Predicting the Structural Impact of Human Alternative Splicing

Yuxuan Song[1], Chengxin Zhang[1], Gilbert S. Omenn[1], Matthew J. O'Meara[1,2]*, Joshua D. Welch[1,3]*

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
[2]Department of Medicinal Chemistry, University of Michigan, Ann Arbor, MI, USA
[3]Department of Computer Science and Engineering, University of Michigan, Ann Arbor, MI, USA

*Address correspondence to maom@umich.edu and welchjd@umich.edu

ORCiD IDs:
YS:0000-0002-5883-1208
CZ: 0000-0001-7290-1324
GSO: 0000-0002-8976-6074
MJO: 0000-0002-3128-5331
JW: 0000-0002-5869-2391

ABBREVIATIONS:
scRNA-seq: single-cell RNA sequencing
TM-Score: template modeling score
pLDDT: predicted local distance difference test
RMSD: root mean squared deviation
MSA: multiple sequence alignment
PCC: pearson correlation coefficients
ANOVA: analysis of variance
PTM: post translational modification
AS: alternative splicing
ES: exon skipping
ADS: alternative donor site or alternative 5′ splice site,
AAS: alternative acceptor site or alternative 3’ splice site
MXE: mutually exclusive exons
IR: intron retention
AFE: alternative first exon
ALE: alternative last exon
MXE-AFE: mutually exclusive exon-alternative first exon
MXE-ALE: mutually exclusive exon-alternative last exon
ANOVA: analysis of variance
IDR: intrinsic disorder region
PPI: protein-protein interaction

SASA: solvent accessible surface area
RSA: relative solvent accessibility
GO: gene ontology

Summary

Protein structure prediction with neural networks is a powerful new method for linking protein sequence, structure, and function, but structures have generally been predicted for only a single isoform of each gene, neglecting splice variants. To investigate the structural implications of alternative splicing, we used AlphaFold2 to predict the structures of more than 11,000 human isoforms. We employed multiple metrics to identify splicing-induced structural alterations, including template matching score, secondary structure composition, surface charge distribution, radius of gyration, accessibility of post-translational modification sites, and structure-based function prediction. We identified examples of how alternative splicing induced clear changes in each of these properties. Structural similarity between isoforms largely correlated with degree of sequence identity, but we identified a subset of isoforms with low structural similarity despite high sequence similarity. Exon skipping and alternative last exons tended to increase the surface charge and radius of gyration. Splicing also buried or exposed numerous post-translational modification sites, most notably among the isoforms of *BAX*. Functional prediction nominated numerous functional differences among isoforms of the same gene, with loss of function compared to the reference predominating. Finally, we used single-cell RNA-seq data from the Tabula Sapiens to determine the cell types in which each structure is expressed. Our work represents an important resource for studying the structure and function of splice isoforms across the cell types of the human body.

Keywords:

Protein Structure, Alternative Splicing, AlphaFold2, Isoform function, Single-Cell RNA-seq

1 Introduction

Eukaryotic cells achieve remarkable functional diversity from relative compact genomes. This is achieved in part through alternative splicing, where for 90% of genes different combinations of pre-mRNA exons are spliced and ligated together. In humans, alternative splicing expands ~20,000 genes into over 100,000 protein products[1–3]. Alternative splicing can exert diverse biological effects on proteins, including alterations in their binding properties, subcellular localization, and stability[4]. Moreover, aberrant alternative splicing can disrupt normal biological processes and lead to Duchenne muscular dystrophy[5], cardiovascular disease and multiple types of cancers[6]. Alternative splicing has been extensively studied from the perspectives of

regulation[7], evolution[8], and expression profile[9]. In contrast, the impact of alternative splicing on protein structure has been studied much less. Previous studies have shown that alternative splicing can induce unstable protein conformations[10], change protein localization[4], alter transmembrane domains[11], and create variations in repeat regions[12,13]. However, these studies have been relatively constrained in scope[13], and a thorough and systematic investigation of how splicing affects structure is imperative.

For decades, obtaining experimental protein structures has been a laborious and time-consuming process. As a result, structures of multiple spliced isoforms from the same gene have not often been experimentally solved. Computational protein structure prediction methods like Rosetta and I-TASSER offer a more practical means of studying isoform structures on a large scale[14,15]. Most recently, neural network approaches such as AlphaFold2 and RoseTTAFold have enabled high-precision protein structure prediction[16,17]. These tools provide an exciting new opportunity to study how alternative splicing affects protein structure. Recently, Sommer et al. utilized ColabFold to predict structures for more than 127,000 human spliced isoforms annotated from RNA-seq experiments[18,19]. They used protein structure confidence scores (pLDDT) to nominate a reference isoform for each gene based on the assumption that predicted structures for loss-of-function isoforms will have lower confidence. However, Sommer et al. did not investigate the structural differences among isoforms any further, and the details of how splicing changes specific structural properties of proteins remain unexplored. Additionally, the functional implications of such structural changes remain unclear.

Another important question concerns the cell-type-specific expression of isoforms. Single-cell RNA sequencing (scRNA-seq) has enabled us to map the expression profiles of individual genes and isoforms at cellular resolution[20]. Prior investigations utilizing scRNA-seq have delineated differential isoform usage in contexts such as the adult brain[21] and muscle cell maturation[22]. Multiple studies have profiled nearly the entire human body with scRNA-seq[23–25], raising the exciting possibility of mapping the cells within which each predicted structure is present. This would be an important step toward understanding how alternative splicing changes the functions of proteins within different cellular contexts.

In this study, we addressed these questions by using AlphaFold2 to predict structures of more than 11,000 human splice isoforms annotated in UniProt. We then analyzed the structures of our isoforms plus the 127,000 isoforms folded by Sommer et al. using a variety of metrics to identify splicing-induced structural changes. We identified numerous ways in which splicing affected structure, including changes in secondary structure, surface charge, protein compactness, and the surrounding environment of post-translational modification sites. Additionally, we integrated the predicted structural information with expression data from the Tabula Sapiens, which includes scRNA-seq experiments spanning the whole human body. Finally, we used structure-based function prediction to evaluate the functional consequences of alternative splicing. Our workflow is summarized in **Figure 1**.

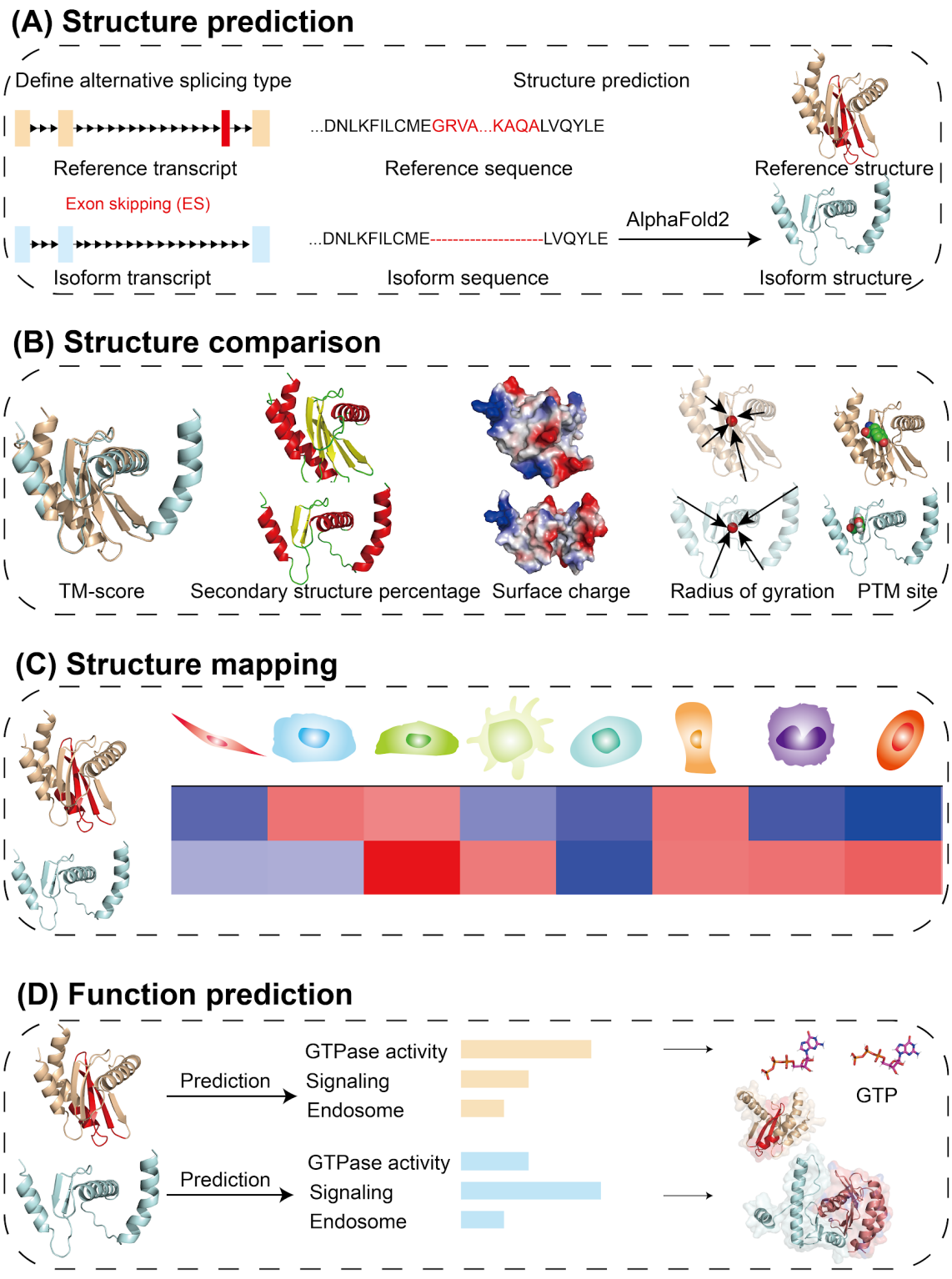

Figure 1: Workflow for exploring the structural effects of alternative splicing. The workflow is composed of three parts: **(A)** Structure prediction: we used AlphaFold2 to predict the structure of 11,161

human alternative spliced isoforms from 5,966 genes. The structures for the "reference" isoform of each gene are publicly available. For each isoform, we annotated the alternative splicing type based on the pattern of spliced exons relative to the reference isoform. **(B)** Structure comparison: we compared the structures of isoforms for each gene using five different metrics. We calculated template-matching score (TM-score), secondary structure percentage, surface charge, radius of gyration, and solvent-accessible area of post translational modification sites. **(C)** Structure mapping: We quantified the expression for each isoform from the Tabula Sapiens scRNA-seq dataset using Kallisto and identified isoform expression differences across human cell types. **(D)** Function prediction: We used COFACTOR to predict protein functions based on their structures and compared the predicted gene ontology (GO) terms for reference and isoform.

2 Results

2.1 Assessment of AlphaFold2 structure prediction for human alternative spliced isoforms

Recent studies have successfully predicted the three-dimensional structures of the human proteome[26,27]. These works capture only one structure per gene, however, neglecting the complexity of alternative splicing where multiple isoforms can be expressed. Recently, Sommer et al. (2022) predicted the structures for over 127,000 human transcripts from the CHESS database of gene annotations. The goal of the Sommer study was to use predicted structures to nominate a canonical or "reference" isoform of each gene by identifying well-folded gene products[19,28]. A key question is left unanswered in that work, however: What is the overall structural impact of human alternative splicing? To begin to answer this question, we identified all human splice isoforms in SwissProt. After filtering to a maximum of 600 amino acids and removing the "reference" isoforms which have already been folded, we obtained 11,159 isoforms. We then compared our predicted isoform structures with the publicly available structures for the 5,966 isoforms annotated as "reference" isoforms in the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/). The advantage of using isoforms from SwissProt is that many annotations are available, such as the locations of post-translational modification sites. We also incorporated the structures predicted by Sommer et al. in our analyses (see below).

While the structures predicted by AlphaFold2 are generally high-quality with well-calibrated confidence scores[29], AlphaFold2 predictions remain inaccurate in certain contexts. For example, AlphaFold-Multimer and AlphaMissense were developed to address limitations in the prediction of protein complexes and the structural impact of point mutations[30–32]. A key insight from these and other works is that the quality of the structure prediction in deep-learning based structure prediction methods like AlphaFold2 depends on the abundance and quality of multiple sequence alignments (MSAs) for the targets to be folded[33]. Because alternative splicing is more prevalent in higher eukaryotes[34], the conservation across variable exons in alternatively spliced isoforms may vary. This variation in MSA depth could influence the quality of the predicted isoform structures and potentially confound overall interpretation of the structural impact of alternative splicing.

We thus investigated how variability in MSA information affects the prediction of structures for alternatively spliced transcripts. To do this, we extracted the MSA depth for each residue across the isoforms we folded and mapped the different alternative splicing regions marked as “replaced” or “missing” and non-AS regions with their MSA counts (**Figure 2A**). As expected, non-AS regions exhibited a higher MSA depth compared to AS regions (ANOVA p-value: 0.0). We also compared the MSA depth of exons that were missing from the reference (“missing”) or replaced with a difference sequence not found in the reference (“replaced”) (**Figure S1B**). The replaced regions in the isoforms displayed the lowest MSA counts, lower than the missing regions, which in turn had fewer MSA counts than the regions that were shared between reference and alternate isoforms (“reference other” and “isoform other”). The structure confidence score (predicted local distance difference test, pLDDT) values for isoform structures reflected the differences in MSA depth, with replaced regions having the lowest pLDDT, missing regions slightly higher, and unchanged regions having the highest pLDDT. However, some structures in each category were folded with high confidence. Importantly, the pLDDT scores depended on the MSA depth in a similar way regardless of whether the sequence was a reference or alternate isoform (**Figure 2B**, Pearson correlation of 0.475 for reference isoforms and 0.529 for alternate isoforms). Consequently, the AS regions demonstrated lower structural quality compared to non-AS regions (**Figure 2C**). While the overall distribution of pLDDT scores was broadly similar, the average pLDDT was slightly lower for predicted isoform structures compared with reference structures (**Figure 2D**). This trend was also observed by Sommer et al 2023 when they folded isoforms from the CHESS database[19] (**Figure S2A**). In summary, we find that MSA depth does influence the structural prediction quality of AS regions, but pLDDT reflects this dependence for both reference and alternate isoforms. Thus, it should be possible to use pLDDT to filter out isoforms whose structures cannot be accurately predicted due to low MSA depth.

