frontiers in Molecular Biosciences

TYPE Brief Research Report
PUBLISHED 03 November 2026
DOI 10.3389/fmolb.2026.1877678

Check for updates

OPEN ACCESS

EDITED BY
Bruno Hay Mele,
Università telematica San Raffaele, Italy

REVIEWED BY
Käthe M. Dahlström,
Åbo Akademi University, Finland
Ashish Shrivastava,
Translational Health Science and Technology Institute, India

*CORRESPONDENCE
Taylor Richards,
taylor.richards@wlv.ac.uk
Paraskevi Goggolidou,
p.goggolidou@wlv.ac.uk

RECEIVED 10 May 2026
REVISED 03 August 2026
ACCEPTED 28 September 2026
PUBLISHED 03 November 2026

CITATION
Richards T and Goggolidou P (2026) A computational report on Fibrocystin structure via AlphaFold2: considerations and limitations for ARPKD research. *Front. Mol. Biosci.* 13:1877678. doi: 10.3389/fmolb.2026.1877678

COPYRIGHT
© 2026 Richards and Goggolidou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A computational report on Fibrocystin structure via AlphaFold2: considerations and limitations for ARPKD research

Taylor Richards* and Paraskevi Goggolidou*

Faculty of Science and Engineering, University of Wolverhampton, Wolverhampton, United Kingdom

Autosomal Recessive Polycystic Kidney Disease (ARPKD) is a rare, genetic disease that often occurs during early childhood and can be lethal at its most severe presentation. ARPKD is caused by biallelic variants in *PKHD1* which encodes Fibrocystin, a large receptor protein. The disease presentation and progression of ARPKD is highly variable, the reason for this is currently unknown. To address this, AlphaFold2 was employed to generate a model of Fibrocystin which was used to establish spatial relationships between compound heterozygous pairs in ARPKD patients. AlphaFold2 was used to generate a model of full-length human Fibrocystin which was assessed for structural similarities to known protein structures. Compound heterozygous variant pairs and disease outcomes were acquired from previously published work and annotated onto the three-dimensional Fibrocystin model to identify spatial relationships. AlphaFold2 predicted the full-length of Fibrocystin with varying confidence. Globally, average modelling confidence was low, with more confident modelling around the G8 and PBH1 domains. Confidence was lowest around the transmembrane region and intracellular tail, which modelled poorly. FoldSeek identified structural similarities to other proteins; however, these had similar structures to those already predicted present in Fibrocystin. No new structural elements were detected. Spatial relationships between compound heterozygous variants did not correlate with disease outcomes. There are limitations to AlphaFold2 which limits its ability to accurately model a protein like Fibrocystin. These limitations may hinder the tools application in determining genotype-phenotype relationships in ARPKD. Currently, Fibrocystin models produced by AlphaFold2 are unable to identify genotype-phenotype relationships.

KEYWORDS
AlphaFold2, ARPKD, fibrocystin, genotype-phenotype, PKHD1

1 Introduction

Autosomal Recessive Polycystic Kidney Disease (ARPKD) is a severe, rare genetic disease characterised by the presence of kidney cysts and hepatic fibrosis (Guay-Woodford et al., 2014; Bergmann et al., 2018; Goggolidou and Richards, 2022). ARPKD is a ciliopathy with an early onset, frequently manifesting at or prior to birth (Bergmann et al., 2018; Goggolidou and Richards, 2022). The disease can progress rapidly to severe outcomes, including intrauterine death and pulmonary insufficiency (Bergmann et al., 2018). In individuals surviving beyond the first year of life, progressive cystic kidney disease and hepatic fibrosis can be debilitating, with progressive organ injury ultimately leading to kidney and liver failure (Bergmann et al., 2003; Bergmann et al., 2004;

Bergmann et al., 2005). Despite this critical early phenotype, the clinical presentation of ARPKD is highly variable. Some individuals present with milder disease; being diagnosed in childhood or adulthood (Adeva et al., 2006). Furthermore, the kidney and liver manifestations can be discordant, with pronounced severity observed in one organ while the other follows a comparatively mild disease course (Gunay–Aygun et al., 2013). The age at which end-stage kidney or liver disease is reached varies considerably between patients and intra-familial variability has been reported, with affected siblings demonstrating markedly different disease severities (Bergmann et al., 2005). This clinical uncertainty causes significant distress for patients and their families and presents challenges for healthcare institutions in identifying and prioritising the most at-risk individuals.

Mutations in the *Polycystic Kidney and Hepatic Disease 1* (*PKHD1*) gene are the primary cause of ARPKD, accounting for greater than 90% of cases (Bergmann et al., 2003; Bergmann et al., 2004; Bergmann et al., 2005; Adeva et al., 2006; Gunay-Aygun et al., 2010; Melchionda et al., 2016; Lu et al., 2017; Burgmaier et al., 2021; Richards et al., 2024). Fibrocystin (FPC) represents the longest open reading frame of *PKHD1*, spanning exons 1–67 and encoding a large protein of 4,074 amino acids (AAs) (Onuchic et al., 2002; Ward et al., 2002).

Structural characterisation of FPC using conventional experimental techniques has proven challenging, likely due to the protein's considerable size. Based on homology modelling, FPC possesses a large extracellular region, a single transmembrane domain and a short intracellular tail and is conventionally represented as a long, linear protein (Supplementary S1) (Ward et al., 2002; Bergmann et al., 2018; Bannell and Cockburn, 2024). The extracellular region is rich in IPT/TIG-like domains, complemented by a single PA14, multiple PBH1 and two G8 domains (Ward et al., 2002; He et al., 2006; Bannell and Cockburn, 2024). The intracellular tail harbours nuclear and primary cilium localisation sequences and interaction motifs for Polycystin-2 and CAML (Supplementary S1) (Nagano et al., 2005; Hiesberger et al., 2006; Kim et al., 2008; Follit et al., 2010). On the basis of this structure, FPC is proposed to function as a receptor protein at the primary cilium (Ward et al., 2002; Ward et al., 2003). To date, however, no ligand for FPC has been identified.

ARPKD arises from biallelic *PKHD1* variants distributed across the 67 exons encoding FPC. Missense variants represent the most frequently reported variant type; however, with the exception of a small number of recurrent variants, the majority are family-specific or of low prevalence in the broader ARPKD population (Bergmann et al., 2005). Additionally, many ARPKD patients exhibit compound heterozygosity (Bergmann et al., 2005). Given the hundreds of reported pathogenic variants, establishing robust genotype-phenotype correlations remains a considerable challenge.

The presence of two chain-terminating variants has previously been associated with the most severe, early clinical outcomes, while the presence of at least one non-chain-terminating variant has been suggested to confer a reduction in disease severity (Bergmann et al., 2003; Bergmann et al., 2005). Nevertheless, severe clinical outcomes can occur in individuals carrying two missense variants and this classification does not adequately account for the observed variability in kidney and liver disease progression (Bergmann et al., 2003; Bergmann et al., 2004; Bergmann et al., 2005). Variant location has also been proposed as a modifier of organ-specific outcomes; however, this only covers some compound heterozygous variants (Furu et al., 2003; Qiu et al., 2020; Burgmaier et al., 2021).

