


LoCoHD: a metric for comparing local environments of proteins

Received: 29 July 2023

Zsolt Fazekas^{1,2}, Dóra K. Menyhárd^{1,3} & András Perczel^{1,3}✉

Accepted: 22 April 2024

Published online: 13 May 2024

 Check for updates

Protein folds and the local environments they create can be compared using a variety of differently designed measures, such as the root mean squared deviation, the global distance test, the template modeling score or the local distance difference test. Although these measures have proven to be useful for a variety of tasks, each fails to fully incorporate the valuable chemical information inherent to atoms and residues, and considers these only partially and indirectly. Here, we develop the highly flexible local composition Hellinger distance (LoCoHD) metric, which is based on the chemical composition of local residue environments. Using LoCoHD, we analyze the chemical heterogeneity of amino acid environments and identify valines having the most conserved-, and arginines having the most variable chemical environments. We use LoCoHD to investigate structural ensembles, to evaluate critical assessment of structure prediction (CASP) competitors, to compare the results with the local distance difference test (IDDT) scoring system, and to evaluate a molecular dynamics simulation. We show that LoCoHD measurements provide unique information about protein structures that is distinct from, for example, those derived using the alignment-based RMSD metric, or the similarly distance matrix-based but alignment-free IDDT metric.

The Research Collaboratory of Structural Bioinformatics Protein Data Bank (RCSB-PDB)¹ currently contains more than 217,000 experimentally determined protein structures. As the function of a protein is closely linked to its structure, this database provides valuable information on biological processes related to evolution, development, disease progression, drug design or agriculture, to name a few. In order to understand the behavior of these 3D atomic arrangements, computational tools and algorithms have been developed for their numerical analysis. In silico methods, such as molecular dynamics protocols²⁻⁴, molecular modeling software⁵⁻⁷, AI-based systems⁸⁻¹⁰, and de novo protein design platforms^{11,12} have also become increasingly popular and accessible, generating a vast amount of structural data concerning bio-macromolecular systems.

The comparison of different conformations of the same protein, or of the structures of similarly folded but different proteins, can be realized either by metrics, by generalized metrics (which are less

restricted), or by similarity measures (for an overview of some of the available methods¹³⁻³¹, please refer to Supplementary Note 1 and Supplementary Table 1). However, if the intention is to compare two proteins based on the chemical environment of their components, the commonly used measures do not provide focused information. The appearance or disappearance of different side-chain interactions (or interaction networks), changes in salt bridges, hydrogen bonds, π -cation interactions, polar-polar contacts, hydrophobic cores are all vital information. Since the nature of these environments dictates how proteins fold, move, or interact with each other, it should be critical to develop a method that provides an objective measure for their characterization.

With this in mind, we developed the local composition Hellinger distance (LoCoHD) metric presented here, which measures the chemical and structural difference between two local environments in proteins. We aim to provide a highly flexible scheme for objective

¹Laboratory of Structural Chemistry and Biology, Institute of Chemistry, ELTE Eötvös Loránd University, Budapest, Hungary. ²ELTE Hevesy György PhD School of Chemistry, ELTE Eötvös Loránd University, Budapest, Hungary. ³HUN-REN-ELTE Protein Modeling Research Group, ELTE Eötvös Loránd University, Budapest, Hungary. ✉e-mail: perczel.andras@ttk.elte.hu

comparison of two arbitrary atomic arrangements within a protein, while keeping the evaluation simple, intuitive, and relatively fast.

Results

Distribution of LoCoHD scores

In order to assign meaning to the absolute size of a score, and to decide whether a given score is “large” or “small”, it is important to know the underlying distribution from which the score is sampled. Therefore, we set out to determine how the LoCoHD scores for the FA+Cent and CG+Cent typing schemes are distributed when the uniform weight function is used between 3 Å and 10 Å (see the Methods section, Description of the LoCoHD Algorithm subsection for clarification). We chose these schemes for this initial investigation because the residue centroid primitive atoms can serve as anchors when comparing any residue type with any other. This feature is necessary for the random sampling protocol described in the Methods section (Random LoCoHD distribution generation subsection). The distributions of these random samples are shown in Supplementary Fig. 1.

Although theoretically LoCoHD scores can range from 0 to 1, it can be seen that even random residue-pairs do not frequently achieve values greater than 40%. The resulting experimental distributions can be modeled with β -distributions with parameters $\alpha=10.52$ and $\beta=33.48$ (p -value is 2.22×10^{-12} according to the Kolmogorov-Smirnov test) for typing scheme FA+Cent, and parameters $\alpha=12.99$ and $\beta=35.48$ (p -value= 1.70×10^{-9}) for typing scheme CG+Cent. For FA+Cent, the mean LoCoHD value is around 23.98% with a standard deviation (StDev) of 6.37%, and data ranging from 5.38% to 79.23%. For CG+Cent on the other hand, the mean LoCoHD value is around 26.83% with a StDev of 6.27%, and data ranging from 5.25% to 62.28%. It is important to keep in mind, that these distributions come from sampling random residue-pairs, which means that usually not the same amino acid types are paired and compared, likely resulting in a higher average LoCoHD score. The residue-pairs showing the lowest and highest LoCoHD scores in case of the FA+Cent typing were also extracted from the process. The lowest value (5.38%) belongs to the residue pair of PDB ID 2IJX (a PCNA3 monomer³²), chain C, residue Ala¹⁷ and PDB ID 1XHK (an ATP-dependent Lon protease³³), chain B, residue Ala⁵⁰¹. Both of these residue environments are hydrophobic cores, containing mostly aliphatic side-chains from valines, leucines, and isoleucines. The highest value (79.23%) belongs to the residue pair of PDB ID 6JV7 (a rat complement protein³⁴), chain B, residue Gly²⁸ and PDB ID 3PLO (a PF10014 dioxygenase³⁵), chain A, residue Ile¹⁴³. The isoleucine's environment (10 Å around the residue) in 3PLO contains a significant amount of aromatic carbon primitive types coming from 10 different aromatic residues. This environment also has a relatively low charged primitive atom content, coming from only 4 residues. In contrast, the environment of glycine from 6JV7 contains 3 disulfide bridges and charged primitive types coming from 8 different residues. This way, the primitive type distribution of the two environments highly differ, resulting in the extremely high LoCoHD score.

Statistical descriptors for the different residue type pairs were also extracted from the random samples. The residue type pairs with the highest and lowest average LoCoHD scores are shown in Table 1 for the primitive typing scheme FA+Cent. A t-distributed stochastic neighborhood embedding (tSNE) was also performed using the mean LoCoHD of hetero-residue pairs (i.e. where the residue types are not the same), the result of which is shown in Fig. 1. Using this technique, we were able to map the 20 proteinogenic residues into a two-dimensional space, while preserving the topology dictated by their environmental similarities.

Our analysis shows that the LoCoHD score can distinguish between environments surrounding residues with different physico-chemical properties and different environment-organizing behaviors. The residue pair Val-Thr is the first hetero-residue pair in Table 1, i.e. it has the lowest average LoCoHD score. This means that, on average, the

environments of Val and Thr are very similar both in composition and arrangement. This phenomenon is due to the isoelectronic and isosteric relationship between these two amino acids. The homo-residue pair with the highest average LoCoHD score is arginine, an indication of the diversity of its environments; the arginine side-chains can be solvated in the bulk solvent, can participate in H-bonds, salt-bridges, and π -cation interactions through their guanidino groups, and can participate in hydrophobic interactions through the C β -C γ -C δ aliphatic chain. Arginine also has the highest average LoCoHD scores calculated against all other amino acids. The amino acid with the most similar environments to arginine environments is lysine, followed by glutamine and tyrosine. The high environmental similarity between arginine and lysine is easily explained by their positive charge (see Fig. 1, box C). For further examples, see sections Supplementary Note 2 and Supplementary Figs. 2 and 3 of Supporting Info.

The tSNE analysis in Fig. 1 provides us a visual aid for noticing patterns in the 20 by 20 residue-residue average LoCoHD matrix. Points on this scatter plot represent individual amino acid types, while inter-point distances correlate with the average residue-residue environment dissimilarities. Besides the previously mentioned patterns, other, otherwise intuitive relationships can also be observed. Some residues stay close together, like the residue-sets Glu-Gln-Lys-Arg, Ile-Leu-Phe-Tyr-Trp, or Met-His. It is easy to find common physicochemical patterns in these close amino acids: Lys and Arg are both long, positively charged residues, Ile-Leu-Phe-Tyr-Trp are all hydrophobic residues, with Phe-Tyr-Trp forming a sub-cluster and having aromatic side-chains, and Met-His are common metal-complexing residues. Another noticeable pattern is formed by the Gly-Pro-Ala triplet, far away from all other amino acids. These are the residues with the smallest relative surface areas (with Ser included)³⁶ and are known to disrupt secondary structural elements.

To further validate the connection between the LoCoHD score of two residue environments and the presence/absence of secondary chemical interactions these residues participate in, we wanted to connect these properties through a machine learning model. It is evident that a good performing model can only be constructed, if there is a learnable connection between the inputs and the desired output. We trained a Siamese feedforward neural network that inputs the interaction fingerprints of two residues and tries to output the LoCoHD belonging to those residue environments. An interaction fingerprint vector contains the type of the residue (out of the 20 possibilities) and the number of different interactions it forms (6 counting-dimensions: H-bonds, van der Waals interactions, disulfide bonds, salt bridges, π - π stacking and π -cation interactions). It is important to note that this vector does not count the interactions that are present in the environment but are not formed by the central residue. Using these inputs the model was able to approximate the LoCoHD score (typing scheme: FA+Cent) of the two residue environments with a root mean squared error of 5.57% (validation: 5.57%) and a mean absolute error of 4.45% (validation: 4.39%). The final SpR between the predicted and true LoCoHD values came to be 0.454 (Supplementary Fig. 4). This result shows that a strong connection can be drawn between the LoCoHD score and the interaction fingerprint of the two central residues, but the information provided by the fingerprints is far from complete for an accurate score prediction.

Based on LoCoHD scores it was also possible to assess the environmental changes caused by functionally significant single mutations (and the absence of such in case of benign substitutions, see Supplementary Note 3-4 and Supplementary Figs. 5-8).

