*Nucleic Acids Research*, 2025, **53**, D331–D339
https://doi.org/10.1093/nar/gkae1018
Advance access publication date: 12 November 2024
**Database issue**

# ASpdb: an integrative knowledgebase of human protein isoforms from experimental and AI-predicted structures

**Yuntao Yang[1], Himansu Kumar[1], Yuhan Xie[2], Zhao Li[1], Rongbin Li[1], Wenbo Chen[1], Chiamaka S. Diala[1], Meer A. Ali[1], Yi Xu[1], Albon Wu[3], Sayed-Rzgar Hosseini[1], Erfei Bi[4], Hongyu Zhao[2], Pora Kim[1,*] and W. Jim Zheng[1,*]**

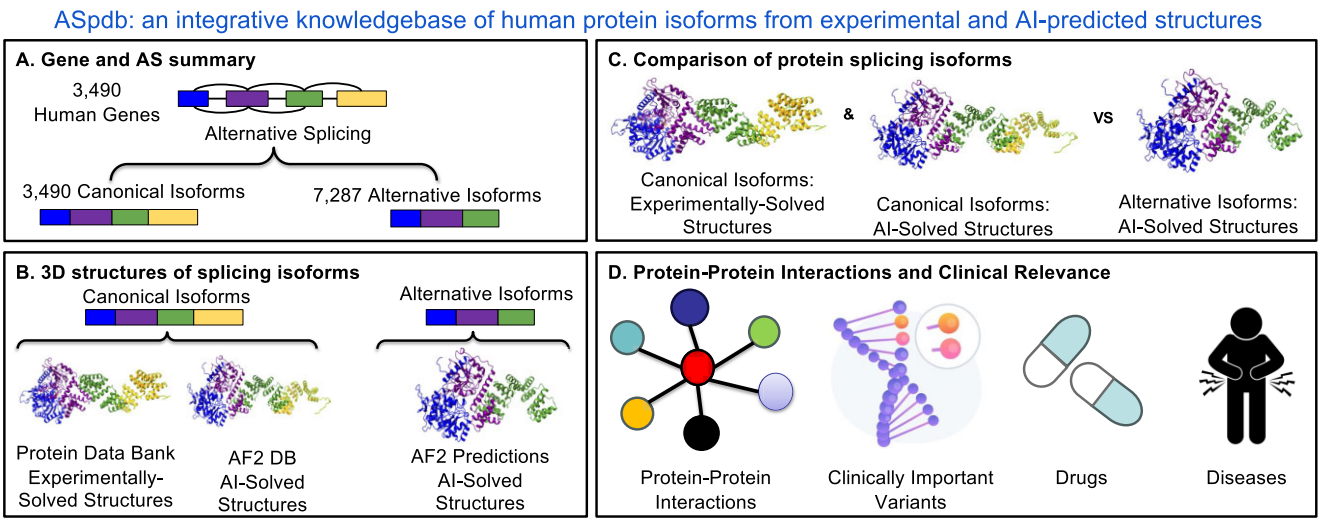
[1]McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin Street, Suite 600, Houston, TX 77030, USA
[2]Department of Biostatistics, Yale University School of Public Health, 300 George Street, Set 503, New Haven, CT 06511, USA
[3]Department of Computer Science and Engineering, University of Michigan, 2260 Hayward Street, Ann Arbor, MI 48109-2121, USA
[4]Department of Cell and Developmental Biology, University of Pennsylvania Perelman School of Medicine, Room 1156, BRB II/III, 421 Curie Boulevard, Philadelphia, PA 19104-6058, USA

[*]To whom correspondence should be addressed. Tel: +1 713 500 3641; Email: Wenjin.J.Zheng@uth.tmc.edu
Correspondence may also be addressed to Pora Kim. Email: Pora.Kim@uth.tmc.edu

## Abstract

Alternative splicing is a crucial cellular process in eukaryotes, enabling the generation of multiple protein isoforms with diverse functions from a single gene. To better understand the impact of alternative splicing on protein structures, protein–protein interaction and human diseases, we developed ASpdb (https://biodataai.uth.edu/ASpdb/), a comprehensive database integrating experimentally determined structures and AlphaFold 2-predicted models for human protein isoforms. ASpdb includes over 3400 canonical isoforms, each represented by both experimentally resolved and predicted structures, and >7200 alternative isoforms with AlphaFold 2 predictions. In addition to detailed splicing events, 3D structures, sequence variations and functional annotations, ASpdb uniquely offers comparative analyses and visualization of structural alterations among isoforms. This resource is invaluable for advancing research in alternative splicing, structural biology and disease mechanisms.

## Graphical abstract




## Introduction

Alternative splicing (AS) is a cellular mechanism that generates diverse mRNA and protein isoforms from a single gene (1), impacting up to 95% of human genes with multiple exons (2). This process leads to the production of various splicing isoforms, including canonical and alternative variants. Canon-

**Received:** August 15, 2024. **Revised:** October 13, 2024. **Editorial Decision:** October 15, 2024. **Accepted:** October 16, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/nar/article/53/D1/D331/7893317 by guest on 19 February 2025

ical isoforms, as defined by UniProtKB (3), are typically the most prevalent, conserved, highly expressed and longest, while alternative isoforms encompass all other variants that differ from the canonical form. These splicing isoforms often have distinct protein structures, which can result in different functions, playing crucial roles in a wide range of biological processes and diseases (4–6). AS and other genomic regulatory mechanisms are intricately interconnected, leading to the complex regulation of biological functions (7). Notably, abnormal AS is frequently linked to tumor progression, resulting in the generation of diverse isoforms and an expanded protein expression profile (8). This dysregulation of AS can lead to the production of oncogenic isoforms, which may promote tumorigenesis by altering metabolism, inhibiting apoptosis and inducing angiogenesis (9). Moreover, the aberrant expression of these isoforms can contribute to the development of drug resistance in cancer therapies, complicating treatment strategies (10). Understanding the mechanisms of AS in diseases (11) is therefore essential for identifying new therapeutic targets and enhancing patient outcomes.