To date, all genotype-phenotype analyses have been conducted with reference to the linear homology model of FPC, which may not accurately reflect the true three-dimensional conformation of the protein. This limitation arises primarily from the absence of complete three-dimensional structural data for FPC. To address this gap, AlphaFold2 was employed to generate a three-dimensional model of FPC. The highest confidence predicted structure was subsequently used to identify new and verify existing domain structures in FPC and to examine the spatial relationships between compound heterozygous variant pairs within the context of a three-dimensional structural model.

2 Methods

2.1 Data access

An assessment of previously modelled FPC structures was conducted using the RCSB Protein Data Bank (RCSB PDB; RCSB.org), the Protein Data Bank in Europe (PDBe; ebi. ac.uk/pdbe), SWISS-Model (swissmodel.expasy.org) and the AlphaFold Protein Structure Database (AlphaFold PSD; alphafold. ebi.ac.uk) (Berman et al., 2000; Waterhouse et al., 2018; Armstrong et al., 2020; Jumper et al., 2021; Fleming et al., 2025). The following search terms were applied across all databases: Fibrocystin, Polyductin, *PKHD1* and P08F94.

The amino acid sequence for *Homo sapiens* FPC (GRCh38, Fibrocystin Isoform X1) used as input for AlphaFold2 modelling and was retrieved from the National Center for Biotechnology Information (NCBI) (Sayers et al., 2024).

Variant data was obtained for 286 patients with a confirmed genetic diagnosis and documented clinical presentation, sourced from previously published cohort studies including (Adeva et at., 2006 and Bergmann et al., 2003; Bergmann et al., 2004; Bergmann et al., 2005).

2.2 AlphaFold2 structure prediction

Three-dimensional structural modelling of FPC was performed using the AlphaFold2 implementation (v2.3.2+Galaxy0) available through the Galaxy Bioinformatics platform (Jumper et al., 2021; The Galaxy Community, 2024). The following parameters were applied: reduced database, monomer model preset and a model limit of five. The highest confidence model, as determined by predicted Local Distance Difference Test (pLDDT) scoring, was selected for all downstream analyses.

2.3 FoldSeek structural homology search

The PDB file corresponding to the highest confidence AlphaFold2 prediction was submitted to the FoldSeek web server (accessed: 23 April 2026) to identify structural homologues and characterise domains (van Kempen et al., 2024). FoldSeek was run using the following parameters: 3Di/AA mode, iterative search,

TABLE 1 A systematic review of experimentally determined and three-dimensionally predicted structures of Fibrocystin across the RCSB PDB, PDBe, SWISS-Model and AlphaFold PSD databases. Search terms included the *PKHD1* gene name as well as alternative protein names (Fibrocystin/ Polyductin) and Uniprot Human ID for Fibrocystin (P08F94). Initial terms represent total count from searching the above terms. Filtered are Fibrocystin specific predictions that are either experimentally determined or predicted by homology or AlphaFold.

Terms		RCSB PDB	PDBe	Swiss-model	AlphaFold PSD
Initial	PKHD1	0	0	803	34
	Fibrocystin	0	0	1,369	33
	Polyductin	0	0	327	0
	P08F94	0	0	1	0
Filtered	PKHD1	0	0	1	19
	Fibrocystin	0	0	0	32
	Polyductin	0	0	0	0
	P08F94	0	0	1	0

with a taxonomic filter restricted to Mammalia. The search was conducted against the following databases: AFDB-Proteome, AFDB-SwissProt, AFDB50 and PDB100.

2.4 Protein visualisation and annotation

AlphaFold2 models of FPC were visualised and structurally annotated using UCSF ChimeraX (v1.11.1) (Meng et al., 2023). Residue distance measurements were calculated in ChimeraX by measuring the distance between alpha carbons for target residues.

2.5 Statistical and data analysis

All data processing and visualisation was performed in R (v4.5.3) within the RStudio integrated development environment (v2026.01.2 + 418), utilising the tidyverse package (v2.0.0).

3 Results

3.1 Database survey of fibrocystin structural data reveals an absence of full-length FPC structures

Four publicly available databases were systematically searched to characterise the current structural landscape of FPC. The RCSB PDB and PDBe were queried for experimentally determined structures, SWISS-Model was queried for homology models and AlphaFold PSD searched for available computational models.

Neither the RCSB PDB nor the PDBe returned significant results for any of the search terms applied (Table 1). The SWISS-Model repository returned a substantial number of results for *PKHD1* (803), Fibrocystin (1,369) and Polyductin (327); however, these results collectively corresponded to a single homology model, identified under both *PKHD1* and P08F94 search terms, with no models detected under Fibrocystin or Polyductin. This model consists of the region at approximately AAs 2,700–4,074 and contains the structures of 1 G8, 3 PBH1 and a transmembrane domain.

Multiple AlphaFold2 models have been previously generated for FPC. Searching for *PKHD1* returned 34 models, 19 of which were attributed to FPC specifically (Table 1; Supplementary S2). A search for FPC returned 33 models, 32 of which were directly linked to FPC (Table 1; Supplementary S2). The available AlphaFold2 models were predominantly derived from non-human orthologues. A single human model was listed (AF-L0R5C3-F1-v6), spanning AAs 1–43 and corresponding to the previously annotated signal peptide domain of FPC (Supplementary S2).

Given their established use as model organisms in ARPKD research, the structural models of FPC in *Mus musculus* and *Rattus norvegicus* was also examined. No AlphaFold2 models were available for *R. norvegicus*. Two models have been previously generated for *M. musculus* (Table 1): a 151 AA TIG-like domain segment predicted with high confidence (AF-Q8K3J6-F1-v6; average pLDDT: 94.94) and an 187 AA segment encompassing the signal peptide domain (AF-Q3UNU5-F1-v6; average pLDDT: 69.31).

Taken together, these findings demonstrate that experimental structural data for FPC remain severely limited, with neither the RCSB PDB nor the PDBe returning structures for the protein. While AlphaFold2 models exist, these are restricted to segments of the protein and are predominantly derived from non-human orthologues. Crucially, no complete structural model of human FPC was available.

3.2 AlphaFold2 structural modelling of fibrocystin

To address the lack of a complete, full-length structure of FPC, AlphaFold2 was used to produce a *de Novo* computational model of the complete human structure of FPC. AlphaFold2 produced 5 models for FPC, with varying average pLDDT scores. The five models were ranked 1–5 based on average pLDDT, with the lowest pLDDT ranked 5 and the highest ranked 1 (Figure 1, Supplementary S3). Models 2 to 5 had an average pLDDT less than 60, with model 5 averaging 49.87 and model 2 averaging 58.26. Model 1 had an average pLDDT of 61.65.