Comparison of CASP14 contestants through LoCoHD and IDDT

We tested the performance of five CASP14 contestants against the LoCoHD scoring system and compared these results to IDDT scoring, one of the scores used in the competition. 31 target structures were collected from the CASP14 archive. We compared the environment of

Table 1 | Rows show the residue-type pairs having the smallest 5 and largest 5 mean LoCoHD scores

Type pair	Number of samples	Mean	Median	StDev	Confidence interval	Minimum	Maximum
Val-Val	2507	17.47%	16.76%	4.97%	0.16%	6.44%	38.74%
Val-Thr	3875	18.53%	17.95%	5.27%	0.14%	6.80%	45.36%
Ala-Ala	3105	18.60%	17.92%	5.45%	0.16%	5.38%	41.35%
Ile-Ile	1633	18.90%	18.16%	5.29%	0.22%	7.50%	43.40%
Val-Ile	3928	18.94%	18.41%	5.08%	0.13%	6.78%	40.22%
... 200 additional rows...							
Asp-Arg	3074	29.31%	28.98%	6.05%	0.18%	13.48%	52.51%
Gly-Arg	3710	29.43%	28.94%	6.35%	0.17%	11.69%	55.33%
Ser-Arg	3281	29.56%	29.30%	6.37%	0.18%	13.39%	56.21%
Pro-Arg	2289	29.81%	29.36%	6.34%	0.22%	13.58%	63.09%
Cys-Arg	762	29.94%	29.58%	6.08%	0.36%	11.81%	48.34%

Primitive typing scheme FA+Cent with residue centroid anchors and a uniform 3 Å to 10 Å weight function were used. Two sided confidence intervals were calculated with a 95% upper confidence bound. Source data are provided as a Source Data file.

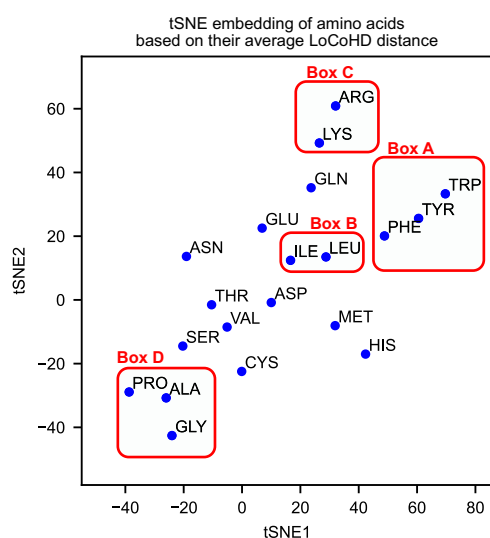


Fig. 1 | Average LoCoHD-based residue embedding. The t-distributed stochastic neighborhood embedding (tSNE) of the 20 proteinogenic residues based on the average LoCoHD scores of their environments for the primitive FA+Cent typing scheme. It can be seen that LoCoHD scores can cluster together environments around residues with similar physicochemical properties. In the embedding the large aromatic residues (Phe, Tyr, and Trp) form a cluster (box A), the highly similar Leu and Ile are close together (box B), the two positive residues Arg and Lys form a close pair (box C), and the small, secondary structure disrupting residues (Gly, Ala, Pro) also form a close triple (box D). Source data are provided as a Source Data file.

each true (experimental) and predicted residue for each contestant and target structure. This was done using the primitive FA+Cent typing scheme with a uniform weight function between 3 Å and 10 Å. Only hetero-residue contacts were allowed. The geometric center of each residue was used as the anchor atom.

We generated a dataset for each predicted structure, where each data point in the dataset belongs to one residue. The data points are two-dimensional vectors, with the residue's LoCoHD score as its first coordinate and the residue's IDDT score as its second coordinate. Different statistical descriptors were then generated for each dataset, namely the per-residue median of the LoCoHD and IDDT values (denoted prm-LoCoHD and prm-IDDT, respectively), and also their Spearman's rank correlation coefficient (SpR). The full statistics of these descriptors are reported in Supplementary Table 2. Here, we focus only on the median values, which can be found in Table 2. We

Table 2 | IDDT and LoCoHD scoring statistics for the first five CASP14 contestants

	AlphaFold2 (TS427)	BAKER (TS473)	BAKER-experimental (TS403)	FEIG-R2 (TS480)	Zhang (TS129)
Median SpR(IDDT, LoCoHD)	-0.6788 (±0.1106)	-0.5550 (±0.1239)	-0.5257 (±0.1279)	-0.5071 (±0.1632)	-0.4847 (±0.1583)
Median prm-IDDT	0.8410 (±0.0884)	0.6000 (±0.1344)	0.5860 (±0.1304)	0.5350 (±0.1471)	0.5210 (±0.1347)
Median prm-LoCoHD	0.0814 (±0.0177)	0.1311 (±0.0216)	0.1340 (±0.0211)	0.1425 (±0.0215)	0.1482 (±0.0213)

In this table the median Spearman's correlation coefficient (SpR), per-residue median (prm-) IDDT and prm-LoCoHD values are reported over all datasets, i.e. protein structures. The standard deviation of these values are presented between parentheses. It can be observed, that the IDDT and LoCoHD scoring systems agree on the order of the five contestants (rows median prm-IDDT and prm-LoCoHD). Also, it can be seen that as the quality of the prediction decreases (IDDT decreases, LoCoHD increases) the magnitude of the median SpR value decreases (row median SpR). Source data are provided as a Source Data file.

also collected all data points into two-dimensional histograms, one for each predictor. These are depicted in Fig. 2.

Our first results show an agreement between the contestant-order set by the median IDDT and LoCoHD values, proving that LoCoHD is able to separate (true, predicted) structure pairs with high, but different similarities. This is to be expected from a proper scoring system, as similarity scores should converge to their maximum values as the similarity between the structure pair increases, while dissimilarity scores should tend towards their minimum value. These tendencies inherently create correlations between different scoring systems, with higher absolute correlations closer to similarity extremities. This effect is clearly reflected in the median SpR values in Table 2. For the best-performer AlphaFold2, the median SpR value among the different structures is around -0.679, indicating a relatively high correlation. As we progress downwards on the list however, the SpR steadily decreases down to -0.485, supporting the aforementioned statement.

We also inspected some special cases individually. Since for every CASP14 predictor there are five predicted structures belonging to one target (experimental) structure, we can select target structures for which the predictions are outstandingly different. Namely, for every predictor, we identified the target structures for which the predicted structures show the largest SpR, median LoCoHD, and median IDDT gaps. These can be seen in Supplementary Table 3.

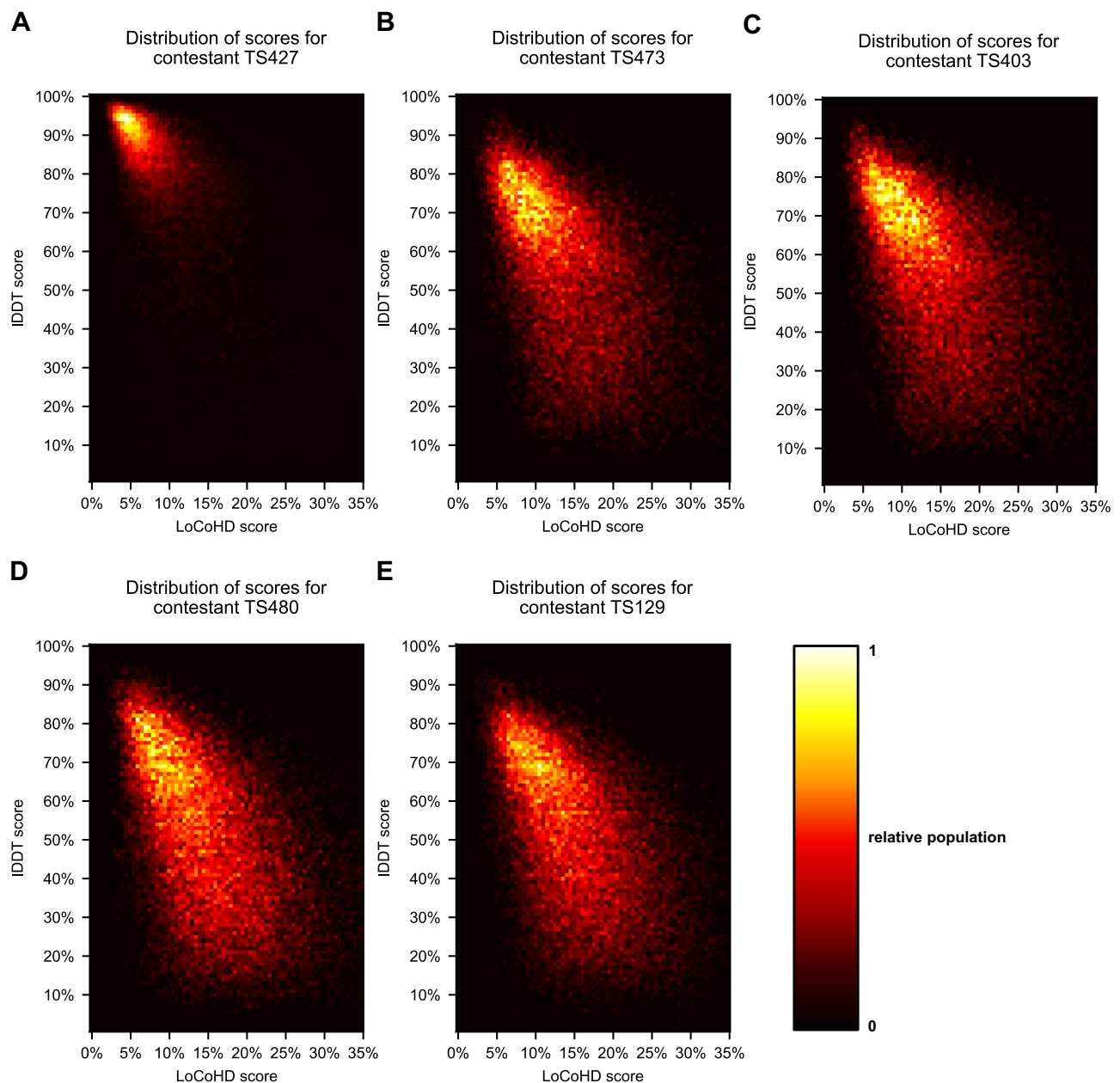


Fig. 2 | The LoCoHD-IDDt relationship. Visualizing the two-dimensional distributions of the per-residue LoCoHD-IDDt pairs for each CASP14 contestant examined. On panels **A–E** the histograms for the structure predictors AlphaFold2,

BAKER, BAKER-experimental, FEIG-R2, and Zhang can be seen, respectively. Histograms are depicted as heat maps, with warmer colors indicating higher populations in the corresponding area. Source data are provided as a Source Data file.

