Toward a comprehensive profiling of alternative splicing proteoform structures, interactions and functions

Elodie Laine, Maria Inés Freiberger

Toward a comprehensive profiling of alternative splicing proteoform structures, interactions and functions

Elodie Laine[1,2,*] and Maria Inés Freiberger[1]

1 Sorbonne Université, CNRS, IBPS, Laboratory of Computational and Quantitative Biology (LCQB), UMR 7238, 75005 Paris, France

2 Institut universitaire de France (IUF)

* Corresponding author: elodie.laine@sorbonne-universite.fr

Abstract

The mRNA splicing machinery has been estimated to generate 100,000 known protein-coding transcripts for 20,000 human genes (Ensembl, Sept. 2024). However, this set is expanding with the massive and rapidly growing data coming from high-throughput technologies, particularly single-cell and long-read sequencing. Yet, the implications of splicing complexity at the protein level remain largely uncharted. In this review, we describe the current advances toward systematically assessing the contribution of alternative splicing to proteome function diversification. We discuss the potential and challenges of using artificial intelligence-based techniques in identifying alternative splicing proteoforms and characterising their structures, interactions, and functions.

Alternative splicing and proteome diversity

The recent advances in high-throughput sequencing, imaging, and proteomics have revealed an incredible complexity behind the classical protein sequence-structure-function paradigm [1]. In particular, in multicellular organisms, alternative splicing (AS), together with alternative promoter usage and alternative polyadenylation, can produce multiple mature messenger RNAs, or *transcripts*, from a single gene [2] (**Figure 1A**). Some of these transcripts will lead to different protein isoforms, or *proteoforms* [3], that may adopt 3D structures with different shapes [4], interact with distinct cellular partners [5], and perform divergent or specialised functions [6,7]. For example, a 10-amino acid (aa) substitution between two proteoforms of the protein kinase JNK1 changes its binding partner preferences, thus triggering different stress responses [8]. Similarly, while a shorter clathrin proteoform self-assembles into spherical coats in neurons, a 7-aa longer one forms flat plaques in muscle cells [9] (**Figure 1B**). The plethora of scenarios in which AS modulates protein functions and interactions [10] play essential roles in muscle fibre diversification [11], nervous system development [12], and innate immunity [13]. Moreover, the combinatorial expression of various proteoforms can influence disease susceptibility [14] and signalling outcomes in response to drugs [15], and AS misregulation is often linked to various diseases, including cancer [16,17].

Experimentally determining how much of the splicing complexity uncovered by RNA-seq [18] contributes to protein diversity remains a long-standing challenge [19]. Higher AS rates are typically observed for species with lower effective population sizes, suggesting that they result from genetic drift of the splicing machinery [20]. Along this line, analysing mass spectrometry data on a large scale initially suggested that most highly expressed human genes have only one dominant proteoform [21]. However, improved analysis protocols using custom peptide databases or integrating long-read transcriptomics reported many more

proteoforms [22–23]. Furthermore, a major fraction of alternative transcripts are engaged by ribosomes [24] and a recent deep-coverage mass spectrometry study revealed evidence that most frame-preserving alternative transcripts are translated [25].

Emerging high-throughput computational methods efficiently leveraging large amounts of protein-related data represent an opportunity for complementing experimental evidence, toward refining the definition of gene structures, quantifying the alternative usage of exons, and improving our understanding of AS impact on protein 3D structures, interactions, and functions (**Figure 2**).

