Using 3D protein models to uncover genetic risk in ALS

Name	Julia Hermens
Student number	6480349
Master Program	Bioinformatics & Biocomplexity
Supervisor	Kevin Kenna
Second examiner	Jan Veldink
Date	17 November 2023

ABSTRACT

Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disease that is characterized by the involvement of rare genetic variants. Disease association of single rare variants can often not be tested. Therefore, variants are usually grouped together to perform burden tests. In numerous proteins, disease associated variants are located at specific regions (hotspots). Hotspots can be identified by clustering variants together to define groups for burden testing. We extend previous ALS hotspot identification study designs, which clustered variants based on distances within linear sequences, to include 3D spatial clustering methods. We aim to determine if these 3D methods can be used to identify known ALS hotspots or whether specific limitations prevent the application of these methods in ALS hotspot detection. In order to examine this, three well known ALS proteins (*SOD1*, *FUS* and *TARDBP*) that each represent a different use case (no hotspot, localised hotspot and elongated hotspot respectively) will be studied with spherical clustering and protein-structure based scan (PSCAN) methods. Our PSCAN results resemble previous findings for the three use cases, while spherical clustering methods are not able to replicate expectations for the elongated hotspot use case at all. PSCAN is thus an improvement over spherical clustering, for our intended application, as no predefined window sizes or shapes are used. Still, PSCAN has notable limitations. The most important limitation, especially for neurodegenerative disease, lies in the use of AlphaFold2 models which do not adequately represent intrinsically disordered regions. Hotspots which occur in disordered regions can not be identified with 3D methods until this limitation is resolved. Therefore, current 3D spatial clustering methods should only be used for ALS hotspot detection in ordered regions of proteins.

KEYWORDS

Spatial clustering, AlphaFold2, rare variants, association testing, Amyotrophic Lateral Sclerosis

LAYMAN'S SUMMARY

Amyotrophic Lateral Sclerosis (ALS) is a fatal disease in which patients gradually lose control of muscle movements. Discovery of genetic rare variants that cause this disease is difficult. Therefore, studies often try to find groups of rare variants, instead of single variants, that are associated with ALS. In biology, disease causing variants often lie close together. This makes the identification of disease associated regions relevant. These regions can be discovered based on 1D distances between variants within either protein or gene sequences. However, some relevant information might be missed as proteins are folded into 3D structures. Because of this, variants that are far apart linearly can be situated close together. In this study, 1D methods are extended, to include 3D information to identify ALS associated regions. We do this for proteins with known ALS associated regions so that we can verify our findings with existing knowledge. Based on this, we determine whether 3D information should be included to study ALS. Our findings show that 3D clustering methods were able to adequately replicate expected results. However, we also observed that 3D methods are limited by the input structures, as structure prediction tools assume that protein coordinates stay fixed. Since no biological reliable coordinates are available for dynamical regions, 3D clustering can only be used to

study disease association of non dynamical regions. Indeed, until alternative sources for coordinate data have been developed, ALS association of dynamical regions can not be studied in 3D.

1. INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a relatively rare, but fatal neurodegenerative disease. In Europe, 1 out of 350 individuals are diagnosed with ALS in their lifetime (Feldman et al., 2022; Goutman et al., 2022). Patients progressively lose their upper and lower motor neurons (Hardiman et al., 2017), which ultimately leads to death as a result of respiratory failure 2-4 years after the diagnosis (Feldman et al., 2022; Goutman et al., 2022). In most patients, this disease is characterized as sporadic as there is no familial history of ALS (Goutman et al., 2022; Hardiman et al., 2017). In both sporadic and familial cases, genetic factors are thought to be involved (Hardiman et al., 2017; Al-Chalabi et al., 2017). However, current genetic findings explain only 70% of familial and 15% of sporadic cases (Chia et al., 2018). To date, the exact mechanism behind disease manifestation is not yet understood (Hardiman et al., 2017). In the past few years, advances in genomics techniques have shed light on the role of rare variants within numerous diseases (Pierre & Génin, 2014). Indeed, these rare variants, and not the well-studied common variants, are important for ALS disease risk (Al-Chalabi et al., 2016; Chen et al., 2022; van Rheenen et al., 2016). Discovering which rare variants are involved in ALS can improve our understanding of the underlying disease mechanism.

As their name suggests, rare variants occur less frequently compared to common variants. Usually a Minor Allele Frequency (MAF) cutoff value below 1% is used when referring to rare variants (Chen et al., 2022; Pierre & Génin, 2014). Genome-Wide Association Studies (GWAS), which examine disease association of single variants, can only discover (relatively) common variants (Asimit & Zeggini, 2010; Auer & Lettre, 2015; Pierre & Génin, 2014). To identify rare variants with this technique, an incredibly large sample size would be necessary. This is especially difficult to obtain for relatively rare diseases such as ALS (Asimit & Zeggini, 2010; Chen et al., 2022; Lee et al., 2014). An additional problem, the multiple testing burden, arises due to the detection of large amounts of unique rare variant positions during single variant analyses (Chen et al., 2022). All in all, GWAS is not very powerful when it tries to identify variants with low MAF values. In order to deal with the shortcomings of GWAS, Rare-Variant Association Study methods, such as the burden test, have been developed (Auer & Lettre, 2015).

During burden tests, association with disease is not tested on individual variants, but on sets of variants instead. The number of minor alleles that occur in each set is counted. This is represented as a summary score for each set. This score can be allelic, which reflects the presence of variants on both alleles, or binary. Since association with the entire set is tested, variants that do not increase disease risk decrease the power of the test when they occur in the set (Auer & Lettre, 2015; Chen et al., 2022; Lee et al., 2014). These sets, or units, in which variants are grouped together have to be determined beforehand. This can be a biological unit such as genes, exons or domains. Alternatively, sliding windows (Chen et al., 2022) or clustering methods can be used. Previously, Loehlein Fier et al. (2017) devised a method to cluster variants together within sliding windows by introducing breaks in between variants that are relatively far apart on the linear sequence. This improves test power and is biologically relevant as germline disease causing variants, irrespective of disease origin, have a tendency to “cluster” together (Sivley et al., 2018). The method is also relevant for ALS research as clusters, or “hotspots”, with ALS associated variants have been observed in the C-terminal domain of ALS genes *FUS* and *TARDBP* (Lattante et al., 2013; Zou et al., 2017). In our study we define a hotspot as the specific region within a protein that is associated with disease. While the 1D spatial clustering method has yielded promising findings for ALS (Zonneveld, 2022), it neglects that proteins are not just linear sequences. Instead, they are folded into functional 3D structures. Variants that are close together in space, but far apart in the sequence will never be clustered together by 1D methods (Fig.

1AB). Because of this, the hotspots detected using 1D clustering do not necessarily represent all biologically relevant hotspots. Indeed, Sivley et al. (2018) show that relevant hotspot regions can be introduced due to protein folding as less than 40% of the spatial patterns identified using protein structures were also retrieved from the linear sequences. Including structural information within variant testing has already helped to discover potential new hotspots within cancer and Alzheimer research (Jin et al., 2022; Niu et al., 2016). Discovering which regions of the protein are affected in disease deepens our understanding and could help in the prediction of future risk variants. Therefore, identification of hotspots using spatial clustering methods that take 3D structures into account can improve hotspot discovery and provide more insight into disease.

To our knowledge no study has tried to identify ALS hotspots with 3D spatial clustering methods yet. The absence of experimental structures for important ALS proteins might explain this. In order to perform 3D spatial clustering, structures of the folded proteins are necessary. Previously, this would have been a problem for many studies as only a small percentage of proteins have a representative experimental structure within the Protein Data Base (PDB). With advances of AlphaFold2 (AF2), this problem may be solved as structures can now be predicted based on the protein sequence (Tunyasuvunakool et al., 2021). Therefore, protein models predicted by AF2 can be used for proteins with and without previously defined experimental structures. The advances in AF2 have encouraged studies on protein structures using both newly developed and already existing algorithms. For analysis of variants within tumours, a wide variety of clustering methods has already been designed (Gao et al., 2017; Martinez-Ledesma et al., 2020; Tokheim et al., 2016). However, the majority of these methods were not specifically made to define units for burden testing. We focus on a 3D sliding window approach, because 1D sliding window methods have proven to be reliable for the discovery of ALS hotspots. A direct translation of 1D to 3D sliding windows, which we refer to as spherical clustering in this study, uses spheres with predefined radii to slide across the protein structure. Each variant is subsequently used as the centre of their own sphere and all variants that occur within this sphere are clustered together (Gao et al., 2017). According to Tang et al. (2020), this method has some shortcomings for application in burden testing, such as the use of a predefined window shape. To deal with this limitation, they developed an alternative clustering method, namely protein-structure-based scan (PSCAN). This algorithm iterates over a selection of automatically determined window sizes to constructs graphs, in which variants are represented as nodes. All variant pairs with distances below the window size are connected by edges (Fig. 1C). Next, the connected components in the resulting graph are clustered together. These clusters are subsequently used as units during burden testing in order to identify hotspots associated with disease (Tang et al., 2020).

The aim of our study is to determine whether 3D spatial clustering methods can be used to identify ALS associated hotspots. The insights that we obtain can be used during future studies to find new hotspots in genes that have already been associated with disease and to empower novel disease gene discovery. In order to consider the application and limitations of 3D spatial clustering for ALS research, we apply spherical clustering and PSCAN to replicate known ALS hotspots. For our purposes, we have chosen to study three well-known ALS proteins: *SOD1*, *FUS* and *TARDBP*. These three proteins are relevant to answer our research question since they represent three types of hotspot results. First, *SOD1* represents a protein where disease associated variants occur across the entire protein. Therefore, no hotspot is present in this protein. Second, *FUS* represents a protein with a small localised hotspot (which spans < 10% of the protein) that covers the majority but not all relevant variants. Third, proteins with a large and potentially irregular shaped hotspot (which spans > 30% of the protein) are represented by *TARDBP* (Lattante et al., 2013; Zou et al., 2017). Furthermore, *FUS* and *TARDBP* are relevant for our study since these proteins do not have existing experimental structures. Therefore, we can investigate whether AF2 models solve the limitation caused by the

absence of experimental structures. Following the replication analysis, we also perform an exploratory analysis on four candidate ALS genes (*KIF4A, UTP14C, UNC13C* and *TTC3*) to test the intended application of our method further. All in all, we intend to gain better insight in the current possibilities for 3D spatial clustering of rare variants and determine in which directions development is still necessary.

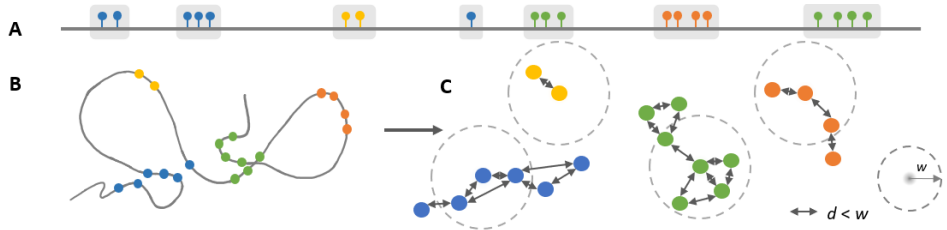


Figure 1. Spatial clustering of genetic variants. A comparison of results obtained with 1D clustering (A) and 3D spatial clustering (BC). Clusters obtained by 1D clustering are displayed in light grey boxes (A). Variants are visualized as small coloured circles. The colour represents clusters obtained from one iteration of PSCAN (Tang et al., 2020) (ABC). For each of these PSCAN clusters, a comparison with spherical clustering is displayed to visualize the ability of each method to discover elongated hotspots. Spherical clusters are represented by large circles (C). During clustering, connections (dark grey lines) are made when the pairwise distance (*d*) between two variants is below the window size (*w*). For PSCAN, all connected components in the constructed graph become one cluster (C).

