A Computational Pipeline to Identify Potential Drug Targets and Interacting Chemotypes in SARS-CoV-2

Sarah H. Sandholtz[a,*], Jeffrey A. Drocco[a], Adam T. Zemla[b], Marisa W. Torres[b], Mary S. Silva[b], and Jonathan E. Allen[b]

[a]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California, United States of America
[b]Computing Directorate, Lawrence Livermore National Laboratory, Livermore, California, United States of America
[*]Corresponding author: Sarah H. Sandholtz, sandholtz3@llnl.gov

March 21, 2022

1 Abstract

Minimizing the human and economic costs of the COVID-19 pandemic and of future pandemics requires the ability to develop and deploy effective treatments for novel pathogens as soon as possible after they emerge. To this end, we introduce a unique, computational pipeline for the rapid identification and characterization of binding sites in the proteins of novel viruses as well as the core chemical components with which these sites interact. We combine molecular-level structural modeling of proteins with clustering and cheminformatic techniques in a computationally efficient manner. Similarities between our results, experimental data, and other computational studies provide support for the effectiveness of our predictive framework. While we present here a demonstration of our tool on SARS-CoV-2, our process is generalizable and can be applied to any new virus, as long as either experimentally solved structures for its proteins are available or sufficiently accurate homology models can be constructed.

2 Introduction

To minimize the human and economic costs of a pandemic, effective treatments for novel pathogens like SARS-CoV-2 must be developed as quickly as possible. While drug target selection has improved over time, it remains difficult, and poor target selection often leads to early failures in the drug development process. An accurate and efficient computational method for identifying and characterizing protein binding pockets (*i.e.* potential drug target sites), in emerging pathogens would thus serve a critical purpose in shortening the timeframe for drug development by providing more viable starting points for experimental screenings. To enable a timely and rapid response to a new pathogen, such target identification techniques must be adaptable to unfamiliar disease agents.

Molecular docking simulations, Molecular Mechanics Generalized Born Surface Area (MM/GBSA) methods, and molecular dynamics simulations have traditionally been employed to gauge the nature and quality of protein-ligand binding and have been performed broadly in the study of SARS-CoV-2 proteins.[1–7] The

webserver D3Targets-2019-nCoV, for example, uses molecular docking to predict potential drug targets and drug compounds specifically for COVID-19.[8] Though such simulations provide detailed insight of physicochemical properties at the atomistic level, they are computationally costly and limited to short timescales. A variety of computational platforms have been designed to minimize or circumvent these drawbacks.

Existing platforms for binding pocket detection in proteins rely on a range of strategies, broadly based on templates, geometry, physicochemical properties, or machine learning. The PDBspheres program[9] draws on experimental measurements of protein-ligand complex structures from Protein Data Bank (PDB) to detect and assess potential binding pockets through superposition of structural models. The resulting PDBspheres data includes the set of protein residues predicted to be in contact with each ligand as well as a set of descriptive metrics. Other hybrid approaches combine methods from the four broad categories above. For instance, dPredGB[10] couples a geometric detection method with the calculation of desolvation properties, while PocketMatch[11] and SiteHopper[12] integrate a representation of the surface shape with a description of chemical properties. The "binding response descriptor" is based on a sphere method, refined with clustering, and energy and geometry calculations from docking simulations.[13] Similarly, Pockets 2.0[7] features the direct integration of docking simulations into Fpocket. Methods like these that incorporate physicochemical properties through calculations or simulations include an additional perspective in their description of protein binding pockets, but they also depend on predictive *in silico* data, which have not necessarily been validated experimentally and whose quality and accuracy may vary across different pockets. Still other approaches, including WaveGeoMap,[14] DeeplyTough,[15] and Site2Vec[16] make use of neural networks. While such deep learning techniques enable scalability to enormously large data sets, they create a layer of abstraction between input data and results, making them less interpretable. These platforms lay a foundation for protein binding pocket detection tasks but also leave room for new hybrid approaches that offer both template/geometry-based and machine learning-based methods – roots in purely structural experimental data with a capability to ingest large quantities of input data in a straightforward manner.

Building on this body of work, we introduce a unique computational framework for the rapid identification and characterization of binding pockets in the proteins of SARS-CoV-2 and other novel pathogens. Our new quantitative and automated approach takes into account all available experimental structural data from PDB to uncover the combination of significant residues and core ligand components. Our goal is to distinguish and characterize what we call "consensus pockets," a term we put forward both as a general concept and a more narrow, practical definition for the specific task in this work. Conceptually, consensus pockets could refer to structurally similar pockets from different organisms (structural consensus across organisms) or to pockets within a single organism that bind similar classes of ligands (consensus of ligand features), for example. The overarching notion of consensus pockets is relevant to a variety of applications and when tailored appropriately, defines the key factors of interest in a given application. For instance, structural consensus across different organisms would be important for identifying and avoiding off-target effects of prospective drug candidates, while structural or ligand consensus within a single organism would be critical for developing multi-target drugs. For this particular application, we define consensus pockets as those binding pockets with the strongest consensus of contacts across different protein-ligand complexes, based on the residues in contact with experimentally observed or computationally predicted binding ligands and certain features of those ligands (described in more detail in the next section).

To this end, we first combine molecular-level

structural modeling of protein-ligand binding and pocket detection from PDBspheres[9] with a hierarchical clustering procedure to determine and describe relevant binding pockets. While protein-ligand binding data from any pocket detection platform could serve as input to our framework, we choose to use data from PDBspheres because the metrics it reports enable more comprehensive and refined pocket characterization, the primary goal of this work. We then employ cheminformatics tools and density-based clustering to uncover the pocket-specific, key chemical substructures involved in binding. Though our framework relies on access to protein structural data which may vary in quality, we account for such differences by incorporating into our analysis a measure of the quality of our structural models. We also apply several strategies for substantiating our results and find strong qualitative support. Based on the capabilities and findings of our pipeline, we propose two simple guidelines for which consensus pockets are more promising drug targets. A SARS-CoV-2 binding pocket may be more effective as a drug target 1) the more similar its structure is to that of other viral binding pockets and 2) the more drug-like the molecules that bind.

By quickly identifying and describing binding sites in unfamiliar pathogens and the core chemical components with which these sites interact, our pocket characterization method provides a starting point for drug discovery efforts. Capable of producing results in a matter of days, our framework is designed to act as a rapid response tool in the face of an emerging pathogen. Our novel approach enables cross-organism analysis of pocket structural similarities and can reveal which, if any, binding pockets in a new pathogen are more conserved. The combination of residue and ligand clustering provides a more well-rounded perspective and is a notable feature of our process. In establishing a link between protein pockets (and even pocket subregions) and key chemical features, it allows us to make focused suggestions of molecular structures that are better suited for specific binding pockets. Additionally, the relative simplicity of our system makes it more straightforward and interpretable. Taking advantage of the fact that structural and physicochemical information is inherent in PDB entries and hence in PDBspheres data, our pipeline uses minimal parameters. Most importantly, our framework is generalizable and can be applied to any new virus, provided there are experimentally solved structures or homology models of sufficient quality.