All five models poorly modelled (pLDDT < 50) the initial sequence AAs of FPC (AAs 1–23) and AAs 3,851–4,074

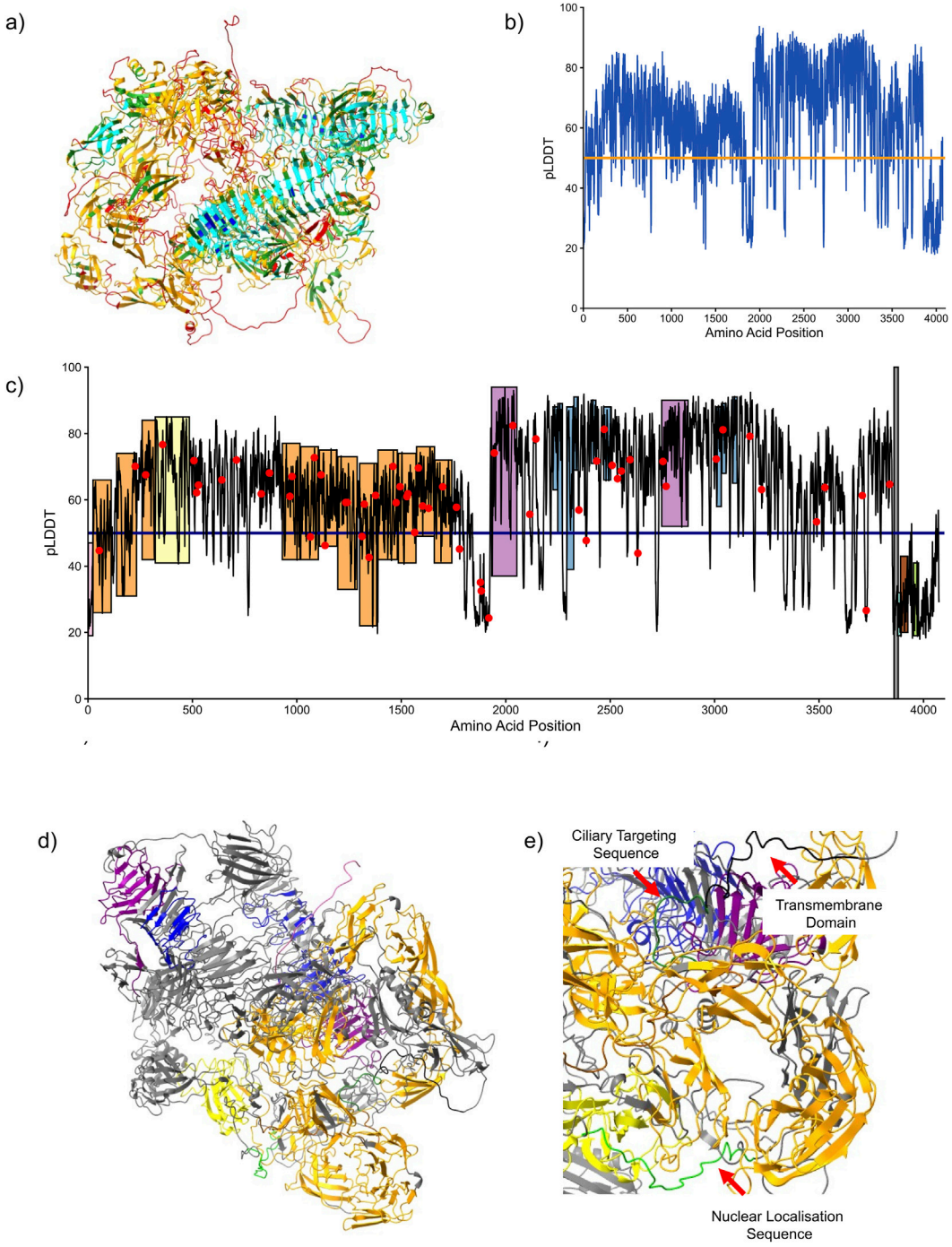

FIGURE 1
To gain a better understanding of the structure of FPC, AlphaFold2 was utilised to model the three-dimensional structure. **(a)** The best ranked three-dimensional structure of FPC determined by AlphaFold2. Colours represent pLDDT values. Very low (<50, Red), Low (50–70, Orange), Mid (70–80, Green), High (80–90, Cyan) and Very High (90–100, Blue). **(b)** A graphical representation of the pLDDT scores for the best ranked AlphaFold2 model of FPC, plotting Amino Acid position against pLDDT. **(c)** A graphical representation of Amino Acid Position and pLDDT annotated with homology predicted structures and glycosylation sites (Red points). Annotations include the signal region (Pink), IPT/TIG domains (Orange), PA14 domain (Yellow), G8 domains (Purple), PBH1 domains (Blue), the Transmembrane Region (Black), Ciliary targeting Sequence (Dark Green), Nuclear Localisation Sequence (Light Green) and Disordered regions (Brown). **(d,e)** An annotated three-dimensional model of FPC with previously identified domains based on homology. **(d)** Represents the full FPC structure. **(e)** Is a zoomed in model annotated with regions and domains within the cytoplasmic tail of FPC. Colour annotations represent the signal region (Pink), IPT/TIG domains (Orange), PA14 domain (Yellow), G8 domains (Purple), PBH1 domains (Blue), the Transmembrane Region (Black), Ciliary targeting Sequence (Dark Green), Nuclear Localisation Sequence (Light Green) and Disordered regions (Brown).

(Figures 1a, b; Supplementary S2a-h). The region between AAs 2,000–3,000 was the most confidently predicted region in all models, often seeing the highest per-residue pLDDT scores (Figures 1a, b; Supplementary S2a-h). A second smaller, distinct region was also observable, preceding the high peak between AAs 2,000–3,000. The smaller peak, between AAs 250–1,800, varied in size between the five predicted models and was poorly predicted in models 4 and 5 (Figures 1a, b; Supplementary S3a-h). Model 1 had the highest average pLDDT in this region (61.9), supporting a low–moderate predictive strength.

Compared to the homology model of FPC, the regions with highest predictive confidence were the G8 and PBH1 domains in model 1, with an average pLDDT of 76 and 78, respectively (Figure 1c). The IPT/TIG domains had an average pLDDT of 59 and suffered from low pLDDT confidence (Figure 1c). The best predicted IPT/TIG domain was 3, with an average 69. TIG domains 4, 6, 10, 11, 12 all featured an average pLDDT ≥60. The lowest IPT/TIG domains were 1 and 8, at 50 and 52, respectively. The PA14 domain was predicted with moderate confidence, with an average pLDDT of 67. The signal domain (AAs 1–23), the Transmembrane domain (AAs 3,859–3,879) and the entirety of the intracellular tail (AAs 3,880–4,074) were all predicted with very low confidence (pLDDT < 50, Figure 1c). This included critical functional components such as the Ciliary targeting Sequence (AAs 3,880–3,893) and Nuclear Localisation Sequence (AAs 3,952–3,976; Figures 1c, d).

FPC is typically illustrated as a long, linear receptor protein with a long, straight extracellular component and short, intracellular component (Supplementary S1). The five predicted models did not predict a structure representative of this model (Figures 1a, e; Supplementary S2a, c, e, f). Instead, the AlphaFold2 predicted models of FPC appeared to highlight a spiral/ curve around the TIG domains, creating an internally hollow structure with curved IPT/TIG domains and 2 parallel sets of G8 and PBH1 domains (Figure 1e). The curved nature of the IPT/TIG domains was present in all FPC models but most evident in models 1, 2 and 3. Poor predictive strength for the intracellular tail resulted in AlphaFold2 being unable to predict the structure of the region, leaving its relationship to the greater extracellular region unknown (Figure 1e).