2. METHODS

2.1 Obtaining genetic variant datasets

The whole exome sequencing variant dataset used during this study was retrieved from Project MinE (Project MinE ALS Sequencing Consortium, 2018) as genome database (GDB) files, the adapted SQLite based data storage format used in the Rare Variant Analysis Toolkit (RVAT) package (Kenna & Hop, 2023). The cohort included for association testing included 13,138 ALS patients and 69,775 controls (for full details on sample and variant quality control see van Rheenen et al. (2021)). All clustering and burden testing were restricted to non-synonymous SNVs that were annotated as missense variants within the canonical transcripts (Table S1) of the ALS proteins of interest (*SOD1, FUS, TARDBP, UNC13C, UTP14C, TTC3* and *KIF4A*). A second GDB was constructed for a separate reference analysis on 2 cancer proteins, which were previously reported to exhibit striking hotspot localisations (*RAC1* and *MAP2K1*) (Gao et al., 2017). We generated this GDB by combining variant data of MinE controls with genotype count data of patients described in the ICGC data portal (Zhang et al., 2019). Only variants annotated as “substitutions” and “missense” within the data portal were downloaded (On July 3th and 5th for *RAC1* and *MAP2K1* respectively). This data contained the number of patients (in each cohort) with each mutation and the cohort sizes. Clustering of cancer variants was restricted to the patient specific variants for comparability to Gao et al. (2017). For both GDBs, variant annotation, which includes chromosomal position, was obtained for all variants that occurred at least once in the cohorts and that had a corresponding allele frequency below 0.001. In parallel to our rare variant analysis, we also performed an ultrarare variant analysis for our ALS proteins. Ultrarare variants were further characterized by a maximal amount of 5 carriers.

2.2 Mapping of genetic variants to predicted protein structures

All protein structure predictions were retrieved from the AF2 database (Jumper et al., 2021; Varadi et al., 2022). We chose structures, based on the UniProt ID that corresponded to the canonical transcript defined in Ensembl (Table S1) (Martin et al., 2023). For these transcripts, Ensembl data were also retrieved for the chromosomal start and end position of each exon and UTR. We determine which chromosomal positions occur both within an exon and outside of the UTRs, to obtain each position that corresponds to a codon. Because of this, chromosomal positions can be translated directly to protein positions. Based on this, the chromosomal positions were added to the AF2 files,

which already included protein positions. These chromosomal positions were used to combine AF2 and annotation files in order to obtain atom coordinates of all rare variants.

2.3 Simulation of randomized variant positions and randomized 3D protein structures
In parallel to the main pipeline, we also perform permutation testing using randomised variant positions to determine if predicted hotspots are significant or if they could have been obtained by chance instead. For each proteins, we perform 100 iterations in which the translation between chromosome and protein position is randomised. Next, clustering and burden testing is performed for each iteration. To obtain a significance threshold for each protein, the association score of the cluster most strongly associated with ALS is selected for each iteration.

In addition, we perform a parallel analysis with randomised structures instead of AF2 models. For each protein, 100 randomised backbone structures, of equal length to the original AF2 prediction, are obtained (Methods S1). These structures do not contain sidechains and do not have biological plausible foldings. The randomised structures are relatively more elongated compared to biological structures, but they do contain folded regions. For our purposes, determining how much predicted hotspots depend on the exact AF2 structure, these non-biological structures are adequate. Again, clustering and burden testing is performed for each iteration. Based on the association scores of the most informative clusters, we obtain a 95% Confidence Interval (CI) for each position.

2.4 Spatial clustering of genetic variants with PSCAN and spherical clustering algorithms
The two clustering methods described in the introduction, spherical clustering and PSCAN, are performed based on the x, y and z coordinates of each atom of the variant positions. A distance matrix is constructed, which contains the smallest distance between each variant position. For spherical clustering we use this matrix and a fixed window sizes (or radius) of 5 angstrom (Å) as described by Gao et al. (2017). We also repeat the analysis with 10 Å when specified. The general use of these radii was validated by Hicks et al. (2019). Each variant is subsequently used as the centre of a spherical window. All variants that occur within the window become clustered together. Since overlapping clusters are not combined, spherical clustering results in a number of clusters equal to the number of variant positions.

The PSCAN algorithm, as described by Tang et al. (2020), differs from spherical clustering in two aspects. First, the algorithm iterates over multiple window sizes. All unique values from the distance matrix, rounded on 1 decimal, are used as window sizes. Rounding is done to decrease the number of clustering iterations while retaining relevant differences in distances. Second, clusters are combined by the use of adjacency-matrix based graphs in which variants are represented as nodes and proximity is represented by edges. Adjacency matrices are constructed by changing the distance matrix values, that are lower than the window size, to 1 and by changing the other values to 0. All connected components of the graph become clusters.

2.5 Association testing of spatial clusters
In order to determine whether each cluster is significantly associated with ALS, a burden test was performed using the assocTest function of the RVAT package (Kenna & Hop, 2023). During testing, each cluster is considered as a unit. The statistical test “firth” and the genetic model “allelic” are used. The covariates previously used by the same lab were also applied within this study. All p-values obtained during testing are capped at 10^{-16}, which means that no lower p-values can be obtained.

Three additional analyses were performed on the burden testing results of PSCAN clusters. These analyses were only performed on clusters with a cumulative minimum allele count of 5 since association scores are not reliable when the allele count is too low (Tang et al., 2020). First, a Spearman correlation test was performed to determine whether cluster size (number of protein

positions that make up the cluster) is correlated with the p-value. In our study, we refer to this test as the cluster size correlation test. Second, the omnibus test ACAT-O, previously described by Liu et al. (2019), was used to obtain a combined p-value for each protein. Third, the "most informative" division of clusters was determined, as described by Tang et al. (2020), to obtain the most significant non-overlapping clusters. The clusters with smallest p-values are iteratively selected. Every time a cluster is selected, all clusters with overlapping variants are removed from the group that is used during the next iteration of the most informative division search. In case multiple overlapping clusters obtain the same significance value, only the largest cluster is selected.

2.6 Analysis on the reliability of AlphaFold2 wildtype and mutant structure predictions

In addition to burden testing based on 3D units, we also perform three analyses on AF2 models. First, we compare predicted local distance difference test (pLDDT) scores to the locations of intrinsically disordered regions (IDRs) and annotated domains, within the canonical transcripts, according to the MobiDB (Necci et al., 2017) and Interpro database (Hunter et al., 2009) respectively. Second, *SOD1* experimental and AF2 structures (Table S2) were compared. Experimental structures were chosen based on their overall structure quality in the Protein Data Bank (PDB), from RCSB.org. Since the structures are in different orientations, they were first aligned with the pairwise alignment tool on the same website before downloading them (Berman et al., 2000; RCSB Protein Bank, 2023). For each pair of aligned structures, a deformation score, the distance between Cα atoms at each position, was calculated. Mean scores are used to compare overall deformation between pairs of structures.

Third, in order to investigate whether AF2 can predict reliable mutant structures, the deformation score is used to compare predicted mutant and wildtype structures as similar metrics are considered to be reliable predictors for mutation effect (McBride et al., 2023). Following recommendations for mutant structure predictions, we used AF2 Google Colab (AF_{colab}) as this version of AF2 does not use templates (Jumper et al., 2021; Reynisdottir et al., 2022; McBride et al., 2023). The AF_{colab} algorithm uses a smaller reference database compared to the general AF2 algorithm. Sequences for wildtype predictions are obtained from Ensembl (Martin et al., 2023) and mutations are added manually. The mutations were chosen based on their impact in ALS (van Deerlin et al., 2008; van Rheenen et al., 2016; van Rheenen et al., 2021; Yamashita & Ando, 2015; Zhou et al., 2020). Predictions were made twice for the *SOD1*, *FUS* and *TARDBP* wildtypes, once for the A5V *SOD1* mutant, the P525L *FUS* mutant and the G298S *TARDBP* mutant and five times for the *CFAP410* wildtype and the V58L *CFAP410* mutant. In addition, wildtype models are obtained from the AF2 database (Jumper et al., 2021; Varadi et al., 2022). The predictions for the same protein were compared with each other to study the effect of mutations and the AF2 version. For *SOD1*, corresponding experimental structures (Table S2) were also used within the comparison. All alignments were performed based on the entire protein. For *FUS*, *TARDBP* and *CFAP410* additional alignments, based on high confidence regions (280-370 and 420-455 for *FUS* and 1-78 and 105-260 for *TARDBP* and 1-145 and 210-256 for *CFAP410*), were also made. Together, the three analyses give insight into the reliability of AF2 predictions.

3. RESULTS

3.1 Low confidence scores of AlphaFold2 models indicate disorder

We downloaded the structures of *SOD1*, *FUS* and *TARDBP* from the AF2 database (Jumper et al., 2021; Varadi et al., 2022). We consider the prediction quality of these structures to determine if they are appropriate alternatives to experimental structures. The pLDDT confidence score of the *SOD1* structure is high for the entire protein (Fig. 2A). Moreover, the predicted structure differs as much from experimental structures as these experimental structures differ from each other (Fig. 3). Strikingly, the majority of the *TARDBP* and *FUS* structures are predicted with low confidence. Visually, these regions seem to be unfolded (Fig. 2BC). In order to investigate whether unconfidently predicted

regions are supposed to be unfolded or whether low confidence scores reflect a poor prediction, we compared the predicted structures to known annotations (Fig. 2D). Results show that low pLDDT scores overlap with IDRs, while high pLDDT scores overlap with ordered domains. We conclude that low confidence scores do not necessarily refer to bad predictions. Instead, these low scores imply genuine biological uncertainty. Based on these preliminary findings, AF2 models appear to be good approximations of biological structures and can be used reliably instead of experimental structures.

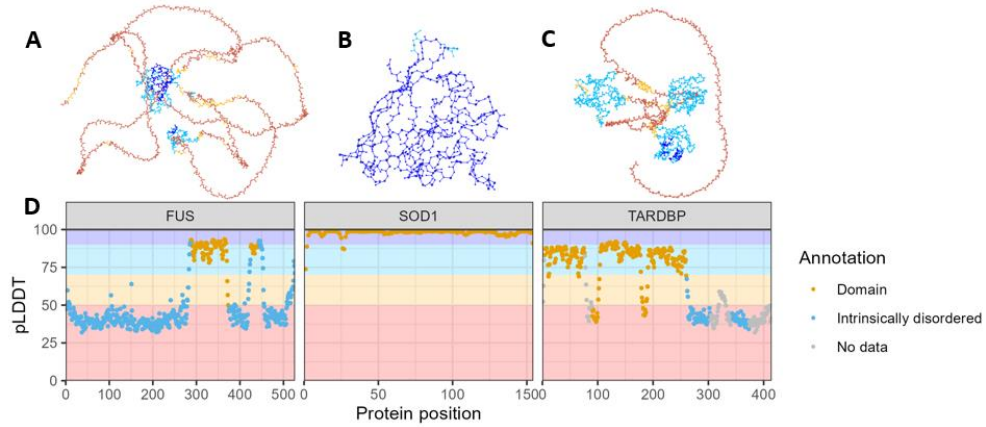


Figure 2. Confidence score of AlphaFold2 structure predictions for FUS (AD), SOD1 (BD) and TARDBP (CD). The structures are coloured according to the pLDDT score calculated by AlphaFold2 (A-C). These scores and known annotation are plotted for each position in the proteins. Annotation refers to either Interpro domains (orange) or MobiDB intrinsic disorder predictions (blue). Some positions have not been annotated with either domains or disorder (grey) (D).