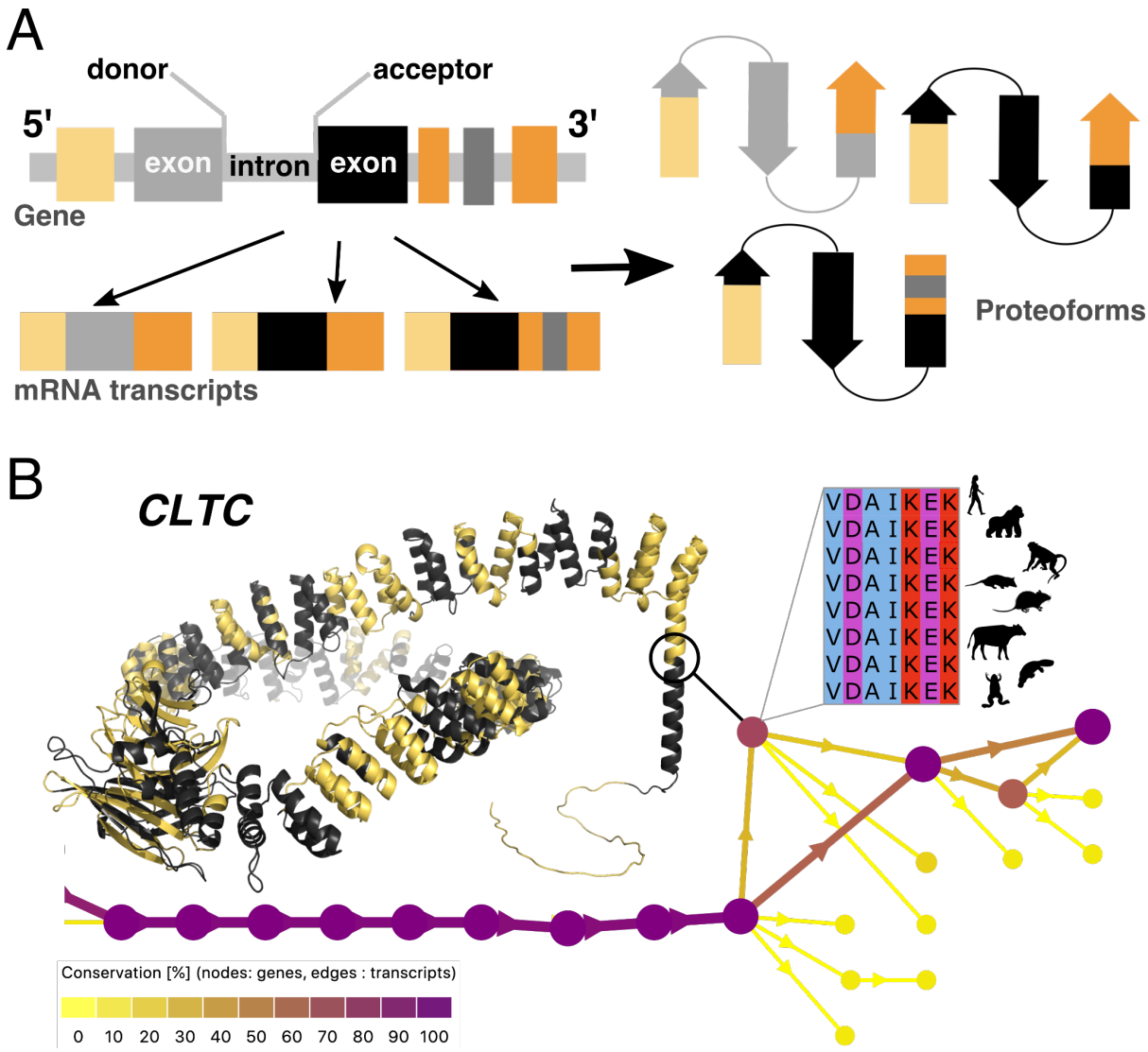

Figure 1. Basics of alternative splicing and an illustrative example. (A) Schematic representation of a eukaryotic gene, focusing on the protein-coding region. The high boxes are the exons, separated by thin boxes depicting the introns. The donor and acceptor splice sites are at the exon-intron and intron-exon boundaries. Three mRNA transcripts corresponding to different combinations of exons are shown. They may be translated into proteoforms adopting different shapes (arrows: β-sheet, rectangle: α-helix). **(B)** The human gene *CLTC* encodes the 1675-residue long clathrin heavy chain 1 whose 3D model (https://alphafold.ebi.ac.uk/search/text/Q00610 [26]) is displayed as cartoons. The evolutionary splicing graph at the bottom recapitulates the alternative proteoforms observed over eleven species from human to zebrafish (http://www.lcqb.upmc.fr/Ases http://www.lcqb.upmc.fr/Ases/results?jobid=KXFyXXbHm3 [27]), focusing on the C-terminal protein region. The nodes or s-exons

are coloured according to their conservation level (species fraction) on the graph and delineated in black and yellow on the 3D structure. The most conserved event, present in human, gorilla, macaque, rat, cow, opossum, platypus, and frog, is a 7-aa insertion in the protein C-terminal trimerization domain. This insertion, supported by transcriptomic and proteomic data, triggers a switch from spherical clathrin-coated pits to flat clathrin lattices during muscle cell differentiation [9].

Deciphering the splicing code

Computationally recognising the genomic signals determining which protein-coding segments will be spliced together by the spliceosome is a fundamental step for describing proteoform diversity. Donor (or 5') splice sites, at the exon–intron junctions, typically feature a GT dinucleotide, and acceptor (or 3') splice sites, at the intron–exon junctions, an AG dinucleotide (**Figure 1A**). Nevertheless, not all GT-AG pairs signify splicing, some splice sites may feature non-canonical patterns, and other environmental factors and regulatory signals come into play, making the task challenging [28]. While early splicing code models relied on putative regulatory features [29], the most recent predictors recognize splice sites directly from raw genomic or pre-mRNA nucleotide sequences [30–37] (**Figure 2**). They borrow deep learning architectures from image classification like convolutional neural networks (CNN) or from natural language processing like transformers.

Among these next-generation splicing predictors, the ultra-deep residual CNN-based model SpliceAI [30] analyses pre-mRNA genomic sequences to compute the probability of each residue being a splice donor, splice acceptor, or neither. It has proven effective in predicting splicing alterations, exon skipping, and splicing rescue through cryptic site activation [38]. SpliceAI performance is matched by large language models (LLMs) pre-trained to reconstruct masked or corrupted genomic sequences at scale [35]. SpliceAI and LLMs evaluate thousands of nucleotides around the position of interest, up to 32 kb with HyenaDNA [39]. However, accounting for wider contexts does not necessarily translate into improved accuracy [35] and may not reflect what the spliceosome can recognize in the cell [31-32]. Several predictors reach state-of-the-art accuracy by focusing on shorter sequences, and a few further constrain their architectures to more closely mimic the splicing process and improve interpretability [32,37]. For instance, the SAM splice site predictor is explicitly informed with knowledge about sparse RNA-binding protein motifs [37].

Models trained on one species typically exhibit low generalisation capability to other species [32]. Scalzitti and co-authors explicitly addressed this issue by training the Spliceator model [31] on a carefully curated benchmark set encompassing a hundred phylogenetically diverse organisms. In addition, using the predictors to assess the impact of alterations in the input sequence on the splicing outcome requires choosing appropriate thresholds that may depend on factors not explicitly modelled, such as splice site strength or exons' baseline inclusion rates [40]. Pangolin [33] and TrASPr [36] make a step forward by quantifying AS splice site usage and events under specific conditions (*e.g.*, tissue), opening the way to design genomic sequences tuned to desired splicing outcomes (**Figure 2**).

