International Journal of
Molecular Sciences

MDPI

Article

Structural Dynamics Predominantly Determine the Adaptability of Proteins to Amino Acid Deletions

Anupam Banerjee [1,*] and Ivet Bahar [1,2,*]

[1] Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794, USA
[2] Department of Biochemistry and Cell Biology, Stony Brook University, Stony Brook, NY 11794, USA
* Correspondence: anupam.banerjee@stonybrook.edu (A.B.); bahar@laufercenter.org (I.B.)

Abstract: The insertion or deletion (indel) of amino acids has a variety of effects on protein function, ranging from disease-forming changes to gaining new functions. Despite their importance, indels have not been systematically characterized towards protein engineering or modification goals. In the present work, we focus on deletions composed of multiple contiguous amino acids (mAA-dels) and their effects on the protein (mutant) folding ability. Our analysis reveals that the mutant retains the native fold when the mAA-del obeys well-defined structural dynamics properties: localization in intrinsically flexible regions, showing low resistance to mechanical stress, and separation from allosteric signaling paths. Motivated by the possibility of distinguishing the features that underlie the adaptability of proteins to mAA-dels, and by the rapid evaluation of these features using elastic network models, we developed a positive-unlabeled learning-based classifier that can be adopted for protein design purposes. Trained on a consolidated set of features, including those reflecting the intrinsic dynamics of the regions where the mAA-dels occur, the new classifier yields a high recall of 84.3% for identifying mAA-dels that are stably tolerated by the protein. The comparative examination of the relative contribution of different features to the prediction reveals the dominant role of structural dynamics in enabling the adaptation of the mutant to mAA-del without disrupting the native fold.

Keywords: adaptability; deletion mutations; folding stability; positive-unlabeled learning classifiers; structural dynamics; elastic network models

Citation: Banerjee, A.; Bahar, I. Structural Dynamics Predominantly Determine the Adaptability of Proteins to Amino Acid Deletions. *Int. J. Mol. Sci.* **2023**, *24*, 8450. https://doi.org/10.3390/ijms24098450

Academic Editors: Alexandre G. De Brevern and Jean-Christophe Gelly

Received: 24 March 2023
Revised: 1 May 2023
Accepted: 6 May 2023
Published: 8 May 2023

1. Introduction

The insertion/deletion (indel) of nucleotides is an important source of genetic variation. Those occurring at coding regions via triplets of nucleotides, i.e., non-frame-shifting indels, result in amino acid indels (AA-indels). AA-indels constitute a mechanism of protein evolution alongside point mutations [1–4]. While point substitutions alter the side chains of residues, indels alter the protein backbone. Such modifications can have a wide range of effects, from disease conditions [5] to benefits in terms of structure or function [6,7]. While most point mutations are neutral, indels and especially those involving multiple amino acids (AAs), called mAA-indels, may introduce significant leaps in the fitness landscape [1] and drive the evolutionary adaptation to new functions [8–10]. AA-indels contribute to approximately one-fourth of disease-causing mutations in humans [5,11] and are responsible for several Mendelian disorders [12] and different types of cancers [13]. They are also responsible for the functional divergence between homologous protein structures [14]. For example, a polybasic insert near the S1/S2 cleavage site functionally distinguishes the SARS-CoV-2 spike protein from its orthologs in the SARS subfamily of betacoronaviruses [10].

There have been many experimental studies on the effects of pre-defined AA-indels on catalytic specificities [15,16] or on the binding affinities of engineered antibodies [17,18]. Despite the recognized importance of accurately assessing the effects of AA-indels on

structure and function, the use of machine learning (ML) tools for accomplishing this goal has been limited, and backbone modifications for protein engineering purposes remain relatively less explored [19].

AA-indels frequently occur at loops/coils, as these regions more readily sustain variations than buried or structured regions. Such observations directed attention to investigating the thermodynamic stability or native state entropy of AA-indel mutants [20,21], the role of compensatory mutations in supporting structural stability [1,22], the effect of point mutations and backbone modifications on the evolution of structure [23], and the pathogenicity of indels [5,24–26]. However, a computational framework to differentiate between destabilizing and neutral AA-indels remains to be established.

The primary impediment to constructing ML-based predictors has been the sparsity of learning data. However, recent studies have succeeded in compiling datasets and developing classifiers that estimate the foldability of proteins containing single-point [27] and multiple (contiguous) deletions [28]. In these studies, hundreds of AA-indels of various sizes (up to n_{AA} = 23 amino acids) have been identified that retain the wild-type (wt) fold in the absence of compensatory mutations.

In the present study, we build on this recent progress to establish the molecular basis for the ability of proteins to accommodate or resolve mAA-dels with minor, if any, changes in wt structure. As will be shown below, the intrinsic dynamics of the protein, not considered in previous studies, emerge as a major determinant of adaptability to mAA-dels and help discriminate between sustainable mAA-dels and those that would compromise the native fold.

Intrinsic dynamic refers to the collective modes of molecular motions evolutionarily optimized and uniquely encoded by the native fold, which usually enable protein–protein interactions, allosteric signaling, or other activities [29–31]. The structure-based modeling of protein dynamics has been successfully incorporated into previous ML-based algorithms for inferring the mechanisms of protein function [32]. The efficient evaluation of intrinsic dynamics using elastic network models (ENMs) [33,34] also proved useful in the genome-scale characterization of biomolecular dynamics [34], the ensemble analysis of protein families [35], or the ML-based prediction of pathogenicity for single-amino-acid variants [36–38]. The latter highlighted the role of intrinsic dynamics in eliciting or avoiding a pathogenic response.

The present study shows that the inclusion of dynamics-based features in a new positive-unlabeled (PU)-learning classifier yields a high recall rate of 84.3%. Notably, among various sequences, structures, and dynamics features considered in the algorithm, the dynamics features predicted by ENMs contribute by 72.3% to classification; and among ENM-based features, the collective motions of the native fold at the two ends of the spectrum (i.e., the lowest and the highest frequency modes) are distinguished as major determinants of the effect of mAA-dels. Furthermore, the involvement in allosteric communication is noted as another important feature that precludes the adaptation to the mAA-del. Overall, this analysis points to the intrinsic dynamics of the overall wt protein, and not that of the local structure or the chemical features at the AA-indel alone, as the major determinant of adaptability to mAA-dels.

2. Results

2.1. Dataset

We adopted a dataset [28] previously compiled for assessing the stability of deletion mutants using a PU classifier, *Profound*. As described in the methods, the dataset contains data for 153 proteins deposited in the Protein Data Bank (PDB) as both wt and mutants containing mAA-dels of length $2 \leq n_{AA} \leq 23$ (called the subset of positive mAA-dels), as well as a curated set of conformers (7649 of them) derived from existing PDB structures where indels were randomly introduced (called the subset of unlabeled mAA-dels). In the former subset, the mutant structure (deposited in the PDB) exhibits at least 70% fold similarity to that of the wt protein as computed by TM-align [39]. Unlabeled mAA-dels, on the other hand, refer to the cases where the effects of indels on the fold are unknown.

2.2. Features Describing the Intrinsic Dynamics of the Protein

We considered six types of features based on intrinsic dynamics, all evaluated for the wt protein using the *ProDy* interface [40,41] (Figure 1): (i) the mean-square fluctuations (MSFs) of residues along the softest collective motions (*global modes*) predicted by the Gaussian network model (GNM) [42]; softest modes here refer to the 2% of the *N*-1 GNM modes lying at the lowest frequency end of the mode spectrum accessible to a protein of *N* residues; these are robustly defined by the overall architecture of the protein; (ii) MSFs along the same number of highest frequency GNM modes (*local modes*); (iii) *sensitivity* of amino acids to perturbations; (iv) *effectiveness* to transduce signals within the structure—the latter two provide a measure of the role of deleted residues in sensing or transmitting allosteric signals prior to deletion, and they are deduced from perturbation response scanning (PRS) analysis [43] as previously described [44,45]—(v) *mechanical stiffness* of mAA-del as measured by the effective resistance to uniaxial tension [46]; and (vi) *essentiality* of mAA-del prior to deletion, as predicted by the essential site scanning analysis (*ESSA*) [47]. A higher *ESSA* score means a higher capacity to alter the global dynamics if bound to a ligand.