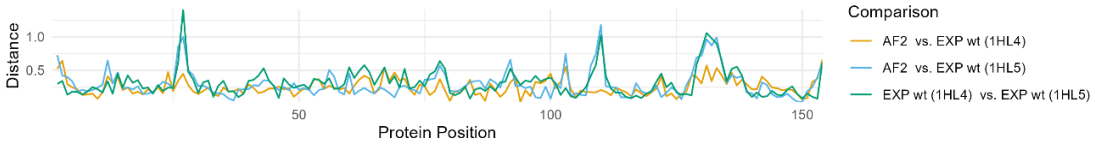


Figure 3. Deformation between SOD1 experimental and computational structures. Distance between AlphaFold2 wildtype SOD1 structure and two experimentally determined wildtype structures are displayed for each position in the protein.

3.2 AlphaFold2 should not be used to predict mutant structures

The presence of variants affects protein structures and alters the pairwise distance between positions used during clustering. It is useful to know whether AF2 mutant structures can be used for variant analyses as the majority of mutant structures have not been resolved experimentally. We use AF_{colab} for predictions as this algorithm does not use templates that would introduce wildtype bias (Reynisdottir et al., 2022; McBride et al., 2023). In addition to *SOD1, FUS* and *TARDBP*, the model of *CFAP410* was analysed as well as this protein is known for an ALS associated variant outside of its disordered region. The number of reference sequences used during prediction changed when mutations were added to the input sequence (Table S3). Deformation scores were calculated between each structure pair. For *FUS*, *TARDBP* and *CFAP410*, large deformation scores did not correspond to mutation placement. Instead, these scores occurred at low confidence regions (Fig. 4A, Fig. S1-S4). This was not due to stochasticity as only minor deformation scores were obtained between AF_{colab} models made from the same sequence (Fig. 4A, Fig. S4). When our AF_{colab} models are compared to AF2 database models, which were predicted using a larger reference set, we observe a similar effect on the low confidence region (Fig. 4A). This shows that both changes to the input sequence and AF2 algorithm have a large effect on the predicted structure for disordered regions.

The potentially biologically relevant changes in the protein structure, as a result of mutations, are overshadowed by the presence of disordered regions. Therefore, we specifically aligned the ordered regions of *CFAP410* to investigate the direct effect of mutations. These alignments resulted in minor deformation scores with a small peak on the mutation location (Fig. 4B). Similar results were obtained

for *SOD1* structures, which only contain high confidence regions. Indeed, the only notable change in the AF_{colab} models of *SOD1* occurs (at the end of the protein) close to the mutation (Fig. 5). This change is not observed in the experimental structures. Furthermore, experimentally observed differences between wildtype and mutant structures are not present in the AF_{colab} models (Fig. 5). The mutant AF_{colab} model resembles the wildtype structures more than the experimental mutant structure (Fig. S5). Our results indicate that AF2 is not capable of finding reliable mutant structures. Therefore, only the wildtype models should be used to obtain coordinates.

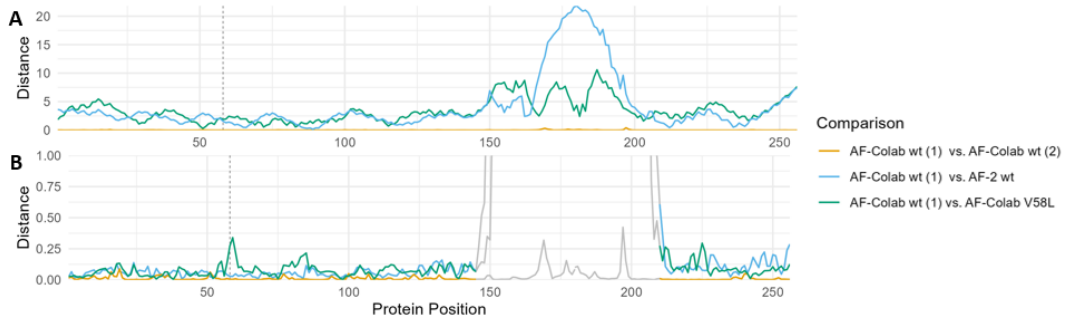


Figure 4. Deformation scores between AlphaFold2 structures of *CFAP410*. The scores are based on full structure alignments (A) and localised (only folded regions) alignments to filter out the deformation caused by disordered regions (B). The deformation score is the distance at each position between structure pairs. The position of the mutation is displayed by a vertical line.

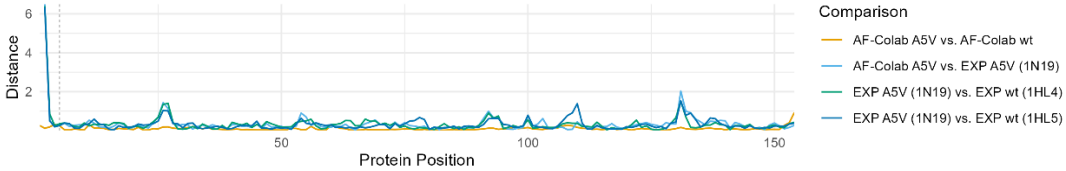


Figure 5. Deformation scores between computational and experimental *SOD1* structures. The score represents the distance at each position between aligned structure pairs. The position of the mutation is displayed by a vertical line.

3.3 Hotspot identification with spherical clustering is limited by predefined window sizes

Spherical clustering was used to find units for burden testing. Potential hotspots consist of the clusters with the highest level of association. A permutation test is performed to determine whether the same level of association can be obtained when variants are randomly distributed. When significance is reached according to permutation testing, all additional significant clusters are also considered as potential hotspots. In order to test our spherical clustering method, we aimed to reproduce *MAP2K1* and *RAC1* cancer "hotspots" that Gao et al. (2017) previously identified with a similar method. All but one of our predicted hotspots overlapped with these expected hotspots (Fig. S6). However, none of the clusters reached significance during permutation testing (p-value = 0.37 and p-value = 1 respectively) (Fig. S7). This may be an artifact of p-value capping as multiple clusters reach the maximal association value. After this preliminary analysis, we decided to apply this method to ALS hotspot detection. First we consider if our *SOD1* results match with expectations. Subsequently, we compare our *FUS* and *TARDBP* results with previously detected hotspots.

Previous studies did not find hotspots within *SOD1* (Zou et al., 2017). Therefore, we expect to find no significant potential hotspot during our analysis. We find multiple rare variant clusters that are maximally associated with ALS (Fig. 6A, Fig. S8A), while ultrarare variant clusters reach relatively lower association scores (Fig. S8B**,** S9B). Both rare and ultrarare variant analysis match expectations as no significance is reached (p-value = 1 and p-value = 0.23 respectively) (Fig. 6B, Fig. S10A). This indicates that permutation testing can help to distinguish between hotspot presence and absence.

In case a hotspot is present, our hotspot detection method should also be able to discover the correct location. Therefore, we compare the location of known and predicted hotspots in *FUS* and *TARDBP*. In *TARDBP,* we expect to find a C-terminal hotspot that covers a large region (260-400) (Lattante et al., 2013). However, neither rare or ultrarare variant analysis identifies any significant hotspot (p-value = 0.39 and p-value = 1 respectively) (Fig. 6B, Fig. S8, S10A). Similar results are obtained when we increase the window size to 10Å (p-value = 0.69 and p-value = 1) (Fig. 6CD, Fig. S10B, S11). The non-significant rare variant cluster, with the highest level of association, does occur within the expected region (383-384). However, we are not able to obtain any cluster that resembles the known *TARDBP* hotspot as no sphere can cover this entire region (Fig. 6AC, Fig. S9C). Based on this, we hypothesize that spherical clustering can only be used to replicate localised hotspots that cover a smaller region. To investigate this further, we consider the known localised hotspot (490-526) in *FUS* (Lattante et al., 2013; Zou et al., 2017). Our analysis does indeed identify a significant rare variant hotspot, which covers variants at positions 517 and 521, within the expected region (p-value = 0.01) (Fig. 6AB). However, the association of this cluster is only driven by the rare variants at position 521 (Fig. S12). Moreover, no significant ultrarare variant cluster can be obtained as most of the association driving variants at position 521 are not included during ultrarare variant testing (p-value = 0.77) (Fig. S9A, S10A). Together, this indicates the presence of a hotspot position rather than a hotspot region. Since our 5Å spheres were not able to cover the entire expected hotspot region, we consider whether we can detect a hotspot region when radii are increased to 10Å (Fig. 6C). This parameter change improved hotspot detection as now both rare and ultrarare variant analyses yield significant hotspot regions (p-value = 0 and p-value = 0) that cover a part (517-526) of the established hotspot (Fig. 6CD, Fig. S10B, S11). Thus, our results for both *FUS* and *TARDBP* show that the selection of an appropriate window size is important for reliable hotspot detection.

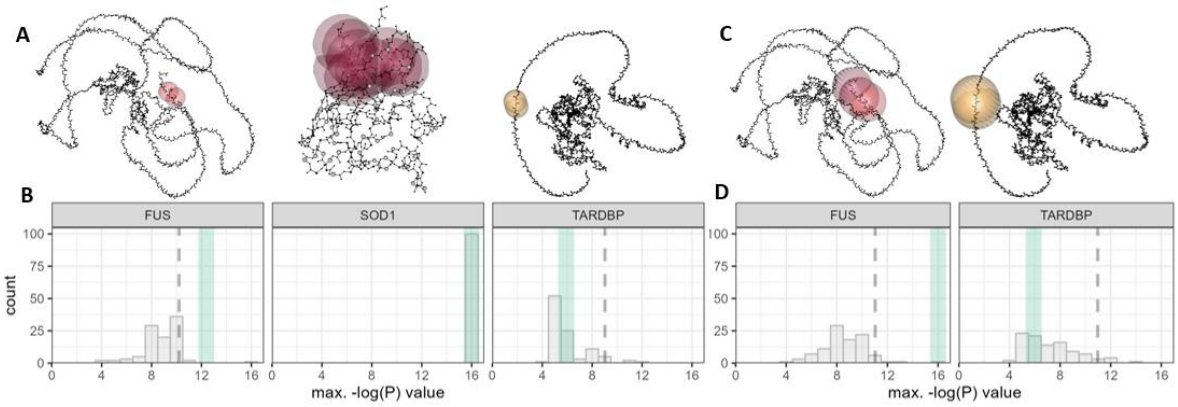


Figure 6. Predicted rare variant hotspots obtained with spherical clustering analysis. The location of the 5Å (AB) and 10Å (CD) spheres are displayed in 3D. The colour of the spheres indicates relative association values (yellow for low values and dark red for high values). For *FUS*, all clusters that reach significance according to permutation testing are displayed. For *SOD1* and *TARDBP*, only the spheres with the highest level of association are displayed (AC). Permutation test for spherical clustering results for each protein. Grey histogram bars represent counts of maximal association scores obtain in each permutation. The grey dashed line represents significance threshold (p-value = 0.05) and the green line represents the burden test result for the potential hotspot (BD).