3 Methods

The foundation of our approach lies in molecular-level knowledge of protein-ligand binding, which we obtain through experimental data and protein structural modeling from PDBspheres.[9] The Global Distance Calculations (GDC) metric,[17,18] which takes both the backbone and side chains into account, is a measure of the structural similarity between pairs of PDBspheres structural models. For a given protein-ligand binding complex, the higher the calculated GDC score, the greater the structural similarity between the PDBspheres models of the SARS-CoV-2 protein of interest and the reference ligand binding pocket observed in experimentally solved structures deposited in PDB. Using the GDC metric as a means to evaluate our data and understand our results, we leverage the residue-level resolution of PDBspheres data to identify relevant binding pockets in SARS-CoV-2 and the corresponding site-specific chemical components, as shown in the schematic of our workflow in Fig. 1.

To begin, we filter the PDBspheres data to minimize noise from lower quality homology models and irrelevant ligands and to better focus on significant binding events. Based on the modeling quality assessments from Ref. 9, we use only PDBspheres data for which the PDBspheres parameters that meet the thresholds specified below. We also apply four ligand-specific exclusion criteria.

From the remaining PDBspheres residue-ligand contact data, we use NetworkX[19] to con-

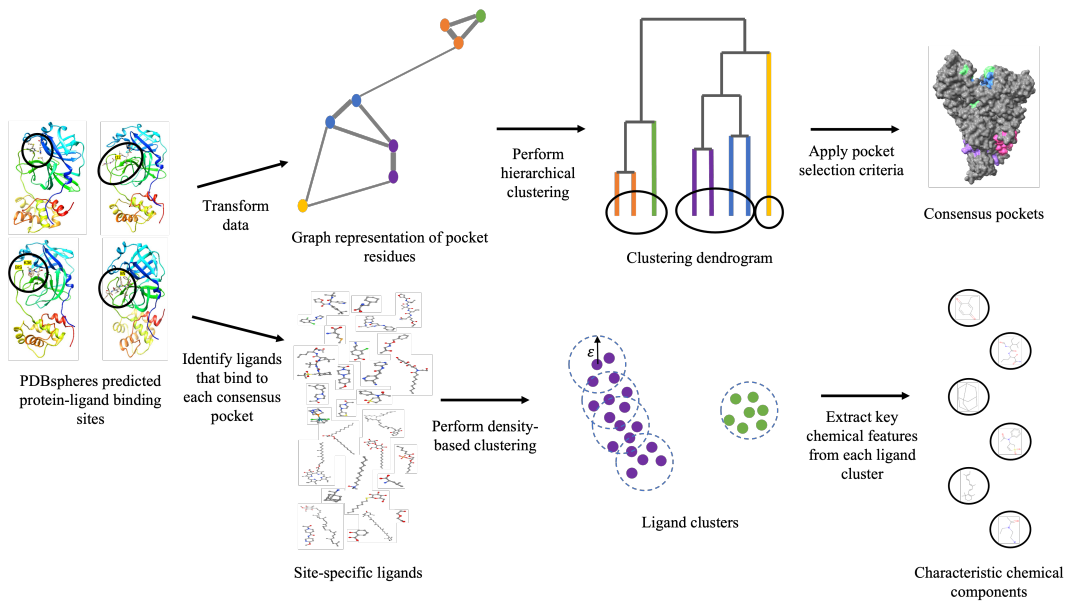


Figure 1: Schematic of our process for identifying relevant binding pockets (top row) and site-specific chemical components (bottom row).

struct a weighted, undirected graph for each SARS-CoV-2 protein, in which each node corresponds to a protein residue and each edge indicates that the two connected residues bind one or more of the same ligands (Fig. 1).

We then perform complete linkage hierarchical clustering to group protein residues. The result is a full dendrogram, like that in Fig. 1, indicating which residues or residue groups are clustered together and at what distances those clusters occur. Different distance cutoffs in the dendrogram correspond to different numbers and identities of clusters (*i.e.* putative binding pockets). Given that binding pockets may range in size, as ligands do, they may appear at different length scales in the dendrogram, and we implement a selection process (described below) to identify all relevant binding pockets.

The specific steps we take are as follows:

1. Filter the PDBspheres data based on the criteria below:
   - $GDC \geq 60$
   - Number of conserved residues $Nc \geq 15$
   - Number of residues forming an interface with the ligand $N4 \geq 4$
   - Number of residues potentially clashing with the ligand $cl = 0$
   - For practical reasons, exclude data for ligands whose SMILES string is missing or cannot be parsed.
   - To aid in distinguishing between regions of the protein with different binding behavior, exclude data for non-specific ligands that bind over a third of the residues in a given viral protein.
   - Exclude data for ligands that do not contain at least eight non-hydrogen atoms, which are too small to be relevant.
   - To focus on ligands whose composition may be more relevant for drug discov-

ery, exclude ligands that do not contain at least two of the following atoms: carbon, nitrogen, and oxygen.

2. Construct a weighted, undirected graph for each SARS-CoV-2 protein, in which each node corresponds to a protein residue and each edge indicates that the two connected residues bind one or more of the same ligands.

   - The edge E_{rs}, connecting residues r and s, will have an edge weight w given by
   $$w(E_{rs}) = \frac{1}{\sum_k \delta_{kr}\delta_{ks}} \quad (1)$$
   where δ_{kr} and δ_{ks} indicate the existence of a contact between ligand k and residue r or residue s, respectively.
   - The summation is performed over all ligands in PDBSpheres.
   - The criterion for a ligand-residue contact is defined as follows:
   $$\delta_{kr} = \Theta(C - \min_{i,j} |\mathbf{x}_k^i - \mathbf{x}_r^j|) \quad (2)$$
   where Θ is the Heaviside step function, $\mathbf{x}_k^i$ denotes the position of atom i in ligand k, and $\mathbf{x}_r^j$ denotes the position of atom j in residue r.
   - For this calculation, water atoms are excluded from consideration, and the contact threshold $C = 4.5\mathring{A}$.

3. Perform complete linkage hierarchical clustering. The distance between two residues is given by Eqn. 3.

   - The distance between two residues is given by the minimum sum of edge weights over all possible paths connecting them. That is, if P is a path containing a set of edges connecting residues r and s, then the distance between residues r and s is given by
   $$d(r,s) = \min_P \sum_{E \epsilon P} w(E). \quad (3)$$

4. Make sequential cuts over the resulting dendrogram from bottom to top in distance increments of 0.001. Along the way, select as consensus pockets the largest unique clusters that meet the following two criteria:

   - Contain a mean number of contacts per residue of at least 90. This threshold is based on the observed level of noise in the distribution of contacts per residue (Fig. S1)
   - Consist of a biophysically reasonable number of residues - at least 10 and no more than the maximum number of residues in contact with any ligand for the given viral protein