The AlphaFold2 predicted structures T1064TS427_1 and T1064TS427_5 show large SpR, median LoCoHD, and median IDDT differences (first row, Supplementary Table 3). The experimental structure behind target T1064 is the SARS-CoV-2 ORF8 accessory protein (PDB ID: 7JTL³⁷). Structures can be seen in Fig. 3, score-correlations, and per-residue LoCoHD scores can be seen in Supplementary Fig. 9. Although T1064TS427_1 has a much lower median LoCoHD than T1064TS427_5, the former structure has an outlier residue Lys⁹⁴ with an extremely high LoCoHD score at around 37% (residue numbering is according to the structure 7JTL) while this residue is by no means an outlier according to its IDDT score. In the experimentally determined structure, this lysine does not participate in interactions with any other residues and its N ζ faces the solvent bulk. In model T1064TS427_1, this lysine is H-bonded to the backbone of Leu¹¹⁸, while being close to Tyr⁷⁹ and Phe¹²⁰, which increase the aromatic content around Lys⁹⁴. In T1064TS427_5 the two aromatic

residues are farther apart than in the first structure, lowering the LoCoHD score to 27%. This high environmental difference is not reflected in the IDDT scores belonging to Lys⁹⁴ (37% in T1064TS427_1 and 38% in T1064TS427_5). The residue Glu¹⁰⁶ has the largest LoCoHD score in T1064TS427_5, but it does not appear in the top 10 residues with the largest LoCoHD scores in T1064TS427_1. In the experimental structure, Glu¹⁰⁶ participates in a relatively isolated H-bonding interaction with Tyr³⁸. Although in T1064TS427_1 the two aforementioned residues are too far away for this H-bond to form, they are still close together, creating a similar environment around the glutamic acid. In T1064TS427_5 Glu¹⁰⁶ faces the solvent bulk, while Tyr³⁸ forms a π - π interaction with Tyr¹⁰⁵, which creates a highly different environment, than in the true structure. This example also shows that LoCoHD is able to differentiate chemically significant changes in protein structures that might be overlooked when using purely structure-based comparisons.

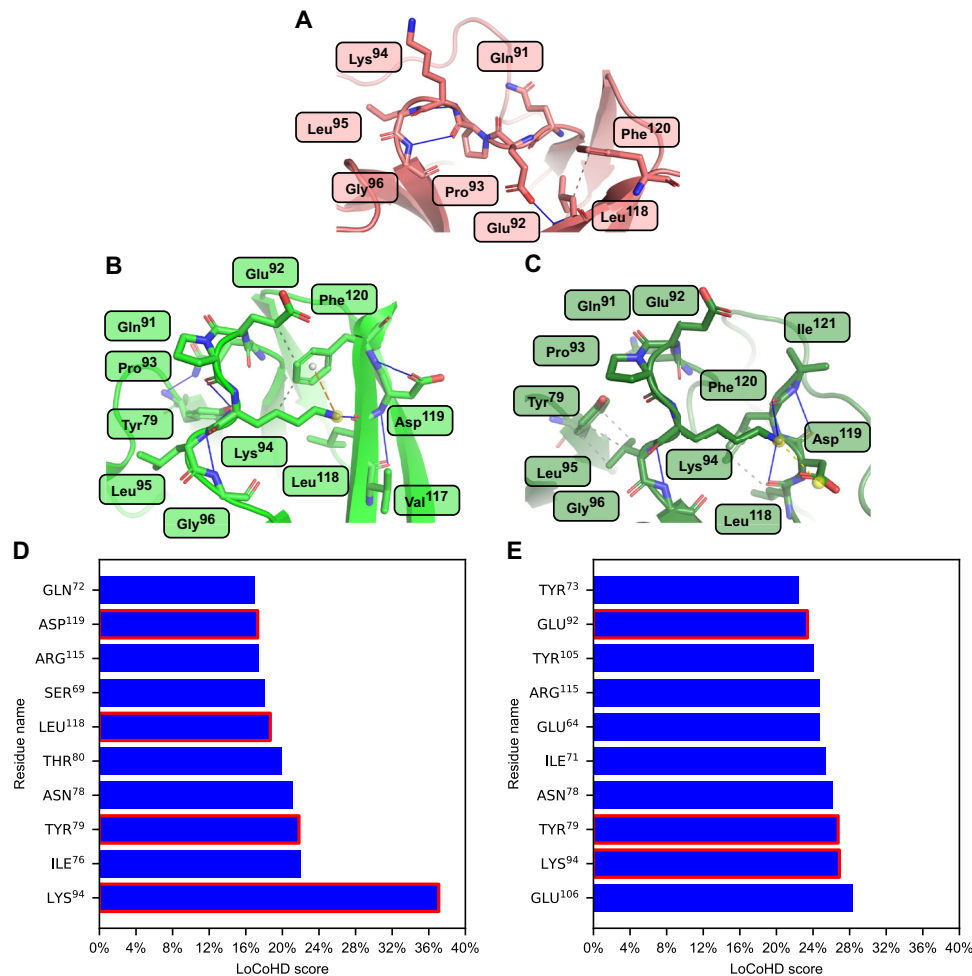


Fig. 3 | Case study of the prediction target T1064. Summarizing the residue environmental differences between the AlphaFold2 predicted structures T1064TS427_1 and _5 and the true, experimental SARS-CoV-2 ORF8 accessory protein structure (T1064). Panels A–C show the environments of the residue Lys⁹⁴ in the experimental structure, in T1064TS427_1 and in T1064TS427_5, respectively. Low opacity yellow spheres denote ionic interaction centers, yellow dashed lines denote ionic interactions, orange dashed lines denote pi-cation interactions, full blue lines denote H-bridges, gray spheres denote pi interaction centers and short-

dashed gray lines denote van der Waals contacts. It can be seen that while in the true structure this lysine faces the solvent, in the predicted structures this residue participates in several inter-residue interactions. On panels D and E the first ten residues can be seen having the largest LoCoHD scores in T1064TS427_1 and in T1064TS427_5, respectively. For the residues that are present either on panel B or C, the bars are highlighted with a red contour. Source data are provided as a Source Data file.

Similar analysis was performed on some of the CASP15 contestants and their predicted structures. Global IDDT, LoCoHD and CAD-score³¹ statistics for contestants TS229 (Yang-Server), TS278 (PEZY-Foldings), TS439 (Yang) and TS074 (DFolding) are presented in Supplementary Table 4 and in Supplementary Fig. 10. The former three contestants all achieve a median prrm-IDDT of about 0.85, a performance similar to that of AlphaFold2 in CASP14. This performance is also reflected by the low median prrm-LoCoHD values and again, the two scoring systems agree on the contestant order. Meanwhile, Supplementary Figs. 11–14 show the IDDT-LoCoHD analysis of two target structures and their predictions; H1166TS278_1 and _5, which is a human Fab S24-188 in complex with the N-terminal domain of the SARS-CoV-2 Nucleocapsid protein (to be published, PDB ID: 7SUE), and H1144TS278_1 and _5, which is a mouse/alpaca CNPase-Nb8d nanobody-antigen complex (to be published). The environmental differences in these complex structures, which were highlighted by LoCoHD, are chemically intuitive and are helpful in pointing out how a predictor weighs the relevance of the inter-residue interactions during the reconstruction of a complex.

Comparison of structure ensembles through LoCoHD and RMSD

For the analysis of the in-house determined NMR structure ensembles we chose the tryptophan cage fold extended by 5 residues, the so-called E5 miniprotein. The structural ensembles of this protein, each containing 50 different conformations, were previously determined at five different temperatures ranging from 277 K to 321 K³⁸. With increasing temperature, the protein gradually loses its well-defined tertiary structure, which is reflected in the diversity of residue conformations (Fig. 4A, B). Supplementary Fig. 15A, B show the 50 by 50 LoCoHD distance matrices averaged over all primitive atoms and plotted as heatmaps. These scores were calculated using the FA typing scheme, resulting in 197 environment comparisons per structure pair. Each cell in this matrix describes the relationship of two structures within the ensemble, i.e. it is the average LoCoHD score of the primitive atom environments computed for the two structures in question. Meanwhile, Supplementary Fig. 15C, D shows the LoCoHD scores of each primitive atom. These scores are the averaged LoCoHD values over all structure comparisons. For the analysis of the E5 ensembles, these views convey orthogonal information, one about the overall

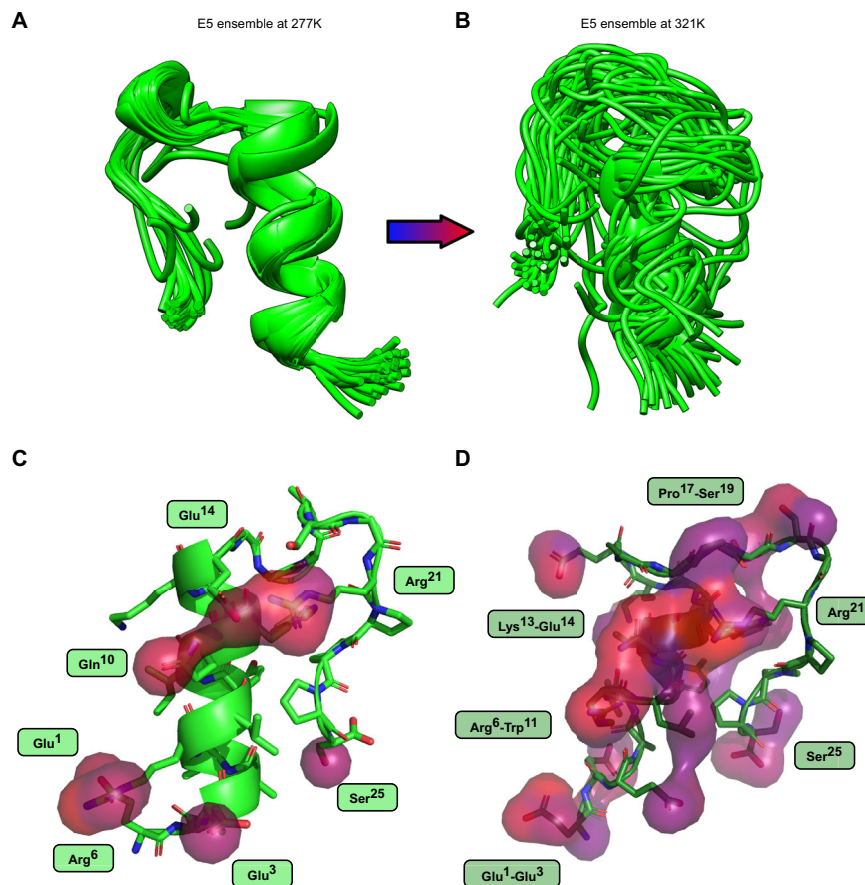


Fig. 4 | Comparison of the structures within the E5 ensembles. Panels **A** and **B** depict the structural ensembles of 50 conformers at temperatures of 277 K and 321 K, respectively. The E5 miniprotein consists of 25 residues, which fold into an N-terminal alpha helix, followed by a 3_{10} helix, and a C-terminal polyproline II helix. At higher temperatures these secondary structural elements disappear and the protein loses its well-defined structure. Panels **C** and **D** highlight the regions as surfaces within the 277 K and 321 K structures, respectively, where the ensemble-

average LoCoHD scores of the primitive atoms are above 20%. In the low mobility “cold” structure only a few primitive atoms are present that reach this threshold. In contrast, in the high mobility “hot” structure a lot of high LoCoHD score regions are highlighted. Residues or residue-intervals owning these high LoCoHD primitive atoms are also indicated in boxes. Surfaces are colored according to primitive atom LoCoHD scores, with warmer colors indicating higher values (from blue, through purple, to red). Source data are provided as a Source Data file.

dissimilarity of the ensemble elements (different conformations), and one about the environmental variability of the primitive atoms.