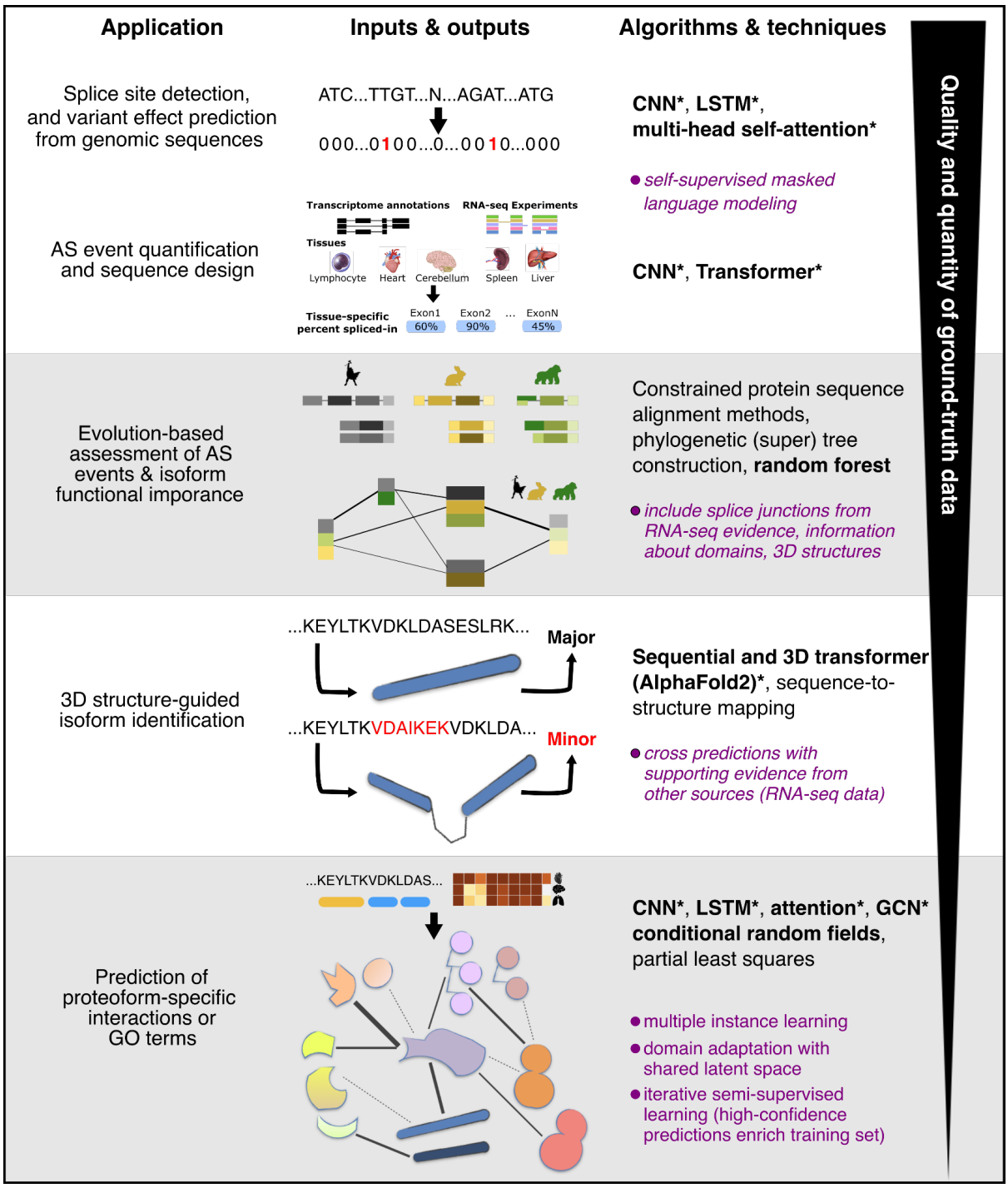

Figure 2. Overview of methods and applications for shedding light on alternative splicing contribution to proteome diversification. We mostly focus on approaches developed in recent years, which often integrate one or more of the mentioned algorithms and techniques. The latter are classified into heuristic (plain text), classical machine learning (bold) and deep learning (starred bold). The bullet points in italics and purple indicate strategies for coping with lack of ground-truth data.

Leveraging evolutionary conservation

Evolutionary conservation often serves as a reliable indicator of function, suggesting that natural variations induced by AS, and selected through evolution, likely fulfil important

functional roles under physical and environmental constraints. For instance, mutually exclusive tandem duplicated exons (MXE) are an example of ancient AS events that have critical functional significance [41,42]. Substitutions in these exons have likely contributed to tissue and organ evolution in metazoans and have clinical implications in humans [41]. More broadly, cross-species conservation is the most discriminating feature for state-of-the-art prediction of transcript biological relevance at the protein level [43]. Reciprocally, AS variations disrupting conserved active sites and functional domains are unlikely to result in functional translated products.

To accurately assess the evolutionary conservation of AS events, it is necessary to match exons, splice junctions, or transcripts/proteoforms across species. Early methods have relied on genomic sequence alignments to identify orthologous exons [44]. However, difficulties arise from large indels, ambiguities between highly similar or short sequences, or lack of plausible matches for highly divergent sequences. A few recent methods address these challenges by adopting an end-product perspective, working with the amino acid sequences of the putative proteoforms enriched with knowledge about the gene structure [45,46].

In particular, evolutionary splicing graphs provide a compact representation summarising the full proteoform diversity observed for a set of orthologous genes [45] (**Figure 1B**). By extending the concept of splicing graphs [47] to several species, they allow for identifying (sub-)exon orthogroups (nodes), quantifying splice junction usage (edges), and investigating exon co-occurrence (paths). Building such graphs from annotations and RNA-seq data across a dozen species spanning 800 million years of evolution showed a clear link between conservation, tissue regulation, and functional relevance of alternative transcripts [45]. Furthermore, disentangling orthologous from paralogous relationships between entire transcripts/proteoforms [48,49] and simulating or reconstructing transcript phylogenies [8,50] can help to infer evolutionary scenarios explaining AS-induced protein function diversification.

Modelling proteoform 3D structures

While only a few tens of alternative splicing proteoforms have experimentally resolved 3D structures, the advent of high-throughput deep learning-based protein structure prediction methods, which achieve near-experimental accuracy, has enabled systematic probing of AS impact on protein folds and structural stability [51,52]. Sommer and colleagues [52] proposed using AlphaFold2 [26] average predicted local distance difference test (pLDDT) score as a measure of "biological functionality" for genome annotation (**Figure 2**). One should be cautious with such an approach because short, well-folded fragments from larger proteins often display higher pLDDT scores than the full-length protein, potentially misleading functional interpretations. Reciprocally, a proteoform with a longer inter-domain disordered linker would be penalised in terms of pLDDT while it may acquire the ability to translocate to another cellular compartment or bind to new partners. The authors partially addressed these issues by applying a series of filtering criteria based on proteoform length, pLDDT distribution and RNA-seq expression data [52].