After selecting consensus pockets in each SARS-CoV-2 protein, we perform another clustering routine to identify the site-specific, core chemical components involved in binding. For the ligands in the PDBspheres data, we extract SMILES strings from PDBE Chem and chemical taxonomy information (chemical Kingdom, Superclass, Class, and Subclass) from the ClassyFire database.[20] Using techniques for natural language processing, we then create for each ligand a word vector embedding to further identify similarities between ligands. We retrieved the PubMed Central Open Access Subset of biomedical literature[21] and joined frequent n-grams using the word2phrase tool in four iterations.[22] We then used the Gensim implementation of Word2vec to train a word embedding with 3,600 dimensions, a 6-token window of optimization, and a minimum token frequency of 5 instances within the corpus.[23] We compute a vector representing each PDBe ligand using the mean embedding of all synonyms for the chemical compound as listed in the PubChem database.[24] An embedding was successfully obtained for 1,156 of the 5,180 small molecule ligands used in our PDBspheres analysis; other ligands did not have sufficient frequency in the corpus to estimate an embedding. From the same filtered PDBspheres data mentioned above, we pick out the ligands that bind to each consensus pocket and calcu-

late the following three distances for each pair of ligands a and b: the Tanimoto distance T_{ab} (calculated using RDKit [25]), the chemical taxonomy distance C_{ab}, and the word vector embedding distance V_{ab}. We define $C_{ab} = 1 - \frac{c_{ab}}{4}$, where c_{ab} is the number of taxonomic levels in common between ligands a and b and $c_{ab} \in [0, 4]$. The word vector embedding distance is the cosine distance between vectors A and B for ligands a and b, or $V_{ab} = 1 - \frac{A \cdot B}{\|A\|\|B\|}$. We construct an overall distance metric W_{ab} as the sum of the available, normalized distances, weighted evenly. For example, if T_{ab}, C_{ab}, and V_{ab} are all available for ligands a and b, then $W_{ab} = \frac{1}{3}T_{ab} + \frac{1}{3}C_{ab} + \frac{1}{3}V_{ab}$. If only T_{ab} and C_{ab} are available, then $W_{ab} = \frac{1}{2}T_{ab} + \frac{1}{2}C_{ab}$. We cluster the ligands using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [26] and our precomputed distance metric. For each binding pocket, we select the parameters ϵ and *min_points* by performing a grid search over the values $\epsilon \in [1, 1.05, 1.10, 1.15, ..., 15]$ and *min_points* $\in [1, 2, 3, ..., 10]$ and choosing the combination that yields the optimal clustering, as evaluated by the silhouette score.

4 Data and Software Availability

The complete PDBspheres library of data is available at https://proteinmodel.org/AS2TS/PDBspheres. Code and input data for the residue and ligand clustering procedures is located https://github.com/LLNL. All other software is open source and cited throughout the paper.

5 Results and Discussion

Characteristics of our results provide confirmation that our method performs as intended and produces reasonable and high quality results. First, groups of residues obtained from the residue clustering procedure appear to colocalize in 3D space, as expected. Second, the following features of the ligand clustering silhouette plots (Fig. S2), which are typical of our results, indicate reasonable separation between ligand clusters and support their validity: mostly positive individual silhouette coefficients for members of each cluster, a positive average silhouette coefficient, and an average coefficient that crosses each of the clusters.

Using the PDBspheres data and residue clustering procedure described above, we identify as relevant binding sites the consensus pockets visualized in Fig. 2 for nsp5, nsp12, and Spike. Lists of the constituent residues for each consensus pocket can be found in the table in Fig. 3 for nsp5 and in Tables S2 and S7 in the Supplemental Section for nsp12 and Spike, respectively. The number of consensus pockets and the nature of those pockets can be used to characterize each viral protein. The bar charts in Fig. 3 (for nsp5) and Fig. S3 (for nsp12 and Spike) provide such an initial characterization, showing the number of consensus pockets identified in each protein, the number of residues comprising each consensus pocket (blue), and the number of ligands that bind to each consensus pocket (orange). For example, nsp5 contains one moderately-sized pocket that binds to an extraordinarily high number of ligands (top panel of Fig. 3).

Previously published results from both experimental and computational studies corroborate the consensus pockets we identify in SARS-CoV-2 proteins. To ensure that the assessment of our method's performance is fair and draws on different data than was used in its development, we exclude as reference points published pockets derived from PDB entries that are used as SARS-CoV-2 protein models in PDBspheres. For Spike, we identify the same pocket as in Ref. 27. The pocket reported in Ref. 28 and Ref. 2 is only identified by our method when the threshold for the required mean number of contacts per residue is lowered significantly. We find that out of 28 residues in Spike-pocket6 and 14 residues in

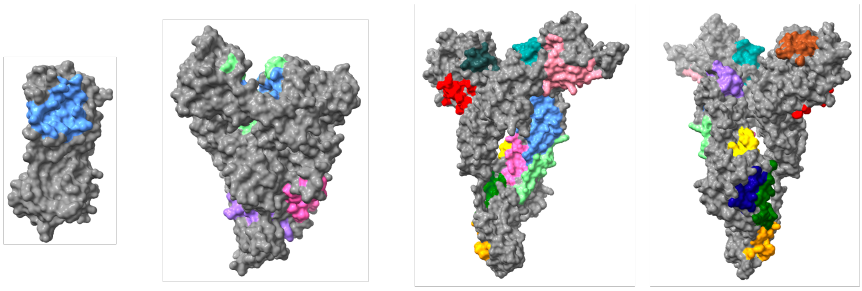
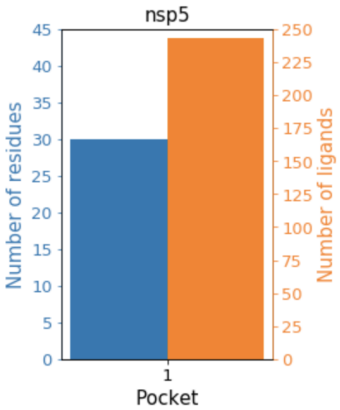


Figure 2: Consensus binding pockets for nsp5 (far left), nsp12 (center left), and Spike (center right and far right). The protein surfaces are colored gray, and within a given protein, each consensus pocket is highlighted in a different color.



Pocket Number	Residues
1	T24, T25, T26, L27, H41, C44, T45, S46, M49, Y54, F140, L141, N142, G143, S144, C145, H163, H164, M165, E166, L167, P168, H172, V186, D187, R188, Q189, T190, A191, Q192

Figure 3: On the top, the bar chart shows the number of constituent residues (blue) and number of ligands (orange) that bind at least 50% of the residues for the consensus pocket in nsp5. On the bottom, consensus pockets for nsp5. All residues belong to the A chain.

the pocket in Ref. 27, nine match. For nsp5, 26 of the 30 residues in pocket 1 appear in at least one instance of the corresponding nsp5 pocket in other papers. Specifically, residues in nsp5-pocket1 coincide with 21 of 24 residues in the pocket in Ref. 5, all 18 residues in the pocket in Ref. 29, and all 16 residues in the pocket in Ref. 30. For nsp12, we note that of the 31 residues in pocket 3, 13 are in common with the 26-residue pocket in Ref. 4, 5 are in common with the 18-residue pocket in Ref. 7, and 17 are in common with the 36-residue pocket in Ref. 31. In addition, of the 18 residues in nsp12-pocket2 and 36 residues in Ref. 31, nine overlap. Of the 25 residues in nsp16-pocket1 and 8 residues in Ref. 32, eight are the same.

Several of the drug binding pockets reported in Ref. 33 overlap with our consensus pockets, providing further support for our results. Although we do not detect the drug binding pockets they observe in nsp3, nsp13, and nsp16, we do capture most of the same pockets for nsp5, nsp12, nsp14, nsp15, and Spike. We find correspondence between our nsp5-pocket1 and the nsp5 catalytic site, our nsp12-pocket4 and the nsp12 NIRAN ADP binding site, our nsp12-pockets 2 and 3 and the nsp12 catalytic site, our nsp14-pocket2 and the nsp14 methyltransferase site, our nsp15-pocket1 and the nsp15 catalytic site, and our Spike-pocket6 and the Spike LA binding pocket.[33] Results from PDBspheres indicate that several ligands are predicted to bind to nsp9, and residues in contact with those ligands overlap with 13 pocket residues reported for nsp9 in Ref. 33; our workflow described in Methods, however, does not detect any consensus pockets in nsp9 at the currently selected thresholds.