At 277 K the structure of E5 is well ordered. The overall environmental variability of the ensemble is low, as only 3-4 structures are significantly different from the others (i.e. structures with approximately 15-17% LoCoHD away from the other structures), and the median LoCoHD of the primitive atoms is 7.4%. These can be observed on the corresponding distance matrix (Supplementary Fig. 15A), where only the bottom rows and leftmost columns are shown in warm colors, and on the corresponding growth plot (Supplementary Fig. 15C), where most of the values are between 3.7% and 14.5%. In contrast, at 321 K the miniprotein shows high disorder and a broad conformational distribution. The distance matrix of this ensemble (Supplementary Fig. 15B) contains high LoCoHD values, mostly above 15%, with only about 5 structures showing some similarity (upper left blue patch). The median LoCoHD of the primitive atoms (Supplementary Fig. 15D) also shifted from 7.4% to 15.1%. This leaves the one sigma range of primitive atoms between 11.5% and 21.3% at 321 K.

The scatter-plots of the LoCoHD-RMSD pairs are depicted in Supplementary Fig. 15E, F for ensembles at 277 K and 321 K, respectively. In these plots, each point belongs to a structure-structure comparison, resulting in 1225 points per plot. The LoCoHD and primitive atom RMSD values show high correlation at all temperatures in the E5 ensembles. SpR values range from 0.74 (at 310 K) to 0.88 (at 277 K) with no obvious connection between the SpR and the

temperature of the ensemble. At low temperatures, these points form visually three (277 K) or two (288 K and 299 K) clusters, while at higher temperatures this behavior is not observed, leaving only one point cloud. The separation of these clusters mainly happens due to the RMSD metric, since it produces obvious boundaries, while the LoCoHD scores do not distinguish such sharp separating values in these cases. For example, at 277 K there are three “visually obvious” RMSD boundaries at approximately 2 Å, 3 Å and 3.75 Å. Performing agglomerative clustering using the RMSD matrix (with a distance threshold of 2 Å and complete linkage) results in five clusters. The first cluster consists of 46 structures and the remaining structures form clusters by themselves, indicating that these are outlier structures. However, despite the absence of an obvious LoCoHD boundary, when the same clustering procedure is applied on the LoCoHD matrix, but with the number of clusters set to five, the same outlier structures are identified. Also, when the LoCoHD-RMSD point cloud at 321 K is inspected, it can be noted that between the narrow range of LoCoHD scores of 17.1% and 18.7%, the points have a large spread along the RMSD axis (StDev = 1.0 Å and a maximum distance difference of 5.6 Å). This indicates the existence of a large structural difference-range within a small environmental composition difference-range.

An NMR-derived ensemble of the W316A, M317A mutant Gag-Pol polyprotein of HIV-1 (between residues Pro¹³³-Val³⁵³) was downloaded from the PED (PED ID: PED00072e000, residues numbered by UniProt ID: P12493, PDB ID: 2M8P) and subjected to similar analyses³⁹. These

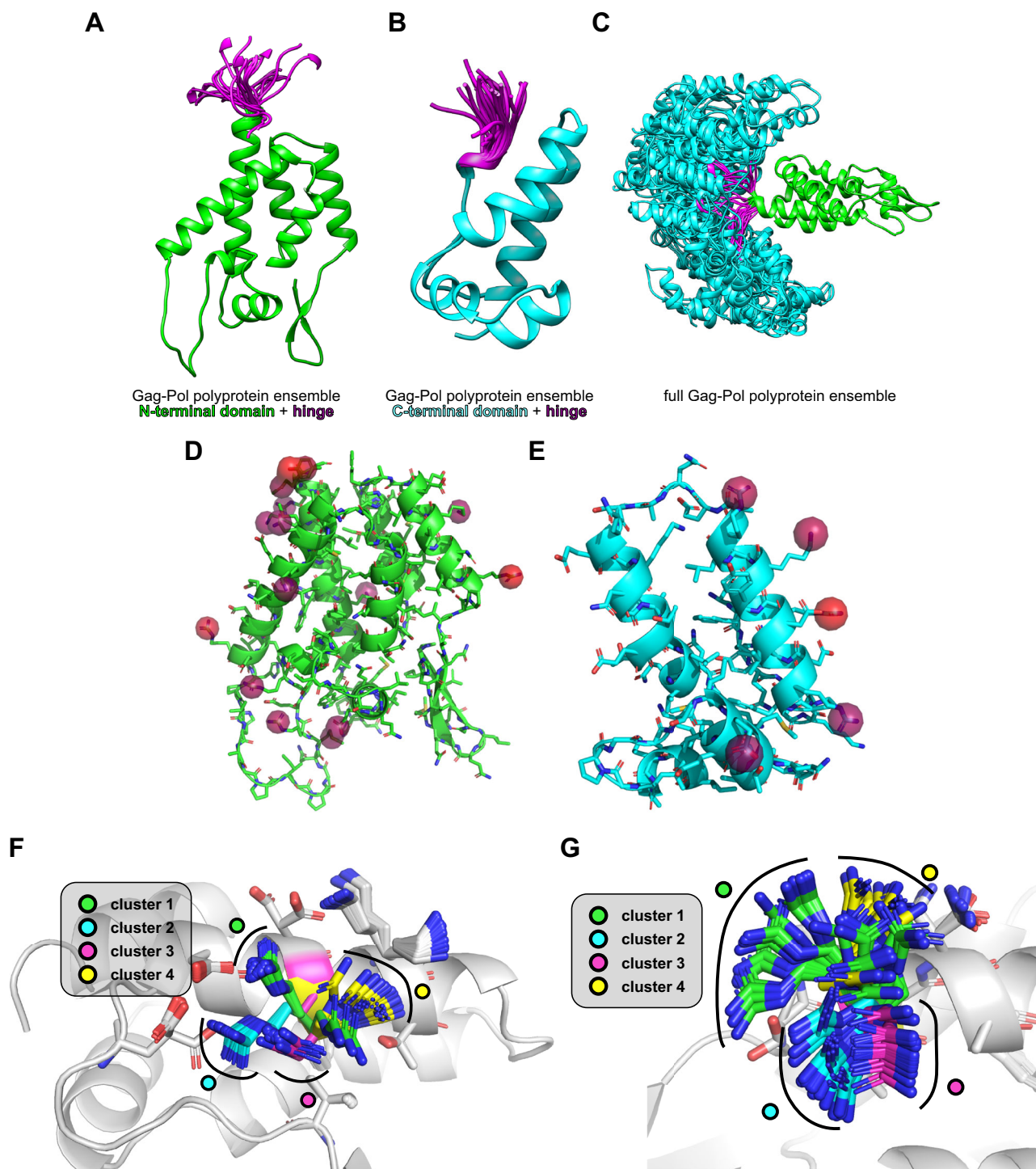


Fig. 5 | Case study using the HIV-1 Gag-Pol polyprotein. Panels **A–C** show the N- and C-terminal domains, and the full ensemble of the HIV-1 Gag-Pol polyprotein, respectively. Both panels show an ensemble of aligned structures. It can be seen, that while the domains themselves are very rigid, the hinge region connecting them (colored magenta) is highly mobile. Panels **D** and **E** show the regions in the N- and C-terminal domains with the highest (greater than 20%) average LoCoHD scores highlighted as semitransparent surfaces. These surfaces are also colored according

to the average LoCoHD score in that region. The relatively low number of such primitive atoms indicate low chemical variability. Panels **F** and **G** show realizations of LoCoHD clusterings with a cutoff value of 7%. These panels focus on residues Arg²⁹⁹ and Arg¹⁵⁰, respectively, along with their environments, which can (at least partially) explain the separation between the four clusters. Source data are provided as a Source Data file.

structures consist of two rigid domains, connected by a flexible hinge region between residues Ser²⁷⁸-Leu²⁸³ as depicted in Fig. 5A, B. The structural diversity of this ensemble is mainly due to the different conformational states of this hinge region (Fig. 5C). In ensembles like this, the RMSD after an all-atom alignment usually produces large and

uninformative values, since a global alignment cannot be performed optimally on the two domains at the same time.

