Mapping protein conformational landscapes from crystallographic drug fragment screens

*Ammaar A. Saeed[1], Margaret A. Klureza[2], and Doeke R. Hekstra[1,3]**

[1]Department of Molecular & Cellular Biology, Harvard University, Cambridge, MA 02138
[2]Department of Chemistry & Chemical Biology, Harvard University, Cambridge, MA 02138
[3]School of Engineering & Applied Sciences, Harvard University, Cambridge, MA 02138
*Corresponding author. Email: doeke_hekstra@harvard.edu

Keywords: Conformational landscape, PCA, crystallographic drug fragment screen, PTP-1B, MPro

Abstract

Proteins are dynamic macromolecules. Knowledge of a protein's thermally accessible conformations is critical to determining important transitions and designing therapeutics. Accessible conformations are highly constrained by a protein's structure such that concerted structural changes due to external perturbations likely track intrinsic conformational transitions. These transitions can be thought of as paths through a conformational landscape. Crystallographic drug fragment screens are high-throughput perturbation experiments, in which thousands of crystals of a drug target are soaked with small-molecule drug precursors (fragments) and examined for fragment binding, mapping potential drug binding sites on the target protein. Here, we describe an open-source Python package, COLAV (COnformational LAndscape Visualization), to infer conformational landscapes from such large-scale crystallographic perturbation studies. We apply COLAV to drug fragment screens of two medically important systems: protein tyrosine phosphatase 1B (PTP-1B), which regulates insulin

signaling, and the SARS CoV-2 Main Protease (MPro). With enough fragment-bound structures, we find that such drug screens also enable detailed mapping of proteins' conformational landscapes.

Introduction

While often shown as single structures, proteins exhibit dynamic behavior necessary for their function[1-3], e.g. binding and releasing ligands[4], modulating activity[5], and reversibly shielding the active site[6]. Hence, proteins are better thought of as populating ensembles of structural states or conformations. Individual protein molecules transition frequently between these conformations through the concerted motions of their amino acids. For many proteins, there are only a handful of accessible backbone conformations at physiological temperatures, all separated by distinct concerted motions[2,7].

Consequently, proteins can often be thought of as residing on a conformational landscape that describes metastable conformations and the concerted motions necessary to transition between them[8]. Ideally, conformational landscapes would be inferred from experimental structures and would succinctly recapitulate the known conformational diversity of a target protein. Additionally, these empirical landscapes would suggest thermally accessible concerted motions between conformations—probable temporal sequences of conformational change sometimes referred to as conformational reaction coordinates or transition paths[9-11]. Such conformational landscapes for validated protein drug targets would suggest particular conformations to (de)stabilize to enhance or inhibit functional activity. These conformations can then be targeted

by the design of a small molecule that binds the drug target within the active site (orthosteric) or elsewhere (allosteric).

Existing biophysical methods can experimentally characterize aspects of a protein's conformational landscape, e.g., by nuclear magnetic resonance (NMR) spectroscopy[12], fluorescence resonance energy transfer spectroscopy[13], electron paramagnetic resonance spectroscopy[14], and room-temperature X-ray crystallography[6,15]. These techniques probe the equilibrium distribution of a desired conformational ensemble. However, such measurements generally reflect the ground state of the protein and only provide limited insight into the presence and/or nature of any alternate, higher-energy conformations. For large proteins and protein complexes, cryogenic electron microscopy (cryo-EM) and electron tomography (cryo-ET) can capture small populations of metastable conformations directly[16], and machine learning methods are beginning to pave the way for the identification of these rare protein states[17,18]. Yet, determining high-resolution structures of metastable states through cryo-EM or cryo-ET remains an ongoing challenge, due to the need for a vast quantity of correctly classified particle images.

An alternative approach to studying these excited states is to directly perturb the protein of interest. These perturbations alter the conformational landscape, stabilizing otherwise short-lived excited states. Common methods to introduce perturbations include mutation of the protein and addition of substrate/transition-state analogs. Once the protein has been perturbed, the stabilized states can be examined via standard biophysical techniques. Though the efficacy of this approach has been demonstrated in a variety of model systems[19-21], designing individual perturbations can be time-consuming and may only explore a limited portion of the conformational landscape.

An ideal approach to mapping protein conformational landscapes would be to subject the protein of interest to a large number of distinct perturbations that are just strong enough to bias the energetics of particular conformations by a few k_BT and then determine the structure of the protein under each perturbation[22,23]. Crystallographic drug fragment screens constitute an intriguing approximation to this ideal experiment: in these high-throughput crystallographic screens, many crystals of the same drug target are each soaked with a unique drug fragment and are then subjected to the standard X-ray crystallography pipeline. Advances in automation at the Diamond Light Source[24] and elsewhere, paired with novel data processing software[25,26], have enabled these screens to solve thousands of protein structures within days, some of which contain bound drug fragments. Importantly, these drug fragment screens may yield information valuable for drug design beyond the immediate identification of drug fragment/binding site pairs: a comprehensive exploration of the protein's conformational landscape.

To test this idea, we developed a software package known as COLAV (COnformational LAndscape Visualization) that calculates three different representations of protein structure—dihedral angles, pairwise distances, and strain—to quantify structural change across a group of crystal structures. COLAV is an open-source, Python-based software, freely available at https://github.com/Hekstra-Lab/colav. Using COLAV, we show that sets of crystal structures can be used to construct a map of a protein's conformational landscape and infer correlated regions within the protein. We then ask whether the conformational landscape constructed from structures obtained only from a crystallographic drug fragment screen is consistent with a map of the landscape based on structures obtained using a variety of perturbations (e.g., mutants,

substrate analogs, and inhibitors) available from the Protein Data Bank (PDB)[27]. We find that the drug fragment-derived map provides a partial view of the conformational landscape that is consistent with the landscape derived from the complete dataset. The drug fragment-derived map becomes substantially more complete with increasing scale of the crystallographic drug fragment screen.

Methods

Structural representations

We implemented three methods to represent a protein structure in COLAV: backbone dihedral angles (ϕ, ω, and ψ), pairwise distances between Cα atoms, and strain. We implemented these methods on top of the Scientific Python stack (NumPy[28], SciPy[29], and BioPandas[30]). Dihedral angles and distances were calculated according to standard methods, and strain was calculated according to previously published frameworks[31,32] described briefly below. To ensure consistent features across each protein dataset, we truncated structures at the N and C termini and then removed any structures missing backbone atoms between the truncated endpoints. For PTP-1B, we calculated representations between residues 7 and 279 (inclusive). For "focused PCA" of the PTP-1B L16 loop, we only used representations between residues 236 and 244 (inclusive). For MPro, we calculated representations between residues 3 and 297 (inclusive). If alternate conformations had been modeled for any atoms, then we included only the "A" conformer in our calculations. In our strain implementation, we calculated three different variants of strain: strain tensor, shear tensor, and shear energy. We used the off-diagonal elements of the shear tensor as inputs for principal component analysis (PCA). Use of COLAV is illustrated in the accompanying Jupyter Notebooks available at https://github.com/Hekstra-Lab/colav.

Data analysis

We analyzed these structural representations using the Scikit-Learn implementation of PCA, using 10 principal components (PCs) and otherwise default parameters[33]. Because of the inherent periodicity present in dihedral angles, we linearized these features by calculating the sine and cosine of each angle and using the resulting tuple as the input feature for PCA. To determine a per-residue measure of importance for each method ("residue contributions"), we transformed the coefficients of the principal components as follows. For dihedral angles, we first summed the absolute values of the sine and cosine coefficients of the same dihedral angle to determine a per-angle, per-residue measure. We also summed the absolute values of these per-angle measures into a single per-residue measure. For the pairwise distance representation, we summed the absolute value of all coefficients pertaining to each residue. For the strain-based representation, we summed the absolute value of the off-diagonal elements of the shear matrix for each residue.

We also analyzed these structural representations using the Scikit-Learn implementation of t-distributed Stochastic Network Embedding[34] (t-SNE) and the Umap-Learn implementation of Uniform Manifold Approximation and Projection[35] (UMAP). We initialized both of these latter methods randomly; we did not observe major differences in the clustering of structures when using different seeds. To identify groupings of structures similar to each other in the MPro dataset, we used the Scikit-Learn implementation of the *k*-means algorithm with default settings[33]. In our assessment of the role of dataset size, we generated MPro datasets of varying size by sampling the complete MPro dataset (without replacement) each time.

To establish the coupling between regions of PTP-1B, we performed Fisher exact tests for independence (https://www.socscistatistics.com/tests/). This test asserts as a null hypothesis that the variables used are independent and as an alternative hypothesis that there is a dependence structure among the variables. We tested for conditional independence by adding the chi-square statistics of two two-way tests and comparison to the null distribution (chi-square with two degrees of freedom) as described in Ch. 5, "Analysis of Discrete Data", (https://online.stat.psu.edu/stat504/book/).

Dataset construction

For PTP-1B, we retrieved 165 structures of the human enzyme from the Protein Data Bank (PDB) in March 2022 with a sequence identity of 90% or higher compared to wild-type PTP-1B. We also retrieved 187 structures of PTP-1B bound to fragment ligands from a crystallographic drug fragment screen[36] that were identified either by Pan-Dataset Density Analysis (PanDDA)[25] alone or after tandem processing by cluster4x[26] and PanDDA. We retrieved all PTP-1B files in the PDB file format (hereafter .pdb).

For MPro, we retrieved all 1,830 crystallographic drug fragment screen structures in March 2022 from the Fragalysis database[37-41]. We retrieved all 1,015 other MPro structures from the PDB in July 2023. We excluded MPro structures from an ensemble refinement study of MPro at multiple temperatures (7MHL, 7MHM, 7MHN, 7MHO, 7MHP, 7MHQ)[42]; these temperature-induced effects dominated the analysis, masking the native conformational landscape of MPro. Several MPro structures were too large to download in the .pdb format, so we downloaded them in the

mmCIF file format. We subsequently converted them to the .pdb format using an online GEMMI tool[43].

Before feature extraction, we aligned structures of PTP-1B or MPro using THESEUS v3.3.0[44], as superposing structures of the same protein was crucial for proper strain calculations. Where noted, we also idealized the backbone dihedral angles of each structure separately using Representation of Protein Entities (RoPE)[45].

Results and Discussion

A framework for examining conformational change

COLAV offers three different structural representations to summarize differences between conformations, each with a distinct emphasis (Table S1 summarizes the functions available in COLAV). Dihedral angles and pairwise distances are internal coordinates, meaning that they are measures calculated from atomic coordinates regardless of the orientation of the protein. Therefore, these calculations can be performed on individual structures and do not require alignment of protein structures. Dihedral angles efficiently summarize local backbone dynamics of individual residues or loops by capturing these motions in only a few features, while pairwise distances better capture global protein dynamics, such as breathing motions[6].

In contrast, strain analysis is a directional measure of the structural deformations accompanying conformational transitions. Using the strain analysis framework of previous studies[31,32], all the structures must be aligned and compared to a designated reference structure. Here, the notion of continuous strain is discretized, instead focusing on individual atoms and their surrounding

atomic neighborhoods—nearby atoms within 8 Å. By comparing the atomic neighborhoods in the working and reference structures, discrete analogs to continuous strain can be estimated, which then describe directional deformations of the desired structure relative to the reference. Notably, strain measurements pick up on regions with relative motion, for example around hinge points, while ignoring rigid-body-like motion, e.g., within subdomains.

COLAV representations distinguish between known PTP-1B conformations

We applied all three methods implemented in COLAV to infer the conformational landscape of protein tyrosine phosphatase 1B (PTP-1B) from crystal structures. PTP-1B is a validated drug target for type II diabetes[46] and breast cancer[46,47], and has been implicated in Alzheimer's disease[48]. Although there has been major pharmacological interest in PTP-1B, no drugs targeting PTP-1B have successfully made it through stage II clinical trials[49]. One major reason is that the PTP-1B active site is highly conserved across the protein tyrosine phosphatase family, making it difficult to design competitive inhibitors without off-target effects *in vivo*[50,51]. The PTP-1B active site is also charged, limiting the effective availability of charged competitive inhibitors that must cross a cell's plasma membrane[51]. For these reasons, there has been widespread interest in allosterically targeting and modulating PTP-1B activity[52]. It is of particular interest, then, to discover surface sites allosterically coupled with the active site[36,53,54].

To do so, we first analyzed a set of 352 crystal structures of PTP-1B obtained from the PDB (165 individual structures and 187 structures from a drug fragment screen performed by Keedy *et al.*[36]). Using principal component analysis (PCA), we found that each structural representation of conformational change implemented in COLAV separated the conformations into the same four

clusters of distinct, known conformations (Fig. 1). These four conformations are described by the conformational states of the WPD and L16 loops (WPD loop/L16 loop): open/open (Fig. 1a top-left), open/closed (Fig. 1b bottom-left), closed/open (Fig. 1c top-right), and closed/closed (Fig. 1d bottom-right). For dihedral angles and strain, the first two PCs clustered these conformations (Fig. 1a, c); for pairwise distances, the first and third PCs clustered these conformations (Fig. 1b; PC2 determines regions with large motions relative to the rest of PTP-1B). We also applied two non-linear dimensionality reduction methods, t-distributed stochastic network embedding (t-SNE) and uniform manifold approximation and projection (UMAP), to the structural representations. These methods similarly clustered PTP-1B structures (Fig. S1), indicating that the PCA clusters were representative of the major groupings in the PTP-1B structures. We next asked whether inconsistent refinement practices for the deposited structures and/or deviations from ideal geometry in individual structures could explain the observed structural heterogeneity. To examine this possibility, we repeated the analysis after applying Representation of Protein Entities (RoPE)[45] to all the PTP-1B structures to idealize and standardize the bond distances and bond angles across the dataset. In RoPE, the backbone dihedral angles of the structures are adjusted to match the original atomic coordinates. PCA identified the same PTP-1B clusters after pre-processing the data (Fig. S2a, b, e), confirming that individual refinement artifacts did not meaningfully affect the results.