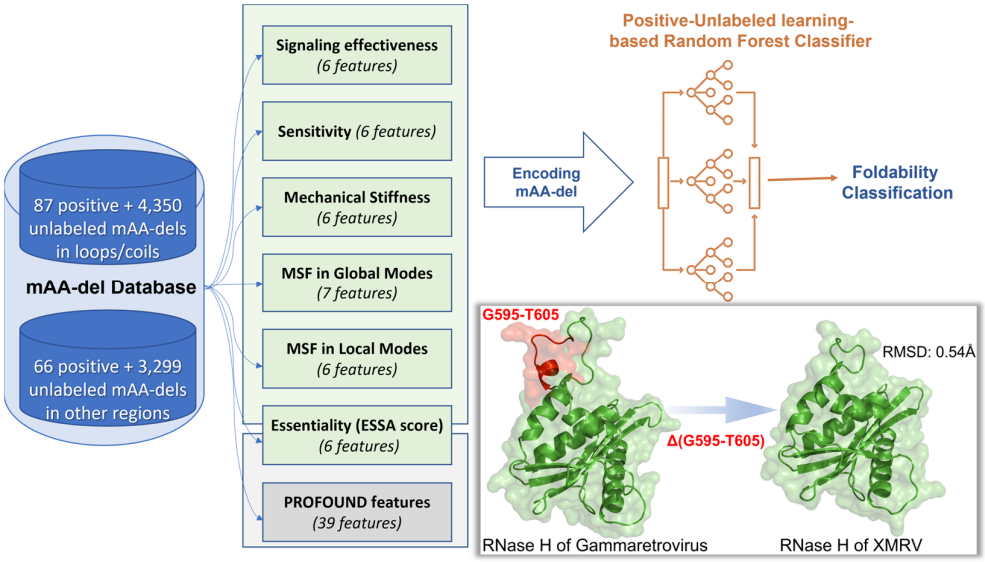

Figure 1. Schematic of the proposed method illustrating the database and features used to construct the PU-learning classifiers. The inset shows an example of a positive mAA-del in which the protein (RNase H of Gammaretrovirus, PDB ID: 4E89, Chain A, containing the Gly595–Thr605 stretch) and its homolog with the mAA-del at that particular stretch (RNase H of XMRV, PDB ID: 3V1Q, Chain A) exist both in nature. An RMSD of 0.54 Å between the two structures shows that the deletion of the Gly595–Thr605 stretch does not affect the folded conformation.

For each of the six features, we computed six values: the minimum and maximum values among the mAA-del residues, and the mean, in addition to the corresponding minimum and maximum *Z-scores* and the mean *Z-score*, <*Z-score*>. Additionally, using the *global mode* shape, we estimated whether the mAA-del is co-localized with a hinge site. This led to 37 features for each mutant, which were positive or unlabeled.

2.3. Dynamics-Based Features Discriminate between Positive and Unlabeled mAA-dels

Figure 2 displays the <*Z-scores*> computed for the six dynamics-based properties listed above for the subsets of positive and unlabeled deletions. More detailed distributions broken down by loop and non-loop regions are presented in the Supplementary Figure S1. The counterparts of Figure S1 for the maximum and minimum *Z-scores* are presented in the respective Supplementary Figures S2 and S3. The distributions for the unlabeled mAA-dels act as a control for the statistical significance of the positive mAA-del features. We found that all dynamics-based features considered here, except *essentiality*, significantly differ

between the positive and unlabeled/control mAA-dels. This is evidenced by the *p*-values reported in Figure 2, as well as the results from Student's *t*-test, Welch two-sample *t*-test, and two-sample *Z*-test (Tables S1–S4). *Essentiality* also contributes to the classification of mAA-dels, as will be shown below; however, its contribution is relatively smaller. The violin plots in Figures S1–S3 further show that the loop and non-loop regions exhibit distinctive dynamics.

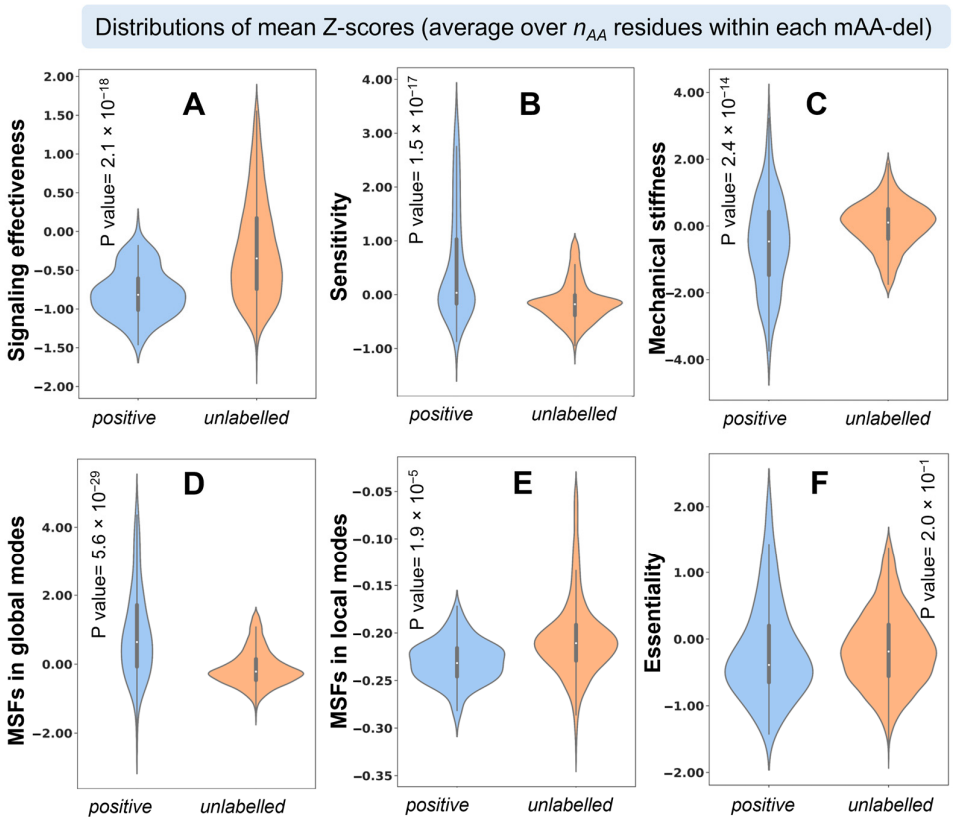

Figure 2. Dynamics-based features are differentially distributed in the positive and unlabeled mAA-del subsets. Violin plots show the distribution of <*Z-scores*> (averaged over n_{AA} residues for each mAA-del) for (**A**) *effectiveness*, (**B**) *sensitivity*, (**C**) *mechanical stiffness*, (**D**) *MSFs* in *global modes*, (**E**) *MSFs* in *local modes*, and (**F**) *essentiality* (*ESSA score*) for the subsets of positive (blue) and unlabeled (orange) mAA-dels.

2.4. Enhanced Mobilities in Global Modes Underlie the Adaptation of wt Structure to mAA-dels

The above analysis reveals the significance of the fluctuations in global modes in distinguishing between positive and unlabeled mAA-dels. Figure 3A–C display the notched box plots for the MSFs of the n_{AA} mAA-del residues in the global modes (prior to deletion) averaged over the n_{AA} residues. Residues belonging to positive mAA-dels exhibit a broad range of fluctuations, and generally exhibit higher fluctuations in global modes compared to those belonging to the unlabeled mAA-dels. This means that deletions at regions that enjoy relatively large movements in the global modes are more readily accommodated by the protein structure. The same effect is apparent in both loop and non-loop regions. The respective median values for positive mAA-dels are 0.83 and 0.44; and those for the unlabeled mAA-dels are −0.07 and −0.25. Furthermore, the fluctuations in unlabeled deletions are narrowly distributed, which further indicates the higher constraints experienced by these deletions compared to those of the positive mAA-dels. The Welch two-sample *t*-test and Student's *t*-test and *Z*-test results (Table S2) further corroborate the significance of conformational mobility in global modes as a determinant of the adaptability of the 3D structure to the deletion.