To further investigate the accuracy of AlphaFold2 predictions for isoform structures, we searched the PDB for proteins with experimentally determined structures for multiple isoforms. We identified 11 such proteins with experimental structures for at least 2 distinct isoforms and for which the structures include the alternatively spliced region (**Table S1**). Of the 22 experimentally-determined structures meeting these criteria, 20 were resolved using X-ray crystallography, one (UBE2V1) was determined using Nuclear Magnetic Resonance spectroscopy, and one (ELOC) was determined by Electron Microscopy. We excluded ELOC and UBE2V1 from the comparison because of the low resolution (8.20 Å) of the ELOC isoform structure and lack of resolution information of the UBE2V1 structure. The other structures determined by X-ray crystallography have resolutions ranging from 1.45 Å to 3.72 Å. Reassuringly, the AlphaFold2 predictions matched the experimentally determined structures equally well for both reference and alternate isoforms (**Figure 2E-F**). The TM-score and RMSD for each experimental-predicted structure pair were not significantly different between isoform and reference structures, with paired T-test p-values of 0.83 and 0.56 (**Figure 2E-F**). We note that AlphaFold2 was trained on the PDB, likely including these structures. We folded these sequences using a template-free version of AlphaFold2 to ensure that the PDB structures were not directly used in the prediction. It is possible that the results of our comparison are overly

optimistic in assessing the quality of isoform structure prediction because the structures were in the AlphaFold2 training dataset. However, the analyses in the AlphaFold2 paper indicate that the model is robust to small changes in training data[16], so these predictions would likely still be similar if AlphaFold2 were re-trained without the structures included. Overall, we take this as promising but not definitive evidence that AlphaFold2 can accurately predict isoform structures.

Given the broad overall similarity in structural quality between reference and alternate isoforms and trusting the pLDDT as a measure of structural quality, in the next sections we proceed to explore the structural differences between 4,450 reference and 7,631 isoform structures with high prediction quality (pLDDT >= 70), while being aware of the limitation of the analysis that differences in MSA depth in alternatively spliced regions may lead to lower structural quality in some cases.

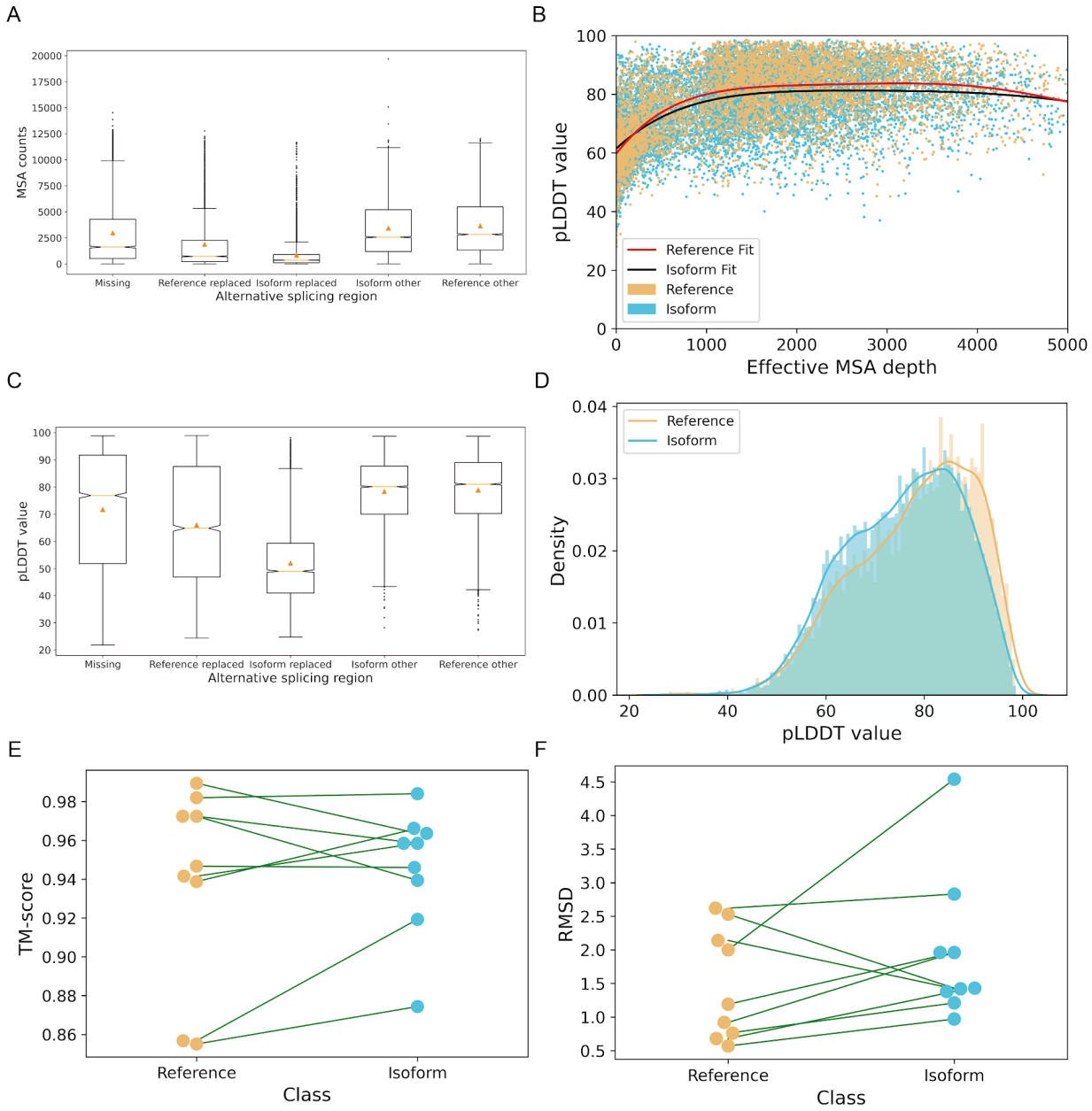

Figure 2: Reliability assessment of AlphaFold2 predictions for human alternative spliced isoforms. (**A**) Box plots (orange triangle: mean, box: 25-75% quantile range, dots: outlier values) of MSA counts across alternatively spliced vs. constitutive regions. “Missing”: sequence present in the reference but not alternate isoform. “Reference replaced”: sequence present in the reference but not alternate isoform, but this region is replaced by a new region in alternate isoform. “Isoform replaced”: sequence present in the alternate isoform but not the reference. “Isoform other” and “Reference other”: sequence present in both the reference and alternate isoforms. A diagram explaining the differences among these types of splicing is shown in **Figure S1B**. (**B**) Scatter plot of pLDDT value by effective MSA depth for reference and alternative isoform structures. Polyfit trend lines for reference and isoform are colored in black and red. (**C**) Box plots of pLDDT values across alternatively spliced vs. constitutive regions. ANOVA shows that differences in mean msa counts and pLDDT among missing, reference replaced, isoform replaced, isoform other and reference other regions are significant (p-value < 1.33e-301 for msa and p-value <

5.04e-295 for pLDDT). (**D**) Density plot of prediction quality for 5,966 reference and 11,161 isoform structures predicted by AlphaFold2 from sequences annotated in SwissProt. Distribution of prediction quality for CHESS dataset refers to **Figure S2A**. There are 9 genes that have experimental structures with resolution greater than 3 Å deposited in Protein Data Bank for both the reference and an alternate isoform. For these pairs, (**E**) shows the TM-score (paired t-test p-values: 0.76) and (**F**) shows the RMSD (paired t-test p-values: 0.32) comparing the AlphaFold2 predicted structures against the ground truth experimental structures.

2.2 Structural similarity of isoforms generally reflects sequence similarity, but some structures diverge despite similar sequences

A sensible initial hypothesis about the effects of alternative splicing on protein structure is that removing or replacing large parts of the reference sequence should have large effects on the structure. Conversely, mutually exclusive exons with similar size or small changes in exon boundaries should generally result in smaller structural changes. To explore these effects, we examined the relationship between the sequence length of the alternatively spliced region and the template matching score (TM-score) for each isoform. We annotated each of the splicing events in our set of isoforms using the following splicing events: exon skipping (ES), alternative donor site (ADS), alternative acceptor site (AAS), mutually exclusive exons (MXE), intron retention (IR), alternative first exon (AFE), alternative last exon (ALE), mutually exclusive exon-alternative first exon (MXE-AFE) and mutually exclusive exon-alternative last exon (MXE-ALE). Note that multiple types of splicing events may occur in a given isoform. To quantify the degree of sequence change, we calculated the length of each splicing event. Our findings indicate a negative correlation between the TM-score and length of alternatively spliced sequence across all nine types of alternative splicing, with the strongest relationship for MXE-ALE, with a PCC of -0.547 (**Figure 3A**). This indicates that, as expected, larger sequence changes due to splicing tend to cause larger structural changes.

We additionally calculated percent sequence identity based on sequence alignment between isoform and reference. We observed a positive correlation between sequence identity and TM-score (PCC=0.570) (**Figure 3B**). This indicates again that the isoforms of a gene generally have more similar structures when they share a higher sequence identity. We confirmed a similar relationship in the isoforms of the CHESS dataset (**Figure S2B**). Some isoforms can have very low sequence identity compared to the reference isoform (<30%), such as truncated isoforms that are much shorter compared to their references. In such cases, the TM-score between reference and alternate isoform structure is always correspondingly low.

However, this analysis highlighted a subset of isoforms that have low structural similarity despite high sequence similarity. These examples are particularly interesting to examine in our study, because the structural effects of alternative splicing cannot be predicted from the length of the spliced portion of the isoform and thus require structures to detect. 328 isoforms were characterized by high sequence identity (>70%) but low TM-score (<0.5) (**Table S2**). Among our high-quality structures, we identified 53 isoforms where the structured domains were altered owing to alternative splicing, and thus may have functional implications. We noticed that the

structural differences for the 53 isoforms largely fall into three groups: different helix orientation (*CENPX*, *SNAP23* and *IL32,* etc.), loss of repeat domain (*SIGLEC8*, *CFHR3* and *BTN3A1,* etc.), and change in structural compactness (*LIAS*, *NDRG2* and *MDH2,* etc.) (**Figure 3C**). For example, the CFHR3 reference isoform contains five repeated Suchi domains, which mediate cytokine binding[35], while CFHR3 isoform 2 lacks one Suchi domain, which may affect its normal function. One particularly interesting case is the Synaptosomal-associated protein 23 (*SNAP23*) (**Figure 3D**), which has a sequence identity of 74.41% but a TM-score of 0.474 between the reference (SNAP-23a) and the isoform (SNAP-23b). An exon skipping event at exon 5 causes a loss of 53 residues in SNAP-23b. Lack of a long flexible loop completely reverses the orientation of a long helix between SNAP-23a and SNAP-23b (**Figure 3D** top). Consequently, the N- and C-terminal residues are adjacent in the SNAP-23a structure but at opposite ends in SNAP-23b (**Figure 3D** bottom). The remaining 275 isoforms contain long unstructured loop regions and cause the low TM-score between the reference and isoform. Because loops can be highly flexible, their precise structures are difficult to predict computationally and thus it is difficult to draw many conclusions about them. Gene ontology (GO) enrichment analysis for the genes associated with these 275 isoforms suggests that their functions can be broadly categorized into immune response (CD244, CD1E, CD276, etc.) and protein transportation (CTLA, SYT4, SYT15, etc.) (**Figure S3**). For example, the structures of the immune response receptors are presented in a similar pattern: an Ig-like extracellular domain, a single-helix transmembrane domain followed by an unstructured cytoplasmic region; variation in the linker regions or the cytoplasmic region could result in different orientation of domains which will lead to a low TM-score in the structural alignment.

