Supplemental Material for
"AlphaFold2 can predict single-mutation effects"

John M. McBride,[1,*] Konstantin Polev,[1,2] Amirbek Abdirasulov,[3] Vladimir Reinharz,[4] Bartosz A. Grzybowski,[1,5,†] and Tsvi Tlusty[1,5,‡]

[1] *Center for Soft and Living Matter, Institute for Basic Science, Ulsan 44919, South Korea*

[2] *Department of Biomedical Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea*

[3] *Department of Computer Science and Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea*

[4] *Université du Québec à Montréal, Canada*

[5] *Departments of Physics and Chemistry, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea*

Contents

1. PDB Structure Data

A. Full

We curate a set of structures from the PDB (downloaded 10 August 2020) to study the effect of mutations on protein structure. We first select all proteins from the PDB that have equal sequence length, for which there are multiple structures whose sequences differ by no more than 3 mutations; we include proteins with identical sequences as a control group. We only include proteins of length $50 \leq L \leq 500$. Binding can result in large structural change, so we control for this: we exclude all protein complexes, or proteins bound to RNA or DNA since large interactions lead to large deformations with relatively poor reproducibility across repeat measurements; we only match pairs of proteins if they are bound to the exact same types of small molecules, since these result in small, more reproducible deformations. We exclude NMR structures for simplicity, avoiding the need to determine additional cutoffs to infer disorder, or to choose a representative structural model; we also found that NMR structures were much less reliable (higher deformation between repeat measurements) than crystal structures. We only consider pairs that were prepared at a similar pH (within ± 0.5). We exclude pairs of structures that are almost identical (*e.g.*, from time-resolved crystallography experiments); *i.e.*, pairs with a RMSD smaller than 0.001 Å. This leaves us with 3,901 PDB structures, and ∼90,000 pairs; ∼70,000 of these pairs derive from a set of 485 endothiapepsin structures, so we remove most of these until we are left with 17,813 pairs.

B. Differences in crystallographic group

In about 10 % of pairs, the crystallographic groups are different. We find that this does not affect the PDB-AF defor-

mation correlations. We do find that it affects the absolute values of deformation, so we do not include these pairs in the distribution of S_i in the main text Fig. 1F.

C. Non-redundant sets

To create a non-redundant sample, we cluster protein sequences using CD-hit (Li and Godzik, 2006), with a 90 % sequence identity threshold. From the total pool of pairs, we create sub-samples with no more than 10 examples per group, for each value of M, the mutation number (in total, 2,636 pairs); to estimate sampling error, we re-run analyses with 1,000 sub-samples.

D. Structures not used in training AF

A major limitation of evaluating PDB-AF correlations is that AF was trained on the PDB structures that we use in evaluation. In our dataset, we found only 211 out of 17,813 cases where all structures in a matched set were deposited in the PDB after the cutoff used for the training set (28 August 2019). Out of these, 153 are structures of bovine trypsin. Hence, it is not currently possible to evaluate AF's performance on structures that it was not trained on, due to insufficient examples in this dataset.

2. High-throughput Phenotype Data

A. eqFP611

There are two variants of the fluorescent protein epFP611, mKate2 and mTagBFP2, which exhibit respectively red and blue fluorescence. These two proteins differ by only 13 mutations. A previous study measured blue and red fluorescence for all 2^{13} sequences that account for both mKate2, mTagBFP2, and all intermediate sequences (Poelwijk *et al.*, 2019). When comparing deformation with phenotype, we assign mKate2 to be the wild-type (WT) for red fluorescence, and mTagBFP2 to be the WT for blue fluorescence. We then calculate deformation of each variant compared to the WT, for each fluorescence colour. In the main text, we report deformation calculated using single structures using DeepMind's implementation of AF (6).

B. GFP

The GFP dataset is a subset (2,312 sequences) of the full dataset (51,715 sequences) published in (Sarkisyan *et al.*, 2016). The WT sequence was used to create a library of variants via random mutagenesis, and green fluorescence was measured for WT and all variants. The highest number of missense mutations is $M = 15$, although the average is much lower (3.7). To create a subset of the full dataset, we first grouped sequences by the number of mutations M, and picked at most 200 variants per value of M. We next chose variants in order to maximise the variance in fluorescence: We marked fluorescence values on a 200-point, equally-spaced grid from the minimum to the maximum fluorescence (for each value of M), and picked the variants with fluorescence closest to each grid point. Due to the sparsity of fluorescence values between 2 and 2.5, many grid points were assigned the same variant. We only included each variant once, and randomly chose variants to make up the remainder if there were less than 200 unique variants. When comparing deformation with phenotype, we calculate deformation of each variant compared to the WT. In the main text we report deformation calculated using averaged AF structures (6).

C. PafA

The PafA dataset contains the WT phosphate-irrepressible alkaline phosphatase (PafA) of Flavobacterium, and 1036 mutants (Markin *et al.*, 2021). All possible single-mutants involving substitutions to glycine or valine were created; if the native amino acid was glycine or valine, then the substitution was to alanine; one double mutant at the active site was also produced. For each protein, the catalytic rate constant k_{cat}, Michaelis constant K_{m}, and catalytic efficiency $k_{\mathrm{cat}}/K_{\mathrm{m}}$ were measured, alongside expression levels (using a GFP tag). To help infer whether mutations affected protein folding, the authors measured the kinetic constants at different temperatures and Zn concentrations, for different substrates, and with inhibitors. They then separated the kinetic measurements into components of both catalytic effects and folding effects (which they term the fraction of active enzymes). When comparing deformation with phenotype, we calculate deformation of each variant compared to the WT. In the main text, we report deformation calculated using averaged AF structures (6).

3. Structure Prediction

A. AlphaFold: DeepMind

We predict structures using AF with a default template cutoff date (14 May 2020) and a reduced genomic database. We run AF using all five pre-trained models. Models 1 and 2 use structural templates as input; see Sec. 12 for more detail on differences between models.

B. AlphaFold: ColabFold

After initially using the DeepMind's implementation of AlphaFold, we switched to using the ColabFold implementation as it is faster and allows more control over the internal parameters of the algorithm (Mirdita *et al.*, 2022). We used ColabFold to produce five repeat predictions of structures for each sequence, for all five AF models. This resulted in 25 ColabFold structures per sequence in addition to the 5 AlphaFold structures per sequence. We ran ColabFold without templates, and used 6 recylces per structure.

C. DMPfold

To put the results of AF in context, we also predict structures using DMPfold (Kandathil *et al.*, 2022). We chose DMPfold primarily because it is fast. Note that DMPfold is a deterministic algorithm so there is no repeat-prediction variability.

4. Deformation metrics

To study the effect of mutations on structure, we need an appropriate metric of structural change, or *deformation*. We can differentiate between measures by whether they measure local or global changes, and whether the measures are absolute, or normalized scoring functions. We will explain the types of measures that have been used, and why they

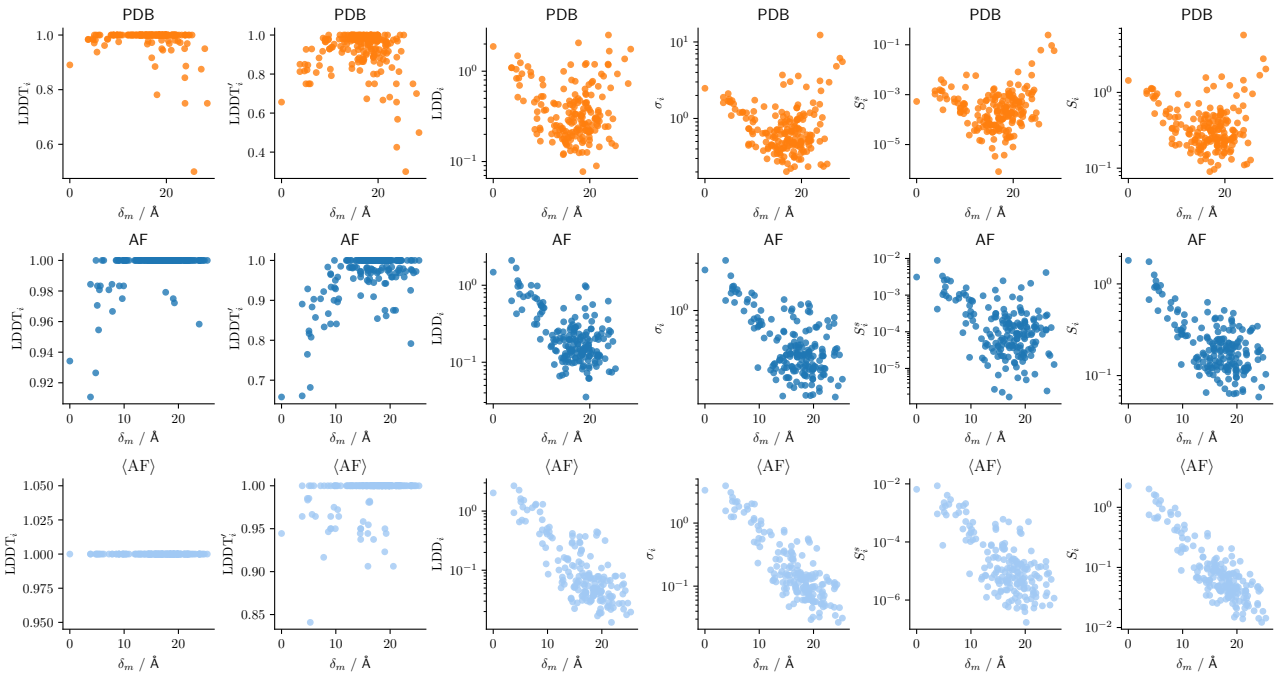

FIG. 1 **Examples of deformation metrics.** Deformation as a function of distance from nearest mutated site for deformation upon mutation from WT CypA (`6U5C_A`) to a double mutant (S99T, C115S, `6BTA_A`). Deformation is shown for 6 metrics, from left to right: LDDT; LDDT′ with more precise cutoffs ζ_k; LDD; neighborhood distance, σ_i; shear strain, S^s; effective strain (ES), S_i. For all metrics, $\gamma = 13$ Å. Deformation is calculated for PDB structures (top), AF-predicted structures (middle), and averaged AF-predicted structures (⟨AF⟩, bottom, see Sec. 6).

are not appropriate for our purpose, and discuss the pros and cons of these alongside several other measures.

Historically, in the field of protein structure prediction, the focus has been on measuring similarity instead of differences, and finding a well-behaved metric that can score similarity so that algorithm performance can be easily evaluated (Kufareva and Abagyan, 2012).

The most commonly used metric, although problematic, is the Root Mean Squared-Deviation (RMSD), which is the root-mean-square-deviation of atomic positions between a target structure (with positions $\mathbf{r}'_i$) and a reference structure (positions $\mathbf{r}_i$) (Kufareva and Abagyan, 2012),

$$\mathrm{RMSD} = \frac{1}{L}\sqrt{\sum_{i=1}^{L} \left|\mathbf{r}'_i - \mathbf{r}_i\right|^2}\,, \tag{1}$$

where L is the sequence length. The first step in calculating RMSD is to align the two structures via translation and rotation using the Kabsch algorithm (Kabsch, 1976). This is problematic since parts of proteins can undergo rigid-body motion, in which there is little local deformation yet the global positions undergo large-scale rearrangements. Therefore RMSD can be high in proteins that barely deform. Additionally, RMSD is an absolute metric, which results in sensitivity to large deviations due to outliers (*e.g.*, in flexible loops and tails).

RMSD is still widely used due to its simplicity, but more robust global metrics have been developed – such as the template modelling score (TM-score) (Zhang and Skolnick, 2004) and the global distance test (GDT) (Zemla, 2003) – which align subsets of atoms rather than the whole protein, and produce scores between 0-1 so that effects of outliers are minimized. More recently, the local distance different test (LDDT) was developed (Mariani *et al.*, 2013). LDDT is a score per residue, which compares neighbor distances in a target structure and a reference structure, and measures the fraction of corresponding distances that are within some threshold value of each other.

Our goal is to measure local deformation, which requires a completely different type of metric. First, we need a metric capable of distinguishing between deformation at different residues, since we expect mutation effects to be primarily (but not necessarily) local. This means global metrics like RMSD, TM-score and GDT are unsuitable. Second, we want an absolute metric so that large deformation is taken for what it is, and not subject to diminishing returns. This implies that LDDT is not suitable, which we show in Figures 1, 2, 3 and 4.

In the following, we consider several metrics of local deformation that all share similar aspects, but differ in their origins: the local distance difference (LDD) and neighborhood distance (σ_i) are mathematically related to LDDT, but result in absolute metrics rather than scores; we investigate three measures of strain (shear strain, non-affine strain, and effective strain (ES)) based on finite strain theory (Lubliner, 2008). Note that we exclude AF-predicted residues with pLDDT < 70 from calculations, as we treat these as disordered residues which would always lead to high deformation.