The extracellular region of FPC is predicted to be heavily modified, containing 64 asparagine glycosylation sites; however, the positioning of these residues, in context to the three-dimensional structure of FPC, remains largely unknown. In the "best" FPC model (Model 1), the asparagine residues were located towards the external surface of their respective structures, this includes both the exterior surface and the exposed interior of the curvature (Supplementary S4). pLDDT values for these residues ranged from 19 to 93, with a median of 65, suggesting that some residues lie within poorly predicted or disordered regions of FPC.

Taken together, the current AlphaFold2 predicted models of FPC show that the protein forms a spiral-like arrangement, with a hollow internal compartment surrounded by a curving IPT/TIG domains and supported by 2 parallel structures consisting of a single G8 and multiple PBH1 domains. As Predicted Aligned Error (PAE) information was unavailable, we could not assess the confidence in the relative positions of the domains as they relate to each. However, given the low global confidence it is likely the observed models have poor positional relationships. This is highly evident round the IPT/TIG region, the region with the most evident curved domain but with an average confidence score of 59. The regions with the highest predictive confidence were G8 and PBH1 domains with an average pLDDT > 75. Poor predictive confidence resulted in poor modelling of the intracellular region, which contains critical localisation sequences for FPC's primary functions.

3.3 FoldSeek structural homology analysis of fibrocystin demonstrated concordance with previous homology models

To identify known structural domains within FPC, the top ranked AlphaFold2 model was submitted to the FoldSeek web server to query the predicted structure against publicly available protein structure databases. To narrow down the search criteria the databases compared to were the AFDB-Proteome, AFDB-Swissprot, AFDB50 and the PDB100 focusing on models for mammalian proteins with an E-value < 0.001. In total, 830 models were identified, with 206 being significant matches (Figure 2a).

Annotating the 206 hits revealed a common trend in many of the detections, allowing the hits to be grouped into 9 categories (Supplementary S5). Hits were either from *PKHD1* or *PKHD1L1* models from different species and accounted for the most hits (42). Some hits were proteins containing domains known to be in FPC, including G8 domains (19 hits), IPT/TIG domains (28 hits), PA14 domains (17 hits) and PBH1 domains (14 hits). Many of these similar hits were associated with TMEM2 (20 hits) which contain G8 and PBH1 domains. Some hits were linked to proteins with Immunoglobulin-Like (IG-Like) and Fibronectin Type III domains (32 hits) or were otherwise Glycosylation site heavy, similar to FPC (14 hits). 5 hits had no discernible features that could be ruled as similar to already known domains in FPC, these included 2 right-handed beta helix domains, a UBR-type domain and similarity to receptor protein-tyrosine kinases (Supplementary S5). 15 of the 206 hits were from uncharacterised proteins and were removed from the analysis.

Many of the significant hits (E-value < 0.001) with the highest fitted scores (>1,000) came from existing *PKHD1* models, G8 domain, PA14 domain, TMEM2 and glycosylation site heavy proteins (Figure 2b). Proteins linked to IPT/TIG, PBH1 and IG-Like or Fibronectin Type III domains had lower fitting scores on average (<1,000) (Figure 2b).

An assessment of the type of hits by position revealed a mixture of fragment overlaps for the FPC and FPC-Like protein models, which overlaid with known homology predicted domains throughout the FPC model (Figure 2c). Additionally, predicted structural similarity with known domains was evident, including the PA14, G8 and PBH1 domains (Figure 2c). Glycosylation site heavy proteins overlaid with the G8-PBH1 region of FPC which is known to be Glycosylation site heavy (Figure 2c; Supplementary S1). Unlike the homology model, which observes IPT/TIG domains as early as AA 24–335 (IPT/TIG 1–3), structural overlaps were present only for TIG domains 4–12, this includes IG-Like domains (Figure 2c).

Taken Together, the top ranked AlphaFold2 predicted model of FPC compared to known AlphaFold protein models showed considerable consensus with the already predicted homology model of FPC. This included similar structural placement for G8, PBH1, PA14 and IPT/TIG domains. However, no new structural domains

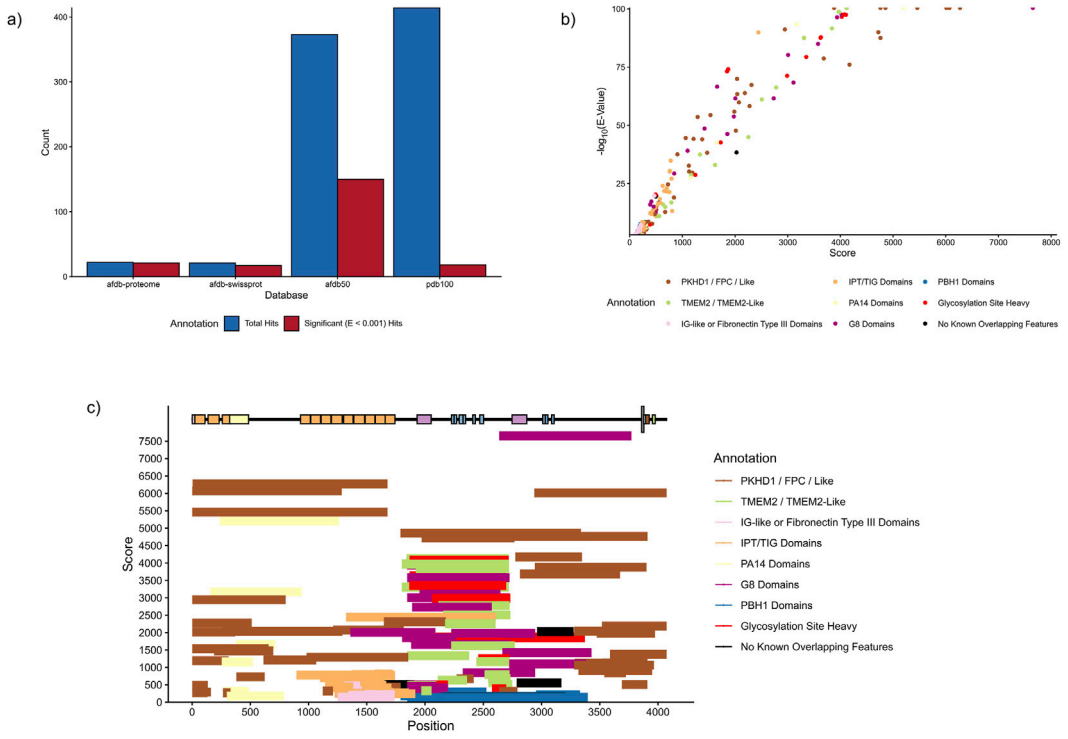

FIGURE 2
To identify the presence of domains based on the three-dimensional structure of FPC, the FoldSeek Web server was queried using the AFDB-Proteome, AFDB-Swissprot, AFDB50 and PDB100 database sets and limited to searches against mammalian proteins. **(a)** A bar chart representing the number of total hits (Blue) and hits with an E-value < 0.001 (Red) per database set. **(b)** A scatter plot showing the distribution of FoldSeek hits comparing Score to E-Value. Coloured annotations represent the type of structures within matches (e.g., IPT domains). **(c)** A bar chart showing the position of a hit and its Score, against the relative location of known FPC domains. A schematic of FPC is annotated across the top of the chart. Coloured Segments of FPC represent the signal region (Pink), IPT/TIG domains (Orange), PA14 domain (Yellow), G8 domains (Purple), PBH1 domains (Blue), the Transmembrane Region (Dark Grey), Ciliary targeting Sequence (Dark Green), Nuclear Localisation Sequence (Light Green) and Disordered regions (Brown).

were discovered as a result of these comparisons and comparisons failed to highlight structural overlaps between IPT/TIG domains 1–3 and other IPT/TIG containing proteins that are not FPC or FPC-Like 1.