For the Gag-Pol polyprotein, we used the CG typing scheme instead of the FA scheme, due to the large number of atoms present in the system and the resulting high computational time. Also, rather

than comparing the full ensemble of 100 structures, we only compared the first 50 structures in the ensemble. The molecular surface colored according to the LoCoHD values is depicted in Supplementary Fig. 16A, while the first twelve primitive atoms with the largest LoCoHD values are listed in Supplementary Table 5. The structure-structure LoCoHD distance matrix was also calculated and is shown in Supplementary Fig. 16B, along with the ordered primitive atom LoCoHD values in Supplementary Fig. 16C. When compared with the distance matrices of E5 (Supplementary Fig. 15A, B), a different pattern emerges. Blocks with low LoCoHD scores can be easily distinguished, suggesting high structural clusterability. This can also be seen in the LoCoHD - RMSD scatter-plot (Supplementary Fig. 16D), where an obvious gap can be seen between low and high LoCoHD values, separating the points into two clusters. In this case the RMSD values (calculated for all structure-structure pairs) are between 0 Å and 20 Å and do not discriminate clusters. The SpR between the LoCoHD values and the RMSD values is 0.24, which is lower than for E5 at any temperature. When agglomerative clustering is applied to the LoCoHD distance matrix with a distance threshold of 7% (a value in the gap between the two LoCoHD score clusters), 4 distinct clusters are produced. Performing the clustering on the RMSD matrix with a cluster number of 4 does not produce the same clusters. Again, this is in contrast to the case of E5, where clustering on the LoCoHD and RMSD matrices produced the same result. Based on the LoCoHD analysis, we can conclude that the fluctuation in relative domain positions causes little change in the chemical environment within each domain (Fig. 5D, E). The largest changes in LoCoHD indicate that, with the exception of the N-terminal 2 residues (Tyr²⁷⁷, Ser²⁷⁸) of the loop connecting the two domains, the structure is not perturbed by the domain movements - the most significant chemical changes within the domains are observed by three Arg residues occupying different niches as they rotate on the surface of the protein (Fig. 5F, G), independent of the large domain fluctuations.

Interestingly, one of these, Arg¹⁵⁰ (or Arg¹⁸, according to a different numbering convention), was shown crucial for the formation of the hexameric capsid of HIV-1^{40,41,42}. Mutations at this site result in distinct morphological variation of the viral assembly without causing conformational changes discernible by solid state NMR. LoCoHD identified this residue as being able to detect conformational fluctuations of the matrix - as would be expected of a residue that recognizes the presence of interaction partners and guides the assembly process.

An additional ensemble LoCoHD-RMSD comparison analysis can be found in Supplementary Note 5 and in Supplementary Fig. 17.

Using LoCoHD for the analysis of an MD simulation

Molecular dynamics simulations generate trajectories of proteins or protein complexes that represent the conformational changes of the systems under study. These trajectories are hundreds of thousands of time-correlated samples from a structural ensemble. Since a thorough visual inspection of these trajectories is problematic due to the size of these datasets, several numerical tools have been developed to plot the time dependence of different descriptors (such as RMSD, solvent accessible surface area, principal components, etc.). Here, by analyzing the MD trajectory of the dimeric form of a structural protein of the renal filtration barrier, podocin (UniProt ID: Q9NP85)⁴³, we show that the time-dependent LoCoHD score of different residues can pinpoint structurally important changes of the simulated protein. The CG+Cent typing scheme was used with the uniform weight function between 3 Å and 10 Å and the residue centroid primitive atoms as anchor atoms. The trajectory was analyzed between 600 ns and 1600 ns with 2.5 ns intervals. Each frame was compared to the first frame at 600 ns anchor atom by anchor atom, and the time dependence of the LoCoHD scores was recorded.

After visual inspection of the LoCoHD score vs. time plots, it became clear that some residues fluctuated between two or more

different environmental compositions and arrangements. To objectively select these residues from the 344-residue long homodimer, we calculated Sarle's bimodality coefficient (denoted by β)⁴⁴ for the LoCoHD distribution of each residue;

$$\beta = \frac{\gamma^2 + 1}{k}$$

where γ is the sample skewness of the distribution and k is the sample kurtosis. The higher this number is, the more likely it is that the distribution of these scores is bimodal. The value $\beta = 0.555$ is a good reference, since it belongs to the uniform distribution. Any distribution above $\beta = 0.555$ is likely to be bimodal. Using this procedure, we identified six residues - His²⁷⁶ (chain A, $\beta = 0.75$), Gly²⁷³ (chain B, $\beta = 0.66$), Met¹⁹⁷ (chain A, $\beta = 0.65$), Asp²⁶⁷ (chain B, $\beta = 0.64$), His²⁷⁶ (chain B, $\beta = 0.63$), and Phe¹⁷⁶ (chain B, $\beta = 0.63$) - as the residues with the most bimodal LoCoHD score distributions. The LoCoHD score vs. time plots for these residues are depicted in Supplementary Fig. 18.

The two environmental states of the residue His²⁷⁶ can be easily characterized by the X_1 angle (N-C α -C β -C γ dihedral angle) of the histidine. This dihedral angle takes on values from two angle-ranges, one between (+155°, -165°) (minor form), and one between (+40°, +90°) (major form). When His²⁷⁶ takes on the former conformation, the histidine side chain faces the bulk solvent and allows the backbone carbonyl group of Gly²⁷³ - another highly bimodal residue - to form a H-bond with the backbone NH group of Ser²⁷⁷, extending a short α -helix (Fig. 6A, B). However, when His²⁷⁶ is in its major form, it is positioned between Gly²⁷³ and Ser²⁷⁷, blocking the formation of the aforementioned H-bond and shortening the helix. This behavior is more dominant in the case of His²⁷⁶ in chain A. In the case of His²⁷⁶ in chain B, the histidine also moves away from the Gly²⁷³-Ser²⁷⁷ pair, but the formation of the Gly²⁷³-Ser²⁷⁷ backbone H-bond appears to be more sporadic than in the case of chain A.

In the case of the residue Met¹⁹⁷ in chain A, the two states are realized by the orientation of the Met side chain with respect to a hydrophobic core (Fig. 6C, D). In one setting, the C ϵ atom stays close to Val¹⁶⁵, Asp¹⁶⁶, Leu¹⁶⁷, Asn¹⁹⁹, and Ala²⁰⁰. This state is similar to the starting state ($t = 600$ ns) and has a low LoCoHD score (-8-10%). At about 1000 ns, the methionine side chain moves away from these residues and fills a previously unoccupied hydrophobic cavity, created by Val¹⁶⁵, Gln¹⁷⁰, Tyr¹⁹⁵, Leu²⁰³, Val²¹⁰, and Ile²⁵⁸. This state is different from the initial state and has higher LoCoHD scores (-16-20%). Here, the saturability of environments can be easily observed, when the LoCoHD time dependence of Ala²⁰⁰ and Leu²⁰³ are inspected (Fig. 6E, F). Ala²⁰⁰ has relatively few neighbors besides Met¹⁹⁷, and thus the same sharp change in LoCoHD score can be observed at 1000 ns, since its environmental composition is mainly determined by the primitive atoms from the methionine. In contrast, Leu²⁰³ is constantly surrounded by other residues (Leu²⁰⁴, Leu²⁰⁵, Leu²⁰⁷, Val²¹⁰, Ile²⁵⁸), and although Met¹⁹⁷ comes close to it at 1000 ns, no visible correlation between the two LoCoHD score time dependencies can be observed.

The 273-277 segment is in the critical hinge region of the podocin monomer, the flexibility of which was suggested to influence the effect that pathological mutations exert⁴³. Thus, recognizing that two interaction-wise different orientations are sampled by His²⁷⁶ may carry functional significance.

Discussion

The in silico study of protein structures requires precise mathematical and computational techniques capable of distinguishing between different conformational states and residue-residue interaction network topologies. Existing methods that aim to measure differences between atomic arrangements focus mainly on atomic coordinates or interatomic distances, ignoring important physicochemical differences between the states under study. To overcome these limitations, we

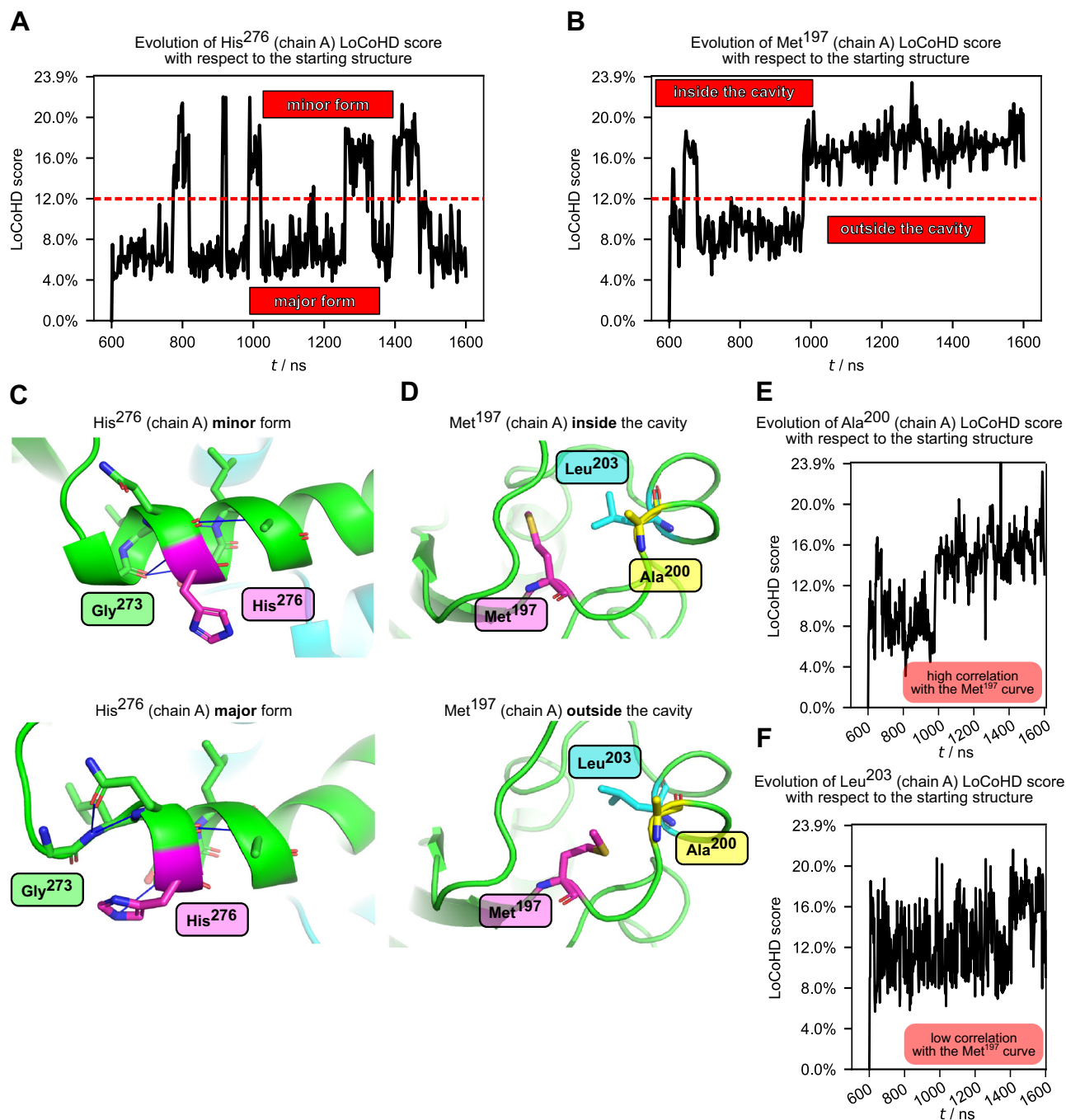


Fig. 6 | Use of LoCoHD for an MD analysis. Panel A shows the time dependence of the LoCoHD score of the highly bimodal residue His²⁷⁶, while also showing the clear separation of the two modes. Panel C shows the atomic representation of these modes, distinguishing a long (minor) and a short (major) helical form. Full blue lines denote H-bridges. The important interacting partners are highlighted by sticks. His²⁷⁶ is shown in purple. Panels B and D show the same representations for Met¹⁹⁷. Here the two forms are the occupied and empty cavity forms. The system starts

with the methionine sidechain outside the cavity and then fills the cavity at about 1000 ns due to the conformational change of the aforementioned sidechain. This behavior can also be followed from the time dependence of the LoCoHD score of the initially close residue Ala²⁰⁰ in panel E. In contrast, in the case of Leu²⁰³, which also becomes close to Met¹⁹⁷ after filling the cavity, this correlation is not present due to the environmental saturation effect (panel F). Source data are provided as a Source Data file.