Figure 4 provides a visualization of the comparison between our consensus pockets and those reported in the literature. Binding pocket residues are aggregated across all published papers mentioned in the preceding two paragraphs and are colored according to whether they appear only in the literature (blue), only in our results (purple), or in both (green). The moderately sized green regions in nsp12 and Spike show substantial overlap between our results and those in previously published papers. Notably, the pocket in nsp5 is almost entirely green with only small spots of blue and purple, indicating an especially high degree of overlap between our results and others' and thus reinforcing the effectiveness of our approach. Overall, while the performance of our consensus binding pocket identification is somewhat mixed for Spike, we clearly and consistently identify the same pockets for nsp5 and nsp12 that are found in other work.

It is important to note that in performing the residue clustering procedure, we excluded PDB data for known SARS-CoV-2 drugs that are either currently being tested or have already been approved, saving this data as a test set. Even so, our consensus pockets still include many of the same residues that are in contact with those drugs. Table 1 shows the fraction of PDBspheres protein-ligand binding conformations for which at least one residue belongs to one of our consensus pockets. Our consensus pockets cover at least 42% of the conformations for all ten templates in Table 1 and often cover a much higher percentage, up to 100% in two templates. The fact that our method can discern significant binding pockets without some of the original data provides another layer of evidence that our method is performing well and could be reasonably trusted to identify significant pockets in other new viruses.

Comparison of our consensus pockets against the protein domain families in the Pfam database[34] provides additional support for our method. The consensus pocket we identify in nsp5 corresponds to the single Pfam family (Peptidase_C30) for the protein. Pfam designates two families for nsp12 (RdRP_1 and CoV_RPol_N), and two of our four consensus pockets (pockets 2 and 3) match RdRP_1 while the other two (pockets 1 and 4) match CoV_RPol_N. For Spike, all consensus pockets overlap with a Pfam family: 23% of the consensus pockets overlap with bCoV_S1_N, 23% with bCoV_S1_RBD, and 61.5% with CoV_S2. We note that the sum is greater than 100% because one of the consensus pockets maps to two Pfam families; all other

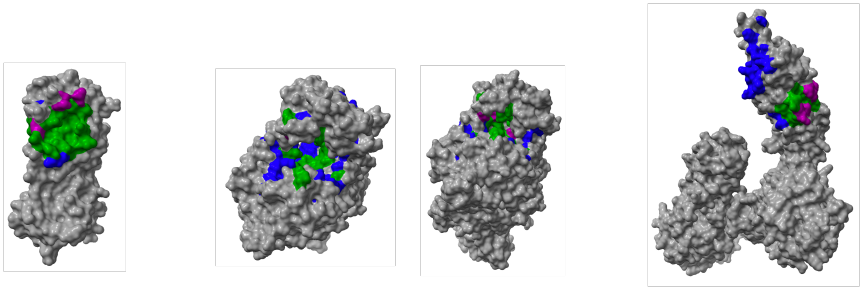


Figure 4: Overlap between published pockets and our consensus pockets for nsp5-pocket1 (far left), nsp12-pocket3 (center left and center right), and Spike-pocket6 (far right). Residues colored blue, purple, and green appear only in published pockets, only in our consensus pocket, or in both, respectively. The results from multiple papers are aggregated for each published pocket. The two images for nsp12 give different views of the same pocket.

PDB Template	Protein	Fraction of PDBspheres protein-ligand binding conformations	Drug Name
7AAP	nsp12	5806/8040 = 0.722	Favipiravir-RTP
7B3B	nsp12	12/20 = 0.6	Remdesivir
7B3C	nsp12	12/20 = 0.6	Remdesivir
7BV2	nsp12	24/32 = 0.75	Remdesivir-RTP
7CTT	nsp12	16/24 = 0.667	Favipiravir
7L1F	nsp12	34/34 = 1.0	Remdesivir monophosphate
7OZU	nsp12	6/14 = 0.429	Molnupiravir-NHC
7OZV	nsp12	6/14 = 0.429	Molnupiravir-NHC
7SI9	nsp5	4/4 = 1.0	PF-07321332
7VH8	nsp5	9/13 = 0.692	PF-07321332

Table 1: For each PDB template corresponding to a known SARS-CoV-2 drug (either currently undergoing testing or already approved), we report the fraction of PDBspheres binding conformations for which at least one residue belongs to one of our consensus pockets.

pockets map to a single family. None of the consensus pockets align with the fourth Pfam family for Spike. These results indicate that our consensus pockets are associated with some Pfam families more than others and do not represent a random distribution of Pfam families. Likewise, the distribution of Pfam families for previously published pockets (mentioned above) is also not random, consistent with our results. Based on these observations, it is possible that certain Pfam families are more or less likely to be significant binding sites and successful drug targets. The fact that our consensus pockets align with known protein domain families reaffirms that our procedure identifies relevant pockets. In addition, as noted in Pfam,[34] the S1 N-terminal domain and the S1 receptor binding domain play a role in binding to host receptors. That many of our consensus pockets for Spike map to these two integral Pfam domains is further evidence that our process selects relevant pockets with an important viral function.

We test the sensitivity of our consensus binding pocket identification approach to the input data by rerunning our residue clustering procedure using different subsets of the PDBspheres data and comparing the resulting consensus pockets. We filter the PDBspheres data by date of deposition of the original data in PDB and the GDC score, which allows us to examine how the passage of time (*i.e.* data availability for a new pathogen) and quality of homology modeling affect the performance of our approach. The adjusted rand index (ARI) serves as a metric for evaluating the similarity of the consensus pockets across different subsets of the data. An ARI of one indicates that two sets of clusters (*i.e.* consensus pockets) are identical, while an ARI of zero indicates that the clusterings have no clusters with common elements (*i.e.* residues). A plot of the ARI across different date cutoffs in Fig. 5 shows that our predicted consensus pockets are very stable for nsp5 over different input data but more variable for nsp12 and Spike. Images on the right side of Fig. 5 illustrate the overlap between pairs of clusterings and thus provide a physical context for interpreting the ARI scores in the plot. Although the ARI is noticeably lower for the earliest two date cutoffs for nsp12, it covers a more modest range over the rest of the dates and is relatively consistent from 06/23/2020 through 10/05/2021. While the ARI for Spike drops consistently from more recent to older date cutoffs, our process still identifies the same literature-confirmed pocket at every date cutoff except 07/09/2019 and 03/10/2020, when data on SARS-CoV-2 was more limited. Since more data on a new pathogen will be available at later dates, we would naturally expect that predictions from our data-driven approach will improve over time and that there will be some moderate time-sensitivity in our results. Yet even early on in the pandemic, when data on SARS-CoV-2 was limited, there seems to be enough data from other related organisms for our framework to detect reasonably comparable binding pockets. Together, the findings from these internal validation tests lend support for the effectiveness of both our structural modeling methods and consensus pocket identification pipeline.