They identified 940 alternative human proteoforms with pLDDT scores suggesting they might be more functionally active than those annotated as primary in the MANE (Matched Annotation from NCBI and EMBL-EBI) database [53]. Evolutionary wise, these alternative proteoforms span a wide range of conservation levels (**Figure 3A**). Some of them, like the mu opioid receptor OPRM1 proteoform lacking the N-terminus and first helix, are much less

conserved than their MANE counterpart (**Figure 3A-B**). These observations suggest that cross-species conservation could be useful to refine the approach.

In addition, the suitability of AlphaFold2 for predicting some alternative proteoform structures is questionable. For instance, AlphaFold2 tends to model AS-induced large deletions in well-folded domains through cut-and-stitch with low confidence scores assigned to the stitched region (**Figure 3C**). A recent study highlighted how this comparative modelling-like behaviour produces physically unrealistic 3D models for alternative proteoforms where patches of hydrophobic residues are exposed to the solvent [54]. This limitation, also shared by protein Language Model(pLM)-based protein structure predictors, emphasises the need for methods tailored to extract signals from AS-induced sequence variations.

Variations on the same theme

AS of duplicated protein regions enables fine-tuning of protein function without altering the protein fold [42,55]. Systematically mapping the MXEs identified in 5 high-quality Metazoan genomes to the CATH-derived protein domain functional families (FunFams) revealed that MXE-specific residues are mostly located on the protein surface and cluster at or near protein functional sites [42]. A parallel study focusing on human and relaxing the criterion of mutual exclusivity confirmed these findings on AlphaFold-predicted 3D models and further showed that MXE amino acid substitutions tend to affect disordered residues or residues that do not directly bind ligands [55].

Beyond MXE pairs, the combinatorics of AS-induced topological rearrangements of similar exonic sequences can be more complex [56]. For instance, AS produces four different combinations of calmodulin-binding motifs in myosin 1b's lever arm [45], thereby modulating the protein’s ability to sense mechanical forces and hence to pull membranes [57]. The giant skeletal muscle protein nebulin gives a more extreme example where over 100 protein regions, each corresponding to one or more ~35 aa-long nebulin-like motifs, are subject to 47 inclusion/exclusion events across a dozen species, including two primates, six other mammals, zebrafish and frog [56]. In the leucine-rich-repeat containing G-protein coupled receptor 5, the number of repeats is modulated by three inclusion/exclusion AS events, among which two are conserved from human to frog. When the repeats have well-defined structures, the exon-intron gene structure and the AS-induced variations tend to preserve their integrity and can help refine their boundaries [58].

Jointly analysing the proteoforms of tandem repeat-containing proteins observed in several species can help to gain insights into how these protein regions have evolved essential functions [59] and more broadly, into the relationship between AS and gene duplication [60].