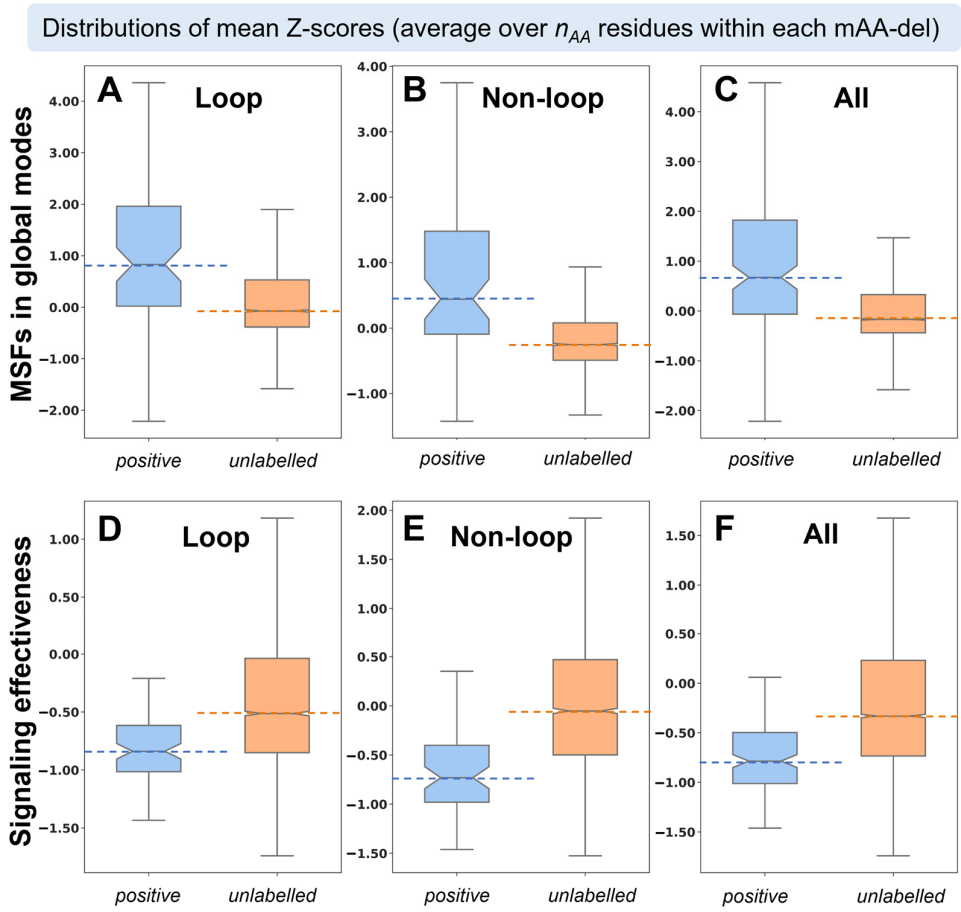

Figure 3. *MSFs* in *global modes* and signaling *effectiveness* distinguish the subsets of positive and unlabeled mAA-dels. Notched box plots show the distribution of <*Z-scores*> corresponding to *MSFs* in *global modes* for mAA-dels in (**A**) loop, (**B**) non-loop, and (**C**) all regions for positive (blue) and unlabeled (orange) mAA-dels. The distributions for the *effectiveness* of mAA-del residues in (**D**) loop, (**E**) non-loop, and (**F**) all regions further demonstrate the differences between the positive and unlabeled subsets.

2.5. Deletion of Effectors of Allosteric Communication Impairs the Adaptation to mAA-dels

This is evidenced by the higher occurrence of signaling *effectiveness* in unlabeled mAA-dels than positive mAA-dels (Figure 3D–F). Figure 3D shows that the signaling *effectiveness* (<*Z-score*> median of −0.84) of positive mAA-dels is much lower than that (−0.51) of the unlabeled mAA-dels for the loop regions, and the differences are further pronounced (−0.71 vs. −0.05) in the non-loop regions. We observe a similar trend for the consolidated set (Figure 3F) where the median values for positive and unlabeled mAA-dels are −0.79 and −0.33, respectively. This means that the residues participating in positive mAA-dels exhibit a lower tendency to transmit/propagate allosteric signals across the structure than other residues, which may explain their accommodation without necessitating an alteration in 3D fold. The Welch two-sample *t*-test (and Student's *t*-test and *Z*-test) data (Table S2) further support this result. Additionally, we observe that the third quartile of the positive instances in the loop, non-loop, and 'all' regions in Figure 3D–F are well below the zero value. All these data robustly establish that mAA-dels that are sustained without alteration in the native structure are minimally involved, if any, in signal transmission, and hence these non-influential (in allostery) residues can be resolved without the need for introducing changes in the native fold.

2.6. Relatively Lower Mechanical Stiffness at mAA-del Site Assists in Adaptation

Mechanical stiffness is a metric that helps us quantify the effective resistance of residue pairs to uniaxial tension [46]. Residues belonging to positive mAA-dels are observed to be broadly distributed and show overall lower *mechanical stiffnesses* compared to those belonging to unlabeled deletions (see Figure 2C). This is consistent with the fact that a higher *mechanical stiffness* would confer a stronger resistance to accommodate the deletion. The median of the *mechanical stiffness* <*Z-score*> for the positive mAA-dels in loops (−0.65) is significantly weaker than that of the unlabeled mAA-dels (0.04), and similar trends (−0.46 vs. 0.11 and −0.49 vs. 0.08) have been observed in the non-loop and 'all' regions, respectively. Similarly, the Welch two-sample *t*-test and *Z*-test (Table S2) indicate significant ($p < 0.001$) differences. These data confirm that lower resistance to uniaxial tension (or lower stiffness) is a discriminating feature of positive mAA-dels.

2.7. Deletion of Residues Involved in High-Frequency (Local) Motions Disrupts the Native Fold

High-frequency modes usually induce highly localized fluctuations in the most tightly packed (e.g., core) regions of proteins. These regions, also referred as kinetically hot sites (due to the localization of high vibrational energy), have been proposed to serve as folding nuclei [48,49], and would be expected to resist mutations, including deletions. The peaks in the residue-profile of *MSFs* in *local modes* thus help us identify such highly conserved and tightly packed residues. The median values for <*Z-score*> associated with *MSFs* in *local modes* were −0.23 (loop), −0.22 (non-loop), and −0.23 (all) for positive mAA-dels, which are all lower than their counterparts for unlabeled mAA-dels (Figure S1E). Therefore, positive mAA-dels seldom contain kinetically hot residues. The Welch two-sample *t*-test (Table S2) also yields a significant difference between the two subsets regardless of their location (loop or non-loop). The deletions of such segments involved in high-frequency fluctuations are therefore not tolerated due to their frequent participation in core interactions that stabilize the structure.

2.8. Classifiers Exclusively Trained on Intrinsic Dynamics Yield High Recall and Low Fall-Out Rates

To quantify the predictive ability of classifiers that distinguish between positive mAA-dels and others, we performed three sets of computations, using (a) the dynamics-based features computed using *ProDy*; (b) the sequence and structure-dependent features adopted in *Profound*; and (c) the combination of the two sets. In all cases, we trained three classifiers specific to loop, non-loop, and all regions. The stratified 10-fold (and 5-fold) cross-validation of the PU-learning classifiers ensured the proportionate representation of mAA-dels from the positive and unlabeled class labels to train and test the classifiers. In each cross-validation, we evaluated the recall and fall-out rates as metrics to evaluate the performance of the classifier and the probabilistic occurrence of positive mAA-dels within the unlabeled dataset. The recall rate informs us of the percentage of correctly identified positive mAA-del instances; whilst the fall-out rate measures the percentage of unlabeled mAA-dels that are predicted to be positive. In the absence of information regarding the true positive and true negative instances among the unlabeled mAA-dels, the fall-out rate provides an estimate of the expected fraction of positive mAA-dels among randomly selected mAA-dels.