3.4 PSCAN methods can identify both localised and elongated hotspots

The same analysis was repeated with a PSCAN-based burden test to determine whether this method is an improvement over spherical clustering as stated by Tang et al. (2020). In this analysis, only the largest clusters that are most strongly associated with ALS are considered potential hotspots. The PSCAN method was not able to find any cluster that resemble the known hotspots for *RAC1* or *MAP2K1* (Fig. S13). Permutation testing further indicates that no hotspot is predicted for these proteins (p-value = 1 and p-value = 1). These results imply that the two clustering methods might be

suitable in different use cases. We apply PSCAN in ALS hotspot detection to determine whether the method is appropriate in the use cases represented by *SOD1*, *FUS* and *TARDBP*.

First, we consider our PSCAN results for *SOD1*. Corresponding to our expectations for this protein, we find no potential hotspot region during our (ultra)rare variant analysis. Indeed, clusters that contain all tested variants reach maximal association scores (Fig. 7A, Fig. S14, S15). Randomly distributing the variants across the protein does not change this. Therefore, the same association score is reached in each permutation (p-value = 1) (Fig. 7B, Fig. S16). We extend our PSCAN analysis to include cluster size correlation testing to gain a potential additional method to distinguish between presence and absence of hotspots. A hotspot cluster should only cover a part of the tested variants. Therefore, adding non-hotspot variants decreases the association significance. Because of this, we expect a relatively low degree of correlation between association score and cluster size for a protein with an actual hotspot. For that reason, the absence of hotspots in *SOD1* is further substantiated by a relatively high degree of correlation (ρ = 0.5, p = 5.3x10^{-4}) (Table S4). All in all, similar to spherical clustering, PSCAN results clearly show that, as expected, there is no hotspot within *SOD1*.

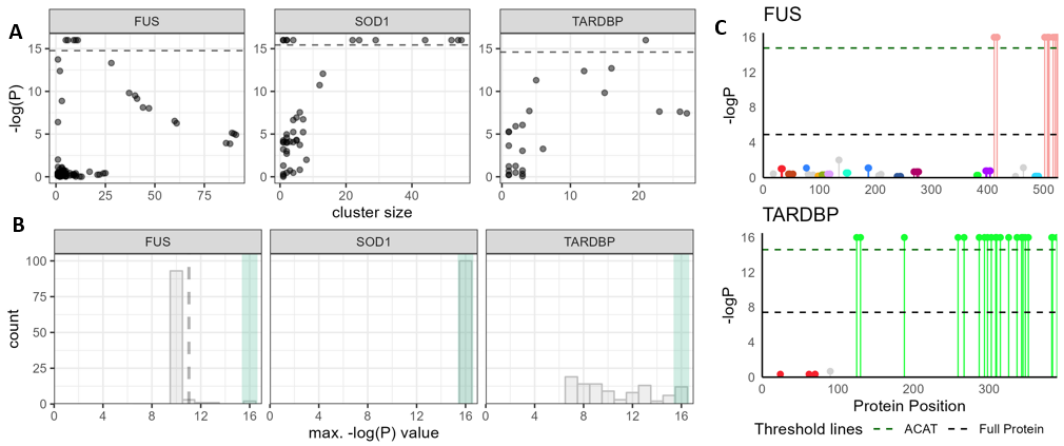


Figure 7. Association scores of PSCAN clusters and significance of potential hotspots from rare variant analysis. The size of the cluster is plotted against the association score (-log(p-value)). The horizontal dashed line represents the -log(P) value obtained with ACAT-O (A). Grey bars represent a histogram of the maximal -log(P) in each iteration of the permutation test. A green line is added to the plot to show the -log(P) value of the potential hotspot, i.e. the cluster with the highest -log(P) value obtained in the non-randomised variant distribution. The grey dashed line represents the significant threshold for p-value = 0.05, based on the permutation test. This line is only displayed when the threshold could be calculated (B). Rare variant hotspot results obtained for *FUS* and *TARDBP*. Each colour represents a separate cluster and the most significant clusters represents the potential hotspot. Grey coloured positions represent positions that have not been clustered (C).

In contrast to *SOD1*, hotspots are expected within both *FUS* and *TARDBP*. Spherical clustering was only able to identify the localised *FUS* hotspots, as no prediction matched with the known large *TARDBP* hotspot. PSCAN might be reliable in a wider range of use cases as there is no manual selection of window sizes. Indeed, preliminary PSCAN analysis indicates hotspot presence in both proteins as clusters maximally associated with ALS do not cover the full set of tested variants (Fig. 7A, Fig. S14). In line with this, the ACAT score substantiates the presence of a hotspot for both proteins as this score reaches a higher level of association compared to full protein clusters (Fig. 7AC). Consistent with our expectations of a localised and elongated hotspot in *FUS* and *TARDBP* respectively, the maximally associated clusters in *FUS* contain relatively few variants, while the potential hotspot cluster contains the majority of variants in *TARDBP*. The difference in hotspot size is also reflected in the cluster size correlation testing, as we obtain a low degree of correlation (ρ=0.28, p=0.01) for *FUS* and a high degree of correlation (ρ=0.67, p=2.4x10^{-4}) for *TARDBP* (Table 4). This shows that a high degree of correlation does not necessarily indicate the absence of hotspots as we assumed beforehand. Based on our analysis, the correlation test can not distinguish between a protein with a

large hotspot cluster, such as *TARDBP*, and a protein that is associated with disease in its entirety, such as *SOD1*. Preliminary inquiry into the association scores of the different cluster sizes do indicate that PSCAN can retrieve the expected hotspot results for both localised and elongated hotspots.

We analyse our PSCAN results further by comparing the predicted and known hotspot location. PSCAN is more appropriate for our use cases than spherical clustering, if the predicted locations match with expectations for both proteins. Similar to our spherical clustering results, we obtain significant rare and ultrarare hotspots for *FUS* (p-value = 0.01 and p-value = 0.01), while no significance is reached for *TARDBP* (p-value = 0.08 and p-value = 0.18) (Fig. 7B, Fig. S16). Still, we investigate whether PSCAN and burden testing were able to identify the expected region in both proteins. The exact rare variant hotspot region can not be determined for *FUS* as multiple clusters reach the same maximal association score due to p-value capping (Fig. 7A). In order to circumvent this problem, we perform an additional rare variant analysis with SKAT burden testing for this protein (Fig. S17). Our *FUS* analyses yield a 3D hotspot that overlaps with a larger part (507-526) of the expected region (490-526) compared to spherical clustering (517-526). However, the predicted hotspot also contains additional variant positions that do not match with our expectations (412-417) (Fig. 7C, Fig. S18-S19). Similar results are obtained for *TARDBP* as PSCAN predicts a hotspot (125-405) that covers an unexpected large part of the protein. Still, the prediction matches well with our expectations as the majority of the rare and ultrarare variants (all but three and four respectively) in the predicted hotspot occur within the expected region (264-400) (Fig. 7C, Fig. S18A). While PSCAN predicts hotspots that contain unexpected variants, the method is an improvement over spherical clustering as approximate regions can be replicated for both localised and elongated hotspots.

The inclusion of the unexpected variants is investigated further to determine whether they are indeed non-hotspot variants. To this end, we use SKAT burden testing to compare the association scores of expected regions (500-526 and 259-390) with the association scores of predicted regions for both proteins. We obtained slightly lower levels of association when the unexpected variants are included (Fig. S20). This indicates that the unexpected variants do not contribute to ALS association. We hypothesise that the variants may have been included in prediction due to their proximity with expected hotspot variants. This is supported by randomised structure predictions which, generally speaking, lack the unexpected variants (Fig. S21). The predicted regions primarily occur within IDRs that appear unfolded in the AF2 model (Fig. 2A, Fig. S22). Our results give no evidence that the unexpected variants are part of the biological hotspot. The results further shows that PSCAN is limited by the exact 3D coordinate prediction of IDRs. A second potential limitation has been observed as well. Indeed, we obtain a slightly different hotspot region, which does not include the unexpected variants, when we remove position 521 prior to clustering. This indicates that the hotspot prediction is affected by the exact set of positions included within clustering (Fig. S19). Both observations thus point to potential limitations of PSCAN that we should consider further to develop an improved, even more reliable, hotspot detection method.

3.5 Pilot analysis on candidate ALS proteins identifies a potential 3D hotspot

The PSCAN and spherical clustering methods are meant to help during discovery of new hotspots. Therefore, we apply these methods to four proteins (*KIF4A*, *TTC3*, *UTP14C* and *UNC13C*), which have recently been associated with ALS (Fig. S23-S29). No significance in either PSCAN (p-value = 0.56, p-value = 0.4 and p-value = 0.09) or spherical clustering (p-value = 0.27, p-value = 0.94 and p-value = 1) analysis is obtained for *KIF4A, UTP14C* and *UNC13C* (Fig. S23-S25, Table S5). Moreover, no cluster is clearly more associated with ALS compared to the other clusters for any of these proteins (Fig. S26-S29). According to our methods, there is no hotspot present within any of these three proteins. In contrast, the PSCAN based method does identify a significant rare variant hotspot for *TTC3* (p-value =

0.01) (Fig. 9AB), which occurs at the interface between two structured regions (Fig. 9C). Therefore, we conclude that our PSCAN based method is able to identify a potential 3D ALS hotspot in the structured part of *TTC3*.

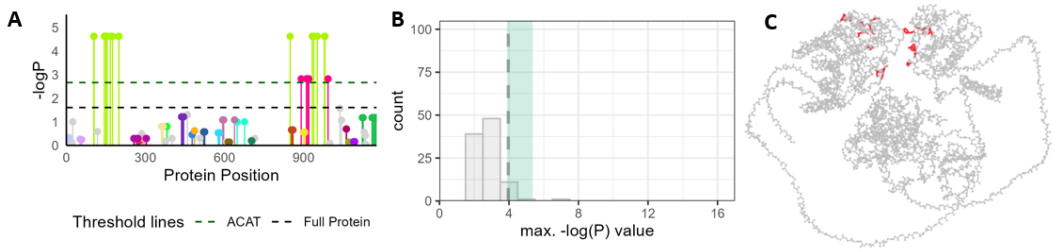


Figure 9. Potential hotspot in *TTC3* based on rare variant PSCAN analysis. The burden test results, (A) the corresponding permutation test results (B) and location in 3D (C) are displayed. Each colour represents a separate PSCAN cluster while grey coloured positions represent variants that have not been clustered (A). Grey bars represent a histogram of the maximal -log(P) in each iteration of the permutation test. A green line is added to the plot to show the -log(P) value of the potential hotspot, i.e. the cluster with the highest -log(P) value obtained in the non-randomised variant distribution. The grey dashed line represents the significant threshold for p-value = 0.05, based on the permutation test. This line is only displayed when the threshold could be calculated (B). The 3D location of the hotspot is displayed by colouring the variants in red (C).

4. DISCUSSION & CONCLUSION

The aim of this study was to determine whether 3D spatial clustering methods could be used in the study of ALS associated genetic hotspots. In order to consider this, we applied both PSCAN and spherical clustering to replicate known ALS hotspots. We observed that these methods were able to replicate known hotspots relatively well. However, some key limitations to both methods prevented optimal hotspot identification. To make 3D spatial clustering a viable alternative to 1D spatial clustering techniques, it is necessary to specify these limitations and describe potential solutions that can be used to improve upon the existing methods. We will conclude whether spatial clustering methods should take 3D structural information into account for application in ALS hotspot discovery.