A. Local Distance Difference Test (LDDT)

LDDT is a score from 0-1 that measures the similarity between two structures, with a value of 1 being maximally similar. LDDT_i is calculated for each residue i by first defining a set N_i of $n_i = |N_i|$ neighbors, $j \in N_i$, using a distance cutoff, γ. Residues are considered neighbors if the

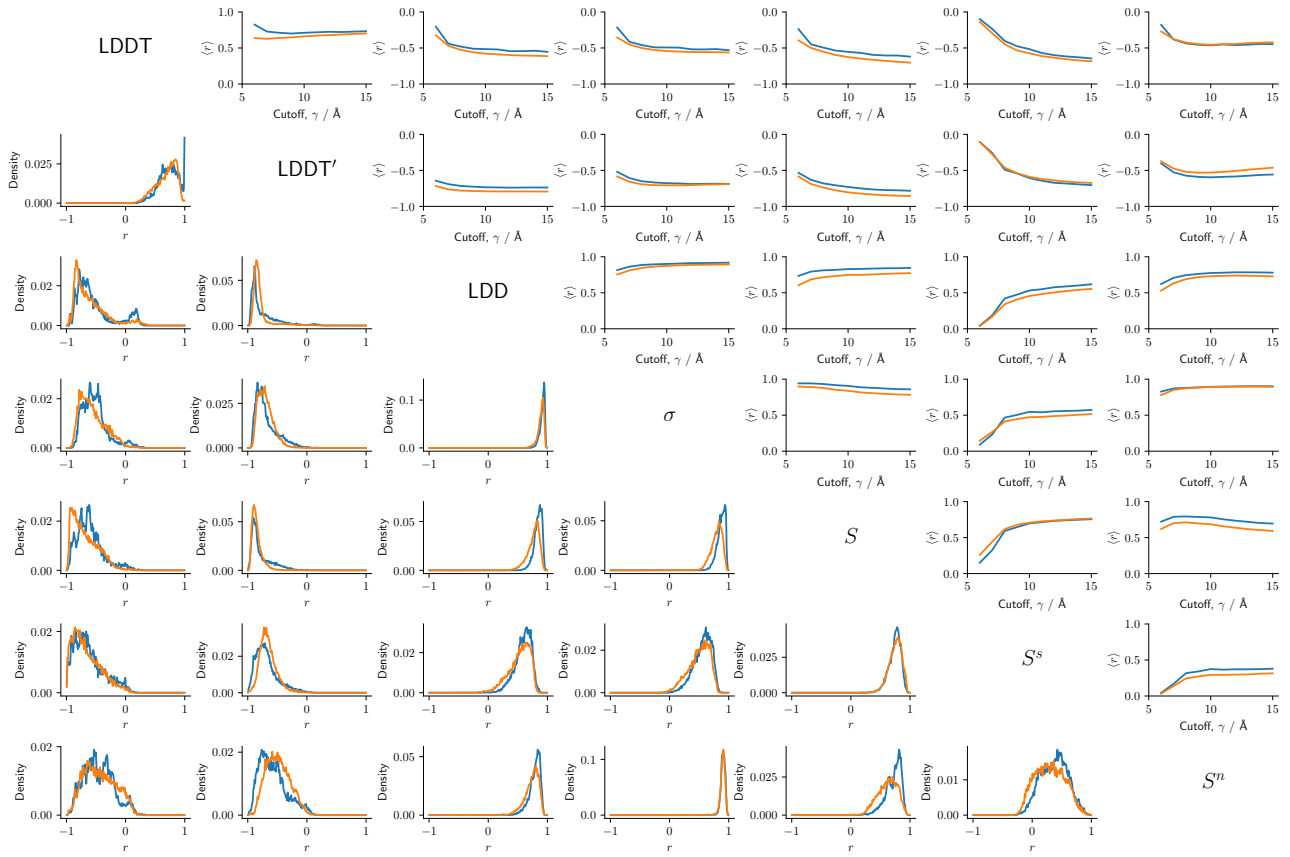

FIG. 2 **Correlations between deformation metrics.** Pearson's correlation coefficient, r, is calculated for each pair of proteins in our PDB dataset (17,813 pairs). Distribution of r for each combination of metrics (Bottom, $\gamma = 13$ Å), and mean correlation $\langle r \rangle$ as a function of γ (Top). Data is shown for both PDB (orange) and AF-predicted (blue) structures.

positions of their C_α atoms, $\mathbf{r}_j$, are closer than γ; *i.e.*, if $r_{ij} = |\mathbf{r}_{ij}| = |\mathbf{r}_i - \mathbf{r}_j| \leq \gamma$. For each neighbor, we calculate the distances between neighbors in both the reference r_{ij} and the target structures r'_{ij}, and calculate the difference $\Delta r_{ij} = r'_{ij} - r_{ij}$. LDDT_i is defined as the fraction of distance differences that are within a set of k cutoffs ζ_k,

$$\mathrm{LDDT}_i = \frac{1}{4n_i} \sum_{j \in N_i} \sum_k \theta(\Delta r_{ij}, \zeta_k) \,, \tag{2}$$

where θ is the Heaviside step function,

$$\theta(\Delta r_{ij}, \zeta_k) = \begin{cases} 0, & \Delta r_{ij} > \zeta_k \\ 1, & \Delta r_{ij} \leq \zeta_k \,. \end{cases}$$

LDDT is typically calculated with $\zeta_k \in \{0.5, 1, 2, 4\}$ Å. It is not possible to discriminate between small amounts of deformation with these cutoffs (since small deformation always gives $\mathrm{LDDT}_i = 1$; Fig. 1), so we also measure a more precise alternative, LDDT′, using $\zeta_k \in \{0.125, 0.25, 0.5, 1\}$ Å.

B. Local Distance Difference (LDD)

As an alternative to LDDT, we propose to skip the step where distances are compared with cutoff values, and instead directly compare neighbor distances in reference and target structures. For each residue i we calculate the local distances for each of its n_i neighbors, $r_{ij} = |\mathbf{r}_{ij}| = |\mathbf{r}_i - \mathbf{r}_j|$, and the change in the distances between the structures $\Delta r_{ij} = r'_{ij} - r_{ij}$. Then, LDD_i is the sum of the squared distance change,

$$\mathrm{LDD}_i = \sqrt{\sum_{j \in N_i} (\Delta r_{ij})^2} \,. \tag{3}$$

One clear benefit of LDD is that does not require an alignment since it measures differences in scalar distances, yet this also means that it neglects deformation due to differences in rotations between neighbors. Beyond this, it is conceptually simple, easy to measure, and similar to LDDT.

C. Neighborhood Distance

LDDT and LDD depend only on the magnitude of the distances $\mathbf{r}_{ij}$, not their orientation. For a more robust measure of deformation that takes into account deformation due to rotation of distance vectors as well as changes in their magnitude, we propose the *neighborhood distance*. For each residue i we define a $n_i \times 3$ neighborhood tensor, $\mathbf{D}_i$, as the tensor of the C_α distance vectors $\mathbf{r}_{ij}$ for all $j \in N_i$, where rows correspond to neighbors, and columns correspond to the distances in three dimensions, x, y, z. We get neighborhood tensors for both the reference structure, $\mathbf{D}_i$, and the target structure, $\mathbf{D}'_i$. Then, the neighborhood distance is the norm of the change in distances, $\Delta \mathbf{r}_{ij} = \mathbf{r}'_{ij} - \mathbf{r}_{ij}$, between neighborhoods.

$$\sigma_i = \left\| \mathbf{D}_i - \mathbf{D}'_i \right\| = \sqrt{\sum_{j \in N_i} \left| \Delta \mathbf{r}_{ij} \right|^2} \,. \tag{4}$$

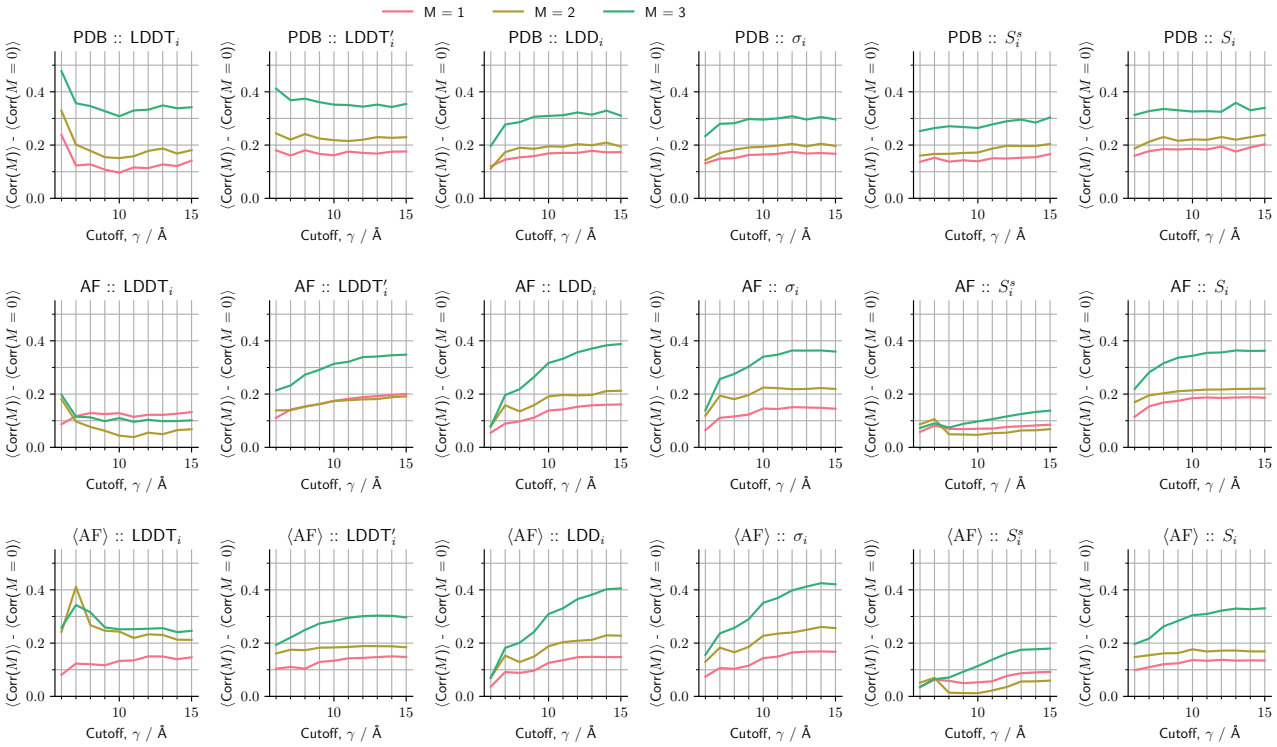

FIG. 3 **Correlations between PDB and PDB (top), AF (middle), and ⟨AF⟩ deformation vectors for different metrics.** Residual correlation due to mutation, $\langle \mathrm{corr}(M)\rangle - \langle \mathrm{corr}(M=0)\rangle$, extracted from the distributions of correlation between PDB and AF-predicted deformation vectors, $\mathrm{corr}(M)$. Residual correlation is shown as a function of neighbor cutoff γ. Results are shown for 6 metrics, from left to right: LDDT; LDDT′ with more precise cutoffs ζ_k; LDD; neighborhood distance, σ_i; shear strain, S^{s}; effective strain (ES), S_i. For non-averaged structures (AF) we report the model generated by AlphaFold that has the highest pLDDT. We construct the average structures (⟨AF⟩) using all 5 AlphaFold models, along with 25 ColabFold-generated structures (5 models, 5 repeats).

This method requires that the target and reference neighborhoods are aligned. We achieve this by rotating (without translating) the target neighborhood using the Kabsch algorithm (Kabsch, 1976).

We note that this metric is similar to the frame-aligned-point-error (FAPE) – which forms part of the loss function used by DeepMind to train AlphaFold (Jumper *et al.*, 2021). σ_i and FAPE differ by the normalization factor, n_i, and in the frame of reference used to align the two neighborhood tensors; in AlphaFold, the reference frame is defined by the positions of the three heavy atoms (N, C_α, and C) in the backbone.

D. Shear Strain

We follow the standard treatment of finite-strain theory as implemented in (Eckmann *et al.*, 2019). To calculate strain at residue i, we first calculate the deformation gradient tensor, $\mathbf{F}_i$, where

$$\mathbf{D}'_i = \mathbf{F}_i\,\mathbf{D}_i\ . \tag{5}$$

The matrix equation (5) is usually overdetermined, so one finds $\mathbf{F}_i$ using least-squares regression over the residual $\|\mathbf{D}'_i - \mathbf{F}_i\,\mathbf{D}_i\|^2$. We then get the Lagrangian finite strain tensor, $\mathbf{E}_i$,

$$\mathbf{E}_i = \frac{1}{2}(\mathbf{F}_i^{\mathsf{T}}\mathbf{F}_i - \mathbf{I})\ ,$$

where $\mathbf{I}$ is the identity matrix. Finally, as a measure of the magnitude of the shear strain, we use

$$S_i^{\mathrm{s}} = \mathrm{Tr}(\mathbf{E}_i\cdot\mathbf{E}_i) - \frac{1}{3}[\mathrm{Tr}(\mathbf{E}_i)]^2\ . \tag{6}$$

E. Non-Affine Strain

The non-affine strain is the deformation that is not due to any of the affine transformations – isotropic volume expansion/contraction, or shear/twist motion. A simple estimate of the non-affine component of deformation is the residual left after solving Eq. (5) for the affine deformation gradient tensor, $\mathbf{F}_i$,

$$S_i^{\mathrm{n}} = \left\|\mathbf{D}'_i - \mathbf{F}_i\,\mathbf{D}_i\right\|^2 = \sum_{j\in N_i}\left|\mathbf{r}'_{ij} - \mathbf{F}_i\,\mathbf{r}_{ij}\right|^2\ , \tag{7}$$

where j is the neighbor index. S_i^{n} has been previously used as an effective measure of the non-affine strain (Falk and Langer, 1998). We use a modified version of the "atomic_strain.py" implementation from (Kermode and Pastewka, 2023).