The leftmost pair of bars in each of the three panels A–C of Figure 4 show that classifiers exclusively trained on dynamics features (labeled as *ProDy*) report a recall of 78.0% and a fall-out rate of 17.0% (during stratified 10-fold cross-validation) for mAA-dels in loops (panel A); the respective percentages are 83.8% and 21.1% in non-loop mAA-dels (panel B); and 78.0% and 19.8% in the combined dataset (panel C). The standard deviations are lower in the combined set compared to those in panels A and B. The counterparts of these bars for stratified 5-fold cross-validations confirm the same trend (Table S3). These data reveal the ability of classifiers exclusively trained on dynamics-based features to distinguish between positive and unlabeled mAA-dels. Their performance is only slightly below that of the

state-of-the-art classifier *Profound* (middle pairs of bars in panels A–C of Figure 4) despite the use of a large and diverse set of features in the latter, including sequence-, structure-, environment-dependent, and physicochemical features.

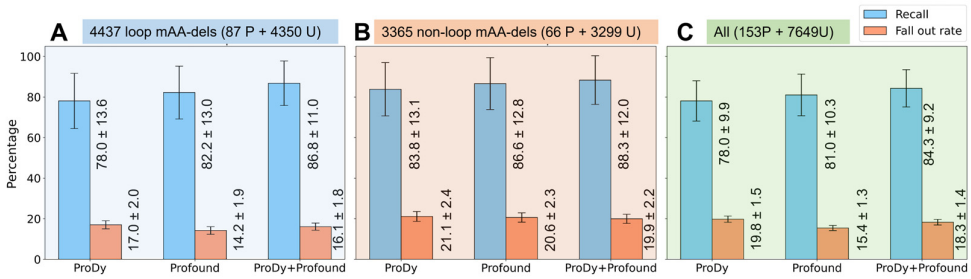

Figure 4. High performance of PU-learning classifiers for predicting positive mAA-dels. Results are presented from stratified 10-fold cross-validations for loops, non-loop regions, and all regions in the respective panels (**A**–**C**). The recall (blue bars) and fall-out rates (orange bars) are displayed for the classifiers trained on dynamics-only (*ProDy*) features (left bar), *Profound*-only features (middle bar) and consolidated (*ProDy* + *Profound*) features (right bar) for mAA-dels in loops (**A**), non-loop regions (**B**), and all regions (**C**).

2.9. Inclusion of Dynamics Features Improves the Ability of State-of-the-Art Random Forest (RF) Classifier to Predict the Effect of mAA-dels on Fold Stability

Next, we examined whether the PU-learning classifiers constructed with the consolidated feature set (referred to as *ProDy* + *Profound*) outperformed those based on either set of features. To have a common set of consolidated features, we omitted two loop-specific features used in the *Profound* classifier specific to loop mAA-dels. We find that, on stratified 10-fold cross-validation, the PU-learning classifiers trained on the consolidated feature set report a recall of 86.8% for mAA-dels in loops, and 88.3% in non-loop regions; the respective fall-out rates are 16.1% and 19.9 (rightmost bars in Figure 4A,B); and in the case of all mAA-dels (Figure 4C), the two percentages and their standard deviations are 84.3 $\pm$ 9.2 and 18.3 $\pm$ 1.4. In all three cases, the consolidated (*ProDy* + *Profound*) classifiers outperform the individual classifiers, indicating that the merged set of features in the RF enables a more accurate identification of positive mAA-dels. The recall rate exceeds those of the separate classifiers by 2–8 percentage points (depending on the subsets of mAA-dels).

The fall-out rate provides important information—that of success (in maintaining the native fold) by randomly occurring mAA-dels during evolution. A value of 15–20% is computed regardless of the classifier, suggesting that one out of 5–6 randomly selected mAA-dels is likely to be accommodated by the protein. In other words, the observed positive mAA-dels in the PDB presumably represent 1/5–1/6 of all possible mAA-dels, the large majority of which have not survived, presumably giving rise to fold destabilization.

Overall, the inclusion of dynamics-based features in the combined dataset helps us construct classifiers with improved predictive power (regarding the effect of mAA-dels on folding stability) and provide estimates of the fraction of mAA-dels (80–85%) that are not evolutionarily sustained. We next examine which features dominate the outcomes when all are used to train the RF classifier.

2.10. Dynamics-Based Features Predominantly Determine the Change in Folding Stability Caused by mAA-dels

The use of decision tree-based RFs enables us to quantify the contribution of individual features to classification. We present in Table S4 the contributions to both 5-fold and 10-fold cross-validations for PU-learning classifiers constructed for mAA-dels in loops, non-loop, and all regions, based on *ProDy* and on *ProDy* + *Profound* features.

First, we examine the relative contributions of dynamics-based features. Figure 5A,B display the corresponding results from 10-fold cross-validation. Panel A displays the contributions of features broken down by their various values/scores, as listed along the

abscissa, and the pie chart in panel B displays the agglomerated contribution of each feature. The MSFs of mAA-del residues in the global (19.7%) and local (20.2%) modes of the wt protein contribute almost 40% to foldability classification, followed by *effectiveness* to propagate signals/interactions (19.7%) and *sensitivity* to signals/interactions (17.4%). Even *ESSA* scores contribute about 11%, despite their relatively high *p*-values (see Figure 2). These dominant features are invariably distinguished in either the *ProDy* only or *ProDy* + *Profound* classifiers (Tables S4 and S5). The distributions of the minimum *sensitivity*, *MSF* (*local modes*) <*Z-score*>, *effectiveness* <*Z-score*>, and the *MSF* (*global modes*) <*Z-score*> for positive and unlabeled mAA-dels in the merged dataset are presented in Figure S4A–D, respectively.

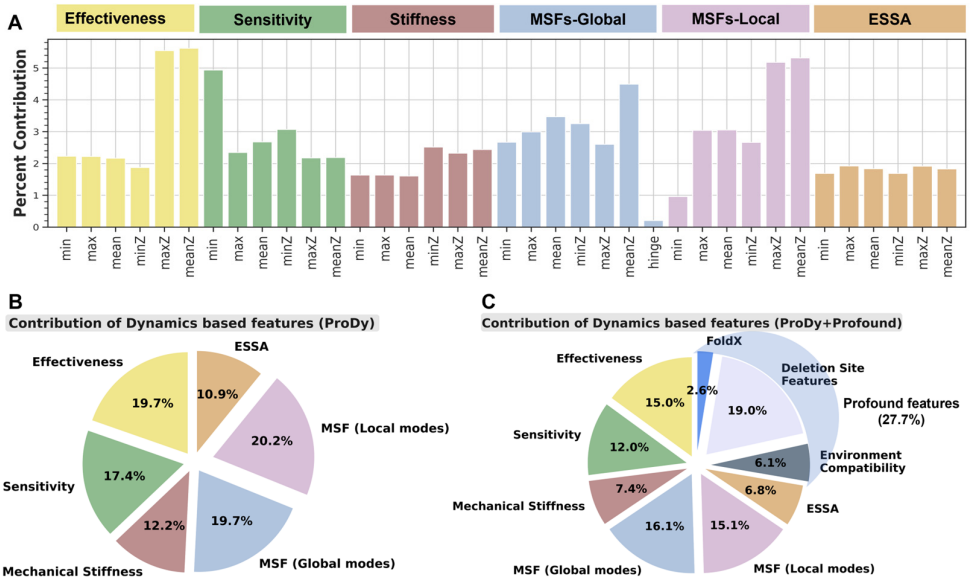

Figure 5. Dynamics-based features are major determinants of the ability of mAA-del-containing mutants to retain the native fold. (**A**) The percent contribution of individual features to stability classification as assessed by the stratified 10-fold cross-validation of the PU-learning classifier exclusively trained on dynamics-based features. The attributes are color-coded based on the feature. (**B**,**C**) Agglomerated contribution of different features when the classifier is trained only on dynamics-based features (**B**), or on the merged feature set (**C**).