The three different structural representations implemented in COLAV can each capture different aspects of conformational change. It is conceivable that local conformational changes take place without much global change and are therefore primarily detectable by monitoring dihedral angles. Another possibility is that global change can be related to only a few dihedral angles,

e.g., in hinge motion, but be detectable elsewhere as changing distances to other parts of the protein. Lastly, it is possible that coupled conformational changes are separated by regions of almost imperceptible change—possibly a common case for proteins[55-57]. To compare the conformational changes revealed by each representation, we calculated residue contributions (RCs) from the coefficients of each of the principal components (PCs), combining per residue the contributions of the sines and cosines of the dihedral angles (for the dihedral angle representation), of distances to all other residues (for the Cα pairwise distance representation), or off-diagonal components of the shear matrix (for the strain representation), respectively, as described in the Methods. By calculating the correlation between these RCs for each pair of representations (Figs. 1d-f, S3), we find that the residue contributions underlying PC1 and PC2 ("RC1" and "RC2") for dihedral angles and for strain are strongly correlated (0.79 comparing RC1s and 0.74 comparing RC2s), respectively. Both RC1 and RC2 of these two representations show a correlation with the residue contributions underlying PC1 and PC3 for pairwise distances (Fig. 1d,f). As expected, however, the residue contributions are not perfectly correlated, indicating differences in the aspects of conformational change captured by each representation.

The PCs distinguish conformational clusters by the states of the WPD loop (Fig. 2a, b; residues 176-188) and L16 loop (Fig. 2a-c; residues 237-243). The active-site WPD loop participates in the PTP-1B catalytic mechanism, while the L16 loop is located ~15 Å away (Fig. 1a). Both loops can take on open and closed states, and all four possible combinations of their states are present in the existing crystal structures. These loops account for most of the conformational heterogeneity present in the PTP-1B dataset (dihedral angles: 36.1% of total variance captured by the first two principal components, Cα pairwise distances: 66.6%, and strain analysis: 67.0%).

In the WPD loop-open state, the loop is positioned such that the active-site pocket is exposed, facilitating substrate access and product release (Fig. 1a-left). In the WPD loop-closed state, the loop binds the substrate and covers the active site pocket, facilitating catalysis[4] (Fig. 1a-right). The L16 loop states differ most saliently by the position of lysine 239 (K239)[36]. In the open state, the sidechain atoms of K239 interact primarily with the solvent (Fig. 1a-top). In the closed state, the sidechain atoms of K239 interact with other residues in the protein (Fig. 1a-bottom). By distinguishing the states of the WPD and L16 loops, PCA captures the major conformational heterogeneity present in crystal structures of PTP-1B.

Could this conformational clustering be caused by crystal packing interactions, rather than the effects of perturbations introduced in individual structures? The most common space group of PTP-1B crystals in our dataset is the $P3_121$ space group, with 293 structures. The space groups of other PTP-1B crystals are $P2_12_12_1$ (29), $P12_11$ (9), C121 (9), $P3_221$ (7), and $P4_12_12$ (2). As we show in Figure S4, the set of structures from crystals in the $P3_121$, $P2_12_12_1$, and $P12_11$ space groups each encompasses all four major conformational clusters. Only the two structures from crystals in the $P4_12_12$ space group take on only a single conformation (closed/open). Since PTP-1B molecules across diverse space groups adopted different conformations, we conclude that crystal packing artifacts cannot account for the conformational clusters highlighted by PCA. Instead, these crystal structures represent semi-random samples from the PTP-1B conformational landscape.

COLAV enables detection of correlated regions in PTP-1B

Although the crystal structures deposited in the PDB for any protein do not, together, constitute a valid thermodynamic ensemble, there is a long history of interpreting frequencies observed in crystal structures in thermodynamic terms[58-61], most recently extending to the interpretation of AlphaFold parameters in energetic terms[62,63]. In this spirit, the statistical correlations observed as principal components can be interpreted as (rough) energetic couplings. Since the conformational landscapes determined by PCA were equivalent for all structural representations, we focus here on the dihedral angle representation (Fig. 1a). We interpreted the first principal component (PC), which accounts for 29.7% of the total variance, to indicate a coupling between the WPD loop and L16 loop (Fig. 2b). Indeed, previous experimental studies using multi-temperature X-ray crystallography[36] and NMR[53,54] have strongly suggested that these two loops are allosterically coupled. We interpreted the second PC, which accounts for 6.3% of the variance, to indicate additional motion of the L16 loop independent of the WPD loop (Fig. 2c). This observation suggests two possibilities. Either the L16 loop undergoes two distinct motions—one coupled to the WPD loop and another decoupled from the WPD loop—or the L16 loop undergoes a single motion that is not always coupled to the WPD loop. To differentiate between these possibilities, we performed a focused PCA on the dihedral angles of the L16 loop (Fig. 2d). We find that the L16 loop has a single dominant motion (Fig. 2d-f) that distinguishes between the open and closed states of the loop; this motion accounts for 63.5% of the variance in this focused PCA. Thus, the L16 loop undergoes a single motion that is not always coupled to the WPD loop.

To examine this coupling more closely, we considered the confounding effect of the C-terminal α7 helix, which has previously been implicated in allosteric coupling within PTP-1B[53] and forms contacts with both loops in their respective closed states. We had initially excluded the α7 helix

from our analysis to avoid missing values, as the α7 helix can transition between an ordered, folded helix state and a disordered state that is not crystallographically observable. However, we noticed that the α7 helix typically takes on the ordered state when at least one of the WPD or L16 loops takes on their respective closed conformations (Table 1). We hypothesized that the exclusion of the α7 helix had led to the observed inconsistencies in the coupling of the two loops. Within the PTP-1B dataset, we find that the presence of an ordered α7 helix greatly increases the probability of finding the closed state of each loop (~40x for the L16 loop and ~50x for the WPD loop). This suggests a cooperative mechanism in which binding of a ligand or inhibitor in the active site can drive concerted loop closure and ordering of the α7 helix.

To formally test for a coupling between the three regions of PTP-1B, we performed a three-way chi-square test of independence (Table 1; treating structures as independent observations), finding strong evidence that these regions are not independent ($p \sim 10^{-158}$). To assess the role of the α7 helix, we next tested how the correlation between the states of the WPD loop and L16 loop depends on the state of this helix (by Fisher's exact test). Given a disordered α7 helix, we find no significant evidence for coupling of the WPD and L16 loops (however, the L16 loop is rarely in the closed state when the α7 helix is disordered, limiting the power of this test). Given an ordered α7 helix, the states of the two loops are strongly coupled to each other ($p = 0.006$; Fisher's exact test). We can, in addition, reject the hypothesis that the state of the α7 helix solely specifies the state of each loop, as the loop states are not conditionally independent given the state of the α7 helix ($p = 0.005$; chi-squared test). Moreover, ligands are not necessary for the protein to visit states with closed WPD and L16 loops and an ordered α7 helix. For instance, apo structures collected at temperatures above 100 K (6B8E, 6B8T, 6B8X) show electron densities

consistent with both states at each of these regions. In addition, several mutations can stabilize apo PTP-1B with the WPD and L16 loops in their closed states and an ordered α7 helix (1PA1, 6OLQ, 6OMY, 6PFW, 7KEN). The two loops are therefore coupled to each other and to the α7 helix, although the exact molecular mechanism remains unclear.

Detailed analysis of the COLAV results further showed active site deformation consistent with oxidation of the active-site catalytic cysteine residue Cys215 (Fig. 3a,b). Oxidation dynamics of this residue play a critical role in its function[64-67] through a self-regulatory mechanism in PTP-1B[65] and (when fully oxidized) degradation (Fig. 3a,b)[68]. The most striking of several oxidized states is a cyclized state in which a sulphenyl-amide bond between the Sγ atom of Cys215 and the backbone nitrogen atom of Ser216 forms a five-membered ring. Structures of oxidized conformations (1OEM and 1OES) show deformations at active site loops, matching RC4 (accounted for 3.5% of total variance) and RC5 (accounted for 2.9% of total variance) of the dihedral angle representation (Fig. 3d,e). Only six PTP-1B structures present in the dataset (~2%) have oxidized cysteine states modeled, and PCA distinguishes these structures from structures in the native, reduced state (Fig. 3c, top-right corner). However, it is possible that low levels of oxidation in PTP-1B crystals are present more widely in the structures[36], impacting the average electron density and, therefore, structure coordinates. Overall, applying PCA to COLAV results successfully identified these rare conformations.

In the analysis of these oxidized structures, we further noticed a strong signal from a region of PTP-1B distant from the active site and distinct from the L16 loop (green shaded box in Figure 3d,e). This spike in signal corresponds to a short loop including residues 59-66. Intriguingly, this

loop is near Ser50 and contains Tyr66, two known phosphorylation sites of PTP-1B[69,70]. Furthermore, a computational analysis of PTP-1B by CryptoSite[71] indicated that this loop is directly adjacent to a cryptic binding site capable of accommodating a small molecule. These observations point to a potential regulatory role of this loop in PTP-1B and perhaps a more direct role in the regulation of oxidized PTP-1B. Speculatively, recent work has shown that the E3-ligase Cullin1 is known to interact with oxidized (sulfonated Cys215) PTP-1B, but the mechanism of this molecular recognition is unclear. The putative coupling suggested by our analysis implies that oxidation of Cys215 triggers concerted motions in this loop, which may allow for recognition and ubiquitination by Cullin1.

Drug fragment screen structures recapitulate the PTP-1B conformational landscape

Could structures from only the PTP-1B crystallographic drug fragment screen[36] suffice to infer the same conformational landscape as the complete PTP-1B dataset or the (non-screen) PTP-1B structures deposited in the PDB ("PDB-only")? To address this question, we again used the dihedral angle representation to map the conformational landscape of PTP-1B based solely on either the fragment screen or the PDB structures (Fig. 4). We first quantified the similarity of the fragment screen-only dataset and the PDB-only dataset using matching and coverage scores[72,73]. The matching score reports on how similar the datasets are by RMSD (root-mean-square deviation) and the score ranges from 0 (each structure has an identical match in the other dataset) to infinity. The coverage score reports on the relative diversity between the datasets and ranges between 0 and 1. Because these scores compare individual structures between datasets, comparing either the fragment screen-only or the PDB-only datasets to the complete dataset would yield perfect scores (matching score of 0 and coverage score of 1) because they contain

the same structures, so we compared the fragment screen-only dataset and PDB-only dataset. We calculated the matching score to be 0.493 Å and the coverage score to be 0.963 with an RMSD similarity cutoff of 1.0 Å, which indicated that the fragment screen-only dataset resembles the PDB-only structures both in terms of containing similar ("matching") structures and in the overall coverage of the conformational landscape.

To determine the relationship between the inferred conformational landscapes more carefully, we compared RCs for PCs from each dataset by calculating correlation coefficients. We found that most key RCs from the complete dataset were also clearly identifiable from the fragment screen-only dataset (Fig. 4a). We found similar results when we compared RCs of the fragment screen-only dataset and the PDB-only dataset (Fig. 4b). This mapping suggests similar structural interpretations for the complete, fragment screen-only, and PDB-only datasets. Indeed, the fifth and seventh fragment screen RCs resemble the first and second RCs of the complete dataset, again indicating a coupling between the WPD loop and L16 loop (Fig. 4c-e), albeit with different proportions of the major states. We note that since refinement of partial-occupancy states, typical for drug fragment screens, tends to be biased towards the unbound state, closed-loop conformations are likely underreported. Effects of catalytic cysteine oxidation were more prominent in the drug fragment screen than in the whole dataset, as observed by Keedy *et al*.[36], such that the second and third fragment screen RCs correlated well with the fourth and fifth RCs of the complete dataset. As discussed above, the fourth and fifth RCs of the complete dataset report on active site deformation due to Cys-215 oxidation (Fig. 4f-h). We note that the first PC of the fragment-only dataset partially reports on a coupling between the L16 loop and the K loop, another active-site loop, that receives little weight in the PDB-only dataset (Fig. S5b). These

comparisons show that the PTP-1B fragment screen conformational landscape matches that of the complete PTP-1B dataset, albeit with a different order of the PCs. This reordering reflects the relative prevalence of the different conformations in the fragment screen dataset.

Continuous motions in the SARS-CoV-2 linker may be coupled to distant surface sites

We next applied the representations implemented in COLAV and PCA to the SARS-CoV-2 main/3CL protease (MPro). MPro is a component of a polyprotein translated from the positive-sense SARS-CoV-2 RNA genome. Through its protease activity, MPro cleaves itself and other functional proteins from this polyprotein, making MPro essential for viral replication[74]. Consequently, MPro is a validated drug target for coronavirus disease caused by SARS-CoV-2 infection (COVID-19). The protein consists of three subdomains: domains I and II form a β-barrel catalytic core, and domain III forms an α-helical bundle unit that facilitates MPro obligate homodimerization (Fig. 5a,b)[75,76]. MPro is the subject of an intense research effort, with several crystallographic drug fragment screens and many other structural studies capturing the homodimer bound to a variety of ligands[37-41]. We analyzed 1,830 structures from these fragment screens and 1,015 other structures deposited in the PDB to determine the MPro conformational landscape by PCA.