F. Effective Strain (ES)

For a more comprehensive measure of strain that does not distinguish between isotropic, shear, and non-affine components of strain, we measure the effective strain (ES); this is simply the *average relative change in positions*,

$$S_i = \left\langle \frac{|\Delta\mathbf{r}_{ij}|}{r_{ij}} \right\rangle = \frac{1}{n_i}\sum_{j\in N_i}\frac{\left|\mathbf{r}_{ij} - \mathbf{r}'_{ij}\right|}{r_{ij}}\ . \tag{8}$$

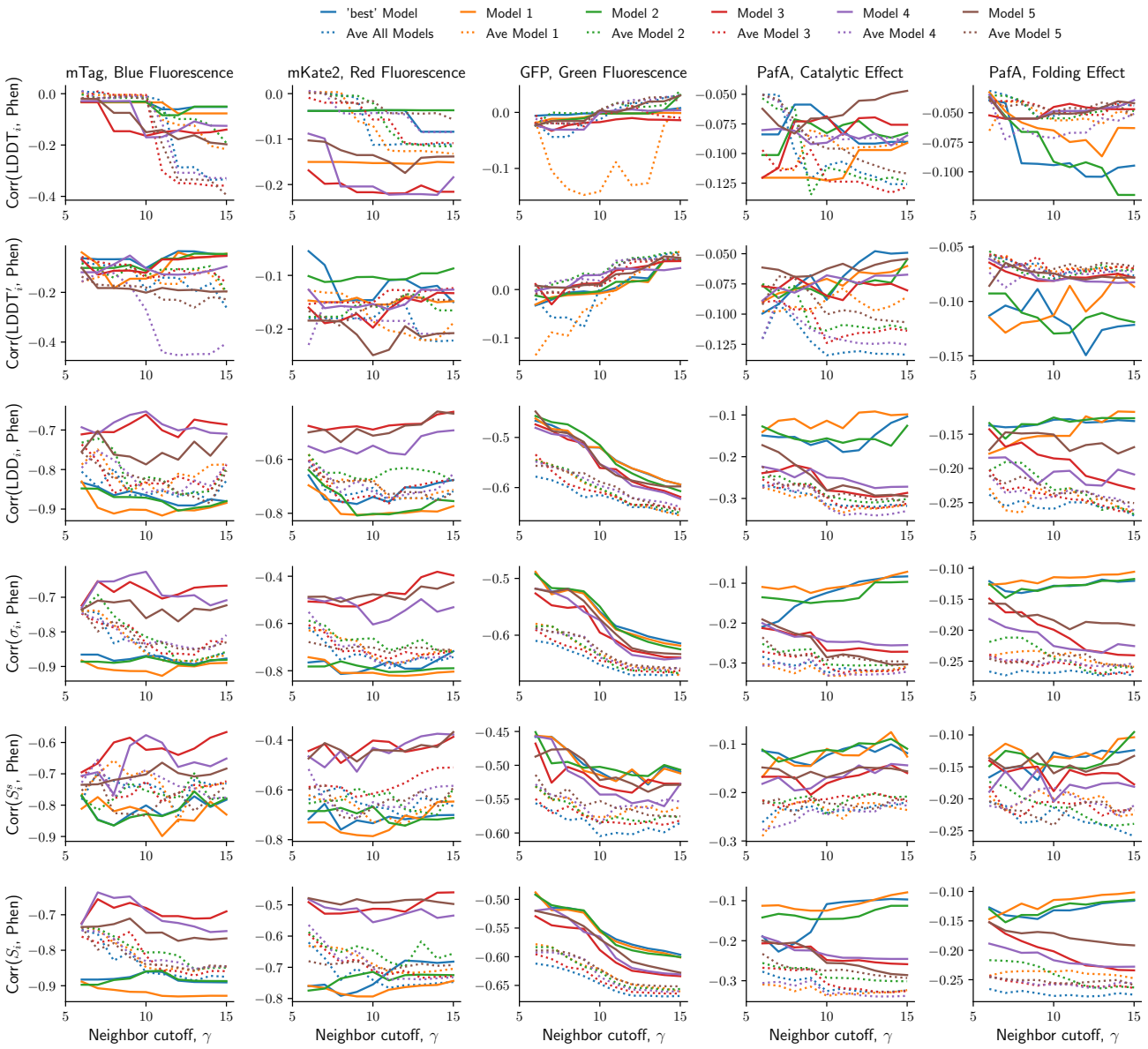

FIG. 4 **Correlations between deformation metrics and phenotype measurements.** Correlations between deformation at residue i and phenotype for 5 sets of phenotype measurements (left to right: mTag, blue fluorescence; mTag, red fluorescence; GFP, green fluorescence; PafA, catalytic effect; PafA, folding effect), for 6 deformation metrics (from top to bottom: LDDT; LDDT$'$ with more precise cutoffs ζ_k; LDD; neighborhood distance, σ_i; shear strain, S^s; effective strain (ES), S_i), for non-averaged and averaged structures, for all AF models. For each case (metric, phenotype, γ, model), we choose the residue i that gives the highest correlation as a simple statistic to compare performance, and report the correlation between deformation at that specific residue with phenotype. The 'best' model is the one with the lowest pLDDT. When averaging across structures, "Ave All Models" refers to averages using the entire set of structures, while the others only averaged over one out of the five AF models; an exception is PafA, where we excluded the AlphaFold-predicted models 1 and 2, since we found that the use of a template in these cases resulted in odd predictions; we included the ColabFold-predicted models 1 and 2, since we disabled the use of templates in these cases.

The ES, S_i, like the neighborhood distance, σ_i, requires that neighborhood tensors are first aligned via rotation, and takes into account rotation-based deformation between neighbors. Additionally, S_i is weighted by distances of neighbors from residue i, so it is a dimensionless property that approximates the local strain. Since far-away residues contribute less to the total S_i, this measure should be more robust to changes in the neighbor cutoff γ.

5. Evaluation of deformation metrics

A. Correlations between deformation metrics

To probe the similarities and differences in the deformation metrics, we measured each metric, using cutoff values of $\gamma \in \{6, 7, ..., 16\}$ Å, for both PDB and AF-predicted pairs of structures in our set of 17,813 matched pairs. For each pair of structures, we calculate Pearson's correlation coefficient, r, between pairs of metrics, and report the distribution of r, and the average $\langle r \rangle$ as a function of γ (Fig. 2).

We find that LDDT has the lowest correlation with all

of the other metrics, compared to the other metrics. We attribute this mainly due to the low resolution of LDDT (at best 0.5 Å) compared to the average difference in backbone distances (PDB, 0.15 Å; AF, 0.07 Å, Fig. 18). Accordingly, we find that using lower LDDT cutoffs ζ_k (LDDT′) results in higher correlations with other metrics.

Importantly, we find that *the LDD, σ_i, and effective strain (ES), S_i, are all extremely well correlated*, as may be expected given they are mathematically similar. Out of the metrics related to strain, we see that the ES and non-affine strain are highly correlated, while shear strain has a much lower correlation with all of the metrics. This implies that the deformation in proteins, whether due to dynamical fluctuations ($M = 0$) or due to mutations ($M > 0$), has a significant non-affine component. This may not be true for large-scale deformation due to functional motion, such as that which occurs in enzymes and motor proteins.

B. PDB-AF correlations for different deformation metrics

In order to compare the performance of different deformation metrics for measuring mutation effects, we compare three sets of data: For PDB-PDB comparisons we identify groups of three or four PDB structures whose sequences differ by $M \in \{0,1,2,3\}$ mutations. For example, for $M = 0$, we find three or four structures with identical sequences and calculate deformation vectors between the metrics computed for two of the pairs, and calculate Pearson's correlation coefficient, r. If there are only three available structures, then one structure is shared between the two pairs, and each pair has one unique structure; if there are four available structures, then all structures are unique. We also compare sets of PDB-AF, and PDB-$\langle$AF$\rangle$ (averaged AF-predicted structures; Sec. 6) structures.

For example, for $M = 1$ we take two sequences that differ by one mutation, for which there are corresponding structures in the PDB, and calculate a deformation vector; we use AF to predict structures for the two sequences and calculate a deformation vector; we then calculate r between these two vectors. For average AF-predicted structures, we follow the procedure in Sec. 6 to get a pair of averaged structures, calculate the deformation vector between these, and then calculate r with respect to the PDB deformation vector. To calculate the residual correlation, $\langle \text{corr}(M) \rangle - \langle \text{corr}(M = 0) \rangle$, we calculate the average correlation for each $M \in \{1,2,3\}$, and subtract this from the average correlation for all $M = 0$. In this way, we estimate the average proportion of the correlations that is due to mutations, rather than correlated fluctuations. Correlations are calculated between LDDT vectors on a linear scale, since LDDT exhibits little variation. For all other metrics we calculate correlations on a log scale, since they can vary over several orders of magnitude.

In Fig. 3, we show the residual correlation as a function of γ for all deformation metrics, for the three sets of data (PDB-PDB, PDB-AF, PDB-$\langle$AF$\rangle$). The highest correlations are found in the region of $10 \leq \gamma \leq 15$ Å, for LDD, σ_i, and S_i. **Spurious** high correlations are found for LDDT and LDDT′: most values are LDDT $= 1$, and for over half of proteins all residues have LDDT $= 1$; the high correlations reported in Fig. 3 for LDDT are due to relatively few values that are LDDT < 1.

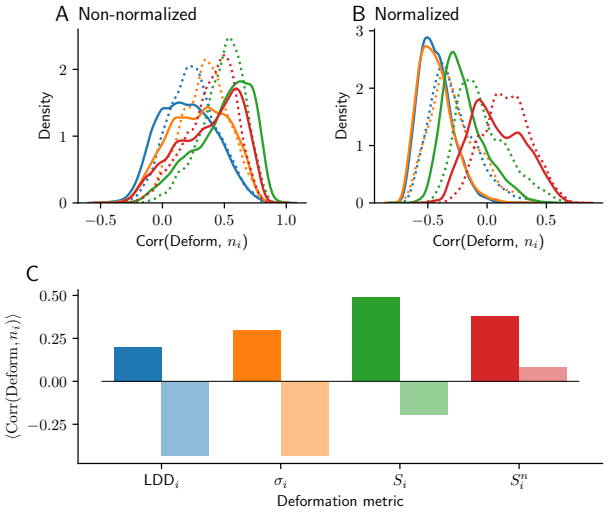

FIG. 5 **Effect of normalization.** A-B: Distribution of correlations between deformation (LDD, blue; neighborhood distance, orange; ES, green; non-affine strain, red) and number of neighbors n_i, over all matched pairs of proteins ($0 \leq M \leq 3$) for PDB (solid line) and AF-predicted structures (dotted line). Distributions are shown for non-normalized metrics (A) and normalized metrics (B). C: Average correlation for different metrics. Results are shown for non-normalized metrics in bold and for normalized metrics in pastel color.

C. Deformation-Phenotype correlations for different metrics

In order to compare the performance of different deformation metrics for predicting the phenotypic effects of mutations, we computed correlations between phenotypic measurements and deformation at residue i, separately for each i (Sec. 2, Sec. 9). In Fig. 4 we compare performance for: all 5 phenotypes, 6 deformation metrics, different AF models, and either non-averaged or averaged structures; we show results for deformation at whichever residue i has the highest correlation with phenotype. With regards to the different deformation metrics, we find that *the metrics that show the highest correlations between PDB-AF structures (LDD, σ_i, and S_i), also show the highest deformation-phenotype correlations.* LDDT results in especially poor correlations with phenotype, and using the more precise distance cutoffs ζ_k (LDDT′) does not help. Shear strain S_i^s is only slightly worse than effective strain S_i.

D. Normalizing by number of neighbors

In our definitions of LDD$_i$, σ_i and S_i, there is the choice to normalize by the number of neighbors, n_i, which gives us the average deformation per neighbor. There is no inherently correct choice, as the two types of metric (normalized or non-normalized) simply report different information: Normalizing results in a metric that describes mean deformation, while not normalizing results in a metric that also takes into account the number of neighbors. Thus, one might suspect that non-normalized metrics are correlated with the number of neighbors. We find this to be true, however we also find that normalized metrics are anti-correlated with the number of neighbors (Fig. 5). We attribute this to the increased flexibility and variance across repeat predictions that we observe in surface residues that have fewer neighbors.