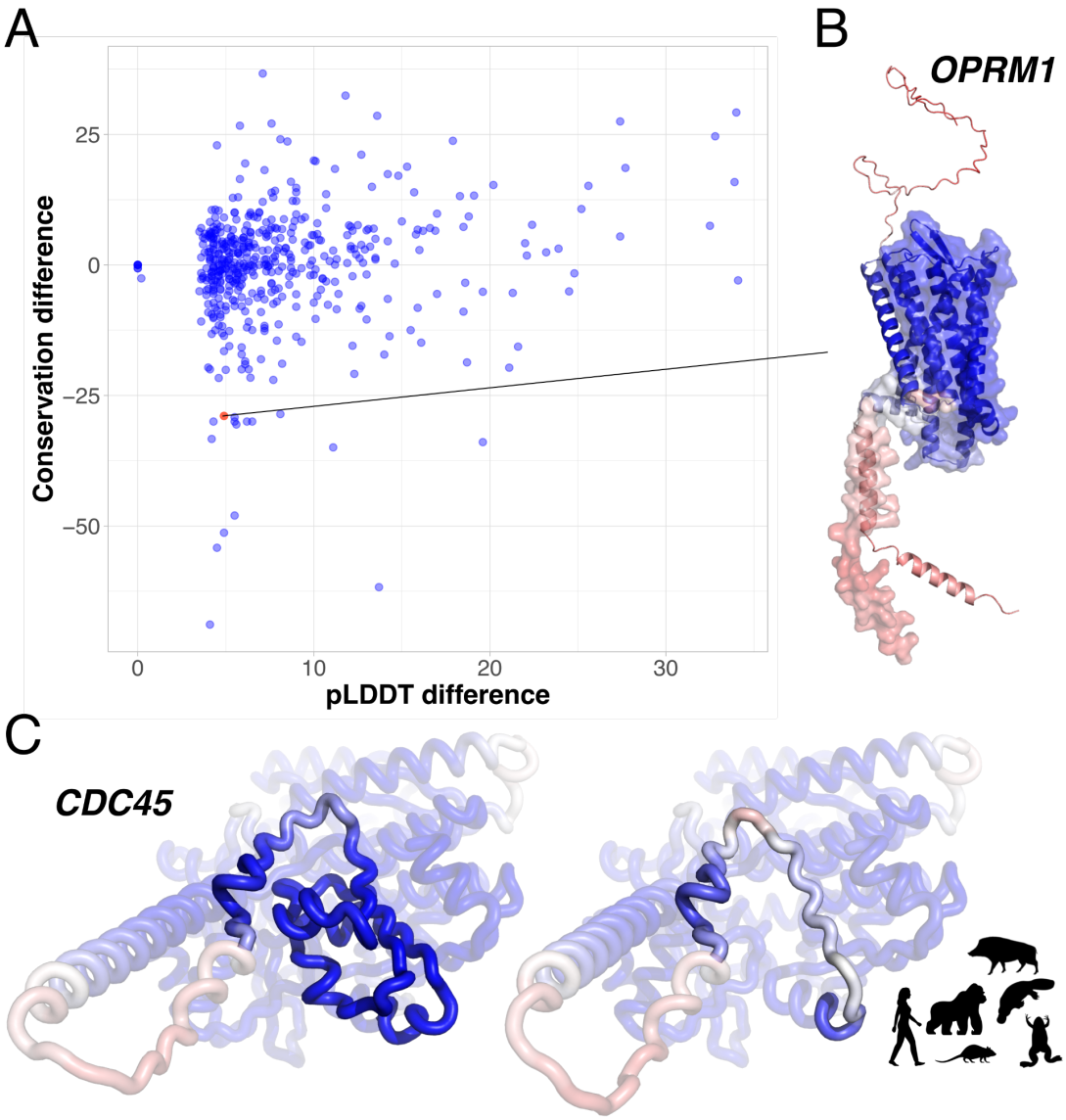

Figure 3. Structural modelling and evolutionary conservation of alternative splicing events. (A) The alternative proteoforms identified in [52] as having more stable structures (higher pLDDT, x-axis) than MANE-annotated primary proteoforms are not necessarily more conserved in evolution (y-axis). ThorAxe estimated evolutionary conservation as the average splice junction usage across a dozen species, from human to nematode (averaged transcript fraction mean in [45]). **(B-C)** AlphaFold2-predicted 3D models (from isoform.io v3.1 [52]), coloured according to the pLDDT, from red (low) to blue (high), for OPRM1 and CDC45 proteoforms. **(B)** The mu opioid receptor OPRM1 MANE proteoform (in cartoon) represents a full-length canonical G protein-coupled receptor (GPCR) structure with an extracellular N-terminus followed by seven transmembrane alpha-helices. The candidate alternative proteoform (in surface) lacking the N-terminus and the first transmembrane helix has a higher pLDDT but much lower conservation. This truncated form does not retain the function of the full-length receptor, while it could modulate the function of other GPCRs [61]. **(C)** AlphaFold2 modelled exon 4 skipping in CDC45 as cut-and-stitch. The exon seems essential for function since it encodes a part of the RecJ nuclease-orthologue's DHH domain. Yet, exon 3-5 junction is expressed in low levels in several tissues based on GTEx mRNA data [62], and ThorAxe detected it in several species, from human to frog, based on Ensembl annotations (https://www.ensembl.org/).

Unveiling AS impact on interactions and functions

AS events can rewire protein-protein interaction (PPI) networks by altering functional motifs in intrinsically disordered regions, gaining or losing entire structured domains, or inducing small changes in their interacting surfaces [5,6,10,44,63-64]. To move forward in assessing the functional role of AS on a system biology level, researchers have systematically mapped known human protein interactions and functional annotations on genomic exons [65–67]. The Domain Interaction Graph Guided ExploreR (DIGGER) database even enables an exon-centric exploration of human PPIs by exploiting information about physical contacts between residues from experimental 3D complexes [65]. Nevertheless, experimental data about exon-exon interactions cover only about 5% of all known human PPIs, stressing the need for producing and analysing high-quality 3D models [68].

Machine learning approaches have emerged to predict proteoform-specific interactions and functions, but they face challenges such as the scarcity of ground-truth data and the heterogeneity of proteoform-related data [69] (**Figure 2**). Multiple instance learning (MIL) algorithms address the first issue by integrating genomic and protein-level information. In these frameworks, a gene is conceptualised as a *bag* containing its proteoforms, which are treated as *instances* within the bag. Functional annotations are initially assigned at the gene level (the bag) and then propagated to its proteoforms (the instances), with refinements made to ensure proteoform-specific accuracy [70–72]. The attention mechanism can be used to increase the difference between isoform pairs from the same gene bag in the context of interaction prediction [71] or to account for the fact that two or more proteoforms from the same gene can work together to accomplish the same function [73]. Semi-supervised learning offers an alternative approach to MIL in which high-confidence predictions generated from unlabeled samples are iteratively added to the training set to refine the model [74].

In recent years, modular deep learning architectures have been proposed to deal with the second issue, the heterogeneity of the input data [69,75]. They typically consider nucleotide and/or aa sequences, RNA-seq data and optionally domain composition or domain-domain interactions. For instance, the proteoform function predictor DIFFUSE combines features extracted with convolutions from proteoform aa sequences, transcript co-expression RNA-seq data aggregated with probabilistic graphical models, and information about evolutionary conserved domains treated with LSTM [69]. Further developments aim at a more unified integration, either through end-to-end deep learning architectures [73] or by projecting the input data into a common latent space with partial least squares [76].

Beyond protein-protein complexes, AS events affecting protein interactions with nucleic acids, lipids, carbohydrates and small molecules have also been documented [10]. Protein-ligand interactions established through binding interfaces found in specific proteoforms expressed in specific tissues provide new opportunities for drug design and targeting strategies, as exemplified by the GPCR superfamily [15].

Concluding remarks and future outlook

Emerging deep learning paradigms show promise in clarifying how AS contributes to protein functional diversification. Unlike traditional methods, and provided sufficient compute and data, deep learning techniques excel at automatically extracting meaningful features directly from raw data, eliminating the need for labour-intensive feature engineering. However, a lack

of high-quality ground-truth data poses significant challenges for training and evaluating machine learning, especially deep learning approaches. Gene predictions frequently contain errors [77], and alternative splicing proteoform-specific functional annotations and interactions are scarce, partial, and likely noisy [74]. Defining reliable negative training sets is extremely difficult because it is almost impossible to demonstrate that splice variants do not have any function or cannot interact with one another [43]. Traditional train/test splitting may lead to training sets that are not representative enough. Data augmentation is often impractical because we do not know or control the impact of small changes on protein interactions and functions.

Self-supervised representation learning could allow for overcoming some of these limitations. Large foundation models, in particular, enable the transfer of knowledge across genes, proteins, and species by leveraging universal representation spaces. This approach enhances robustness and reduces sensitivity to errors and ambiguities in input data preprocessing, such as those arising from multiple sequence alignments. However, not all learned representations may achieve high quality due to uneven coverage and biases in the training data. Looking ahead, multimodal generative models like ESM-3 (https://github.com/evolutionaryscale/esm), which build upon the breakthroughs in protein structure prediction, are starting to jointly reason over protein sequences, structures, and functions. These models open exciting possibilities for finely capturing proteoform differences and even designing or tweaking specific proteoforms.