4.1 PSCAN rather than spherical clustering should be used to identify 3D hotspots

We consider the ability of both spherical clustering and PSCAN-based association tests to reproduce known hotspots. Prior to performing a cluster based association test to identify new hotspots, the shape and size of the hotspots are not known. Therefore, the test that is used should be able to find both small localised and large elongated hotspots. Both methods were able to detect a significant localised hotspot in *FUS* that matched with the approximate location that previous studies identified (Lattante et al., 2013; Zou et al., 2017). The choice of window size was crucial to obtain this prediction for spherical clustering as the result is limited by the variants that fit within the spherical window. PSCAN is not limited by predefined window size and thus predicts relatively larger hotspot regions. The effect of this limitation of spherical clustering is even more apparent for *TARDBP*. Indeed, spherical clustering was not able to identify a cluster that resembled the expected hotspot as, again, spheres were not able to cover enough variants from the expected region. On the other hand, PSCAN clustering and burden testing results matched with the known hotspot location. Despite the fact that these clusters were consistent with the hotspot described by Lattante et al. (2013), permutation testing results did not match with the presence of any hotspot in *TARDBP*. In contrast, this test was able to correctly detect hotspot presence in *FUS* and absence of hotspots in *SOD1*. This shows that finding a significant result might be more challenging for large hotspots that cover the majority of variants as we see for *TARDBP*. It should be explored whether the test consistently lacks statistical power for large hotspots or whether the non significant result was primarily caused by capped association scores. In the former case, additional methods that can deal with both localised and elongated hotspots are necessary to determine hotspot presence. However, these permutation test results do not change our findings for the two clustering methods. Indeed, we conclude that PSCAN is

the more appropriate clustering method as it can find reliable results in all tested use cases, while this is not the case for spherical clustering. Therefore, PSCAN should be used to identify 3D hotspots during future analyses.

Three caveats have to be made to this preliminary recommendation. First, the best clustering method depends on the exact definition of a hotspot. PSCAN is an excellent method to find specific protein regions that are associated with disease. This is reflected by the assumption that adding non-hotspot regions to the hotspot-cluster decreases the significance of association. However, some studies use a different hotspot definition. For instance, the hotspots that Gao et al. (2017) discovered in *RAC1* and *MAP2K1* refer to a part of the protein with a relatively large amount of mutations. In this case, almost all other mutations are also associated with disease. Therefore, adding non-hotspot variants to the hotspot-cluster does not decrease significance. For this hotspot definition, PSCAN assumptions are not met and spherical clustering should be used instead. Second, our study does not cover the entire variety of 3D clustering methods and it is possible that alternative, more appropriate methods exist for either hotspot definition (Tunyasuvunakool et al., 2021). Third, some notable use cases exist for which protein structure based clustering methods can not be used at all. No relevant coordinates can be assigned to untranslated variants as they do not occur within the protein structure. However, they can be crucial to understand disease association of genes. This is reflected by *UNC13C*, for which no hotspots could be identified, as it is mostly known for an ALS associated intronic variant (Daoud et al., 2010; Willemse et al., 2023). Similarly, *KIF5A*, which is related to *KIF4A*, is known for a hotspot at the C-terminal that involves splice-site variants (Brenner et al., 2018). To study these variants and potential hotspots, pre-mRNA based 1D clustering methods should be used instead.

4.2 Limitations of PSCAN to identify reliable hotspots

The PSCAN method solves the main limitation of spherical clustering as hotspots of different shapes and sizes can be discovered without prior information. Still some limitations to PSCAN exist. A notable limitation is caused by the firth statistical test, which caps association scores due to limited precision beyond p-values of 10^{-16}. When multiple clusters reach capped scores, as we observed for *FUS*, it is not possible to predict the exact hotspot region. Analysis with the SKAT statistical test solved this limitation as the cluster with the highest level of association could be identified. This statistical test could also be used to investigate whether significant *TARDBP* hotspots can be obtained if no capping occurs. In our study, we also identified unresolved limitations that have to be addressed.

A thus far unresolved limitation of PSCAN involves the focus on detection of a region most associated with disease, rather than the identification of regions with high variant density. This is relevant for hotspot discovery as disease associated variants tend to cluster together (Sivley et al., 2018). Predicted hotspots for both *FUS* and *TARDBP* contain variants that do not occur within known hotspot regions and that do not contribute to ALS association. These non-hotspot variants occur (linearly) relatively far away from the known hotspot region. In contrast, 1D spatial clustering analysis specifically finds the expected regions by considering variant density (Zonneveld, 2022). Therefore, this 1D method results in predictions that match expectations for *FUS* and *TARDBP* better, compared to the predictions obtained with PSCAN. This observation could be due to the input coordinates, which we will discuss later on. However, it should also be noted that the 1D method performs an additional processing step which removes variants that occur relatively far away from the other clustered variants (Loehlein Fier et al., 2017). Extending PSCAN with a similar processing step could result in an improved method that favours biologically relevant clusters. However, this step is less straightforward for 3D methods as different distributions of distances apply. Graph-based clustering methods have been used as an alternative technique to prioritise 3D regions, which are highly dense with variants, within some cancer related studies (Kumar et al., 2019; Niu et al., 2016). These graph-

based clustering methods, such as Girvan-Newman or Markov clustering, can be used to find a module of well connected variants within PSCAN graphs (Girvan & Newman, 2002; van Dongen, 2000). This removes less connected variants that lie relatively far away. Future studies should test whether informative variant modules can be obtained and test whether hotspot detection is improved. Similarly, label propagation, using relative occurrence in patients versus controls or pathogenicity scores as labels, should be tested as it can help to find the most relevant modules (Dimitrakopoulos et al., 2018; Leiserson et al., 2015). Indeed, graph-based clustering methods could improve hotspot detection as they prioritise biologically relevant regions that are densely mutated with risk variants.

Graph-based clustering methods could also resolve two additional PSCAN limitations. First, these methods improve computational efficiency as a reduced number of clusters have to be tested for association. Second, they may improve robustness with respect to variants included in the study. Currently, adding or removing variant positions to the analysis can affect the result of clustering. Namely, introducing an additional variant not only adds this single position to the cluster, but could in some cases also lead to the union of two separate clusters by bridging the interconnecting distance. Because of this, the exact predicted hotspot region can differ slightly between analyses on the same protein. This is not desirable since the underlying biological hotspot does not change. Therefore, it is important to consider whether extending PSCAN as described above would resolve this limitation or whether additional adjustments to the analyses method are necessary.

4.3 Limitations of AlphaFold2

The coordinates that the PSCAN algorithm uses can originate from experimentally determined structures. However, for many proteins, computational methods such as AF2 are necessary to obtain structures. In our study, we assumed that AF2 models are a good representation of biologically relevant folding. However, there are some important downfalls of AF2 that we should consider. These limitations are not specific to PSCAN or spherical clustering, but apply to all clustering methods that rely on protein structures. First, AF2 does not deal well with different protein conformations (Perrakis & Sixma, 2021). For example, only the apo conformation of *SOD1* is represented by the AF2 prediction (Strange et al., 2003a). Since only one conformation of the protein may be predicted (Saldaño et al., 2022), some conformation specific hotspots could be missed. On top of that, it should be noted that proteins might have a relevant change in conformation upon complex formation. For *FUS*, a change from compact to elongated conformation has been observed upon interaction with RNA (Hamad et al., 2020). This completely changes the distance between variants and could therefore influence hotspot identification. While advances have been made for the prediction of protein-protein complexes, this not is not case for complexes that include RNA or DNA (Perrakis & Sixma, 2021). Because of that, *FUS* and *TARDBP* hotspots that are relevant during complex formation can not be studied with AF2 models. In order to find these conformation specific hotspots, either experimental structures or structures obtained with computational modelling techniques have to be used instead (Allison, 2020; Thomasen & Lindorff-Larsen, 2022). Moreover, it should be noted that AF2 does not predict structures based on fundamental driving forces of folding (Buel & Walters, 2022; Perrakis & Sixma, 2021). Instead, predictions reflect structures that could be present within the PDB (Jumper et al., 2021). This shows a bias towards the structures that occur within the PDB. This has important implications for the prediction of both mutant and IDR structures as we describe below.

We performed a pilot analysis to determine whether AF2 can reliably predict the effect of variants. Variant pathogenicity was previously predicted based on deformation scores calculated from AF2 mutant structures (McBride et al., 2023) and based on the output of a new algorithm that builds on the existing AF2 architecture (Cheng et al., 2023). These methods might already be able to improve

hotspot detection as predicted pathogenicity scores can be used for either filtering out benign and neutral variants prior to clustering or for label propagation during clustering. This leads to the question whether the underlying structural prediction can also be trusted. These structural predictions could lead to a better understanding of disease manifestation by showing the direct effect of mutations. In our pilot analysis, the direct effect of mutations on AF2 structures was mostly overshadowed by IDR artefacts on which we will focus in the next section. Therefore, we consider the effect of mutations within ordered regions of *CFAP410* and *SOD1*. For both proteins, the structure is only effected at the location of the mutation. McBride et al. (2023) also observed a strictly local effect of mutations that gradually declines as distance to the mutation increases. We hypothesize, based on our *SOD1* findings, that predicted mutation effects do not match with biological effects of mutations. Indeed, previous studies showed that predicted mutant structures are not reliable as destabilizing mutations, that cause unfolded structures in nature, did not lead to an unfolded protein prediction (Buel & Walters, 2022). This makes sense as the AF2 algorithm assumes that mutations do not affect the structure (Yang et al., 2023). Furthermore, the inability of AF2 to predict reliable mutant structures may be caused by a innate bias to wildtype structures since these are overrepresented in the PDB.

4.4 Challenges of structure prediction for intrinsically disordered regions

The expected and predicted hotspots for *FUS* and *TARDBP* occur within low confidence regions of the AF2 structures. It is important to consider whether we can trust the coordinates for these regions as they influence our clustering result. Indeed, our randomised structure analysis indicated that the inclusion of non-hotspot variants within hotspot predictions depended on the coordinates predicted by AF2. Low confidence scores are considered good predictors for disorder (Jumper et al., 2021; Tunyasuvukanool et al., 2021). Indeed, we observe that low confidence regions in *FUS* and *TARDBP* overlap with known IDRs. This raises the question whether IDR structures predicted by AF2 can be trusted. The structure of IDRs can usually not be captured experimentally, because of which reliable computational predictions are necessary (Punta et al., 2015). However, absence of IDRs within the PDB also impacts AF2 prediction quality. Our analysis with AF_{colab} models demonstrated this as predicted IDR structures completely change depending on AF2 algorithm or mutations in the query sequence. Both directly affect the number of reference structures that is taken into account. Moreover, it has been hypothesized that the number of potential references is negatively impacted by the fast evolving nature of IDRs (Ruff & Pappu, 2021). Since IDRs are mostly absent from reference structures, small changes in the reference set can have a large impact on the predicted structure. This explains why these regions are always predicted with low confidence. These results also show that the predicted coordinates for IDRs do not represent actual biology. Indeed, while IDRs are unfolded in AF2 models, this is not necessarily the case in nature. Namely, they adopt transient conformations that facilitate interactions with multiple binding partners (Punta et al., 2014; Van Der Lee et al., 2014). This high level of conformation diversity further contributes to a decreased prediction quality (Saldaño et al., 2022). Moreover, neither AF2 or experimental structures can capture the dynamic behaviour and conformational diversity of IDRs. We conclude that AF2 can not predict reliable IDR coordinates.

