AlphaFold2 can predict single-mutation effects on structure and phenotype

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*

AlphaFold2 (AF) is a promising tool, but is it accurate enough to predict single mutation effects? Here, we report that a measure for localized structural deformation between protein pairs differing by only 1-3 mutations is correlated across 4,645 experimental and AF-predicted structures. Furthermore, analysis of ∼11,000 proteins shows that the local structural change correlates with various phenotypic changes. These findings suggest that AF can predict the magnitude of single-mutation effects in many proteins, and we propose a method to identify those proteins for which AF is most predictive.

Alteration of one or few amino acid residues can affect structure [1–3] and function [4] of a protein and, in extreme cases, be the difference between health and disease [5, 6]. Understanding structural consequences of point mutations is important for drug design [7, 8] and could also accelerate optimization of enzymatic function via directed evolution [9, 10]. In these and other applications, theoretical models [11] could be of immense help, provided they are sufficiently accurate. In this context, AlphaFold2 [12] has recently made breakthroughs in predicting global protein structure from sequence with unprecedented precision. Notwithstanding, it is not yet known whether AF is sensitive enough to detect small, local effects of single mutations. Even if AF achieves high accuracy, the effect of a mutation may be small compared to the inherent conformational dynamics of the protein – predicting static structures may not be particularly informative [13–15]. Furthermore, as accuracy improves, evaluating the quality of predictions becomes increasingly complicated by the inherent noise in experimental measurements [15–22]. So far, no study has evaluated whether AF can accurately measure structural changes due to single mutations, and there are conflicting reports as to whether AF can predict the effect of a mutation on protein stability [23–27]. Furthermore, recent evidence suggests that AF learns the energy functional underlying folding, raising the question of whether the inferred functional is sensitive enough to discern the subtle physical changes due to a single mutation [28]. We aim to resolve this issue by comparing AF predictions with extensive data on protein structure and function.

We examine AF predictions in light of structural data from a curated set of proteins from the Protein Data Bank (PDB) [29], and phenotype data from high-throughput experiments [30–32]. We find that AF can detect the effect of a mutation on structure by identifying local deformations between protein pairs differing by 1-3 mutations. The deformation is probed by the effective strain (ES) measure. We show that the ES measure computed between a pair of PDB structures is correlated with the ES computed for the corresponding pair of structures predicted by AF. Furthermore, analysis of ∼10,000 proteins whose function was probed in three high-throughput studies shows significant correlations between local deformation in AF-predicted structures, as measured by ES, and three categories of phenotype (fluorescence, folding, catalysis) across three experimental data sets [30–32]. These sets of correlations suggest that AF can predict the range and magnitude of single-mutation effects. We provide new tools for computing deformation in proteins, and a methodology for increasing the precision of AlphaFold predictions. Altogether, these results indicate that AF can be used to predict physicochemical effects of missense mutations, undamming vast potential in the field of protein design and evolution.

AF can predict local structural change.— We illustrate our approach by analyzing wild-type (`6BDD_A`) and single-mutant (`6BDE_A`, A71G) structures of H-NOX protein from *K. algicida* (Fig. 1D). [33] To quantify local deformation, we calculate the effective strain (ES) per residue S_i (See App. A) for, respectively, experimental and AF-predicted pairs of structures (Fig. 1A). The ES is the mean relative change in distance from C_α of residue i to neighboring C_α positions within a range of 13 Å. ES provides a robust estimate of the magnitude of local strain, which accounts also for non-affine deformation in addition to affine deformation [34, 35]. Like the frame-aligned-point-error (FAPE) measure used in training AF [12], ES is invariant to alignment. In H-NOX, we observe that the S_i is highest at, and decays away from the mutated site (Fig. 1B). Furthermore, S_i is correlated with distance from the mutated site (Fig. 1B). We find that S_i is correlated across PDB and AF structures (Fig. 1C,E). Taken together, these correlations suggest that S_i is a sensitive measure of local structural change, and that AF is capable of predicting such structural change upon mutation.

Experimental measurement variability limits evaluation.— Before exploring AF predictions in more detail, we first examine variation within experimental structures by comparing repeat measurements of the same protein. In Fig. 1F we show the distribution of S_i calculated for all residues in all pairs (Supplemental Material (SM) Sec. 1A [36]) of protein structures with identical sequences (number of mutations, $M = 0$); we excluded pairs where the crystallographic group differed (SM Sec. 1B [36]). Protein structures vary considerably between repeat measurements (average ES is $\langle S_i \rangle = 0.019$, and the average Root Mean Square Deviation is RMSD = 0.27 Å), while differences between repeat predictions of AF are much lower ($\Delta S_i =$

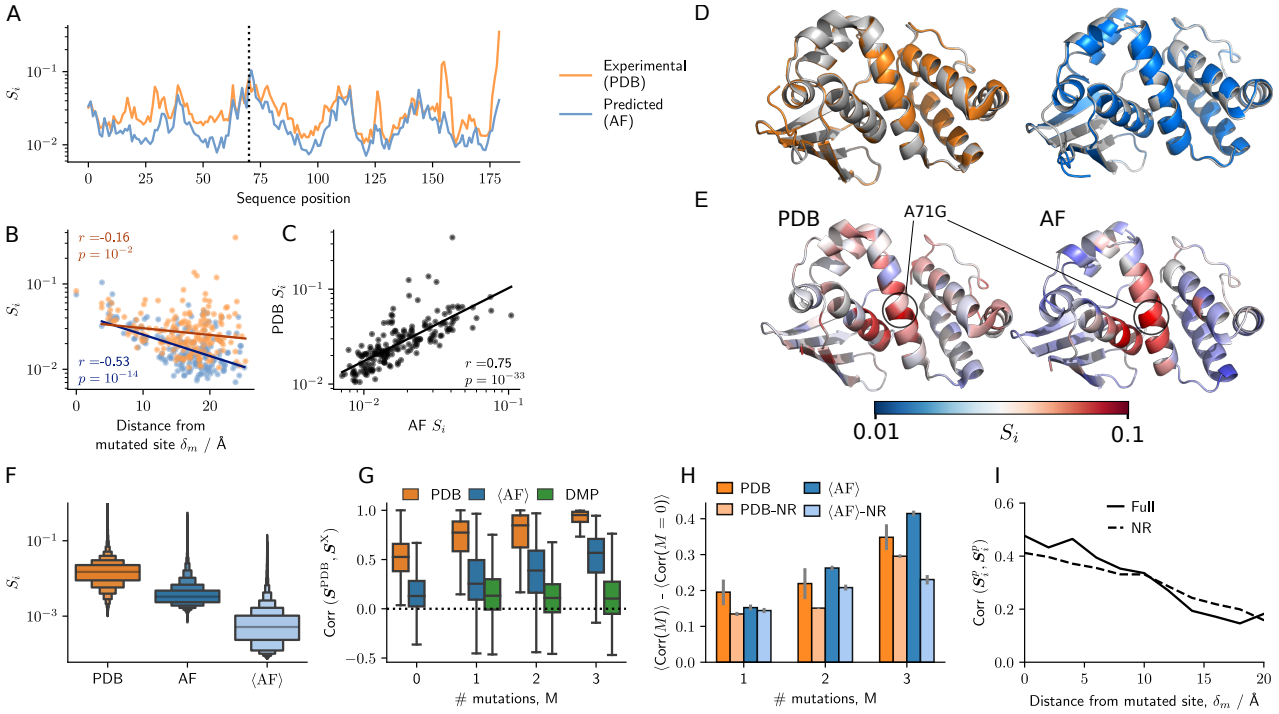

FIG. 1. A: Local deformation per residue measured by strain, S_i, between wild-type and mutant (A71G) H-NOX protein, for experimental (orange) and AF-predicted (blue) structures. Dotted line indicates mutated residue. B: S_i vs distance from the nearest mutated site, δ_m. C: Comparison of S_i obtained from experimental and predicted structures. D: Overlaid wild type (grey, `6BDD_A`) and mutant (colour, `6BDE_A`), experimental (orange) and predicted (blue) structures. E: Wild type protein with residues coloured by S_i; location of A71G mutation is shown. F: Distribution of S_i between matched pairs of structures with the same sequence ($M = 0$), for PDB, AF, and averaged AF ($\langle$AF$\rangle$) structures. G: Distribution of correlation between PDB strain fields and equivalent vectors from PDB, AF and DMPfold, shown for different numbers of mutations, M. H: Residual correlation that is due to mutations, shown for the full dataset and a non-redundant version (NR); whiskers show bootstrapped 95 % confidence intervals. I: Correlation between PDB and $\langle$AF$\rangle$ strain fields, $\mathbf{S}_i^p$, across all pairs p and residues i that are within a distance δ_m from a mutated site, shown for the full dataset and a non-redundant version (NR).

0.005, RMSD =0.06 Å). For comparison, the experimental RMSD between WT and mutant H-NOX is 1.6 Å, while the AF-predicted RMSD is 0.3 Å. We can refine AF predictions further by making multiple repeat predictions and averaging over the local neighborhoods ($\langle$AF$\rangle$ in Fig. 1F, App. B), which results in even lower differences ($\Delta S_i = 0.001$). We find that averaging decreases deformation away from mutated residues, while preserving deforamtion in mutated areas, and it typically leads to higher correlations with empirical data; thus we henceforth report results for averaged structures, except where noted. This variation between experimental measurements might mask the deformation due to mutation, and therefore limits our ability to evaluate AF predictions.

Mutation effects are measurable in PDB structures.— To quantify how well we can measure mutation effects from PDB structures, we compare deformation between two matching pairs of PDB structures with identical ($M = 0$) and non-identical ($M > 0$) sequences (SM Sec. 1A [36]) of length L (number of residues). For each pair, we calculate the strain fields, $\mathbf{S} = (S_1, \ldots, S_L)$, which record ES values for all residues, and we calculate Pearson's correlation coefficient r as in Fig. 1C. We find that even among protein structures with identical sequences, strain fields are highly correlated (Fig. 1G). This occurs because the magnitude of positional fluctuations depends on local flexibility; more flexible regions exhibit higher strain in repeat measurements (App. B). Thus, a portion of the $\mathbf{S}$ correlation in Fig. 1C is due to effects other than mutation. Despite this, we find that correlations are much higher when comparing pairs of structures that differ by one or more mutations ($M > 0$), and correlations increase with M (Fig. 1G). Thus, the strength of PDB-PDB deformation correlations is partly due to differences in local flexibility, and partly also due mutations.

Mutation effects are correlated across PDB and AF structures.— To evaluate the performance of AF in prediction muatation effects, we calculate correlations between PDB and AF-predicted strain fields, $\mathbf{S}^\mathrm{PDB}$ and $\mathbf{S}^\mathrm{AF}$, calculated for all matched pairs of proteins (SM Sec. 1A [36]). The PDB-$\langle$AF$\rangle$ correlations between pairs of structures with identical sequences ($M = 0$) are lower than PDB-PDB correlations (Fig. 1G), as are the correlations for non-identical sequences ($M > 0$). Nonetheless, the correlations are significant and they increase with M. To put this result in context, the PDB-AF correlations are considerably higher than correlations obtained by using another algorithm to predict protein structure (DMPfold2) [37]. To compare the degree of correlation that is due to mutation effects, we plot the mean correlation for non-identical sequences $\langle \mathrm{corr}\big(M \in \{1,2,3\}\big)\rangle$ subtracted from the mean correlation that can be attributed to fluctuations, $\langle \mathrm{corr}\big(M = 0\big)\rangle$. Fig. 1H shows that the degree of correlation due to mutations is as high for AF-PDB comparisons as it is

for PDB-PDB comparisons. Since many protein families are overrepresented in the PDB, we repeat the analyses on non-redundant sets of proteins (SM Sec. 1C [36]), finding that AF-PDB correlations are still comparable to PDB-PDB correlations (NR in Fig. 1H).