Exciting progress is also being made in top-down proteomics [78] and nanopore protein sequencing [79] aided by machine learning for signal processing and recognition. They let us envision the possibility of comprehensively identifying proteoforms generated by AS and post-translational modifications. Beyond identification, emerging RNA-targeting strategies for programmable manipulation of AS – such as synthetic splicing factors or recruitment of the endogenous splicing machinery – open the way to systematic functional exon screening [80]. Although challenges remain before achieving high resolution and throughput, these experimental techniques are starting to peel back the layers of complexities of protein states and functioning in the cell.

Funding: Funded/Co-funded by the European Union (ERC, PROMISE, 101087830). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Acknowledgements: For the purpose of Open Access, a CC-BY public copyright licence has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission.

Bibliography

[1] Parisi G, Palopoli N, Tosatto SCE, Fornasari MS, and Tompa P. “Protein” no longer means what it used to. Current Research in Structural Biology 2021;3:146–52.

[2] Graveley, Brenton R. Alternative splicing: increasing diversity in the proteomic world. TRENDS in Genetics 2001 17: 100-107.

[3] Smith LM, Kelleher NL. Proteoforms as the next proteomics currency. Science 2018; 359:1106–1107.

[4] Birzele F, Csaba G, Zimmer R. Alternative splicing and protein structure evolution.

Nucleic Acids Res 2007;36:550–8.

[5] Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell 2016;164:805–17.

[6] Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol 2017;18:437–51.

[7] Tapial J, Ha KCH, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. Genome Res 2017;27:1759–68.

[8] Ait-Hamlat A, Zea DJ, Labeeuw A, Polit L, Richard H, Laine E. Transcripts' Evolutionary History and Structural Dynamics Give Mechanistic Insights into the Functional Diversity of the JNK Family. J Mol Biol 2020;432:2121–40.

[9] Moulay G, Lainé J, Lemaître M, Nakamori M, Nishino I, Caillol G, et al. Alternative splicing of clathrin heavy chain contributes to the switch from coated pits to plaques. J Cell Biol 2020;219.

[10] Kjer-Hansen P, Weatheritt RJ. The function of alternative splicing in the proteome: rewiring protein interactomes to put old functions into new contexts. Nat Struct Mol Biol 2023;30:1844–56.

[11] Nakka K, Ghigna C, Gabellini D, Dilworth FJ. Diversification of the muscle proteome through alternative splicing. Skelet Muscle 2018;8:1–18.

[12] Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, et al. A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. Cell 2014;159:1511–23.

[13] Schaub A, Glasmacher E. Splicing in immune cells—mechanistic insights and emerging topics. Int Immunol 2017;29:173–81.

[14] Park E, Pan Z, Zhang Z, Lin L, Xing Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am J Hum Genet 2018;102:11–26.

[15] Marti-Solano M, Crilly SE, Malinverni D, Munk C, Harris M, Pearce A, et al. Combinatorial expression of GPCR isoforms affects signalling and drug responses. Nature 2020;587:650–6.**
Comprehensive study of human GPCR alternative splicing proteoforms, their structural and signaling properties, their combinatorial expression in specific tissues. This work proposes a context-specific view of GPCR signaling and sketches strategies for developing drugs exploiting proteoform selectivity.

[16] Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet 2016;17:19–32.

[17] Climente-González H, Porta-Pardo E, Godzik A, Eyras E. The Functional Impact of Alternative Splicing in Cancer. Cell Rep 2017;20:2215–26.

[18] Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. Nat Methods 2024;21:1349–63.

[19] Light S, Elofsson A. The impact of splicing on protein domain architecture. Curr Opin Struct Biol 2013;23:451–8.

[20] Bénitière F, Necsulea A, Duret L. Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans 2024. https://doi.org/10.7554/eLife.93629.**
Large-scale evaluation of the correlation between effective population size and genome-wide AS rates across 53 metazoan species (vertebrates and insects). The results support the drift-barrier hypothesis according to which AS rate variations reflect the limits set by drift on the capacity of selection to prevent gene expression errors.

[21] Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML. Most highly expressed protein-coding genes have a single dominant isoform. J Proteome Res 2015;14:1880–7.

[22] Miller RM, Jordan BT, Mehlferber MM, Jeffery ED, Chatzipantsiou C, Kaur S, et al. Enhanced protein isoform characterization through long-read proteogenomics. Genome Biol 2022;23:69.

[23] Agosto LM, Gazzara MR, Radens CM, Sidoli S, Baeza J, Garcia BA, et al. Deep profiling and custom databases improve detection of proteoforms generated by alternative splicing. Genome Res 2019;29:2046–55.

[24] Weatheritt RJ, Sterne-Weiler T, Blencowe BJ. The ribosome-engaged landscape of alternative splicing. Nat Struct Mol Biol 2016;23:1117–23.

[25*] Sinitcyn P, Richards AL, Weatheritt RJ, Brademan DR, Marx H, Shishkova E, et al. Global detection of human variants and isoforms by deep proteome sequencing. Nat Biotechnol 2023;41:1776–86.
Global assessment of the impact of genomic variants and AS at the protein level across six different cell lines with six different proteases. This study identified 17,717 unique proteins with high median coverage and showed that most frame-preserving alternative transcripts are translated.

[26] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021; 596: 583-589.

[27] Zea DJ, Richard H, Laine E. ASES: visualizing evolutionary conservation of alternative splicing in proteins. Bioinformatics 2022;38:2615–6.

[28] Wilkinson ME, Charenton C, and Nagai K. RNA splicing by the spliceosome. *Annual review of biochemistry* 2020;89:359-388.

[29] Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature 2010;465:53–9.

[30] Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell 2019;176:535–48.e24.