have developed the local composition Hellinger distance (LoCoHD) metric and demonstrated that it is able to discriminate between different chemical compositions of residue environments. This metric is highly flexible, offering several customizable points in its workflow, and opens up an unexplored area of composition-based residue environment investigations. Specification of the task is completed in three conceptually separate steps: the choice of anchor atoms provides global spatial resolution and helps focusing on the region of

interest; the choice of weighting scheme provides local spatial resolution, allowing the specification of what the user considers to be the environment of an anchor atom; and the choice of the primitive typing scheme provides the chemical resolution, a chance to specify which atoms (or atom groups) should be differentiated based on their nature.

First, we were able to assert the environmental similarity of all amino acid type pairs between randomly selected and uncorrelated residue environments. In this way, it was possible to construct the full

distribution of LoCoHD scores, which provided comparative information with other LoCoHD measurements. It is important to note that this distribution is dependent on the primitive typing scheme and weight function and should be recalculated for each new setup. Nevertheless, for the two different typing schemes used for this task, we observed a β distribution of LoCoHD values in both cases. These had averages of 23.92% and 26.83%, providing good benchmarks for deciding whether a score is considered large or small. The average LoCoHD values of specific residue type and category pairs were also compared to these global values and to each other. These results respected the chemical intuition regarding the behavior of amino acids exceptionally well.

Secondly, the correlation between the LoCoHD score and the IDDT score, and the correlation between the LoCoHD score and the RMSD score were also examined. While these correlations were high for protein pairs with high structural similarity, the different information content of these descriptors became more apparent as the structures became more dissimilar. LoCoHD was able to rank the top five CASP14 competitors in the same order as IDDT. However, the absolute correlation between the residue IDDT scores and the residue LoCoHD scores gradually decreased as the predictive power of the competitors decreased. In the case of the RMSD score, protein ensembles with high internal structure-structure similarity showed a high LoCoHD-RMSD absolute correlation, but this absolute correlation also decreased with lower internal similarities, as in the case of multi-domain proteins or IDPs.

Finally, we demonstrated the use of the LoCoHD score in a molecular dynamics setup. This descriptor, used in a time-dependent manner, was able to pinpoint structurally important residues within the podocin dimer simulation. Highly bimodal LoCoHD score distributions corresponded to bimodal environmental states. We propose that inspection of time-dependent LoCoHD graphs can suggest trajectory convergence, highlight regions where residues undergo interaction mode changes, or - when compared to different energetically optimal environmental arrangements - even provide a distance measure from local optima.

Methods

Description of the LoCoHD algorithm

To characterize local chemical differences by calculating LoCoHD scores, two protein structures must be provided, which are then treated as labeled point clouds. In theory, this initial labeling can contain as much information as desired, but during the development and testing of LoCoHD we simply considered atoms to be centers of interest and labeled them with their standard PDB name and the name of the source residue to which the atom belongs. Next, the initial point clouds of both proteins are mapped to new point clouds, for which the new labels are chosen from a finite set called the “primitive type set”. These primitive types should preferably contain chemical quality descriptive information. For example, one may map the glutamic-acid O ϵ 1 and O ϵ 2 atoms to the negative oxygen primitive type (O $_{\text{neg}}$), while the serine O δ atom to the neutral oxygen primitive type (O $_{\text{neu}}$), discriminating the two chemically different oxygen-atom types. During the mapping from the initial atom cloud to the primitive atom cloud, any number of atoms can be omitted. Thus, in all of our calculations we only considered heavy atoms and ignored all H-atoms. Furthermore, virtual sites can also be introduced into the primitive atom cloud, like specific atom-set centroids or center-of-mass sites, with their own designated primitive types.

Once the primitive atom clouds have been created, a certain subset of primitive atoms must be selected from both clouds in such a way that for each primitive atom in one subset must have at least one corresponding primitive atom in the other. These atoms are called “anchor” atoms and they form the basis of the LoCoHD comparisons. For each corresponding anchor atom pair, our algorithm computes a

LoCoHD score, which reflects the difference in the primitive type composition between the environments of the anchor atoms. Since the selection of the anchor atoms defines not only the global spatial resolution and the focus area of the comparison but also its resource-efficiency, the applied anchoring-scheme has to be adapted to the task at hand.

The LoCoHD score for a given anchor atom pair (i, j) is calculated as follows:

$$\text{LoCoHD}_{ij} = \int_0^{\infty} w(r) H(\Phi_i(r), \Phi_j(r)) dr$$

where $w(r) \geq 0$ is a weight function whose integral on $(0, \infty)$ is 1, H is the Hellinger distance between the two probability mass functions (PMFs), $\Phi_i(r)$ and $\Phi_j(r)$, and $\Phi_i(r)$ is the distance-dependent environmental composition (DDEC) of the i th anchor atom: a vector with positive entries and an L1 norm of 1, and it has as many dimensions as many primitive atom types are used. $\Phi_i(r)$ contains the fraction of occurrence of each primitive type within a sphere around the i th anchor atom with a radius of r .

As an example, suppose a primitive type set of {A, B, C} is used and the environment of the anchor atom is described by the set {(A, 0 Å), (A, 1 Å), (B, 3 Å), (B, 5 Å)}, in which each entry is a (primitive type, distance from anchor atom) pair. This means that the anchor atom, which has a primitive type of ‘A’, is surrounded by 3 other primitive atoms. $\Phi(r = 2 \text{ Å})$ in this case would be (1, 0, 0), since only the primitive type ‘A’ is present in a 2 Å sphere around the anchor, while $\Phi(r = 4 \text{ Å})$ would be (0.66, 0.33, 0) (two ‘A’ and one ‘B’ type inside the sphere), and $\Phi(r = 7 \text{ Å})$ would be (0.5, 0.5, 0).

The Hellinger distance⁴⁵ of two PMFs (here, \mathbf{p} and \mathbf{q}) is given by:

$$H(\mathbf{p}, \mathbf{q}) = \sqrt{\frac{1}{2} \sum_i (\sqrt{p_i} - \sqrt{q_i})^2}$$

which guarantees a result between 0 and 1, with 0 meaning total PMF similarity, and 1 meaning total PMF dissimilarity. Since the weight function is chosen so that it satisfies the properties of a probability density function (PDF), the LoCoHD integral is a weighted average of Hellinger distances, also resulting in a value between 0 and 1.

Due to the discrete nature of atomic positions, the Hellinger distance between the two DDEC functions is constant for specific (r_n, r_{n+1}) intervals, with values denoted by H_n . This means, that the LoCoHD integral can be simplified into an easily computable form:

$$\text{LoCoHD} = \sum_{n=0}^{N-1} H_n \int_{r_n}^{r_{n+1}} w(r) dr = \sum_{n=0}^{N-1} H_n (W_{n+1} - W_n)$$

Here, we omitted the indices i and j of the anchor atoms for the purpose of readability. In these equations r_0 is considered 0 Å, while r_N is considered ∞ Å. The values W_n are the antiderivatives of $w(r)$ evaluated at r_n .

In addition to omitting all initial (i.e. non-primitive) atoms, it is also possible to omit primitive atom types from an environment depending on the central, anchor atom. This is an important feature, since an atom from a particular amino acid is always surrounded by the other atoms from the same amino acid. For small distances this will make the DDEC functions more similar, i.e. this will add a systematic, residue-type dependent bias into the LoCoHD scores. Therefore, it is advantageous to ignore primitive atoms that belong to the same residue as the anchor atom. This feature is referred to as “using only hetero-residue contacts”.