AF predicts the range of mutation effects.— Fig. 1G-H shows that within matched protein pairs, deformation is correlated between PDB and AF, although the magnitude of deformation can differ (Fig. 1F). This indicates that AF is at least correctly predicting the range and the relative strength of the effect of a mutation. On average, AF predicts that mutations can produce changes in structure up to 16-18 Å (SM Sec. 7 [36]), whereas the average range in the PDB data is only14 Å due to the higher measurement variance in the PDB. This suggests that AF correctly predicts the range of a mutation's effect on structure.

AF predicts the relative magnitude of mutation effects.— It is essential to be able to predict whether a mutation will lead to a big or small effect on structure. While the previous analysis did not show this, we directly address this problem by examining whether predicted effects correlate with empirical effects *across* proteins. To do this, we group S_i values from all matched pairs p by distance (in bins of 2 Å) from the nearest mutated residue, δ_{m}, to get sets of S_i^p for both PDB and $\langle \mathrm{AF} \rangle$ pairs of structures. This allows us to compare ES magnitudes across proteins, by calculating the correlation between S_i^p for PDB and $\langle \mathrm{AF} \rangle$. At mutated sites, the correlation is quite high, and decreases away from the mutated site as expected (Fig. 1I); this is also true for the non-redundant sample. Hence, AF is capable of distinguishing between mutations that have relatively large or small effects on structure.

Phenotypic change correlates with AF-predicted structural change.— An orthogonal test of whether AF can predict the effect of a mutation is to study correlations between the effective strain (ES), S_i, and phenotypic change. This approach avoids the pitfalls associated with noisy PDB measurements, and allows us to test predictions of structures that AF was not directly trained on. However, the link between structure and function is often unknown, and likely quite complex. Therefore, a lack of a correlation between S_i and phenotype is not strong evidence that the structure is incorrect, as there maybe be a non-trivial mapping between structure and function. On the other hand, observation of correlations between S_i and phenotype is strong evidence that AF can be predictive in estimating the effect of mutations. We study three data sets from high-throughput experiments, covering three distinct phenotypes (SM Sec. 2 [36]): (i) blue and red fluorescence is measured for 8,192 sequences linking mTagBFP2 (blue) and mKate2 (red) [31]; (ii) green fluorescence is measured for 2,312 GFP sequences [30]; (iii) folding (fraction of active enzymes) and catalytic (k_{cat}) effects of mutations are measured for PafA [32] (SM Sec. 2C [36]).

We find significant correlations between phenotype and AF-predicted ES (compared to wild-type) for all phenotypes (Fig. 2). It is possible to predict blue, red and green fluorescence (Pearson's $r = -0.93$, $r = -0.76$, $r = -0.67$) by measuring the ES at residues Y65, A218 and L59, respectively (Fig. 2C,E,F). There are many residues at which deformation measured by ES is predictive of fluorescence (Fig. 2A-B), and these residues are found to be closer to residue Y65 (Fig. 2D), which covalently binds to a chromophore. We also find weaker, yet significant correlations between ES and the empirical effects of mutations on folding and catalytic activity (Fig. 2G-H). For catalytic activity, we measure mean deformation at the active site; for the folding effect, we measure mean ES between the 50 residues that correlate best with the folding effect (SM Sec. 9 [36]).

In contrast, we do not find consistent correlations with RMSD, a standard estimate of AF accuracy[12], indicating that local deformation, as measured by the ES, is more appropriate for measuring mutational effects (SM Sec. 10). In some cases, performance is heavily dependent on which pre-trained model (SM Fig. 4 [36]) is used: surprisingly, we found that using the highest ranked (by pLDDT; see SM Sec. 3 & 11 [36]) models resulted in worse performance for phenotypic change (SM Fig. 4 [36]), and average performance for structural change (SM Sec. 12 [36]). Taken together, these results provide evidence that AF can be used to predict the structural effect of a single mutation.

ES correlates with phenotypic change for wild-type proteins.— It is quite unexpected that the ES S_i should be a good predictor of phenotypic change, even if AF can accurately predict structure. We suspect that the correlation is strong because the structures are always compared to the wild-type proteins, where the structure is adapted for function through evolution – any deviation from this optimal structure is likely to diminish protein function. To test this hypothesis, we calculate the correlation between S and phenotype change across all possible pairs of proteins (*i.e.*, comparing mutants with mutants). We find that the correlation is much weaker, or completely disappeared (SM Sec. 9D [36]) for the divergent fluorescent proteins, but unchanged for PafA sequences (which are all within $M \leq 2$ of WT). Thus we conclude that S is a good predictor of phenotypic change from native protein sequences. For studying phenotypic change away from optima in phenotype landscapes, a more sophisticated mapping from structure to function is needed.

Discussion.— We have shown that AF is capable of predicting structures with sufficient accuracy and that it can pick up changes as small as those resulting from a single missense mutation. Direct validation of predicted mutational effects on structure is limited by the accuracy of empirical structures (Fig. 1F), and further hindered by the lack of sequence pairs that are suitable for comparison (SM Sec. 1 [36]). Likewise, predicting phenotypic change from structure alone ought to be challenging, to say the least. Despite these steep hurdles, we have shown, using effective strain (ES) as a measure of deformation, that differences between AF-predicted structures do correlate with both structural (Fig. 1) and phenotypic changes (Fig. 2) in empirical data. Examining individual pairs of PDB structures, mutation effects are masked by fluctuations, but this inherent noise is filtered by analyzing the statistics of many pairs, demonstrating that AF is accurate. The difficulties in assembling sufficient data for validation highlight that the age of experimental protein structure identification is far from over [38], despite the success of AF and RoseTTAFold [12, 39].

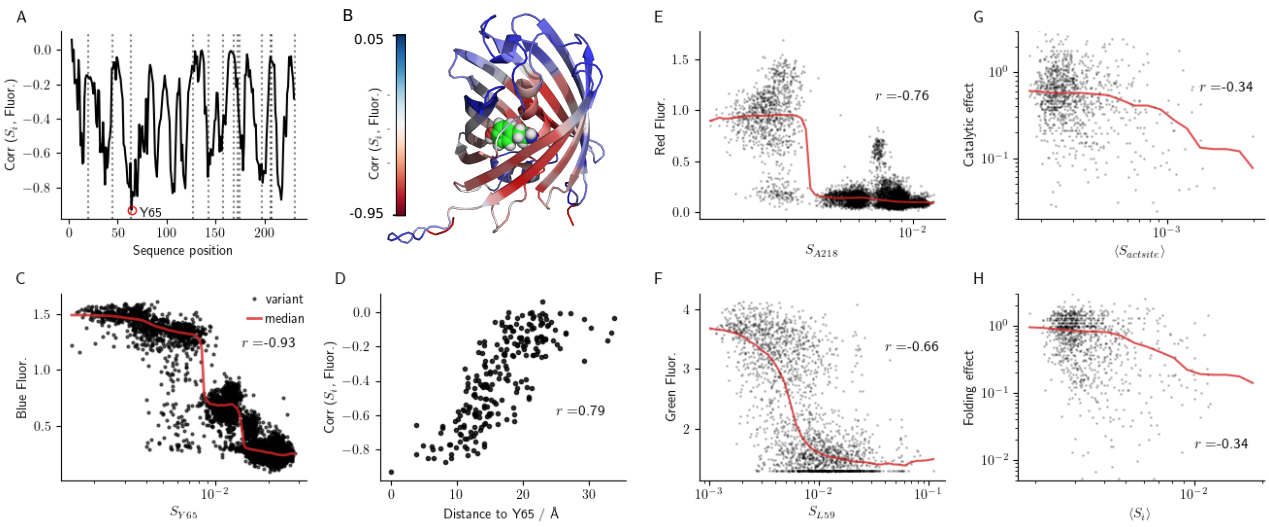

FIG. 2. A: Correlation (Pearson's r) between blue fluorescence (mTagBFP2) and AF-predicted effective strain (ES), S_i, between WT and 8,191 variants for all sequence positions i; positions of mutated residues are shown by dotted lines; chromophore site (Y65) is indicated (red circle). B: Structure of BFP, with each residue colored according to corr(S_i, Fluor.) (A); Y65 atoms are shown as spheres. C: Strain at residue Y65 vs. fluorescence for mTagBFP2 variants. D: Fluorescence-strain correlation per residue vs. distance from residue i to Y65. E-H: Correlations between: S_{A218} and red fluorescence (mKate2); S_{L59} and green fluorescence (GFP); catalytic activity and S at the active site (PafA); folding ability (fraction of active enzymes) and average strain, $\langle S_i \rangle$, of the 50 residues that correlate best with folding ability (PafA).

Advice for using AF to study mutations.— We find higher correlations between AF and PDB when mutations are in less flexible regions of proteins, and when mutations have large effects (App. C). One can quickly estimate flexibility using pLDDT (AF's confidence in a residue's predicted position, or a proxy measure of rigidity; SM Sec. 11 [36]), but it is more robust to measure the variance of AF predictions by predicting multiple structures (App. B). Depending on the flexibility, and mutation effect size, one can achieve much more reliable estimates of mutation effects by averaging across many repeat structures (SM Sec. 6 [36]). We advise against using templates in predictions (used by default in AF models 1 and 2), since this appears to offer at best negligible increases in accuracy, and we found one example where including templates made the predictions much worse (SM Fig. 4 [36]).

AF predicts structure, not folding.— We need to emphasise that AF is only trained to predict structures of stable proteins, and we make no claims about whether the proteins will indeed fold into the predicted structure. Given the marginal stability of most proteins, mutations may easily destabilize a protein so that its melting temperature falls below room temperature. The process of protein folding is carefully tuned *in vivo* for folding on the ribosome, and through interactions with chaperones, and mutations that do not change structure may retard folding through other mechanisms. To see whether pLDDT is predictive of whether a protein will fold or not, we studied a set of 147 WW-domain-like sequences, of which 40 were found to fold *in vitro*. Although more sophisticated methods may perform better, mean pLDDT by itself proved insufficient to sort folding from non-folding proteins (SM Sec. 11B [36]). Now that one question – what structure will a protein likely fold into? – has been seemingly solved, at least partially, it is crucial to next answer the question of *whether* a protein will spontaneously fold.

Local deformation should be used to measure mutation effects.— Placing the current results in a broader context, we note that the evidence in support of AF's capacity to predict the effect of a mutation has so far been mixed. Some studies suggest that AF and RoseTTTaFold can be indirectly used to predict phenotype, but not by comparing structures. [25–27] Two studies have reported negative results [23, 24], which we attribute primarily to their use of pLDDT and RMSD, which are much less precise measures of mutational effects compared to strain (SM Sec. 10-11 [36]). In one study, the authors found only weak correlations between pLDDT and fluorescence using the same GFP dataset used here. Although we do not expect pLDDT to strongly correlate with fluorescence, we do find higher correlations than those reported in [23] (SM Sec. 11A [36]). In another analysis [24], the authors appear to assume that structure-disrupting mutations should result in a large change in predicted structure or pLDDT [24]. We first note that this paper only studied three proteins, limiting our ability to draw general conclusions. We also see that the deformation due to mutations in one of these proteins is higher than 95 % of mutation effects in our PDB sample (SM Sec. 13 [36]); it is possible that such large deformations are predictive of destabilizing mutations, and testing this is a promising future direction. Ultimately, we think the present study has demonstrated that deformation (measured by SE) is a more robust measure of structural change upon mutation.

In summary, we showed here that AF predictions of local structural change can be used to study missense mutations in proteins. These analyses suggest that AF can, indeed, be a powerful tool, if used in the right context and backed up by appropriate analyses. Using AF, we can bridge the gap between sequence and function

in high-throughput deep-mutational scan experiments, guide directed evolution studies [9], and design drugs *in silico* [10]. For example, on a smaller scale, AF can be used to screen potential mutants, and in costly experiments where the number of mutations is limited, one can select mutations with strong or weak effects in desired regions of the protein. Overall, it appears that AF provides a step change in our ability to study and guide protein evolution.