It seems that due to these opposing effects – few neighbors leads to higher flexibility and higher mean deformation, and many neighbors leads to higher overall deformation due to summing over many neighbors – there is no clear overall effect of normalization. Importantly, we find that there is no effect of normalization on deformation-phenotype correlations. For PDB-AF deformation correlations, we find that normalizing results in higher correlations when the deformation metric uses the L1 norm (*e.g.*, S_i) instead of the L2 norm. We find that not normalizing results in higher correlations when the deformation metric uses the L2 norm instead of the L1 norm. The normalized and un-normalized versions of each metric also correlate highly with each other. Ultimately, we find that the results are not particularly sensitive to the exact form of the deformation metric, and exclusively report ES in the main manuscript.

6. Averaging AF-predicted structures

We observed that deformation between repeat measurements (structures with same sequence, $M = 0$), whether PDB or AF, tends to depend on local flexibility (main paper, App. 2B). Thus, we consider that it may be possible to smooth out these structural fluctuations by averaging over many AF predictions. We take into account the fact that proteins can undergo large-scale, rigid-body transformations, and thus do not average over the entire *global* protein structure. This would almost certainly lead to extreme, unphysical average configurations. Instead, we average over *local* neighborhoods $\mathbf{D}_i$. For each residue i, we extract the local neighborhoods per repeat prediction k, $\mathbf{D}_i^k$, using a distance cutoff γ. We choose one neighborhood k_0 as a reference neighborhood. We then remove neighbors that are not contained in all k neighborhoods, such that the neighborhood is defined by the neighbors that are within a distance of γ from residue i in all predicted structures k. We then rotate all neighborhoods $k \neq k_0$ with respect to neighborhood k_0 using the Kabsch algrithm (Kabsch, 1976), and average over neighborhood k_0 and the rotated neighborhoods $k \neq k_0$ (a total of n_k neighborhoods) to get an average neighborhood $\langle \mathbf{D}_i \rangle$,

$$\langle \mathbf{D}_i \rangle = \frac{1}{n_k} \sum_k \mathbf{D}_i^{k*} \, ,$$

where $\mathbf{D}_i^{k*}$ is the neighborhood of residue i from repeat prediction k after rotation. We refer to these averaged structures (sets of $\langle \mathbf{D}_i \rangle$) using the notation $\langle \mathrm{AF} \rangle$.

We note that averaging in this way can lead to odd configurations, that can only be accessed through bond angles that are unphysical. Thus, these averaged configurations should not be used to study structure. Instead, this is a viable approach to studying differences in structure, as the effect of averaging is to get a better estimate of a summary statistic (the 'average structure'), which can be used to measure small differences.

To illustrate how averaging can reduce the strain that arises from fluctuations rather than mutations (also see Fig. 1), we show deformation against sequence position for two examples (PafA WT vs Y174V, and mTag WT vs I175L; Fig. 6A-D). In Fig. 6A, there are many high-strain regions that are far away from the mutated site when using the non-averaged structures (Fig. 6B). Averaging over neighborhoods reduces the strain in these areas and reveals that strain is much more dependent on the distance from the mutated site (Fig. 6B). To show how this effect generalizes, we calculate the correlation between strain and distance from the nearest mutated site for all mutated protein pairs in the PDB dataset (Fig. 6I), and find that averaging tends to increase this correlation. This indicates that *averaging is able to smooth out the effects of non-mutation fluctuations*.

We also note that it is possible for averaging to result in increased strain, as can be seen for residue 158 in Fig. 6C. These cases can arise due to the fact that conformations can exist in discrete populations (*e.g.*, rotamers). To illustrate this, we investigate the neighborhood of residue 158 in more detail, by comparing neighborhoods from different predicted structures. We first rotate all neighborhoods $\mathbf{D}_{158}^k$ to an arbitrary reference neighborhood k_0. This allows us to examine the variance of individual components of the neighborhoods, and how they differ across predictions. We see that the x, y and z components of $\mathbf{D}_{158}^k$ corresponding to neighbor $j = 15$ and different repeat predictions k exhibit bimodal distributions (Fig. 6G). By looking at the distribution of the variances of each component (each neighbor j, and each spatial dimension), we see that this is quite common – the rightmost peak in Fig. 6H corresponds to components with bimodal distributions. This stands in contrast to similar results for the neighborhood of residue 79, where there are some bimodal distributions amongst the components of the neighborhood tensor (Fig. 6E), but there are far fewer of these overall (Fig. 6F). We have thus learned that due to the presence of discrete local conformations, averaging can fail to reduce the strain across structures if too few repeats are used, but in general this is a promising strategy to achieve more precise predictions of mutation effects.

Since some proteins (and some parts of proteins) are more flexible than others, it is useful to have an estimate of how many structures one would need to average over to smooth out the non-mutation fluctuations. AlphaFold's predicted confidence score, pLDDT, is a suitable candidate for this. Using the above examples (Fig. 6A,C), we compare pLDDT with the change in deformation after averaging, $S_i^{\mathrm{AF}} - S_i^{\langle \mathrm{AF} \rangle}$ (Fig. 6J,K). We find that pLDDT is only correlated with the change in deformation for one out of two cases. We suspect that this is due to pLDDT being a low-resolution metric, since most residues have similar pLDDT scores. An alternate approach is to predict many repeat structures, and to calculate the variance between n_k repeat measurements of deformation, $\langle S_i \rangle$,

$$\langle S_i \rangle = \frac{1}{n_k} \sum_k S_i^k \, .$$

This is a more direct measurement of the variance in repeat predictions, and it is a much better predictor of the change in deformation due to averaging. The repeat-prediction variance can thus be used to guide decisions about how many structures one should average over to get a good estimate of mutation effects. If the repeat-prediction variance is low, then few repeats are sufficient; if the repeat-

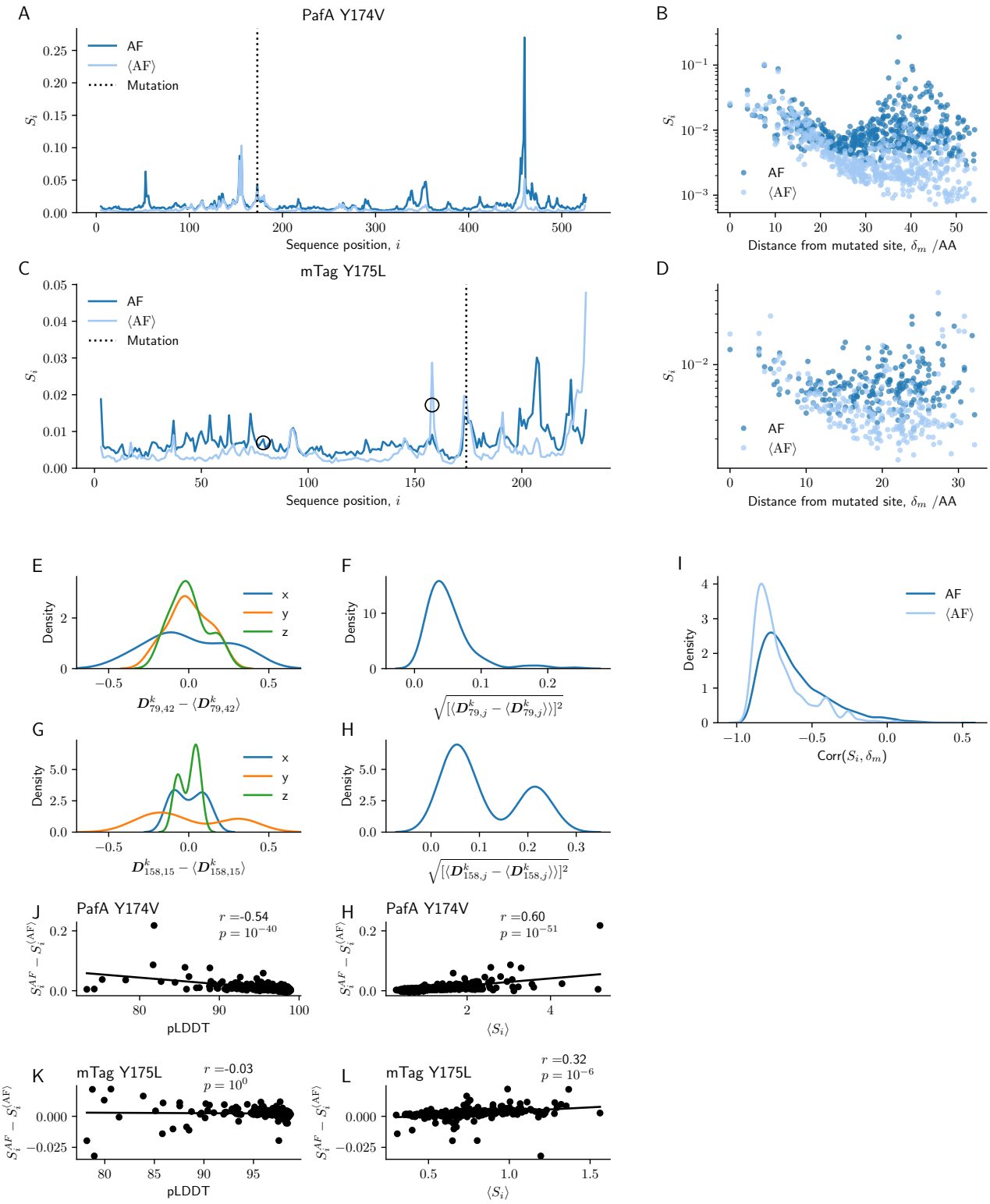

FIG. 6 **Averaging across multiple AF predictions.** A-D: Examples of strain per residue S_i for PafA (A: WT vs Y174V) and mTag (C: WT vs I175L) calculated using single AF structures (AF) and averaged structures ($\langle$AF$\rangle$). Dotted line indicates the mutated residue; circles indicate residues 79 and 158 (C, mTag). Effective strain (ES) per residue against distance from the mutated site, $\delta_{m,i}$, for each residue i for PafA (B) and mTag (D). E-H: Distributions of x, y and z components of the (neighborhood-aligned) vectors $\mathbf{D}^k_{79,42}$ (E) and $\mathbf{D}^k_{158,15}$ (G) across all repeat predictions k of the mTag WT and I175L, normalized by the average $\langle \mathbf{D}^k_{ij}\rangle$ across k. Distribution of standard deviations of elements ($\delta x_{i,j}$, $\delta y_{i,j}$ and $\delta z_{i,j}$) in aligned neighborhood tensors across all repeat predictions of the mTag WT and I175L for residues 79 (F) and 158 (H). I: Distribution of correlations between strain per residue S_i and distance from the nearest mutated site $\delta_{m,i}$ (I) for all pairs of AF-predicted structures ('best' model) in our PDB dataset that are mutated ($M > 0$). J-L: Difference in deformation between AF and $\langle$AF$\rangle$ structures as a function of pLDDT (J, K) and repeat-prediction variability $\langle S_i\rangle$ (H, L), for PafA WT vs variant Y174V (J,H) and mTag WT vs variant Y175L (K, L).

prediction variance is very high, then perhaps AlphaFold will not be useful at predicting mutation effects in that region.

A. Effect of averaging on PDB-AF correlations

When we compare the PDB-AF correlations with PDB-$\langle$AF$\rangle$ correlations (Fig. 3), we typically no difference for the best deformation metrics (LDD, σ_i, and S_e). We attribute this to the high levels of repeat-measurement deformation in PDB structures; increasing precision of AF predictions cannot increase correlations since they are limited by PDB imprecision.

B. Effect of averaging on AF-phenotype correlations

To understand whether averaging structures can produce more accurate predictions, we also look at the effect of averaging on deformation-phenotype correlations. In Fig. 4 we compare performance for: all 5 phenotypes, 6 deformation metrics, different AF models, and either non-averaged or averaged structures. In the majority of cases, averaging results in correlations that are stronger by about $|\Delta r| = 0.05 - 0.1$. Excluding the results for LDDT (which performs poorly in general), averaging across all models tends to produce the best results ("Ave All Models", blue dotted line), although there is often little difference between the results obtained for averaging over specific models (other dotted lines). In comparison, for non-averaged structures, we see much higher variability in performance depending on the AF model used. This suggests that averaging across many models is a useful way to measure mutation effects, and can be used instead of simply choosing the model with the highest pLDDT.

The results for the protein mTag (blue and red fluorescence) are an exception. They show that for two models, 1 and 2 (in this case the 'best' model almost always corresponds to either of these two models), the non-averaged structures perform better than any of the averaged structures. We investigated why this may be the case, and found that it is due to differences in the variability of predictions using DeepMind's AlphaFold versus ColabFold. We calculate the average effective strain (ES) per pair of WT and mutant structures for each high-throughput dataset (mTag, GFP, PafA), for both AlphaFold implementations,

$$\langle \mathbf{S} \rangle = \frac{1}{L'} \sum_i S_i \, ,$$

where L' is the length of the protein, minus the number of disordered residues (*i.e.*, where pLDDT < 70). We then calculate the difference in the average ES between implementations, and show the distribution for each AF model in Fig. 7. In the majority of cases, the DeepMind implementation produced lower average strain. The major exception to this is that DeepMind models 1 and 2 produced extremely high strain for PafA, which we consider an anomalous result that arises due to the use of structural templates in prediction (Sec. 12). We also see that DeepMind models 3, 4 and 5 occasionally produced high strain for mTag variants.