3.4 Structural mapping of pathogenic fibrocystin variant pairs did not identify genotype-phenotype relationships

ARPKD is a clinically, variable disease ranging from a requirement for critical care shortly after birth, to severe kidney or liver disease during childhood, adolescence and early adulthood. To explore the clinical relevance of the AlphaFold2-predicted FPC model, known pathogenic variant pairs identified from the literature were mapped onto the AlphaFold2 structure to assess their spatial relationships.

Patient data was taken from Adeva et al., 2006 and Bergmann et al., 2003, 2004 and 2005. Across these 4 studies, 8 Adult (20+), 7 Adolescent (13–19 years), 45 Children (1–13 years) and 16 Severe (<1 Year Critical Care) cases were reported with 2 missense variants per person (Figure 3a). This list was further refined to isolate paediatric cases that did not reach adulthood prior to reaching end-stage kidney disease or prematurely dying, were not homozygous for the same missense variant and variants were not close to each other in the linear AA sequence (>100 AA). This criteria isolated 4 Adult, 5 Child and 6 Severe cases (Figure 3a; Supplementary S6).

Spatial closeness was determined as domain structures that were close to each other in the three-dimensional AlphaFold2 model but were otherwise distant in previous homology models (Supplementary S1). A comparison of the spatial relationship between variants pairs did not reveal a common trend between how close two variants were and the severity of ARPKD. Patients in the severe group featured variant pairs that were both close and distant in the three-dimensional structure (Figures 3b-d). This trend was also observable in the childhood end-stage kidney group (Figures 3e-g) and adult group (Figures 3h,i). Additionally, the short-listed cohort showed evidence of intra-familial variability, were two family members had the same compound heterozygous variants but different disease outcomes, further suggesting poor correlation between the variants three-dimensional position and disease outcomes (Supplementary S6).

To quantitively determine whether their was a significant relationship between disease outcome and the position of the variant pairs in three-dimensional space a one-way ANOVA was performed on the distance between the alpha carbons of impacted residues (Supplementary S6). No significant relationship was identified with $P > 0.691$ (sums of squares = 482, F-statistic = 0.381 and 2 degrees of freedom).