Next, we turn our attention to the contributions of all features to the classifiers generated for the consolidated feature set (Figure 5C). Strikingly, dynamics-based (*ProDy*) features contribute 72.3%. These are by far the largest contributors to the classification process, even in the presence of *Profound* features. This further calls for attention to the significance of intrinsic dynamics attributes as the major determinant of the adaptability of protein folds to mAA-dels. Notably, in Figure 5C, the relative contributions of the six dynamics-based features exhibit similar trends to those observed in panel B, indicating the robustness of the relative importance of these features. As regards *Profound* features, a group designated as 'deletion site features' makes the largest (19%) contribution. The group includes information on hydrogen bonds, salt bridges, solvent accessibility, dihedral angles, end-to-end distance, and amino acid frequencies relative to natural occurrences, specific to the residues lying within the mAA-del.

3. Discussion

Multiple ML methods have been proposed in the last decade for predicting the pathogenicity associated with indels, including the SIFT Indel [26], DDIG-in [50,51], VEST-Indel [52], and MutPred-LOF/-Indel [25,53] methods trained using the HGMD [11]. These methods are based on gene and/or sequence information (structural features, when used, are predicted from sequence). On the other hand, structural data are essential to make

inferences on biophysical effects. Indel PDB [54] and IndelFR [55] are structural databases of indels identified from the sequence alignments of highly similar proteins found in the PDB, but the aligned sequences are not necessarily sequentially identical and may contain compensating mutations [1], which may complicate the interpretation of the response to AA-dels. Here, we considered pairs of protein structures whose sequences are identical, except for the mAA-del, which enabled us to sort out which properties of the wt protein predominantly underlie the adaptation to mAA-dels.

Our study highlights the importance of the protein intrinsic dynamics in defining the adaptability to mAA-dels. Strikingly, classifiers *exclusively* trained on dynamic properties (using *ProDy*) achieve a recall rate of 78%. When combined with other features (39 of which are adopted in *Profound*, including the sequence, structure, and environmental features), the improvement is relatively modest (to 84%). However, the dynamics-based features make a major contribution (72.3%) to the prediction. A dichotomy is the success (81%) of *Profound*, even though its features contribute only 27.7% to the predictor when used together with *ProDy* features. Given that intrinsic dynamics are themselves dependent on structure, which is also encoded by sequence, it is conceivable that *ProDy* supersedes many features otherwise included in *Profound*, but not all of them. However, the significantly higher weights assigned to *ProDy* features when all features are included in training the RF predictor point to the relatively stronger power of intrinsic dynamics for distinguishing the mAA-dels that can be tolerated vs. others. In this context, it is important to note that the proposed classifiers might fall short of identifying the impact of mAA-dels in intrinsically disordered proteins (IDPs) as elastic network models are unable to accurately describe and quantify the intrinsic dynamics of IDPs.

Previous studies have demonstrated that intrinsic dynamics provide the mechanisms for accomplishing biological function, adapting to protein–protein interactions, or defining the response to single amino acid variants. The present study further shows that intrinsic dynamics underlie the adaptability to short deletions (up to 23 amino acids included herein). Not only do dynamics-based features differentiate between positive and unlabeled mAA-dels, but they also play a dominant role in predicting positive mAA-dels (see Figure 5C). Note that the dynamics-based features considered herein are agnostic to sequence and are purely based on the 3D structure of the wt protein modeled as an elastic network, without any knowledge of the specific interactions and or energetics. They are purely defined by the topology of inter-residue contacts in the native structure. As such, they reflect the preferences driven by conformational entropy in the native state. The computed MSFs and other ENM-dependent properties represent unique solutions that comply with the entropy maximization principle for the collective fluctuations of all residues near native state conditions [56].

Finally, the present study was possible despite the sparsity of data on mAA-del containing proteins/mutants because of the use of a PU-learning-based classifier. PU-learning-based classifiers have recently been used in several biological applications, including the prediction of drug–drug interactions [57,58] and the identification of RNA disease associations [59]. The present study demonstrates their utility in assessing the effect of deletions on folding stability. The database was constructed by extracting from the PDB pairs of proteins or chains/subunits (in the case of multimeric structures) that shared the same sequence except for the mAA-del in one of them. However, it is important to note that, in certain cases, the extracted subunit might not fold in the absence of the other subunits of the multimeric protein or complex; and consequently, the structural and dynamic features evaluated for such cases might not be in compliance with the adaptability to deletions. To assess the extent to which such effects might have affected the results, we thoroughly examined whether the monomers/subunits used in our dataset were sufficiently stable to exist as monomers. Our extensive survey, compiled in Supplementary Table S6, showed evidence of existence as monomers for 86% of our positive mAA-dels. Furthermore, on the flip side, the inclusion of the effect of the interacting units in the environment might make it difficult to discern the effects of mAA-dels themselves on the folding properties of

individual proteins/subunits. Furthermore, one major constraint in studying mAA-dels has been the sparsity of data. The inclusion of a small fraction of subunits that may not be stable in isolation is a compromise to increase the population of positive mAA-dels. With the increase in structural data on deletion mutants and the possibility of using a dataset containing the wt and mutant in the same multimerization or complexation state, we anticipate the performance of the PU-learning classifier to be higher.

Another interesting piece of information provided by this study was the fall-out rates of 15–20%, which provides us with an estimate of the naturally occurring fraction of positive mAA-dels. This type of information cannot be observed since proteins subject to negative mAA-dels are not evolutionarily sustained, even if such deletions occur 4–5 times more frequently than the positive mAA-dels.

Finally, we recognize that the introduction of AI-driven structure prediction methods such as AlphaFold2 [60] and RoseTTAFold [61] has been revolutionary. However, as recently pointed out by Buel and Walters, AlphaFold2 is insensitive to structure-disrupting mutations in an input sequence as there are no databases that provide structural information on mutations, and hence, the predictions are largely based on wt or homologous sequences [62]. In silico predictors of the effects of genome variants provide new therapeutic opportunities in personalized medicine [63]. These studies focus on different types of mutations and their various effects from folding stability to disease-causing properties. Here, we focused on deletions and developed a framework for predicting their effects on native fold stability. Such predictors, which will only improve with increasing data, can help open new avenues for protein engineering and molecular therapeutics.

4. Materials and Methods

4.1. Dataset

We adopted the dataset [28] previously compiled to prepare the PU classifiers in *Profound*, accessible at https://cse.iitkgp.ac.in/~pralay/resources/PROFOUND/, accessed on 5 May 2023. The dataset contains 153 positive and 7649 unlabeled mAA-dels. Unlabeled mAA-dels obey the same distributions (for n_{AA}, the fraction of loop residues being deleted n_{AA}/n_{loop}, where n_{loop} is the number of residues in the loop/coil, and the location of a deleted segment with respect to *N*- or *C*-termini) as the proteins in the subset of positive mAA-dels. Deletions take place at any position along the protein sequence except for the *N*- and *C*-termini. The positive mAA-dels belonged to proteins from 42 different species and 5 different SCOP classes. We also noted that 132 out of 153 (~86%) positive mAA-dels belong to proteins that have been observed as monomeric units (see Table S6). The wt and mutant proteins for the remaining 21 mAA-dels both belong to identical multimeric assemblies. We further examined the subsets of mAA-dels located in loops (87 positive and 4350 unlabeled) and other regions (66 positive and 3299 unlabeled).