In contrast to PTP-1B, the MPro conformational landscape is dominated by a continuous band of structures along PC1 rather than by distinct clusters (Fig. 5c); along PC2, there is a distinct cluster of structures. We cautiously interpreted this to mean that the most common motions in MPro are continuous in the protein: the most flexible regions of the protein do not take on distinct, individual states. However, structures that are related in our conformational landscape, a

reduced-dimensional space, may be more dissimilar in the higher-dimensional space considering all dihedral angles. To test this interpretation, we determined similar groups of MPro structures using the *k*-means algorithm ($k = 8$) for the full high-dimensional dihedral angle representation of each structure, yielding groups that are similar in the high-dimensional space. This proximity is well preserved in the low-dimensional space of the first two principal components (Fig. 5c). As for PTP-1B, PCA determined similar results for the three structural representations according to the *k*-means groups (Fig. S6; coloring of the structures matches between panels; t-SNE and UMAP analysis in Figure S6). From these analyses, we concluded that the dominant concerted motion in MPro is a gradual deformation.

To further investigate the motions of MPro and its correlated regions, we examined the residue contributions, again focusing on the dihedral angle representation. We interpreted the RCs corresponding to PC1 and PC2, respectively accounting for 14.4% and 7.0% of the total variance, as indicative of motion in the linker between domains II and III (Fig. 5d,e). Molecular dynamics simulations and ensemble refinement of MPro structures have shown that this region of the protein is flexible[42,77]. In addition, the motion corresponding to the first PC indicates that this linker is correlated with residues 148-154 and residues 215-227 (Fig. 5d,e). These regions are located approximately 20 Å and 30 Å away from the linker, respectively, in both a single protomer and the homodimer (Fig. 5a,b), indicating an allosteric coupling between these regions. Because the linker abuts the MPro active site, these regions may be suitable targets for drug design.

Next, we asked again whether the drug fragment screen recapitulates the conformational landscape inferred from either the complete MPro dataset or the non-fragment screen ("PDB-only") dataset, as we did for PTP-1B above. We similarly find that the fragment screen-only dataset is nearly as conformationally diverse as the PDB-only dataset, with a coverage score of 0.925 using a RMSD threshold of 1.0 Å; a matching score of 0.466 Å shows that the structures of the fragment screen-only dataset closely match those of the PDB-only dataset. Likewise, we similarly find that the residue contributions to the different PCs have close matches between the fragment screen-only dataset and the whole dataset or the PDB-only dataset (Fig. S7). Therefore, as in PTP-1B, COLAV analysis of the MPro crystallographic drug fragment screen mapped the MPro conformational landscape efficiently and thoroughly.

We have found that conformational landscapes inferred from drug fragment screens alone recapitulate the main features of the conformational landscapes that can be inferred from larger ensembles of structures present in the PDB, often including deliberately designed mutants or targeted ligands. The stronger correspondence found for MPro (Figure 5) than for PTP-1B (compare Figure 4 to Figures 1-3) suggests that the sheer number of fragment-bound structures is an important parameter. To test this idea, we generated random samples from the MPro drug fragment screening dataset without replacement. We then compared the inferred conformational landscapes (based on dihedral angles) to the complete dataset by calculating correlation coefficients between RCs (Fig. S7). Compared to the complete dataset, we found that a reduced dataset of 135 structures was sufficient to broadly capture the top 5 RCs of the complete dataset (Fig. S7e). Most of the top 10 RCs were strongly recapitulated in the reduced datasets of 270 and

458 540 structures (Fig. S7c, d), matching the visual appraisal that the inferred conformational
459 landscape looks like that of the complete dataset.
460
461 *Ordering protein structures by PC score exposes potential transition pathways*
462 The PCA results showed several apparent conformational transitions in both PTP-1B and MPro.
463 To examine these transitions more closely, we used the PC scores to order the structures of either
464 PTP-1B or MPro for both PC1 and PC2 using the dihedral angle representation (Fig. 6). Doing
465 so with PC1 for PTP-1B showed a distinct transition of the WPD loop between the open and
466 closed state (Fig. 6a), while the same for PC2 described the transition of the L16 loop from a
467 closed to open state (Fig. 6b). For MPro, the transitions between most conformations for PC1
468 and PC2 are more subtle (Fig. 6c, d), except for a distinct transition between MPro
469 conformations in the linker along PC2 (Fig. 6d). Ordering structures by PC scores is especially
470 informative when analyzing structures from crystallographic drug fragment screens, as
471 conformations can be paired with the fragment ligands that stabilize them. Those fragment
472 ligands that stabilize particular conformations of the target protein are then readily identifiable as
473 the basis for targeted rational drug design.
474
475 **Conclusions**
476 Crystallographic drug fragment screens provide rich data, not only concerning the binding sites
477 of fragments on drug targets but also on how protein conformations change in response to such
478 binding. In this respect, drug fragment screens approximate an ideal experiment in which the
479 structure of a protein is determined in the presence of each of many random perturbations. We
480 introduced an open-source software package, COLAV, to facilitate inference of empirical

481 conformational landscapes from such drug fragment screening data using three different
482 representations of conformational change. We find that the results are insensitive to the choice of
483 representation and largely robust under the choice of method for dimension reduction, indicating
484 that the discovered conformational clustering is intrinsic to the conformational ensembles
485 studied. Moreover, we found that the conformational landscapes determined this way resemble
486 those inferred from the larger universe of previously determined structures and that the
487 correspondence improves with the number of fragment-bound structures. Altogether, these
488 findings lay the foundation for the systematic use of crystallographic drug fragment screens to
489 map the accessible states of proteins of interest and a roadmap for steering proteins toward
490 desirable conformations. The tools introduced in COLAV are general and may perform equally
491 well for other protein structural ensembles, as the highly constrained nature of protein dynamics
492 will leave its fingerprints on any such dataset.
493

494 **Author Contributions**

495 AAS, MAK, and DRH conceived the approach. AAS developed the code and performed the
496 analysis with feedback from MAK and DRH. All authors contributed to the manuscript.
497

498 **Funding Sources**

499 This work was supported by the Harvard College Research Program (to AAS) and the NIH
500 Director's New Innovator Award (DP2-GM141000 to D.R.H.).
501

502 **Acknowledgement**

We thank Dr. Daniel Keedy and Dr. Helen Ginn, and members of the Hekstra lab for fruitful discussions. We thank Dennis Brookner for assistance in making COLAV available as a package on https://github.com/Hekstra-Lab/colav and PyPI.

Declaration of Interests

The authors declare that no competing interests exist.

Data and Code Availability

All data and code used in this study to generate the figures can be found at https://github.com/Hekstra-Lab/colav. Figures were prepared using PyMOL v2.5.4, available from Schrödinger, LLC.

Figures and Tables

Figure 1 caption on next page

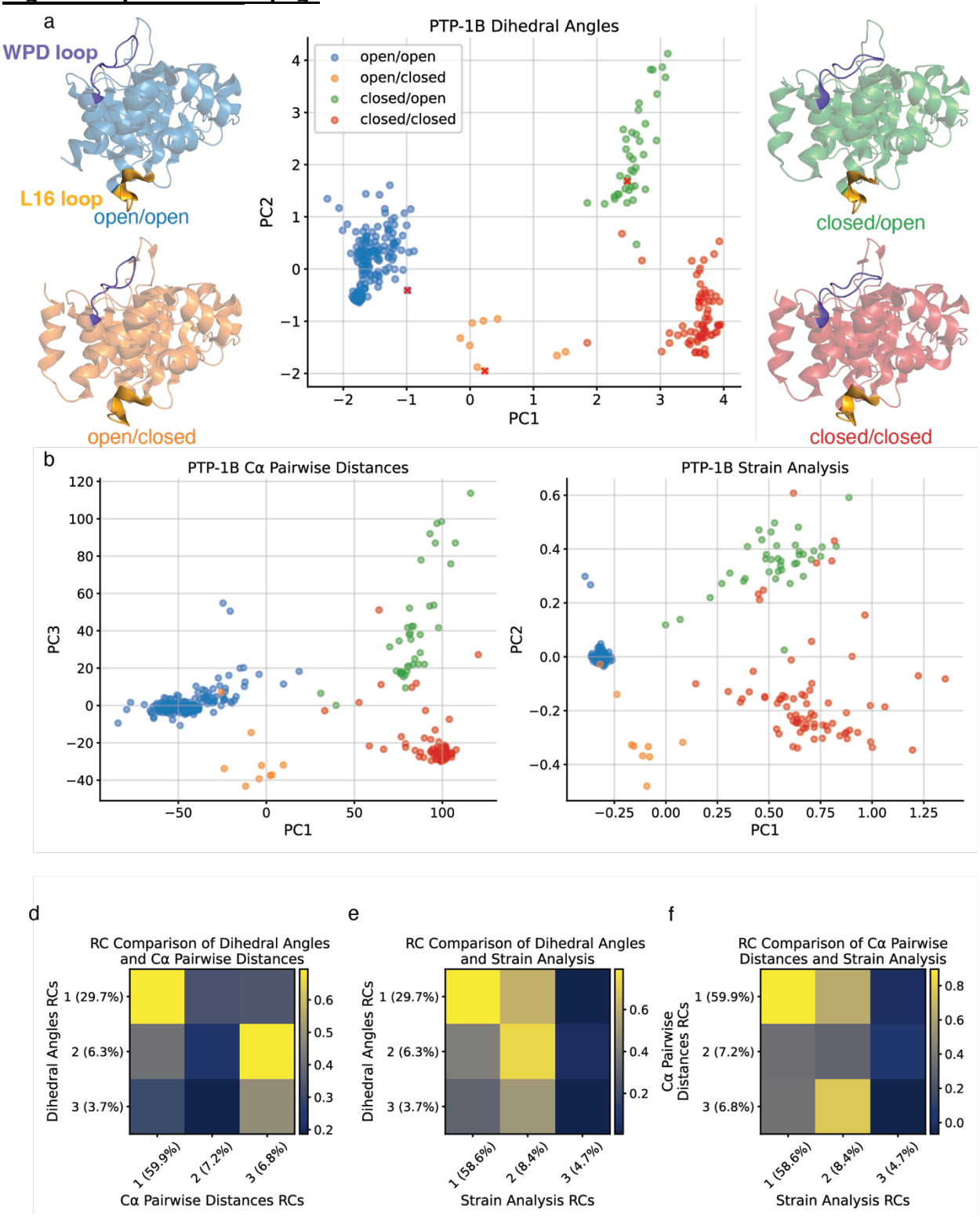


Figure 1: Conformational landscape of PTP-1B inferred using three different structural representations and colored by conformation.

(a) PTP-1B conformational landscape by dihedral angles, flanked by representative PTP-1B structures of the four major conformations labeled by the conformational state of the WPD loop (purple) and L16 loop (yellow): (open/open: 1NWL, open/closed: 4QBW, closed/open: 1PXH, closed/closed: 1SUG). (b) PTP-1B conformational landscape by Cα pairwise distances; note that PC3 is shown on the *y*-axis. (c) PTP-1B conformational landscape by strain analysis. (d-f) Correlation coefficient matrix comparing RCs 1-3 for (d) dihedral angles and Cα pairwise distances; (e) dihedral angles and strain; (f) Cα pairwise distances and strain.

519

520

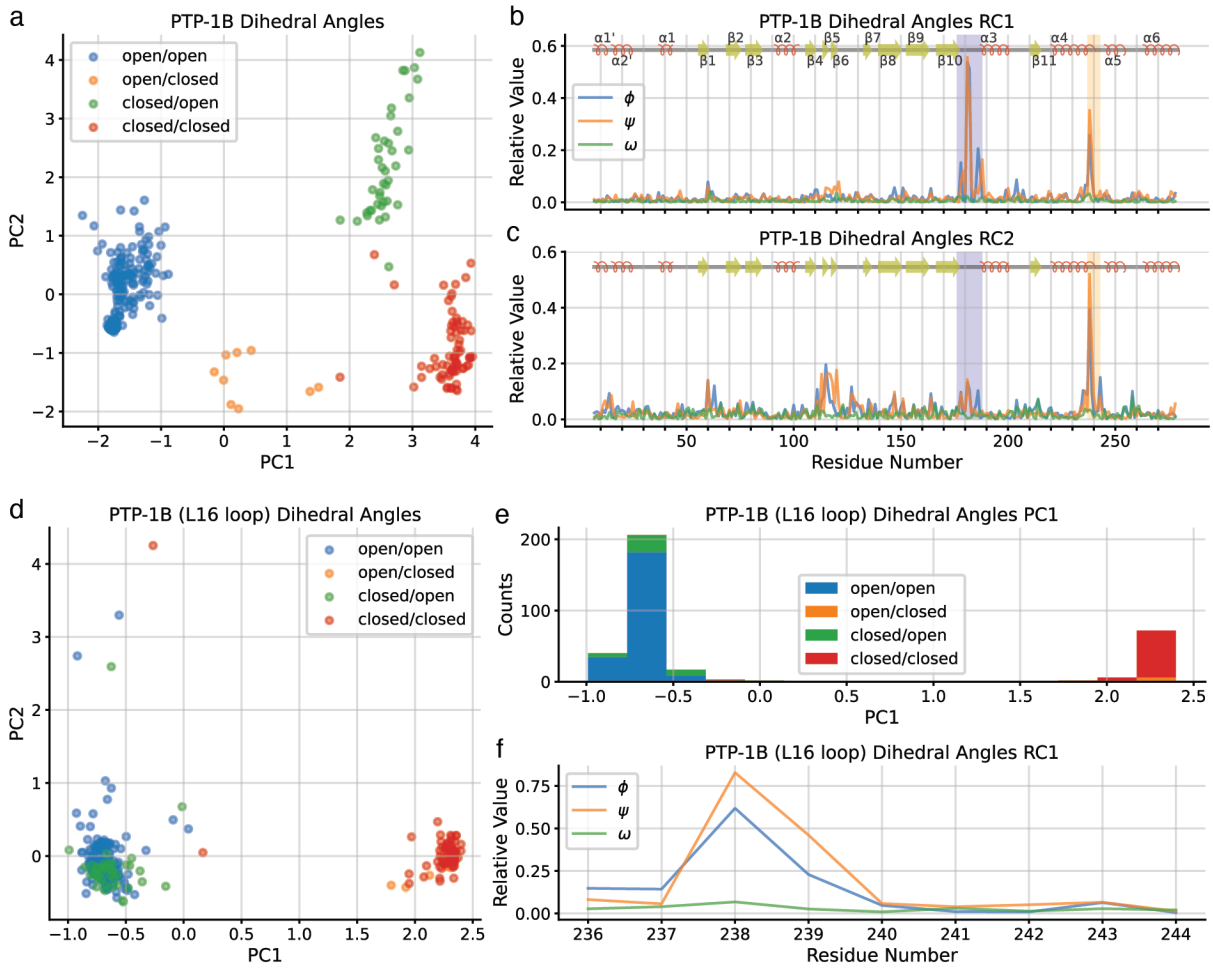


Figure 2: The dihedral angle representation distinguishes between conformations of PTP-1B based on the conformations of the WPD loop and L16 loop.