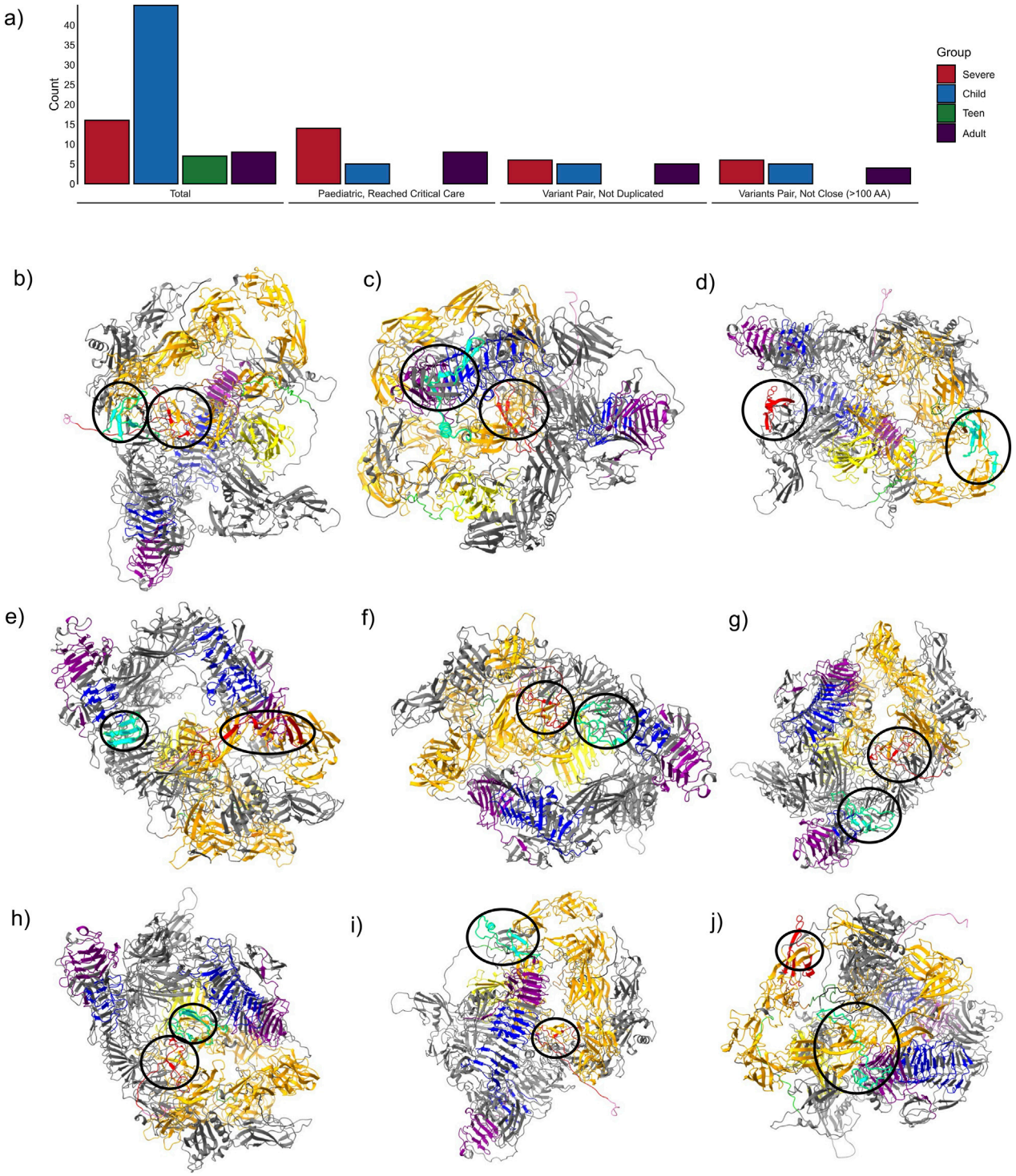

FIGURE 3
To assess whether a relationship between an individuals *PKHD1* variants and their relative position to each other in three-dimensional space was present missense variants were mapped to the highest confidence AlphaFold2 model of FPC. **(a)** A breakdown of the number of individuals with two missense variants in (Adeva et at., (2006) and (Bergmann et al. 2003, 2004, Bergmann et al.,2005) after filtering for suitable matches. Group represents approximate age bracket at the patients final recorded examination. Severe (<1 year), Child (<10 years), Teen (<20 years) and Adult (20+ years). **(b-j)** Annotated prediction models of FPC, mapped with recessive variant pairs. **(b-d)** Severe ARPKD group. **(e-g)** Child ARPKD group. **(h-j)** Adult ARPKD group. Variants are colour coded Red for the earliest variant in the Amino Acid chain and Cyan for the second variant. Coloured variants represent an Amino Acid region 50 Amino Acids surrounding the missense variant (±25 AA) to improve visibility. Additionally, variants are circled with a black circle. Additional Annotations include the signal region (Pink), IPT/TIG domains (Orange), PA14 domain (Yellow), G8 domains (Purple), PBH1 domains (Blue), the Transmembrane Region (Black), Ciliary targeting Sequence (Dark Green), Nuclear Localisation Sequence (Light Green) and Disordered regions (Brown).

4 Discussion

4.1 AlphaFold2 reveals a curved IPT/TIG structure supported by parallel G8-PBH1 domains

FPC is a large, membrane-bound protein with a considerably long extracellular domain that spans the majority of the proteins length. The exact mechanisms by which the protein functions is unknown; however, it is thought to play the role of a receptor protein, within the limited cell types the protein is expressed (Onuchic et al., 2002; Ward et al., 2002; Ward et al., 2003). No experimentally determined structural models exist for FPC, despite the proteins' clinical relevance to ARPKD. Additionally, no experimentally determined models exist for the structurally similar FPC-like 1, for which to base an FPC model on. To date, no cryo-electron microscopy, X-ray crystallography, or NMR structures have been solved for either the full-length proteins or their isolated ectodomain fragments; the oligomeric state of FPC likewise remains undetermined. Given the significant lack of structural data for FPC and the potential gains from identifying the proteins three-dimensional structure, e.g., residue pair analysis, we used AlphaFold2 to predict the three-dimensional structure of FPC.

A systematic review of publicly available structural databases confirmed the absence of full-length human FPC models and revealed only a limited number of partial structures. The AlphaFold2 predicted model in this study represents a complete, full-length three-dimensional structural predication of human FPC. The average per-residue local confidence of the top ranked model in this study was low, with variable per-residue confidence scores throughout the protein. Moderate-to-high confidence was present throughout the region attributed to G8 and PBH1 domains, suggesting these regions are well modelled. In contrast, the IPT/TIG domains generally showed low per-residue confidence, with specific TIG domains showing low-to-moderate confidence. The transmembrane region and intracellular tail were predicted with very low confidence; however, AlphaFold2 is known to struggle with the residues within transmembrane domains and the low confidence of the intracellular tail could be attributed to its disordered nature (Bertoline et al., 2023; Bannell and Cockburn, 2024; Li et al., 2026).

The structural domains identified within the predicted model was broadly consistent with the previously described linear homology model of FPC, with no novel structural domains identified (Bannell and Cockburn, 2024). This overlap between domain predictions in this three-dimensional model and linear homology model highlights the models predictive strength at representing FPC. Interestingly, IPT/TIG domains 1–3 were not specifically detected by FoldSeek as IPT/TIG domains or as proteins containing such domains. This could be due to the low per-residue confidence in this region; however, IPT/TIG domain 3 has the highest average confidence of all the IPT/TIG domains. Despite the lack of specific detections, domains were predicted within this region, resulting from matches to other AlphaFold2 FPC/FPC-like 1 models. This could represent general uncertainty in the assigned IPT/TIG domain or a more unique variant of such a structure to FPC.

FPC is typically illustrated as a long linear, membrane-bound protein with a long extracellular and short intracellular regions (Bergmann et al., 2018; Bannell and Cockburn, 2024). The predicted AlphaFold2 model does not share this same trait. Instead, the IPT/TIG domains are arranged in a spiral-like loop within the predicted structure. Additionally, the G8-PBH1 pairs run parallel to each other, as opposed to linearly leading to the transmembrane domain. Although, this curving nature of the FPC protein could be a result of low confidence in this region, it has been suggested that FPC IPT/TIG pattern could result in a curved ectodomain due to similar curves in other proteins rich in IPT/TIG domains, like plexins (Bannell and Cockburn, 2024). Although the predicted model provided by AlphaFold2 highlights some potential structural insights into FPC, the low global confidence of the "best" model means additional experimental modelling will be required to validate the observed AlphaFold2 structures.

There are several limitations that modelling a protein like FPC will have in AlphaFold2. The first, AlphaFold2 can model the whole protein structure; however, FPC undergoes considerable post-translational modification that could change the proteins conformation, of which, AlphaFold2 can not predict (Bertoline et al., 2023; Li et al., 2026). In FPC these post-translational modifications are throughout the extracellular region and located in IPT/TIG and G8 domains and regions of the protein with no marked domain structure and likely play a major role in the final confirmation of the protein. Additionally, FPC is a membrane-bound protein with a disordered intracellular tail, two additional factors AlphaFold2 may struggle with (Bertoline et al., 2023; Li et al., 2026). In the former, the training data for membrane-bound proteins is low and can not account for the influence of lipids and other proteins-bound to the membrane (Bertoline et al., 2023; Li et al., 2026). For the latter, the low pLDDT confidence attributed to disordered regions can result in poorly placed domains (Li et al., 2026). In the context of FPC this results in a tail that loops back towards to the main protein body with no understanding of its function. Finally, a high percentage of residues with low pLDDT scores, like FPC, can result in a low accuracy assignment of the protein backbone structure, reducing confidence in domain assignment (Li et al., 2026); however, confidence in the positioning of the domains relative to each other in the model could not be assessed since PAE information was unavailable. Despite this, the FPC structure predicted in this study by AlphaFold2 is unlikely to be representative of the final functional protein.

4.2 Three-dimensional spatial clustering of missense variants in fibrocystin do not influence disease outcomes

A major outstanding issue with ARPKD is understanding why some patients have worse outcomes than others. Generally speaking, patients with two chain-terminating variants in *PKHD1* show the worse disease outcomes, while the presence of a single missense variant can reduce disease severity and improve survival beyond the perinatal period (Bergmann et al., 2003; 2005). However, this only explains the presentation of a small subset of patients. Many ARPKD patients can have severe outcomes and are compound heterozygous for *PKHD1* missense variants (Bergmann et al., 2003; Bergmann et al., 2005). Additionally, kidney and liver disease severity can differ significantly between patients, even siblings (Bergmann et al., 2003; Bergmann et al., 2004; Bergmann et al., 2005;

Guay-Woodford et al., 2014). What is currently uncertain is whether the position of these variants influence the presentation and progression of ARPKD. Prior attempts have been made to establish this and have drawn some correlations (Furu et al., 2003; Qiu et al., 2020; Burgmaier et al., 2021). A recent example suggests the that the presence of two missense variants localised between AAs 709–1,837 result in slower kidney disease progression, while variants between AAs 1,838–2,624 have been suggested to be protective against significant liver complications (Burgmaier et al., 2021).