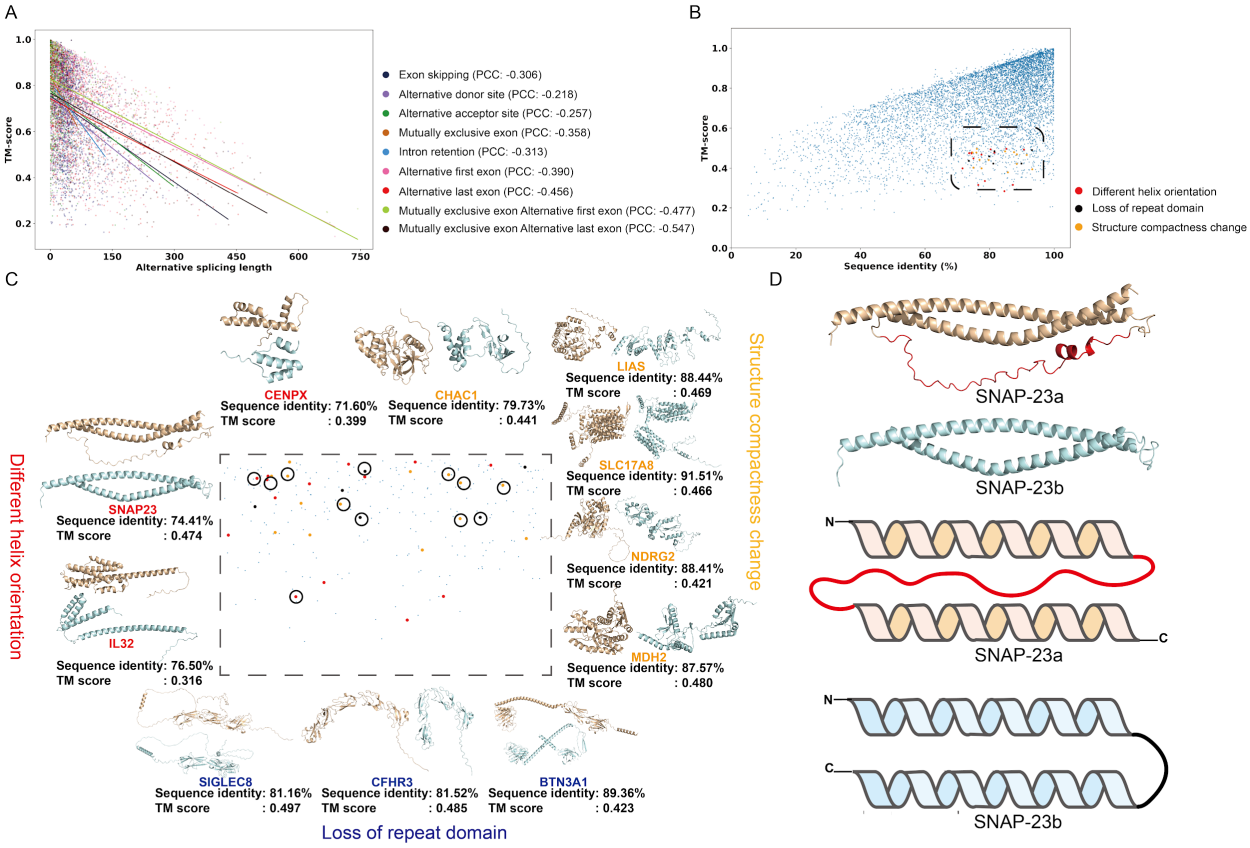

Figure 3: **Structural similarity of isoforms generally reflects sequence similarity, but some structures differ despite similar sequences.** (**A**) Scatter plot of length of alternative splicing region (x-axis) vs. TM-score between reference and alternate isoform structure (y-axis) colored by alternative splicing type, fitted lines are presented for each alternative splicing type. (**B**) Scatter plot of percent sequence identity between reference and alternate isoform (x-axis) vs. TM score (y-axis). (PCC=0.565). Proteins with high TM-score but low sequence identity are shown in the box and colored by type of structural change. The same analysis for CHESS dataset is shown in **Figure S2B**. (**C**) Examples of isoforms with high sequence identity (>70%) but low TM score (<0.5) categorized into three classes: structure compactness change, different helix orientation and loss of repeat domain. The isoforms predominantly influenced by structural irregularities in loop regions are illustrated in **Figure S3**. The observed variability in the loop regions may arise either from inherent structural disorder or as a consequence of AlphaFold2 predictions characterized by lower accuracy.. (**D**) Structures and simplified diagrams for *SNAP23*. The reference (SNAP-23a) is shown in wheat color, and the isoform (SNAP-23b) is shown in cyan. Alternatively spliced regions are colored in red.

2.3 Alternative splicing changes secondary structure, surface charge, and radius of gyration

Our predicted structures allow us to investigate structural changes that are not obvious from sequence changes but have important implications for protein function. To investigate how splicing may change such properties, we calculated the surface charge and radius of gyration for our predicted structures and quantified differences in these metrics between reference and alternate isoforms. We also investigated how splicing changed secondary structures (helix,

sheet and loop) by comparing the percentage of each secondary structure for each alternative splicing type (**Figure S4A**). For most alternative splicing types, the distribution of secondary structure percentage between isoform and reference is similar; however, exceptions occurred at the beginning and end of proteins. For example, alternative first exon events tend to create more helix and less loop in the isoform. This phenomenon is even more obvious in the CHESS dataset (**Figure S4B**).

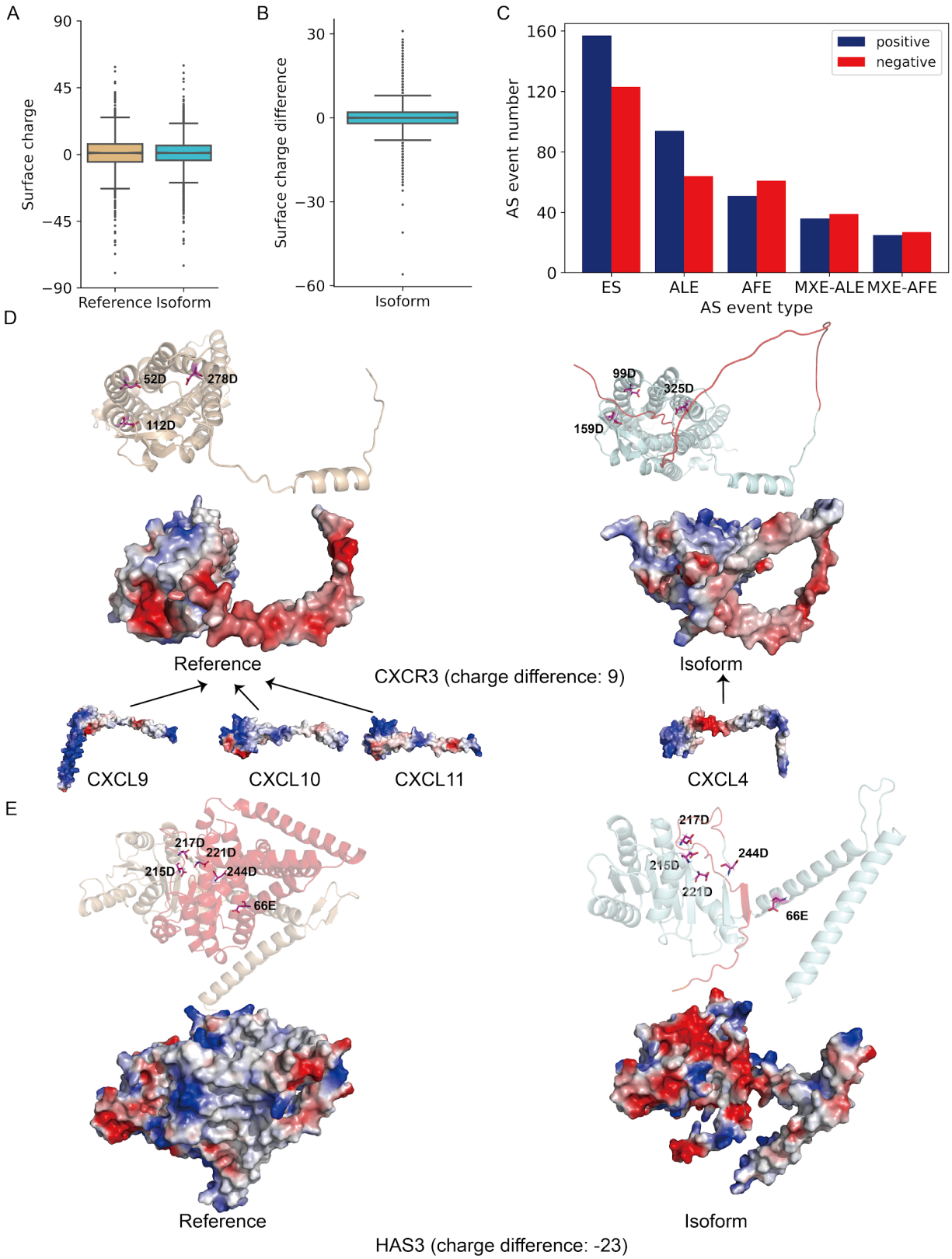

Figure 4: Alternative splicing changes surface charge. (**A**) Overall surface charge distribution between reference and isoform structures (p-value: 0.304 Mann–Whitney U test). Surface charge

distribution for the CHESS dataset is shown in **Figure S2C**. (**B**) Differences of surface charge for the SwissProt dataset. Difference of surface charge is calculated by using the isoform surface charge minus the reference surface charge. The difference of surface charge for the CHESS dataset is shown in **Figure S2D**. (**C**) The five most frequent alternative splicing events in the positive and negative surface charge outliers. For the CHESS dataset, it is presented in **Figure S2E**. (**D**) Example of positive surface charge outliers: CXCR3. The reference isoform binds different ligands (CXCL9, CXCL10 and CXCL11) compared to the ligand that binds the alternative isoform (CXCL4). The three aspartic acid residues are labeled in pink. (**E**) Example of a negative surface charge outliers: HAS3, acidic amino acids are labeled in pink sticks.

In contrast to structural similarity and secondary structure percentage, which predominantly influence protein stability, structural attributes such as surface charge and radius of gyration may carry greater significance in shaping alterations in protein function. For example, the location of charged residues is crucial for electrostatic interactions in which proteins participate. Surface charge thus plays a pivotal role in various processes, including protein ion binding and protein localization[36,37]. Additionally, the overall compactness or looseness of a structure can dramatically alter protein functions such as transport or catalysis. The radius of gyration serves as a metric for evaluating protein compactness and may serve as an indicator of potential intrinsically disordered regions[38,39].

We calculated the total surface charge by first identifying surface residues based on the relative solvent accessibility (RSA) for each residue of the structure, then summing the charges of the surface residues. The overall distribution of surface charge does not show significant difference between reference and alternate isoforms (Mann–Whitney U test p-value: 0.304) (**Figure 4A**), with a median of 0.99 in the reference group and 1.00 in the isoform group. We observe the same result in the CHESS dataset (**Figure S2A**). However, if we compute the difference in surface charge between each isoform structure and its corresponding reference structure, we observe outliers (defined as Q1 - 1.5 IQR and Q3 + 1.5 IQR) where the isoform surface charge increased or decreased significantly (**Figure 4B**). We observe 231 positive surface charge outlier isoforms, where alternate isoforms have from 9 to 31 more units of surface charge than the reference isoforms. There are 214 negative surface charge outlier isoforms, where isoforms have between -56 and -9 units of surface charge compared to the references. The five most frequent alternative splicing types among isoforms with significant surface charge changes are ES, ALE, AFE, MXE-ALE, and MXE-AFE (**Figure 4C**). Interestingly, positive and negative surface charge changes are about equally likely to occur for AFE, MXE-ALE, and MXE-AFE, but ES and ALE more often increase surface charge. A notable example among these positive outlier isoforms is C-X-C chemokine receptor type 3 (*CXCR3*). An AAS event in isoform 2 causes a 47 residue insertion in the N-terminal portion of the protein; the extended loop inserts into the binding pocket of CXCR3 and buries 99D, 159D and 325D, which are all positively charged aspartic acid residues (**Figure 4D** top). In addition, this insertion causes a much more positively charged surface in CXCR3 isoform 2, which could explain the binding preference between the reference and the isoform, where the ligands (*CXCL9*, *CXCL10* and *CXCL11*) for the reference have much a more positively charged ‘head’ than the ligand (CXCL4) binding for the isoform[40] (**Figure 4D** bottom). Hyaluronan synthase 3 (*HAS3*) is an interesting example from among the negative outlier isoforms. An MXE-ALE event in the isoform removes a large amount

of the structure (272 residues) and exposes five negatively charged residues: 66E, 215D, 217D, 221D, and 244D, creating a more negatively charged surface in the isoform (**Figure 4E**).

The overall distribution of radius of gyration is similar between reference and alternate isoforms, but the median radius is slightly lower for alternate isoforms (28.59 Å vs 27.46 Å) (**Figure 5A**). Thus, the distribution of the difference of radius of gyration (isoform radius minus reference radius) is negatively skewed, with a median of -0.73 Å, with outliers (defined as Q1 - 1.5 IQR and Q3 + 1.5 IQR) occurring at both ends (**Figure 5B**). This trend likely reflects a general loss of sequence and shorter overall protein length in alternate isoforms relative to the reference isoform. The five most frequent alternative splicing types among isoforms that are outliers in terms of radius of gyration changes are ES, AFE, ALE, MXE-ALE, and MXE-AFE, and there are more negative outliers than positive outliers (**Figure 5C**). ES and ALE are especially likely to decrease the radius of gyration. Negative radius of gyration outliers often have decreased radius due to shorter overall sequence and loss of domains; examples include *IL1R2*, *P2RX7*, *VAPB* and *ITM2A* (**Figure S5**). In contrast, positive outliers often have a more unfolded, less compact structure compared to the reference, like mitochondrial Lipoyl synthase (*LIAS*). An alternate isoform of LIAS lacks 50 residues from the protein core compared to the reference, leading to a less compact structure whose radius of gyration increases by 10.708 Å (**Figure 5D**). One consequence of such loose structure is that one 4Fe-4S cluster binding site is close to the binding cluster in reference (352S); in the isoform this binding site is retained, although it is away from other binding sites (309S) (**Figure 5D**). The 4Fe-4S cluster binding site engages in the binding of the 4Fe-4S cluster, which functions as the electron donor during the half-reaction of reduction of S-adenosylmethionine (SAM) [41,42], and the resulting radical will attack the sulfide of 4Fe-4S cluster, leading to lipoic acid synthesis followed by another similar half-reaction. In the LIAS alternate isoform, the 309S binding site is spatially distinct from other 4Fe-4S cluster binding sites, potentially influencing electron transfer, SAM cleavage, and consequently, the synthesis of lipoic acid.