(a) PTP-1B conformational landscape by dihedral angles by PC1 and PC2. (b) Residue contributions to principal component 1 (PC1), with WPD loop (residues 176-188) indicated by a purple box and L16 loop (residues 237-243) in a yellow box. (c) Residue contributions to PC2, with WPD loop in purple box and L16 loop in yellow box. (d) PTP-1B L16 loop conformational landscape by dihedral angles colored by conformation. (e) Histogram of PTP-1B structures according to PC1 of the focused PCA. (f) Residue contributions to PC1 of the focused PCA.

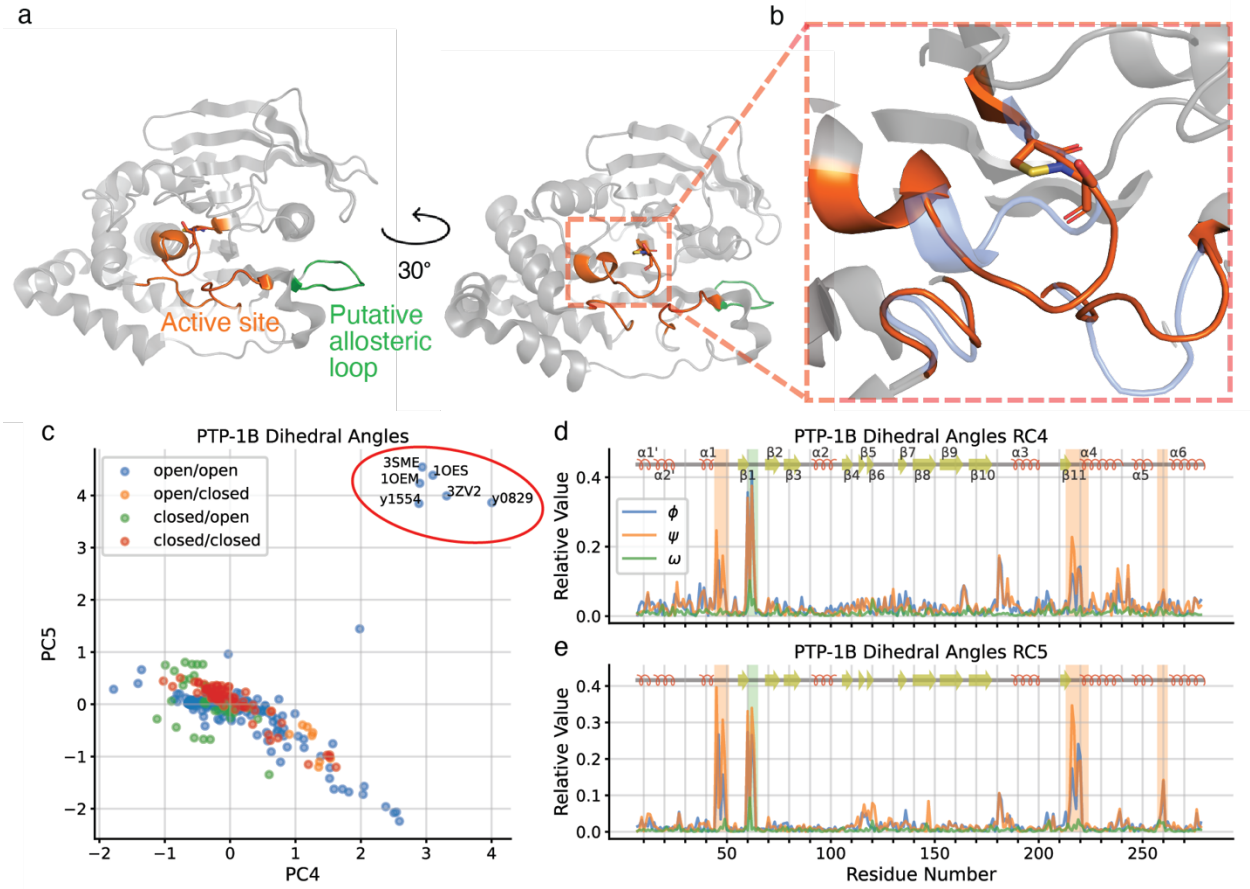


Figure 3: PTP-1B conformational change due to oxidation states of Cys215.

(a) Cartoon representations of oxidized PTP-1B conformation (1OES), highlighting active site loops (orange) and putative allosteric loop (green). (b) Cartoon representation of the oxidized PTP-1B active site conformation (1OES; orange), with sulphenyl-amide ring shown in sticks, and the reduced PTP-1B active site conformation (1SUG; blue) for comparison. (c) PTP-1B conformational landscape by dihedral angles by PC4 and PC5; structures showing oxidized PTP-1B conformation as in panels (a) and (b) are circled in red. (d) Residue contributions to PC4, with active site loops in orange box and putative allosteric loop (residues 59-66) in green box (coloring matches panel (a)). (e) Residue contributions to PC5, with loop coloring as in panel (d).

523

524

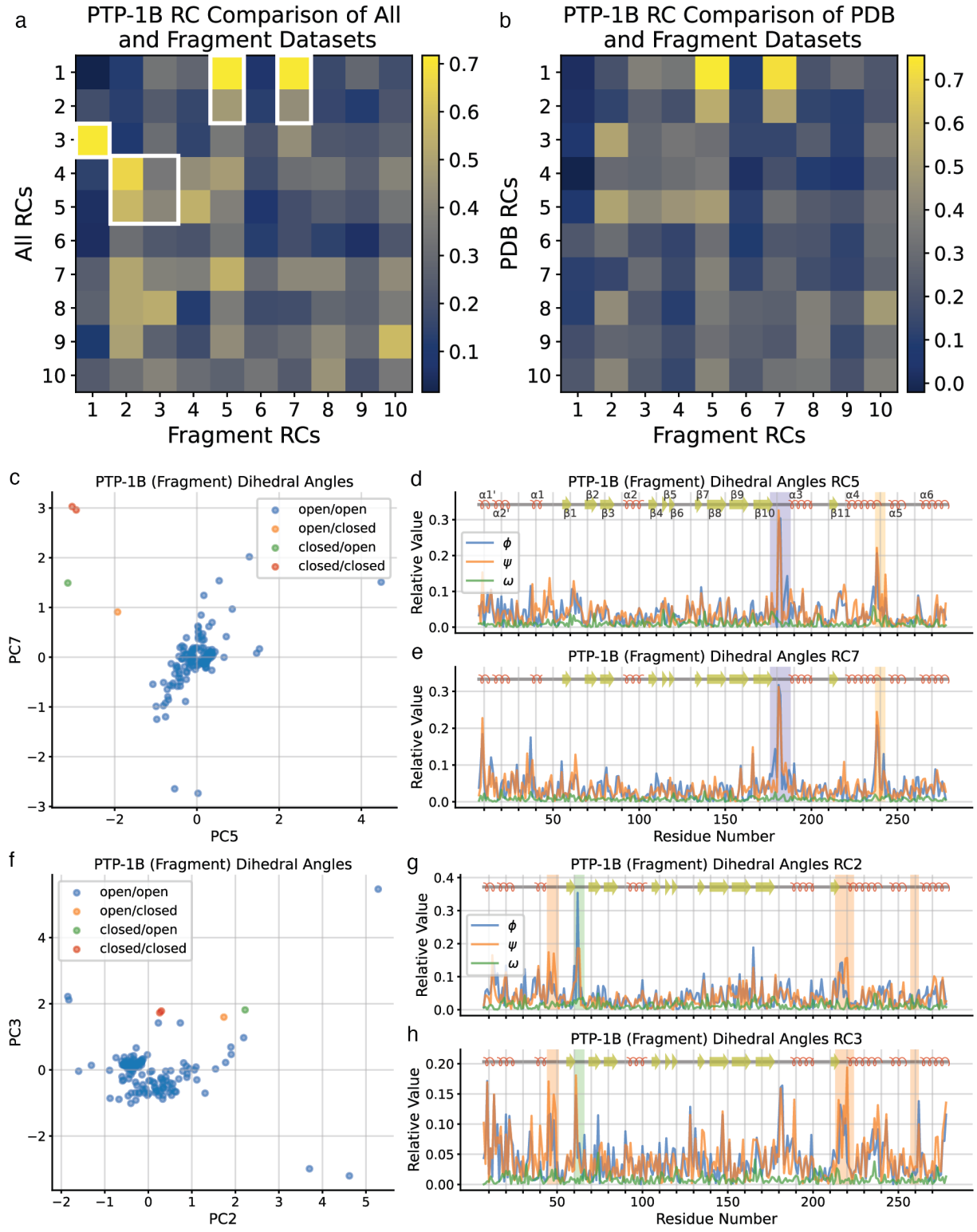


Figure 4 caption on next page

Figure 4: COLAV analysis of the PTP-1B crystallographic drug fragment screen recapitulates key aspects of the conformational landscape.

(a, b) Correlation coefficient matrix comparing residue contributions (RCs) of (a) the complete PTP-1B dataset to those of the fragment screen-only PTP-1B dataset, and (b) the PDB-only PTP-1B dataset to those of the fragment screen-only PTP-1B dataset. Correlations discussed in the text are highlighted using white edges. (c) Fragment screen PTP-1B conformational landscape by dihedral angles, emphasizing similarities of PC5 and PC7 with PC1 and PC2 of the complete PTP-1B conformational landscape. (d) Residue contributions to PC5, with WPD loop in purple box and L16 loop in yellow box. (e) Residue contributions to PC7, with coloring as in panel (d). (f) Fragment screen PTP-1B conformational landscape by dihedral angles, emphasizing similarities of PC2 and PC4 with PC4 and PC5 of the complete PTP-1B conformational landscape. (g, h) Residue contributions to (g) PC2 and (h) PC4, with active site loops in orange box and putative allosteric loop in dark blue box.

531

Figure 5 caption on next page

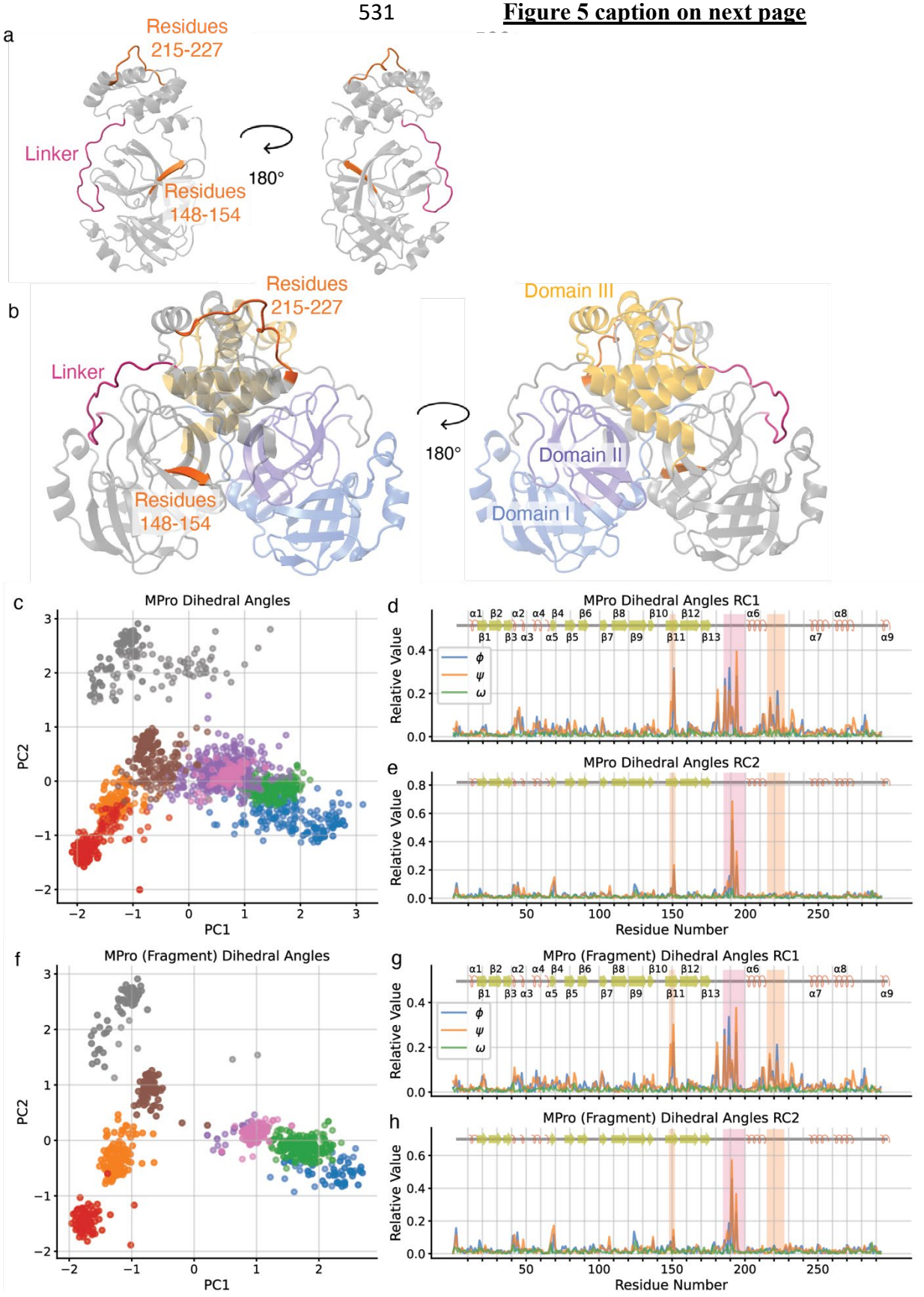


Figure 5: The MPro complete dataset and fragment-screen-only dataset generate similar conformational landscapes

(a) Cartoon representations of single MPro protomer (7AR5), highlighting linker (residues 185-200) in magenta and putative allosteric regions (residues 148-152 and 215-227) in orange. (b) Cartoon representations of MPro homodimer (7AR5), highlighting subdomain I in blue, subdomain II in purple, and subdomain III in yellow on protomer 1 and linker and putative allosteric regions colored as in (a). (c) MPro conformational landscape by dihedral angles. (d) Residue contributions to PC1, with linker in magenta box and putative allosteric loops in orange box (coloring matches panels (a) and (b)). (e) Residue contributions to PC2, with loop coloring as in panel (d). (f) Fragment screen MPro conformational landscape by dihedral angles. (g, h) Residue contributions to fragment screen (g) PC1 and (h) PC2, with loop coloring as in panel (d).