We thank Jacques Rougemont and Jean-Pierre Eckmann for discussions and for providing code to calculate shear. This work was supported by the Institute for Basic Science, Project Code IBS-R020-D1.

Appendix A: Calculating local deformation.— As a measure of local deformation, we compute the effective strain (ES), S_i. ES is simply the mean relative change of the interparticle distances around a given residue and is partially correlated with shear strain (SM Sec. 4-5 [36]). To calculate S_i per residue i, we first define a neighborhood N_i that includes the $n_i = |N_i|$ residues $j \in N_i$ whose C_α positions $\mathbf{r}_j$ are within 13 Å of $\mathbf{r}_i$, the C_α position of residue i. We obtain a $3 \times n_i$ *neighborhood* tensor $\mathbf{D}_i$ whose n_i rows are the distance vectors, $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$. We calculate, respectively, $\mathbf{D}_i$ and $\mathbf{D}'_i$ for the two structures we are comparing (*e.g.*, wild-type and mutant), and rotate $\mathbf{D}'_i$ to maximize overlap between the tensors. The ES is the average over the n_i neighbors of the relative change in the distance vectors,

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

We have evaluated several other local metrics, similar in nature to ES, finding that the conclusions are not very sensitive to the specific choice of metric or neighborhood cutoff - only to (SM Sec. 4-5 [36]).

Appendix B: Averaging local neighborhoods increases accuracy.— Since AF predictions are stochastic, repeat predictions vary. We find that deformation between repeat predictions of the same protein leads to non-negligible ES (Fig. 1G). The ES is higher in flexible regions, which is indicated by higher B-factor, solvent accessibility (RSA), and lower pLDDT (Fig. 3). It is possible to obtain more reliable estimates of mutation effects by averaging across local neighborhoods, $\mathbf{D}_i$, in repeat predictions (SM Sec. 6 [36]). Our average structures ($\langle$AF$\rangle$) are typically averaged over all 5 AF models, with one set of predictions from DeepMind's AF implementation, and five sets of predictions from ColabFold's AF implementation [40]. Averaging typically increases PDB-AF structure correlations (SM Sec. 6A [36]), and also deformation-phenotype correlations (SM Sec. 6B [36]). One exception is the mTagBFP2/mKate2 dataset, where DeepMind's implementation of AF produces a better correlation than the average; we find that this is due to the ColabFold implementation performing poorly on this specific protein (SM Sec. 6B [36]).

Appendix C: When do AF predictions correlate with PDB data? — Here we assess why AF sometimes predicts mutation effects similar to those measured in experimental structures (Fig. 1G). Across all proteins, AF-PDB correlations are higher for mutant pairs of proteins in two situations

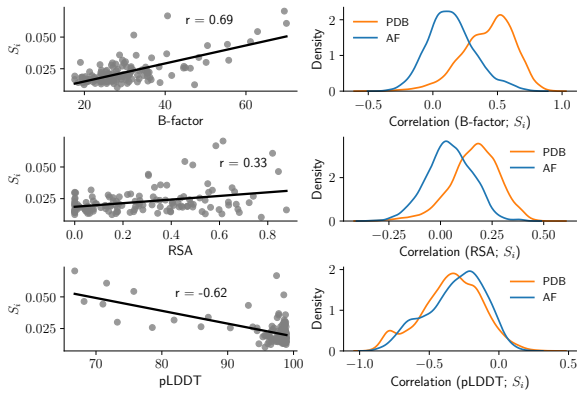

FIG. 3. **Fluctuations are greater in flexible regions.** Deformation (ES) between experimental hen lysozyme structures (`194L_A` and `6RTA_A`), S_i is correlated (Pearson's r) with B-factor, relative solvent accessibility (RSA), and pLDDT (left). Distributions (kernel density estimates) of correlations for all proteins (right).

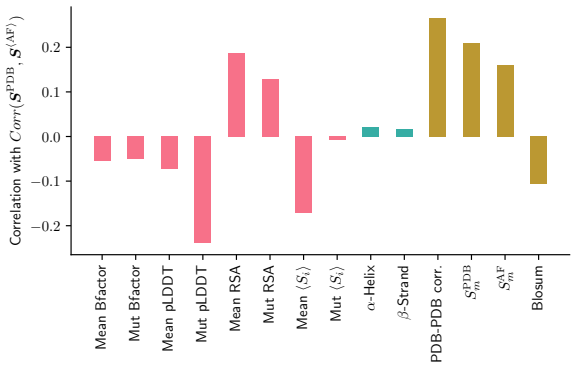

FIG. 4. **Correlations are higher if mutations have large effects in rigid regions.** A: Pearson's correlation between PDB-AF **S**-correlation and: mean and mutated residue values of flexibility (B-factor, RSA, pLDDT, $\langle S_i \rangle$); fraction of secondary structure (α-helix or β-sheet); magnitude of mutation effect (PDB-PDB **S** correlation, ES at mutated site in PDB and AF, S_m^{PDB} and S_m^{AF}, BLOSUM score).

(Fig. 4, SM Sec. 8 [36]): when flexibility is low (low B-factor, low RSA, high pLDDT, high repeat ES across repeat predictions $\langle S_i \rangle$); and when mutations have large effects that are easier to measure (high PDB-PDB correlation, high deformation at mutated site S_m, BLOSUM score). There was no significant effect of the secondary structure type.

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

[1] N. Tokuriki and D. S. Tawfik, Stability effects of mutations and protein evolvability, Curr. Opin. Struc. Biol. **19**, 596 (2009).

[2] M. Lorch, J. M. Mason, R. B. Sessions, and A. R. Clarke, Effects of mutations on the thermodynamics of a protein folding reaction: Implications for the mechanism of formation of the intermediate and transition states, Biochemistry-us. **39**, 3480 (2000).

[3] M. Zhang, D. A. Case, and J. W. Peng, Propagated perturbations from a peripheral mutation show interactions support-

ing ww domain thermostability, Structure **26**, 1474 (2018).

[4] G. Yang, N. Hong, F. Baier, C. J. Jackson, and N. Tokuriki, Conformational tinkering drives evolution of a promiscuous activity through indirect mutational effects, Biochemistry-us. **55**, 4583 (2016).

[5] N. Sahni, S. Yi, M. Taipale, J. Fuxman Bass, J. Coulombe-Huntington, F. Yang, J. Peng, J. Weile, G. Karras, Y. Wang, I. Kovács, A. Kamburov, I. Krykbaeva, M. Lam, G. Tucker, V. Khurana, A. Sharma, Y.-Y. Liu, N. Yachie, Q. Zhong, Y. Shen, A. Palagi, A. San-Miguel, C. Fan, D. Balcha, A. Dricot, D. Jordan, J. Walsh, A. Shah, X. Yang, A. Stoyanova, A. Leighton, M. Calderwood, Y. Jacob, M. Cusick, K. Salehi-Ashtiani, L. Whitesell, S. Sunyaev, B. Berger, A.-L. Barabási, B. Charloteaux, D. Hill, T. Hao, F. Roth, Y. Xia, A. Walhout, S. Lindquist, and M. Vidal, Widespread macromolecular interaction perturbations in human genetic disorders, Cell **161**, 647 (2015).

[6] V. M. Prabantu, N. Naveenkumar, and N. Srinivasan, Influence of disease-causing mutations on protein structural networks, Frontiers in Molecular Biosciences **7**, 10.3389/fmolb.2020.620554 (2021).

[7] H. Jacquier, A. Birgy, H. L. Nagard, Y. Mechulam, E. Schmitt, J. Glodt, B. Bercot, E. Petit, J. Poulain, G. Barnaud, P.-A. Gros, and O. Tenaillon, Capturing the mutational landscape of the beta-lactamase tem-1, P. Natl. Acad. Sci. Usa. **110**, 13067 (2013).

[8] A. T. Albanaz, C. H. Rodrigues, D. E. Pires, and D. B. Ascher, Combating mutations in genetic disease and drug resistance: Understanding molecular mechanisms to guide drug design, Expert Opin. Drug Dis. **12**, 553 (2017).

[9] F. H. Arnold, Design by directed evolution, Accounts Chem. Res. **31**, 125 (1998).

[10] L. Cao, B. Coventry, I. Goreshnik, B. Huang, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E. M. Strauch, N. D. DeBouver, A. Pires, A. K. Bera, S. Halabiya, B. Hammerson, W. Yang, S. Bernard, L. Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, and D. Baker, Design of protein binding proteins from target structure alone, Nature 10.1038/s41586-022-04654-9 (2022).

[11] A. Stein, D. M. Fowler, R. Hartmann-Petersen, and K. Lindorff-Larsen, Biophysical and mechanistic models for disease-causing protein variants, Trends Biochem. Sci. **44**, 575 (2019).

[12] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, Highly accurate protein structure prediction with alphafold, Nature **596**, 583 (2021).

[13] S. J. Fleishman and A. Horovitz, Extending the new generation of structure predictors to account for dynamics and allostery, J. Mol. Biol. **433**, 167007 (2021).

[14] A. S. Morgunov, K. L. Saar, M. Vendruscolo, and T. P. J. Knowles, New frontiers for machine learning in protein science, J. Mol. Biol. **433**, 167232 (2021).

[15] G. Masrati, M. Landau, N. Ben-Tal, A. Lupas, M. Kosloff, and J. Kosinski, Integrative structural biology in the era of accurate structure prediction, J. Mol. Biol. **433**, 167127 (2021).

[16] M. Andrec, D. A. Snyder, Z. Zhou, J. Young, G. T. Montelione, and R. M. Levy, A large data set comparison of protein structures determined by crystallography and nmr: Statistical test for structural differences and the effect of crystal packing, Proteins. **69**, 449 (2007).

[17] A. A. Rashin, A. H. L. Rashin, and R. L. Jernigan, Protein Flexibility: Coordinate Uncertainties and Interpretation of Structural Differences, Acta Crystallogr. D. **65**, 1140 (2009).

[18] B. Venkatakrishnan, M. L. Palii, M. Agbandje-McKenna, and R. McKenna, Mining the protein data bank to differentiate error from structural variation in clustered static structures: An examination of hiv protease, Viruses **4**, 348 (2012).

[19] J. K. Everett, R. Tejero, S. B. K. Murthy, T. B. Acton, J. M. Aramini, M. C. Baran, J. Benach, J. R. Cort, A. Eletsky, F. Forouhar, R. Guan, A. P. Kuzin, H. W. Lee, G. Liu, R. Mani, B. Mao, J. L. Mills, A. F. Montelione, K. Pederson, R. Powers, T. Ramelot, P. Rossi, J. Seetharaman, D. Snyder, G. V. T. Swapna, S. M. Vorobiev, Y. Wu, R. Xiao, Y. Yang, C. H. Arrowsmith, J. F. Hunt, M. A. Kennedy, J. H. Prestegard, T. Szyperski, L. Tong, and G. T. Montelione, A community resource of experimental data for nmr / x-ray crystal structure pairs, Protein Sci. **25**, 30 (2016).

[20] G. L. Hura, C. D. Hodge, D. Rosenberg, D. Guzenko, J. M. Duarte, B. Monastyrskyy, S. Grudinin, A. Kryshtafovych, J. A. Tainer, K. Fidelis, and S. E. Tsutakawa, Small angle x-ray scattering-assisted protein structure prediction in casp13 and emergence of solution structure differences, Proteins. **87**, 1298 (2019).

[21] M. L. Lynch, M. F. Dudek, and S. E. J. Bowman, A searchable database of crystallization cocktails in the pdb: Analyzing the chemical condition space, Patterns **1**, 100024 (2020).

[22] N. J. Fowler, A. Sljoka, and M. P. Williamson, The accuracy of nmr protein structures in the protein data bank, Structure **29**, 1430 (2021).

[23] M. A. Pak, K. A. Markhieva, M. S. Novikova, D. S. Petrov, I. S. Vorobyev, E. S. Maksimova, F. A. Kondrashov, and D. N. Ivankov, Using alphafold to predict the impact of single mutations on protein stability and function, PLOS ONE **18**, 1 (2023).