We also observed some isoforms that have nearly the same length as their reference structures but a much higher radius of gyration. In such cases, the change in radius of gyration often reflects a change in the position of the protein's center of mass. For example, as shown in **Figure 5E**, the center of mass for Calcium homeostasis modulator protein 4 (*CALHM4*) and C-type lectin domain family 10 member A (*CLEC10A*) isoforms is outside their structure in both cases, while the center of mass for the reference is inside the structure, which leads to a much higher radius of gyration in the CALHM4 and CLEC10A alternate isoforms. Some isoforms undergo even more extreme changes, like one Noelin-3 (*OLFM3*) alternate isoform, where ES and MXE-ALE events cause the isoform to form a single helix, very different from the reference, which contains a globular structure composed largely of sheets (**Figure 5F** top). As another example, an isoform of Sodium-dependent phosphate transport protein 1 (*SLC17A1*) undergoes an ES event that causes the protein to form two helical bundles in the isoform rather than a single transport channel as in the reference structure (**Figure 5F** bottom).

To summarize the relationship between alternative splicing types and structural properties examined so far, we performed a regression analysis to predict the values of TM-score, surface

charge difference, secondary structure percentage difference, and radius of gyration difference as a function of the number of residues changed by each alternative splicing type. This regression model provides a principled way to control for confounding factors, including total protein length and structure quality. The effect size represents the per-residue effect of each alternative splicing type on TM-score, secondary structure percentage (helix, sheet, loop), surface charge and radius of gyration.

The regression analysis indicates that most alternative splicing types induce a statistically significant decrease in TM-score, with the exception of MXE (**Figure S6**). This could be because MXE often results in relatively small sequence changes. Furthermore, ES, ALE, and MXE-ALE stand out as the three alternative splicing events exhibiting the most pronounced and statistically significant negative per-residue impact on the TM-score. Alternative splicing has little per-residue effect on the secondary structure percentage; only splicing events at the beginning and end of the protein significantly affect the secondary structure (**Figure S6**). Interestingly, intron retention in the alternate isoform shows a significant positive effect on surface charge. We also found that alternative splicing types leading to loss of residues, like exon skipping, exert a positive per-residue influence on the difference of radius of gyration. This makes sense, because isoforms with fewer residues will generally have a lower radius of gyration.

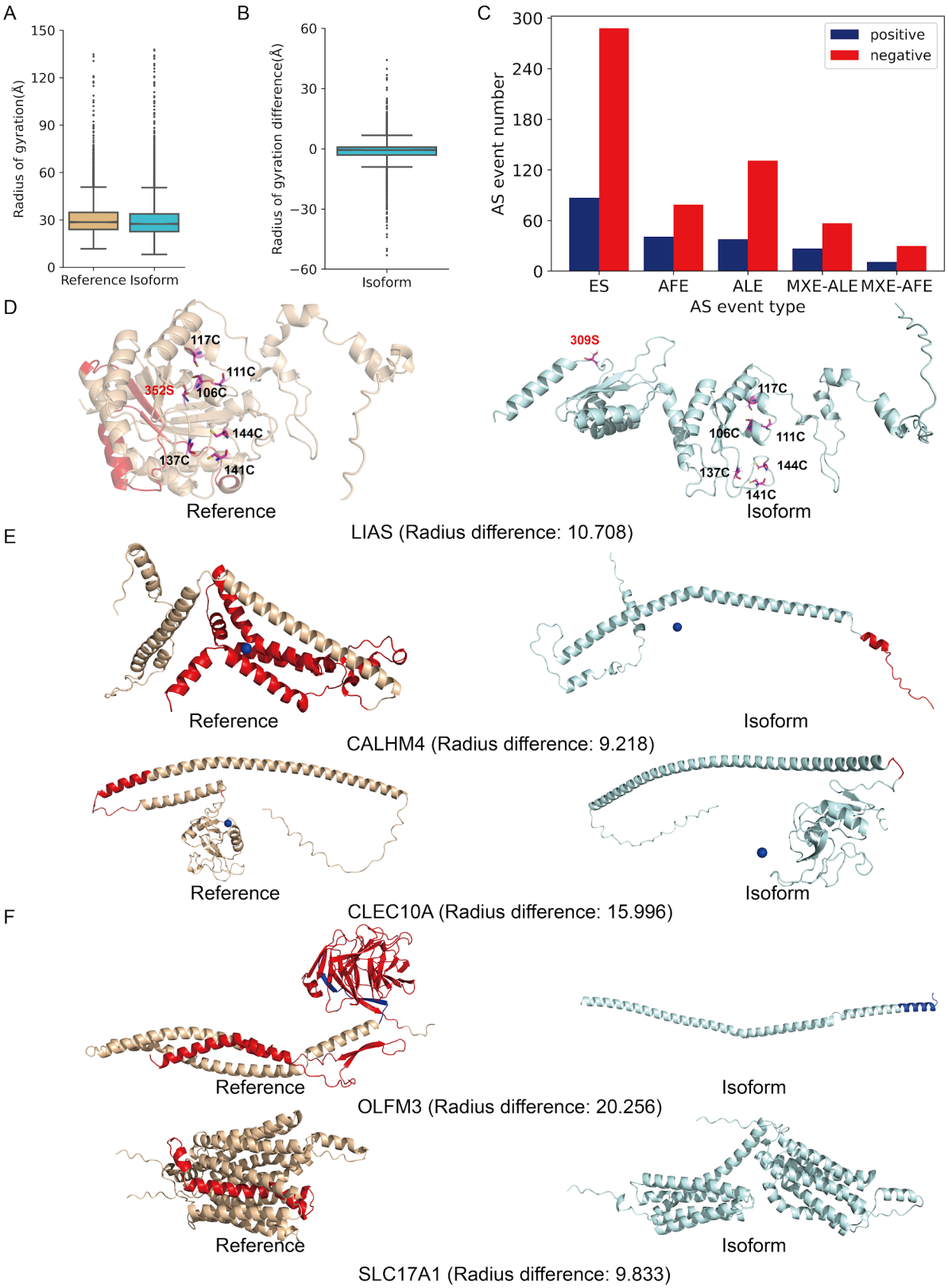

Figure 5: Alternative splicing changes protein compactness as measured by radius of gyration.
(**A**) Box plot of radius of gyration values for reference and alternate isoform structures (p-value: 2.876e-11

Mann–Whitney U test). The median difference between reference and alternate isoform radius is -0.73 Å. Radius of gyration distribution for the CHESS dataset is shown in **Figure S2F**. (**B**) Differences in radius of gyration for the Swissprot dataset. Results for the CHESS dataset are shown in **Figure S2G**. (**C**) The five most frequent alternative splicing events in positive and negative outliers. Results for the CHESS dataset are shown in **Figure S2H**. (**D**) Reference and alternate structure for LIAS isoforms. The Fe-S binding site residues are shown in pink sticks. The binding site residues which change position (352S in reference and 309S in isoform) relative to other binding site residues are labeled in red text. Binding site positions are labeled in red. Alternatively spliced portions of the structure are colored red. (**E**) Examples of radius of gyration changes caused largely by a shift in the center of mass (CALHM4 and CLEC10A). The center of mass for each structure is indicated with a blue dot. Alternatively spliced portions of the structure are colored red. (**F**) Radius of gyration differences indicate dramatic structural changes between reference and alternate isoforms for OLFM3 and SLC17A1. Alternative first exon of the structure is colored in red. Mutually exclusive regions are colored in blue.

2.5 Alternative splicing buries and exposes post translational modification sites

Protein function is regulated by the addition of various post translational modifications (PTMs) to amino acid residues, including phosphorylation, ubiquitination, acetylation, methylation and glycosylation[43]. Modifications can lead to significant conformational changes, particularly in the case of phosphorylation. Furthermore, PTMs are added by enzymes that must be able to access the residue to be modified; for example, phosphorylation requires contact with specific kinases. Thus, phosphorylation sites buried inside the protein are more difficult for kinases to access compared to PTM sites that are solvent-accessible. If structure changes induced by alternative splicing modify the positions of PTM sites, these changes could have significant functional implications. Such changes are especially interesting in this study, because they cannot be directly predicted from sequence differences between reference and alternate isoform but can readily be predicted from structure. We therefore searched our predicted structures for evidence that alternative splicing can cause structural changes that bury or expose PTM sites.

To quantify changes in PTM site location, we calculated the relative solvent-accessible area (RSA) for each PTM site and determined whether the residue was solvent-accessible (exposed) or inaccessible (buried). We then grouped the PTM sites into five classes based on how they changed between reference and alternate isoforms: unchanged, spliced out, spliced in, buried to exposed and exposed to buried. We identified a total of 587 exposed PTM sites in the reference structures that become buried in the alternate isoform and 1,358 PTM sites that are exposed to the solvent in the reference but buried in the alternate isoform (**Table S3**). For example, in the Cyclin-dependent kinase 1 (*CDK1*) isoform, the 15T phosphorylation site is buried in the reference (RSA: 20.8 $Å^2$), but is exposed (RSA: 175.6 $Å^2$) due to the loss of a helix in the alternate isoform (**Figure 6A**). Similarly, the 200T phosphorylation site is exposed (RSA: 79.1 $Å^2$) in the TGF-beta receptor type-1 (*TGFBR1*) reference, but the loop orientation buries (RSA: 16.5 $Å^2$) this site inside the *TGFBR1* alternate isoform (**Figure 6B**).

Such changes in the surrounding environment of PTM sites could be a mechanism by which alternative splicing regulates protein function. For example, the apoptosis regulator BAX

(encoded by gene *BAX*) induces apoptosis under stress conditions, while phosphorylation at 184S by RACα serine/threonine-protein kinase (AKT1) causes BAX to prevent apoptosis[44]. In the BAX reference isoform (BAXα), 184S is located inside a helix and the side chain is oriented toward the inside of the helix with a RSA of 1.18 $Å^2$. However, in the structure for the BAXδ isoform, the corresponding residue 135S is exposed to the solvent with an RSA of 55.5 $Å^2$ (**Figure 6C**). Moreover, the 184S site in BAXα can form polar contacts with the residues from a parallel helix, while the corresponding residue (135S) in BAXδ only has polar contacts with the residues in the same helix, which makes this site highly flexible in BAXδ but rigid in BAXα (**Figure 6C**). In summary, the PTM site is almost completely inaccessible in BAXα but highly accessible in BAXδ. Considering the critical role that phosphorylation of 184S plays in apoptosis regulation, this splicing-induced change in the structural context of the PTM site could have significant functional implications. Interestingly, we identified several other differences in PTM site locations among BAX isoforms, including BAXζ and BAXσ (**Figure S7**). The structure of BAXζ is similar to BAXδ: 106S, which corresponds to 184S in BAXα, has an RSA of 51.6 $Å^2$, and the polar contact is between the residues within the same helix (**Figure S7A**). While BAXσ has a more compact structure, the PTM site (171S) is still exposed (RSA: 53.5 $Å^2$) and forms polar contacts within the helix (**Figure S7B**). Because of the protein's prominent role in regulating apoptosis, differences among BAX isoforms are especially interesting in the context of cancers. We investigated the expression of BAX isoforms across human cancer types using the TSVdb (TCGA Splicing Variants DB)[45]. We stratified donors based on high vs. low expression of the BAXδ isoform across several cancer types **(Figure 6D).** Compared with the low expression group, we observe significantly higher survival for kidney renal clear cell carcinoma, liver hepatocellular carcinoma and mesothelioma but lower survival for ovarian serous cystadenocarcinoma in high expression BAXδ isoform group (**Figure 6D**).

The statistics for all metrics discussed so far including sequence identity, TM-score, secondary structure percentage, surface charge, radius of gyration and PTM analysis are listed in **Table S4.**