537
538

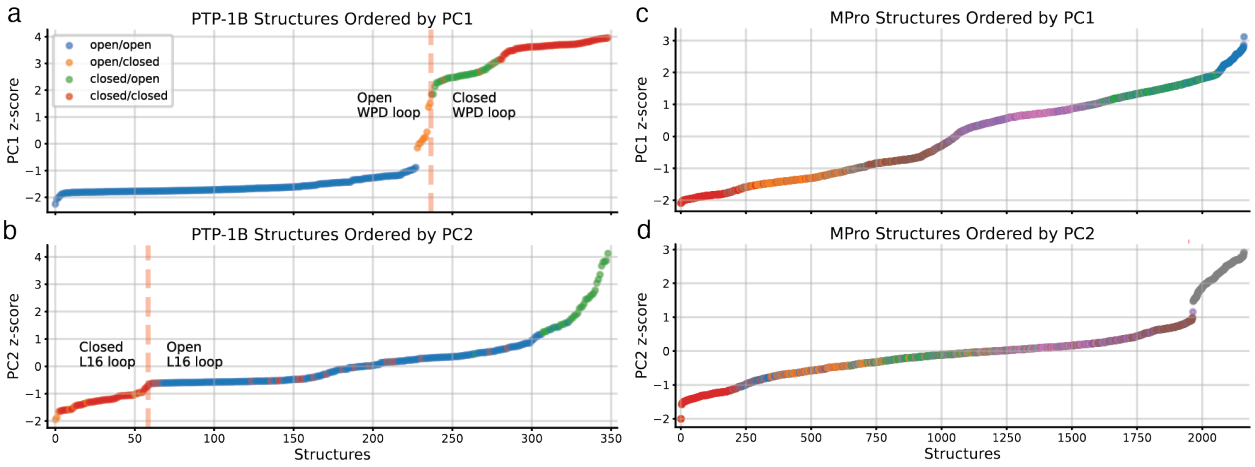


Figure 6: Ordering structures of PTP-1B and MPro by PC scores marks distinct conformational transitions.

(a) PTP-1B structures ordered by dihedral angle score along PC1, with the transition from open WPD loop to closed WPD loop highlighted. (b) PTP-1B structures ordered by dihedral angle score along PC2, with the transition from closed L16 loop to open L16 loop highlighted. (c) MPro structures ordered by dihedral angle score along PC1. (d) MPro structures ordered by dihedral angle score along PC2. Coloring of datasets for both proteins matches preceding figures.

539
540

Disordered α7 Helix

	Open L16	Closed L16	**total**
Open WPD	**221**	4	225
Closed WPD	23	1	24
total	244	6	249

P-value: 0.40

Ordered α7 Helix

	Open L16	Closed L16	**total**
Open WPD	7	5	12
Closed WPD	16	**72**	88
total	23	77	100

P-value: 0.006

Table 1: Assessing the correlations of the WPD loop, L16 loop, and α7 helix through χ^2 test of independence. Contingency table comparing PTP-1B conformations of the WPD loop, the L16 loop, and the α7 helix. Calculated p-values are based on a Fisher exact test.

Protein	Matching Score (Å)	Coverage Score
PTP-1B	0.493	0.963
MPro	0.466	0.925

Table 2: Matching and coverage scores comparing PDB-only and fragment screen-only structures for PTP-1B and MPro. The matching score reports on the similarity of the datasets by RMSD, and a smaller score implies that the datasets are more similar. The coverage score reports on diversity of structures between datasets, and the highest score of 1 implies that the datasets are similarly diverse.

556 **Supplementary Figures and Table**

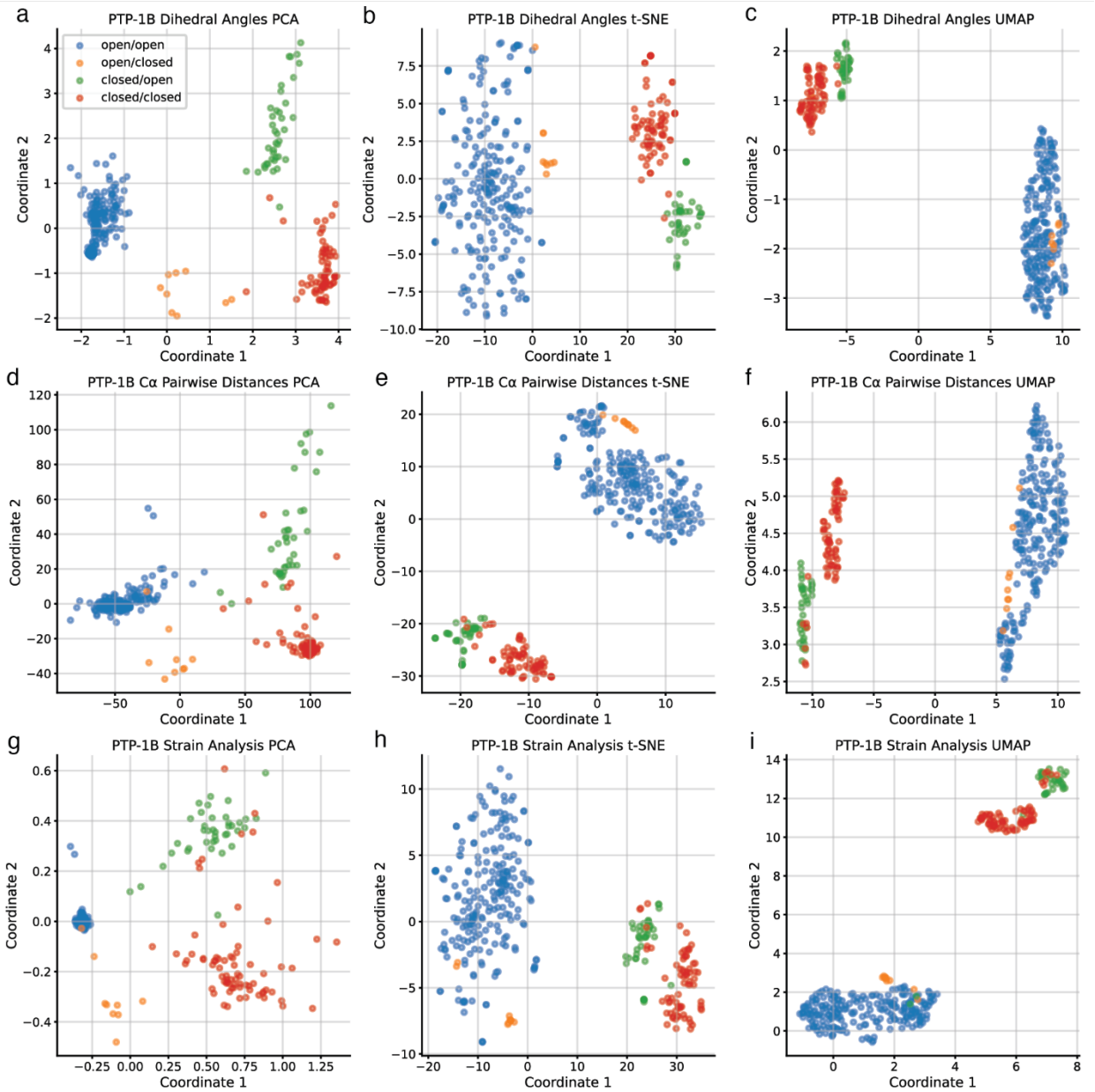
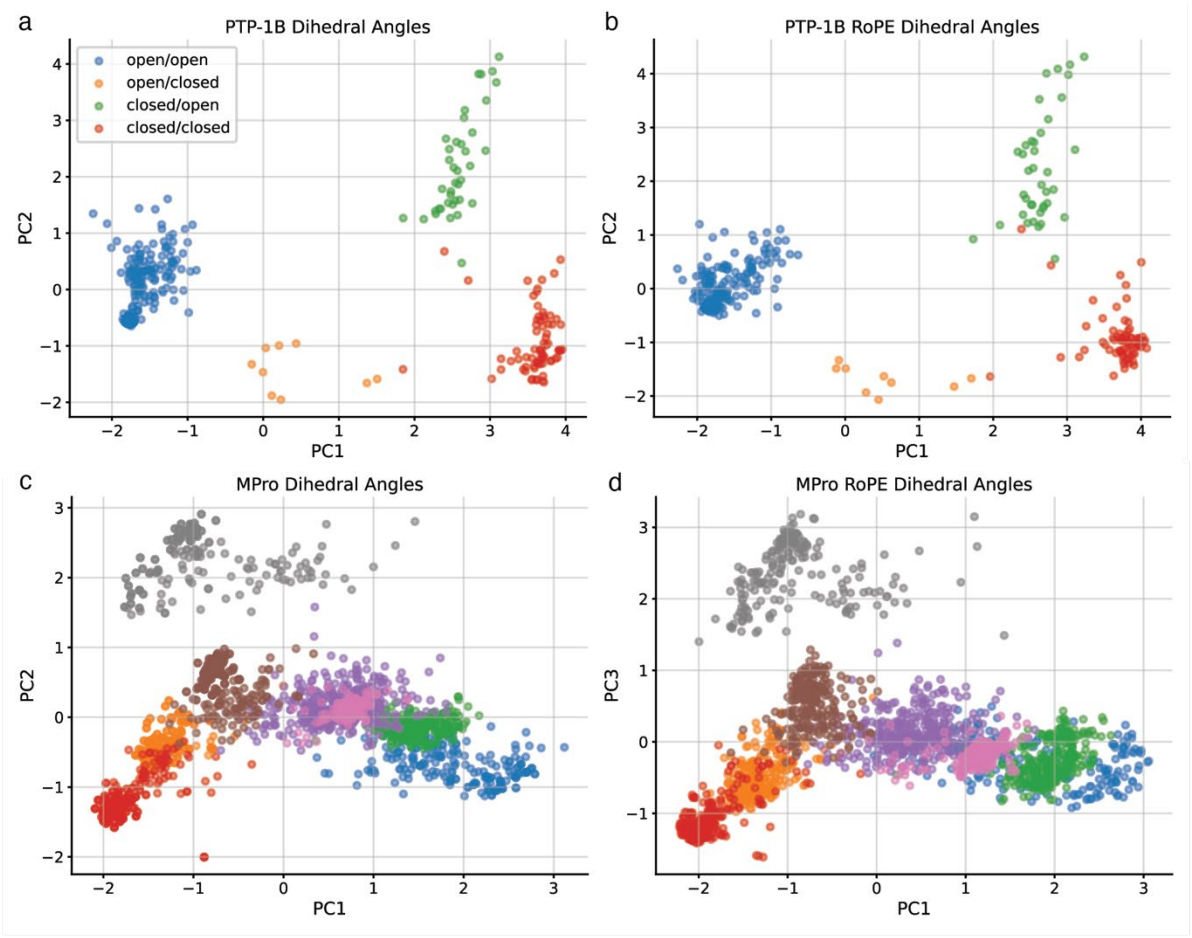


Figure S1: Comparison of PCA, t-SNE, and UMAP applied to all three structural representations of PTP-1B colored by conformation.

(a-c) PTP-1B conformational landscape based on dihedral angles with dimensionality reduction by (a) PCA, (b) t-SNE, and (c) UMAP. (d-f) PTP-1B conformational landscape based on Cα pairwise distances, analyzed using (d) PCA, (e) t-SNE, and (f) UMAP. (g-j) PTP-1B conformational landscape based on strain analysis and (g) PCA, (h) t-SNE, and (i) UMAP. Coloring of structures is consistent among all panels.