[24] G. R. Buel and K. J. Walters, Can alphafold2 predict the impact of missense mutations on structure?, Nature Structural & Molecular Biology **29**, 1 (2022).

[25] M. Akdel, D. E. V. Pires, E. P. Pardo, J. Jänes, A. O. Zalevsky, B. Mészáros, P. Bryant, L. L. Good, R. A. Laskowski, G. Pozzati, A. Shenoy, W. Zhu, P. Kundrotas, V. R. Serra, C. H. M. Rodrigues, A. S. Dunham, D. Burke, N. Borkakoti, S. Velankar, A. Frost, J. Basquin, K. Lindorff-Larsen, A. Bateman, A. V. Kajava, A. Valencia, S. Ovchinnikov, J. Durairaj, D. B. Ascher, J. M. Thornton, N. E. Davey, A. Stein, A. Elofsson, T. I. Croll, and P. Beltrao, A structural biology community assessment of alphafold2 applications, Nature Structural & Molecular Biology **29**, 1056 (2022).

[26] Y. Zhang, P. Li, F. Pan, H. Liu, P. Hong, X. Liu, and J. Zhang, Applications of alphafold beyond protein structure prediction, bioRxiv 10.1101/2021.11.03.467194 (2021).

[27] S. Mansoor, M. Baek, D. Juergens, J. L. Watson, and D. Baker, Accurate mutation effect prediction using rosettafold, bioRxiv 10.1101/2022.11.04.515218 (2022).

[28] J. P. Roney and S. Ovchinnikov, State-of-the-art estimation of protein model accuracy using alphafold, Phys. Rev. Lett. **129**, 238101 (2022).

[29] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, The Protein Data Bank, Nucleic Acids Res. **28**, 235 (2000).

[30] K. S. Sarkisyan, D. A. Bolotin, M. V. Meer, D. R. Usmanova, A. S. Mishin, G. V. Sharonov, D. N. Ivankov, N. G. Bozhanova, M. S. Baranov, O. Soylemez, N. S. Bogatyreva, P. K. Vlasov, E. S. Egorov, M. D. Logacheva, A. S. Kondrashov, D. M. Chudakov, E. V. Putintseva, I. Z. Mamedov, D. S. Tawfik, K. A. Lukyanov, and F. A. Kondrashov, Local fitness landscape of the green fluorescent protein, Nature **533**, 397 (2016).

[31] F. J. Poelwijk, M. Socolich, and R. Ranganathan, Learning the pattern of epistasis linking genotype and phenotype in a

protein, Nat. Commun. **10**, 4213 (2019).

[32] C. J. Markin, D. A. Mokhtari, F. Sunden, M. J. Appel, E. Akiva, S. A. Longwell, C. Sabatti, D. Herschlag, and P. M. Fordyce, Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics, Science **373**, eabf8761 (2021).

[33] C. W. Hespen, J. J. Bruegger, Y. Guo, and M. A. Marletta, Native alanine substitution in the glycine hinge modulates conformational flexibility of heme nitric oxide/oxygen (h-nox) sensing proteins, Acs Chem. Biol. **13**, 1631 (2018).

[34] M. R. Mitchell, T. Tlusty, and S. Leibler, Strain analysis of protein structures and low dimensionality of mechanical allosteric couplings, P. Natl. Acad. Sci. Usa. **113**, E5847 (2016).

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

[36] See supplemental material at url for detailed information about methods used in this letter. (2023).

[37] S. M. Kandathil, J. G. Greener, A. M. Lau, and D. T. Jones, Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins, P. Natl. Acad. Sci. Usa. **119**, e2113348119 (2022).

[38] T. C. Terwilliger, D. Liebschner, T. I. Croll, C. J. Williams, A. J. McCoy, B. K. Poon, P. V. Afonine, R. D. Oeffner, J. S. Richardson, R. J. Read, and P. D. Adams, Alphafold predictions: Great hypotheses but no match for experiment, bioRxiv 10.1101/2022.11.21.517405 (2022).

[39] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, and D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network, Science **373**, 871 (2021).

[40] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, Colabfold: making protein folding accessible to all, Nature Methods **19**, 679 (2022).

Supplemental Material for
"AlphaFold2 can predict single-mutation effects on structure and phenotype"

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 1
- A. Full 1
- B. Differences in crystallographic group 1
- C. Non-redundant sets 2
- D. Structures not used in training AF 2

2.. High-throughput Phenotype Data 2
- A. eqFP611 2
- B. GFP 2
- C. PafA 2

3.. Structure Prediction 2
- A. AlphaFold: DeepMind 2
- B. AlphaFold: ColabFold 2

4.. Deformation metrics 2
- A. Local Distance Difference Test (LDDT) 3
- B. Local Distance Difference (LDD) 4
- C. Neighborhood Distance 4
- D. Shear Strain 4
- E. Non-Affine Strain 5
- F. Effective Strain (ES) 5

5.. Evaluation of deformation metrics 5
- A. Correlations between deformation metrics 5
- B. PDB-AF correlations for different deformation metrics 6
- C. Deformation-Phenotype correlations for different metrics 8

6.. Averaging AF-predicted structures 8
- A. Effect of averaging on PDB-AF correlations 9
- B. Effect of averaging on AF-phenotype correlations 9

7.. Range of mutation effects 10

8.. When do AF predictions correlate with PDB data? 10

9.. Deformation-phenotype correlations 11
- A. mKate2 11
- B. GFP 11
- C. PafA 11
- D. Comparing mutants with mutants 12

10.. RMSD 12

11.. pLDDT 14
- A. Correlations between changes in pLDDT and phenotype 14
- B. Is pLDDT a good predictor of protein folding? 14

12.. AlphaFold Models 16

13.. Scale of mutation effects 16
- A. Deformation between PDB structures is typically small 16
- B. Example of large deformation 16
- C. UBA1 of hHR23a 17

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 very large, comparatively variable deformations; 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 (high repeat-measurement deformation) 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 4,645 PDB structures, and 92,868 pairs; 70,554 of these pairs derive from a set of 485 endothiapepsin structures, so we remove most of these until we are left with 22,821 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 deformation 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, 2,815 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 22,281 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.

B. GFP

The GFP dataset is a subset (2,312 sequences) of the full dataset (51,715 sequences) published in (Sarkisyan *et al.*, 2016). A specific 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, 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.

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; two double mutants at the active site were 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 the deformation, as measured by the ES, S_i, of each variant compared to the WT.

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.

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.

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 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

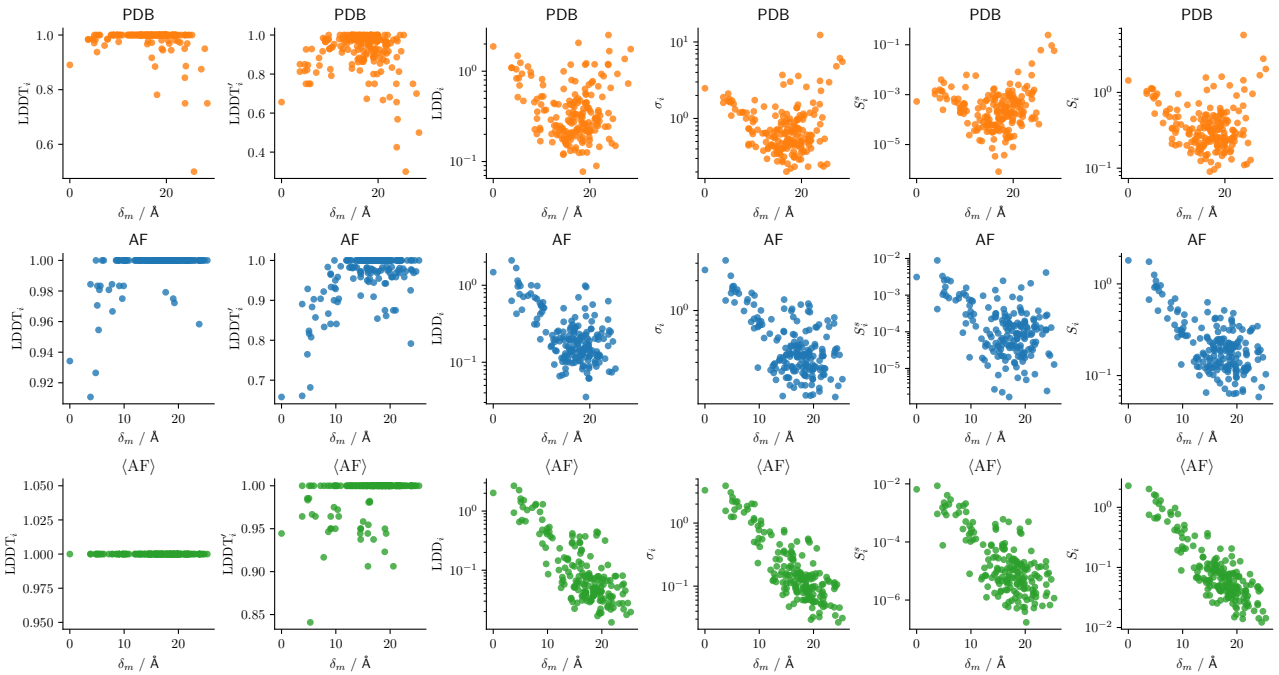

FIG. 1 **Examples of deformation metrics.** Deformation as a function of distance from nearest mutated site for deformation upon mutation from WT CypA (`6U5C_A`) to a double mutant (S99T, C115S, `6BTA_A`). Deformation is shown for 6 metrics, from left to right: LDDT; LDDT′ with more precise cutoffs ζ_k; 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 (bottom, see Sec. 6.).

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 absolutely 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 modifications of LDDT that 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).

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 using a distance cutoff, γ. Residues are considered neighbors if the positions of their C_α atoms, $\mathbf{r}_j$, are closer than γ; *i.e.*, if $r_{ij} = |\mathbf{r}_{ij}| = |r_i - r_j| \leq \gamma$. 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.