4.2. Intrinsic Dynamics-Based Attributes

MSFs of residues in the global and local modes. We used the Gaussian Network Model (GNM) [42,64] to compute the MSFs driven by the global modes, using 2% of the N-1 GNM modes at the lowest frequency end of the mode spectrum (*MSF global modes*). Secondly, the MSFs driven by the same number of highest frequency GNM modes (*MSF local modes*) were computed. In the GNM, the cross-correlation between the fluctuations $\Delta \boldsymbol{R_i}$ and $\Delta \boldsymbol{R_j}$ in the position vectors $\boldsymbol{R_i}$ and $\boldsymbol{R_j}$ of C^{α}-atoms i and j scales with the ijth element of the pseudoinverse of the Kirchhoff connectivity matrix $\boldsymbol{\Gamma}$, as-

$$\langle \Delta \boldsymbol{R_i} \cdot \Delta \boldsymbol{R_j} \rangle = (k_B T/\gamma) \left[\boldsymbol{\Gamma}^{-1} \right]_{ij} \quad (1)$$

where k_B is the Boltzmann constant, T is the absolute temperature, and γ is the uniform force constant between all nodes of the GNM representing the 3D structure [65]. The ijth element of $\boldsymbol{\Gamma}$ is -1 if the distance between the C^{α}-atoms i and j is less than a cutoff distance (usually 10 Å) of direct interaction; and the ith diagonal element is the coordination number

of that residue (or the degree of the network node at the *i*th C^{α}-atom). The MSFs of residue *i*, $\langle(\Delta \boldsymbol{R_i})^2\rangle$, are obtained by using $i = j$ in Equation (1), and may be broken down into the contribution of the normal modes *k* as

$$\langle(\Delta \boldsymbol{R_i})^2\rangle = (k_B T/\gamma)\left[\boldsymbol{\Gamma}^{-1}\right]_{ii} = (k_B T/\gamma)[\textstyle\sum_{k=1}^{N-1} \lambda_k^{-1} \boldsymbol{u_k} \boldsymbol{u_k}^T]_{ii} \quad (2)$$

where $\boldsymbol{u_k}$ and λ_k designate the k^{th} eigenvector and eigenvalue of $\boldsymbol{\Gamma}$. The modes are usually organized in ascending order, such that mode 1 refers to the lowest frequency (most global) motion, and mode *N*-1 describes the highest frequency (most local) motion. In a protein of $N = 200$ residues, for example, modes $1 \leq k \leq 4$ are included in eq 2 to calculate the *MSFs* in *global modes*, and modes $196 \leq k \leq 199$ define the *MSFs* in *local modes*.

Ability to serve as sensors and effectors of allosteric communication. We used the PRS module [44,66] of the *ProDy* API to compute the *sensitivity* and *effectiveness* of residues with regard to the sensing and transmitting signals which typically propagate through coupled fluctuations in residue positions. PRS uses the linear response theory [67] to sequentially apply directed forces on each residue and compute the resulting change in position of all residues. In principle, for a network of elastic springs, the force–displacement relation is given by $\mathbf{F} = \mathbf{H}\,\Delta\mathbf{R}$, where $\mathbf{F}$ is a 3*N*-dimensional force vector, $\mathbf{H}$ is the Hessian matrix corresponding to the anisotropic network model (ANM), and $\Delta\mathbf{R}$ is the *3N*-dimensional fluctuation $(\Delta \boldsymbol{R}_1\ \Delta \boldsymbol{R}_2\ \ldots\ \Delta \boldsymbol{R}_N)^{\mathrm{T}}$; conversely, by pre-multiplying both sides by $\mathbf{H}^{-1}$, the response $\Delta\mathbf{R}$ to $\mathbf{F}$ becomes

$$\Delta\mathbf{R} = \mathbf{H}^{-1}\,\mathbf{F} \quad (3)$$

In the PRS, $\mathbf{F}$ is the perturbation and $\Delta\mathbf{R}$ is the response. The response to a perturbation at residue *i* may be expressed as $\Delta \boldsymbol{R}^{(i)} = (\Delta r_{1x}^{(i)}\ \Delta r_{1y}^{(i)}\ \Delta r_{1z}^{(i)} \ldots \ldots\ \Delta r_{Nz}^{(i)})^{T}$. The response (*sensitivity*) of the residue *j* to the perturbation of residue *i* (effector) is organized in a $N \times N$ PRS matrix, $\mathbf{S_{PRS}}$, the *i*th row of which describes the *effectiveness* of the *i*th residue in transmitting signals and the *j*th column corresponds to the *sensitivity* of a residue *j* to signals from other residues. The response to unit perturbation at each site is obtained by dividing each row of $\mathbf{S_{PRS}}$ by its diagonal element. The average of row *i* elements in this normalized matrix represents the *effectiveness* of residue *i*, and the average of column *i* represents the *sensitivity* of residue *i* [44].

Mechanical stiffness defines the effective resistance of residue pairs to uniaxial tension. The *mechanical stiffness* was computed using the *Mechstiff* module of *ProDy* based on the theory introduced by Eyal and Bahar [46]. We generate an $N \times N$ matrix for the effective stiffness (or effective resistance or force constant, $< \kappa_{ij} >$) for each residue pair (*i*, *j*) in response to uniaxial tension using

$$\langle \kappa_{ij} \rangle = \frac{\sum_{(k)} d_{ij}^k \gamma \lambda_k}{\sum_{(k)} d_{ij}^k} \quad (4)$$

where $d_{ij}^{(k)}$ is the deformation along the *k*th mode in response to the tension

$$d_{ij}^{(k)} = (k_B T/\gamma\lambda_k)^{\frac{1}{2}}\ cos\alpha_{ij}{}^{(k)} |\boldsymbol{u}_i^{(k)} - \boldsymbol{u}_j^{(k)}| \quad (5)$$

Here, $\alpha_{ij}{}^{(k)}$ is the angle between the direction of the external force and that of the change $\Delta \boldsymbol{R}_{ij}^{(k)}$ in the inter-residue distance induced by mode *k*.

Essential site-scanning analysis (*ESSA*) provides a measure of the change in global dynamics in response to ligand binding [47]. Ligand binding to a given residue is mimicked by crowding the neighborhood of that residue upon the inclusion of additional nodes at its side chain atoms' positions. The *ESSA* score for each residue corresponds to the Z-score of the percent shift in the eigenvalues of the softest 10 modes due to this crowding. Higher *ESSA* scores suggest essential sites that alter the global dynamics.

Z-scores. For each of the above six features, we computed six values: the minimum, maximum, mean, minimum *Z-score*, maximum *Z-score*, and <*Z-score*>, considering all residues in the mAA-del segment. The *Z-score* for each feature f of each mAA-del residue i is computed using

$$Z\text{-}score_i^f = \frac{f_i - \mu_f}{\sigma_f} \quad (6)$$

where μ_f and σ_f denote the mean and standard deviation of f over all residues in the wt protein.

Hinge sites. Hinge sites in a given mode are residues distinguished by their minimal motions, if any, in that particular mode. They serve as anchors between substructures that concertedly move around them, and as such, they play a critical mechanical role. In the GNM analysis, they are readily identified by plotting the eigenvectors as a function of residue index and examining the zero-crossover points. Here, we focused on the global hinges, i.e., considering 2% of GNM modes at the lowest frequency end of the mode spectrum. We used the *calcHinges* function of *ProDy* with the default parameters and protocol to compute the hinge sites corresponding to these global modes.

4.3. Construction of PU Learning-Based Classifiers

We adhered to the PU learning-based classifier introduced by Elkan and Noto [68]. In previous work (*Profound*), we considered separate subsets to train classifiers for loop or non-loop regions. Here, we constructed a more robust RF classifier using the merged dataset composed of all mAA-dels (loops and non-loops) to train the classifier. Previous work [28] considered 39 (41 for loop mAA-dels) features composed of fold attributes, environment-specific properties, and deletion site-specific properties to train the classifier; here, we additionally included intrinsic dynamics-dependent attributes. We used the scikit-learn [69] implementation of the RF method, and our API *ProDy* [40,41] to evaluate dynamics-based features for all members of our dataset. We constructed nine PU learning-based classifiers (the performance of which on 10-fold stratified cross-validations is illustrated in Figure 4) corresponding to mAA-dels in the loop, non-loop, and combined datasets, each constructed with only intrinsic dynamics-based attributes, attributes used in *Profound*, and a combination of all attributes. The feature vectors corresponding to the positive and unlabeled mAA-dels in each dataset were considered to construct each PU learning-based classifier. The PU classification framework was constructed with the help of two random forest classifiers (RFCs). The first RFC consults the distribution of the different features in the positive mAA-dels to identify the likelihood of each unlabeled mAA-dels belonging to a positive and negative class. Subsequently, along with the positive mAA-dels, the unlabeled mAA-dels weighted by these probabilities are used to train the final RFC. The detailed algorithm describing the construction of the PU learning classifiers can be found in our previous work [28].