The observation that IDRs are not predicted reliably explains the occurrence of non-hotspot variants within the PSCAN predictions for both *FUS* and *TARDBP*. In *FUS,* the predicted hotspot contains variants from two separate IDRs which are in close proximity within the AF2 model. Similar effects occur in *TARDBP* as the IDR, which contains the known hotspot, circles around the ordered regions. This shows that IDR predictions interfere with optimal hotspot discovery. Consistent with our observations, AF2 recommends to not interpret the structure of low confidence regions. However, only taking high confidence regions into account during clustering is not a suitable solution. The

presence of hotspots in the IDRs of *FUS* and *TARDBP* indicates that these regions are not trivial. Indeed, IDRs are important in neurodegenerative diseases such as ALS (Babu, 2016; Coskuner-Weber et al., 2018; Santamaria et al., 2017). Variants within these regions are expected to contribute to disease by promoting an irreversible change in conformation that can lead to aggregation of proteins (Lim et al., 2016; Patel et al., 2015; Seera & Nagarajaram, 2022). Based on the involvement of these regions in ALS, we conclude that the reliable prediction of IDRs is the most important limitation of AF2 based spatial clustering methods.

In order to prevent unreliable 3D clustering based on IDR predictions, the disordered regions could be considered as linear. A hybrid 1D-3D clustering method, in which 3D clustering is only applied to ordered regions, could be used. This way, potential 3D hotspots in ordered regions, as we have observed in *TTC3*, can still be detected while it also allows for reliable 1D hotspot detection in disordered regions. Despite the importance of IDRs in ALS, variants within ordered regions of *SOD1* and *CFAP410* (van Rheenen et al., 2021; Yamashita & Ando, 2015) show that it is still relevant to study these regions as well. Still, it would be beneficial to find a method in which transient conformations of IDRs are taken into account. These conformations can be 3D as exemplified by the interaction between N- and C-terminal IDRs of *FUS* (Hamad et al., 2020; Loughlin & Wilce, 2019). This brings the known hotspot in close proximity to N-terminal variants that might together form a relevant 3D hotspot. Therefore, only considering IDRs as separate linear regions, with hybrid 1D-3D methods, could ignore disease relevant interactions and 3D hotspots. Recently developed machine learning approaches might be a solution for this problem. Grazioli et al. (2019) developed a method to study transient conformations of IDRs based on molecular dynamics simulations. All conformations that corresponded with a local energy minimum were used to construct graphs. These graphs, in which residues that are in contact with each other are connected, could also be used to obtain 3D variant clusters. This novel technique could lead to discovery of biologically relevant 3D hotspots within IDRs. New insights into these important, but relatively less studied, regions can contribute to a better understanding of ALS.

4.5. Conclusion

In our study we observed that spatial clustering methods can rediscover known ALS hotspots. The PSCAN method solves limitations of spherical clustering techniques. However, further improvements are still necessary. Graph based clustering techniques could help in the discovery of the most relevant variant dense regions as well as improve robustness. Furthermore, molecular dynamics simulations could be used to obtain reliable data on spatial variant proximity within intrinsically disordered regions as current AlphaFold2 models can not be used for this purpose. All in all, we conclude that the current 3D spatial clustering methods can only be used to identify ALS hotspots within ordered protein regions. This can already lead to the discovery of new hotspots relevant in ALS. However, improvements in 3D spatial clustering methods still have to be explored to study hotspots within IDRs, which are abundant in neurodegenerative diseases such as ALS. Indeed, the most important direction for improvement in ALS hotspot identification techniques lies in mapping IDR interactions and transient conformations.

5. REFERENCES

Al-Chalabi, A., van den Berg, L. H., & Veldink, J. (2017). Gene discovery in amyotrophic lateral sclerosis: implications for clinical management. *Nature Reviews. Neurology, 13*(2), 96-104. https://doi.org/10.1038/nrneurol.2016.182

Allison, J. R. (2020). Computational methods for exploring protein conformations. Biochemical Society Transactions, 48(4), 1707-1724. https://doi.org/10.1042/BST20200193

Asimit, J., & Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annual Review of Genetics, 44*, 293-308. https://doi.org/10.1146/annurev-genet-102209-163421

Auer, P. L., & Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine, 7*(1), 16. https://doi.org/10.1186/s13073-015-0138-2

Babu, M. M. (2016). The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. Biochemical Society Transactions, 44(5), 1185-1200. https://doi.org/10.1042/BST20160172

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1), 235-242. https://doi.org/10.1093/nar/28.1.235.

Brenner, D., Yilmaz, R., Müller, K., Grehl, T., Petri, S., Meyer, T., ... & German ALS network MND-NET Weyen Ute Hermann Andreas Hagenacker Tim Koch Jan Christoph Lingor Paul Göricke Bettina Zierz Stephan Baum Petra Wolf Joachim Winkler Andrea Young Peter Bogdahn Ulrich Prudlo Johannes Kassubek Jan. (2018). Hot-spot KIF5A mutations cause familial ALS. Brain, 141(3), 688-697. https://doi.org/10.1093/brain/awx370

Buel, G. R., & Walters, K. J. (2022). Can AlphaFold2 predict the impact of missense mutations on structure?. Nature Structural & Molecular Biology, 29(1), 1-2. https://www.nature.com/articles/s41594-021-00714-2

Cardoso, R. M., Thayer, M. M., DiDonato, M., Lo, T. P., Bruns, C. K., Getzoff, E. D., & Tainer, J. A. (2002b). Insights into Lou Gehrig's disease from the structure and instability of the A4V mutant of human Cu, Zn superoxide dismutase. Journal of molecular biology, 324(2), 247-256. https://doi.org/10.1016/S0022-2836(02)01090-2

Cardoso, R.M.F., Thayer, M.M., DiDonato, M., Lo, T.P., Bruns, C.K., Getzoff, E.D., Tainer, J.A. (2002a). Structure of the HSOD A4V mutant. https://doi.org/10.2210/pdb1n19/pdb

Chen, W., Coombes, B. J., & Larson, N. B. (2022). Recent advances and challenges of rare variant association analysis in the biobank sequencing era. *Frontiers in Genetics, 13*, 1014947. https://doi.org/10.3389/fgene.2022.1014947

Cheng, J., Novati, G., Pan, J., Bycroft, C., Zemgulyte, A. Applebaum, T., … & Avsec, Z. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science, 0, eadg7492 https://doi.org/10.1126/science.adg7492

Chia, R., Chiò, A., & Traynor, B. J. (2018). Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. *The Lancet. Neurology*, 17(1), 94-102. https://doi.org/10.1016/S1474-4422(17)30401-5

Coskuner-Weber, O., & Uversky, V. N. (2018). Insights into the molecular mechanisms of Alzheimer's and Parkinson's diseases with molecular simulations: understanding the roles of artificial and pathological missense mutations in intrinsically disordered proteins related to pathology. International journal of molecular sciences, 19(2), 336. https://doi.org/10.3390/ijms19020336

Daoud, H., Belzil, V., Desjarlais, A., Camu, W., Dion, P. A., & Rouleau, G. A. (2010). Analysis of the UNC13A gene as a risk factor for sporadic amyotrophic lateral sclerosis. Archives of neurology, 67(4), 516-517. https://doi.org/10.1001/archneurol.2010.46

Dimitrakopoulos, C., Hindupur, S. K., Häfliger, L., Behr, J., Montazeri, H., Hall, M. N., & Beerenwinkel, N. (2018). Network-based integration of multi-omics data for prioritizing cancer genes. Bioinformatics, 34(14), 2441-2448. https://doi.org/10.1093/bioinformatics/bty148

Feldman, E. L., Goutman, S. A., Petri, S., Mazzini, L., Savelieff, M. G., Shaw, P. J., & Sobue, G. (2022). Amyotrophic lateral sclerosis. Lancet (London, England), 400(10360), 1363-1380. https://doi.org/10.1016/S0140-6736(22)01272-7

Gao, J., Chang, M. T., Johnsen, H. C., Gao, S. P., Sylvester, B. E., Sumer, S. O., Zhang, H., Solit, D. B., Taylor, B. S., Schultz, N., & Sander, C. (2017). 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Medicine, 9*(1), 4. https://doi.org/10.1186/s13073-016-0393-x

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), 7821-7826. https://doi.org/10.1073/pnas.122653799

Goutman, S. A., Hardiman, O., Al-Chalabi, A., Chió, A., Savelieff, M. G., Kiernan, M. C., & Feldman, E. L. (2022). Emerging insights into the complex genetics and pathophysiology of amyotrophic lateral sclerosis. The Lancet. Neurology, 21(5), 465-479. https://doi.org/10.1016/S1474-4422(21)00414-2

Grazioli, G., Martin, R. W., & Butts, C. T. (2019). Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods. Frontiers in molecular biosciences, 6, 42. https://doi.org/10.3389/fmolb.2019.00042

Hamad, N., Watanabe, H., Uchihashi, T., Kurokawa, R., Nagata, T., & Katahira, M. (2020). Direct visualization of the conformational change of FUS/TLS upon binding to promoter-associated non-coding RNA. Chemical Communications, 56(64), 9134-9137. https://doi.org/10.1039/d0cc03776a

Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E. M., Logroscino, G., Robberecht, W., Shaw, P. J., Simmons, Z., & van den Berg, L. H. (2017). Amyotrophic lateral sclerosis. *Nature Reviews. Disease Primers, 3*, 17071. https://doi.org/10.1038/nrdp.2017.71

Hart, P.J., Liu, H., Pellegrini, M., Nersissian, A.M., Gralla, E.B., Valentine, J.S., Eisenberg, D. (1997). FAMILIAL ALS MUTANT G37R CUZNSOD (HUMAN). https://doi.org/10.2210/pdb1azv/pdb

Hicks, M., Bartha, I., di Iulio, J., Venter, J. C., & Telenti, A. (2019). Functional characterization of 3D protein structures informed by human genetic diversity. Proceedings of the National Academy of Sciences, 116(18), 8960-8965. https://doi.org/10.1073/pnas.1820813116

Hough, M. A., Grossmann, J. G., Antonyuk, S. V., Strange, R. W., Doucette, P. A., Rodriguez, J. A., ... & Hasnain, S. S. (2004b). Dimer destabilization in superoxide dismutase may result in disease-causing properties: structures of motor neuron disease mutants. Proceedings of the National Academy of Sciences, 101(16), 5976-5981. https://doi.org/10.1073/PNAS.0305143101

Hough, M.A., Grossmann, J.G., Antonyuk, S.V., Strange, R.W., Doucette, P.A., Rodriguez, J.A., Whitson, L.J., Hart, P.J., Hayward, L.J., Valentine, J.S., Hasnain, S.S. (2004a). I113T mutant of human SOD1. https://doi.org/10.2210/pdb1uxl/pdb

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., ... & Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic acids research*, 37(suppl_1), D211-D215. https://doi.org/10.1093/nar/gkn785

Ivanov, Y. D., Taldaev, A., Lisitsa, A. V., Ponomarenko, E. A., & Archakov, A. I. (2022). Prediction of monomeric and dimeric structures of CYP102A1 using AlphaFold2 and AlphaFold multimer and assessment of point mutation effect on the efficiency of intra-and interprotein electron transfer. Molecules, 27(4), 1386. https://doi.org/10.3390%2Fmolecules27041386

Jin, B., Capra, J. A., Benchek, P., Wheeler, N., Naj, A. C., Hamilton-Nelson, K. L., Farrell, J. J., Leung, Y. Y., Kunkle, B., Vadarajan, B., Schellenberg, G. D., Mayeux, R., Wang, L., Farrer, L. A., Pericak-Vance, M. A., Martin, E. R., Haines, J. L., Crawford, D. C., & Bush, W. S. (2022). An association test of the spatial distribution of rare missense variants within protein structures identifies Alzheimer's disease–related patterns. *Genome Research, 32*(4), 778-790. https://doi.org/10.1101/gr.276069.121

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2

Kenna K, Hop P (2023). rvat: Rare variant analysis toolkit. R package version 0.2.0.