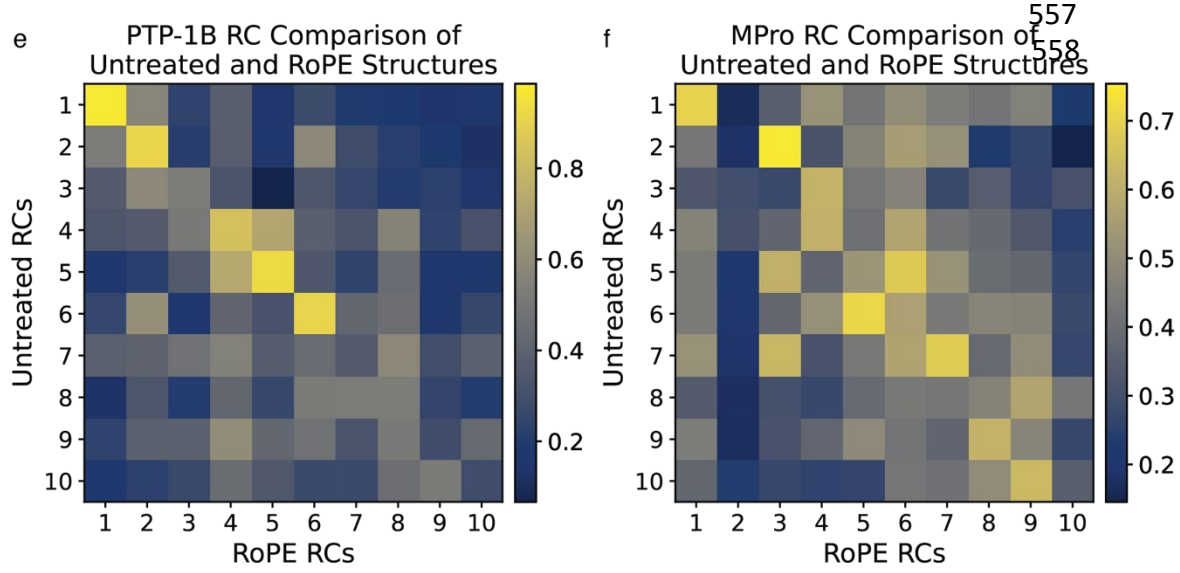


Figure S2 caption on next page

Figure S2: Dihedral angles with and without idealization by RoPE reveal similar conformational landscapes.

(a) PTP-1B conformational landscape by dihedral angles calculated by COLAV. (b) PTP-1B conformational landscape by dihedral angles idealized by RoPE. (c) MPro conformational landscape by dihedral angles calculated by COLAV. (d) MPro conformational landscape by dihedral angles calculated by RoPE. (e) Correlation coefficient matrix comparing PTP-1B RCs of the untreated (COLAV) dihedral angles and RoPE dihedral angles. (f) Correlation coefficient matrix comparing MPro RCs of the untreated (COLAV) dihedral angles and RoPE dihedral angles.

565

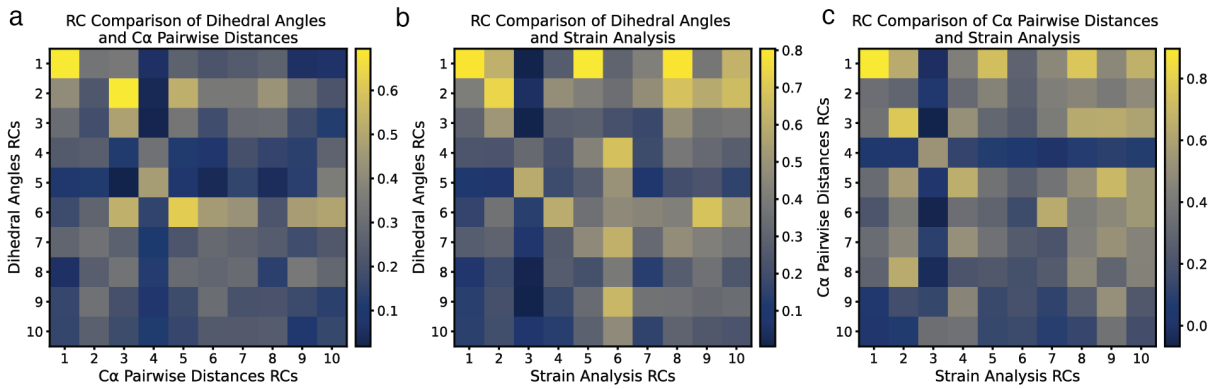


Figure S3: Comparison of residue contributions for structural representations of PTP-1B.

Correlation coefficients comparing PTP-1B residue contributions (RCs) for (a) dihedral angles and Cα pairwise distances, (b) dihedral angles and strain, and (c) Cα pairwise distances and strain.

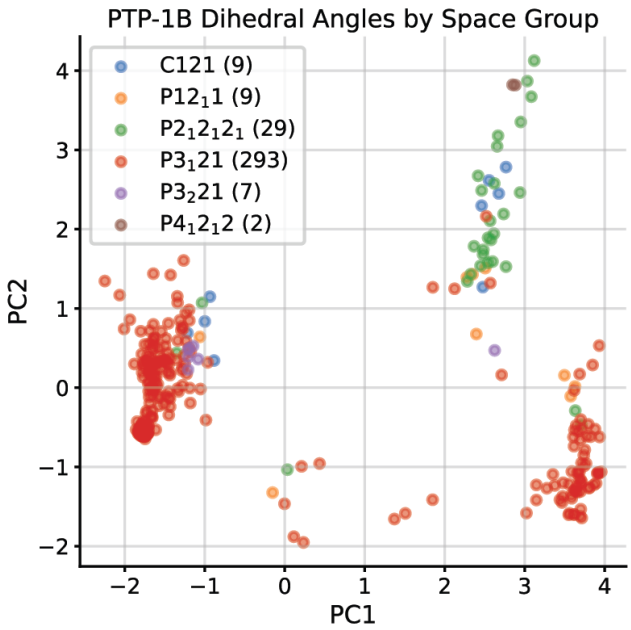


Figure S4: PTP-1B conformations are found across space groups.

Distribution of structures of PTP-1B after PCA of their dihedral angles. Structures are colored by the space group of their crystal forms.

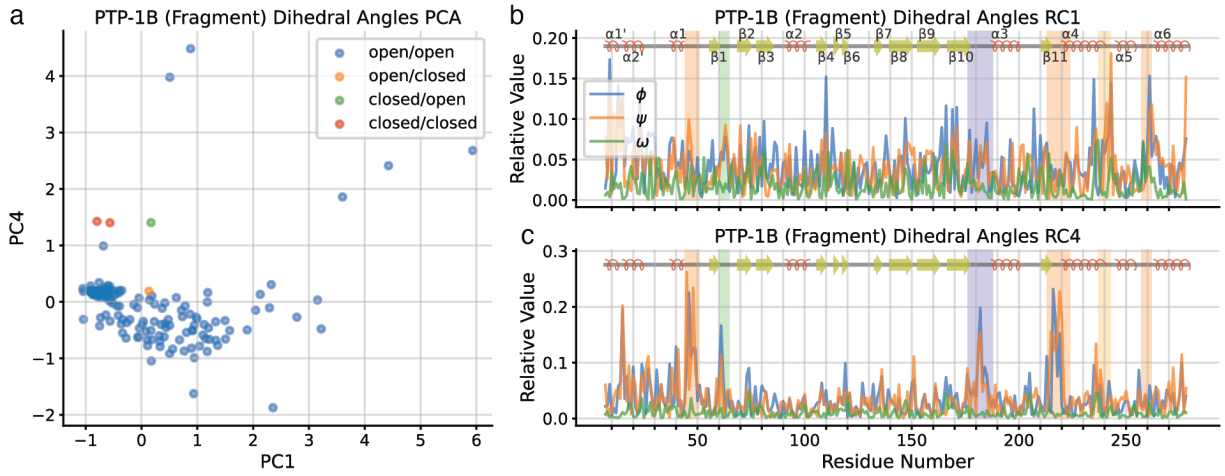


Figure S5: Additional dimensions of the PTP-1B conformational landscape inferred from crystallographic drug fragment screen structures.

(a) Fragment screen PTP-1B conformational landscape by dihedral angles, using PC1 and PC4. (b) Residue contributions to PC1, with active site loops in orange box, putative allosteric loop in dark blue box, WPD loop in purple box, and L16 loop in yellow box. (c) Residue contributions to PC4, with coloring as in panel (b).

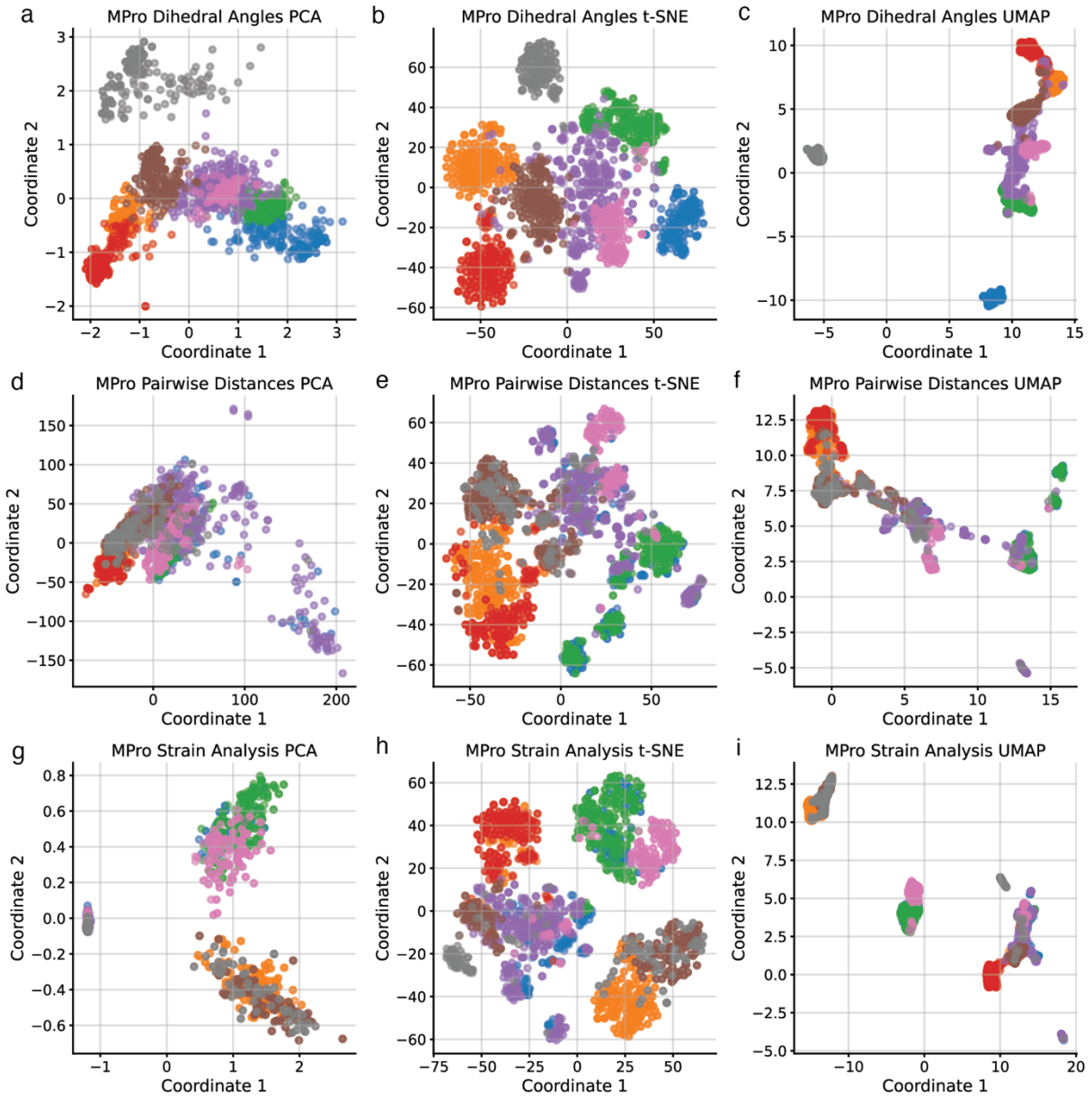


Figure S6: Comparison of PCA, t-SNE, and UMAP applied to all three structural representations of MPro.

(a-c) MPro conformational landscape based on dihedral angles with dimensionality reduction by (a) PCA, (b) t-SNE, and (c) UMAP. (d-f) MPro conformational landscape based on Cα pairwise distances, analyzed using (d) PCA, (e) t-SNE, and (f) UMAP. (g-j) MPro conformational landscape based on strain analysis and (g) PCA, (h) t-SNE, and (i) UMAP. Coloring of structures is consistent among all panels.

576

577

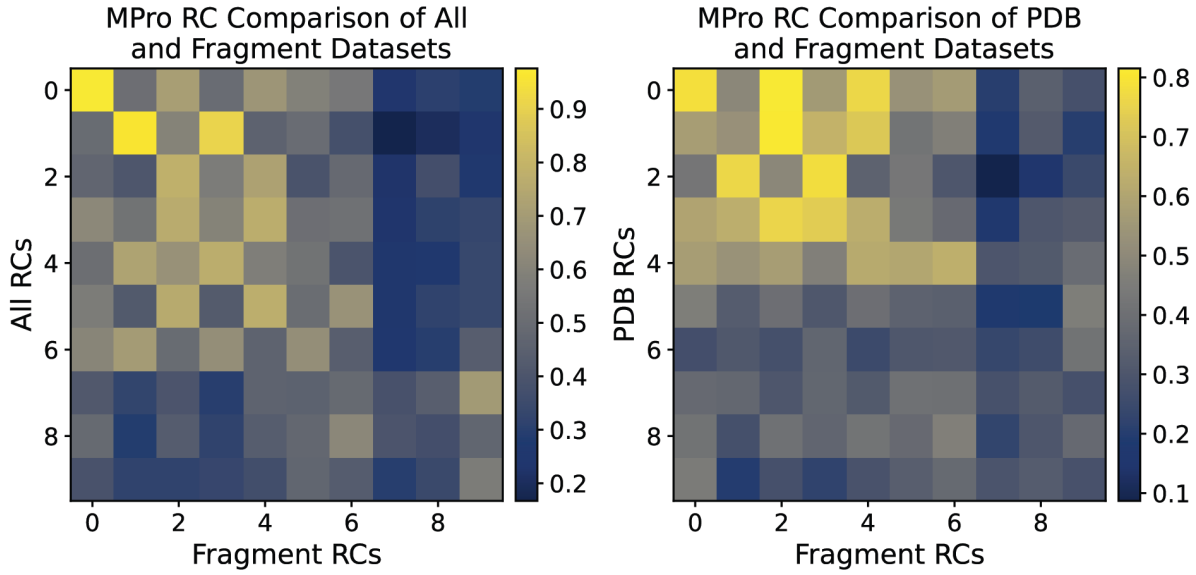


Figure S7: Comparison of dihedral angles residue contributions for MPro datasets.