[31*] Scalzitti N, Kress A, Orhand R, Weber T, Moulinier L, Jeannin-Girardon A, et al. Spliceator: multi-species splice site prediction using convolutional neural networks. BMC Bioinformatics 2021;22:561.
Convolutional neural network for *ab initio* prediction of eukaryotic multi-species splice sites, trained on a high-quality manually curated benchmark containing genomic sequences from human to protists.

[32] Chao K-H, Mao A, Salzberg SL, Pertea M. Splam: a deep-learning-based splice site predictor that improves spliced alignments. bioRxiv 2023. https://doi.org/10.1101/2023.07.27.550754.

[33] Zeng T, Li YI. Predicting RNA splicing from DNA sequence using Pangolin. Genome Biol 2022;23:103.

[34] Liu Q, Fang H, Wang X, Wang M, Li S, Coin LJM, et al. DeepGenGrep: a general deep learning-based predictor for multiple genomic signals and regions. Bioinformatics 2022;38:4053–61.

[35] Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Carranza NL, Grzywaczewski AH, Oteri F, et al. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. bioRxiv 2023:2023.01.11.523679. https://doi.org/10.1101/2023.01.11.523679.

[36] Wu D, Jha A, Jewell S, Maus N, Gardner JR, Barash Y. Generative modeling for RNA splicing code predictions and design 2023. https://openreview.net/forum?id=UZTpkfw0aC

[37*] Gupta K, Yang C, McCue K, Bastani O, Sharp PA, Burge CB, et al. Improved modeling of RNA-binding protein motifs in an interpretable neural model of RNA splicing. Genome Biology 2024;25:23.
Modular deep learning architecture enforcing sparsity for an interpretable end-to-end prediction of splice sites. The model predicts the motifs establishing physical interactions with RNA-binding proteins as intermediate output. Contrasting predictions from the different modules produces a putative regulatory landscape for any exon.

[38] Strauch Y, Lord J, Niranjan M, Baralle D. CI-SpliceAI-Improving machine learning predictions of disease causing splicing variants using curated alternative splice sites. PLoS One 2022;17:e0269159.

[39] Nguyen E, Poli M, Faizi M, Thomas A, Birch-Sykes C, Wornow M, et al. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. ArXiv 2023.

[40] Smith C, Kitzman JO. Benchmarking splice variant prediction algorithms using massively parallel splicing assays. Genome Biol 2023;24:1–22.

[41] Martinez Gomez L, Pozo F, Walsh TA, Abascal F, Tress ML. The clinical importance of tandem exon duplication-derived substitutions. Nucleic Acids Res 2021;49:8232–46.

[42*] Lam SD, Babu MM, Lees J, Orengo CA. Biological impact of mutually exclusive exon switching. PLoS Comput Biol 2021;17:e1008708.
First large-scale, structure-based analysis of the biological impact of homologous mutually exclusive exons in five high-quality Metazoan genomes. The findings suggest an implication of these exons in interactions with other proteins or ligands (no disruption of CATH domains, high solvent exposure, proximity to functional sites).

[43*] Pozo F, Martinez-Gomez L, Walsh TA, Rodriguez JM, Di Domenico T, Abascal F, et al. Assessing the functional relevance of splice isoforms. NAR Genom Bioinform 2021;3:lqab044.
Machine learning predictor of the functional importance of AS proteoforms validated on unbiased data from large-scale mass spectrometry proteomics experiments. The most important features are conservation-based.

[44] Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science 2012;338:1587–93.

[45] Zea DJ, Laskina S, Baudin A, Richard H, Laine E. Assessing conservation of alternative splicing with evolutionary splicing graphs. Genome Res 2021;31:1462–73.**
The first method efficiently estimating AS conservation across many species. This work provides a formal generalisation of splicing graphs to many genes/species, an efficient heuristic to build *evolutionary* splicing graphs, a new curated set of functional AS events, and granular estimates of AS evolutionary conservation at the human protein-coding genome scale.

[46] Márquez Y, Mantica F, Cozzuto L, Burguera D, Hermoso-Pulido A, Ponomarenko J,

et al. ExOrthist: a tool to infer exon orthologies at any evolutionary distance. Genome Biol 2021;22:239.

[47] Heber S, Alekseyev M, Sze S-H, Tang H, Pevzner PA. Splicing graphs and EST assembly problem. Bioinformatics 2002;18 Suppl 1:S181–8.

[48] Guillaudeux N, Belleannée C, Blanquart S. Identifying genes with conserved splicing structure and orthologous isoforms in human, mouse and dog. BMC Genomics 2022;23:216.

[49] Ouedraogo WYDD, Ouangraoua A. Orthology and Paralogy Relationships at Transcript Level. J Comput Biol 2024;31:277–93.

[50] Ouedraogo WYDD, Ouangraoua A. SimSpliceEvol2: alternative splicing-aware simulation of biological sequence evolution and transcript phylogenies. BMC Bioinformatics 2024;25:235.

[51] Osmanli Z, Falgarone T, Samadova T, Aldrian G, Leclercq J, Shahmuradov I, et al. The Difference in Structural States between Canonical Proteins and Their Isoforms Established by Proteome-Wide Bioinformatics Analysis. Biomolecules 2022;12:1610.

[52*] Sommer MJ, Cha S, Varabyou A, Rincon N, Park S, Minkin I, et al. Structure-guided isoform identification for the human transcriptome. Elife 2022;11.
Large-scale evaluation of over 230,000 human alternative splicing proteoforms with 3D structure predictions generated by AlphaFold. This study proposes to use the average pLDDT as a proxy for functional importance and identifies hundreds of alternative proteoforms apparently more stable than their canonical counterparts.

[53] Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature 2022;604:310–5.**
Community-wide effort to define a set of single representative transcripts for every protein-coding genes, for standardizing clinical genomics and research. Criteria include expression levels and conservation.

[54] Zhang Z, Wayment-Steele HK, Brixi G, Wang H, Peraro MD, Kern D, et al. Protein language models learn evolutionary statistics of interacting sequence motifs. PNAS, 121.