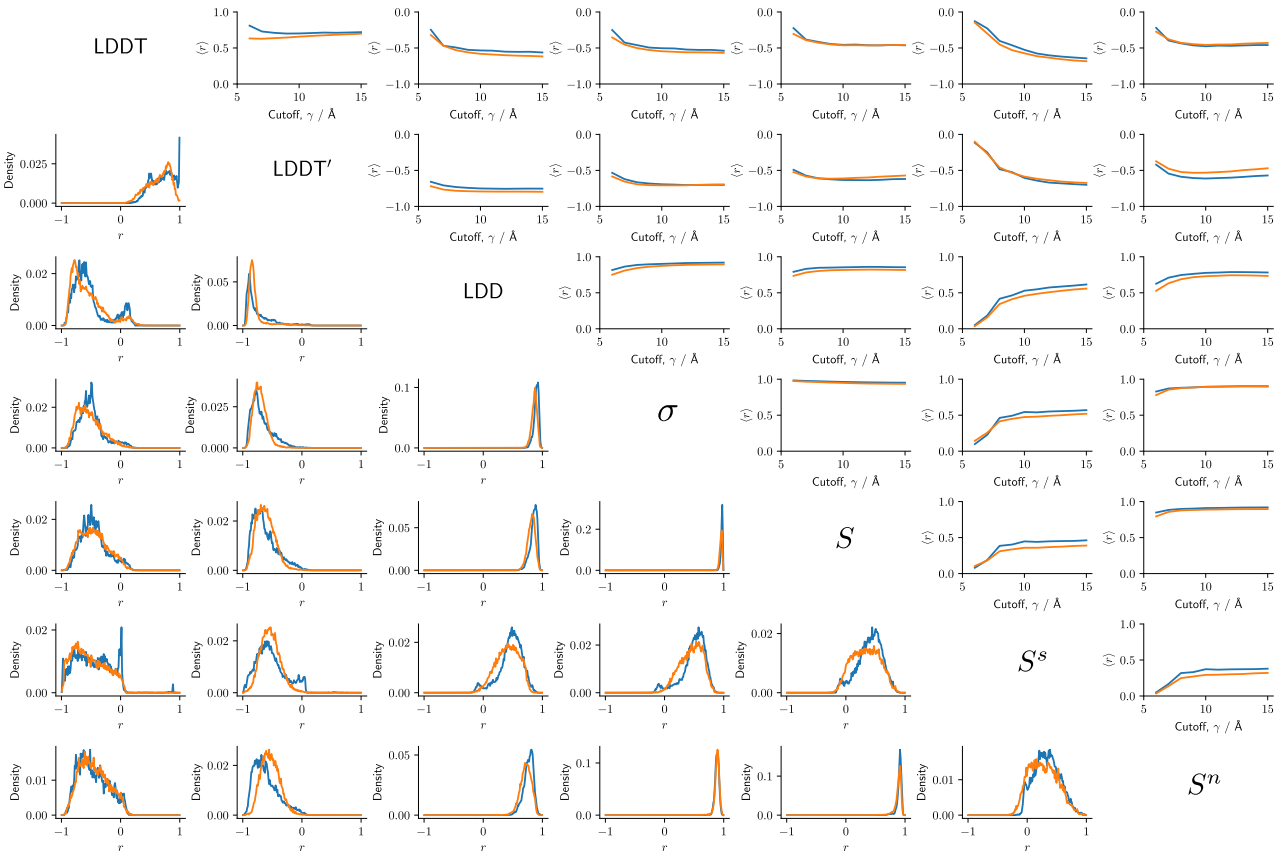

FIG. 2 **Correlations between deformation metrics.** Pearson's correlation coefficient, r, is calculated for each pair of proteins in our PDB dataset (22,281 pairs). Distribution of r for each combination of metrics (Bottom, $\gamma = 13$ Å), and mean correlation $\langle r \rangle$ as a function of γ (Top). Data is shown for both PDB (orange) and AF-predicted (blue) structures.

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 root-mean-squared distance change,

$$\mathrm{LDD}_i = \frac{1}{n_i}\sqrt{\sum_{j \in N_i} (\Delta r_{ij})^2}\,. \tag{3}$$

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 $3 \times n_i$ 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 average norm of the change in distances, $\Delta\mathbf{r}_{ij} = \mathbf{r}'_{ij} - \mathbf{r}_{ij}$, between neighborhoods.

$$\sigma_i = \frac{1}{n_i}\left\|\mathbf{D}_i - \mathbf{D}'_i\right\| = \frac{1}{n_i}\sqrt{\sum_{j \in N_i} \left|\Delta\mathbf{r}_{ij}\right|^2}\,. \tag{4}$$

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 differs from frame-aligned-point-error (FAPE) – which forms part of the loss function used by DeepMind to train AlphaFold (Jumper *et al.*, 2021) – only 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. To calculate strain at residue i, we first calculate the deformation gradient tensor, $\mathbf{F}_i$, where

$$\mathbf{D}'_i = \mathbf{F}_i\,\mathbf{D}_i\,,$$

and then get the Lagrangian finite strain tensor, $\mathbf{E}_i$,

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

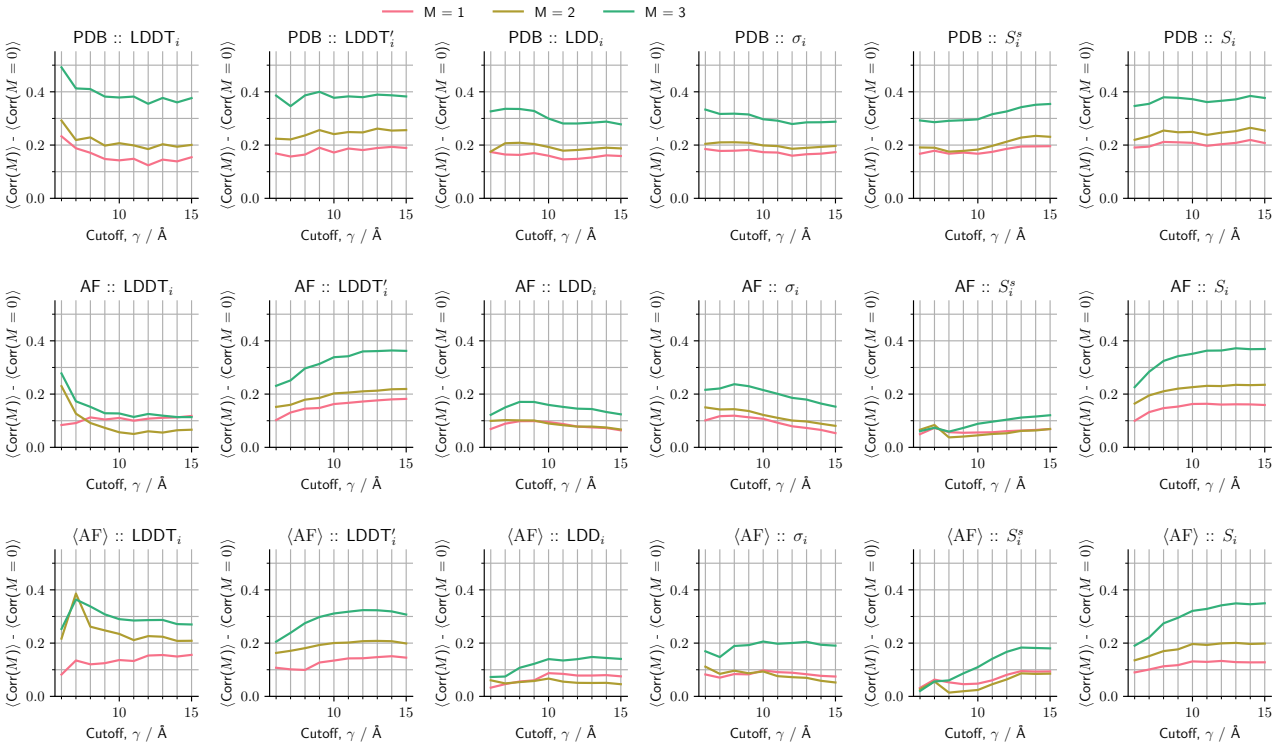

FIG. 3 **Correlations between PDB and PDB (top), AF (middle), and ⟨AF⟩ deformation vectors for different metrics.** Residual correlation due to mutation, $\langle \mathrm{corr}(M)\rangle - \langle \mathrm{corr}(M=0)\rangle$, extracted from the distributions of correlation between PDB and AF-predicted 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).

where $\mathbf{I}$ is the identity matrix. We find $\mathbf{F}_i$ using least-squares regression. Finally, as a measure of the magnitude of the shear strain, we use

$$S_i^{\mathrm{s}} = \mathrm{Tr}(\mathbf{E}_i \cdot \mathbf{E}_i) - \frac{1}{3}[\mathrm{Tr}(\mathbf{E}_i)]^2 \,. \tag{5}$$

In practice, we follow the implementation given in (Eckmann *et al.*, 2019).

E. Non-Affine Strain

To measure non-affine strain – deformation that is not due to isotropic volume expansion/contraction, or shear/twist motion – we measure the residual of the fit of the deformation gradient tensor, $\mathbf{F}_i$, summed over all neighbors $j \in N_i$,

$$S_i^{\mathrm{n}} = \sum_{j \in N_i} \left| \mathbf{r}'_{ij} - \mathbf{F}_i\, \mathbf{r}_{ij} \right|^2 , \tag{6}$$

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*,

$$S_i = \left\langle \frac{|\Delta \mathbf{r}_{ij}|}{|\mathbf{r}_{ij}|} \right\rangle = \frac{1}{n_i} \sum_{j \in N_i} \frac{\left|\mathbf{r}_{ij} - \mathbf{r}'_{ij}\right|}{|\mathbf{r}_{ij}|} \,. \tag{7}$$

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 $\gamma \in \{6, 7, \ldots, 16\}$ Å, for both PDB and AF-predicted pairs of structures in our set of 22,281 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. 16). 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

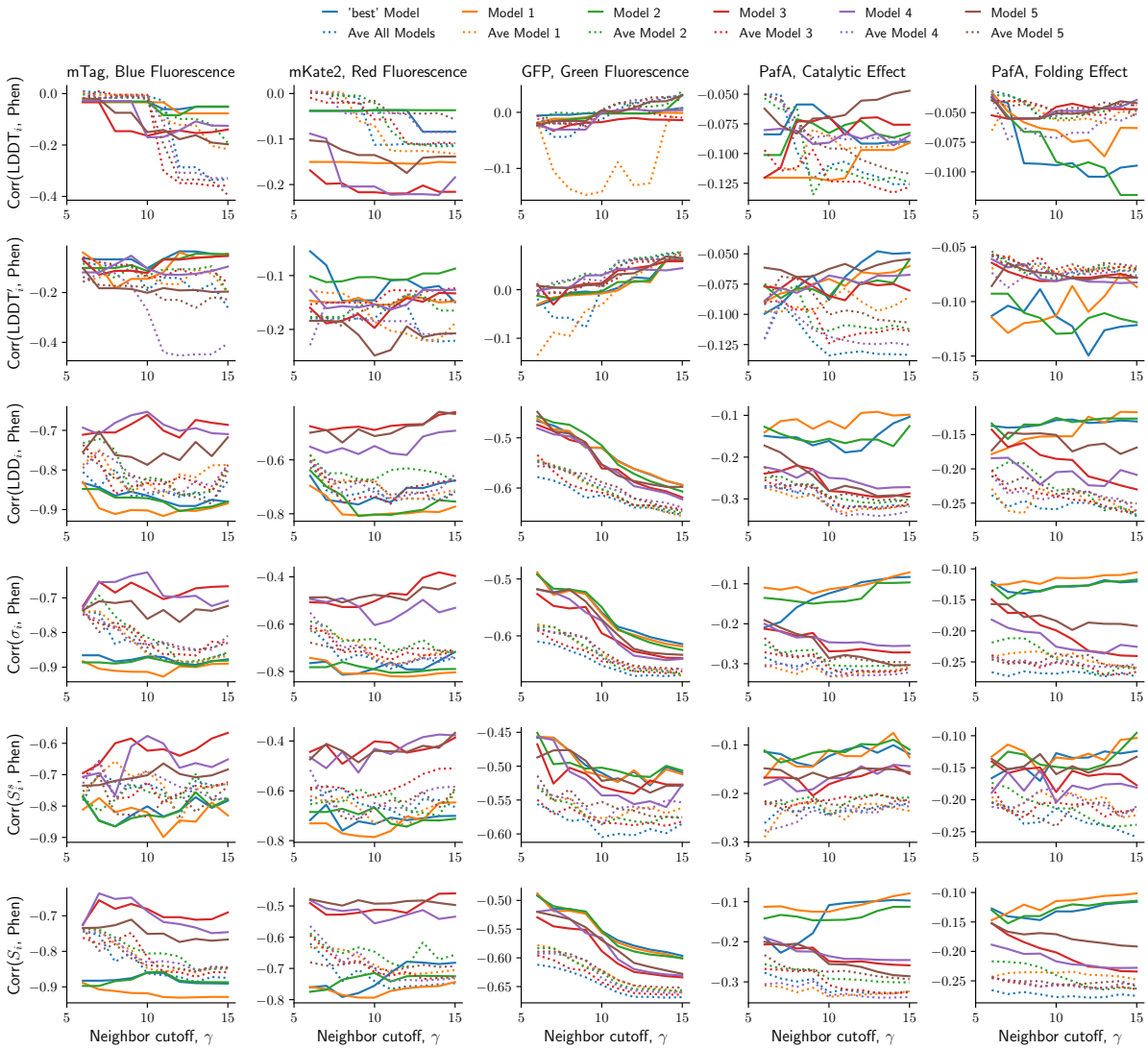

FIG. 4 **Correlations between deformation metrics and phenotype measurements.** Correlations between deformation at residue i and phenotype for 5 sets of phenotype measurements (left to right: mTag, blue fluorescence; mTag, red fluorescence; GFP, green fluorescence; PafA, catalytic effect; PafA, folding 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.