Comprehending the structural variations among splicing isoforms is crucial for devising new therapeutic strategies for diseases linked to AS. Several databases have explored the relationship between AS and human diseases. OncoSplicing, for instance, focuses on clinically relevant AS in cancer, providing survival analysis, differential splicing analysis and annotations of splicing-associated transcripts across 33 cancer types (12). Similarly, ASCancer Atlas is a comprehensive knowledgebase dedicated to aberrant AS in human cancers, featuring 2006 experimentally validated cancer-associated splicing events, a splicing regulatory network and multi-scale analysis tools for investigating splicing dysregulation (13). Advances in protein structure prediction, particularly with AlphaFold2 (AF2) (14), have enabled the exploration of the relationship between AS and human diseases at the protein structure level. Databases like CHESS (15) now include structural data for 130 700 distinct protein isoforms, while APPRIS has also integrated AF2-predicted protein structures to enhance functional annotations of isoforms (16). However, while these databases provide valuable insights, OncoSplicing and ASCancer Atlas do not include any protein structures or structural analysis. In contrast, databases like CHESS and APPRIS provide protein structural data, but they lack statistical analysis to infer structural changes caused by AS events. Additionally, they do not offer clinical information, such as associations with drugs and human diseases. Moreover, none of these databases offer comparative visualization of structural alterations between canonical and alternative isoforms.

To address these gaps, we developed ASpdb, a specialized resource for documenting and analyzing human splicing isoforms. ASpdb uniquely provides both experimentally solved structures and AF2-predicted models for AS isoforms, offering comprehensive coverage of isoform structures. Additionally, we provide comprehensive reliability assessments of AF2-predicted structures using multiple evaluation metrics. To further advance the understanding of AS-induced structural changes, ASpdb employs statistical tests to infer structural alterations between isoforms. The database also integrates information on drugs and human diseases associated with AS. Combined with tools for comparative visualization of canonical and alternative isoforms, ASpdb is an invaluable resource for researchers investigating the structure-function implications of AS in human diseases. Table 1 presents a comparison of ASpdb with other AS databases, including OncoSplicing (12), ASCancer Atlas (13), CHESS (15) and APPRIS (16).

## Materials and methods

### Gene summary and AS summary

We extracted gene information, protein functions and sequences from the reviewed UniProtKB resource (release-2022_03) (3) for 22 349 canonical isoforms and their 39 536 alternative isoforms. Canonical isoforms were selected based on the availability of experimentally solved structures obtained through X-ray crystallography and cryo-electron microscopy; structures determined by NMR were excluded due to the presence of multiple conformations for a single protein. To analyze sequence changes resulting from AS events, we cataloged the start and end points of AS regions, identified the specific types of splicing events (such as substitutions or deletions), and detailed the exact sequence modifications for both canonical and alternative isoforms. For a thorough comparison, we performed multiple sequence alignments (MSAs) of all splicing isoforms for each protein using MUSCLE (17). Additionally, we integrated isoform-level gene expression data from the TCGA (https://www.cancer.gov/tcga) and GTEx (18) databases, visualizing expression patterns across various tissues and conditions through heatmaps.

### 3D structures of human protein isoforms

We extracted 60 878 experimentally solved structures for canonical isoforms from the Protein Data Bank (PDB) (19). To ensure alignment with their corresponding UniProtKB sequences, we renumbered the residues using PDBrenum (20). When multiple structures were available for a single isoform, we selected those with the largest coverage and highest resolution. Additionally, we retrieved AF2-predicted structures for canonical isoforms from the AF2 DB (21). For alternative isoforms, we predicted their 3D structures using the AF2 workflow. To overcome AF2's limited parallel computing limitations, we enhanced efficiency with ParaFold (22), a parallel version of AF2, within a high-performance computing environment. Specifically, we deployed the CPU component on computer clusters at the Texas Advanced Computing Center and the GPU component on Nvidia GPU servers (23). In ASpdb, we successfully predicted structures for 7287 alternative isoforms and also included structures for 3490 canonical isoforms from both PDB (19) and AF2 DB (21), using 2021-07-15 as the maximum template date for AF2 predictions.

We utilized the AlphaPickle program (24) to extract per-residue pLDDT scores, which represent the confidence in AF2-predicted structures, from the 'result_model_.pkl' files generated by AF2. These scores were incorporated into 3D structural models and converted into CIF format using the PDB_EXTRACT tool (25). The CIF files were then visualized using Mol* program (26) to evaluate the accuracy of the predictions at the residue level. We also used scatter plots to visualize pLDDT score distributions across the structures and generated Ramachandran plots to assess conformational angles and structural integrity. Scores above 90 denote high accuracy suitable for detailed applications, scores between 70 and 90 represent good backbone predictions, scores between 50 and 70 suggest low confidence requiring caution and scores below 50 often indicate unstructured regions that should not

Downloaded from https://academic.oup.com/nar/article/53/D1/D331/7893317 by guest on 19 February 2025

**Table 1.** Comparison of ASpdb with existing AS databases

| Database | AS isoform structural data | Comparative analysis of AS isoform structures | Clinical relevance |
|---|---|---|---|
| ASpdb | Experimentally solved & AF2-predicted structures | Superimposed structures, statistical analysis and 3D web visualization | Drugs & disease associations |
| OncoSplicing | No | No | Cancer-specific splicing |
| ASCancer Atlas | No | No | Cancer-specific splicing |
| CHESS | AF2-predicted structures | No | No |
| APPRIS | AF2-predicted structures | No | No |

be interpreted. The pLDDT scores above 90 denote high accuracy, 70–90 indicate good backbone predictions, 50–70 suggest low confidence and scores below 50 often reflect unstructured regions. In addition, we extracted the MSA information from the 'features.pkl' file and generated heatmaps to display the sequence coverage and identity of all sequences mapped to the input sequence in AF2.