From the previous example, the environment can be expanded with residue-source information; the original set becomes {(A, 0 Å, X), (A, 1 Å, X), (B, 3 Å, Y), (B, 5 Å, Z)}, where each third component (X, Y, Z) denotes the residue-source. Since the anchor atom (A, 0 Å, X) is

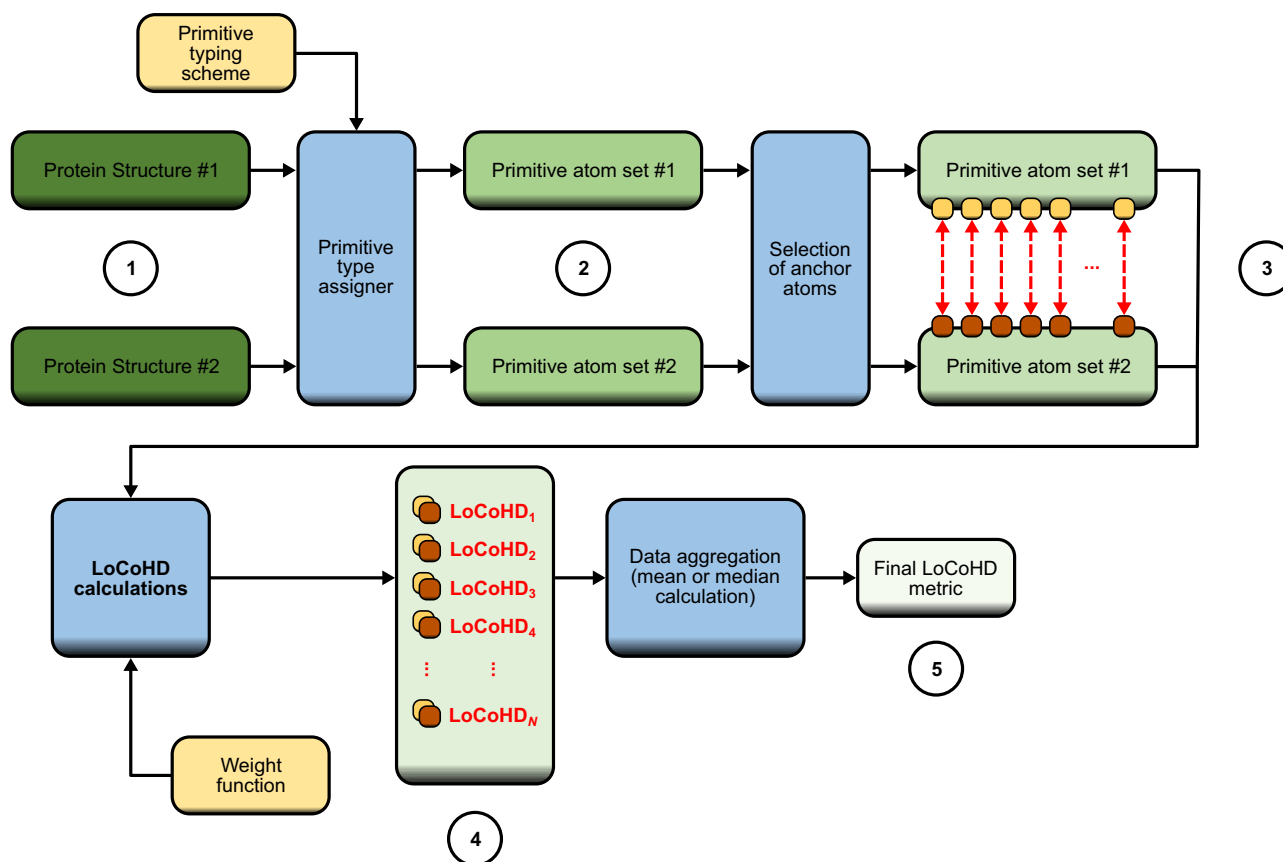


Fig. 7 | General flowchart for the LoCoHD algorithm. Starting from the protein structures, the procedure first maps the set of real atomic coordinates and names (stage 1) to primitive atoms (stage 2). How this mapping is done can be set through the primitive typing scheme. Then, a subset of these primitive atoms is selected as anchor atoms (light and dark brown squares, stage 3). The figure emphasizes the need for a surjective correspondence between these anchor atoms (red dashed

arrows). For each pair of anchor atoms a LoCoHD calculation is performed and the results of the environment comparisons are obtained (stage 4). The LoCoHD calculations are dependent on the shape of the weight function employed. One can use the set of these LoCoHD scores directly, or perform a final average (or median) calculation that yields a single number describing the structural similarity of the two protein structures (stage 5).

from the residue X , the primitive atom ($A, 1 \text{ \AA}, X$) would be omitted from the environment if only hetero-residue contacts are allowed.

The whole procedure is depicted in Fig. 7 and in Supplementary Fig. 19.

Primitive typing schemes

During our work, we investigated several different primitive typing schemes. These schemes can be characterized by either being “full atom” (FA) or “coarse grained” (CG) in nature, or whether they contain residue centroid primitive atoms or not (Cent: a virtual atom at the geometric center of every residue). These different primitive typing schemes are referred to as FA, CG, FA+Cent and CG+Cent.

In the FA typing scheme, all heavy atoms in the original structures are mapped to primitive atoms. The primitive types of these atoms are assigned from the following primitive type set: negative oxygen (O_{neg}), neutral oxygen (O_{neu}), positive nitrogen (N_{pos}), neutral nitrogen (N_{neu}), aliphatic carbon (C_{ali}), aromatic carbon (C_{aro}), and sulfur (S). In FA+Cent an additional residue centroid primitive atom is used, with a primitive type of Cent, and coordinates set by the geometric center of the residue’s heavy atoms. Thus, the DDEC functions produce 7-dimensional vectors in FA, and 8-dimensional vectors in FA+Cent.

In the CG typing scheme the following primitive types are distinguished: amide group carbon atoms (AmideC), alcoholic OH group oxygen atoms (OH), positive centers (Pos), negative centers (Neg), aromatic centers (Aro), aliphatic centers (Ali), and sulfur atoms (S).

Note, that some of these primitive atoms are not mapped from one atom, but rather from the geometric centers of certain atom-groups. Examples are the O δ 1-O δ 2 atom group of Asp for a negative center, or the C γ -C δ 1-C δ 2-N ϵ 1-C ϵ 2 atom group of Trp for an aromatic center. The CG+Cent typing scheme also contains the heavy atom centroids of the residues, similar to FA+Cent.

Schemes FA and CG are useful when a one-to-one correspondence can be established between each atom of the two structures, i.e. all resulting primitive atoms can be used as anchor atoms. This is the case when comparing different conformations of the same protein (as in the case of an NMR ensemble or the trajectory of a molecular dynamics simulation), or when comparing the experimental structure of a protein with its predicted structure (as in the case of CASP competitions). In contrast, the FA+Cent and CG+Cent typing schemes are useful when the two proteins to be compared do not have the same primary structure and thus contain different residues and atoms. In these cases, the Cent primitive atoms can serve as anchors, through which the LoCoHD calculations are performed.

Selection of the primitive atoms is again, task dependent. The FA and FA+Cent schemes provide the most chemical resolution. In the case of FA and CG, anchor pairing is only trivial if the two structures to be compared are comprised of the same atoms. Centroid-containing schemes (like FA+Cent and CG+Cent) can be used if residue-sized global spatial resolution suffices, and they also offer a way to reduce the runtime of the metric calculation.

Random LoCoHD distribution generation

To generate these experimental distributions, we used a homology-filtered PDB database from PISCES⁴⁶, culled on 2022.02.22. We used a maximum sequence identity of 25%, a resolution of 2 Å, an R-value of 0.25, and a protein chain length of 300 residues. This resulted in a database with a total of 3444 pdb files. The order of these pdb files was shuffled, and successive pairs of structures were considered using the shuffled order. For each pair of structures, all residues of the structure with the smaller number of residues were randomly paired with residues of the other structure. In this way, we were able to generate random residue pairs with uncorrelated environments. The LoCoHD values of these pairs were then calculated using the FA+Cent and CG+Cent typing schemes and the uniform weighting function between 3 Å and 10 Å. Only hetero-residue contacts were taken into account.

Construction of the LoCoHD predictor neural network

The neural network was constructed, compiled, trained and evaluated using the Python3 TensorFlow 2.15.0⁴⁷ package. A Siamese architecture was used. The network required a pair of 26-dimensional vectors as inputs, which can be split into two parts: a 20-dimensional one-hot encoded vector, designating the central residue type, and a 6-dimensional interaction-count vector, counting the number of interactions for each interaction type the central residue makes (H-bonds, van der Waals interactions, disulfide bonds, salt bridges, π - π stacking and π -cation interactions). These interactions were identified using the RING standalone software⁴⁸ and counted using in-house Python3 scripts. The network first creates internal representations of these vectors through a weight-shared feedforward 2-layered arm-pair with layer sizes of 256 and 128 neurons and ReLU activations. The resulting 128 dimensional vectors are then subtracted from each other and their difference is squared, resulting in a single 128 dimensional vector. Note, that this intermediate result is invariant with respect to the order of the inputs, making the network symmetric. Then, a simple, 3-layered feedforward network processes this further with layer sizes of 128, 64 and 1 neurons, and activations of ReLU, ReLU and sigmoid, respectively. The total number of learnable network parameters came to be 64641. Weight initialization was performed with the uniform Glorot initializer. Training was performed with the Adam optimizer (learning rate = 0.001) and the binary crossentropy loss (since the output can be thought of as a fuzzy binary categorization). Metrics of mean squared error and mean absolute error were used. Training was done on batch sizes of 64 for 3 epochs and with a validation split of 20/80. A total of 409408 environment-pairs were used for training obtained from the random LoCoHD distribution generation.

CASP naming convention

In CASP^{49–51}, each contestant and target structure has an identifier code in the form of TS[contestant ID] and T[target ID], respectively. Targets are sometimes prefixed with H, denoting heteromer target prediction, instead of T, denoting tertiary-structure target prediction. In addition, each contestant provides five predicted structures, which are denoted by the codes T[target ID]TS[contestant ID]_[structure ID]. An example for this kind of notation is T1026TS473_2, which is the second prediction of TS473 (the predictor named BAKER) for the target T1026 (FBNSV capsid protein, PDB ID: 6S44). We used this naming convention when referring to CASP protein structures.

Processing and score calculation for the CASP structures

Structures were downloaded as tar-files from the CASP archive and were preprocessed with in-house Python3 scripts. Briefly, all experimental and predicted structures were loaded into memory with BioPython 1.81⁵², non-canonical amino acids and disordered elements were removed from reference structures, correct chain-name-pairing was sought using sequence alignment based on the experimental

chains, and residues and atoms were removed from predicted structures that were not present in the experimental ones. Next, using all remaining atoms in the predicted structures, they were compared to the experimental structures using the IDDT and CAD score calculation modules of OpenStructure 2.7.0⁵³. These per-residue scores were further extended with our LoCoHD calculations.

Ensemble analyses

To investigate the differences between distance-matrix based LoCoHD calculations and alignment based RMSD calculations, we compared different conformations of protein structures within ensembles using both methods. These ensembles were obtained from previously published in-house NMR measurements³⁸ and also from the Protein Ensemble Database (PED)⁵⁴. The LoCoHD calculations were performed using the FA and CG (centroid-less) typing schemes and the uniform weight function between 3 Å and 10 Å, allowing hetero-residue contacts only. Since an ensemble contains different conformations of the same protein, all primitive atoms were used as anchor atoms. The atomic coordinates of the primitive atom sets were used for the calculation of the optimal rotation matrix in the singular value decomposition (SVD)-based alignment algorithm and also for the calculation of the RMSD values. Within an ensemble, each structure was compared to every other with respect to their primitive atoms, resulting in a total of $M * N * (N - 1) / 2$ comparisons, where M is the number of primitive atoms inside the protein and N is the number of structures within the ensemble. In other words, a symmetric, zero diagonal, N by N distance matrix was obtained for each primitive atom.