To further assess the robustness of our methods, we examine the relationship between the GDC metric in the PDBspheres data and the date cutoff and source organisms of the underlying PDB experimental complexes. For each organism category, we record the GDC scores of all protein-ligand binding conformations in the PDBspheres data with a source organism belonging to that category and plot the distribution of scores. In the case of PDB complexes with more than one source organism, we count the GDC score towards all source organisms, since our approach does not distinguish which parts of a PDB complex originate from which source organism. As shown in Fig. 6, for source organisms SARS-CoV-1 and SARS-CoV-2, the distribution of GDC scores has one peak between 90 and 95 with a long, thin tail at lower GDC values, indicating a high level of structural similarity with the SARS-CoV-2 proteins of interest, as expected. The distribution for other viruses similarly peaks around 90-95, but it carries more

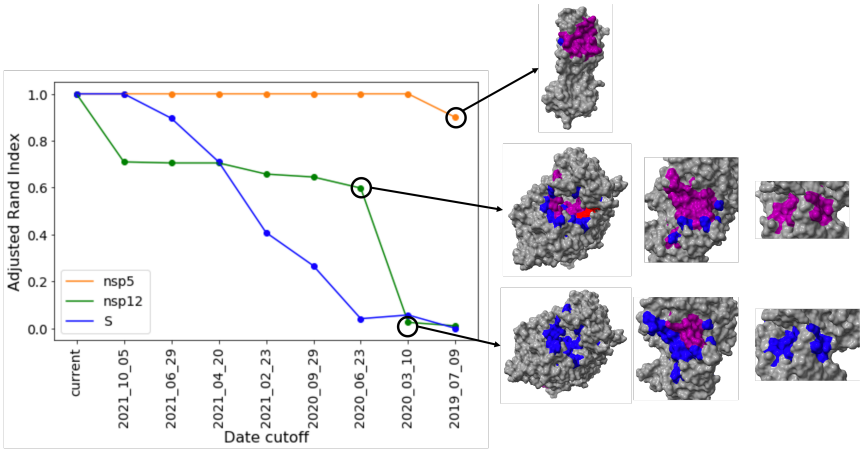


Figure 5: On the left, a plot of the adjusted rand index (ARI) for the consensus pockets in nsp5, nsp12, and Spike as a function of date cutoff in the PDBspheres data. The GDC cutoff is held fixed at 60. An ARI of one indicates that two sets of consensus pockets are identical, while an ARI of zero indicates that the consensus pockets in each set have no residues in common. The reference consensus pockets are those associated with a GDC cutoff of 60 and a current date cutoff. On the right, images of the overlap between the reference clustering and three comparison clusterings noted with a black circle. Residues colored blue, red, and purple appear only in the consensus pockets of the reference point, only in the consensus pockets of the point of comparison, or in both, respectively. The top, middle, and bottom rows illustrate high, medium, and low ARI values.

density across lower GDC values than do the distributions for SARS-CoV-1 and -2. In contrast, for other organism classifications, with the exception of insects, the distribution either peaks around 60-70 or is more uniform across the range of GDC values, signaling a lower degree of structural similarity with SARS-CoV-2 proteins. The peak at high GDC in the insect distribution is an artifact and can be attributed to a single PDB complex with components from both SARS-CoV-2 and the saltans group. As noted above, the GDC score for this complex counts towards both the SARS-CoV-2 category and the insect category, even though the high score is almost certainly a result of the SARS-CoV-2 component and its high degree of similarity with the SARS-CoV-2 nsp3 structural model. These results suggest that the less closely related an organism is to SARS-CoV-2, the less structurally similar its pockets are to those in SARS-CoV-2. This result reinforces why having data on a similar organism is especially valuable: it enables us to more successfully model the proteins of the organism of interest.

The trend between the amount of PDBspheres data and time is also informative. As expected, the number of protein-ligand pairs in PDBspheres with a GDC score above 60 increases over time for all SARS-CoV-2 proteins (Fig. S5, left panel). We note that the number of protein-ligand pairs grows faster and reaches a higher point for Spike than for other proteins simply because there are more unique protein templates for Spike, as seen in the right panel of Fig. S5. We can conclude that pockets detected through PDBspheres are stable over time. That is, once PDBspheres detects a pocket, it continues to detect the pocket even as additional, newer data becomes available. Therefore, we can be confident in the quality of the PDBspheres structural models even early in a pandemic when experimental structural data on a new pathogen is limited. Predictions from our data-driven approach should improve over time as the amount of data about a new pathogen grows.

In-depth analysis of the composition of the consensus pockets in terms of their underlying PDBspheres models both provides additional context for our results and highlights another capability of our approach. For each consensus pocket, we determine which PDBspheres models correspond to the protein-ligand conformations that involve that pocket. For each relevant protein-ligand conformation, we include as contributing models both the model for the target protein of interest and the model for the protein with the observed or predicted ligand binding. For each contributing model, we then retrieve from PDB the source organism for the experimental measurement from which the model was generated. If a single model corresponds to more than one organism, all organisms are counted for that model. Pie charts of the contributions from various organism classifications allow for straightforward visual comparisons across different data filters, as shown in Fig. 7. Observing the evolution of organism composition over time, we see that, by definition, there is no contribution from SARS-CoV-2 in consensus pockets identified from pre-pandemic data and that the SARS-CoV-2 contribution grows over the course of the pandemic as more data on the new virus is collected. The greater the contribution from other viruses to the composition of a given SARS-CoV-2 consensus pocket, the more structurally similar that pocket is to pockets in other viruses. It seems reasonable to think that binding pockets in SARS-CoV-2 that are more structurally similar to binding pockets in other viruses might be more effective as drug targets. Therefore, the source organism composition of consensus pockets could be utilized as a preliminary means of narrowing in on promising drug targets.

Based on this reasoning, we examine the organism composition profiles of the consensus pockets from a range of SARS-CoV-2 proteins and calculate a simple ranking of the pockets' drug target potential. The score S for each pocket p is determined by the size of the viral contribution (including SARS-CoV-1, SARS-CoV-2, and other Viruses) relative to the overall

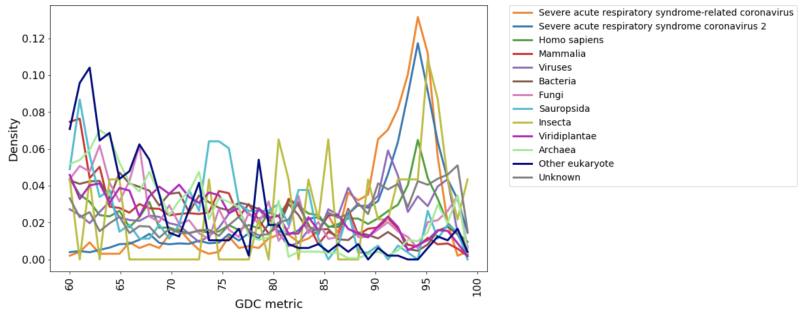


Figure 6: Distributions of the GDC metric for PDBspheres structural models from various organism classifications.

