figure 4. The capsid data set and the generic data set are both dominated by homodimers. There is no pronounced difference between the intra-dimer sequence identity of the two data sets, with the capsid set having slightly a greater proportion of heterodimers.

Figure 5. Different oligomerization states in the two sets. This figure plots the density distribution of the number of simultaneous interacting partners for any protein that is found in the generic set and the capsid set. A capsid protein, on average, has 5.7 neighboring capsid proteins within the same shell, while the average number of interacting partners for a generic protein is only 2.2.

Figure 6. Different shape complementarity in the two data sets. A greater proportion of capsid interfaces have poor atomic packing, measured by the shape correlation statistic.

Figure 7. Statistical significance of the test statistic. Out of 10,000 permutations, no single case results in 418 or fewer structurally similar interfaces identified between a randomly selected set of 551 interfaces and their complement set of 20014 interfaces, which makes the one-tailed p-value of our test statistic less than 0.0001. Hence there is strong statistical evidence supporting the hypothesis that capsid interfaces are unlike generic ones.

Figure 8. Capsid protein-protein interfaces are different from generic ones. Shown here in the empirical cumulative fraction distribution of distances between one set of 551 protein-protein interfaces and their nearest neighbor in the complementary set, we see that capsid protein-protein interfaces are structurally more distant from non-capsid ones, represented by the blue curve, than the random background, represented by 10,000 grey curves. The average empirical cumulative fraction distribution is colored in red. The range of structural distances plotted is from 0.35 to 0.75 to show a better resolved picture, while in theory this can range from 0 to 1.

Figure 9. The connectivity of a generic interface to capsid interface space versus its size. With a correlation of -0.53, the size of a generic interface is largely inversely related to its structural similarity with any inter-subunit interface found within a capsid, based on the nearest neighbor criterion of the 20014 generic interfaces in our data set.

Figure 10. Sizes of the 418 capsid protein-protein interfaces in the overlap region. Majority of these capsid interfaces that structurally resemble some generic interfaces are small in size, with the exception of one interface that has a size of 85 interfacial residues, being a long coiled coil structure.

Figure 11. Examples of similar interfaces from the capsid set and the generic set. In all panels, the viral capsid proteins are colored in cyan and orange, while the cellular protein dimers are colored in blue and red. The monomer proteins are drawn in either cartoon representation or tube representation, the latter for structural models lacking side-chain information. The C α atoms of the aligned interfacial residues are shown in van der Waals representation. (A) The interface formed by the generic protein dimer (PDB: 1XIQ, chains B and F) is aligned to the capsid interface (PDB: 1X9P, chain A in the first IAU and chain A in the 16th IAU). (B) The generic interface between chain A and chain B of PDB: 3M45 is aligned to the capsid interface formed by chain B and chain C within the same IAU in PDB: 1N6G. (C) The generic interface between chain G and chain H of PDB: 3D54 is aligned to the capsid interface between chain F of the first IAU and chain F of the second IAU in PDB: 3MUW.

graphical abstract

Structural space of
capsid interfaces

Structural space of
non-capsid interfaces