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.

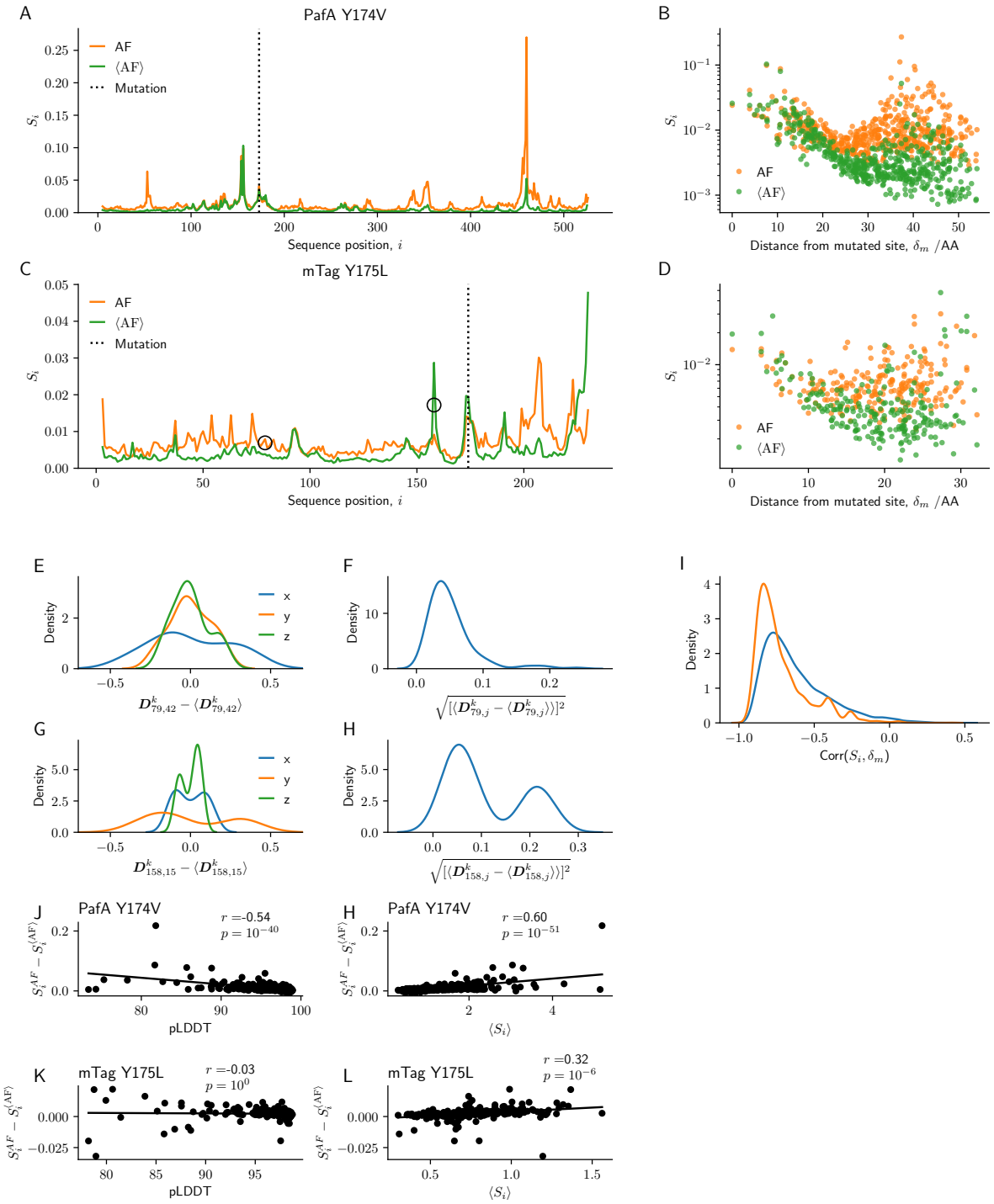

FIG. 5 **Averaging across multiple AF predictions.** A-D: Examples of strain per residue S_i for PafA (A: WT vs Y174V) and mTag (C: WT vs I175L) calculated using single AF structures (AF) and averaged structures ($\langle$AF$\rangle$). Dotted line indicates the mutated residue; circles indicate residues 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).

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, $\langle \mathrm{corr}(M) \rangle - \langle \mathrm{corr}(M = 0) \rangle$, 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 $= 1$.

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. 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^{s} is only slightly worse than effective strain S_i.

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 sequence, 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 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 $\langle$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. 5A-D). In Fig. 5A, there are many high-strain regions that are far away from the mutated site when using the non-averaged structures (Fig. 5B). 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. 5B). 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. 5I), 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. 5C. 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. 5G). 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. 5H 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. 5E), but there are far fewer of these overall (Fig. 5F). 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

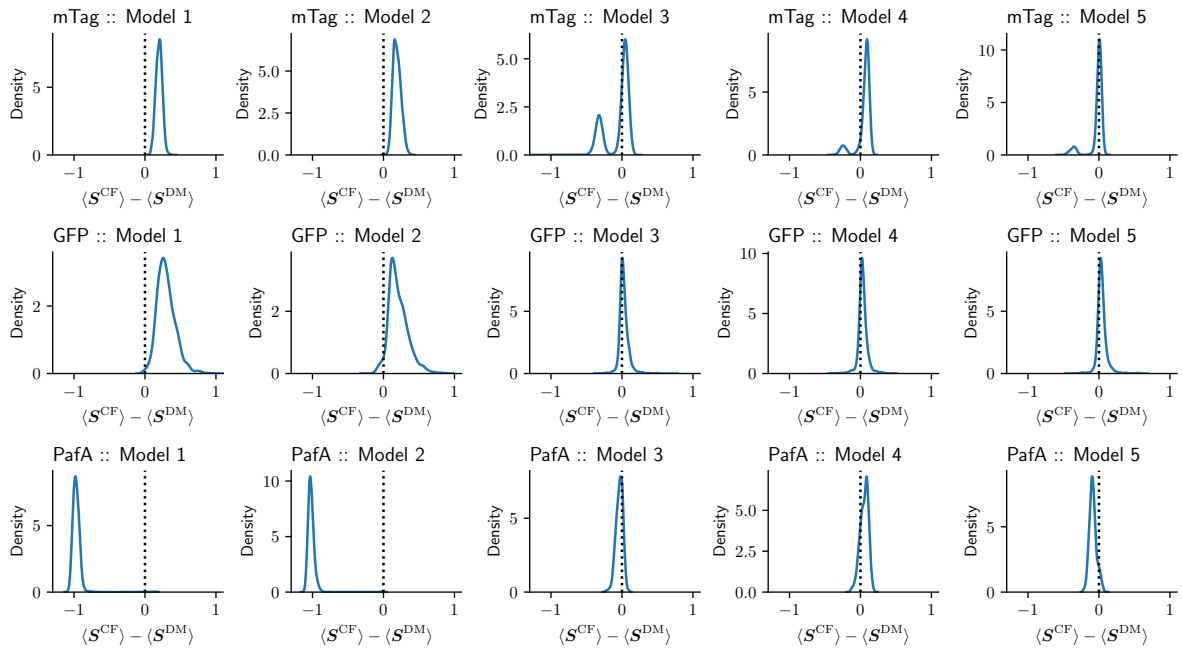

FIG. 6 **Comparing DeepMind and ColabFold implementations of AF.** Difference in the average effective strain (ES) (with $\gamma = 13$ Å) per pair of WT and mutant structure, $\langle S_i \rangle$, for structures predicted using both DeepMind, $\langle S_i^{\mathrm{DM}} \rangle$, and ColabFold, $\langle S_i^{\mathrm{CF}} \rangle$, implementations. Results are shown for each high-throughput dataset (mTag, GFP, PafA), and for each AF model.

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. 5A,C), we compare pLDDT with the change in deformation after averaging, $S_i^{\mathrm{AF}} - S_i^{\langle \mathrm{AF} \rangle}$ (Fig. 5J,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-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 find a modest increase of about ~ 0.05 for the best deformation metrics (LDD, σ_i, and S_e).

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. 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 cal-

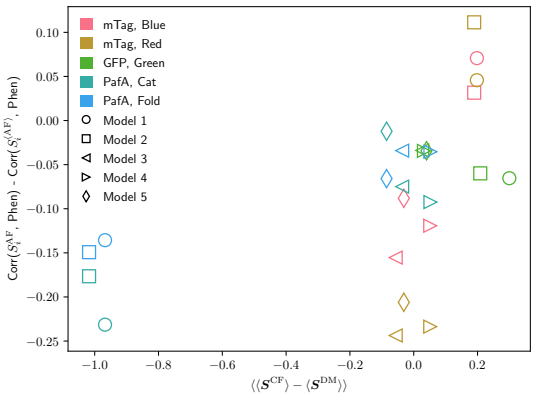

FIG. 7 **Differences between AF implementations correlate with differences in AF vs ⟨AF⟩ accuracy in predicting phenotype.** Average difference in the average strain per pair of WT and mutant structure, $\langle S_e \rangle$, for DeepMind (DM) and ColabFold (CF) implementations , against the difference 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 ($r = 0.51$, $p = 0.019$).

culate 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. 6. 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. We also see that DeepMind models 1, 2 and 3 occasionally produced high strain for mTag variants.

We then compared the average differences between implementations, $\langle \langle \mathbf{S}^{\mathrm{CF}} \rangle - \langle \mathbf{S}^{\mathrm{DM}} \rangle \rangle$, to the differences between strain-phenotype correlations calculated using non-averaged (AF) and averaged (⟨AF⟩) structures. We find a significant correlation between these differences (Fig. 7), 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.

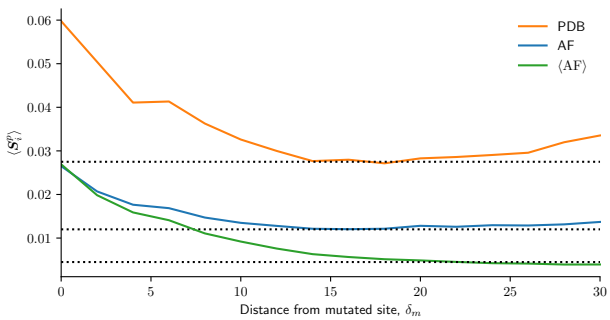

FIG. 8 **Average range of mutation effects.** Average deformation, $\langle S_i^p \rangle$, across all protein pairs p and all residues i within a certain distance from the nearest mutated site δ_m, as a function of δ_m, calculated respectively using PDB, AF and ⟨AF⟩ pairs. Bins of size 2 Å were used. Dotted lines are visual guides.

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, $\langle S_i^p \rangle$, across all mutated ($M > 0$) pairs of proteins p, and all residues i within δ_m bins of size 2 Å (Fig. 8). 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 ⟨AF⟩ pairs the range is about 18 Å. Note that Fig. 8 merely shows the average range of a mutation effect; 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

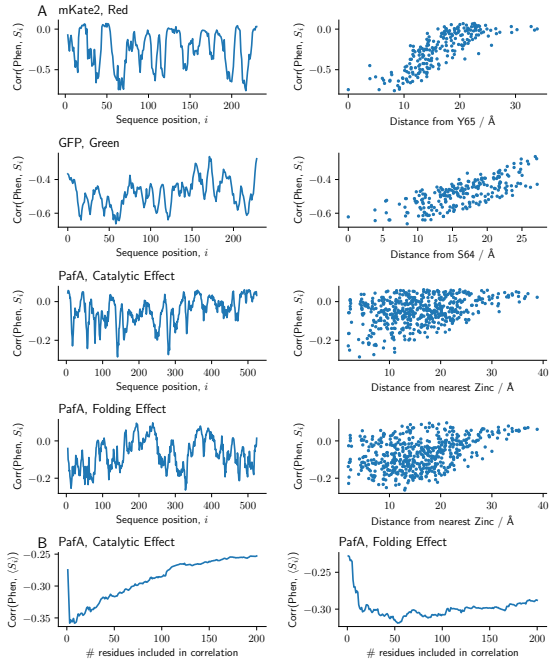

FIG. 9 A: (Left) Correlation between deformation at residue i and phenotype, for all sequence positions i: mKate2, red fluorescence; GFP, green fluorescence; PafA, catalytic effect; PafA, folding effect. (Right) Correlation between deformation at residue i and phenotype as a function of distance from functional residues / co-factors: mKate2, residue Y65; GFP, residue Y58; PafA, distance from the nearest Zinc co-factor. B: Correlation between phenotype and the 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.

surface that is accessible to solvent, compared to the amino acid's total surface area. Residues with higher RSA are less 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).