A possible explanation for the variation in disease outcomes could be that the missense variants sit close to each other within three-dimensional space, such that the region of these variants impact similar folding and function; effectively resulting in both copies of FPC having lost function in the same or similar regions. In a small selection of individuals from previously published literature this was not consistently the case. In all three classification groups (Severe, Childhood and Adulthood ARPKD) individuals with variants spatially close were present. This was also true for variant combinations that were spatially distant, suggesting that the relative closeness of these variants in three-dimensional space does not confer any further negative effects than what has previously been identified with homology models.

To study the relationships between variant pairs in three-dimensional space, this study focused on variants situated within a large distance from each other, which was defined as 100 AAs. This was performed to ensure that the assessed physical position of the two variants was due to the structure of FPC and not their relative closeness in the amino acid chain. A limitation of our variant-pair isolation criteria is that it leaves a gap in knowledge regarding variants within 100 AAs. Although AlphaFold has predicted some structural information regarding FPC, these close amino acids could still play significant roles in the formation of interaction sites, the loss of which could influence disease outcomes. We do not currently know enough about the oligomeric state of FPC. As these tools advance and more information comes to light, additional studies can shed further light on such interaction sites and their relationship to disease outcomes.

Many publications have been released over the years regarding disease outcomes in ARPKD; however, these studies often don't report the same information, making it difficult to group patients across studies regarding disease outcomes. Due to this, in this study the focus was given to a small selection of well established papers within ARPKD genotype-phenotype research. After filtering patient data, only a small selection of patients derived from the four ARPKD publications were suitable for this analysis, such that they were able to be boxed into groups of varying disease presentation with significantly negative outcomes (i.e., end-stage kidney disease). Despite this, data regarding the liver presentation of ARPKD was sparse, so the relationship could not be assessed in this study due to limited patient information. The small number of derived records will have greatly reduced the statistical power of this analysis, potentially hiding genotype-phenotype relationships between the spatial position of the variant and disease outcomes. Future studies will need to look at the spatial relationships on a larger focused population, with the aim of acquiring significant phenotype data over time to help further classify ARPKD patients and reevaluate spatial relationships.

An additional factor that could have negatively impacted the association between genotype and phenotype in this study is the poor pLDDT scores for the residues of many of the variants used. On average, pLDDT for these residues was less than 70, suggesting low confidence placements. Uncertainty in residue placement could suggest that some of these variant pairs are closer or further away in three-dimensional space. As modelling parameters improve and additional experimental validation becomes accessible for FPC, greater confidence in these spatial relationships will become apparent, improving our ability to detect the relationship between variants spatially and their impact on ARPKD disease outcome.

Finally, only the position of the variants in relation to each other was assessed, not the impact of these variants on protein conformation. AlphaFold2 has previously been shown to poorly model missense variants (Buel and Walters, 2022; Bertoline et al., 2023; Pak et al., 2023). As this field advances, it may become possible to assess the specific damage of these variants in FPC and whether they confer similar functional impact, without being spatially close.

5 Conclusion

AlphaFold2 was able to predict the full-length of Fibrocystin with varying degrees of confidence. AlphaFold2 was most confident in the regions surrounding the G8 and PBH1 domains, with low-moderate accuracy in the IPT/TIG domains and very low accuracy in the signal and intracellular tail regions. Assessing spatial relationships between compound heterozygous variant pairs and disease outcomes does not suggest that spatially close missense variants have worsening disease outcomes compared to those further away.

Data availability statement

The data presented in the study are deposited in the Zenodo repository, DOI: 10.5281/zenodo.23146308

Author contributions

TR: Data curation, Formal Analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing. PG: Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review and editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. TR Received funding support from the University of Wolverhampton through the Early Researcher Award Scheme.

Acknowledgments

We are grateful for the tools and databases used in this project which are made freely available to the wider scientific community by the many talented teams who create them.We would also like to express gratitude to the University of Wolverhampton for funding the research project as part of an early researcher award scheme.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2026.1877678/full#supplementary-material

References

Adeva, M., El-Youssef, M., Rossetti, S., Kamath, P. S., Kubly, V., Consugar, M. B., et al. (2006). Clinical and molecular characterization defines a broadened spectrum of autosomal recessive polycystic kidney disease (ARPKD). *Med. Baltim.* 85, 1–21. doi: 10.1097/01.md.0000200165.90373.9a

Armstrong, D. R., Berrisford, J. M., Conroy, M. J., Gutmanas, A., Anyango, S., Choudhary, P., et al. (2020). PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.* 48, D335–D343. doi: 10.1093/nar/gkz990

Bannell, T. A. K., and Cockburn, J. J. B. (2024). The molecular structure and function of fibrocystin, the key gene product implicated in autosomal recessive polycystic kidney disease (ARPKD). *Ann. Hum. Genet.* 88, 58–75. doi: 10.1111/ahg.12535

Bergmann, C., Senderek, J., Sedlacek, B., Pegiazoglou, I., Puglia, P., Eggermann, T., et al. (2003). Spectrum of mutations in the gene for autosomal recessive polycystic kidney disease (ARPKD/PKHD1). *J. Am. Soc. Nephrol. JASN* 14, 76–89. doi: 10.1097/01.asn.0000039578.55705.6e

Bergmann, C., Senderek, J., Schneider, F., Dornia, C., Küpper, F., Eggermann, T., et al. (2004). PKHD1 mutations in families requesting prenatal diagnosis for autosomal recessive polycystic kidney disease (ARPKD). *Hum. Mutat.* 23, 487–495. doi: 10.1002/humu.20019

Bergmann, C., Senderek, J., Windelen, E., Küpper, F., Middeldorf, I., Schneider, F., et al. (2005). Clinical consequences of PKHD1 mutations in 164 patients with autosomal-recessive polycystic kidney disease (ARPKD). *Kidney Int.* 67, 829–848. doi: 10.1111/j.1523-1755.2005.00148.x

Bergmann, C., Guay-Woodford, L. M., Harris, P. C., Horie, S., Peters, D. J. M., and Torres, V. E. (2018). Polycystic kidney disease. *Nat. Rev. Dis. Primer* 4, 50. doi: 10.1038/s41572-018-0047-y

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Bertoline, L. M. F., Lima, A. N., Krieger, J. E., and Teixeira, S. K. (2023). Before and after AlphaFold2: an overview of protein structure prediction. *Front. Bioinforma.* 3, 1120370. doi: 10.3389/fbinf.2023.1120370

Buel, G. R., and Walters, K. J. (2022). Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* 29, 1–2. doi: 10.1038/s41594-021-00714-2

Burgmaier, K., Brinker, L., Erger, F., Beck, B. B., Benz, M. R., Bergmann, C., et al. (2021). Refining genotype-phenotype correlations in 304 patients with autosomal recessive polycystic kidney disease and PKHD1 gene variants. *Kidney Int.* 100, 650–659. doi: 10.1016/j.kint.2021.04.019