We then compared the average differences between implementations, ColabFold and DeepMind, $\langle\langle \mathbf{S}^{\mathrm{CF}} \rangle - \langle \mathbf{S}^{\mathrm{DM}} \rangle\rangle$, for each model, to the differences between strain-phenotype correlations calculated using non-averaged (AF) and averaged ($\langle$AF$\rangle$) structures. We find a significant correlation between these differences (Fig. 8), which suggests that the cases where averaging produced worse correlations (mKate2 and mTagBFP2, models 1 and 2) were due to the higher variability of predictions using ColabFold (since ColabFold structures constitute the majority of the structures used in averaging). We conclude that, in general, averaging structures appears to produce better results, but it depends on the quality of the prediction. When running ColabFold we used 6 recycles per structure, but this might be insufficient for mTag.

7. Range of mutation effects

We calculate the average range of mutation effects by looking at deformation as a function of distance from the nearest mutated site, δ_m. We calculated the average ES, $\langle \mathbf{S}_i^p \rangle$, across all mutated ($M > 0$) pairs of proteins p, and all residues i within δ_m bins of size 2 Å (Fig. 9). When we consider the full set of PDB pairs, we find that deformation reaches a plateau at about 14 Å; for AF pairs, the range appears to be 16 Å, and for $\langle$AF$\rangle$ pairs the range is about 18 Å. Note that while we show allosteric effects in Fig. 9 up to ranges of almost 2 nm, these represent average ranges of mutation effects; this does not preclude the possibility that there are much longer-range allosteric effects due to mutations.

8. When do AF predictions correlate with PDB data?

To understand why some PDB-AF correlations are high and others low, we looked at three *factors*: (i) protein flexibility, (ii) secondary structure, and (iii) effect size of the mutation. For each type of factor, we studied multiple scalar *properties*, listed below. For each property, we measured Pearson's correlation coefficient, r, between the property and the corresponding PDB-AF correlation for a set of protein pairs. We used the non-redundant set, and resampled 1,000 times to get distributions of r to account for sampling variance.

- Protein flexibility
  - **B-factor:** B-factors are obtained from X-ray scattering experiments for each atom. The B-factor is related to the uncertainty in the position of an atom's coordinates, and is known to correlate with local flexbility (Sun *et al.*, 2019).
  - **pLDDT:** pLDDT is predicted by AlphaFold, and is supposed to predict the uncertainty of the structure predictions of individual residues, much like how LDDT scores the accuracy of structure predictions of individual residues. Recent studies have indiciated that it negatively correlates with local flexibility (Guo *et al.*, 2022; Ma *et al.*, 2023).
  - **RSA:** Relative solvent accessibility (RSA) of a protein residue is the amount of an amino acid's surface that is accessible to solvent, compared to the amino acid's total surface area. Residues with higher RSA are less

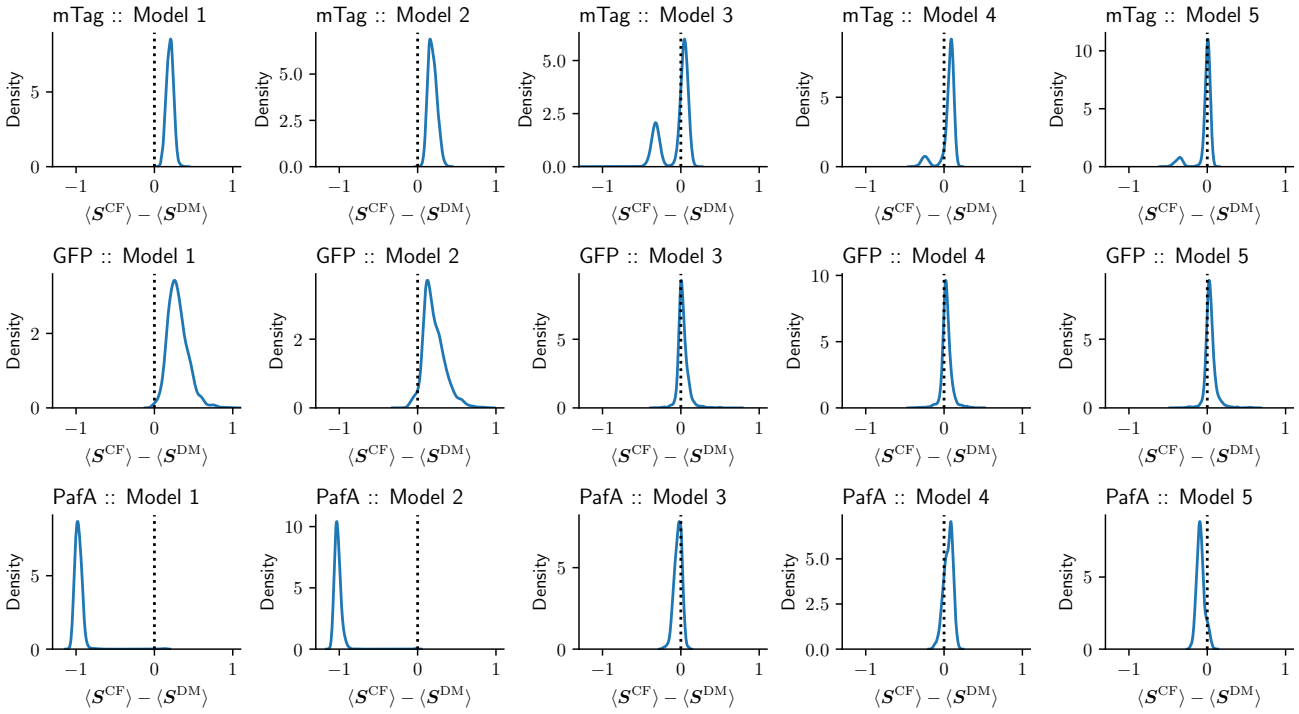

FIG. 7 **Comparing DeepMind and ColabFold implementations of AF.** Difference in the average effective strain (ES) (with $\gamma = 13$ Å) per pair of WT and mutant structure, $\langle S_i \rangle$, for structures predicted using both DeepMind, $\langle S_i^{\mathrm{DM}} \rangle$, and ColabFold, $\langle S_i^{\mathrm{CF}} \rangle$, implementations. Results are shown for each high-throughput dataset (mTag, GFP, PafA), and for each AF model.

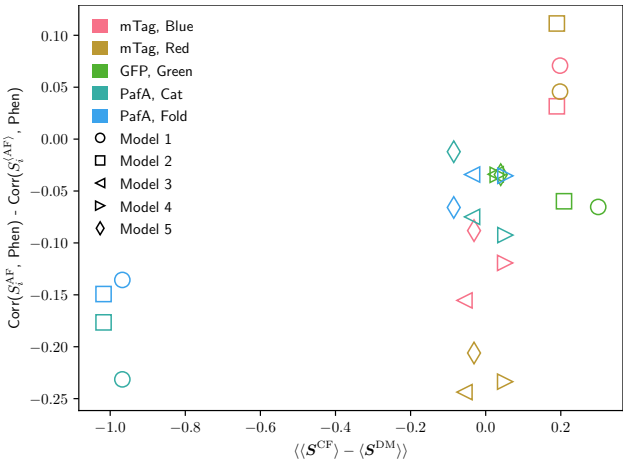

FIG. 8 **Differences between AF implementations correlate with differences in AF vs ⟨AF⟩ accuracy in predicting phenotype.** Average difference in the average strain per pair of WT and mutant structure, $\langle S_e \rangle$, for DeepMind (DM) and ColabFold (CF) implementations, against the difference in correlation between effective strain (ES) (with $\gamma = 13$ Å) and phenotype (blue fluorescence, red fluorescence, green fluorescence, catalytic effect, folding effect), for all AF models. Pearsons's correlation coefficient is calculated for the full 25 points ($r = 0.53$, $p = 0.006$), and also for the subset of points excluding the four points on the left ($r = 0.51$, $p = 0.019$).

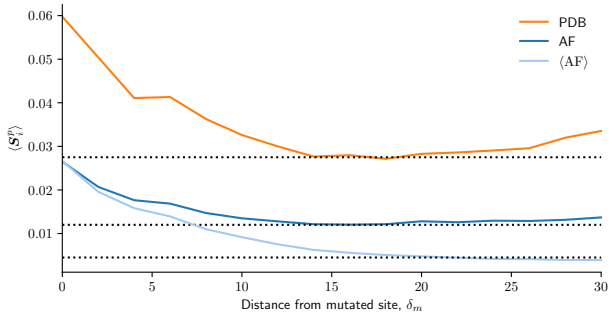

FIG. 9 **Average range of mutation effects.** Average deformation, $\langle \mathbf{S}_i^p \rangle$, across all protein pairs p and all residues i within a certain distance from the nearest mutated site δ_m, as a function of δ_m, calculated respectively using PDB, AF and ⟨AF⟩ pairs. Bins of size 2 Å were used. Dotted lines are visual guides.

sterically constrained by other residues, and thus tend to be more flexible (Zhang *et al.*, 2009).

– **Prediction Variance:** We measure the mean deformation per residue by comparing multiple predictions of the same protein sequence, $\langle S_i \rangle$ (Eq.(6)). AlphaFold is a stochastic algorithm, which has been shown to sample some of the conformational diversity of real proteins (del Alamo *et al.*, 2022; Saldaño *et al.*, 2022). Thus, we expect that the degree of deformation per residue across repeat predictions should be correlated with the degree of conformational flexibility.

• Secondary Structure

– $\alpha-$**Helix:** We calculate the fraction of residues in α-helices per protein (Kabsch and Sander, 1983).

– $\beta-$**Strand:** We calculate the fraction of residues in β-strands per protein (Kabsch and Sander, 1983).

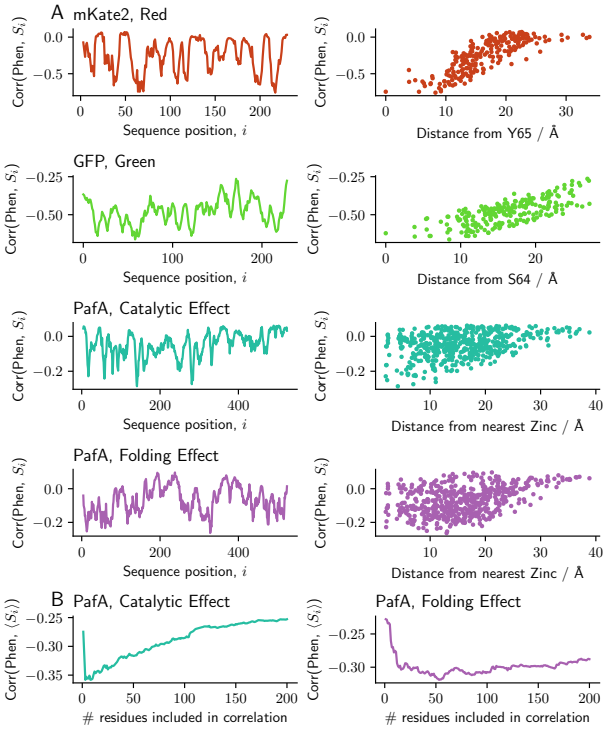

FIG. 10 A: (Left) Correlation between deformation at residue i and phenotype, for all sequence positions i: mKate2, red fluorescence; GFP, green fluorescence; PafA, catalytic effect; PafA, folding effect. (Right) Correlation between deformation at residue i and phenotype as a function of distance from functional residues / co-factors: mKate2, residue Y65; GFP, residue S64; PafA, distance from the nearest Zinc co-factor. B: Correlation between phenotype and the mean deformation at the set of residues i that (individually) correlate best with the phenotype, as a function of the size of the set.

• Magnitude of mutation effect

– **PDB-PDB Correlation:** Some protein pairs tend to have more reliable correlations between their deformation vector, and deformation vectors of other matched pairs. This could be because the mutations have strong effects, which can be reliably measured. Thus we look at the subset of protein pairs for which we can measure the average PDB-PDB correlations with other matched pairs, and see whether this correlates with the PDB-AF correlations. We expect that high PDB-PDB correlations will be predictive of high PDB-AF correlations.

– **Deformation Magnitude:** We calculate the magnitude of the deformation at the mutated site, S_{m}, for both PDB and AF structures. High S_{m} indicates a large mutation effect.

– **BLOSUM Score:** BLOSUM62 scores describe how likely a particular amino acid substitution is, and are calculated from the frequency of substitutions in sequence alignments (Henikoff and Henikoff, 1992). Low BLOSUM scores indicate non-conservative mutations, and are expected to lead to larger mutation effects.