To ensure the accuracy and usability of the predicted models, we further refined these structures using the Protein Preparation Wizard module of Schrödinger (Schrödinger Release 2024-2, Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2024) (27). After refinement, we used Schrödinger's SiteMap module (Schrödinger Release 2024-2: SiteMap, Schrödinger, LLC, New York, NY, 2024) (28,29) to predict active sites within the protein structures. Identifying these active sites is critical for understanding protein function and aids researchers in using our predicted structures for drug discovery and biochemical studies.

### Comparison of protein isoform structures and structural features

To conduct a comprehensive comparison of protein structures, we utilized the TM-score, which generates superimposed structures and quantifies their similarity on a scale from 0 to 1, where 1 indicates perfect alignment (30). We applied this metric in two key comparisons: first, to assess the AF2-predicted canonical structures against the experimentally solved canonical structures, and second, to examine the structural changes between canonical structures and alternative isoforms following AS events. This approach allows us to evaluate the accuracy of the predictions and to identify significant structural variations induced by AS.

Further refining our analysis, we examined structural changes induced by AS at the individual residue level. Using the DSSP program within Biopython (31), we annotated each residue in the AF2-predicted structures with detailed secondary structures and relative accessible surface (RAS) area measurements. To identify changes in the secondary structures due to AS events, we conducted Fisher's exact tests for each type of secondary structure, considering a $P$-value $<0.05$ as indicative of a statistically significant alteration. For RAS area variations, we implemented the Mann-Whitney U test to determine statistically significant differences post-AS events. These detailed analyses allow us to precisely identify and characterize specific structural alterations at the residue level that stem from AS events, providing deeper insights into the molecular impacts of splicing variations on protein function.

### Protein–protein interactions and clinical relevance

To enrich our structural annotations, we integrated functional regions related to protein–protein interactions and AS events from the reviewed resource in UniProtKB (release-2022_03) (3). This integration facilitates a deeper understanding of how splicing variations might affect protein functionality and interaction networks. Additionally, we extracted clinically significant variants for each isoform from the ClinVar database (32), which provides insights into genetic predispositions to diseases. We also incorporated information about drugs targeting these genes from the DrugBank database (version 5.1.12) (33), which helps connect potential therapeutic interventions to specific protein variants influenced by AS events. To explore the association between AS events and human diseases, we conducted a systematic text mining approach using PubMed abstracts. We began by searching for abstracts containing the term 'AS', ensuring that our retrieval captured a wide range of studies, including those focused on individual splicing isoforms. Next, we cross-referenced these abstracts with gene-level information from the 'gene2pubmed' file, available through the NCBI FTP site. This allowed us to link relevant literature to specific genes involved in AS. By incorporating this gene-level data, we were able to associate AS events with a broad range of gene-disease connections. To enhance the disease relevance of our results, we filtered abstracts using human disease-related Medical Subject Headings terms, ensuring that the retrieved articles were directly associated with diseases of interest. This methodical approach enabled us to selectively compile and analyze abstracts directly relevant to the clinical implications of AS events for a specific gene.

This comprehensive integration of interaction data, genetic variants, pharmacological associations and literature-based evidence offers an unparalleled view of each isoform's biological and clinical relevance—filling a critical gap left by other databases covering AS isoforms. By linking structural changes from AS events with functional and clinical data, our database provides researchers with a powerful, multifaceted toolkit for investigating the profound implications of AS in human diseases.

## Results

### Overview of ASpdb

Figure 1 highlights the core features of ASpdb, which include four essential categories: (1) gene and AS summaries, (2) 3D structures of splicing isoforms, (3) comparisons of protein splicing isoforms, including structural comparison between canonical and each of its corresponding AS isoforms and (4) protein–protein interactions with clinical relevance. These comprehensive visualization tools offer invaluable insights into AS, revealing its profound impact on protein structure and enabling direct comparisons with canonical isoforms—a feature not provided by any of the existing

Downloaded from https://academic.oup.com/nar/article/53/D1/D331/7893317 by guest on 19 February 2025