Kumar, S., Clarke, D., & Gerstein, M. B. (2019). Leveraging protein dynamics to identify cancer mutational hotspots using 3D structures. Proceedings of the National Academy of Sciences, 116(38), 18962-18970. https://doi.org/10.1073/pnas.1901156116

Lattante, S., Rouleau, G. A., & Kabashi, E. (2013). TARDBP and FUS mutations associated with amyotrophic lateral sclerosis: summary and update. *Human mutation*, 34(6), 812-826. https://doi.org/10.1002/humu.22319

Lee, S., Abecasis, G., Boehnke, M., & Lin, X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American Journal of Human Genetics, 95*(1), 5-23. https://doi.org/10.1016/j.ajhg.2014.06.009

Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., ... & Raphael, B. J. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics, 47(2), 106-114. https://doi.org/10.1038/ng.3168

Lim, L., Wei, Y., Lu, Y., & Song, J. (2016). ALS-causing mutations significantly perturb the self-assembly and interaction with nucleic acid of the intrinsically disordered prion-like domain of TDP-43. PLoS biology, 14(1), e1002338. https://doi.org/10.1371/journal.pbio.1002338

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., & Lin, X. (2019). ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3), 410-421. https://doi.org/10.1016/j.ajhg.2019.01.002

Loehlein Fier, H., Prokopenko, D., Hecker, J., Cho, M. H., Silverman, E. K., Weiss, S. T., Tanzi, R. E., & Lange, C. (2017). On the association analysis of genome-sequencing data: A spatial clustering approach for partitioning the entire genome into nonoverlapping windows. *Genetic Epidemiology, 41*(4), 332-340. https://doi.org/10.1002/gepi.22040

Loughlin, F. E., & Wilce, J. A. (2019). TDP-43 and FUS–structural insights into RNA recognition and self-association. Current Opinion in Structural Biology, 59, 134-142. https://doi.org/10.1016/j.sbi.2019.07.012

Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A. G., Barnes, I., ... & Flicek, P. (2023). Ensembl 2023. *Nucleic acids research*, 51(D1), D933-D941. https://doi.org/10.1093/nar/gkac958

Martinez-Ledesma, E., Flores, D., & Trevino, V. (2020). Computational methods for detecting cancer hotspots. Computational and Structural Biotechnology Journal, 18, 3567-3576. https://doi.org/10.1016/j.csbj.2020.11.020

McBride, J. M., Polev, K., Reinharz, V., Grzybowski, B. A., & Tlusty, T. (2022). AlphaFold2 can predict single-mutation effects on structure and phenotype. *bioRxiv*, 2022-04. https://www.biorxiv.org/content/10.1101/2022.04.14.488301v2.full

Necci, M., Piovesan, D., Dosztányi, Z., & Tosatto, S. C. (2017). MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*, 33(9), 1402-1404. https://doi.org/10.1093/bioinformatics/btx015

Niu, B., Scott, A. D., Sengupta, S., Bailey, M. H., Batra, P., Ning, J., Wyczalkowski, M. A., Liang, W., Zhang, Q., McLellan, M. D., Sun, S. Q., Tripathi, P., Lou, C., Ye, K., Mashl, R. J., Wallis, J., Wendl, M. C., Chen, F., & Ding, L. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nature Genetics, 48*(8), 827-837. https://doi.org/10.1038/ng.3586

Patel, A., Lee, H. O., Jawerth, L., Maharana, S., Jahnel, M., Hein, M. Y., ... & Alberti, S. (2015). A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. Cell, 162(5), 1066-1077. https://doi.org/10.1016/j.cell.2015.07.047

Perrakis, A., & Sixma, T. K. (2021). AI revolutions in biology: The joys and perils of AlphaFold. EMBO reports, 22(11), e54046. https://doi.org/10.15252/embr.202154046

Project MinE ALS Sequencing Consortium (2018). Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *Eur J Hum Genet*, 26, 1537–1546. https://doi.org/10.1038/s41431-018-0177-4

Punta, M., Simon, I., & Dosztányi, Z. (2015). Prediction and analysis of intrinsically disordered proteins. Structural Proteomics: High-Throughput Methods, 35-59. https://doi.org/10.1007/978-1-4939-2230-7_3

RCSB Protein Data Bank (RCSB.org) (2023). delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research*, 51, D488–D508. https://doi.org/10.1093/nar/gkac1077

Reynisdottir, T., Anderson, K. J., Boukas, L., & Bjornsson, H. T. (2022). Missense variants causing Wiedemann-Steiner syndrome preferentially occur in the KMT2A-CXXC domain and are accurately classified using AlphaFold2. *PLoS genetics*, 18(6), e1010278. https://doi.org/10.1371/journal.pgen.1010278

Ruff, K. M., & Pappu, R. V. (2021). AlphaFold and implications for intrinsically disordered proteins. Journal of Molecular Biology, 433(20), 167208. https://doi.org/10.1016/j.jmb.2021.167208

Saint Pierre, A., & Génin, E. (2014). How important are rare variants in common disease?. *Briefings in functional genomics*, 13(5), 353-361. https://doi.org/10.1093/bfgp/elu025

Saldaño, T., Escobedo, N., Marchetti, J., Zea, D. J., Mac Donagh, J., Velez Rueda, A. J., ... & Parisi, G. (2022). Impact of protein conformational diversity on AlphaFold predictions. Bioinformatics, 38(10), 2742-2748. https://doi.org/10.1093/bioinformatics/btac202

Santamaria, N., Alhothali, M., Alfonso, M. H., Breydo, L., & Uversky, V. N. (2017). Intrinsic disorder in proteins involved in amyotrophic lateral sclerosis. Cellular and Molecular Life Sciences, 74, 1297-1318. https://doi.org/10.1007/s00018-016-2416-6

Seera, S., & Nagarajaram, H. A. (2022). Effect of disease causing missense mutations on intrinsically disordered regions in proteins. Protein and Peptide Letters, 29(3), 254-267. https://doi.org/10.2174/0929866528666211126161200

Sivley, R. M., Dou, X., Meiler, J., Bush, W. S., & Capra, J. A. (2018). Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures. *American Journal of Human Genetics, 102*(3), 415-426. https://doi.org/10.1016/j.ajhg.2018.01.017

Strange, R. W., Antonyuk, S., Hough, M. A., Doucette, P. A., Rodriguez, J. A., Hart, P. J., ... & Hasnain, S. S. (2003a). The structure of holo and metal-deficient wild-type human Cu, Zn superoxide dismutase and its relevance to familial amyotrophic lateral sclerosis. Journal of molecular biology, 328(4), 877-891. https://doi.org/10.1016/S0022-2836(03)00355-3

Strange, R.W., Antonyuk, S., Hough, M.A., Doucette, P., Rodriguez, J., Hart, P.J., Hayward, L.J., Valentine, J.S., Hasnain, S.S. (2003b). The Structure of Apo Type Human Cu, Zn Superoxide Dismutase. https://doi.org/10.2210/pdb1hl4/pdb

Strange, R.W., Antonyuk, S., Hough, M.A., Doucette, P., Rodriguez, J., Hart, P.J., Hayward, L.J., Valentine, J.S., Hasnain, S.S. (2003c). The Structure of Holo Type Human Cu, Zn Superoxide Dismutase. https://doi.org/10.2210/pdb1hl5/pdb

Tang, Z., Sliwoski, G. R., Chen, G., Jin, B., Bush, W. S., Li, B., & Capra, J. A. (2020). PSCAN: Spatial scan tests guided by protein structures improve complex disease gene discovery and signal variant detection. *Genome Biology, 21*(1), 217. https://doi.org/10.1186/s13059-020-02121-0

Thomasen, F. E., & Lindorff-Larsen, K. (2022). Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. Biochemical Society Transactions, 50(1), 541-554. https://doi.org/10.1042/BST20210499

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D. M., Kim, R., Ryan, M., Masica, D., & Karchin, R. (2016). Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Research, 76*(13), 3719-3731. https://doi.org/10.1158/0008-5472.CAN-15-3190

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., ... & Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873), 590-596. https://doi.org/10.1038/s41586-021-03828-1

Van Deerlin, V. M., Leverenz, J. B., Bekris, L. M., Bird, T. D., Yuan, W., Elman, L. B., ... & Yu, C. E. (2008). TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: a genetic and histopathological analysis. *The Lancet Neurology*, 7(5), 409-416. https://doi.org/10.1016/S1474-4422(08)70071-1

Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., ... & Babu, M. M. (2014). Classification of intrinsically disordered regions and proteins. Chemical reviews, 114(13), 6589-6631. https://doi.org/10.1021/cr400525m

Van Dongen, S.M. (2000). Graph clustering by flow simulation [Phd thesis, Utrecht University] https://dspace.library.uu.nl/handle/1874/848

Van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., ... & Kurth, I. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature genetics*, 48(9), 1043-1048. https://doi.org/10.1038/ng.3622

Van Rheenen, W., van der Spek, R. A., Bakker, M. K., van Vugt, J. J., Hop, P. J., Zwamborn, R. A., ... & Mathers, S. (2021). Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nature genetics*, 53(12), 1636-1648. https://doi.org/10.1038/s41588-021-00973-1

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1), D439-D444. https://doi.org/10.1093/nar/gkab1061

Willemse, S. W., Harley, P., van Eijk, R. P., Demaegd, K. C., Zelina, P., Pasterkamp, R. J., ... & van Es, M. A. (2023). UNC13A in amyotrophic lateral sclerosis: from genetic association to therapeutic target. Journal of Neurology, Neurosurgery & Psychiatry. https://doi.org/10.1136/jnnp-2022-330504

Yamashita, S. & Ando, Y. (2015). Genotype-phenotype relationship in hereditary amyotrophic lateral sclerosis. Translational neurodegeneration, 4(1), 1-13. https://doi.org/10.1186/s40035-015-0036-y

Yang, Z., Zeng, X., Zhao, Y., & Chen, R. (2023). AlphaFold2 and its applications in the fields of biology and medicine. Signal Transduction and Targeted Therapy, 8(1), 115. https://doi.org/10.1038/s41392-023-01381-z

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., ... & Ferretti, V. (2019). The international cancer genome consortium data portal. Nature biotechnology, 37(4), 367-369. https://doi.org/10.1038/s41587-019-0055-9

Zhou, B., Wang, H., Cai, Y., Wen, H., Wang, L., Zhu, M., ... & Hong, D. (2020). FUS P525L mutation causing amyotrophic lateral sclerosis and movement disorders. *Brain and behavior*, 10(6), e01625. https://doi.org/10.1002/brb3.1625

Zonneveld, T. (2022). Who bears the burden? Rare-variant burden testing in sub-gene units to identify ALS hotspots (Unpublished master's thesis). Utrecht University, Utrecht, the Netherlands.

Zou, Z. Y., Zhou, Z. R., Che, C. H., Liu, C. Y., He, R. L., & Huang, H. P. (2017). Genetic epidemiology of amyotrophic lateral sclerosis: a systematic review and meta-analysis. *Journal of Neurology*, Neurosurgery & Psychiatry, 88(7), 540-549. https://doi.org/10.1136/jnnp-2016-315018

6. SUPPLEMENTARY MATERIAL

6.1 Supplementary methods

Supplementary methods 1. Algorithm for Randomised protein structures.