– **MSA mut-freq:** A more direct way of measuring the compatibility of certain mutations at specific positions in a protein sequence is to measure the frequency of

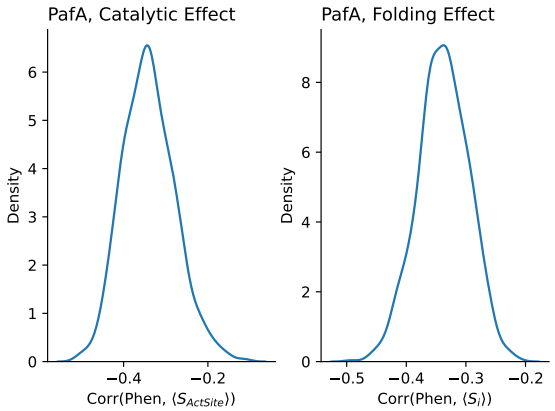

FIG. 11 Distribution of Pearson's correlation between phenotype and deformation obtained by randomly sub-sampling with replacement. Sub-sample size is half of the original sample size; 1,000 sub-samples were drawn.

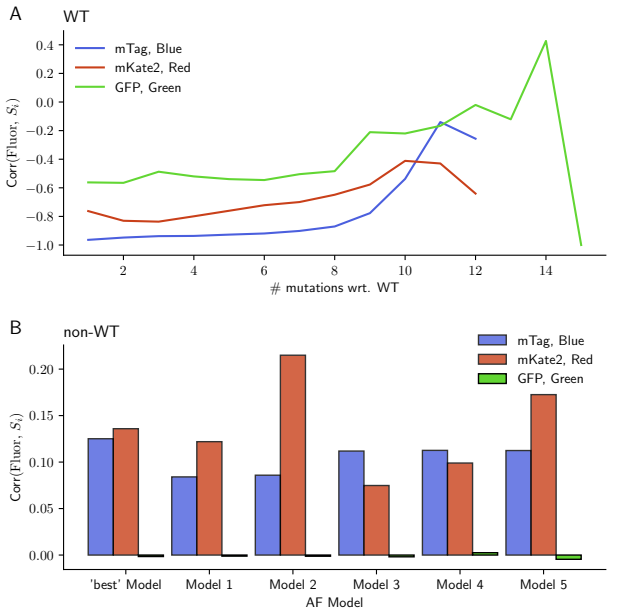

FIG. 12 A: Correlation between phenotype and ES ($\gamma = 13$ Å) as a function of the number of mutations from the WT protein. B: Correlation (Pearson's r) between structural change and the magnitude of phenotypic change between two sequences, i and j: $|\phi_i - \phi_j|$, where ϕ is the relevant phenotype. Correlations are shown for all AF models, for the effective strain at particular residues: mTag, $i = 65$; mKate2, $i = 218$; GFP, $i = 58$. For each data set, correlations are obtained by comparing 10,000 pairs of sequences chosen randomly with replacement.

a mutation in an MSA at the mutated position. One might expect that high-frequency mutations would lead to small changes, and thus may have lower correlations. We actually find the opposite effect, as high-frequency mutations have higher PDB-AF correlations. This may indicate that there is more information in the MSA, which could improve the prediction. Explicit effects of MSA size and coverage ought to be examined in a future study.

9. Deformation-phenotype correlations

A. mKate2

We calculate deformation (*i.e.* the ES, S_i) of all variants with respect to WT mKate2, at all sequence positions i. We then calculate the correlation of S_i at each position i with red fluorescence. Many positions have strong correlations with fluorescence, and the degree of correlation depends on how close each residue is to Y65, which is a site for covalent binding to the chromophore (Fig. 10A). Correlations are robust to choice of metric (Fig. 4).

B. GFP

We calculate deformation of all variants with respect to WT GFP, at all sequence positions i. We then calculate the correlation of S_i with green fluorescence. Many positions have strong correlations with fluorescence, and the degree of correlation depends on how close each residue is to S64, which is a site for covalent binding to the chromophore (Fig. 10A). Correlations are robust to the choice of metric (Fig. 4).

C. PafA

We calculate the deformation S_i of all variants with respect to WT PafA, at all sequence positions i. We then calculate the correlation of S_i with the measured catalytic effect, and folding effect. Many positions are correlated with either the folding effect and/or catalytic effect (Fig. 10A). For the catalytic effect, the strongest correlations are close to the binding site (which contains zinc atoms), while for the folding effect, the strongest correlations are more dispersed. We find higher correlations between deformation and phenotype when we take the mean of deformation values amongst the top λ residues that correlate best with the phenotype (Fig. 10B). In the case of the catalytic effect, we take the first $\lambda = 5$ residues, while for the folding effect, we take $\lambda = 50$. Correlations are robust to the choice of metric (Fig. 4) and robust to subsampling (Fig. 11).

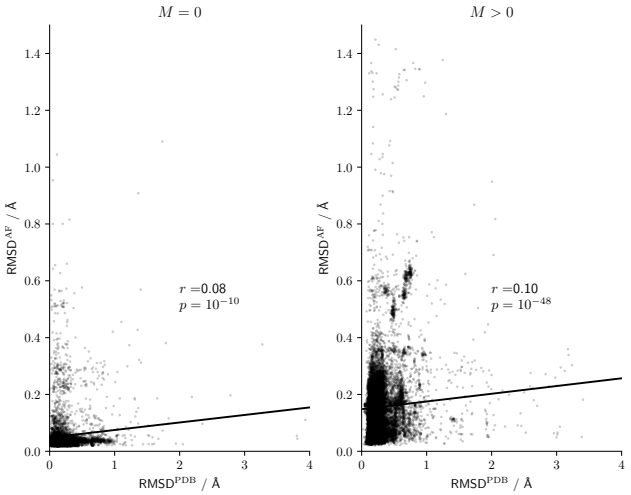

FIG. 13 **Correlations between empirical and predicted RMSD.** RMSD calculated using PDB structures, versus RMSD calculated using AF-predicted structures, for the full set of structure pairs from the PDB; non-mutated pairs (left) and mutated pairs (right) are shown separately. Pearson's correlation coefficient and linear fit are shown.

D. Dependence on number of mutations

We consider the possibility that since deformation likely increases with the number of mutations, the phenotype-deformation correlations may be simply due to an underlying correlation between phenotype and the number of mutations from the WT. To examine this, we measured correlations between phenotype and ES, controlling for the number of mutations (Fig. 12A). We find that the correlations are independent of the number of mutations in the region $M \leq 8$. Even further away from the WT than $M = 8$ we see a drop in correlations. Thus, the phenotype-ES correlations are not due to the number of mutations, but they are only strong for comparing sequences close to the WT.

E. Comparing mutants with mutants

We suspect that the high deformation-phenotype correlations we found are only possible when comparing WT with mutants. Our reasoning is that deformation should, by itself, not be sufficiently informative to predict changes in complex phenotypes. However, if one assumes that WT proteins are locally optimal, then any changes to the structure ought to have a deleterious effect on phenotype. To examine this, we calculate correlations between deformation and absolute differences in phentoype for 10,000 randomly chosen pairs of mutants (Fig. 12B). We find that the correlations are considerably lower than when comparing WT to mutants, in support of our conjecture.

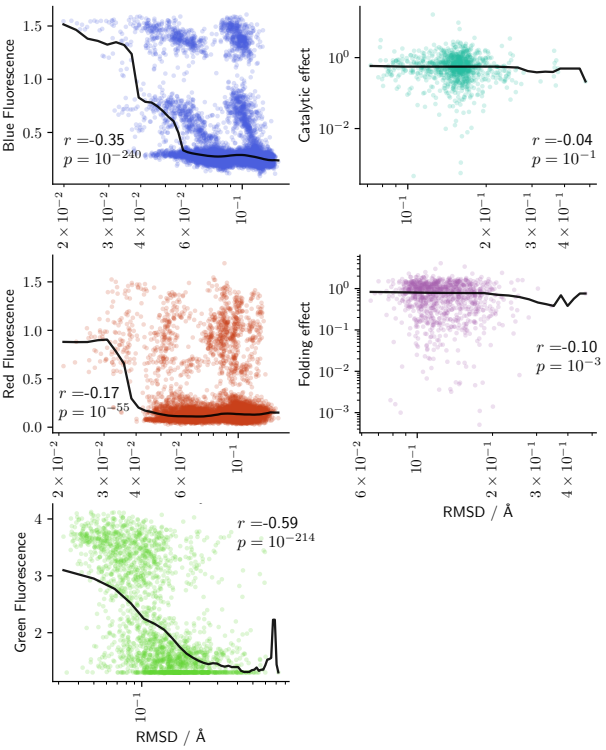

FIG. 14 **Correlations between RMSD and phenotype measurements.** Phenotype as a function of RMSD between WT and mutant. Red line indicates the median, calculated using a sliding window. Pearson's r and corresponding p-values are shown on the graph.

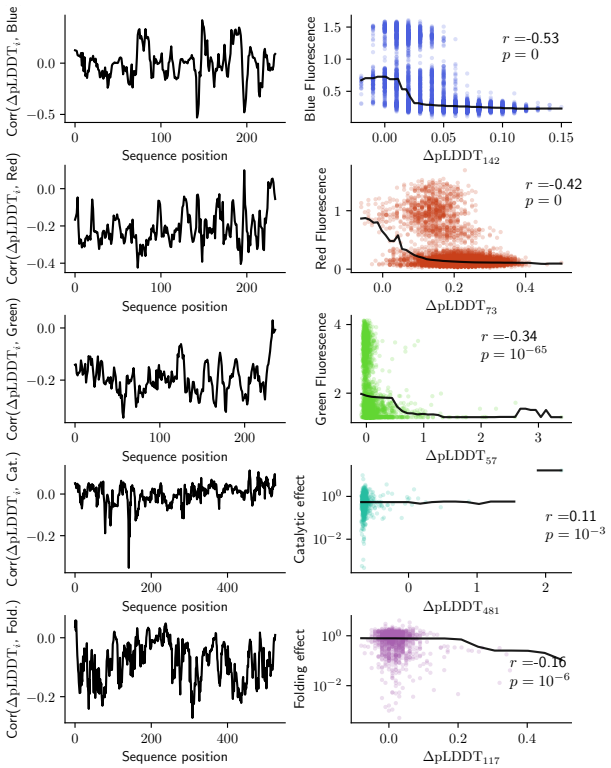

FIG. 15 **Correlations between ΔpLDDT and phenotype measurements.** (Left) Correlations between ΔpLDDT at residue i and phenotype for 5 sets of phenotype measurements (top to bottom: mTag, blue fluorescence; mTag, red fluorescence; GFP, green fluorescence; PafA, catalytic effect; PafA, folding effect) as a function of sequence position i. In each case, results are shown for the AF model with the highest correlation. (Right) Phenotype as a function of ΔpLDDT for the residue with the highest correlation. Red line indicates the median, calculated using a sliding window. Pearson's r and corresponding p values are shown on the graph; zero values are shown for p when p is lower than is possible for 64-bit floating point precision.

10. RMSD

At the outset, we expected that since mutation effects are assumed to be typically localized, it would be prudent to use a local measure of deformation. Here we show that using RMSD, which depends on global alignment, is indeed insufficient to measure the effect of a mutation. We calculate the correlation between RMSD values obtained using PDB vs AF-predicted structures, for all structure pairs in our PDB dataset. Fig. 13 shows that the correlation is quite weak, and does not differ for mutated vs non-mutated pairs. We also show correlations for the best-performing AF models of RMSD between WT and mutant structures, and the corresponding phenotype (Fig. 14). By its nature, RMSD is a global measure that obscures localized deformation effects. Thus, as expected, the correlations for RMSD are much lower than what is observed for local measures of deformation (Fig. 4); one exception is GFP, where RMSD correlates almost as well with green fluorescence ($r = -0.59$).

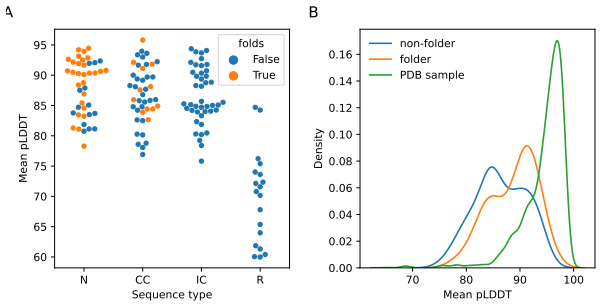

FIG. 16 A: Mean pLDDT for 147 protein sequences based on the WW domain, some of which are found to fold *in vitro*, and some of which do not fold. Proteins are grouped into: natural sequences (N); generated sequences that matches the covariance of the natural sequences (CC); generated sequences that match the statistics of the natural sequences at each position (IC); sequences generated by shuffling the natural sequences (R). B: We compare mean pLDDT of the folding and non-folding WW domain sequences (excluding the shuffled sequences, as these share little sequence identity with the natural sequences), with a sample of mean pLDDT values for proteins in the PDB (same sampling procedure as in the main text). In this case, mean pLDDT appears to be of little use in differentiating folding and non-folding sequences.