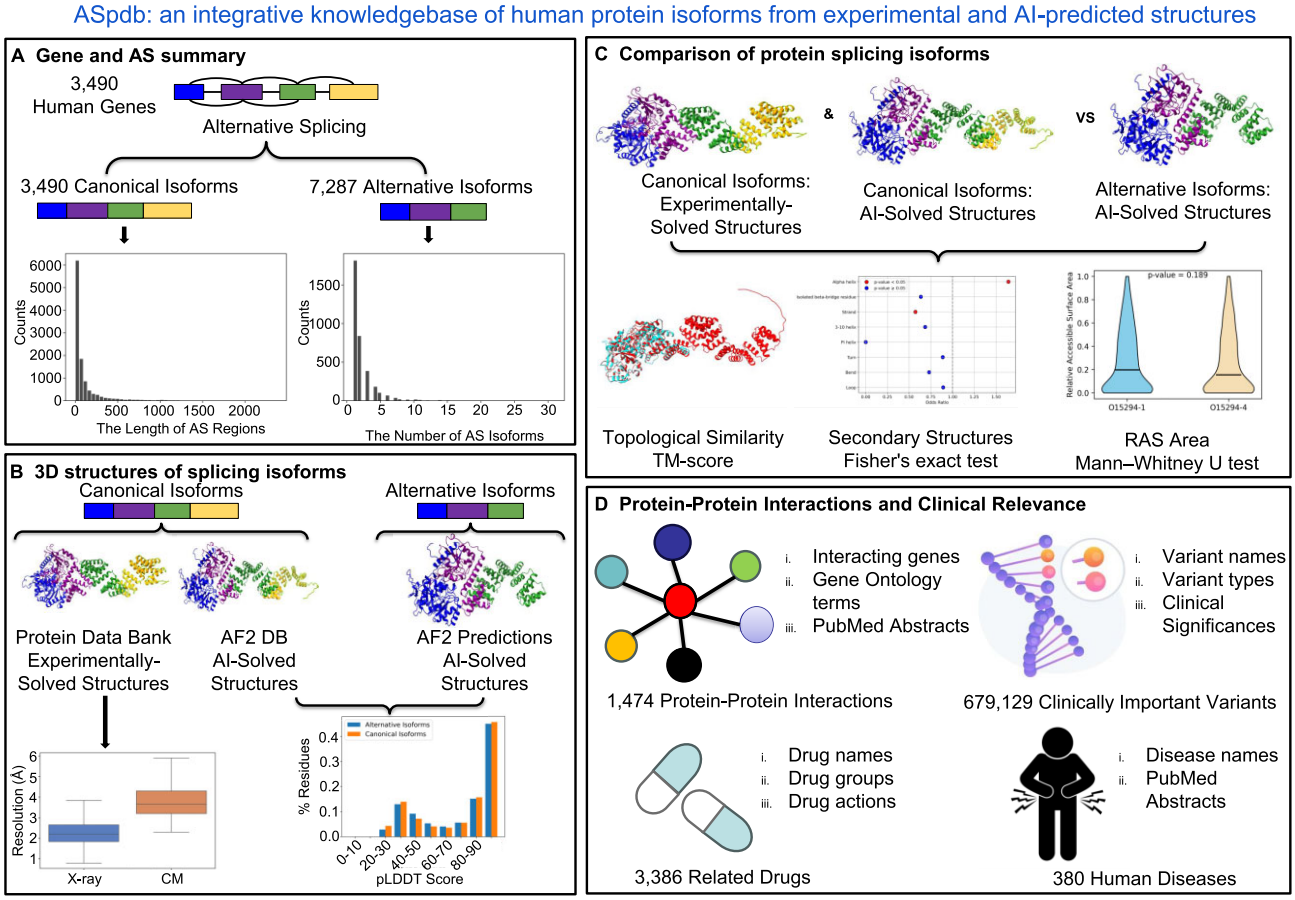



**Figure 1.** Overview of the ASpdb database. (**A**) Summary of gene isoform structure, gene expression and AS information for 3490 human genes. The left bar plot displays the distribution of AS region lengths in canonical isoforms, while the right bar plot illustrates the distribution of the number of AS isoforms per gene. The information provided for gene summary includes: gene overview, gene isoform structure and gene expression across TCGA and GTEx. The information provided for AS summary include AS knowledge, isoform amino acid sequences, MSAs and functional domain retention. (**B**) Experimentally solved and AF2-predicted 3D structures of protein isoforms, including active sites and reliability using pLDDT distribution, MSA heatmap and Ramachandran plot. CM: Cryo-electron microscopy. (**C**) Comparison of 3D structures between canonical and alternative isoforms using TM-scores and statistical tests on changes in structure features. RAS: Relative accessible surface. (**D**) Overview of protein–protein interactions, clinically important variants, related drugs and associated human diseases.

AS databases. Moreover, ASpdb uniquely addresses the influence of AS on protein–protein interactions and their clinical significance—an area overlooked by existing AS databases. Our user-friendly web interface empowers researchers to thoroughly explore these complex relationships, paving the way for new discoveries in the field.

We provide details below about the subcategories of each of these core features:

I. The 'gene and AS summary' category includes four subcategories: 'protein summary', 'AS summary', 'protein functional features' and 'gene isoform structures and expression levels'.
   1. The 'protein summary' subcategory offers a foundational view of the gene and protein, including details such as the gene symbol, ID and description. It also provides Gene Ontologies (GO) associated with the gene (34,35).
   2. The 'AS summary' subcategory provides isoform sequences along with their Ensembl (36) and RefSeq IDs (37), as well as comprehensive AS information such as the start and end points of AS regions, the types of splicing events, and the sequence alterations. Additionally, it facilitates the comparison of isoform sequences through MSA.
   3. The 'protein functional features' subcategory includes the main functions of the protein and highlights the regions where functional features overlap with AS events.
   4. The 'gene isoform structures and expression levels' subcategory includes gene isoform structures in the UCSC Genome Browser (38) and gene isoform expression levels displayed in heatmaps.

II. The '3D structures of splicing isoforms' category includes five subcategories: 'protein structures', 'pLDDT score distribution', 'Ramachandran plot of protein structures', 'heatmap of MSA coverage' and 'potential active site information'.
   1. The 'protein structures' subcategory includes 3D viewers for AF2-predicted isoform structures, with residues colored according to their pLDDT scores.
   2. The 'pLDDT score distribution' subcategory includes scatter plots that visualize the pLDDT score distribution of AF2-predicted isoform structures.

Downloaded from https://academic.oup.com/nar/article/53/D1/D331/7893317 by guest on 19 February 2025

3. The 'Ramachandran plot of protein structures' subcategory includes Ramachandran plots of AF2-predicted isoform structures.
4. The 'heatmap of MSA coverage' subcategory includes heatmaps that display sequence coverage and identity for all sequences mapped to the input sequence in AF2.
5. The 'potential active site information' subcategory provides predicted active sites for each isoform.

III. The 'comparison of protein splicing isoforms' category includes one subcategory: 'protein structure and feature comparison'.
1. The 'protein structure and feature comparison' provides 3D viewers for superimposed isoform structures, bar plots to visualize TM-scores across various comparisons, scatter plots for the Fisher's exact test analysis of secondary structures, and violin plots for the Mann-Whitney U test of RAS area.

IV. The 'protein–protein interactions and clinical relevance' category includes four subcategories: 'protein–protein interaction', 'related drugs', 'related diseases' and 'clinically important variants'.
1. The 'protein–protein interaction' subcategory provides details on 1474 protein–protein interactions, including the names of interacting genes, ontology terms and corresponding PubMed abstracts.
2. The 'related drugs' subcategory includes 3386 drugs targeting 1007 genes, with information on drug names, drug groups (e.g. approved, experimental) and drug actions (e.g. inhibitor, agonist).
3. The 'related diseases' subcategory links the AS of 711 genes to 380 diseases, providing information such as disease names, PubMed titles and abstracts.
4. The 'clinically important variants' subcategory identifies 679 129 clinically significant variants across 3315 isoform genes, providing information on variant names, variant types and clinical significances.

### Example scenarios

To demonstrate the capabilities of ASpdb, we delved into the AS profiles of several extensively researched genes, exploring their regulatory mechanisms and functional impacts in various biological processes and human diseases.