[55] Martinez-Gomez L, Cerdán-Vélez D, Abascal F, Tress ML. Origins and Evolution of Human Tandem Duplicated Exon Substitution Events. Genome Biol Evol 2022;14.

[56*] Szatkownik A, Zea DJ, Richard H, Laine E. Building alternative splicing and evolution-aware sequence-structure maps for protein repeats. J Struct Biol 2023;215:107997.
Robust and versatile computational method for systematically assessing the alternative usage of repeated protein regions across many species and mapping the information on protein 3D structures.

[57] Greenberg MJ, Ostap EM. Regulation and control of myosin-I by the motor and light chain-binding domains. Trends Cell Biol 2013;23:81–9.

[58] Paladin L, Necci M, Piovesan D, Mier P, Andrade-Navarro MA and Tosatto SCE. A novel approach to investigate the evolution of structured tandem repeat protein families by exon duplication. Journal of Structural Biology 2020;212:107608.

[59] Delucchi M, Schaper E, Sachenkova O, Elofsson A, Anisimova M. A New Census of Protein Tandem Repeats and Their Relationship with Intrinsic Disorder. Genes 2020;11.

[60] Abascal F, Tress ML, Valencia A. The evolutionary fate of alternatively spliced homologous exons after gene duplication. Genome Biol Evol 2015;7:1392–403.

[61] Marrone GF, Grinnell SG, Lu Z, Rossi GC, Le Rouzic V, Xu J, et al. Truncated mu opioid GPCR variant involvement in opioid-dependent and opioid-independent pain modulatory systems within the CNS. Proc Natl Acad Sci U S A 2016;113:3663–8.

[62] Schoch K, Ruegg MSG, Fellows BJ, Cao J, Uhrig S, Einsele-Scholz S, et al. A second hotspot for pathogenic exon-skipping variants in CDC45. Eur J Hum Genet 2024;32:786–94.

[63] Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, et al. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol Cell 2012;46:871–83.

[64] Roth JF, Braunschweig U, Wu M, Li JD, Lin ZY, Larsen B, et al.. Systematic analysis of alternative exon-dependent interactome remodeling reveals multitasking functions of gene regulatory factors. Mol Cell. 2023;83:4222-4238.e10.

[65*] Louadi Z, Yuan K, Gress A, Tsoy O, Kalinina OV, Baumbach J, et al. DIGGER: exploring the functional role of alternative splicing in protein interactions. Nucleic Acids Res 2021;49:D309–18.
User-friendly database allowing for exploring AS functional role in protein interactions. The resource provides a residue-level mapping of physical interactions between proteins or protein domains.

[66] Tseng Y-T, Li W, Chen C-H, Zhang S, Chen JJW, Zhou X, et al. IIIDB: a database for isoform-isoform interactions and isoform network modules. BMC Genomics 2015;16 Suppl 2:S10.

[67] Tranchevent L-C, Aubé F, Dulaurier L, Benoit-Pilven C, Rey A, Poret A, et al. Identification of protein features encoded by alternative exons using Exon Ontology. Genome Res 2017;27:1087–97.

[68] Liebold J, Del Moral-Morales A, Manalastas-Cantos K, Tsoy O, Kurtz S, Baumbach J, et al. The power and limits of predicting exon-exon interactions using protein 3D structures. bioRxiv 2024:2024.03.01.582917. https://doi.org/10.1101/2024.03.01.582917.

[69] Chen H, Shaw D, Zeng J, Bu D, Jiang T. DIFFUSE: predicting isoform functions from sequences and expression profiles via deep learning. Bioinformatics 2019;35:i284–94.

[70] Wang J, Zhang L, Zeng A, Xia D, Yu J, Yu G. DeepIII: Predicting Isoform-Isoform Interactions by Deep Neural Networks and Data Fusion. IEEE/ACM Trans Comput Biol Bioinform 2022;19:2177–87.

[71] Zeng J, Yu G, Wang J, Guo M, Zhang X. DMIL-III: Isoform-isoform interaction prediction using deep multi-instance learning method. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE; 2019, p. 171–6.

[72] Yu G, Zeng J, Wang J, Zhang H, Zhang X, Guo M. Imbalance deep multi-instance learning for predicting isoform–isoform interactions. Int J Intell Syst 2021;36:2797–824.

[73] Qiu S, Yu G, Lu X, Domeniconi C, Guo M. Isoform function prediction by Gene Ontology embedding. Bioinformatics 2022;38:4581–8.

[74*] Narykov O, Johnson NT, Korkin D. Predicting protein interaction network perturbation by alternative splicing with semi-supervised learning. Cell Rep 2021;37:110045.
Semi-supervised learning approach, namely iterative self-learning random forest, to predict alternative splicing proteoform-specific interactions. The method exploits general knowledge about protein interactions (statistical potential) and information about known PPIs to estimate AS-induced changes in binding partners.

[75] Yu G, Zhou G, Zhang X, Domeniconi C, Guo M. DMIL-IsoFun: predicting isoform function using deep multi-instance learning. Bioinformatics 2021;37:4818–25.

[76] Li H-D, Yang C, Zhang Z, Yang M, Wu F-X, Omenn GS, et al. IsoResolve: predicting splice isoform functions by integrating gene and isoform-level features with domain adaptation. Bioinformatics 2021;37:522–30.

[77] Meyer C, Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes. BMC Bioinformatics 2020;21:513.

[78] Roberts DS, Loo JA, Tsybin YO, Liu X, Wu S, Chamot-Rooke J, et al. Top-down proteomics. Nat Rev Methods Primers 2024;4. https://doi.org/10.1038/s43586-024-00318-2.

[79] Zhang M, Tang C, Wang Z, Chen S, Zhang D, Li K, et al. Real-time detection of 20 amino acids and discrimination of pathologically relevant peptides with functionalized nanopore. Nat Methods 2024;21:609–18.

[80] Schmok JC and Yeo GW. Strategies for programmable manipulation of alternative splicing. Current Opinion in Genetics & Development 2024, 89:102272.