11. pLDDT

Since AlphaFold has been released, much attention has been given to pLDDT, AlphaFold's predicted (using a neural network) confidence score. pLDDT is, effectively, AlphaFold's attempt to predict the LDDT score between the predicted structure and the ground truth (PDB). Recent studies have shown that pLDDT is a good predictor of disorder (Piovesan *et al.*, 2022), and that it is correlated with measures of flexibility: root-mean-square deviation measured in molecular dynamics simulations (Guo *et al.*, 2022); S2 from NMR (Ma *et al.*, 2023). This makes sense since flexible regions will inevitably lead to lower-than-average LDDT scores across repeat measurements or predictions. Although there is surely more to uncover, the evidence currently points to pLDDT as being a measure of confidence in the prediction of a residue's position, and a proxy measure for local flexibility. In this work, for example, we exclude residues from deformation calculations if they have pLDDT < 70.

A. Correlations between changes in pLDDT and phenotype

It is plausible that changes in pLDDT may correlate with phenotype, as large changes in pLDDT not only indicates changes in flexibility, but may also be an indicator of large deformation. We therefore look at AF-predicted changes in pLDDT for each residue in mutants compared to the wild type. We calculate the correlation between $\Delta\text{pLDDT}_i = \text{pLDDT(WT)} - \text{pLDDT(Mutant)}$ and phenotype for all five sets of phenotype measurements (Fig. 15). We find statistically significant correlations between ΔpLDDT_i and phenotype, but in each case correlations are about half of what can be achieved using other deformation metrics.

Despite using similar data to (Pak *et al.*, 2023), it might seem that we find higher correlations between pLDDT and fluorescence, but there are two important differences: In (Pak *et al.*, 2023) they study 447 single mutants, while we study 2,312 mutants, 200 of which are single mutants. We calculate correlations between ΔpLDDT_i and fluorescence, while (Pak *et al.*, 2023) calculate the difference in pLDDT at the mutated site ΔpLDDT_m ($r = 0.17$), and the difference in mean pLDDT, $\Delta\langle\text{PLDDT}\rangle$ ($r = 0.16$). For a more direct comparison with this study, we calculated correlations between the same quantities on the 200 single mutants that we examined. Like (Pak *et al.*, 2023), we found that these quantities are not very predictive of fluorescence (ΔpLDDT_m, $r = 0.03$; $\Delta\langle\text{PLDDT}\rangle$, $r = 0.06$).

B. Is pLDDT a good predictor of protein folding?

We are quite confident that pLDDT is a good measure of local flexibility, and that values under 70 are strong indicators of disorder. However, it is less clear whether average (across all residues) pLDDT values are useful for predicting whether a protein will fold or not, especially when the average is above 70. To test this, we looked at the mean pLDDT for a set of WW-domain-like proteins from (Socolich *et al.*, 2005) (Fig. 16). We find that proteins that were created from randomly shuffled sequences produced noticeably lower mean pLDDT values than the other, less random sequences. Proteins that had mutations that only took into consideration the positional statistics (IC; *i.e.*, the likelihood of an amino acid at a certain position in the sequence, as determined from a multiple sequence alignment) were not found to fold *in vitro*, yet they have comparable values of mean pLDDT to the WT sequences that were tested (natural sequences, N). This suggests that mean pLDDT alone is not a useful predictor of whether a protein will fold.

12. AlphaFold Models

Since AF was released with 5 sets of trained model weights, there is a choice of which structure to use. We find that no one model produces significantly better correlations than the other models (Fig. 17). We find significant differences between the predictions of different models on correlations with phenotype (Fig. 4), but the scale of these differences are reduced by calculating average structures ($\langle\text{AF}\rangle$). We found one case (PafA) where models 1 and 2 produced poor predictions (Fig. 4). In this case, most of the sequence variants were predicted to have very similar structures. We attribute this odd behavior to the use of structural templates, which are used by models 1 and 2 by default. We find much better results for these two models when using the ColabFold implementation (Fig. 8, Fig. 9), where we were able to run AF without using structural templates. This leads leads us to recommend not using structural templates in AF predictions. In general, the repeat predictions of structures are more similar when produced by the same model (and conversely, different models produce different structures to each other); for this reason we also advise against averaging over structures from different models unless there are multiple repeat predictions from each model (in our case, we used 6 repeat structures to create average structures).

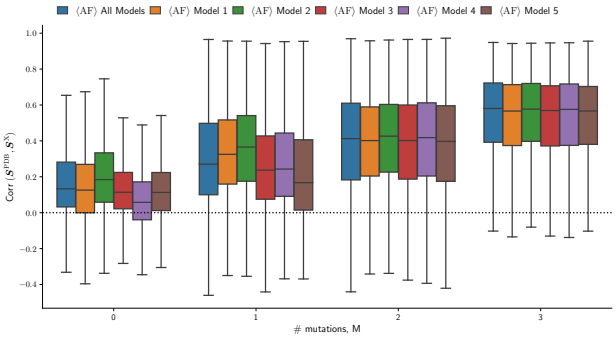

FIG. 17 **Effect of AF model.** PDB - ⟨AF⟩ correlation as a function of M and AF model type.

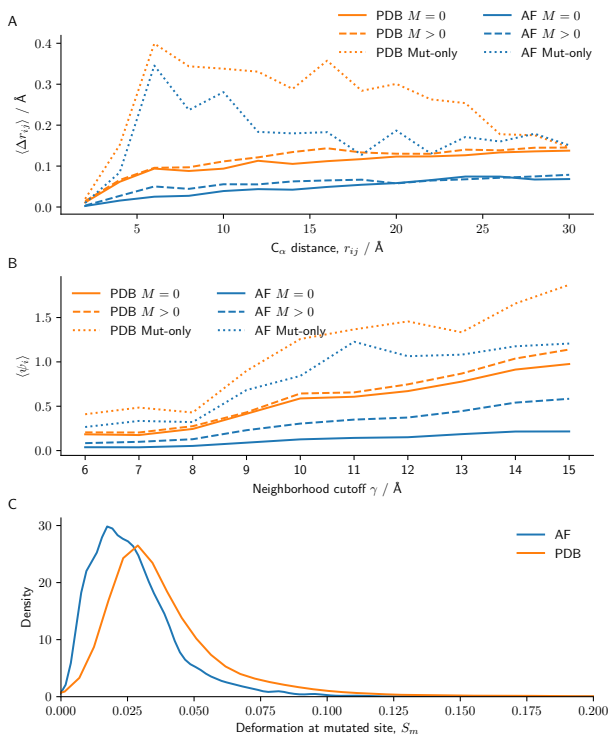

FIG. 18 **Changes in backbone distances, neighborhoods and deformation due to mutation.** A: Average change in C_α distances between structures, as a function of C_α distance. B: Average neighbor mismatch, ψ_i, as a function of neighborhood cutoff γ. Separate lines are shown for PDB and AF structures, and for all pairs of residues in non-mutated pairs ($M = 0$), all pairs of residues in mutated pairs ($M > 0$), and only comparing residues with mutated residues (Mut-only). C: Distribution of mutation effects, S_m, for PDB and AF-predicted pairs of structures.

13. Scale of mutation effects

A. Deformation between PDB structures is typically small

To give an alternative view of the magnitude of deformation in the PDB, we examine changes to distances between residues. We calculate the distance between all C_α positions, r_{ij}, and then get the absolute difference between these distances in a reference structure and a target structure, $\Delta r_{ij} = r_{ij} - r'_{ij}$. In Fig. 18A, we show the mean absolute difference between C_α distances, as a function of C_α distance; we calculate this for all residues in all protein pairs with $M = 0$ and $M > 0$, and also for C_α distances with respect to mutated residues. On average, backbone C_α positions move by less than 0.2 Å, both for PDB and AF-predicted structures. However, when looking at backbone distances only from mutated sites, we see average mutations of up to 0.4 Å within 1 nm from mutated sites. It seems that a few mutations typically lead to rather slight changes to bulk protein structure, yet can have significant effects locally.

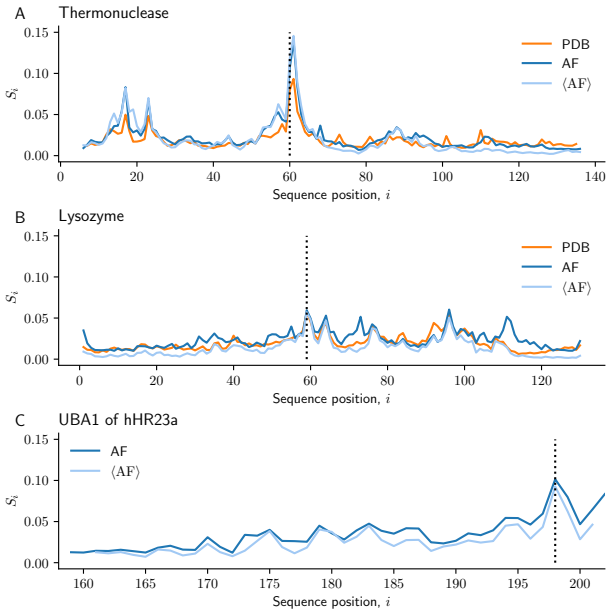

FIG. 19 **Examples of large deformation.** Deformation per residue for three examples: Thermonuclease from *staphylococcus aureus* (A), human Lysozyme (B), and the UBA1 domain of human hHR23a (C). Deformation is shown for AF, ⟨AF⟩, and PDB structures where possible. The location of mutations is indicated by dotted lines.

When we calculate deformation, we include in our calculation residues that are neighbors in both structures. One might suspect that this can lead to problems if deformation leads to big changes in neighborhoods. To test this, for each residue we calculate the symmetric neighbor difference, ψ_i, as the number of residues that are not shared between neighbor sets in a reference, N_i and target structure, N'_i,

$$\psi_i = \left| N_i \ominus N'_i \right| .$$

We calculate the average neighbor difference across all pairs and residues, $\langle \psi_i \rangle$, as a function of neighborhood cutoff γ. We show in Fig. 18B that neighbors typically differ by less than one residue, even when focusing on mutated sites.

To help put mutation effects in context, we plot the distribution of mutation effects (ES at the mutated site, S_m) for both PDB and AF-predicted structures (Fig. 18C). This provides a handy reference for understanding when an effect is relatively large or small.

B. Example of large deformation

Typically, mutation effects in the PDB are quite small, such that they often have only subtle effects even on local structure. Perhaps the best way to illustrate this is to show examples of some of the largest deformations measured. We chose two examples of proteins that differ by one amino

acid, which exhibit some of the highest deformations found in the PDB: (i) thermonuclease (V60H, comparing `3QOL_A` and `5IGC_A`)) and (ii) lysozyme (T59Y, comparing `2MEH_A` and `2MEI_A`). In each case, while deformation is higher (Fig. 19) then what is typical (Fig. 18C), the structural differences are not so severe (Fig. 20). In thermonuclease, the replacement of a valine with histidine results in a slight kink in the α-helix, due to the greater size of the histidine which interferes with how the amino acids pack. In lysozyme, we see a similar effect but this time in a more central location within the protein; replacing threonine with tyrosine results in the neighboring $\alpha-$helix and $\beta-$sheets being pushed apart. Both of these effects are recapitulated in AF predictions (Fig. 20). We note that mutation effects of this magnitude are rare in the PDB, yet they are still difficult to evaluate visually. Nor do we find pLDDT to be a reliable indicator of structure change: thermonuclease, $\Delta\text{pLDDT}_m = -2.5$; lysozyme, $\Delta\text{pLDDT}_m = 0.04$. Instead, *measuring deformation allows us to quantify the effect of a mutation with much higher precision, in a way that can be directly compared with experimental results*.

C. UBA1 of hHR23a

A recent paper studied the UBA1 domain of hHR23a protein (amongst other examples), and suggested that differences in packing, and differences in pLDDT constitute evidence that AF cannot predict the effect of missense mutations (Buel and Walters, 2022). We here show that the effect of the mutation (L198A) is actually quite large compared to average effects (Fig. 19C). The deformation at the mutated site is $S_\text{m} = 0.093$, which is higher than 97 % of all deformation values measured at mutated sites in PDB structures. The mutation in question leads to higher degradation by proteases, which shows that this mutation destablizes the structure. We consider that such large deformations are perhaps likely to destabilize the structure, given how rare they are in the PDB. The PDB is constructed with a sampling bias towards proteins that can fold, so the fact that there are so few mutants with this level of deformation might indicate that high deformation is a good predictor of destabilization. This is an interesting area for future study.