We further used the *feature_importances_* attribute from the RandomForestClassifier class of the sklearn package [69] to compute the contribution of individual features to classification. The importance of each attribute is computed as the mean and standard deviation of the decrease in the accumulation of impurity within each tree across all trees considered in the RFCs.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms24098450/s1.

Author Contributions: I.B. and A.B. conceptualized the study. A.B. carried out the computations, analyzed the data, and prepared the tables and figures under the supervision of I.B. I.B. and A.B. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institutes of Health grant R01 GM139297 and by Human Frontiers Science Program (HFSP) award with reference no. RGP0027/2020.

Data Availability Statement: This paper analyzed publicly available data. Any additional information required to reanalyze the data reported in this paper is available from the corresponding authors upon request (anupam.banerjee@stonybrook.edu and bahar@laufercenter.org).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tóth-Petróczy, A.; Tawfik, D.S. Protein insertions and deletions enabled by neutral roaming in sequence space. *Mol. Biol. Evol.* **2013**, *30*, 761–771. [CrossRef] [PubMed]
2. Chothia, C.; Gough, J.; Vogel, C.; Teichmann, S.A. Evolution of the protein repertoire. *Science* **2003**, *300*, 1701–1703. [CrossRef] [PubMed]
3. Lin, M.; Whitmire, S.; Chen, J.; Farrel, A.; Shi, X.; Guo, J.T. Effects of short indels on protein structure and function in human genomes. *Sci. Rep.* **2017**, *7*, 9313. [CrossRef] [PubMed]
4. Mullaney, J.M.; Mills, R.E.; Pittard, W.S.; Devine, S.E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **2010**, *19*, R131–R136. [CrossRef] [PubMed]
5. Choi, Y.; Sims, G.E.; Murphy, S.; Miller, J.R.; Chan, A.P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **2012**, *7*, e46688. [CrossRef]
6. Britten, R.J. Transposable element insertions have strongly affected human evolution. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 19945–19948. [CrossRef]
7. Hashimoto, K.; Panchenko, A.R. Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 20352–20357. [CrossRef]
8. Grishin, N.V. Fold change in evolution of protein structures. *J. Struct. Biol.* **2001**, *134*, 167–185. [CrossRef]
9. Zhang, Z.; Wang, Y.; Wang, L.; Gao, P. The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. *PLoS ONE* **2010**, *5*, e14316. [CrossRef]
10. Cheng, M.H.; Zhang, S.; Porritt, R.A.; Noval Rivas, M.; Paschold, L.; Willscher, E.; Binder, M.; Arditi, M.; Bahar, I. Superantigenic character of an insert unique to SARS-CoV-2 spike supported by skewed TCR repertoire in patients with hyperinflammation. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 25254–25262. [CrossRef]
11. Stenson, P.D.; Mort, M.; Ball, E.V.; Chapman, M.; Evans, K.; Azevedo, L.; Hayden, M.; Heywood, S.; Millar, D.S.; Phillips, A.D.; et al. The Human Gene Mutation Database (HGMD(®)): Optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **2020**, *139*, 1197–1207. [CrossRef] [PubMed]
12. MacArthur, D.G.; Balasubramanian, S.; Frankish, A.; Huang, N.; Morris, J.; Walter, K.; Jostins, L.; Habegger, L.; Pickrell, J.K.; Montgomery, S.B.; et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **2012**, *335*, 823–828. [CrossRef] [PubMed]
13. Ye, K.; Wang, J.; Jayasinghe, R.; Lameijer, E.W.; McMichael, J.F.; Ning, J.; McLellan, M.D.; Xie, M.; Cao, S.; Yellapantula, V.; et al. Systematic discovery of complex insertions and deletions in human cancers. *Nat. Med.* **2016**, *22*, 97–104. [CrossRef] [PubMed]
14. Jiang, H.; Blouin, C. Insertions and the emergence of novel protein structure: A structure-based phylogenetic study of insertions. *BMC Bioinform.* **2007**, *8*, 444. [CrossRef] [PubMed]
15. Park, H.S.; Nam, S.H.; Lee, J.K.; Yoon, C.N.; Mannervik, B.; Benkovic, S.J.; Kim, H.S. Design and evolution of new catalytic activity with an existing protein scaffold. *Science* **2006**, *311*, 535–538. [CrossRef]
16. Hoque, M.A.; Zhang, Y.; Chen, L.; Yang, G.; Khatun, M.A.; Chen, H.; Hao, L.; Feng, Y. Stepwise Loop Insertion Strategy for Active Site Remodeling to Generate Novel Enzyme Functions. *ACS Chem. Biol.* **2017**, *12*, 1188–1193. [CrossRef]
17. Lamminmäki, U.; Paupério, S.; Westerlund-Karlsson, A.; Karvinen, J.; Virtanen, P.L.; Lövgren, T.; Saviranta, P. Expanding the conformational diversity by random insertions to CDRH2 results in improved anti-estradiol antibodies. *J. Mol. Biol.* **1999**, *291*, 589–602. [CrossRef]
18. Mou, Y.; Zhou, X.X.; Leung, K.; Martinko, A.J.; Yu, J.Y.; Chen, W.; Wells, J.A. Engineering Improved Antiphosphotyrosine Antibodies Based on an Immunoconvergent Binding Motif. *J. Am. Chem. Soc.* **2018**, *140*, 16615–16624. [CrossRef]
19. Emond, S.; Petek, M.; Kay, E.J.; Heames, B.; Devenish, S.R.A.; Tokuriki, N.; Hollfelder, F. Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. *Nat. Commun.* **2020**, *11*, 3469. [CrossRef]
20. Dagan, S.; Hagai, T.; Gavrilov, Y.; Kapon, R.; Levy, Y.; Reich, Z. Stabilization of a protein conferred by an increase in folded state entropy. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 10628–10633. [CrossRef]
21. Gavrilov, Y.; Dagan, S.; Levy, Y. Shortening a loop can increase protein native state entropy. *Proteins* **2015**, *83*, 2137–2146. [CrossRef] [PubMed]
22. Leushkin, E.V.; Bazykin, G.A.; Kondrashov, A.S. Insertions and deletions trigger adaptive walks in Drosophila proteins. *Proc. Biol. Sci.* **2012**, *279*, 3075–3082. [CrossRef] [PubMed]
23. Zhang, Z.; Wang, J.; Gong, Y.; Li, Y. Contributions of substitutions and indels to the structural variations in ancient protein superfamilies. *BMC Genom.* **2018**, *19*, 771. [CrossRef] [PubMed]
24. Kircher, M.; Witten, D.M.; Jain, P.; O'Roak, B.J.; Cooper, G.M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **2014**, *46*, 310–315. [CrossRef]