*Scenario I*. Figure 2 presents a detailed analysis of the NF2 (Merlin) gene, highlighting its transcript-level expressions and associated human diseases. ASpdb includes a comprehensive summary of the NF2 gene, which functions as a well-known tumor suppressor (Figure 2A, Supplementary Table S1). Figure 2B and Supplementary Table S2 illustrates the GO terms associated with the NF2 gene, encompassing highly tumor-related terms such as cell population proliferation, cell–cell adhesion and signal transduction. Importantly, the NF2 gene consists of 10 isoforms and 22 AS events (Figures 2C, Supplementary Table S3) (39), each contributing to the gene's functional diversity. ASpdb also provides detailed tables converting UniProt IDs to their corresponding Ensembl IDs (Figure 3C, Supplementary Table S4) and RefSeq IDs (Figure 3D, Supplementary Table S5). Figure 2E (left) and Supplementary Figure S1 illustrate the gene structure of NF2 isoforms, providing an overview of their exon–intron arrangements. In Figures 2E (right) and Supplementary Figure S2, we identified isoforms P35240-1 (ENST00000338641), P35240-5 (ENST00000361452) and P35240-9 (ENST00000413209) as particularly noteworthy due to their higher expression levels in breast, colon and brain tissues, according to data from the GTEx database. Additionally, we identified P35240-9 (ENST00000413209) as significantly expressed in brain tissues, suggesting its critical role in normal brain function. This finding may contribute to a deeper understanding of the role of NF2 isoforms in various cancers, such as breast cancer (40), colorectal cancer (41) and meningioma (42). Understanding the expression patterns and functions of these isoforms can provide insights into their potential as diagnostic markers or therapeutic targets in these malignancies.

*Scenario II*. Figure 3 presents a detailed analysis of the APH1A (Aph-1 Homolog A) gene, highlighting the structural distinctions and therapeutic implications of its isoforms in Alzheimer's disease. ASpdb provides AF2-predicted structures and their pLDDT distributions for the APH1A isoforms, APH-1aL (Q96BI3-1) and APH-1aS (Q96BI3-2), as shown in Figure 3A and 3B. Both isoforms show high-confidence pLDDT scores (>90) across most of their sequences, with Q96BI3-1 having lower confidence at the C-terminus and Q96BI3-2 showing moderate confidence in the middle regions. The MSA heatmap for APH-1aS (Q96BI3-2) is also included to illustrate the high MSA sequence identity and coverage in AF2 (Supplementary Figure S3). ASpdb also reveals critical structural distinctions between these two APH1A isoforms, as shown by the superimposed structures (Figure 3C), even though their TM-score is nearly 1, with APH-1aL featuring a longer helical tail compared to APH-1aS. The γ-Secretase complexes containing APH1A isoforms demonstrate proteolytic activity essential for processing the amyloid precursor protein and thereby producing amyloid-β, a major factor in Alzheimer's disease pathology (43). Among these, APH-1aL is the predominant isoform in endogenous γ-secretase complexes, playing a vital role in their effective function due to its enhanced stability relative to APH-1aS (44). Notably, structural variations between the isoforms can influence isoform-specific interactions within the γ-secretase complex, impacting substrate specificity and enzymatic activity. Significant alterations in the Pi-helix structure were identified using Fisher's exact test ($P < 0.05$), indicating statistically significant changes in secondary structure due to AS events. However, the Mann-Whitney U test revealed no statistically significant differences in the RAS area post-AS events, further underscoring the specific nature of these structural variations (Figure 3D). Additionally, ASpdb includes information on E-2012, a gamma secretase modulator targeting APH1A, which is being evaluated as a potential treatment for Alzheimer's disease (Figure 3E). In summary, ASpdb showcases the role of APH1A isoforms in Alzheimer's disease, detailing the structural and functional differences of APH-1aL and APH-1aS, their impact on γ-secretase activity, and the therapeutic potential of targeting these isoforms with modulators like E-2012.

## Discussion

Current studies on AS and its associated protein structures are often limited to simply providing isoform structures, lacking the in-depth annotation and functional analysis necessary to truly understand AS's impact on protein structure, interactions and clinical relevance. To bridge this critical gap, we developed ASpdb, a cutting-edge resource that revolutionizes AS and protein structure comparison. By integrating experimentally determined structures with AF2-predicted models,

Downloaded from https://academic.oup.com/nar/article/53/D1/D331/7893317 by guest on 19 February 2025

**A**

| Gene name: NF2 | |
|---|---|
| ASpdb.0 ID: 4771 | |
| | **Gene** |
| **Gene symbol** | NF2 |
| **Gene ID** | 4771 |
| **Gene name** | NF2, moesin-ezrin-radixin like (MERLIN) tumor suppressor |
| **Synonyms** | ACN\|BANF\|SCH\|SWNV\|merlin-1 |
| **Cytomap** | 22q12.2 |
| **Type of gene** | protein-coding |
| **Description** | merlinbilateral acoustic neurofibromatosismoesin-ezrin-radixin likemoesin-ezrin-radixin-like proteinmoesin-ezrin-radizin-like proteinneurofibromin 2 (bilateral acoustic neuroma)neurofibromin-2schwannomerlinschwannomin |
| **Modification date** | 20240407 |
| **UniProtAcc** | P35240 |

**B**

| Partner | Gene | GO ID | GO term | PubMed ID |
|---|---|---|---|---|
| Gene | NF2 | GO:0005634 | nucleus | 10401006\|20178741 |
| Gene | NF2 | GO:0005730 | nucleolus | 9537418 |
| Gene | NF2 | GO:0005737 | cytoplasm | 9537418 |
| Gene | NF2 | GO:0005769 | early endosome | 10861283 |
| Gene | NF2 | GO:0005829 | cytosol | - |
| Gene | NF2 | GO:0005886 | plasma membrane | 10401006 |
| Gene | NF2 | GO:0008285 | negative regulation of cell population proliferation | 12444102\|20178741 |
| Gene | NF2 | GO:0022408 | negative regulation of cell-cell adhesion | 17210637 |
| Gene | NF2 | GO:0042532 | negative regulation of tyrosine phosphorylation of STAT protein | 12444102 |
| Gene | NF2 | GO:0046426 | negative regulation of receptor signaling pathway via JAK-STAT | 12444102 |
| Gene | NF2 | GO:0048471 | perinuclear region of cytoplasm | 17210637 |