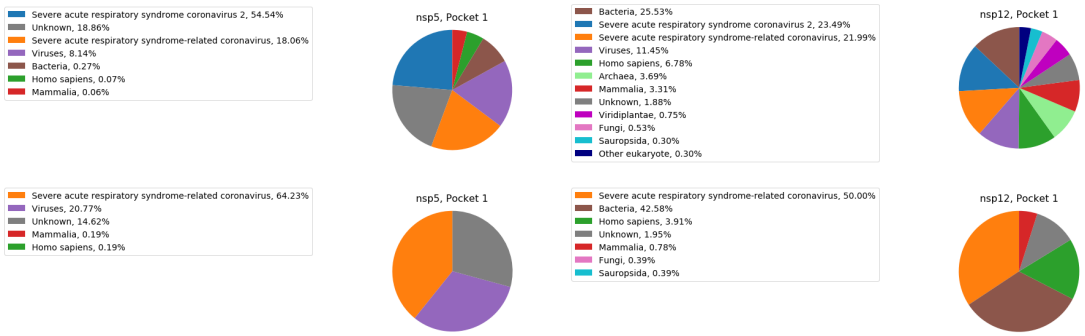


Figure 7: Composition of consensus pockets in terms of the source organisms for the underlying PDBspheres models, shown on a log scale. The left and right columns show corresponding consensus pockets for nsp5 and nsp12, respectively, using currently available data (top row) and pre-pandemic data (bottom row). Pockets were numbered consistently by inspection between iterations of our pipeline across different date cutoffs.

contribution from known organisms and is given by

$$S_p = \frac{(C_{SARS-1} + C_{SARS-2} + C_{Viruses})}{(1 - C_{Unknown})}, \quad (4)$$

where C_i is the fraction of pocket p's pie chart that component i makes up. Pockets are ordered from highest to lowest score, with higher scores indicating a greater degree of specificity for viral sites in comparison to sites from humans or other organisms. Fig. 8 shows the highest scoring pocket for each protein and the range of composition profiles observed across consensus pockets. Proteins without enough PDBspheres data for determination of consensus pockets are excluded from this analysis. Preferred profiles, like those for nsp5-pocket1 and nsp12-pocket2, tend to have fewer different contributing organisms, a larger share of the composition derived from viruses, and minimal contribution from humans (which could, though not necessarily, indicate a potential for undesirable off-target interactions). We note that it is possible for a pocket to have a larger contribution from SARS-CoV-1 than SARS-CoV-2, as is the case with nsp14-pocket1, if the contributing models derived from SARS-CoV-1 are predicted to bind with a larger number of ligands and therefore appear more frequently in the PDBspheres data. Comparing the

composition of all consensus pockets identified for Spike, nsp5, and nsp12, we find that overall, with some exceptions, these three preferable features are more common in the profiles for nsp5 and nsp12 than in Spike, as shown in Fig.S6. Together, these observations suggest that the consensus pockets in nsp5 and nsp12 are more structurally similar to binding pockets in other viruses and are structurally conserved, which may help explain why our method seems to perform better for nsp5 and nsp12 than for Spike. Indeed, RNA Dependent RNA Polymerases like nsp12 are known to be highly conserved across different viruses.[35] Given that half of our nsp12 pockets map to the Pfam RdRp_1 family, the structural similarity we detect between the pockets in nsp12 and other viruses is not only reasonable but expected. If nsp5 and nsp12 share more structural similarities with other viruses, then we may be more successful at modeling them using structural data from other viruses.

Exploring the PDBspheres data from a ligand-centric perspective provides further insight into patterns of protein-ligand binding, as well as an additional indicator of a pocket's drug target potential. Clustering of the ligands that bind to each consensus pocket reveals groups of ligands with clearly identifiable and distinct features, such as functional groups, molecular size, and geometry. Together these features define a unique chemotype for each cluster, as seen in the example ligands for nsp5-pocket1 on the right side of Fig. 9, and the collection of chemotypes for a consensus pocket can be used to characterize it.

Analysis of the fraction of ligands in each cluster that contact each residue in a given pocket enables us to locate subregions of the pocket that correspond to specific chemotypes. As seen in the heatmaps on the left side of Fig. 9, we find that different ligand clusters have similar patterns of contact fractions for the residues in the pocket. We group the ligand clusters according to the pattern of contact fractions and calculate the average contact fraction for each residue over the clusters in the super-cluster. This average contact fraction determines the color of the residue in the 3D images of the protein surface, shown in the center of Fig. 9. As seen on the right side of Fig. 9, we observe that larger subregions correspond to larger chemotypes, a physically reasonable result. Our ligand clustering method thus provides a finer resolution view of our consensus pockets and the ability to hone in on smaller regions of interest for specific chemotypes.

Finally, results from experimental assays[1] lend support for our approach. Using Tanimoto similarity, we compare the structure of compounds tested against nsp5 to the structure of the ligands in each ligand cluster for pocket 1. For each experimental compound, we determine the ligand cluster to which it is closest on average. Then for each cluster, we compare the distributions of average Tanimoto distances for the positive and negative compounds assigned to that cluster. Fig. 10 displays these distributions as well as the results of a t-test on the mean values, omitting clusters for which no experimental compounds were assigned or too few compounds were assigned to statistically compare the positive and negative distributions. P-values above or below the significance level of $\alpha = 0.05$ indicate that the mean Tanimoto distances for positive and negative compounds are statistically equivalent or different, respectively. While the positive and negative compounds are roughly the same distance from two of the clusters, for four other clusters, the positive compounds are on average closer and thus more similar in structure than are the negative compounds. The finding that experimentally observed positive compounds align more closely with our clusters than do negative compounds affirms that our ligand clustering method detects meaningful chemical features associated with effective binding. It also reinforces the hypothesis that the more drug-like the molecules that bind to a pocket, the more promising it may be as a drug target. Comparing such histograms across ligand clusters and identifying those with the greatest degree of separation between the positive and negative distri-

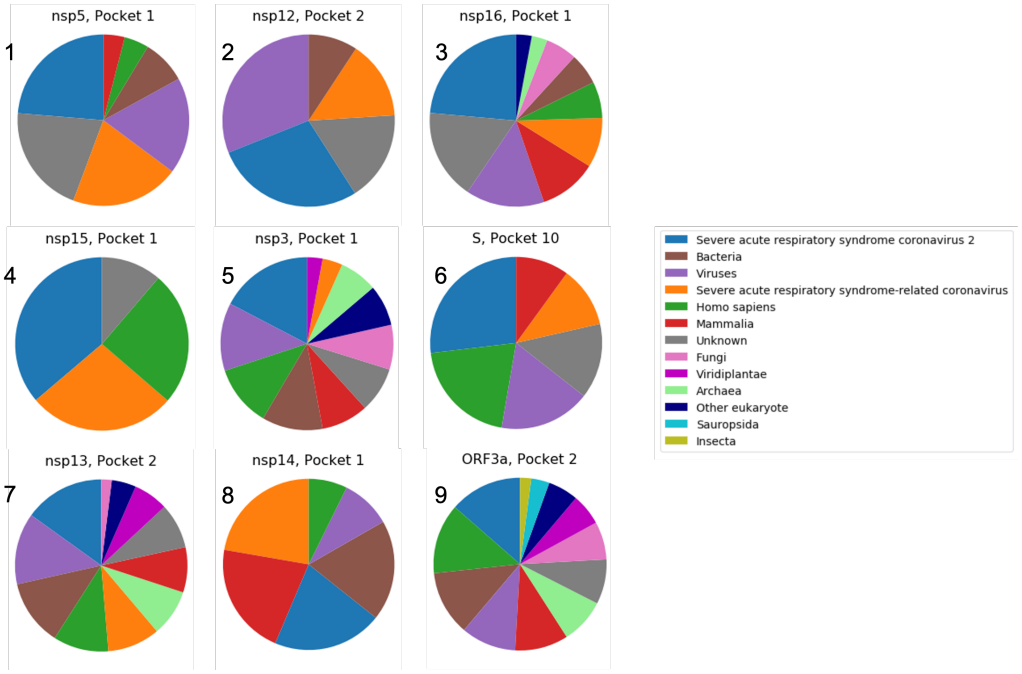


Figure 8: Composition of consensus pockets in terms of the source organisms for the underlying PDBspheres models, shown on a log scale. The top pocket for each protein is shown, as determined by the size of the viral contribution (SARS-CoV-1, SARS-CoV-2, plus other Viruses) relative to the overall contribution from known organisms. Composition profiles are numbered from most to least preferred.