25. Pagel, K.A.; Antaki, D.; Lian, A.; Mort, M.; Cooper, D.N.; Sebat, J.; Iakoucheva, L.M.; Mooney, S.D.; Radivojac, P. Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. *PLoS Comput. Biol.* **2019**, *15*, e1007112. [CrossRef]
26. Hu, J.; Ng, P.C. SIFT Indel: Predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS ONE* **2013**, *8*, e77940. [CrossRef]
27. Banerjee, A.; Levy, Y.; Mitra, P. Analyzing Change in Protein Stability Associated with Single Point Deletions in a Newly Defined Protein Structure Database. *J. Proteome Res.* **2019**, *18*, 1402–1410. [CrossRef]
28. Banerjee, A.; Kumar, A.; Ghosh, K.K.; Mitra, P. Estimating Change in Foldability Due to Multipoint Deletions in Protein Structures. *J. Chem. Inf. Model.* **2020**, *60*, 6679–6690. [CrossRef]
29. Bakan, A.; Bahar, I. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 14349–14354. [CrossRef]
30. Tobi, D.; Bahar, I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 18908–18913. [CrossRef]
31. Zhang, Y.; Doruker, P.; Kaynak, B.; Zhang, S.; Krieger, J.; Li, H.; Bahar, I. Intrinsic dynamics is evolutionarily optimized to enable allosteric behavior. *Curr. Opin. Struct. Biol.* **2020**, *62*, 14–21. [CrossRef] [PubMed]
32. Banerjee, A.; Saha, S.; Tvedt, N.C.; Yang, L.W.; Bahar, I. Mutually beneficial confluence of structure-based modeling of protein dynamics and machine learning methods. *Curr. Opin. Struct. Biol.* **2022**, *78*, 102517. [CrossRef] [PubMed]
33. Bahar, I.; Lezon, T.R.; Yang, L.W.; Eyal, E. Global dynamics of proteins: Bridging between structure and function. *Annu. Rev. Biophys.* **2010**, *39*, 23–42. [CrossRef] [PubMed]
34. Li, H.; Chang, Y.Y.; Lee, J.Y.; Bahar, I.; Yang, L.W. DynOmics: Dynamics of structural proteome and beyond. *Nucleic Acids Res.* **2017**, *45*, W374–W380. [CrossRef]
35. Zhang, S.; Li, H.; Krieger, J.M.; Bahar, I. Shared Signature Dynamics Tempered by Local Fluctuations Enables Fold Adaptability and Specificity. *Mol. Biol. Evol.* **2019**, *36*, 2053–2068. [CrossRef]
36. Ponzoni, L.; Bahar, I. Structural dynamics is a determinant of the functional significance of missense variants. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4164–4169. [CrossRef]
37. Ponzoni, L.; Peñaherrera, D.A.; Oltvai, Z.N.; Bahar, I. Rhapsody: Predicting the pathogenicity of human missense variants. *Bioinformatics* **2020**, *36*, 3084–3092. [CrossRef]
38. Rodrigues, C.H.; Pires, D.E.; Ascher, D.B. DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.* **2018**, *46*, W350–W355. [CrossRef]
39. Zhang, Y.; Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302–2309. [CrossRef]
40. Bakan, A.; Meireles, L.M.; Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **2011**, *27*, 1575–1577. [CrossRef]
41. Zhang, S.; Krieger, J.M.; Zhang, Y.; Kaya, C.; Kaynak, B.; Mikulska-Ruminska, K.; Doruker, P.; Li, H.; Bahar, I. ProDy 2.0: Increased Scale and Scope after 10 Years of Protein Dynamics Modelling with Python. *Bioinformatics* **2021**, *37*, 3657–3659. [CrossRef] [PubMed]
42. Bahar, I.; Atilgan, A.R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des.* **1997**, *2*, 173–181. [CrossRef] [PubMed]
43. Atilgan, C.; Gerek, Z.N.; Ozkan, S.B.; Atilgan, A.R. Manipulation of conformational change in proteins by single-residue perturbations. *Biophys. J.* **2010**, *99*, 933–943. [CrossRef] [PubMed]
44. General, I.J.; Liu, Y.; Blackburn, M.E.; Mao, W.; Gierasch, L.M.; Bahar, I. ATPase subdomain IA is a mediator of interdomain allostery in Hsp70 molecular chaperones. *PLoS Comput. Biol.* **2014**, *10*, e1003624. [CrossRef] [PubMed]
45. Dutta, A.; Krieger, J.; Lee, J.Y.; Garcia-Nafria, J.; Greger, I.H.; Bahar, I. Cooperative Dynamics of Intact AMPA and NMDA Glutamate Receptors: Similarities and Subfamily-Specific Differences. *Structure* **2015**, *23*, 1692–1704. [CrossRef] [PubMed]
46. Eyal, E.; Bahar, I. Toward a molecular understanding of the anisotropic response of proteins to external forces: Insights from elastic network models. *Biophys. J.* **2008**, *94*, 3424–3435. [CrossRef]
47. Kaynak, B.T.; Bahar, I.; Doruker, P. Essential site scanning analysis: A new approach for detecting sites that modulate the dispersion of protein global motions. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1577–1586. [CrossRef]
48. Rader, A.; Bahar, I. Folding core predictions from network models of proteins. *Polymer* **2004**, *45*, 659–668. [CrossRef]
49. Bahar, I.; Atilgan, A.R.; Demirel, M.C.; Erman, B. Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Phys. Rev. Lett.* **1998**, *80*, 2733. [CrossRef]
50. Folkman, L.; Yang, Y.; Li, Z.; Stantic, B.; Sattar, A.; Mort, M.; Cooper, D.N.; Liu, Y.; Zhou, Y. DDIG-in: Detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. *Bioinformatics* **2015**, *31*, 1599–1606. [CrossRef]
51. Zhao, H.; Yang, Y.; Lin, H.; Zhang, X.; Mort, M.; Cooper, D.N.; Liu, Y.; Zhou, Y. DDIG-in: Discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biol.* **2013**, *14*, R23. [CrossRef] [PubMed]
52. Douville, C.; Masica, D.L.; Stenson, P.D.; Cooper, D.N.; Gygax, D.M.; Kim, R.; Ryan, M.; Karchin, R. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). *Hum. Mutat.* **2016**, *37*, 28–35. [CrossRef] [PubMed]

53. Pagel, K.A.; Pejaver, V.; Lin, G.N.; Nam, H.J.; Mort, M.; Cooper, D.N.; Sebat, J.; Iakoucheva, L.M.; Mooney, S.D.; Radivojac, P. When loss-of-function is loss of function: Assessing mutational signatures and impact of loss-of-function genetic variants. *Bioinformatics* **2017**, *33*, i389. [CrossRef] [PubMed]
54. Hsing, M.; Cherkasov, A. Indel PDB: A database of structural insertions and deletions derived from sequence alignments of closely related proteins. *BMC Bioinform.* **2008**, *9*, 293. [CrossRef]
55. Zhang, Z.; Xing, C.; Wang, L.; Gong, B.; Liu, H. IndelFR: A database of indels in protein structures and their flanking regions. *Nucleic Acids Res.* **2012**, *40*, D512–D518. [CrossRef]
56. Lezon, T.R.; Bahar, I. Using entropy maximization to understand the determinants of structural dynamics beyond native contact topology. *PLoS Comput. Biol.* **2010**, *6*, e1000816. [CrossRef]
57. Zhang, Y.; Qiu, Y.; Cui, Y.; Liu, S.; Zhang, W. Predicting drug-drug interactions using multi-modal deep auto-encoders based network embedding and positive-unlabeled learning. *Methods* **2020**, *179*, 37–46. [CrossRef]
58. Zheng, Y.; Peng, H.; Zhang, X.; Zhao, Z.; Gao, X.; Li, J. DDI-PULearn: A positive-unlabeled learning method for large-scale prediction of drug-drug interactions. *BMC Bioinform.* **2019**, *20*, 661. [CrossRef]
59. Wei, H.; Xu, Y.; Liu, B. iPiDi-PUL: Identifying Piwi-interacting RNA-disease associations based on positive unlabeled learning. *Brief. Bioinform.* **2021**, *22*, bbaa058. [CrossRef]
60. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
61. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876. [CrossRef] [PubMed]
62. Buel, G.R.; Walters, K.J. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* **2022**, *29*, 1–2. [CrossRef] [PubMed]
63. Katsonis, P.; Wilhelm, K.; Williams, A.; Lichtarge, O. Genome interpretation using in silico predictors of variant impact. *Hum. Genet.* **2022**, *141*, 1549–1577. [CrossRef] [PubMed]
64. Haliloglu, T.; Bahar, I.; Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **1997**, *79*, 3090. [CrossRef]
65. Tirion, M.M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908. [CrossRef]
66. Atilgan, C.; Atilgan, A.R. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput. Biol.* **2009**, *5*, e1000544. [CrossRef]
67. Ikeguchi, M.; Ueno, J.; Sato, M.; Kidera, A. Protein structural change upon ligand binding: Linear response theory. *Phys. Rev. Lett.* **2005**, *94*, 078102. [CrossRef]
68. Elkan, C.; Noto, K. Learning classifiers from only positive and unlabeled data. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 213–220.
69. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.