- Magnitude of mutation effect

  – **PDB-PDB Correlation:** Some protein pairs tend to have more reliable corrections between their deformation vector, and deformation vectors of other matched pairs. This could be because the mutation(s) has a strong effect, 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.

9. Deformation-phenotype correlations

A. mKate2

We calculate deformation of all variants with respect to WT mKate2 (*i.e.* the ES, S_i), 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. 9A). 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. 9A). 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 GFP, 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. 9A). 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 sum of deformation values amongst the top λ residues that correlate best with the phenotype (Fig. 9B). In the case of the catalytic effect, we take the first $\lambda = 5$ residues, while for the folding effect, we take $\lambda = 50$.

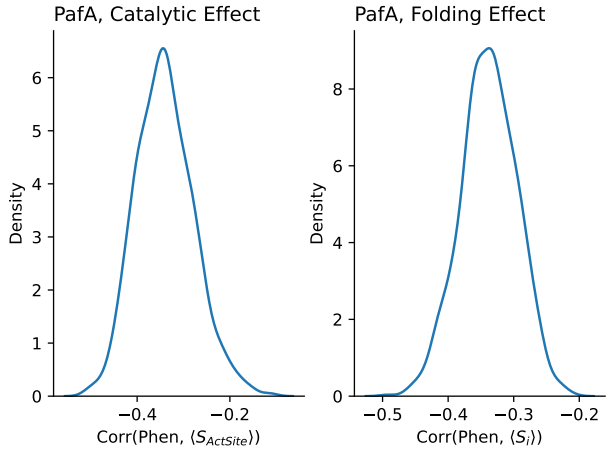

FIG. 10 Distribution of Pearson's correlation between phenotype and deformation obtained by randomly sub-sampling with replacement. Sub-sample size is half of the original sample size; 1,000 sub-samples were drawn.

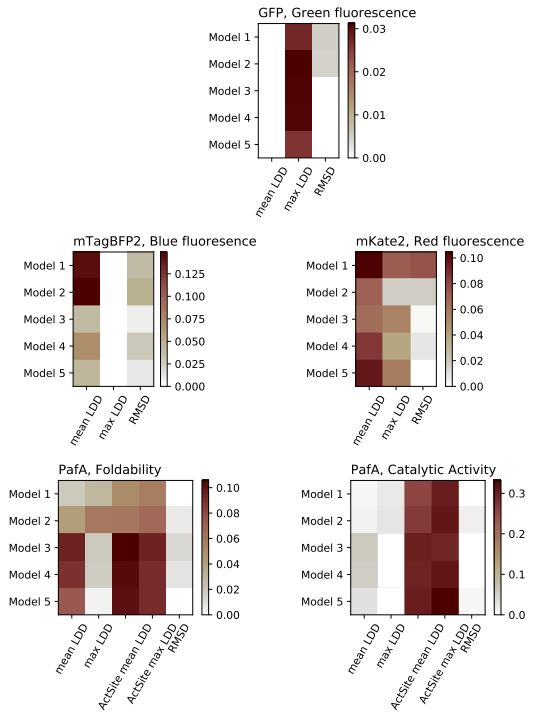

FIG. 11 Correlation (Pearson's r) between structural change and the magnitude of relative phenotypic change between two sequences, i and j: $|\log(\phi_i/\phi_j)|$, where ϕ_i is the relevant phenotype for sequence i. Correlations are shown for all AF models, and for several measures of structural change: mean LDD, max LDD, mean LDD at the active site, max LDD at the active site, and RMSD. For each data set with N sequences, correlations are obtained by comparing all possible $N(N-1)/2$ pairs of sequences.

Correlations are robust to the choice of metric (Fig. 4) and robust to subsampling (Fig. 10).

D. 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

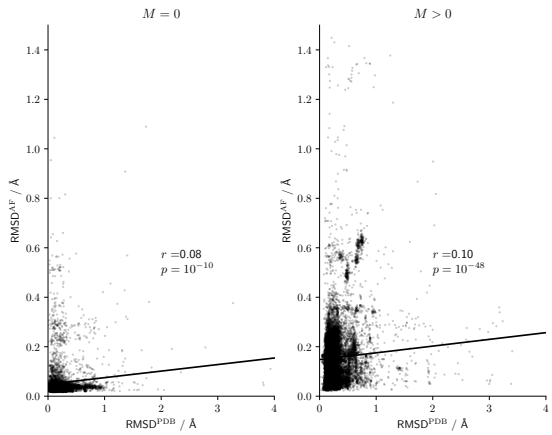

FIG. 12 **Correlations between empirical and predicted RMSD.** RMSD calculated using PDB structures, versus RMSD calculated using AF-predicted structures, for the full set of structure pairs from the PDB; non-mutated pairs (left) and mutated pairs (right) are shown separately. Pearson's correlation coefficient and linear fit are shown.

proteins are locally optimal, then any changes to the structure ought to have a deleterious effect on phenotype. To examine this, we examine correlations between deformation between all mutants, and differences in phentoype between all mutants (Fig. 11). We find that the correlations are considerably lower, or close to zero, in support of our conjecture.

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. 12 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. 13). 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 ($r = -0.59$).

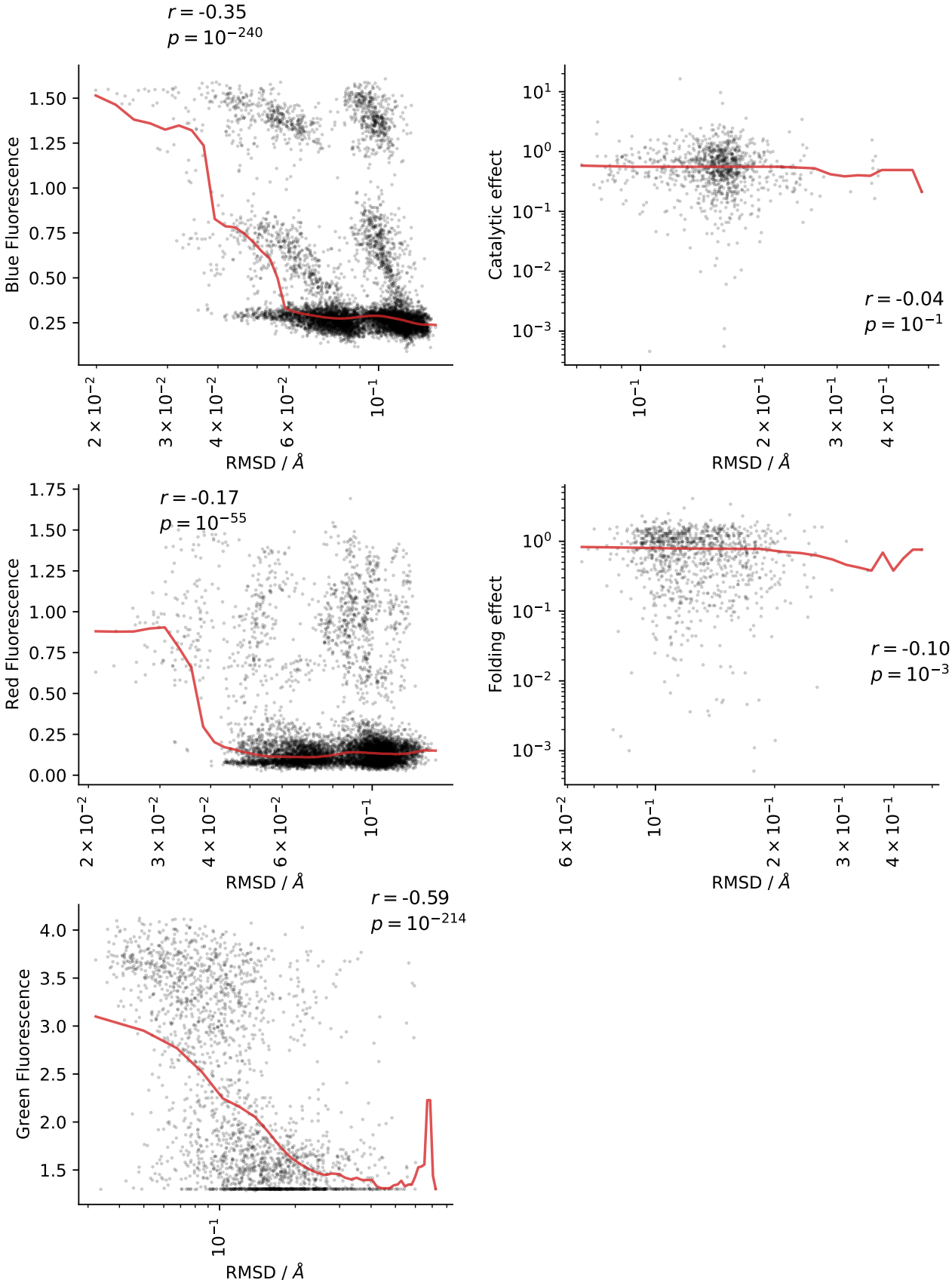

FIG. 13 **Correlations between RMSD and phenotype measurements.** Phenotype as a function of RMSD between WT and mutant. Red line indicates the median, calculated using a sliding window. Pearson's r and corresponding p-values are shown on the graph.

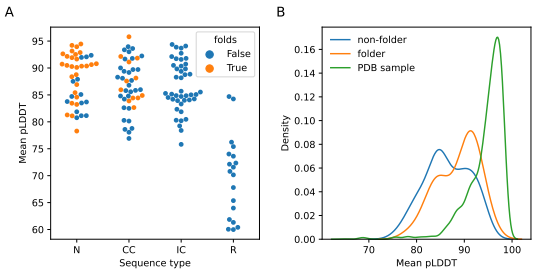

FIG. 14 A: Mean pLDDT for 147 protein sequences based on the WW domain, some of which are found to fold *in vitro*, and some of which do not fold. Proteins are grouped into: natural sequences (N); generated sequences that matches the covariance of the natural sequences (CC); generated sequences that 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 significant correlations between ΔpLDDT_i and phenotype, but in each case correlations are about half of what can be achieved using other deformation metrics.

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. 14). 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.