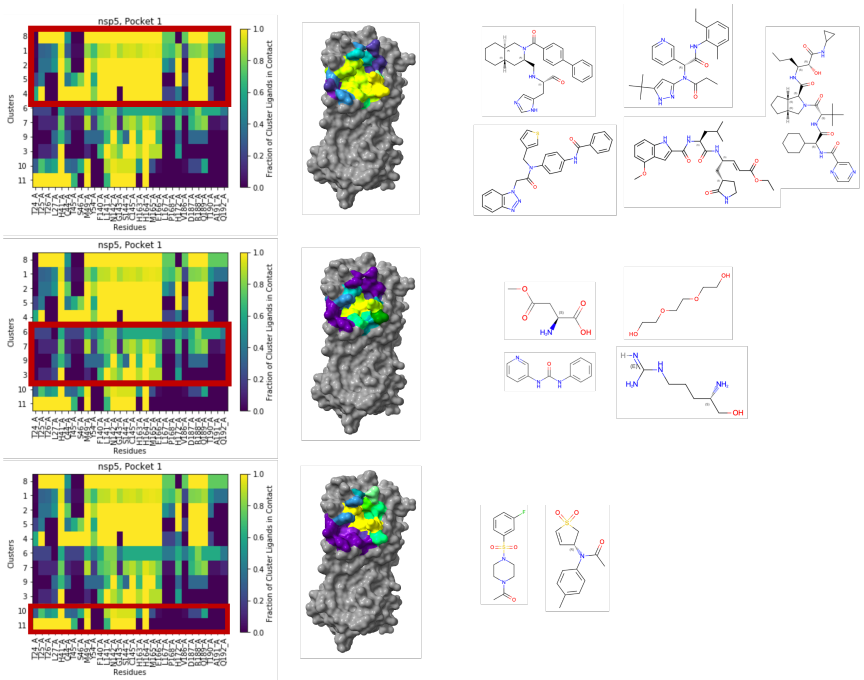


Figure 9: On the left, heatmaps of the fraction of ligands in each cluster with which residues in nsp5-pocket1 come into contact. In the center, 3D images of the surface of nsp5 with pocket residues colored by the average fraction within the supercluster designated in the red box in the corresponding heatmap. On the right, images of one representative molecule from each ligand cluster in the corresponding supercluster.

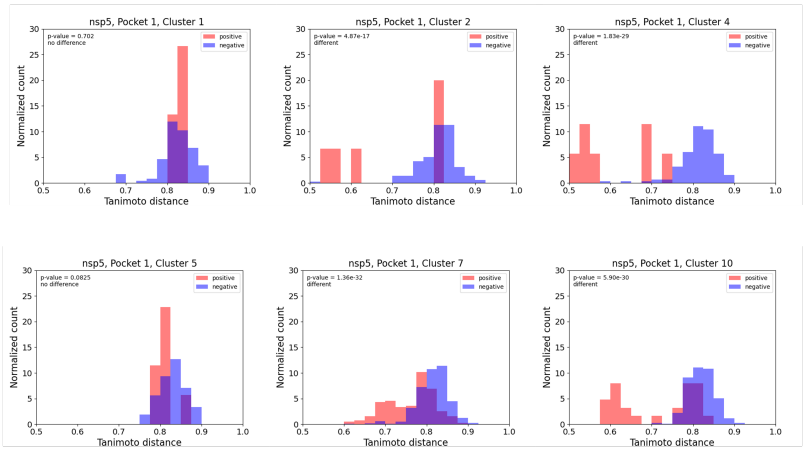


Figure 10: Normalized histograms of the average Tanimoto distance between experimentally-tested compounds and their closest ligand cluster. The distributions for positive and negative compounds are shown in red and blue, respectively. P-values shown on each plot are for the associated t-test comparing the means of the positive and negative distributions.

butions could signal which chemotypes deserve further attention and study. Assessing the drug likeness of site-specific ligands, either through standard drug-likeness tests or by calculating average Tanimoto similarity against a library of known drugs, may serve as an additional means for identifying promising drug targets.

6 Conclusion

Here we present a new approach for detecting and characterizing relevant binding pockets in the proteins of novel pathogens and for uncovering the site-specific, core chemical substructures associated with protein-ligand binding. Our process combines molecular level detail from protein structural modeling with machine learning clustering techniques and cheminformatics tools. Our predictive pipeline is fast, streamlined, and broadly applicable. While the quality of available protein structural models is a limiting factor, we account for this potential issue by considering a quantitative measure of model quality, the GDC score. Only structural models above a desired quality are included in our downstream analysis. Our framework can be deployed as soon as a new pathogen is discovered and can generate results in a matter of days. We find that it performs reasonably well even when experimental structural data on a new pathogen is limited, and our predictions should improve over time as more data becomes available. With a basis in experimental data and few parameters, our system is straightforward and can be readily extended to new pathogens, provided their genomic sequence is available.

In this paper, we focus on SARS-CoV-2 as a case study for explaining our system and assessing its performance. Our predictions of consensus binding pockets in the nsp5, nsp12, nsp14, nsp15, nsp16, and Spike proteins compare favorably with those already published in the literature, including some detected through experiment and others via computational methods. While our consensus pocket identification method is somewhat sensitive to the input data, modest variations like those we observe are reasonable and to be expected. Even though the similarity metric declines going back in time, in many cases we still identify the pockets reported in other published papers. Over time, as more and more data is collected, the performance of our framework should improve.

As expected, we observe that the more closely related an organism is to SARS-CoV-2, the more structurally similar its binding pockets are to those in SARS-CoV-2. Analysis of the composition of source organisms of our consensus binding pockets reveals that nsp5 and nsp12 are more structurally conserved viral proteins than is Spike.

Through our ligand clustering method, we identify characteristic chemotypes for each consensus pocket as well as the subregions of the pocket with which each chemotype interacts. Significantly, we find that positive experimental hits against nsp5 are on average more similar in structure to our ligand clusters than are the negative compounds from the screening – evidence that the core chemical components we detect are in fact relevant.

Based on our results, we propose two simple heuristics for which consensus pockets will make more effective drug targets. A SARS-CoV-2 binding pocket may hold more promise as a drug target 1) the more similar its structure is to that of other viral binding pockets and 2) the more drug-like the molecules that bind. Our pipeline offers a means of assessing these two factors, through source organism composition analysis and ligand clustering. Though they have so far received less attention as potential drug targets and hence have more limited data, our results suggest that pockets in nsp3, nsp13, and nsp16 may also offer some promise and should be studied further, as additional data would provide more clarity and confidence. Overall, our system lays the groundwork for a rapid response tool that can quickly identify promising drug targets and target-specific drug compound components when faced with an emerging pathogen, and ongoing work will enhance these capabilities.

7 Supporting Information

Additional data and figures in support of our methods as well as a list of the constituent residues of the consensus pockets identified for each SARS-CoV-2 protein

8 Acknowledgements

Molecular graphics created with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases. See Ref. 36 and 37 for more details. Financial support for this work is provided by the Laboratory Directed Research and Development program at Lawrence Livermore National Laboratory (20-ERD-062).