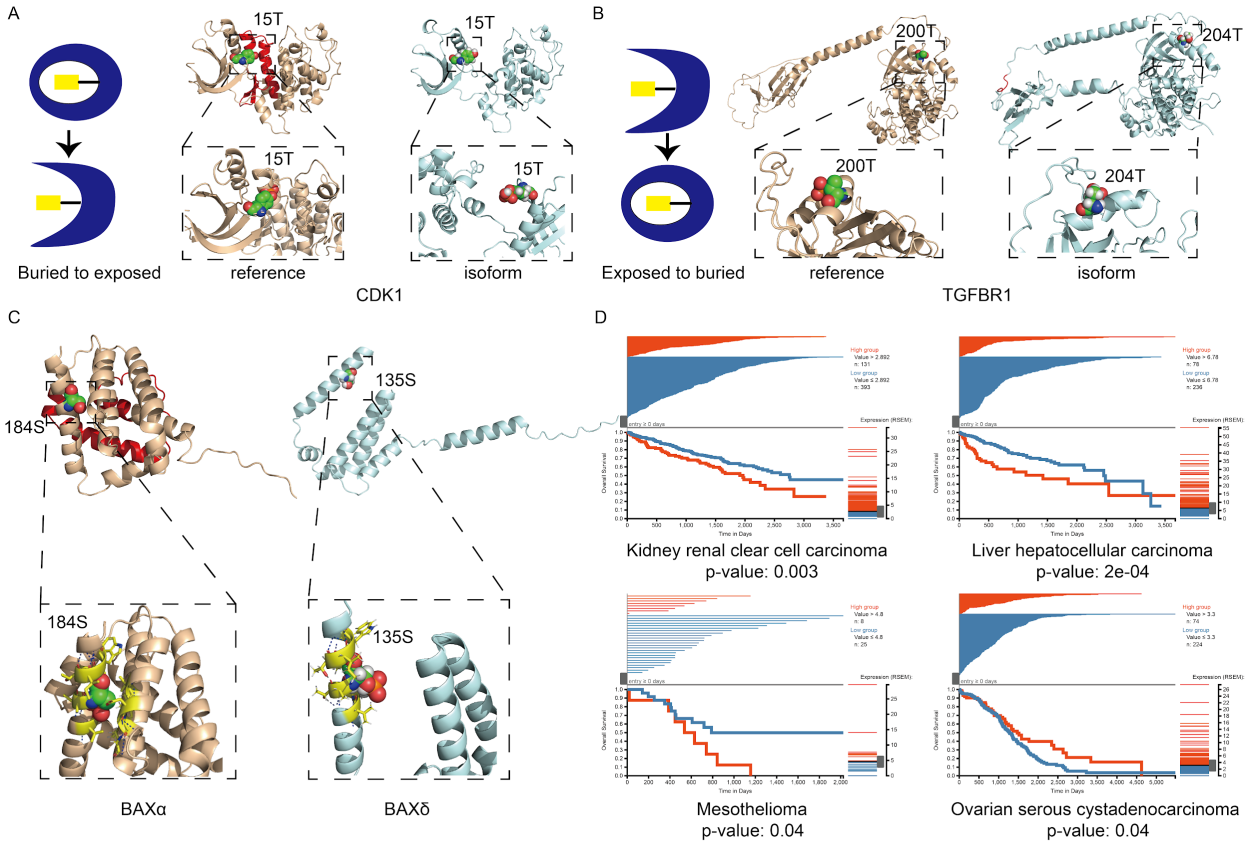

Figure 6: Alternative splicing buries and exposes post translational modification sites. (**A**) PTM sites changed from buried to exposed based on the threshold of relative solvent accessibility (RSA), the 15T in CDK1 reference is buried (RSA: 20.8 Å^2), and in CDK1 isoform the 15T is exposed (RSA:175.6 Å^2). (**B**) PTM sites changed from exposed to buried, the 200T in TGFBR1 reference is exposed (RSA: 79.1 Å^2), but buried in the CDK1 isoform (RSA:16.5 Å^2). (**C**) The 184S buried in BAXα (RSA: 1.18 Å^2), and the corresponding PTM site 135S is exposed in BAXδ (RSA: 53.5 Å^2). The residues within 4Å of the PTM site are represented in yellow sticks, and the polar contacts with those residues are represented in blue lines. (**D**) Survival plot for patients with high vs. low expression of BAXδ in four cancer types. We split the sample into high and low expression groups using the 75th percentile and calculated the p-value using a log-rank test between high and low expression groups.

2.5 Mapping isoform expression across human cell types

Using single-cell RNA sequencing (scRNA-seq) data, differential isoform usage has been discovered from a variety of cellular contexts, such as different mouse neuron types and also different human tissue compartments[21,46]. An intriguing phenomenon lies in the dynamic switching of isoform usage within the same gene across diverse cell types, potentially shedding light on the distinctive functions of cell-type-specific protein isoforms. To investigate these effects, we conducted an analysis to pinpoint the predominant isoform usage within the same gene across 133 cell types, drawing from the Tabula Sapiens dataset. While many isoforms showed cell-type-specific expression, the majority of these involved differences in the expression level of the dominant isoform, rather than a true isoform switch in which an alternate

isoform replaces the reference isoform as the highest-expressed form of a gene. Thus, we used stringent filtering criteria to identify genes showing true isoform switching. This revealed 33 genes, encompassing 71 modeled isoforms, exhibiting a notable shift in isoform usage across various cell types (**Table S5**).

Qualitatively, these genes show clear cell-type-specific isoform switches (**Figure 7A**). Some of the genes are well known examples of alternative splicing. For example, Myosin light polypeptide 6 (*MYL6*) expresses a non-muscle form and a smooth muscle form[47], and owing to its high expression level, it has been validated as differentially expressed between cell types from other scRNA-seq studies[46,48,49]. Analogous to MYL6, a gene associated with muscle function, tropomyosin (*TPM*), whose isoforms exhibit diverse binding dynamics with the actin filament[50], similarly demonstrates distinctive isoform usage. Specifically, we observe variations in isoform utilization concerning three members of the *TPM* family: *TPM1*, *TPM2* and *TPM3* (**Table S5**).

Another example is Superoxide dismutase [Mn], mitochondrial (*SOD2*), where the SOD2 isoform 2 lacks 39 residues compared to the SOD2 reference. We select the cell types which differentially expressed SOD2 reference and isoform 2 in **Figure 7B**. Compared with the expression of SOD2 reference (**Figure 7C**), SOD2 isoform 2 is highly expressed in cell types including muscle cell, stromal cell and endothelial cell of vascular tree (**Figure 7D**). Our single-cell analysis suggests that SOD2 reference is differentially expressed in 72 cell types including mature enterocyte, NK cell and B cell, while isoform 2 is highly expressed in muscle cell, stromal cell and endothelial cell of vascular tree (**Table S5**). The SOD2 reference structure (**Figure 7E**), is quite different from the SOD2 isoform 2 structure (**Figure 7F**), where an ES event causes a helix loss in the isoform. Interestingly, a critical residue 98H manganese ion (Mn^{2+}) binding site is lost in SOD2 isoform, and it is believed to decrease the activity of isoform 2[51]. We also observe that in the reference structure, the four Mn^{2+} binding sites (50H, 98H, 183D and 187H) are clustered together, while in the isoform structure, the remaining binding sites do not form a binding cluster with 50H located away from the 144D and 148H, which could also explain the decrease of superoxide dismutase activity. Additionally, proteins like ER lumen protein-retaining receptor 1 (KDELR1), Ras-related protein Rab-1A (RAB1A) and phospholipid transfer protein (PLTP) also show significant differential expression between isoform and reference transcripts (**Figure S8**). However, for the above isoforms, even though they have distinct expression patterns, without experimentally determined functional annotation, it is hard to link the differential expression, cell type function and protein structure.

Interestingly, we also observed that genes which are differentially expressed among cell types like *TPM3*, *OLFM3* and *SOD2*, also are differentially expressed in the same cell types across different tissues (**Table S6**). We identified six genes with 12 isoforms that have distinct expressions in seven tissues. For example, among classical monocytes from different tissues, TPM3 reference has higher expression in blood, while TPM3 isoform 5 has higher expression in lung. Similarly, the SOD2 reference isoform has higher expression in endothelial cells from spleen, vasculature, trachea and fat, while SOD2 isoform 2 is more highly expressed in skin endothelial cells.

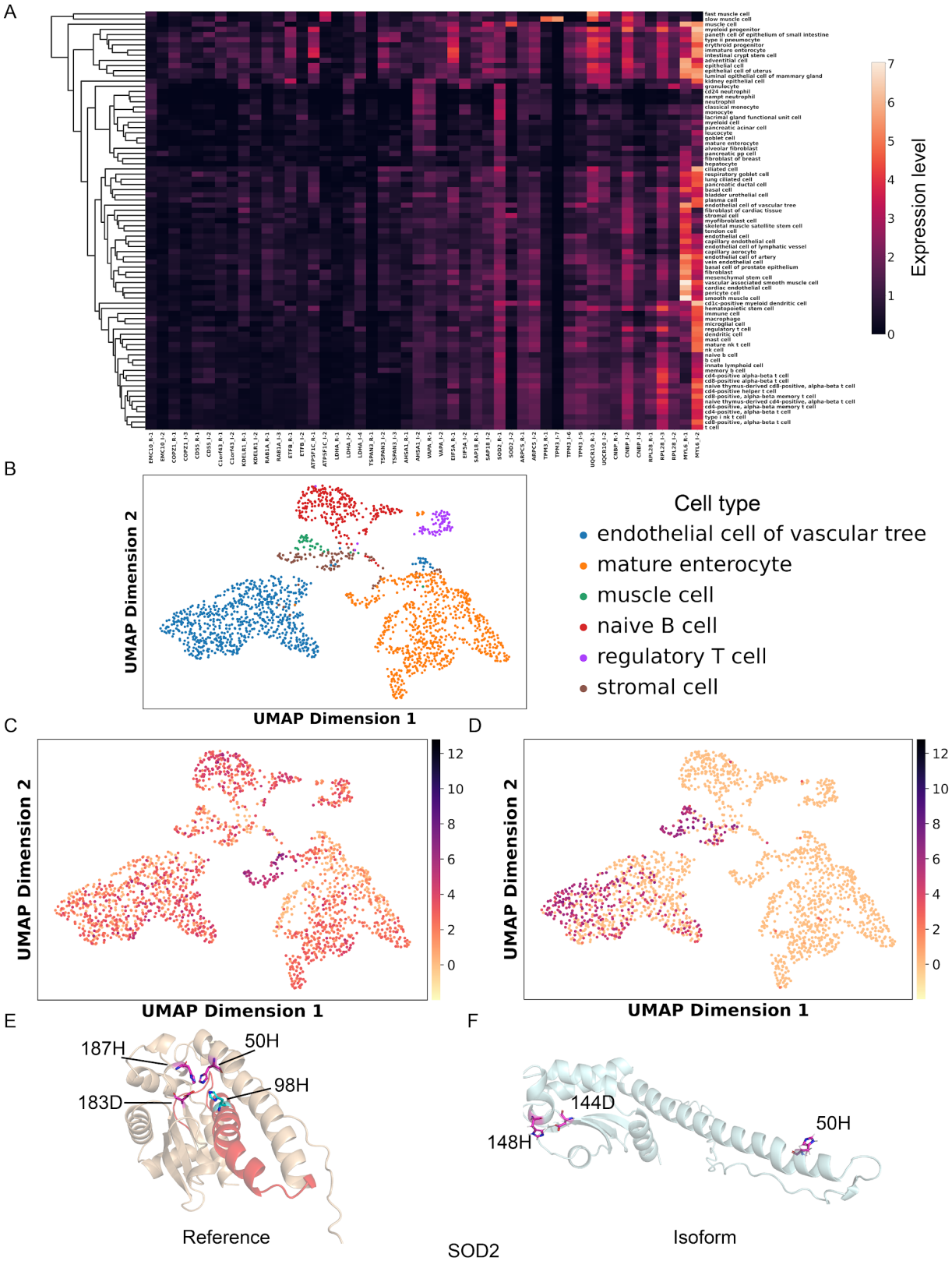

Figure 7: Cell-type-specific expression of alternate isoforms. (**A**) Expression profile for differential expressed reference/isoform across cell types. 'R' in the label stands for each reference and 'I' stands for the alternate isoform. (**B**) UMAP plot of the cell types, with differential expressed spliced isoforms of gene SOD2. (**C**) and (**D**) Expression of SOD2 reference and isoform across five cell types. (**E**) and (**F**) Structure of SOD2 reference and isoform; we label the Mn^{2+} binding sites in SOD2 structures, and the binding sites retained in isoform (500H, 183D and187H) are presented in pink sticks, and the binding site 98H lost in the SOD2 isoform is presented in cyan stick in reference.

2.6 Structure-based function prediction nominates functional changes among spliced isoforms

We used COFACTOR[52] to predict functions for our isoforms structures, allowing us to investigate the impact of alternative splicing on protein function. To do this, we quantified the number of gene ontology (GO) terms confidently predicted for each isoform, comparing these results with their respective reference isoforms. Function prediction shows that in general, reference and alternate isoforms have similar predicted function (**Figure S9A**). In all three subcategories of gene ontology (biological process, cellular component, and molecular function), the reference isoforms exhibit a higher count of confidently predicted GO terms compared to the alternate isoforms, consistent with an overall trend toward loss of function in alternate isoforms. In order to determine if distinct alternative splicing types result in varying functional alterations of spliced isoforms, we conducted a hypergeometric test to assess the gain or loss of GO terms across all alternative splicing types. More GO terms are lost (935) in alternate isoforms compared to the gained (226) GO terms, and this difference is statistically significant (p-value < 0.05, hypergeometric test). Some GO terms are more likely than others to be gained or lost due to alternative splicing. For example ADS is enriched for gaining the organelle membrane term (GO:0031090), while ALE is enriched for gaining the intracellular membrane-bound organelle term (GO:0043231), which both are membrane-related GO terms (**Figure 8A**). ES exhibits enrichment in gaining GO terms like lipid metabolic process (GO:0006629) and gaining of cytoskeleton-related terms including microtubule cytoskeleton (GO:0015630), microtubule organizing center (GO:0005815) and centrosome (GO:0005813). For loss of function, ADS is enriched in losing functions like response to chemical (GO:0042221), while AAS is likely to lose transcription regulator activity (GO:0140110) (**Figure 8B**). Notably, AFE and MXE-ALE significantly lose terms related to the regulation of different processes. Alternative splicing at the beginning and end of the protein seem to have greater effect on predicted protein function, possibly because splicing at these positions leads to greater structural changes. This trend is consistent with the negative PCC between TM-score and sequence identity for alternative splicing types including AFE, AFE, ALE, MXE-AFE and MXE-ALE in **Figure 3A**. Also, AFE, ALE, MXE-AFE and MXE-ALE have longer sequence changes compared to other alternative splicing types (**Table S7**).