(a) Correlation coefficient matrix comparing RCs of the complete MPro dataset to those of the fragment screen-only MPro dataset. (b) Correlation coefficient matrix comparing RCs of the PDB-only MPro dataset to those of the fragment screen-only MPro dataset.

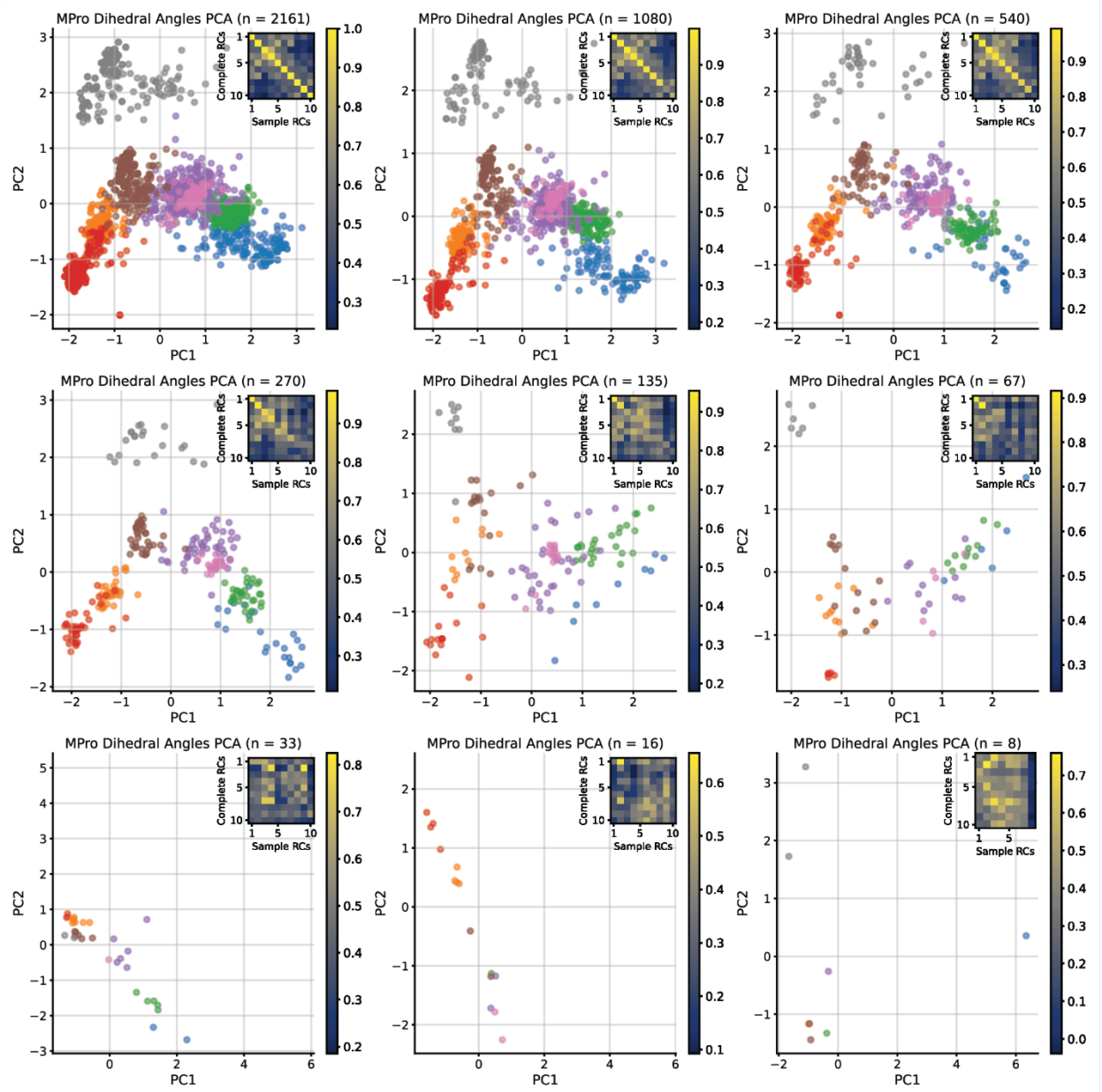


Figure S8: Effect of dataset size on the quality of inferred MPro conformational landscapes.

MPro conformational landscape by dihedral angles were determined for the complete dataset and subsampled datasets of N = 2161 (a), 1080 (b), 540 (c), 270 (d), 135 (e), 67 (f), 33 (g), 16 (h) or 8 (i) structures. In each panel we include an inset of the correlation of the residue contributions inferred for the complete dataset and the sampled dataset.

585
586

COLAV `extract_data` module			
Function	Parameters	Returns	Description
`calculate_coverage_matching_scores`	`reference_strucs, sample_strucs, resnum_bounds, rmsd_threshold, verbose`	`coverage, matching`	Calculates the coverage and matching metrics for a reference set of structures/conformational ensemble compared to a sample set of structures/conformational ensemble.
`calculate_dh_tl`	`raw_dh_loading`	`tranformed_dh_loading`	Adjusts raw dihedral loading for interpretability.
`calculate_pw_tl`	`raw_pw_loading, resnum_bounds`	`transformed_pw_loading`	Adjusts raw pairwise distance loading for interpretability.
`calculate_sa_tl`	`raw_sa_loading, shared_atom_list`	`tranformed_sa_loading`	Adjusts raw strain or shear loading for interpretability.
`generate_dihedral_matrix`	`structure_list, resnum_bounds, no_psi, no_omega, no_phi, save, save_prefix, verbose`	`dh_data_matrix, dh_strucs`	Extracts dihedrals angles from given structures.
`generate_pw_matrix`	`structure_list, resnum_bounds, save, save_prefix, verbose`	`pw_data_matrix, pw_strucs`	Extracts pairwise distances from given structures.
`generate_strain_matrix`	`structure_list, reference_pdb, data_type, resnum_bounds, atoms, alt_locs, save, save_prefix, save_additional, verbose`	`sa_data_matrix, sa_strucs`	Extracts strain tensors, shear tensors, or shear energies from given structures.
`load_dihedral_matrix`	`dh_pkl`	`dh_data_matrix, dh_strucs`	Loads the dihedral data matrix and corresponding structures.
`load_pw_matrix`	`pw_pkl`	`pw_data_matrix, pw_strucs`	Loads the pairwise distance data matrix and corresponding structures.
`load_strain_matrix`	`strain_pkl`	`sa_data_matrix, sa_strucs`	Loads the strain data matrix and corresponding structures.

Table S1: User-accessible COLAV functions for analyzing structural data.
For a more complete description of the COLAV software package and its functionality, visit https://github.com/Hekstra-Lab/colav. Note that "transformed loadings" are referred to in the text as "residue contributions".

References

1. Gao, S., and Klinman, J.P. (2022). Functional roles of enzyme dynamics in accelerating active site chemistry: Emerging techniques and changing concepts. Current Opinion in Structural Biology *75*, 102434. https://doi.org/10.1016/j.sbi.2022.102434.

2. Henzler-Wildman, K., and Kern, D. (2007). Dynamic personalities of proteins. Nature *450*, 964-972. 10.1038/nature06522.

3. Stachowski, T.R., and Fischer, M. (2022). Large-Scale Ligand Perturbations of the Protein Conformational Landscape Reveal State-Specific Interaction Hotspots. Journal of Medicinal Chemistry *65*, 13692-13704. 10.1021/acs.jmedchem.2c00708.

4. Whittier, S.K., Hengge, A.C., and Loria, J.P. (2013). Conformational motions regulate phosphoryl transfer in related protein tyrosine phosphatases. Science *341*, 899-903. 10.1126/science.1241735.

5. Zuccotto, F., Ardini, E., Casale, E., and Angiolini, M. (2010). Through the "Gatekeeper Door": Exploiting the Active Kinase Conformation. Journal of Medicinal Chemistry *53*, 2681-2694. 10.1021/jm901443h.

6. Greisman, J.B., Dalton, K.M., Brookner, D.B., Klureza, M.A., Sheehan, C.J., Kim, I.-S., Henning, R.W., Russi, S., and Hekstra, D.R. (2023). Resolving conformational changes that mediate a two-step catalytic mechanism in a model enzyme. bioRxiv, 2023.2006.2002.543507. 10.1101/2023.06.02.543507.

7. Lewandowski, J.R., Halse, M.E., Blackledge, M., and Emsley, L. (2015). Direct observation of hierarchical protein dynamics. Science *348*, 578-581. doi:10.1126/science.aaa6111.

8. Ramanathan, A., Savol, A., Burger, V., Chennubhotla, C.S., and Agarwal, P.K. (2014). Protein Conformational Populations and Functionally Relevant Substates. Accounts of Chemical Research *47*, 149-156. 10.1021/ar400084s.

9. Noé, F., and Fischer, S. (2008). Transition networks for modeling the kinetics of conformational change in macromolecules. Current Opinion in Structural Biology *18*, 154-162. https://doi.org/10.1016/j.sbi.2008.01.008.

10. Juraszek, J., Vreede, J., and Bolhuis, P.G. (2012). Transition path sampling of protein conformational changes. Chemical Physics *396*, 30-44. https://doi.org/10.1016/j.chemphys.2011.04.032.

11. Hekstra, D.R. (2023). Emerging Time-Resolved X-Ray Diffraction Approaches for Protein Dynamics. Annual Review of Biophysics *52*, 255-274. 10.1146/annurev-biophys-111622-091155.

12. Alderson, T.R., and Kay, L.E. (2021). NMR spectroscopy captures the essential role of dynamics in regulating biomolecular function. Cell *184*, 577-595. 10.1016/j.cell.2020.12.034.

13. Mazal, H., and Haran, G. (2019). Single-molecule FRET methods to study the dynamics of proteins at work. Current Opinion in Biomedical Engineering *12*, 8-17. https://doi.org/10.1016/j.cobme.2019.08.007.

14. McHaourab, H.S., Steed, P.R., and Kazmier, K. (2011). Toward the fourth dimension of membrane protein structure: insight into dynamics from spin-labeling EPR spectroscopy. Structure *19*, 1549-1561. 10.1016/j.str.2011.10.009.

15. Fraser, J.S., van den Bedem, H., Samelson, A.J., Lang, P.T., Holton, J.M., Echols, N., and Alber, T. (2011). Accessing protein conformational ensembles using room-temperature X-

ray crystallography. Proceedings of the National Academy of Sciences *108*, 16247-16252. doi:10.1073/pnas.1111325108.

16. Elmlund, D., Le, S.N., and Elmlund, H. (2017). High-resolution cryo-EM: the nuts and bolts. Current Opinion in Structural Biology *46*, 1-6. https://doi.org/10.1016/j.sbi.2017.03.003.

17. Zhong, E.D., Bepler, T., Berger, B., and Davis, J.H. (2021). CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. Nature Methods *18*, 176-185. 10.1038/s41592-020-01049-4.

18. Punjani, A., and Fleet, D.J. (2023). 3DFlex: determining structure and motion of flexible proteins from cryo-EM. Nature Methods *20*, 860-870. 10.1038/s41592-023-01853-8.

19. Luo, Y., Pfuetzner, R.A., Mosimann, S., Paetzel, M., Frey, E.A., Cherney, M., Kim, B., Little, J.W., and Strynadka, N.C.J. (2001). Crystal Structure of LexA: A Conformational Switch for Regulation of Self-Cleavage. Cell *106*, 585-594. 10.1016/S0092-8674(01)00479-2.

20. Joerger, A.C., Allen, M.D., and Fersht, A.R. (2004). Crystal structure of a superstable mutant of human p53 core domain. Insights into the mechanism of rescuing oncogenic mutations. J Biol Chem *279*, 1291-1296. 10.1074/jbc.M309732200.

21. Wittinghofer, A., and Pal, E.F. (1991). The structure of Ras protein: a model for a universal molecular switch. Trends in Biochemical Sciences *16*, 382-387. 10.1016/0968-0004(91)90156-P.

22. Kondrashov, D.A., Zhang, W., Aranda IV, R., Stec, B., and Phillips Jr., G.N. (2008). Sampling of the native conformational ensemble of myoglobin via structures in different

crystalline environments. Proteins: Structure, Function, and Bioinformatics *70*, 353-362. https://doi.org/10.1002/prot.21499.

23. Buergi, H.B., and Dunitz, J.D. (1983). From crystal statics to chemical dynamics. Accounts of Chemical Research *16*, 153-161. 10.1021/ar00089a002.