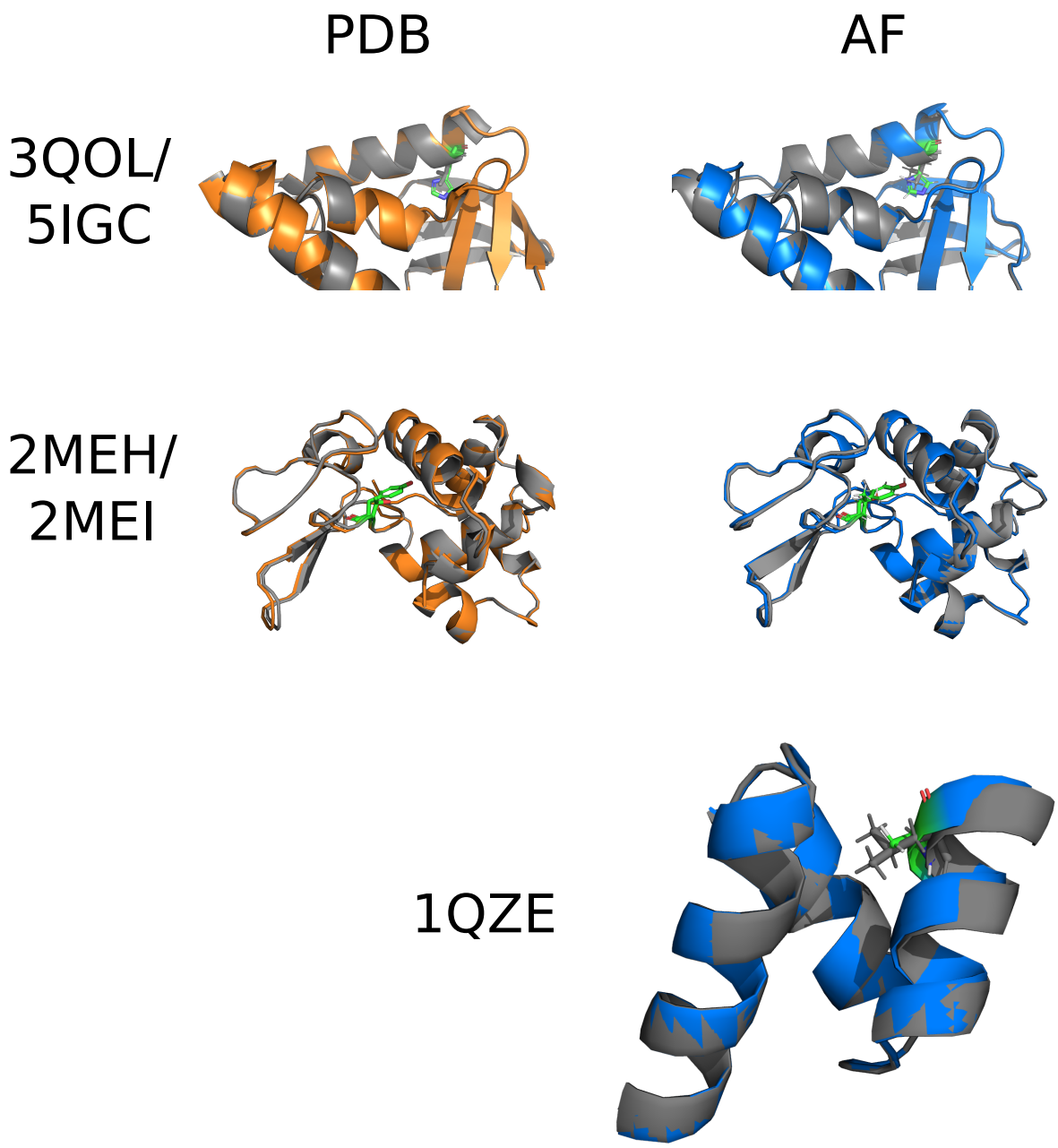

FIG. 20 **Examples of large deformation.** Comparisons of PDB and AF-predicted structures that differ by one amino acid: Thermonuclease from *staphylococcus aureus* (V60H, `3QOL_A` and `5IGC_A`; a close-up of the region affected by the mutation is shown), human Lysozyme (T59Y, `2MEH_A` and `2MEI_A`), and the UBA1 deomain of human hHR23a (L198A). Mutated structures are aligned and shown overlaid as cartoons; atomic positions are shown for the mutated residues. PDB structures are only shown for examples where both proteins are available.

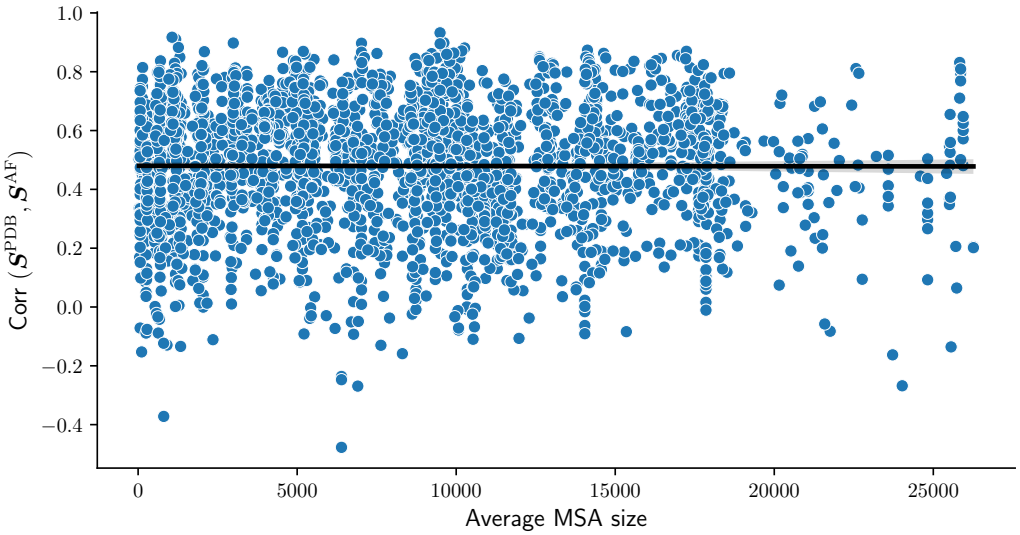

FIG. 21 **Effect of MSA size on PDB-AF correlation.** Average MSA size is the average across the two proteins in each pair of mutants. For protein pairs with no mutation, it is just the size of the MSA. Results are shown for a non-redundant sample.

References

*jmmcbride@protonmail.com
†nanogrzybowski@gmail.com
‡tsvitlusty@gmail.com

del Alamo, Diego, Davide Sala, Hassane S Mchaourab, and Jens Meiler (2022), "Sampling alternative conformational states of transporters and receptors with alphafold2," eLife **11**, e75751.

Buel, Gwen R, and Kylie J. Walters (2022), "Can alphafold2 predict the impact of missense mutations on structure?" Nature Structural & Molecular Biology **29** (1), 1–2.

Eckmann, J P, J. Rougemont, and T. Tlusty (2019), "Colloquium: Proteins: The physics of amorphous evolving matter," Rev. Mod. Phys. **91**, 031001.

Falk, M L, and J. S. Langer (1998), "Dynamics of viscoplastic deformation in amorphous solids," Phys. Rev. E **57**, 7192–7205.

Guo, Hao-Bo, Alexander Perminov, Selemon Bekele, Gary Kedziora, Sanaz Farajollahi, Vanessa Varaljay, Kevin Hinkle, Valeria Molinero, Konrad Meister, Chia Hung, Patrick Dennis, Nancy Kelley-Loughnane, and Rajiv Berry (2022), "Alphafold2 models indicate that protein sequence determines both structure and dynamics," Scientific Reports **12** (1), 10696.

Henikoff, S, and J G Henikoff (1992), "Amino acid substitution matrices from protein blocks." Proceedings of the National Academy of Sciences **89** (22), 10915–10919, https://www.pnas.org/doi/pdf/10.1073/pnas.89.22.10915.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis (2021), "Highly accurate protein structure prediction with alphafold," Nature **596** (7873), 583–589.

Kabsch, W (1976), "A solution for the best rotation to relate two sets of vectors," Acta Crystallographica Section A **32** (5), 922–923.

Kabsch, Wolfgang, and Christian Sander (1983), "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," Biopolymers **22** (12), 2577–2637, https://onlinelibrary.wiley.com/doi/pdf/10.1002/bip.360221211.

Kandathil, Shaun M, Joe G. Greener, Andy M. Lau, and David T. Jones (2022), "Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins," Proceedings of the National Academy of Sciences **119** (4), e2113348119, https://www.pnas.org/doi/pdf/10.1073/pnas.2113348119.

Kermode, J R, and L. Pastewka (2023), "Matscipy: Generic python materials science toolkit," .

Kufareva, Irina, and Ruben Abagyan (2012), "Methods of protein structure comparison," in *Homology Modeling: Methods and Protocols*, edited by Andrew J. W. Orry and Ruben Abagyan (Humana Press, Totowa, NJ) pp. 231–257.

Li, Weizhong, and Adam Godzik (2006), "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," Bioinformatics **22** (13), 1658–1659, https://academic.oup.com/bioinformatics/article-pdf/22/13/1658/484588/btl158.pdf.

Lubliner, Jacob (2008), *Plasticity theory* (Courier Corporation).

Ma, Puyi, Da-Wei Li, and Rafael Brüschweiler (2023), "Predicting protein flexibility with alphafold," Proteins: Structure, Function, and Bioinformatics **91** (6), 847–855, https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26471.

Mariani, Valerio, Marco Biasini, Alessandro Barbato, and Torsten Schwede (2013), "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests," Bioinformatics **29** (21), 2722–2728, https://academic.oup.com/bioinformatics/article-pdf/29/21/2722/18533347/btt473.pdf.

Markin, C J, D. A. Mokhtari, F. Sunden, M. J. Appel, E. Akiva, S. A. Longwell, C. Sabatti, D. Herschlag, and P. M. Fordyce (2021), "Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics," Science **373** (6553), eabf8761, https://www.science.org/doi/pdf/10.1126/science.abf8761.

Mirdita, Milot, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger (2022), "Colabfold: making protein folding accessible to all," Nature Methods **19** (6), 679–682.

Pak, Marina A, Karina A. Markhieva, Mariia S. Novikova, Dmitry S. Petrov, Ilya S. Vorobyev, Ekaterina S. Maksimova, Fyodor A. Kondrashov, and Dmitry N. Ivankov (2023), "Using alphafold to predict the impact of single mutations on protein stability and function," PLOS ONE **18** (3), 1–9.

Piovesan, Damiano, Alexander Miguel Monzon, and Silvio C. E. Tosatto (2022), "Intrinsic protein disorder and conditional folding in alphafolddb," Protein Science **31** (11), e4466, https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4466.

Poelwijk, Frank J, Michael Socolich, and Rama Ranganathan (2019), "Learning the pattern of epistasis linking genotype and phenotype in a protein," Nature Communications **10** (1), 4213.

Saldaño, Tadeo, Nahuel Escobedo, Julia Marchetti, Diego Javier Zea, Juan Mac Donagh, Ana Julia Velez Rueda, Eduardo Gonik, Agustina García Melani, Julieta Novomisky Nechcoff, Martín N Salas, Tomás Peters, Nicolás Demitroff, Sebastian Fernandez Alberti, Nicolas Palopoli, Maria Silvina Fornasari, and Gustavo Parisi (2022), "Impact of protein conformational diversity on AlphaFold predictions," Bioinformatics **38** (10), 2742–2748, https://academic.oup.com/bioinformatics/article-pdf/38/10/2742/49010056/btac202.pdf.

Sarkisyan, Karen S, Dmitry A. Bolotin, Margarita V. Meer, Dinara R. Usmanova, Alexander S. Mishin, George V. Sharonov, Dmitry N. Ivankov, Nina G. Bozhanova, Mikhail S. Baranov, Onuralp Soylemez, Natalya S. Bogatyreva, Peter K. Vlasov, Evgeny S. Egorov, Maria D. Logacheva, Alexey S. Kondrashov, Dmitry M. Chudakov, Ekaterina V. Putintseva, Ilgar Z. Mamedov, Dan S. Tawfik, Konstantin A. Lukyanov, and Fyodor A. Kondrashov (2016), "Local fitness landscape of the green fluorescent protein," Nature **533** (7603), 397–401.

Socolich, Michael, Steve W. Lockless, William P. Russ, Heather Lee, Kevin H. Gardner, and Rama Ranganathan (2005), "Evolutionary information for specifying a protein fold," Nature **437** (7058), 512–518.

Sun, Zhoutong, Qian Liu, Ge Qu, Yan Feng, and Manfred T. Reetz (2019), "Utility of b-factors in protein science: Interpreting rigidity, flexibility, and internal motion and engineering thermostability," Chemical Reviews **119** (3), 1626–1665, pMID: 30698416, https://doi.org/10.1021/acs.chemrev.8b00290.

Zemla, Adam (2003), "LGA: a method for finding 3D similarities in protein structures," Nucleic Acids Research **31** (13), 3370–3374, https://academic.oup.com/nar/article-pdf/31/13/3370/9487387/gkg571.pdf.

Zhang, Hua, Tuo Zhang, Ke Chen, Shiyi Shen, Jishou Ruan, and Lukasz Kurgan (2009), "On the relation between residue flexibility and local solvent accessibility in proteins," Proteins: Structure, Function, and Bioinformatics **76** (3), 617–636, https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.22375.

Zhang, Yang, and Jeffrey Skolnick (2004), "Scoring function for automated assessment of protein structure template quality," Proteins: Structure, Function, and Bioinformatics **57** (4), 702–710, https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.20264.