**C**

| Accession ID | Gene Name | Canonical ID | Alternative ID | Canonical Length | Alternative Length | Canonical Start | Canonical End | Type | Original SEQ | Variation SEQ | Alternative Start | Alternative End |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P35240 | NF2 | P35240-1 | P35240-5 | 595 | 549 | 81 | 121 | Deletion | none | none | 80 | 80 |
| P35240 | NF2 | P35240-1 | P35240-5 | 595 | 549 | 580 | 590 | Substitution | LTLQSAKSRVA | PQAQGRRPICI | 539 | 549 |
| P35240 | NF2 | P35240-1 | P35240-5 | 595 | 549 | 591 | 595 | Deletion | none | none | 549 | 549 |
| P35240 | NF2 | P35240-1 | P35240-9 | 595 | 165 | 150 | 579 | Deletion | none | none | 149 | 149 |

**D**

| UniProt-id | ENSG | ENST | ENSP |
|---|---|---|---|
| P35240-1 | ENSG00000186575.19 | ENST00000338641.10 | ENSP00000344666.5 |
| P35240-5 | ENSG00000186575.19 | ENST00000361452.8 | ENSP00000354897.4 |
| P35240-9 | ENSG00000186575.19 | ENST00000413209.6 | ENSP00000409921.2 |

| UniProt-id | NM ID | NP ID |
|---|---|---|
| P35240-1 | NM_000268.3 | NP_000259.1 |
| P35240-5 | NM_181829.2 | NP_861967.1 |
| P35240-9 | NM_181833.2 | NP_861971.1 |

**E**

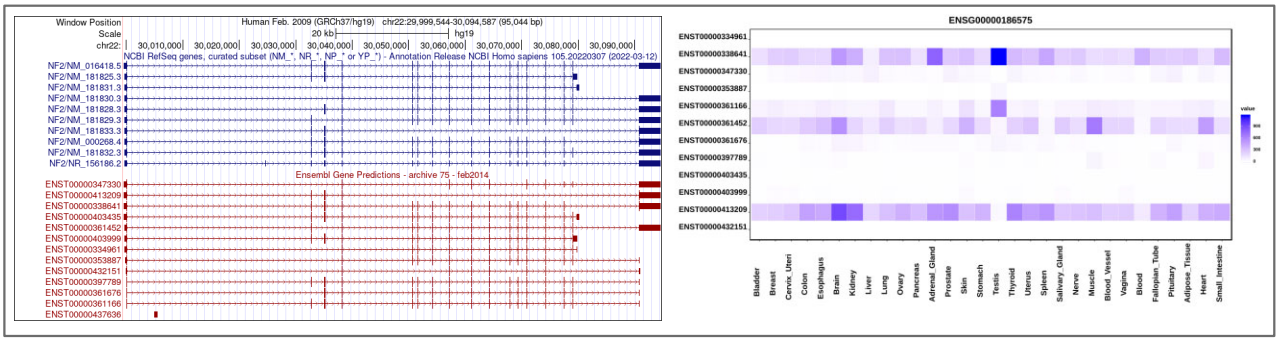
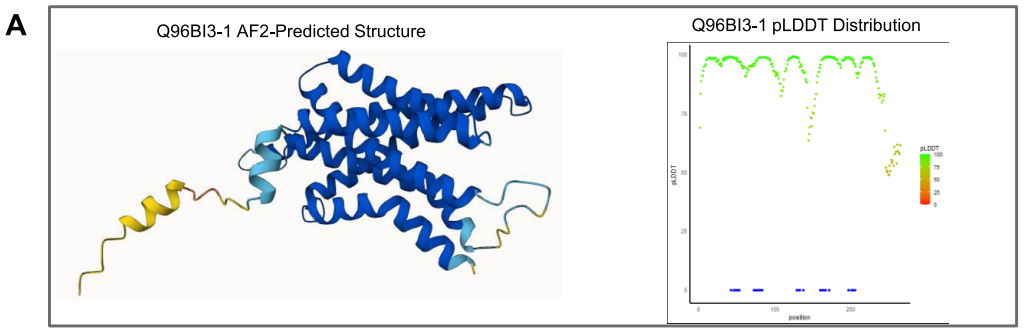



**Figure 2.** An illustration of the comprehensive analysis results of NF2 gene. The detailed information for each panel can be found in Supplementary Tables S1–S5 and Supplementary Figures S1–S2. (**A**) Gene summary, including gene name, ID and description. (**B**) Gene ontology terms with evidence from Entrez and associated PubMed IDs. (**C**) AS and isoform information for three NF2 isoforms, listing canonical and alternative spliced isoforms and modifications. (**D**) Conversion tables of UniProt, Ensembl and RefSeq IDs, specific to the three isoforms. (**E**) Gene structures of canonical and alternatively spliced genes, along with their expression levels across GTEx tissues, visualized as heatmaps.

ASpdb provides an unparalleled repository, offering access to over 3400 canonical isoforms and >7200 alternative isoforms, complete with detailed 3D structures and annotations. This allows researchers to gain valuable insights into how splicing variations influence protein function and contribute to disease pathology. Moreover, ASpdb uniquely enables comparative analyses and visualizations of structural alterations among isoforms, linking structural data with functional annotations and clinical significance. This integration marks a crucial advancement towards personalized medicine, facilitating the identification of disease-associated splicing events and potential therapeutic targets. To ensure the reliability of the AF2-predicted structures, we compared AF2 predictions to three recently determined experimental PDB structures, with an average RMSD of 0.9 Å (Supplementary Table S6), demonstrating AF2's accuracy in predicting AS isoform structures. Additionally, a comparison between AF2 and AlphaFold 3 (AF3) predictions for nine representative isoforms revealed no significant differences in structural quality (45) (Supplementary Table S7), with consistent pLDDT scores and RMSD values (0.3 to 1.1), confirming the reliability of AF2's predictions.