References

[1] Stevenson, G. A. et al. High-Throughput Virtual Screening of Small Molecule Inhibitors for SARS-CoV-2 Protein Targets with Deep Fusion Models. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. New York, NY, USA, 2021; event-place: St. Louis, Missouri.

[2] Deganutti, G.; Prischi, F.; Reynolds, C. Supervised molecular dynamics for exploring the druggability of the SARS-CoV-2 spike protein. *Journal of Computer-Aided Molecular Design* **2021**, *35*, 195–207.

[3] Ortega, J. T.; Serrano, M. L.; Jastrzebska, B. Class A G Protein-Coupled Receptor Antagonist Famotidine as a Therapeutic Alternative against SARS-CoV2: An In Silico Analysis. *Biomolecules* **2020**, *10*, 954.

[4] Mirza, M.; Froeyen, M. Structural elucidation of SARS-CoV-2 vital proteins: Computational methods reveal potential drug candidates against main protease, Nsp12 polymerase and Nsp13 helicase. *J Pharm Anal* **2020**, *10*, 320–28.

[5] Choudhary, M.; Shaikh, M.; tul Wahab, A.; ur Rahman, A. In silico identification of potential inhibitors of key SARS-CoV-2 3CL hydrolase (Mpro) via molecular docking, MMGBSA predictive binding energy calculations, and molecular dynamics simulation. *PLOS ONE* **2020**, *15*, e0235030.

[6] Komatsu, T. S.; Okimoto, N.; Koyama, Y. M.; Hirano, Y.; Morimoto, G.; Ohno, Y.; Taiji, M. Drug binding dynamics of the dimeric SARS-CoV-2 main protease, determined by molecular dynamics simulation. *Scientific Reports* **2020**, *10*, 16986.

[7] Gervasoni, S.; Vistoli, G.; Talarico, C.; Manelfi, C.; Beccari, A.; Studer, G.; Tauriello, G.; Waterhouse, A.; Schwede, T.; Pedretti, A. A Comprehensive Mapping of the Druggable Cavities within the SARS-CoV-2 Therapeutically Relevant Proteins by Combining Pocket and Docking Searches as Implemented in Pockets 2.0. *Int J Mol Sci* **2020**, *21*, 5152.

[8] Shi, Y.; Zhang, X.; Mu, K.; Peng, C.; Zhu, Z.; X. Wang, Z. X.; Zhu, W. D3Targets-2019-nCoV: a webserver for predicting drug targets and for multi-target and multi-site based virtual screening against COVID-19. *Acta Pharmaceutica Sinica B* **2020**, *10*, 1239–48.

[9] Zemla, A.; Allen, J. E.; Kirshner, D.; Lightstone, F. C. PDBspheres - a method for finding 3D similarities in local regions in proteins. *bioRxiv* **2022**,

[10] Schneider, S.; Zacharias, M. Combining geometric pocket detection and desolvation

properties to detect putative ligand binding sites on proteins. *J Struct Biol* **2012**, *180*, 546–50.

[11] Yeturu, K.; Chandra, N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinformatics* **2008**, *9*, 543.

[12] Batista, J.; Hawkins, P. C.; Tolbert, R.; Geballe, M. T. SiteHopper - a unique tool for binding site comparison. *Journal of Cheminformatics* **2014**, *6*, P57.

[13] Zhong, S.; Jr., A. M. Binding response: a descriptor for selecting ligand binding site on protein surfaces. *J Chem Inf Model* **2007**, *47*, 2303–15.

[14] Kupas, K.; Ultsch, A.; Klebe, G. Large scale analysis of protein-binding cavities using self-organizing maps and wavelet-based surface patches to describe functional properties, selectivity discrimination, and putative cross-reactivity. *Proteins* **2008**, *71*, 1288–306.

[15] Simonovsky, M.; Meyers, J. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *J Chem Inf Model* **2020**, *60*, 2356–66.

[16] Bhadra, A.; Yeturu, K. Site2Vec: a reference frame invariant algorithm for vector embedding of protein–ligand binding sites. *Machine Learning: Science and Technology* **2021**, *2*, 015005.

[17] Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res* **2003**, *31*, 3370–74.

[18] Keedy, D.; Williams, C.; Headd, J.; III, W. A.; Chen, V.; Kapral, G.; Gillespie, R.; Block, J.; Zemla, A.; Richardson, D.; Richardson, J. The other 90% of the protein: Assessment beyond the Cas for CASP8 template-based and high-accuracy models. *Proteins* **2009**, *77*, 29–49.

[19] Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference* **2008**, 11–15.

[20] Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. ClassyFire: Automated Chemical Classification With A Comprehensive, Computable Taxonomy. *Journal of Cheminformatics* **2016**, *8*, 61.

[21] PubMed Central Open Access Subset. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/, 2021; 2022-01-12.

[22] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. *Neural information processing systems* **2013**,

[23] GENSIM: topic modeling for humans. https://radimrehurek.com/gensim/, 2021; 2022-01-12.

[24] Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research* **2019**, *47*, D1102–D1109.

[25] RDKit: Open-source cheminformatics. https://www.rdkit.org, 2020; 2022-01-13.

[26] Ester, M.; Kriegel, H. P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* **1996**, *8*, 226–231.

[27] Toelzer, C.; Gupta, K.; Yadav, S.; Borucu, U.; Davidson, A.;

Williamson, M.; Shoemark, D.; Garzoni, F.; O. Staufer, R. M.; Capin, J.; Mulholland, A.; Spatz, J.; Fitzgerald, D.; Berger, I.; Schaffitzel, C. Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. *Science* **2020**, *370*, 725–30.

[28] Trezza, A.; Iovinelli, D.; Santucci, A.; Prischi, F.; Spiga, O. An integrated drug repurposing strategy for the rapid identification of potential SARS-CoV-2 viral inhibitors. *Scientific Reports* **2020**, *10*, 13866.

[29] Dai, W. et al. Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science* **2020**, *368*, 1331–35.

[30] Zhang, C.-H. et al. Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations. *ACS Central Science* **2021**, *7*, 467–75.

[31] Hillen, H.; Kokic, G.; Farnung, L.; Dienemann, C.; Tegunov, D.; Cramer, P. Structure of replicating SARS-CoV-2 polymerase. *Nature* **2020**, *584*, 154–56.

[32] Lin, S.; Chen, H.; Ye, F.; Chen, Z.; Yang, F.; Zheng, Y.; Cao, Y.; Qiao, J.; Yang, S.; Lu, G. Crystal structure of SARS-CoV-2 nsp10/nsp16 2'-O-methylase and its implication on antiviral drug design. *Signal Transduction and Targeted Therapy* **2020**, *5*, 131.

[33] Yazdani, S.; Maio, N. D.; Ding, Y.; Shahani, V.; Goldman, N.; Schapira, M. Genetic Variability of the SARS-CoV-2 Pocketome. *Journal of Proteome Research* **2021**, *20*, 4212–4215.

[34] Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A. Pfam: The protein families database in 2021. *Nucleic Acids Research* **2021**, *49*, D412–D419.

[35] Venkataraman, S.; Prasad, B. V. L. S.; Selvarajan, R. RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. *Viruses* **2018**, *10*, 76.

[36] Goddard, T.; Huang, C.; Meng, E.; Pettersen, E.; Couch, G.; Morris, J.; Ferrin, T. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **2018**, *27*, 14–25.

[37] Pettersen, E.; Goddard, T.; Huang, C.; Meng, E.; Couch, G.; Croll, T.; Morris, J.; Ferrin, T. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **2021**, *30*, 70–82.