We identified several interesting examples where isoforms gained or lost predicted functions compared to the reference isoform. TPTE2 isoform 4 (also called TPIP beta) has a lower prediction probability for terms related to lipid phosphatase activity, including phosphatidylinositol phosphate phosphatase activity (GO:0052866), glycerophospholipid

metabolic process (GO:0006650), phosphatidylinositol metabolic process (GO:0046488) and membrane related terms including membrane-bound organelle (GO:0043227) and intracellular membrane-bound organelle (GO:0043231) compared with the reference (**Figure 8C**). This is consistent with the fact that the fourth isoform of TPTE2 (also called TPIP β) was previously shown to lack phosphatase activity and act in the cytosol[53]. By comparison with the structure of *Ciona intestinalis* TPTE2 protein[54], the structure of human TPTE2 protein can be assigned into three main domains: the voltage sensor domain (VSD) formed by transmembrane helices, the phosphatase domain (PD) in the middle, and the C2 domain formed mainly by beta sheets (**Figure 8C**). The phospholipid phosphatase activity is triggered by voltage change between the plasma membrane[55], while TPTE2 isoform 4 lacks 130 amino acids of the VSD, suggesting that it cannot translocate to the plasma membrane and perform its normal function (**Figure 8C**). Similarly, another previous study found that the subcellular localization of another TPTE2 isoform, TPTE2 isoform 3 (also called TPIP2 α) did not locate to the plasma membrane[55], and TPTE2 isoform 3 only lacks 40 residues in VSD (whereas TPTE2 isoform 4 lacks 130 amino acids). In summary, our prediction that TPTE2 isoform 4 loses function is highly consistent with previous studies.

Interestingly, TPTE2 also has strong homology to the tumor suppressor PTEN[53], where the PTEN is nearly identical to the phosphatase domain and C2 domain. Overexpressed TPTE2 rescues $PTEN^{-/-}$ mutants through the interaction between C2 domain and the PTEN adhesion sites[56]. The TPTE2 isoform 4, marked by the absence of six out of nine beta sheets within its C2 domain, may potentially have forfeited its capability to bind to PTEN adhesion sites. This suggests that the loss of sequence in TPTE2 isoform 4 may have additional functional implications.

Another loss of function example is the TAFAZZIN isoform 7, which lacks exon 5 and exon 7 compared with the full-length reference, and is predicted to have lost transacylase activity[57]. The isoform is predicted to lose functions including 1-acylglycerol-3-phosphate O-acyltransferase activity (GO:0003841), acylglycerol O-acyltransferase activity (GO:0016411), lysophosphatidic acid acyltransferase activity (GO:0042171) and 1-acylglycerophosphocholine O-acyltransferase activity (GO:0047184) (**Figure 8D**). Since the isoform that lost exon 5 is also functional, the loss of function of isoform 7 is likely due to the loss of exon 7, which seems to be important to the substrate-binding[58].

We also identified examples where the isoforms are predicted to gain functions compared to their references. RGN is a gluconolactonase, and we predict that RGN isoform 2 has more GO terms related to ester hydrolase activity (**Figure S9B**). The structure of the RGN reference is composed of five repeat domains, where isoform 2 lacks one domain (**Figure S9B**), which may form a larger binding pocket and thus gain the ester hydrolase activity GO terms. Similarly, in another instance involving TGFB3, we predicted a gain in function, specifically in TGF II/III transforming growth factor beta receptor binding (GO:0007179 and GO:0034714), as a consequence of the loss of certain structural elements (**Figure S9C**). Notably, the region that is spliced out corresponds to the transforming growth factor beta-3 (TGF-beta-3) chains, while isoform 2 retains the structure of the latency-associated peptide and the signal peptide. This

phenomenon underscores the concept that alternative splicing can, in certain instances, produce effects akin to the cleavage of a proprotein. Moreover, we also observed that some isoforms are predicted to gain some functions by inserting some unstructured regions. In examples like TRAF2 isoform 2 (**Figure S9D**) and PCID4 isoform 4 (**Figure S9E**), the gained GO terms are related to cellular response and regulation. The inserted flexible loop may be a key factor of the gained GO terms, since some intrinsic disorder regions can mediate new protein-protein interactions (PPIs)[59].

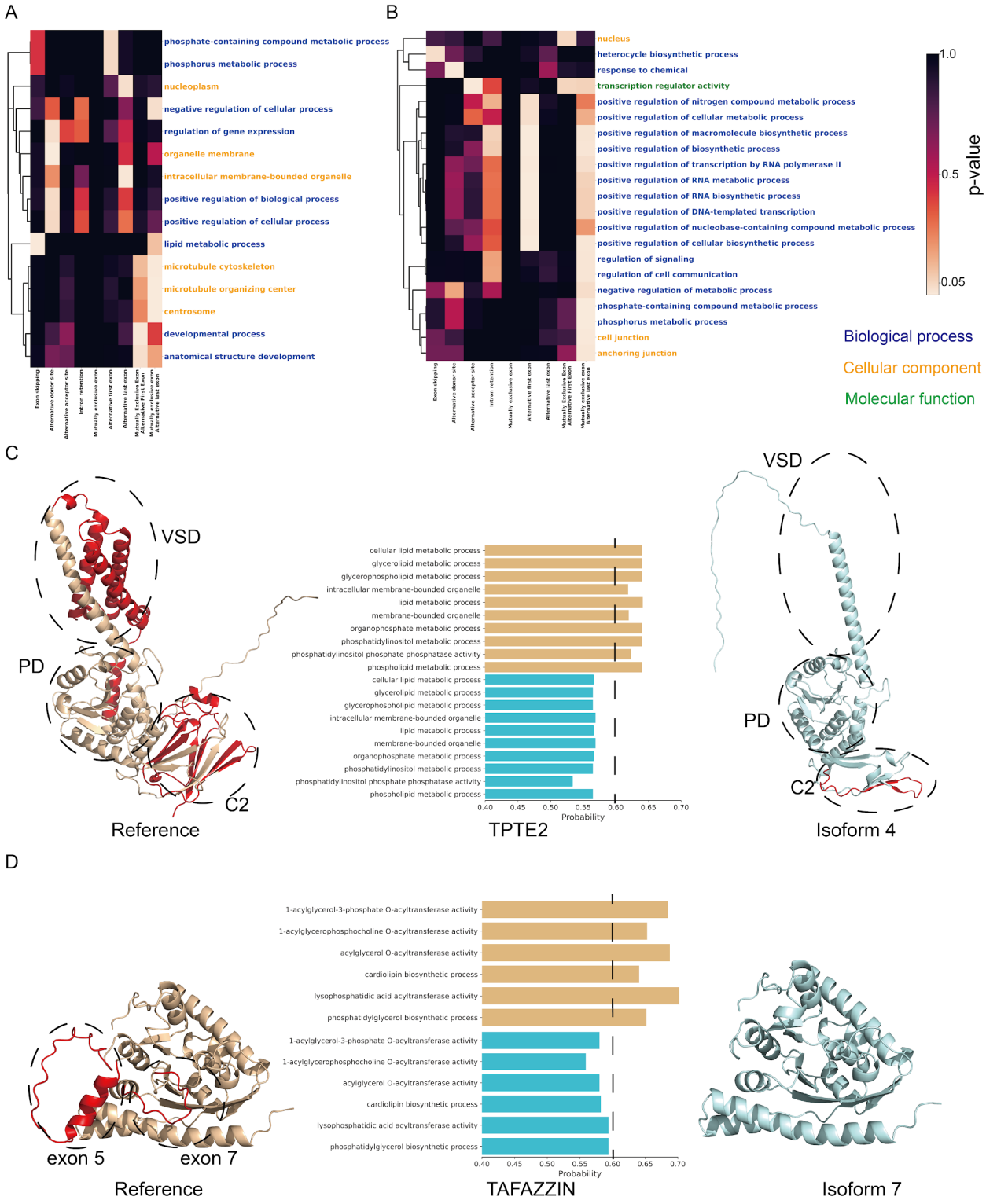

Figure 8: Structure-based prediction of functional changes induced by alternative splicing. Heatmap plot for GO terms frequently gained (A) or lost (B) across nine alternative splicing types, colored by p-value. The text colors for GO terms indicate which sub-ontology they come from. (C) Structures for TPTE2 reference and TPTE2 isoform 4 and predicted GO terms lost by TPTE2 isoform 4. The alternative splicing regions are colored in red. The three domains of TPTE2 are circled in the structure. (D)

Structures for TAFAZZIN reference and TAFAZZIN isoform 7 and predicted GO terms lost by TAFAZZIN isoform 7. The alternative splicing regions are colored in red. The structures encoded by exon 5 and 7 are circled.

3 Methods

3.1 Structure Prediction of human isoforms

We gather the alternative splicing sequences from the 2021_09 release of UniProt database[60], with a particular emphasis on SwissProt (**See Section 2.1**), where the data are manually curated and information for spliced isoforms is available. To enhance computational efficiency, mitigate memory constraints, and we restrict our selection to sequences containing fewer than 600 amino acids. For each gene, we selected the SwissProt canonical sequence as the reference sequence, considering any supplementary sequences listed in the "Sequence & Isoforms" section as isoform sequences. We ran AlphaFold2 locally (version: 2.2.0), utilizing default parameter settings, to generate predictions for spliced isoforms[16]. We obtain reference structures for each protein from the AlphaFold Protein Structure Database, 2022_01 release (https://alphafold.ebi.ac.uk/)[27] as provided by EBI. For both the reference and isoform structures, we computed multiple sequence alignment (MSA) locally, by extracting them from the feature pickle file before the prediction stage of AlphaFold2. The effective MSA depth is calculated for each isoform and reference sequence based on formula (1) [61]:

$$Neff = \sum_{i=1}^{N} \frac{1}{1 + \sum_{j=1, j\neq i}^{N} 1(I_{i,j} \geq 0.8)} \quad (1)$$

Where $Neff$ is the number of effective MSA for each query protein sequence, N is the total number of MSA for the query sequence, $1(I_{i,j} \geq 0.8)$ means the sequence identity will be calculated for each two sequences, and if they have sequence identity higher than 0.8, this value will be 1, otherwise will be 0.

We collected a total of 5,966 reference structures from EBI, and predicted 11,159 isoform structures. Structures are divided into four datasets based on their mean per-residue confidence score (pLDDT). Only structures with high (pLDDT > 90) and confident (70 < pLDDT < 90) quality are utilized for metrics analysis, which results in 4,450 reference structures and 7,631 isoform structures.

For comparison, we also obtained the CHESS (version: 1.1) human protein structure database from (https://www.isoform.io/)[19]. In this database, Sommer et al. employed AlphaFold2 to predict isoform structures at a transcript level, with a maximum sequence length of 1000 amino acids. To determine the reference sequences for the CHESS dataset, we use both the Matched Annotation from NCBI and EMBL-EBI (MANE) GRCh38 v0.95 file and the CHESS v2.2 gene annotation file[28,62]. In total, we gathered 15,727 reference structures and 97,824 isoform structures. Among these, 7,923 isoforms from our SwissProt dataset could be matched to the CHESS dataset, and we excluded 81,229 structures for which we could not decide whether they

were isoforms or references. Similarly, for metrics analysis, only 10,597 reference structures and 57,973 isoform structures with high or confident prediction quality were included.

To validate AlphaFold2's prediction ability on isoform sequences, we employed local Protein BLAST (version: 2.12.0+) to search against all Protein Data Bank (PDB) database sequences. We selected experimental isoform structures with an E-value threshold of < 10e-5. We manually excluded the structures which: (1) included unmodeled splicing regions; (2) were from non-human species; and (3) were low resolution structures from NMR and CryoEM. To assess structural similarity, we used TM-score and RMSD based on US-align (version: 20230609), which included the TM-align to compare AlphaFold-predicted structures with experimental structures[63], and we used the '-seq' option to ensure the sequence aligned correctly before the structure alignment. Furthermore, we conducted a paired t-test to compare reference and isoform structures in terms of TM-score and RMSD. Notably, the AlphaFold structures used for comparison were predicted in a template-free setting of AlphaFold.

3.2 Algorithm for identifying alternative splicing types

To map the exons between each isoform and its respective reference sequence, we use the human genome release 40 (GRCh38.p13)(Ensembl 106) General Feature Format (.gtf) file provided by Gencode. We match the SwissProt accession numbers with their corresponding Ensembl transcript numbers using the SwissProt metadata file from Gencode. It should be noted that 3,252 SwissProt entries (reference or spliced isoform) lack Ensembl accession numbers, and also 484 entries are labeled as nonsense-mediated decay. Consequently, we exclude these isoforms from our analysis of alternative splicing types. We successfully determine the alternative splicing types for 7,685 isoforms from our SwissProt isoform structure dataset and 97,824 isoforms from the CHESS dataset. The exon position information is matched and compared pairwise between each isoform transcript and corresponding reference transcript. Different from the seven alternative splicing events defined in previous study[8], we designed our methods to detect nine alternative splicing events: exon skipping (ES), alternative donor site (ADS) or alternative 5′ splice site (A5S), alternative acceptor site (AAS) or alternative 3' splice site (A3S), mutually exclusive exons (MXE), intron retention (IR), alternative first exon (AFE), alternative last exon (ALE), mutually exclusive exon-alternative first exon (MXE-AFE) and mutually exclusive exon-alternative last exon (MXE-ALE)[64,65]. Given that a prior study observed that alternative first exons can display mutually exclusive behavior[66], we introduced the definitions of mutually exclusive exon-alternative first exon (MXE-AFE) and mutually exclusive exon-alternative last exon (MXE-ALE) to differentiate them from alternative first exon (AFE) and alternative last exon (ALE). The graphical representations of each alternative splicing type, along with illustrative example structures, are depicted in **Figure S1**.

3.3 Metrics for structural analysis

We compute the TM-score using the tmtools Python package (version: 0.0.2) based on TM-align[67]. In each iteration, we perform two alignments of isoform-reference pairs, utilizing either the reference or isoform as the template. The average of the two TM-scores is considered as the final TM-score, accounting for potential disparities between SwissProt selected reference sequence and the predominant functional isoform within the human body.

We assess sequence identity between each isoform sequence and its corresponding reference sequence using pairwise global sequence alignment from the Biopython (version: 1.79) package[68]. Initial global alignments employ the BLOSUM62 substitution matrix and the Needleman-Wunsch algorithm. Subsequently, sequence identity is computed based on the obtained alignments, see formula (2).

$$I_s = N_i / L_a \quad (2)$$

Where I_s is the sequence identity between reference and isoform sequences, N_i is the number of identical residues, L_a is the alignment length.