Despite its strengths, ASpdb has several limitations. First, while ASpdb offers a robust and comprehensive dataset, its accuracy hinges on the predictive power of AF2. AF2 generates highly precise models, particularly for protein isoforms with experimentally solved structures, there remains room to improve confidence and precision, especially for isoforms containing intrinsically disordered regions. Future advancements in AF2 will be key in addressing these challenges. Furthermore, AF2's predictions are primarily derived from protein crystal structures in laboratory settings, which do not fully capture the complexity of *in vivo* conditions. Given that protein folding is influenced by factors such as post-translational modifications and cellular environments, addressing this lim-

Downloaded from https://academic.oup.com/nar/article/53/D1/D331/7893317 by guest on 19 February 2025




E

| UniProt accession | Gene name | DrugBank ID | Drug name | Drug group | Actions |
|---|---|---|---|---|---|
| Q96BI3 | APH1A | DB05171 | E-2012 | investigational | |
| Q96BI3 | APH1A | DB12852 | MK-0752 | investigational | inhibitor |

**Figure 3.** Detailed structural information of APH-1 gene and its isoforms. (**A**) AF2-predicted structure of the long isoform APH-1aL (Q96BI3-1) and its pLDDT distribution. (**B**) AF2-predicted structure of the short isoform APH-1aS (Q96BI3-2) and its pLDDT distribution. (**C**) Superimposed structure of APH-1aL (Q96BI3-1) and APH-1aS (Q96BI3-2) with the corresponding TM-score. (**D**) Comparison of secondary structure and RAS area changes between APH-1aL (Q96BI3-1) and APH-1aS (Q96BI3-2). (**E**) Drugs targeting APH1A isoforms.

Downloaded from https://academic.oup.com/nar/article/53/D1/D331/7893317 by guest on 19 February 2025

itation is critical. Second, incorporating isoform-level interactions with proteins and other molecular complexes—such as nucleic acids, small molecules, ions and modified residues—into ASpdb could significantly enhance our understanding of AS and its functional implications. AF3 offers advanced capabilities for predicting these interactions (45), which could make future versions of ASpdb more comprehensive and insightful for studying the consequences of AS. Lastly, while our text mining approach is not strictly isoform-specific, as it still relies on the availability of literature that explicitly mentions isoforms. Therefore, future refinements could focus on more targeted extraction of isoform-specific insights as more data becomes available. As AI-driven protein structure predictions continue to advance, ASpdb is set to become an essential resource, supporting research in AS, structural biology and translational medicine.

## Data availability

All annotation results are available from the ASpdb website (https://biodataai.uth.edu/ASpdb). Further information and requests should be directed to Dr. W. Jim Zheng (Wenjin.J.Zheng@uth.tmc.edu).

## Supplementary data

Supplementary Data are available at NAR Online.

## Funding

National Institutes of Health (NIH) [1UL1TR003167, 1UM1TR004906-01, 1R01AG066749, 1U24MH130988-01, 5R56AG069880-02, R01GM116876, R35GM138184]; Department of Defense [W81XWH-22-1-0164]; Cancer Prevention and Research Institute of Texas [RP170668]; University of Texas Health Science Center at Houston; NLM Training Program [5T15LM007093-31]. Funding for open access charge: University resource.

## Conflict of interest statement

None declared.

## References

1. Zheng,C.L., Fu,X.D. and Gribskov,M. (2005) Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA*, **11**, 1777–1787.
2. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
3. UniProt,Consortium2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
4. Gallego-Paez,L.M., Bordone,M.C., Leote,A.C., Saraiva-Agostinho,N., Ascensão-Ferreira,M. and Barbosa-Morais,N. (2017) Alternative splicing: the pledge, the turn and the prestige: the key role of alternative splicing in human biological systems. *Hum. Genet.*, **136**, 1015–1042.
5. Sreenivasamurthy,S.K., Madugundu,A.K., Patil,A.H., Dey,G., Mohanty,A.K., Kumar,M., Patel,K., Wang,C., Kumar,A., Pandey,A., *et al.* (2017) Mosquito-borne diseases and omics: tissue-restricted expression and alternative splicing revealed by transcriptome profiling of *Anopheles stephensi*. *OMICS*, **21**, 488–497.
6. Wang,R., Helbig,I., Edmondson,A.C., Lin,L. and Xing,Y. (2023) Splicing defects in rare diseases: transcriptomics and machine learning strategies towards genetic diagnosis. *Brief. Bioinform.*, **24**, bbad284.
7. Wan,J., Oliver,V.F., Zhu,H., Zack,D.J., Qian,J. and Merbs,S.L. (2013) Integrative analysis of tissue-specific methylation and alternative splicing identifies conserved transcription factor binding motifs. *Nucleic Acids Res.*, **41**, 8503–8514.
8. Zhang,Y., Qian,J., Gu,C. and Yang,Y. (2021) Alternative splicing and cancer: a systematic review. *Signal Transduct. Target. Ther.*, **6**, 78.
9. Oltean,S. and Bates,D.O. (2014) Hallmarks of alternative splicing in cancer. *Oncogene*, **33**, 5311–5318.
10. Siegfried,Z. and Karni,R. (2018) The role of alternative splicing in cancer drug resistance. *Curr. Opin. Genet. Dev.*, **48**, 16–21.
11. Brotman,S.M., Raulerson,C.K., Vadlamudi,S., Currin,K.W., Shen,Q., Parsons,V.A., Iyengar,A.K., Roman,T.S., Furey,T.S., Kuusisto,J., *et al.* (2022) Subcutaneous adipose tissue splice quantitative trait loci reveal differences in isoform usage associated with cardiometabolic traits. *Am. J. Hum. Genet.*, **109**, 66–80.
12. Zhang,Y., Yao,X., Zhou,H., Wu,X., Tian,J., Zeng,J., Yan,L., Duan,C., Liu,H. and Li,H. (2022) OncoSplicing: an updated database for clinically relevant alternative splicing in 33 human cancers. *Nucleic Acids Res.*, **50**, D1340–D1347.
13. Wu,S., Huang,Y., Zhang,M., Gong,Z., Wang,G., Zheng,X., Zong,W., Zhao,W., Xing,P. and Li,R. (2023) ASCancer Atlas: a comprehensive knowledgebase of alternative splicing in human cancers. *Nucleic Acids Res.*, **51**, D1196–D1204.
14. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A. and Potapenko,A. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
15. Sommer,M.J., Cha,S., Varabyou,A., Rincon,N., Park,S., Minkin,I., Pertea,M., Steinegger,M. and Salzberg,S.L. (2022) Structure-guided isoform identification for the human transcriptome. *eLife*, **11**, e82556.
16. Rodriguez,J.M., Pozo,F., Cerdán-Vélez,D., Di Domenico,T., Vázquez,J. and Tress,M.L. (2022) APPRIS: selecting functionally important isoforms. *Nucleic Acids Res.*, **50**, D54–D59.
17. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
18. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F. and Young,N. (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
19. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
20. Faezov,B. and Dunbrack,R.L. Jr (2021) PDBrenum: a webserver and program providing Protein Data Bank files renumbered according to their UniProt sequences. *PLoS One*, **16**, e0253411.
21. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G. and Laydon,A. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
22. Zhong,B., Su,X., Wen,M., Zuo,S., Hong,L. and Lin,J. (2022) *International Conference on High Performance Computing in Asia-Pacific Region Workshops*. pp. 1–9.
23. Yang,Y., Li,Z., Shih,D.J. and Zheng,W.J. (2022) AlphaFold 2 Monomer: deployment in an HPC Environment.
24. Arnold,M.J. (2021) AlphaPickle. https://doi.org/10.5281/zenodo.5708709.
25. Yang,H., Guranovic,V., Dutta,S., Feng,Z., Berman,H.M. and Westbrook,J.D. (2004) Automated and accurate deposition of