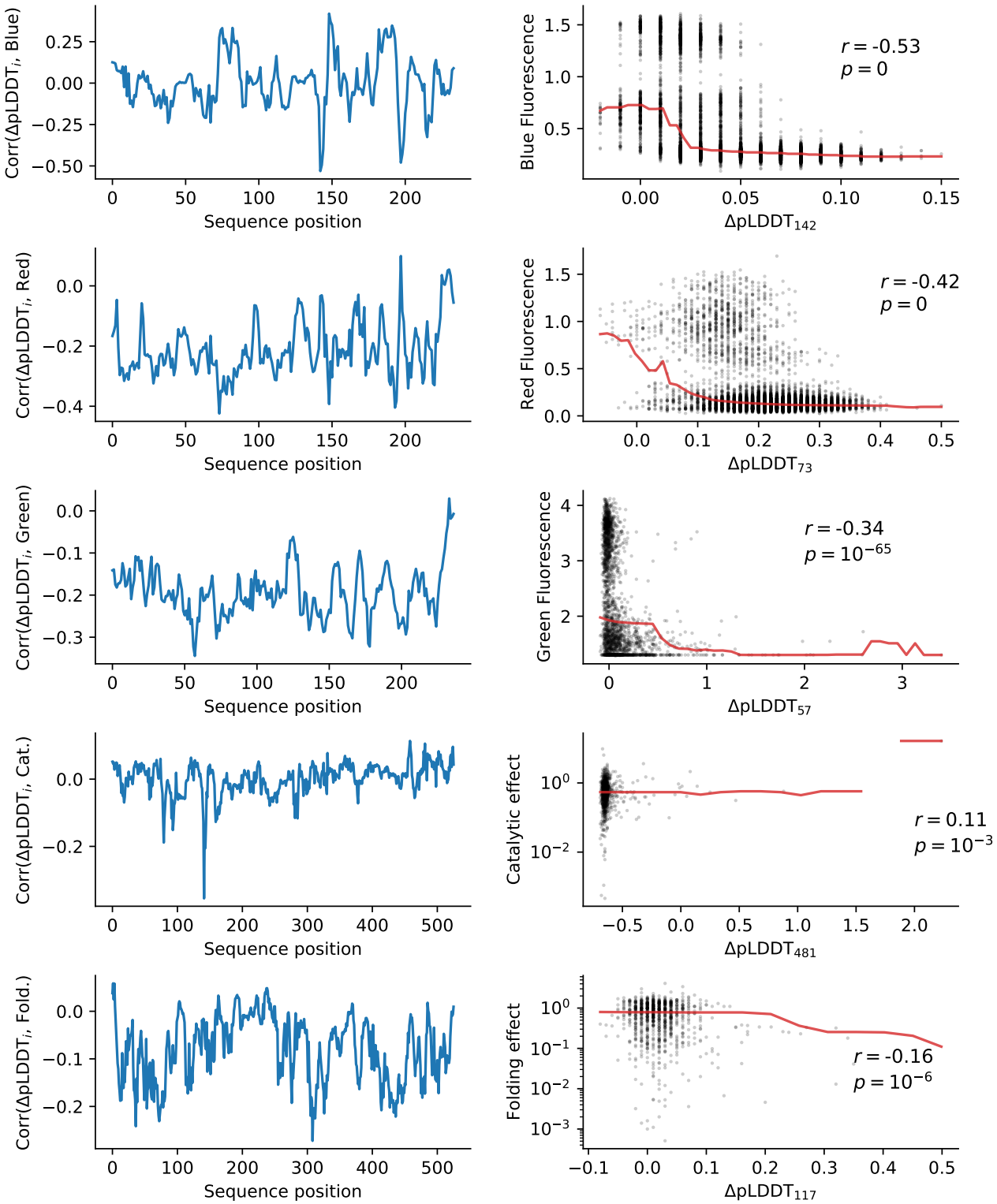

FIG. 15 **Correlations between ΔpLDDT and phenotype measurements.** (Left) Correlations between ΔpLDDT at residue i and phenotype for 5 sets of phenotype measurements (top to bottom: mTag, blue fluorescence; mTag, red fluorescence; GFP, green fluorescence; PafA, catalytic effect; PafA, folding effect) 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.

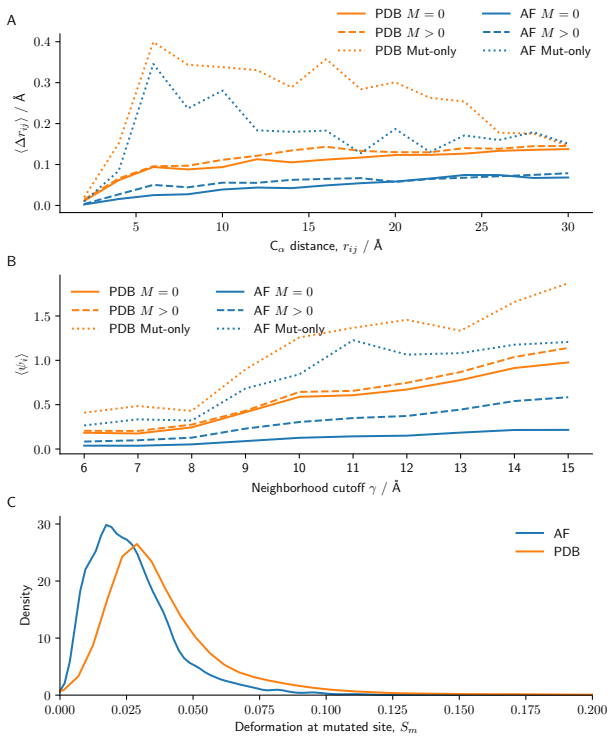

FIG. 16 **Changes in backbone distances, neighborhoods and deformation due to mutation.** A: Average change in C_α distances between structures, as a function of C_α distance. B: Average neighbor mismatch, ψ_i, as a function of neighborhood cutoff γ. Separate lines are shown for PDB and AF structures, 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.

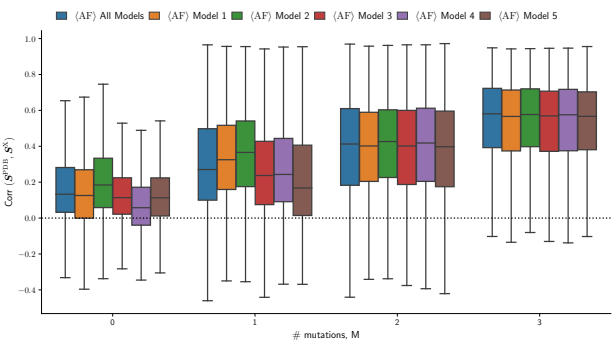

FIG. 17 **Effect of AF model.** PDB - ⟨AF⟩ correlation as a function of M and AF model type.

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).

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, $\mathbf{r}_{ij}$, and then get the absolute difference between these distances in a reference structure and a target structure, $\Delta\mathbf{r}_{ij} = |\mathbf{r}_{ij} - \mathbf{r}'_{ij}|$. In Fig. 16A, 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.

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. 16B 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. 16C). 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 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. 18) then what is typical (Fig. 16C), the structural differences are not so severe (Fig. 19). 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. 19). 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, ΔpLDDT $= -2.5$; lysozyme, ΔpLDDT $= 0.04$. Instead, *measuring deformation allows us to quantify the effect of a*

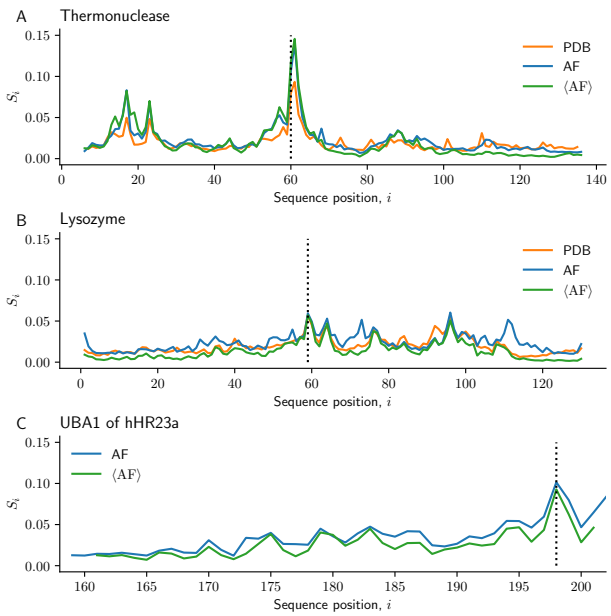

FIG. 18 **Examples of large deformation.** Deformation per residue for three examples: Thermonuclease from *staphylococcus aureus* (A), human Lysozyme (B), and the UBA1 domain of human hHR23a (C). Deformation is shown for AF, ⟨AF⟩, and PDB structures where possible. The location of mutations is indicated by dotted lines.

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. 18C). The deformation at the mutated site is $S_\mathrm{m} = 0.093$, 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.

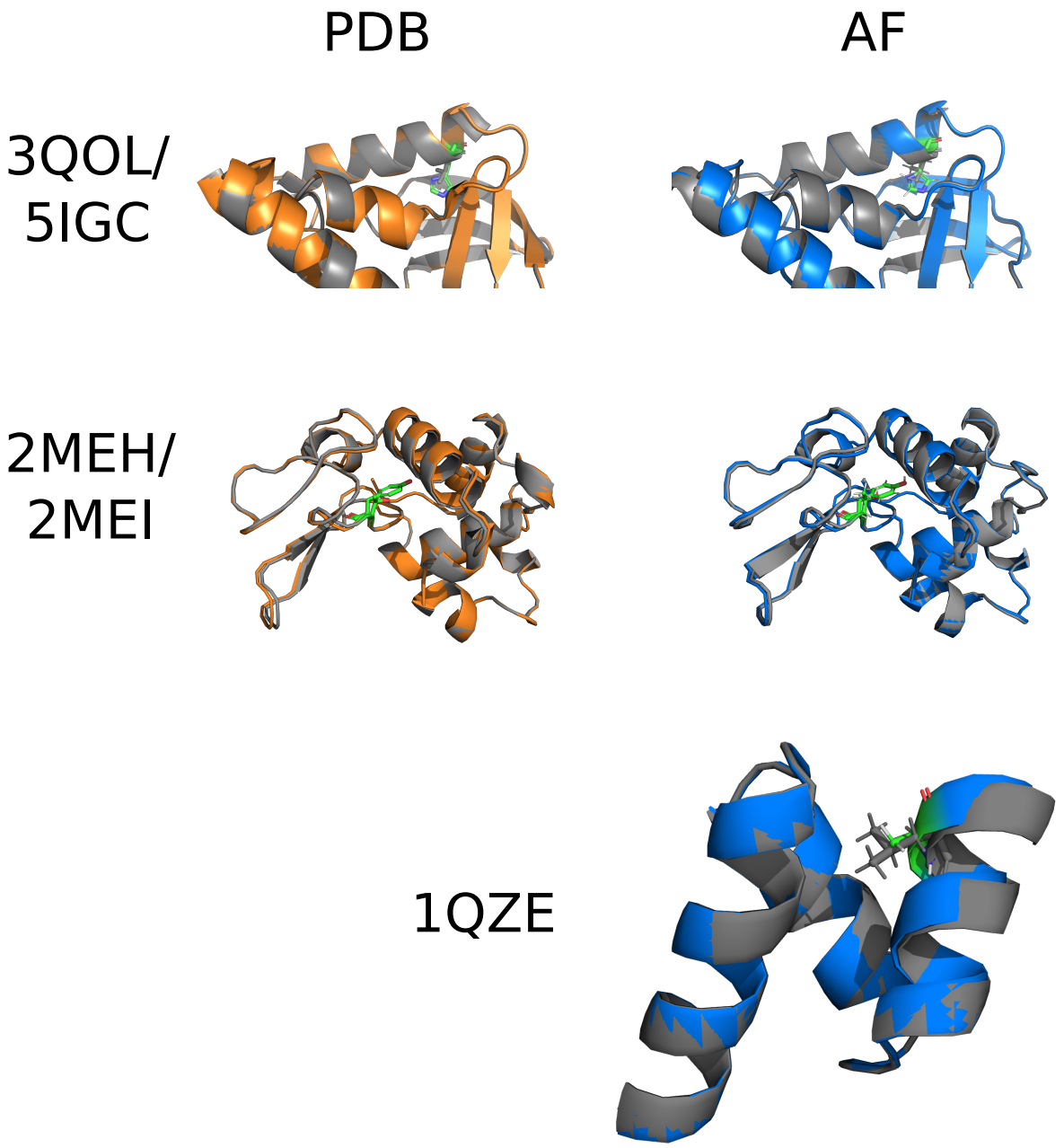

FIG. 19 **Examples of large deformation.** Comparisons of PDB and AF-predicted structures that differ by one amino acid: Thermonuclease from *staphylococcus aureus* (V60H, `3QOL_A` and `5IGC_A`; a close-up of the region affected by the mutation is shown), human Lysozyme (T59Y, `2MEH_A` and `2MEI_A`), 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.

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.

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.

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.