We measure the Pearson correlation coefficients (PCC) between the TM-score and sequence identity for each isoform-reference pair.

We use Pymol (https://pymol.org/2/) (version: 2.4.1) to calculate DSSP secondary structure percentage using the dss command[69]. In our analysis, we only consider the three primary secondary structure types: helix, sheet, and loop.

For surface charge comparison, we first apply the PDB2PQR30 program (version: 3.5.2) with the PARSE force field to assign charges to each residue[70]. Subsequently, to get the surface residues, we calculate the solvent accessible surface area (SASA) using the "measure sasa" command within Visual Molecular Design (VMD) (https://www.ks.uiuc.edu/Research/vmd/) (version: 1.9.2), with the restrict distance set as 1.4 Å. We then use the relative solvent accessibility (RSA) to identify surface residues, which is usually used to decide the exposure content of a residue[71]. RSA is SASA after normalized by the maximum allowed SASA value, and we apply a threshold RSA as defined by Tien et al.[71], where residues with an RSA higher than the threshold are considered surface residues. The surface charge for each protein is determined by summing the charges of all surface residues.

We calculate the radius of gyration for each isoform and reference structure using the VMD according to the formula (3):

$$R(g) = \sum_i m_i \cdot (r_i - r_c)^2 / \sum_i m_i \quad (3)$$

where m_i represents the mass for atom i, r_i represents the radius for atom i, and r_c represents the radius of the center of mass.

We use the lm function in R (version: 4.1.1) to construct a linear regression model to study the effect of specific alternative splicing types. Our features are derived from the sequence length results for each alternative splicing type within each isoform. For alternative splicing types like ES and IR, which could happen in both isoform and reference, we separate them into ES in reference, ES in isoform, IR in reference and IR isoform to distinguish the alternative splicing in reference and isoform. For secondary structure percentage, surface charge, and radius of gyration, we use the difference between the isoform and reference as the outcome variable to

build the linear regression model (formula (5)). In the case of TM-score, which inherently involves a comparison between two structures, we employ the absolute length difference caused by each alternative splicing event as the feature, and since its non-directional, we do not specify the situation where ES and IR occur in reference and isoform (formula (7)). All the predicted metrics are standardized to mean 0 and standard deviation of 1. The coefficient associated with each alternative splicing type was interpreted as the per-residue effect, and the p-value for each coefficient is obtained from the "lm" function.

$$\Delta L_{AS} = \Delta L_{ADS} + \Delta L_{AAS} + \Delta L_{MXE} + \Delta L_{AFE} + \Delta L_{ALE} + \Delta L_{MXE-AFE} + \Delta L_{MXE-ALE} \quad (4)$$

ΔL_{AS} represents the sequence length change (Isoform-reference) from the following alternative splicing types: ADS, AAS, MXE, AFE, ALE, MXE-AFE, MXE-ALE.

$$\Delta_{M} = \Delta_{pLDDT} + \Delta L_{ES(R)} + \Delta L_{ES(I)} + \Delta L_{IR(R)} + \Delta L_{IR(L)} + \Delta L_{AS} \quad (5)$$

Δ_{pLDDT} is the difference of pLDDT. Δ_{M} represents the difference of metrics including secondary structure percentage, surface charge and radius of gyration. $\Delta L_{ES(R)}$, $\Delta L_{ES(I)}$, $\Delta L_{IR(R)}$, $\Delta L_{IR(I)}$ are the sequence length change caused by ES in reference, ES in isoform, IR in reference and IR in isoform.

$$|\Delta L_{AS*}| = |\Delta L_{ADS}| + |\Delta L_{AAS}| + |\Delta L_{MXE}| + |\Delta L_{AFE}| + |\Delta L_{ALE}| + |\Delta L_{MXE-AFE}| + |\Delta L_{MXE-ALE}| \quad (6)$$

$|\Delta L_{AS*}|$ represents the absolute value of the sequence length change from the following alternative splicing types: ADS, AAS, MXE, AFE, ALE, MXE-AFE, MXE-ALE.

$$TM - score = |\Delta_{pLDDT}| + |\Delta L_{ES}| + |\Delta L_{IR}| + |\Delta L_{AS*}| \quad (7)$$

$|\Delta L_{ES}|$, $|\Delta L_{IR}|$ are the absolute value of sequence length caused by ES and IR, respectively.

3.4 Analysis of PTM sites

We collect post translational modification (PTM) site information from PhosphoSitePlus (release 2022.07)[72], which includes seven main PTM types: phosphorylation, ubiquitination, acetylation, methylation, O-GalNAc glycosylation, O-GlcNAcylation and sumoylation. To accurately classify PTM sites, we first map the alternative splicing regions for each isoform based on the information from SwissProt; PTM sites located outside of these regions are classified as 'retained'. Similarly, PTM sites in the regions labeled as 'Missing' are categorized as 'spliced out' PTM sites. And also, there are PTM sites from the PhosphoSitePlus that have records only in isoforms. We reverse map these sites back to their reference counterparts, classifying sites that couldn't be matched to the reference as 'spliced in' PTM sites. For the ‘retained’ PTM sites, we use the RSA to determine the exposure extent for every PTM site, classifying them as either 'buried' or 'exposed'. If the PTM site is consistently buried or exposed between reference and

isoform, it will be labeled as 'unchanged'. In total, we classify the PTM sites into five categories: unchanged, spliced out, spliced in, buried to exposed and exposed to buried. It's worth noting that due to the absence of protein functional annotation for the CHESS isoform structure dataset, PTM site analysis is exclusively applied to our SwissProt isoform structure dataset.

We obtain the RNA-Seq data for BAXα and BAXδ from the TCGA Splicing Variants DB, which measures expression across 33 cancer types[45]. For both BAXα and BAXδ, we split them into high and low expression groups based on their expression in each cancer type by a ratio of 0.25:0.75 (high:low). We use a log-rank test with a threshold p-value of 0.05 between high and low expression groups to determine whether expression of specific BAX isoforms will affect survival or not.

Visualization of the phosphorylation form of PTM sites is simulated by the SwissSidechain (version: 2) [73].

3.5 Preprocessing single-cell data

To reliably identify alternative splicing from single-cell RNA sequencing (scRNA-seq) data, we gather the fastq files for the Smart-seq2 dataset from Tabula Sapiens Consortium (https://tabula-sapiens-portal.ds.czbiohub.org/)[23]. Then, we use the Kb-python (version: 0.27.3) to generate the transcript count matrix from fastq files[74]. We build the index using the "kb -ref" command from kb-python, using the Human genome release 39 (GRCh38.p13) gtf file along with the CHESS2.2 gtf file as the input annotation files. We use the "kb -count" command from kb-python to generate the count matrix. For cell type annotations, we obtain the information for each cell from https://figshare.com/projects/Tabula_Sapiens/100973. Cells without cell type annotations are excluded, resulting in a dataset comprising a total of 133 cell types. We follow the normalization procedure outlined by A. Sina et al.[21], which involves normalizing the expression of each transcript by its length and the library size, calculating the Transcripts Per Million (TPM) for each transcript, and applying a log transformation to the expression data. After preprocessing, we generate two cell-by-transcript matrices for both the Gencode and CHESS transcripts, consisting of 26,748 cells and 164,607 transcripts for Gencode and 26,748 cells and 113,551 transcripts for CHESS, respectively. The cell-by-gene matrix is constructed by aggregating the expression of all transcripts within each gene, resulting in 26,748 cells and 19,980 genes for the Gencode dataset and 26,748 cells and 18,659 genes for the CHESS dataset.

3.6 Isoform usage shift analysis of single-cell data

We specifically subset a SwissProt count matrix from the Gencode dataset (after normalization), including only the isoforms for which we have modeled the protein structures. This subset consists of 26,748 cells and 13,524 isoforms and is used for the analysis of differentially expressed isoforms. For all isoforms within each gene, we conduct either a t-test (for two isoforms) or a one-way analysis of variance (ANOVA) test (for more than two isoforms). The purpose is to assess whether there are significant differences in the expression levels for each isoform within the same gene. We then identify transcripts that exhibit expression specific to particular cell types below the critical p-value threshold of 0.05, after applying the Bonferroni

correction. Different isoforms for the same gene which are up-regulated for the same cell type are excluded, and we also exclude the cell types with fewer than 30 cells; the results are sorted based on fold change. Additionally, we perform the same differential expression analysis on 45 heterogeneous cell types originating from various tissues. This analysis is aiming to identify isoforms that could distinguish expression patterns within the same cell type across different tissues.

3.7 Structure-based function prediction with COFACTOR

The GO terms for the human proteins are predicted by a modified version of COFACTOR[75] optimized for large-scale structure-based function annotation. The pipeline consists of four complementary pipelines based on sequence, structure, protein-protein interactions (PPI), and Pfam domain family.

In the sequence-based pipeline, the query sequence is searched by BLASTp through the UniProt Gene Ontology Annotation (UniProt-GOA) database with default parameters to identify templates with GO annotations. Only annotations validated by experimental or high-throughput evidence, traceable author statement (evidence code TAS), or inferred by curator (IC) are considered. The prediction score of GO term *q* is defined as:

$$Cscore_{sequence}(q) = \frac{\sum_{k-1}^{K(q)} ID_k(q) \cdot bitscore_k(q)}{\sum_{k-1}^{K} ID_k \cdot bitscore_k} \quad (8)$$

Here, *K* is the total number of BLASTp hits; $bitscore_k$ is the bit-score of the *k*-th hit; ID_k is the global sequence identity of the *k*-th hit; $K(q)$ and $bitscore_k(q)$ are the corresponding values for the subset of BLASTp hits with GO term *q*.

In the structure-based pipeline, the query sequence is searched by Foldseek[76] (Accessed 6/13/2023) using parameters "--tmalign-fast 1 -e 10" through the subset of AlphaFold database structures with GO annotations in UniProt-GOA. The identified hits are re-aligned by TM-align[67] to get the TM-score. The prediction score is as:

$$Cscore_{structure}(q) = \frac{\sum_{k-1}^{K(q)} ID_k(q) \cdot TM_k(q) \cdot bitscore_k(q)}{\sum_{k-1}^{K} ID_k \cdot TM_k \cdot bitscore_k} \quad (9)$$

Here, TM_k is the TM-score obtained from TM-align realignment of the *k*-th Foldseek hit; $TM_k(q)$ is the TM-score for Foldseek hit with term *q*.

In the PPI-based pipeline, the query protein is mapped to the STRING [77] PPI database to identify PPI partners with GO annotation. The prediction score can be calculated as:

$$Cscore_{ppi}(q) = \frac{\sum_{k-1}^{K(q)} string_k(q)}{\sum_{k-1}^{K} string_k} \quad (10)$$

Here, $string_k$ is the STRING score of the *n*-th PPI partner, and $string_k(q)$ is the STRING score for the partner with term *q*.

In the Pfam-based pipeline, the query protein is searched by hmmsearch through the Pfam database to get the list of Pfam hits. Only the 7000 most common Pfam families in the UniProt database are considered. The prediction score for term *q* is derived by logistic regression.

$$Cscore_{pfam}(q) = \frac{1}{1+exp(-w_0 - \sum_{m=1}^{7000} w_m \cdot x_m)} \quad (11)$$

Here, x_m indicates the match of the m-th Pfam family; w_m is the weight optimized on a separate training set.

The prediction scores of the sequence-, structure-, PPI- and Pfam-based pipelines are combined with the prediction score from a recent deep learning-based GO predictor [78] and the background frequency of the GO term in the UniProt-GOA as six different input features. These features are fed into a gradient boosted tree model trained by LightGBM [79] to obtain the final consensus GO prediction.

We used as probability of 0.6 as the threshold; only GO terms with above 0.6 predicted probability were considered as confident GO terms. The alternative splicing type-wise GO enrichment hypergeometric test was calculated by the following formula using the scipy python package (version: 1.7.1) on our high and confident dataset:

$$p(k, M, n, N) = \left(\frac{\frac{n!}{k! \cdot (n-k)!} \cdot \frac{(M-n)!}{(N-k)! \cdot ((M-n)-(N-k))!}}{\frac{M!}{N! \cdot (M-N)!}}\right) \quad (12)$$

Where k represents the number of isoforms with gain/lose this GO term and with this AS type, M represents the total number of isoforms, n represents the number of isoforms which gain/lose this GO term, and N represents the number of isoforms with this AS type.

For heatmaps shown in the figures, we removed the go terms with too few gained or lost terms, (minimum of 5 gained GO terms and minimum of 15 lost GO terms), and we also removed general GO terms which have over 5,000 predictions among isoforms and references with high and confident prediction quality.

The UMAP embedding is calculated using the umap-learn (version: 0.5.3) scipy package based on a pre-computed jaccard distance matrix for the confident GO terms each reference/isoform contained.

4 Discussion

Alternative splicing is recognized as a key driver of protein diversity[1,80], but there is an ongoing debate concerning the functional significance of these spliced isoforms[81]. Guided by the "sequence-structure-function" paradigm, we folded human spliced isoforms and applied various structural and functional metrics to reveal the functional implications of alternative splicing from a structural view. Our discoveries contribute to an enhanced comprehension of the sequence determining protein structures. However, it is noteworthy that we observed a generally lower evolutionary conservation suggested by MSAs in splicing regions, which may have a potential impact on structure prediction. This concern is not unique to alternative splicing regions; it pertains to all protein regions with insufficient homologous sequences. It is valuable to explore MSA-enhanced methods and leverage large protein language models to enhance the prediction quality for sequences with limited MSAs, including some alternative splicing regions[82].