Fleming, J., Magana, P., Nair, S., Tsenkov, M., Bertoni, D., Pidruchna, I., et al. (2025). AlphaFold protein structure database and 3D-Beacons: new data and capabilities. *J. Mol. Biol.* 437, 168967. doi: 10.1016/j.jmb.2025.168967

Follit, J. A., Li, L., Vucica, Y., and Pazour, G. J. (2010). The cytoplasmic tail of fibrocystin contains a ciliary targeting sequence. *J. Cell Biol.* 188, 21–28. doi: 10.1083/jcb.200910096

Furu, L., Onuchic, L. F., Gharavi, A., Hou, X., Esquivel, E. L., Nagasawa, Y., et al. (2003). Milder presentation of recessive polycystic kidney disease requires presence of amino acid substitution mutations. *J. Am. Soc. Nephrol. JASN* 14, 2004–2014. doi: 10.1097/01.asn.0000078805.87038.05

Goggolidou, P., and Richards, T. (2022). The genetics of autosomal recessive polycystic kidney disease (ARPKD). *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* 1868, 166348. doi: 10.1016/j.bbadis.2022.166348

Guay-Woodford, L. M., Bissler, J. J., Braun, M. C., Bockenhauer, D., Cadnapaphornchai, M. A., Dell, K. M., et al. (2014). Consensus expert recommendations for the diagnosis and management of autosomal recessive polycystic kidney disease: report of an international conference. *J. Pediatr.* 165, 611–617. doi: 10.1016/j.jpeds.2014.06.015

Gunay-Aygun, M., Tuchman, M., Font-Montgomery, E., Lukose, L., Edwards, H., Garcia, A., et al. (2010). PKHD1 sequence variations in 78 children and adults with autosomal recessive polycystic kidney disease and congenital hepatic fibrosis. *Mol. Genet. Metab.* 99, 160–173. doi: 10.1016/j.ymgme.2009.10.010

Gunay–Aygun, M., Font–Montgomery, E., Lukose, L., Gerstein, M. T., Piwnica–Worms, K., Choyke, P., et al. (2013). Characteristics of congenital hepatic fibrosis in a large cohort of patients with autosomal recessive polycystic kidney disease. *Gastroenterology* 144, 112–121.e2. doi: 10.1053/j.gastro.2012.09.056

He, Q., Liu, X., Li, Q., Studholme, D. J., Li, X., and Liang, S. (2006). G8: a novel domain associated with polycystic kidney disease and non-syndromic hearing loss. *Bioinformatics* 22, 2189–2191. doi: 10.1093/bioinformatics/btl123

Hiesberger, T., Gourley, E., Erickson, A., Koulen, P., Ward, C. J., Masyuk, T. V., et al. (2006). Proteolytic cleavage and nuclear translocation of fibrocystin is regulated by intracellular Ca2+ and activation of protein kinase C. *J. Biol. Chem.* 281, 34357–34364. doi: 10.1074/jbc.M606740200

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-021-03819-2

Kim, I., Li, C., Liang, D., Chen, X.-Z., Coffy, R. J., Ma, J., et al. (2008). Polycystin-2 expression is regulated by a PC2-binding domain in the intracellular portion of fibrocystin. *J. Biol. Chem.* 283, 31559–31566. doi: 10.1074/jbc.M805452200

Li, M. Q. C., Wang, S., Lin, S.-R., Ting, L. E. N., Wan, Z.-H., Xie, G., et al. (2026). Advantages and limitations of AlphaFold in structural biology: insights from recent studies. *Protein J.* 45, 22–38. doi: 10.1007/s10930-025-10310-8

Lu, H., Galeano, M. C. R., Ott, E., Kaeslin, G., Kausalya, P. J., Kramer, C., et al. (2017). Mutations in DZIP1L, which encodes a ciliary transition zone protein, cause autosomal recessive polycystic kidney disease. *Nat. Genet.* 49, 1025–1034. doi: 10.1038/ng.3871

Melchionda, S., Palladino, T., Castellana, S., Giordano, M., Benetti, E., De Bonis, P., et al. (2016). Expanding the mutation spectrum in 130 probands with ARPKD: identification of 62 novel PKHD1 mutations by sanger sequencing and MLPA analysis. *J. Hum. Genet.* 61, 811–821. doi: 10.1038/jhg.2016.58

Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H., et al. (2023). UCSF ChimeraX: tools for structure building and analysis. *Protein Sci.* 32, e4792. doi: 10.1002/pro.4792

Nagano, J., Kitamura, K., Hujer, K. M., Ward, C. J., Bram, R. J., Hopfer, U., et al. (2005). Fibrocystin interacts with CAML, a protein involved in Ca2+ signaling. *Biochem. Biophys. Res. Commun.* 338, 880–889. doi: 10.1016/j.bbrc.2005.10.022

Onuchic, L. F., Furu, L., Nagasawa, Y., Hou, X., Eggermann, T., Ren, Z., et al. (2002). PKHD1, the polycystic kidney and hepatic disease 1 gene, encodes a novel large protein containing multiple immunoglobulin-like plexin-Transcription–Factor domains and parallel beta-helix 1 repeats. *Am. J. Hum. Genet.* 70, 1305–1317. doi: 10.1086/340448

Pak, M. A., Markhieva, K. A., Novikova, M. S., Petrov, D. S., Vorobyev, I. S., Maksimova, E. S., et al. (2023). Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLOS ONE* 18, e0282689. doi: 10.1371/journal.pone.0282689

Qiu, L.-R., Xu, R.-R., Tang, J.-H., and Zhou, J.-H. (2020). Possible PKHD1 hot-spot mutations related to early kidney function failure or hepatofibrosis in Chinese children with ARPKD: a retrospective single center cohort study and literature review. *Curr. Med. Sci.* 40, 835–844. doi: 10.1007/s11596-020-2268-z

Richards, T., Wilson, P., and Goggolidou, P. (2024). Next generation sequencing identifies WNT signalling as a significant pathway in autosomal recessive polycystic kidney disease (ARPKD) manifestation and May be linked to disease severity. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* 1870, 167309. doi: 10.1016/j.bbadis.2024.167309

Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Connor, R., et al. (2024). Database resources of the national center for biotechnology information in 2025. *Nucleic Acids Res.* 53, D20–D29. doi: 10.1093/nar/gkae979

The Galaxy Community (2024). The galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res.* 52, W83–W94. doi: 10.1093/nar/gkae410

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., et al. (2024). Fast and accurate protein structure search with foldseek. *Nat. Biotechnol.* 42, 243–246. doi: 10.1038/s41587-023-01773-0

Ward, C. J., Hogan, M. C., Rossetti, S., Walker, D., Sneddon, T., Wang, X., et al. (2002). The gene mutated in autosomal recessive polycystic kidney disease encodes a large, receptor-like protein. *Nat. Genet.* 30, 259–269. doi: 10.1038/ng833

Ward, C. J., Yuan, D., Masyuk, T. V., Wang, X., Punyashthiti, R., Whelan, S., et al. (2003). Cellular and subcellular localization of the ARPKD protein; fibrocystin is expressed on primary cilia. *Hum. Mol. Genet.* 12, 2703–2710. doi: 10.1093/hmg/ddg274

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. doi: 10.1093/nar/gky427