24. Douangamath, A., Powell, A., Fearon, D., Collins, P.M., Talon, R., Krojer, T., Skyner, R., Brandao-Neto, J., Dunnett, L., Dias, A., et al. (2021). Achieving Efficient Fragment Screening at XChem Facility at Diamond Light Source. JoVE, e62414. doi:10.3791/62414.

25. Pearce, N.M., Krojer, T., Bradley, A.R., Collins, P., Nowak, R.P., Talon, R., Marsden, B.D., Kelm, S., Shi, J., Deane, C.M., and von Delft, F. (2017). A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. Nature Communications *8*, 15123. 10.1038/ncomms15123.

26. Ginn, H. (2020). Pre-clustering data sets using cluster4x improves the signal-to-noise ratio of high-throughput crystallography drug-screening analysis. Acta Crystallographica Section D *76*, 1134-1144. doi:10.1107/S2059798320012619.

27. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Research *28*, 235-242. 10.1093/nar/28.1.235.

28. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357-362. 10.1038/s41586-020-2649-2.

29. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental

algorithms for scientific computing in Python. Nature Methods *17*, 261-272. 10.1038/s41592-019-0686-2.

30. Raschka, S. (2017). BioPandas: Working with molecular structures in pandas DataFrames. Journal of Open Source Software *2*, 279. 10.21105/joss.00279.

31. Gullett, P.M., Horstemeyer, M.F., Baskes, M.I., and Fang, H. (2007). A deformation gradient tensor and strain tensors for atomistic simulations. Modelling and Simulation in Materials Science and Engineering *16*, 015001. 10.1088/0965-0393/16/1/015001.

32. Mitchell, M.R., Tlusty, T., and Leibler, S. (2016). Strain analysis of protein structures and low dimensionality of mechanical allosteric couplings. Proc Natl Acad Sci U S A *113*, E5847-E5855. 10.1073/pnas.1609462113.

33. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825-2830.

34. Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research *9*.

35. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. 10.48550/ARXIV.1802.03426.

36. Keedy, D.A., Hill, Z.B., Biel, J.T., Kang, E., Rettenmaier, T.J., Brandao-Neto, J., Pearce, N.M., von Delft, F., Wells, J.A., and Fraser, J.S. (2018). An expanded allosteric network in PTP1B by multitemperature crystallography, fragment screening, and covalent tethering. Elife *7*. 10.7554/eLife.36307.

37. Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C.D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., et al. (2020). Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. Nature Communications *11*, 5047. 10.1038/s41467-020-18709-w.

38. Zhang, C.-H., Stone, E.A., Deshmukh, M., Ippolito, J.A., Ghahremanpour, M.M., Tirado-Rives, J., Spasov, K.A., Zhang, S., Takeo, Y., Kudalkar, S.N., et al. (2021). 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 *7*, 467-475. 10.1021/acscentsci.1c00039.

39. Qiao, J., Li, Y.-S., Zeng, R., Liu, F.-L., Luo, R.-H., Huang, C., Wang, Y.-F., Zhang, J., Quan, B., Shen, C., et al. (2021). SARS-CoV-2 Mpro inhibitors with antiviral activity in a transgenic mouse model. Science *371*, 1374-1378. 10.1126/science.abf1611.

40. Noske, G.D., Nakamura, A.M., Gawriljuk, V.O., Fernandes, R.S., Lima, G.M.A., Rosa, H.V.D., Pereira, H.D., Zeri, A.C.M., Nascimento, A.F.Z., Freire, M.C.L.C., et al. (2021). A Crystallographic Snapshot of SARS-CoV-2 Main Protease Maturation Process. Journal of Molecular Biology *433*, 167118. https://doi.org/10.1016/j.jmb.2021.167118.

41. Günther, S., Reinke, P.Y.A., Fernández-García, Y., Lieske, J., Lane, T.J., Ginn, H.M., Koua, F.H.M., Ehrt, C., Ewert, W., Oberthuer, D., et al. (2021). X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. Science *372*, 642-646. 10.1126/science.abf7945.

42. Ebrahim, A., Riley, B.T., Kumaran, D., Andi, B., Fuchs, M.R., McSweeney, S., and Keedy, D.A. (2022). The temperature-dependent conformational ensemble of SARS-CoV-2 main protease (Mpro). IUCrJ *9*, 682-694. doi:10.1107/S2052252522007497.

43. Wojdyr, M. (2022). GEMMI: A library for structural biology. Journal of Open Source Software *7*, 4200. 10.21105/joss.04200.

44. Theobald, D.L., and Wuttke, D.S. (2006). THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics *22*, 2171-2172. 10.1093/bioinformatics/btl332.

45. Ginn, H.M. (2022). Torsion angles to map and visualize the conformational space of a protein. bioRxiv, 2022.2008.2004.502807. 10.1101/2022.08.04.502807.

46. Elchebly, M., Payette, P., Michaliszyn, E., Cromlish, W., Collins, S., Loy, A.L., Normandin, D., Cheng, A., Himms-Hagen, J., Chan, C.C., et al. (1999). Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1B gene. Science *283*, 1544-1548. 10.1126/science.283.5407.1544.

47. Krishnan, N., Koveal, D., Miller, D.H., Xue, B., Akshinthala, S.D., Kragelj, J., Jensen, M.R., Gauss, C.M., Page, R., Blackledge, M., et al. (2014). Targeting the disordered C terminus of PTP1B with an allosteric inhibitor. Nat Chem Biol *10*, 558-566. 10.1038/nchembio.1528.

48. Konrad, M.R., Shelly, A.C., Zhaohong, Q., Kaveh, F., Fariba, S., Li, Z., Michael, A.Z., Alexandre, F.R.S., and Hsiao-Huei, C. (2020). Neuronal Protein Tyrosine Phosphatase 1B Hastens Amyloid β-Associated Alzheimer's Disease in Mice. The Journal of Neuroscience *40*, 1581. 10.1523/JNEUROSCI.2120-19.2019.

49. Liu, R., Mathieu, C., Berthelet, J., Zhang, W., Dupret, J.M., and Rodrigues Lima, F. (2022). Human Protein Tyrosine Phosphatase 1B (PTP1B): From Structure to Clinical Inhibitor Perspectives. Int J Mol Sci *23*. 10.3390/ijms23137027.

50. Andersen, J.N., and Tonks, N.K. (2004). Protein tyrosine phosphatase-based therapeutics: lessons from PTP1B. In Protein Phosphatases, J.n. Ariño, and D.R. Alexander, eds. (Springer Berlin Heidelberg), pp. 201-230. 10.1007/978-3-540-40035-6_11.

51. Zhang, Z.Y. (2017). Drugging the Undruggable: Therapeutic Potential of Targeting Protein Tyrosine Phosphatases. Acc Chem Res *50*, 122-129. 10.1021/acs.accounts.6b00537.

52. Wiesmann, C., Barr, K.J., Kung, J., Zhu, J., Erlanson, D.A., Shen, W., Fahr, B.J., Zhong, M., Taylor, L., Randal, M., et al. (2004). Allosteric inhibition of protein tyrosine phosphatase 1B. Nat Struct Mol Biol *11*, 730-737. 10.1038/nsmb803.

53. Choy, M.S., Li, Y., Machado, L., Kunze, M.B.A., Connors, C.R., Wei, X., Lindorff-Larsen, K., Page, R., and Peti, W. (2017). Conformational Rigidity and Protein Dynamics at Distinct Timescales Regulate PTP1B Activity and Allostery. Mol Cell *65*, 644-658 e645. 10.1016/j.molcel.2017.01.014.

54. Cui, D.S., Beaumont, V., Ginther, P.S., Lipchock, J.M., and Loria, J.P. (2017). Leveraging Reciprocity to Identify and Characterize Unknown Allosteric Sites in Protein Tyrosine Phosphatases. J Mol Biol *429*, 2360-2372. 10.1016/j.jmb.2017.06.009.

55. Popovych, N., Sun, S., Ebright, R.H., and Kalodimos, C.G. (2006). Dynamically driven protein allostery. Nature Structural & Molecular Biology *13*, 831-838. 10.1038/nsmb1132.

56. Venkitakrishnan, R.P., Zaborowski, E., McElheny, D., Benkovic, S.J., Dyson, H.J., and Wright, P.E. (2004). Conformational Changes in the Active Site Loops of Dihydrofolate Reductase during the Catalytic Cycle. Biochemistry *43*, 16046-16055. 10.1021/bi048119y.

57. Petit, C.M., Zhang, J., Sapienza, P.J., Fuentes, E.J., and Lee, A.L. (2009). Hidden dynamic allostery in a PDZ domain. Proceedings of the National Academy of Sciences *106*, 18249-18254. 10.1073/pnas.0904492106.

58. Pohl, F.M. (1971). Empirical Protein Energy Maps. Nature New Biology *234*, 277-279. 10.1038/newbio234277a0.

59. Miyazawa, S., and Jernigan, R.L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules *18*, 534-552. 10.1021/ma00145a039.

60. Godzik, A. (1996). Knowledge-based potentials for protein folding: what can we learn from known protein structures? Structure *4*, 363-366. 10.1016/s0969-2126(96)00041-x.

61. Dunbrack, R.L., Jr., and Cohen, F.E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci *6*, 1661-1681. 10.1002/pro.5560060807.

62. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583-589. 10.1038/s41586-021-03819-2.

63. Roney, J.P., and Ovchinnikov, S. (2022). State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. Physical Review Letters *129*, 238101. 10.1103/PhysRevLett.129.238101.

64. van Montfort, R.L.M., Congreve, M., Tisi, D., Carr, R., and Jhoti, H. (2003). Oxidation state of the active-site cysteine in protein tyrosine phosphatase 1B. Nature *423*, 773-777. 10.1038/nature01681.

794 65. Salmeen, A., Andersen, J.N., Myers, M.P., Meng, T.-C., Hinks, J.A., Tonks, N.K., and
795 Barford, D. (2003). Redox regulation of protein tyrosine phosphatase 1B involves a
796 sulphenyl-amide intermediate. Nature *423*, 769-773. 10.1038/nature01680.
797 66. Barrett, W.C., DeGnore, J.P., König, S., Fales, H.M., Keng, Y.-F., Zhang, Z.-Y., Yim,
798 M.B., and Chock, P.B. (1999). Regulation of PTP1B via Glutathionylation of the Active
799 Site Cysteine 215. Biochemistry *38*, 6699-6705. 10.1021/bi990240v.
800 67. Netto, L.E.S., and Machado, L.E.S.F. (2022). Preferential redox regulation of cysteine-
801 based protein tyrosine phosphatases: structural and biochemical diversity. The FEBS
802 Journal *289*, 5480-5504. https://doi.org/10.1111/febs.16466.
803 68. Yang, C.-Y., Yang, C.-F., Tang, X.-F., Machado, L.E.S.F., Singh, J.P., Peti, W., Chen, C.-
804 S., and Meng, T.-C. (2023). Active-site cysteine 215 sulfonation targets protein tyrosine
805 phosphatase PTP1B for Cullin1 E3 ligase-mediated degradation. Free Radical Biology
806 and Medicine *194*, 147-159. https://doi.org/10.1016/j.freeradbiomed.2022.11.041.
807 69. Ravichandran, L.V., Chen, H., Li, Y., and Quon, M.J. (2001). Phosphorylation of PTP1B
808 at Ser50 by Akt Impairs Its Ability to Dephosphorylate the Insulin Receptor. Molecular
809 Endocrinology *15*, 1768-1780. 10.1210/mend.15.10.0711.
810 70. Bandyopadhyay, D., Kusari, A., Kenner, K.A., Liu, F., Chernoff, J., Gustafson, T.A., and
811 Kusari, J. (1997). Protein-Tyrosine Phosphatase 1B Complexes with the Insulin Receptor
812 in Vivo and Is Tyrosine-phosphorylated in the Presence of Insulin*. Journal of Biological
813 Chemistry *272*, 1639-1645. https://doi.org/10.1074/jbc.272.3.1639.
814 71. Cimermancic, P., Weinkam, P., Rettenmaier, T.J., Bichmann, L., Keedy, D.A., Woldeyes,
815 R.A., Schneidman-Duhovny, D., Demerdash, O.N., Mitchell, J.C., Wells, J.A., et al.
816 (2016). CryptoSite: Expanding the Druggable Proteome by Characterization and

Prediction of Cryptic Binding Sites. J Mol Biol *428*, 709-719. 10.1016/j.jmb.2016.01.029.

72. Shi, C., Luo, S., Xu, M., and Tang, J. (2021). Learning Gradient Fields for Molecular Conformation Generation. CoRR *abs/2105.03902*.

73. Xu, M., Luo, S., Bengio, Y., Peng, J., and Tang, J. (2021). Learning Neural Generative Dynamics for Molecular Conformation Generation. CoRR *abs/2102.10240*.

74. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., and Thiel, V. (2021). Coronavirus biology and replication: implications for SARS-CoV-2. Nature Reviews Microbiology *19*, 155-170.

75. Fan, K., Wei, P., Feng, Q., Chen, S., Huang, C., Ma, L., Lai, B., Pei, J., Liu, Y., and Chen, J. (2004). Biosynthesis, purification, and substrate specificity of severe acute respiratory syndrome coronavirus 3C-like proteinase. Journal of Biological Chemistry *279*, 1637-1642.

76. Goyal, B., and Goyal, D. (2020). Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. ACS Combinatorial Science *22*, 297-305. 10.1021/acscombsci.0c00058.

77. Weng, Y.L., Naik, S.R., Dingelstad, N., Lugo, M.R., Kalyaanamoorthy, S., and Ganesan, A. (2021). Molecular dynamics and in silico mutagenesis on the reversible inhibitor-bound SARS-CoV-2 main protease complexes reveal the role of lateral pocket in enhancing the ligand affinity. Scientific Reports *11*, 7429. 10.1038/s41598-021-86471-0.