For large-scale data analysis, the identification of outlier instances proves to be a more valuable approach than the generation of broad, statistically significant conclusions. This preference arises due to the susceptibility of the latter to bias resulting from the abundance of data points exhibiting subtle distinctions between the two categories. Notably, we have successfully pinpointed specific isoforms within our dataset whose structural features underwent substantial alterations due to alternative splicing, as evidenced in **Table S2**. It is worth highlighting that the functional characteristics of the majority of these outlier isoforms have yet to be characterized, thereby implying that our methods may have preemptively selected certain spliced isoforms with the potential for functional divergence from the reference form.

Combined with the isoform-level scRNA-seq expression data, our findings could help to enhance the annotation of the selected canonical/reference sequences from UniProt/SwissProt database. Additionally, it allows for a finer-grained elucidation of expression data at both cell-type and isoform levels, despite the disparities between mRNA and protein expression levels[83]. Furthermore, we have established connections between the predicted structural features and their expression profiles within various cell types and tissues, with the overarching objective of shedding light on the distinctive cellular perspectives offered by protein structural analysis.

In light of recent advancements in computational tools, exemplified by AlphaFold2, this research endeavor stands as a pioneering model for harnessing AlphaFold2's capabilities in exploring pertinent structural themes. To enrich our comprehension of alternative splicing, future investigations may pivot toward the exploration of protein-protein interactions (PPI)[84,85], domain analysis and isoform function prediction[86–88], aimed at facilitating the annotation and elucidation of functional spliced isoforms.

Acknowledgements

We thank members of the Freddolino lab and Matthew Karikomi for helpful discussions and feedback. This work was supported by Chan-Zuckerberg Initiative grant PN-0000000075 to J.D.W. G.S.O. acknowledges support from National Institutes of Health Grants P30 ES017885-11-S1 and U24 CA271037, and M.J.O acknowledges support from National Institutes of Health grant R35GM151129.

Author contributions

YS: Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. CZ: Methodology, Investigation, Formal analysis, Writing - Review & Editing. GSO: Discussion, Review & Editing. MJO: Conceptualization, Methodology, Formal analysis, Writing - Review & Editing, Supervision, Project administration. JW: Conceptualization, Methodology, Formal analysis, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Declaration of interests

The authors have no competing interests to declare.

Data and code availability

All scripts and analysis code are available on GitHub: https://github.com/welch-lab/AF2_scRNA. All predicted structures are available via Figshare: 10.6084/m9.figshare.24891870, processed human scRNA-seq data is available via Figshare: 10.6084/m9.figshare.24843948, function prediction results are available via Figshare: 10.6084/m9.figshare.24891897.

References

1. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
2. Ezkurdia, I. *et al.* Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878 (2014).
3. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
4. Stamm, S. *et al.* Function of alternative splicing. *Gene* **344**, 1–20 (2005).

5. Aartsma-Rus, A. & van Ommen, G.-J. B. Less is more: therapeutic exon skipping for Duchenne muscular dystrophy. *Lancet neurology* vol. 8 873–875 (2009).
6. Climente-González, H., Porta-Pardo, E., Godzik, A. & Eyras, E. The Functional Impact of Alternative Splicing in Cancer. *Cell Rep.* **20**, 2215–2226 (2017).
7. Shenasa, H. & Hertel, K. J. Combinatorial regulation of alternative splicing. *Biochim. Biophys. Acta Gene Regul. Mech.* **1862**, 194392 (2019).
8. Wright, C. J., Smith, C. W. J. & Jiggins, C. D. Alternative splicing as a source of phenotypic diversity. *Nat. Rev. Genet.* **23**, 697–710 (2022).
9. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451 (2017).
10. Melamud, E. & Moult, J. Structural implication of splicing stochastics. *Nucleic Acids Res.* **37**, 4862–4872 (2009).
11. Modrek, B., Resch, A., Grasso, C. & Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **29**, 2850–2859 (2001).
12. Osmanli, Z. *et al.* The Difference in Structural States between Canonical Proteins and Their Isoforms Established by Proteome-Wide Bioinformatics Analysis. *Biomolecules* **12**, (2022).
13. Birzele, F., Csaba, G. & Zimmer, R. Alternative splicing and protein structure evolution. *Nucleic Acids Res.* **36**, 550–558 (2008).
14. Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77**, 363–382 (2008).
15. Yang, J. & Zhang, Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–81 (2015).
16. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
17. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

18. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
19. Sommer, M. J. *et al.* Structure-guided isoform identification for the human transcriptome. *Elife* **11**, (2022).
20. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
21. Booeshaghi, A. S. *et al.* Isoform cell-type specificity in the mouse primary motor cortex. *Nature* **598**, 195–199 (2021).
22. Wen, W. X., Mead, A. J. & Thongjuea, S. MARVEL: an integrated alternative splicing analysis platform for single-cell RNA sequencing data. *Nucleic Acids Res.* **51**, e29 (2023).
23. Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
24. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
25. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
26. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
27. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
28. Pertea, M. *et al.* CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018).
29. Akdel, M. *et al.* A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).
30. Pak, M. A. *et al.* Using AlphaFold to predict the impact of single mutations on protein

stability and function. *PLoS One* **18**, e0282689 (2023).

31. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2022) doi:10.1101/2021.10.04.463034.
32. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
33. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
34. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* **11**, 345–355 (2010).
35. Wei Xq *et al.* The Sushi domain of soluble IL-15 receptor alpha is essential for binding IL-15 and inhibiting inflammatory and allogenic responses in vitro and in vivo. *J. Immunol.* **167**, 277–282 (2001).
36. Linse, S. *et al.* The role of protein surface charges in ion binding. *Nature* **335**, 651–652 (1988).
37. Goldenberg, N. M. & Steinberg, B. E. Surface charge: a key determinant of protein localization and function. *Cancer Res.* **70**, 1277–1280 (2010).
38. Lobanov, M. Y., Bogatyreva, N. S. & Galzitskaya, O. V. Radius of gyration as an indicator of protein structure compactness. *Mol. Biol.* **42**, 623–628 (2008).
39. Ahmed, M. C., Crehuet, R. & Lindorff-Larsen, K. Computing, Analyzing, and Comparing the Radius of Gyration and Hydrodynamic Radius in Conformational Ensembles of Intrinsically Disordered Proteins. *Methods Mol. Biol.* **2141**, 429–445 (2020).
40. Lasagni, L. *et al.* An alternatively spliced variant of CXCR3 mediates the inhibition of endothelial cell growth induced by IP-10, Mig, and I-TAC, and acts as functional receptor for platelet factor 4. *J. Exp. Med.* **197**, 1537–1549 (2003).
41. Roach, P. L. Radicals from S-adenosylmethionine and their application to biosynthesis. *Curr. Opin. Chem. Biol.* **15**, 267–275 (2011).

42. Krishnamoorthy, E. *et al.* Homology modeling of Homo sapiens lipoic acid synthase: Substrate docking and insights on its binding mode. *J. Theor. Biol.* **420**, 259–266 (2017).
43. Beltrao, P., Bork, P., Krogan, N. J. & van Noort, V. Evolution and functional cross-talk of protein post-translational modifications. *Mol. Syst. Biol.* **9**, 714 (2013).
44. Tu, Y.-C., Yeh, W.-C., Yu, H.-H., Lee, Y.-C. & Su, B.-C. Hedgehog Suppresses Paclitaxel Sensitivity by Regulating Akt-Mediated Phosphorylation of Bax in EGFR Wild-Type Non-Small Cell Lung Cancer Cells. *Front. Pharmacol.* **13**, 815308 (2022).
45. Sun, W. *et al.* TSVdb: a web-tool for TCGA splicing variants analysis. *BMC Genomics* **19**, 405 (2018).
46. Olivieri, J. E. *et al.* RNA splicing programs define tissue compartments and cell types at single-cell resolution. *Elife* **10**, (2021).
47. Lenz, S., Lohse, P., Seidel, U. & Arnold, H. H. The alkali light chains of human smooth and nonmuscle myosins are encoded by a single gene. Tissue-specific expression by alternative splicing pathways. *J. Biol. Chem.* **264**, 9009–9015 (1989).
48. Lebrigand, K. *et al.* The spatial landscape of gene expression isoforms in tissue sections. *Nucleic Acids Res.* **51**, e47 (2023).
49. Fu, Y. *et al.* Single cell and spatial alternative splicing analysis with long read sequencing. *Res Sq* (2023) doi:10.21203/rs.3.rs-2674892/v1.
50. Gateva, G. *et al.* Tropomyosin Isoforms Specify Functionally Distinct Actin Filament Populations In Vitro. *Curr. Biol.* **27**, 705–713 (2017).
51. Li, S. *et al.* The novel truncated isoform of human manganese superoxide dismutase has a differential role in promoting metastasis of lung cancer cells. *Cell Biol. Int.* **42**, 1030–1040 (2018).
52. Zhang, C., Lane, L., Omenn, G. S. & Zhang, Y. Blinded Testing of Function Annotation for uPE1 Proteins by I-TASSER/COFACTOR Pipeline Using the 2018–2019 Additions to neXtProt and the CAFA3 Challenge. *J. Proteome Res.* **18**, 4154–4166 (2019).

53. Walker, S. M., Downes, C. P. & Leslie, N. R. TPIP: a novel phosphoinositide 3-phosphatase. *Biochem. J* **360**, 277–283 (2001).

54. Matsuda, M. *et al.* Crystal structure of the cytoplasmic phosphatase and tensin homolog (PTEN)-like region of Ciona intestinalis voltage-sensing phosphatase provides insight into substrate specificity and redox regulation of the phosphoinositide phosphatase activity. *J. Biol. Chem.* **286**, 23368–23377 (2011).

55. Halaszovich, C. R. *et al.* A human phospholipid phosphatase activated by a transmembrane control module. *J. Lipid Res.* **53**, 2266–2274 (2012).

56. Lusche, D. F. *et al.* Overexpressing TPTE2 (TPIP), a homolog of the human tumor suppressor gene PTEN, rescues the abnormal phenotype of the PTEN-/- mutant. *Oncotarget* **9**, 21100–21121 (2018).

57. Xu, Y. *et al.* Characterization of tafazzin splice variants from humans and fruit flies. *J. Biol. Chem.* **284**, 29230–29239 (2009).

58. Hijikata, A., Yura, K., Ohara, O. & Go, M. Structural and functional analyses of Barth syndrome-causing mutations and alternative splicing in the tafazzin acyltransferase domain. *Meta Gene* **4**, 92–106 (2015).

59. Buljan, M. *et al.* Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. *Curr. Opin. Struct. Biol.* **23**, 443–450 (2013).

60. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

61. Wu, T., Hou, J., Adhikari, B. & Cheng, J. Analysis of several key factors influencing deep learning-based inter-residue contact prediction. *Bioinformatics* **36**, 1091–1098 (2020).

62. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).

63. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **19**, 1109–1115

(2022).

64. Sieber, P. *et al.* Comparative Study on Alternative Splicing in Human Fungal Pathogens Suggests Its Involvement During Host Invasion. *Front. Microbiol.* **9**, 2313 (2018).
65. Le, K.-Q., Prabhakar, B. S., Hong, W.-J. & Li, L.-C. Alternative splicing as a biomarker and potential target for drug discovery. *Acta Pharmacol. Sin.* **36**, 1212–1218 (2015).
66. Chen, W.-H., Lv, G., Lv, C., Zeng, C. & Hu, S. Systematic analysis of alternative first exons in plant genomes. *BMC Plant Biol.* **7**, 55 (2007).
67. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
68. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
69. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* **22**, 2577–2637 (1983).
70. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32**, W665–W667 (2004).
71. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilites of residues in proteins. *PLoS One* **8**, e80635 (2013).
72. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–20 (2015).
73. Gfeller, D., Michielin, O. & Zoete, V. SwissSidechain: a molecular and structural database of non-natural sidechains. *Nucleic Acids Res.* **41**, D327–32 (2013).
74. Melsted, P. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.* **39**, 813–818 (2021).
75. Zhang, C., Freddolino, P. L. & Zhang, Y. COFACTOR: improved protein function prediction

by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.* **45**, W291–W299 (2017).

76. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01773-0.
77. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2022).
78. Yuan, Q., Xie, J., Xie, J., Zhao, H. & Yang, Y. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Brief. Bioinform.* **24**, (2023).
79. Lightgbm: A highly efficient gradient boosting decision tree. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.
80. Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103**, 367–370 (2000).
81. Bhuiyan, S. A. *et al.* Systematic evaluation of isoform function in literature reports of alternative splicing. *BMC Genomics* **19**, 637 (2018).
82. Meng, Q., Guo, F. & Tang, J. Improved structure-related prediction for insufficient homologous proteins using MSA enhancement and pre-trained language model. *Brief. Bioinform.* **24**, (2023).
83. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature* **547**, E19–E20 (2017).
84. Yang, X. *et al.* Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164**, 805–817 (2016).
85. Louadi, Z. *et al.* DIGGER: exploring the functional role of alternative splicing in protein interactions. *Nucleic Acids Res.* **49**, D309–D318 (2021).

86. Light, S. & Elofsson, A. The impact of splicing on protein domain architecture. *Curr. Opin. Struct. Biol.* **23**, 451–458 (2013).

87. Hao, Y. *et al.* Semi-supervised Learning Predicts Approximately One Third of the Alternative Splicing Isoforms as Functional Proteins. *Cell Rep.* **12**, 183–189 (2015).

88. Ferrer-Bonsoms, J. A. *et al.* ISOGO: Functional annotation of protein-coding splice variants. *Sci. Rep.* **10**, 1069 (2020).