Downloaded from https://academic.oup.com/nar/article/53/D1/D331/7893317 by guest on 19 February 2025

structures solved by X-ray diffraction to the Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 1833–1839.
26. Sehnal,D., Bittrich,S., Deshpande,M., Svobodova,R., Berka,K., Bazgier,V., Velankar,S., Burley,S.K., Koca,J. and Rose,A.S. (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.*, **49**, W431–W437.
27. Madhavi Sastry,G., Adzhigirey,M., Day,T., Annabhimoju,R. and Sherman,W. (2013) Protein and ligand preparation: parameters, protocols and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.*, **27**, 221–234.
28. Halgren,T.A. (2009) Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.*, **49**, 377–389.
29. Halgren,T. (2007) New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.*, **69**, 146–148.
30. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
31. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
32. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D. and Hoover,J. (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
33. Knox,C., Wilson,M., Klinger,C.M., Franklin,M., Oler,E., Wilson,A., Pon,A., Cox,J., Chin,N.E. and Strawbridge,S.A. (2024) Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Res.*, **52**, D1265–D1275.
34. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. and Eppig,J.T. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
35. Aleksander,S.A., Balhoff,J., Carbon,S., Cherry,J.M., Drabkin,H.J., Ebert,D., Feuermann,M., Gaudet,P. and Harris,N.L. (2023) The gene ontology knowledgebase in 2023. *Genetics*, **224**, iyad031.
36. Harrison,P.W., Amode,M.R., Austine-Orimoloye,O., Azov,A.G., Barba,M., Barnes,I., Becker,A., Bennett,R., Berry,A. and Bhai,J. (2024) Ensembl 2024. *Nucleic Acids Res.*, **52**, D891–D899.
37. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B. and Ako-Adjei,D. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
38. Nassar,L.R., Barber,G.P., Benet-Pagès,A., Casper,J., Clawson,H., Diekhans,M., Fischer,C., Gonzalez,J.N., Hinrichs,A.S. and Lee,B.T. (2023) The UCSC genome browser database: 2023 update. *Nucleic Acids Res.*, **51**, D1188–D1195.
39. Petrilli,A.M. and Fernández-Valle,C. (2016) Role of Merlin/NF2 inactivation in tumor biology. *Oncogene*, **35**, 537–548.
40. Arakawa,H., Hayashl,N., Nagase,H., Ogawa,M. and Nakamura,Y. (1994) Alternative splicing of the NF2 gene and its mutation analysis of breast and colorectal cancers. *Hum. Mol. Genet.*, **3**, 565–568.
41. Čačev,T., Aralica,G., Lončar,B. and Kapitanović,S. (2014) Loss of NF2/Merlin expression in advanced sporadic colorectal cancer. *Cell. Oncol.*, **37**, 69–77.
42. Bachir,S., Shah,S., Shapiro,S., Koehler,A., Mahammedi,A., Samy,R.N., Zuccarello,M., Schorry,E. and Sengupta,S. (2021) Neurofibromatosis type 2 (NF2) and the implications for vestibular schwannoma and meningioma pathogenesis. *Int. J. Mol. Sci.*, **22**, 690.
43. Zhang,H., Ma,Q., Zhang,Y.W. and Xu,H. (2012) Proteolytic processing of Alzheimer's β-amyloid precursor protein. *J. Neurochem.*, **120**, 9–21.
44. Shirotani,K., Edbauer,D., Prokop,S., Haass,C. and Steiner,H. (2004) Identification of distinct γ-secretase complexes with different APH-1 variants. *J. Biol. Chem.*, **279**, 41340–41345.
45. Abramson,J., Adler,J., Dunger,J., Evans,R., Green,T., Pritzel,A., Ronneberger,O., Willmore,L., Ballard,A.J. and Bambrick,J. (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**, 1–3.

Downloaded from https://academic.oup.com/nar/article/53/D1/D331/7893317 by guest on 19 February 2025

**Received:** August 15, 2024. **Revised:** October 13, 2024. **Editorial Decision:** October 15, 2024. **Accepted:** October 16, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.