Visualizations

Structural visualizations were done using PyMol⁵ 2.5.0, while PLIP⁵⁵ 2.3.0 was used for the visualization of residue-residue interactions. Graphs and plots were created with Matplotlib⁵⁶ 3.8.2.

Statistics and reproducibility

Sample sizes in the PISCES dataset analysis were determined by the homology filtering method and the random residue pairing method detailed above. The reported statistical descriptors correspond to a single randomized run on all structures included in this study. No blinding was used for the assessment of the outcomes.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The datasets generated, analyzed and necessary for the reproduction of the case studies (except for the restricted part of the CASP15 dataset) are collected and available in a Figshare repository⁵⁷ with an accession code of <https://doi.org/10.6084/m9.figshare.24885540>. Besides the repository, all structural datasets used in this paper are also freely downloadable from the CASP database (https://predictioncenter.org/download_area/), from RCSB PDB (<https://www.rcsb.org>), from PED (<https://proteinensemble.org>), or from PISCES. The podocin MD trajectory and PDB accession code lists used in this study are also contained within the Figshare repository. Some of the CASP15 structures (and data related to them) are still under embargo by their authors' request and must be requested from the the CASP15 organizers at casp@predictioncenter.org. Specifically, in this manuscript, the CASP15 structures publicly available on 2023.12.31. were used, in addition to the restricted targets for invitees requested on 2024.02.16. The release of the embargoed data can be followed at <https://predictioncenter.org/casp15/targetlist.cgi> (Description column). Source data for the figures and tables are provided with this paper. Source data are provided with this paper.

Code availability

The Rust and Python code for the LoCoHD project, along with dependency descriptions^{52,58} and the Python scripts of the case studies are all available at the GitHub repository https://github.com/fazekaszlo/loco_hd. A release version of v0.1.4⁵⁹ was used for this study. Brief details for the implementation can be found in Supplementary Note 6.

References

1. RCSB PDB. <http://www.rcsb.org>. Accessed 2024-04-09 (2024).
2. Bauer, P., Hess, B. & Lindahl, E. *GROMACS 2022.3 Manual*. <https://doi.org/10.5281/ZENODO.7037337> (2022).
3. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the Amber biomolecular simulation package: Amber biomolecular simulation package. *WIREs Comput. Mol. Sci.* **3**, 198–210 (2013).
4. Eastman, P. et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).
5. *The PyMOL Molecular Graphics System, Version 2.0* Schrödinger, LLC. <https://pymol.org/2/>. Accessed 2023-07-14 (2024).
6. Maestro, S., LLC, <https://www.schrodinger.com/products/maestro> (2021). Accessed 2023-07-14.
7. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
8. Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. & Ascher, D. B. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* **47**, W338–W344 (2019).
9. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
10. Luttrell, J., Liu, T., Zhang, C. & Wang, Z. Predicting protein residue–residue contacts using random forests and deep networks. *BMC Bioinforma.* **20**, 100 (2019).
11. Koga, N. et al. Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
12. Pearce, R., Huang, X., Setiawan, D. & Zhang, Y. EvoDesign: Designing Protein–Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. *J. Mol. Biol.* **431**, 2467–2476 (2019).
13. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Cryst. A* **32**, 922–923 (1976).
14. Kaindl, K. & Steipe, B. Metric properties of the root-mean-square deviation of vector sets. *Acta Crystallogr. A Found. Crystallogr* **53**, 809–809 (1997).
15. Steipe, B. A revised proof of the metric properties of optimally superimposed vector sets. *Acta Crystallogr. A Found. Crystallogr* **58**, 506–506 (2002).
16. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
17. Holm, L. & Sander, C. Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
18. Wohlers, I., Domingues, F. S. & Klau, G. W. Towards optimal alignment of protein structure distance matrices. *Bioinformatics* **26**, 2273–2280 (2010).
19. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).
20. Siew, N., Elofsson, A., Rychlewski, L. & Fischer, D. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* **16**, 776–785 (2000).
21. Ortiz, A. R., Strauss, C. E. M. & Olmea, O. MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Sci.* **11**, 2606–2621 (2009).
22. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
23. Levitt, M. & Gerstein, M. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA* **95**, 5913–5920 (1998).
24. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
25. Chen, C., Chen, X., Morehead, A., Wu, T. & Cheng, J. 3D-equivariant graph neural networks for protein model quality assessment. *Bioinformatics* **39**, btad030 (2023).
26. Hamamsy, T. et al. Protein remote homology detection and structural alignment using deep learning. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-023-01917-2> (2023).
27. Simonovsky, M. & Meyers, J. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *J. Chem. Inf. Model.* **60**, 2356–2366 (2020).
28. Galgonek, J., Hoksza, D. & Skopal, T. SProt: sphere-based protein structure similarity algorithm. *Proteome Sci.* **9**, S20 (2011).
29. Zhou, X., Chou, J. & Wong, S. T. Protein structure similarity from principle component correlation analysis. *BMC Bioinforma.* **7**, 40 (2006).
30. Krasnogor, N. & Pelta, D. A. Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics* **20**, 1015–1021 (2004).
31. Olechnovič, K., Kulberkytė, E. & Venclovas, Č. CAD-score: A new contact area difference-based function for evaluation of protein structural models. *Proteins* **81**, 149–162 (2013).
32. Hlinkova, V. et al. Structures of monomeric, dimeric and trimeric PCNA: PCNA-ring assembly and opening. *Acta Crystallogr D. Biol. Crystallogr* **64**, 941–949 (2008).
33. Im, Y. J. et al. The Active Site of a Lon Protease from *Methanococcus jannaschii* Distinctly Differs from the Canonical Catalytic Dyad of Lon Proteases. *J. Biol. Chem.* **279**, 53451–53457 (2004).
34. Zuo, C. et al. Chimeric protein probes for C5a receptors through fusion of the anaphylatoxin C5a core region with a small-molecule antagonist. *Sci. China Chem.* **62**, 1371–1378 (2019).
35. Xu, Q. et al. Crystal structure of a member of a novel family of dioxygenases (PF10014) reveals a conserved cupin fold and active site: Crystal Structure of PF10014. *Proteins* **82**, 164–170 (2014).
36. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum Allowed Solvent Accessibilities of Residues in Proteins. *PLoS ONE* **8**, e80635 (2013).
37. Flower, T. G. et al. Structure of SARS-CoV-2 ORF8, a rapidly evolving immune evasion protein. *Proc. Natl Acad. Sci. USA* **118**, e2021785118 (2021).
38. Taricska, N. et al. The Route from the Folded to the Amyloid State: Exploring the Potential Energy Surface of a Drug-Like Miniprotein. *Chem. Eur. J.* **26**, 1968–1978 (2020).
39. Ganser-Pornillos, B. K., Cheng, A. & Yeager, M. Structure of Full-Length HIV-1 CA: A Model for the Mature Capsid Lattice. *Cell* **131**, 70–79 (2007).
40. Deshmukh, L. et al. Structure and Dynamics of Full-Length HIV-1 Capsid Protein in Solution. *J. Am. Chem. Soc.* **135**, 16133–16147 (2013).
41. Dick, R. A. et al. Inositol phosphates are assembly co-factors for HIV-1. *Nature* **560**, 509–512 (2018).
42. Lu, J.-X., Bayro, M. J. & Tycko, R. Major Variations in HIV-1 Capsid Assembly Morphologies Involve Minor Variations in Molecular Structures of Structurally Ordered Protein Segments. *J. Biol. Chem.* **291**, 13098–13112 (2016).
43. Tory, K. et al. Mutation-dependent recessive inheritance of NPHS2-associated steroid-resistant nephrotic syndrome. *Nat. Genet.* **46**, 299–304 (2014).
44. Knapp, T. R. Bimodality Revisited. *J. Mod. Appl. Stat. Meth.* **6**, 8–20 (2007).

45. M. S. Nikulin. Hellinger distance. In *Encyclopaedia of Mathematics* 78 (Springer, 2001).
46. Wang, G. & Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589–1591 (2003).
47. TensorFlow Developers. *TensorFlow*. Zenodo <https://doi.org/10.5281/ZENODO.4724125> (2023).
48. Clementel, D. et al. RING 3.0: fast generation of probabilistic residue interaction networks from structural ensembles. *Nucleic Acids Res.* **50**, W651–W656 (2022).
49. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round xv. *Proteins* **91**, 1539–1549 (2023).
50. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round xiv. *Proteins* **89**, 1607–1617 (2021).
51. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins* **87**, 1011–1020 (2019).
52. Biopython. <https://biopython.org/>. Accessed 2023-07-14 (2023).
53. Biasini, M. et al. *OpenStructure*: an integrated software framework for computational structural biology. *Acta Crystallogr. D. Biol. Crystallogr.* **69**, 701–709 (2013).
54. Lazar, T. et al. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.* **49**, D404–D411 (2021).
55. Adasme, M. F. et al. PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. *Nucleic Acids Res.* **49**, W530–W534 (2021).
56. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
57. Fazekas, Z. Datasets for the reproduction of the experiments presented in the LoCoHD article. 2090912615 Bytes Figshare <https://doi.org/10.6084/M9.FIGSHARE.24885540> (2023).
58. The PyO3 user guide. <https://pyo3.rs>. Accessed 2023-07-14 (2023).
59. ZsoltFazekas. LoCoHD: a Metric for Comparing Local Environments Of Proteins. *GitHub* <https://doi.org/10.5281/ZENODO.10848377> (2024).

Acknowledgements

This work was completed in the ELTE Thematic Excellence Programme supported by the Hungarian Ministry for Innovation and Technology. Project no. 2018-1.2.1-NKP-2018-00005 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2018-1.2.1-NKP funding scheme. Project number RRF-2.3.1-21-2022-00015 is implemented with the support of the European Union's Recovery and Resilience Instrument. Supported by the Ministry for Innovation and Technology from the Hungarian NRD Fund (2020-1.1.6-JÖVŐ-2021-00010). All funding were awarded to A.P.

Author contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. Zs.F. proposed the project, designed the algorithms and experiments, implemented the LoCoHD software, collected data, conducted the computations and analyzed the results. D.K.M. and A.P. helped with experiment design, analysis and manuscript proofing. A.P. supervised the project.

Funding

Open access funding provided by Eötvös Loránd University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-48225-0>.

Correspondence and requests for materials should be addressed to András. Perczel.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024