The algorithm used to obtain randomised structures randomly determines new 3D coordinates for each backbone atom (N, Cα and C). In order to find the new positions, the spherical coordinate system, with radius coordinates *r* and angle coordinates ϑ (value between 0 and π) and ϕ (value between 0 and 2π) is used (Steiner, 2008, p.295). In this coordinate system, the radius is based on atom distances in the provided AlphaFold2 model (Jumper et al., 2021). With the spherical coordinates, and the cartesian coordinates of the previous atom, standard conversion formulas for for spherical and cartesian coordinate systems are used to determine the position of each subsequent atom (Formula I).

$$x_i = x_{i-1} + r_i \sin(\theta_i)\cos(\phi_i)$$

$$y_i = y_{i-1} + r_i \sin(\theta_i)\sin(\phi_i)$$

$$z_i = z_{i-1} + r_i \cos(\theta_i)$$

Formula I. Conversion between spherical and cartesian coordinate system. To determine the position of the next atom based distance to previous position (radius), these conversion formulas are used. The position of the previous atom is included in the formulas since the general conversion formulas for spherical to cartesian coordinates assume that the origin is used as middle point, or previous position (Steiner, 2008, p.295).

The first atom (*i = 1*) is always positioned at *(x_1 = 1, y_1 = 1, z_1 = 1)*. The second atom (*i = 2*) is calculated using one random θ_2 value and one φ_2 random value within Formula I. All subsequent points (*i > 2*) are determined by iterating over a couple of steps until the output structure has the same length as the input AlphaFold2 structure. Similar as before, θ_i and φ_i values are used as input to Formula I to obtain the cartesian coordinates of the next position (*i*). Different from before, a wide range of these values are used (ranges with step size of 0.1 within above defined ranges). This results in 2016

different combinations of angle values and therefore 2016 possible positions for atom *i.* The resulting list of potential new positions is filtered using bond angle and atom distance criteria.

New positions have to be at least 2 Angstrom from other atoms (excluding the previous atom) to prevent overlapping atoms and bonds. Furthermore, bond angles (calculated based on positions *i, i-1* and *i-2*) have to match experimentally observed results. If the next atom is a N, this angle is ~110° while the angle should be ~120° when the next atom corresponds to a C or Cα (Pauling et al., 1951; Engh & Huber, 1991). To account for observed angle variety, the calculated angles should be within a 5° range of these values. To account for the planar structure characteristic of proteins, an additional step is performed after filtering on general angles has occurred when the next atom corresponds to a Cα atom. Most amino acids are in *trans*-conformation, this means that we take the new Cα position with maximal distance to the previous Cα atom. A small probability (5.4% for Proline and 0.1% for all other residues) for *cis*-conformation is taken into account based on observations by Joseph et al. (2012). In *cis*-conformation. Only the position with minimal distance to previous Cα atom is chosen. This does not always lead to a maximal planar structure. Since multiple studies show that proteins can deviate quite a bit from this characteristic structure (Matthews, 2016), this is still sufficiently close to known limits of protein structures while taking computational efficiency into account.

From those potential positions that remain, a random new position is chosen and the algorithm continues to the next iteration (*i = i+1)*. When no potential position remains, the algorithm removes the position determined from the past 3 atoms and starts over from this point.

References

Engh, R.A. & Huber, R. (1991). Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement. Acta Cryst, A47, 392-400 https://doi.org/10.1107/S0108767391001071

Joseph, A.P., Srinivasan, N. & de Brevern, A.G. (2012). Cis-trans peptide variations in structurally similar proteins. Amino Acids. 43, 1369-1381 https://doi.org/10.1007/s00726-011-1211-9

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. https://doi.org/0.1038/s41586-021-03819-2

Matthews, B.W. (2016). How planar are planar peptide bonds? Protein Sci. 24(4), 776-777. https://doi.org/10.1002/pro.2901

Pauling, L., Corey, R.B. & Branson, H.R. (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Chemistry, 37(4), 205-211 https://doi.org/10.1073/pnas.37.4.205

Steiner, E. (2008), The Chemistry Maths Book (2nd edition). Oxford University Press, p.295.

6.2 Supplementary tables

Supplementary table 1. Canonical transcript and corresponding identifiers that were obtained from Ensembl and used to find the protein structure of interest from AlphaFold2.

Protein Name / symbol	UniProt ID	Transcript ID	Transcript ID version	AlphaFold2 download
SOD1	P00441	ENST00000270142	ENST00000270142.11	January 1st 2023
FUS	P35637	ENST00000254108	ENST00000254108.12	February 1st 2023
TARDBP	Q13148	ENST00000240185	ENST00000240185.8	April 5th 2023
KIF4A	O95239	ENST00000374403	ENST00000374403.4	May 5th 2023
UTP14C	Q5TAP6	ENST00000521776	ENST00000521776.2	May 5th 2023
TTC3	E9PMP8	ENST00000418766	ENST00000418766.6	May 5th 2023
UNC13C	Q8NB66	ENST00000260323	ENST00000260323.16	May 5th 2023
RAC1	P63000	ENST00000348035	ENST00000348035.9	July 3th 2023
MAP2K1	Q02750	ENST00000307102	ENST00000307102.10	July 3th 2023

Supplementary table 2. Structures obtained from the RCSB. The corresponding PDB ID and references of each structure is summarised in the table. The structures were aligned to each other and AlphaFold Google Colab predictions before they were downloaded.

Protein	Type	Source	PDB ID	References
SOD1	Wildtype	AlphaFold2	AF_AFP00441F1	Jumper et al., 2021; Varadi et al., 2022
	Wildtype apo	Experimental	1HL4	Strange et al., 2003b; Strange et al., 2003a
	Wildtype holo	Experimental	1HL5	Strange et al., 2003c; Strange et al., 2003a
	Mutant A5V	Experimental	1N19	Cardoso et al., 2002a; Cardoso et al., 2002b
	Mutant I114T	Experimental	1UXL	Hough et al., 2004a; Hough et al., 2004b;
	Mutant G38R	Experimental	1AZV	Hart et al., 1997;
FUS	Wildtype	AlphaFold2	AF_AFP35637F1	Jumper et al., 2021; Varadi et al., 2022
TARDBP	Wildtype	AlphaFold2	AF_AFQ13148F1	Jumper et al., 2021; Varadi et al., 2022
CFAP410	Wildtype	AlphaFold2	AF_AFO43822F1	Jumper et al., 2021; Varadi et al., 2022

Supplementary table 3. Number of reference sequences that AlphaFold Colab takes into account during each prediction. The number of reference sequences present in the MSA differs depending on exact structure input. Running the algorithm multiple times for the same sequence did not alter this result.

Protein	Mutation	Unique sequences in MSA
CFAP410	Wildtype	12740
	V58L	13218
	R172W	12722
FUS	Wildtype	10587
	P525L	10572
TARDBP	Wildtype	14141
TARDBP	G298S	14130
SOD1	Wildtype	12105
SOD1	A5V	12111
SOD1	D91A	12081

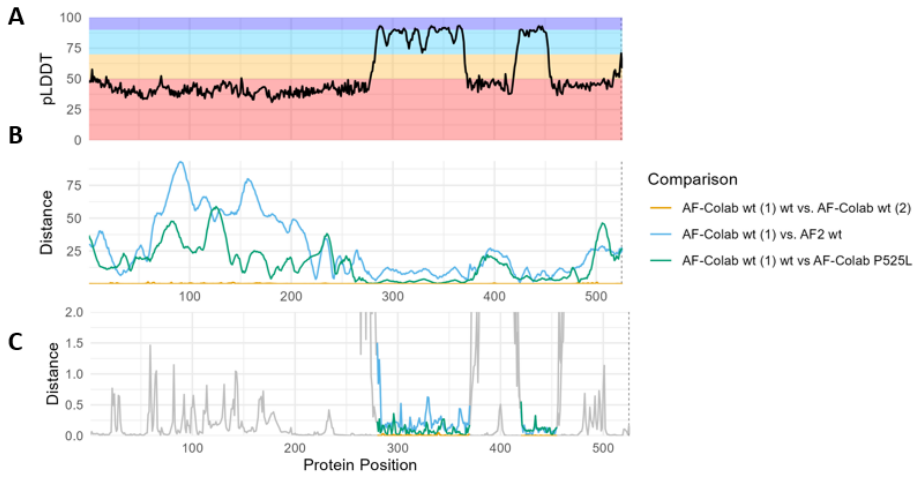
Supplementary table 4. Results of Spearman's correlation test. Correlation test is performed on cluster size versus association test scores from PSCAN analysis. We obtain a p-value and Spearman's rank correlation coefficient (ρ) from the correlation test. A strong correlation is indicated by ρ coefficients close to -1 or +1, while the absence of correlation is indicated by a value close to 0.

	p-value		ρ	
	Rare	Ultrarare	Rare	Ultrarare
SOD1	$5.3x10^{-4}$	$5.0x10^{-6}$	0.5	0.68
FUS	0.01	$3.7x10^{-2}$	0.28	0.26
TARDBP	$2.4x10^{-4}$	$5.0x10^{-5}$	0.67	0.85
KIF4A	$5.5x10^{-8}$	$7.6x10^{-10}$	0.58	0.74
UTP14C	$5.6x10^{-4}$	$1.6x10^{-5}$	0.30	0.49
TTC3	$2.1x10^{-3}$	$7.8x10^{-11}$	0.25	0.61
UNC13C	$4.5x10^{-16}$	$1.6x10^{-5}$	0.38	0.52
RAC1	$1.4x10^{-12}$	-	0.8	-
MAP2K1	$1.8x10^{-17}$	-	0.75	-

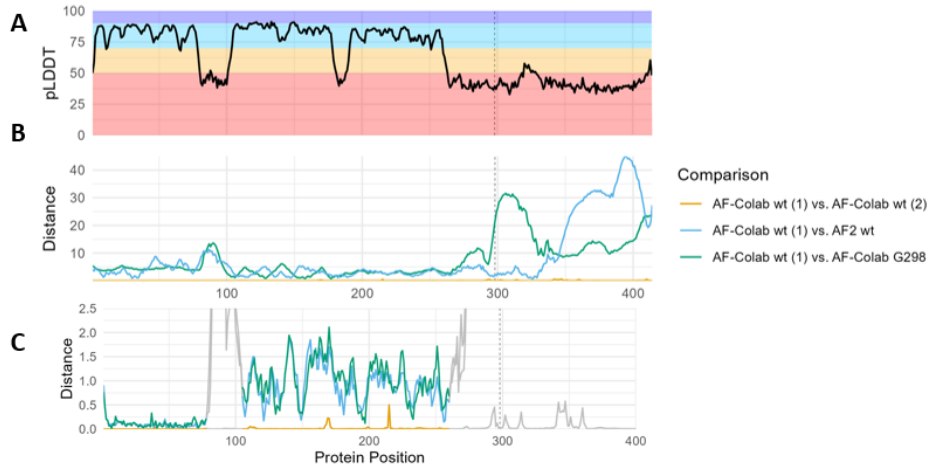
Supplementary table 5. Results of permutation test. Based on the permutation test, a p-values was calculated by dividing the number of iterations that managed to obtained higher -log(P) values than the potential hotspot results and by the total number of permutation iterations (N=100).

	PSCAN		Spherical	
	Rare	**Ultrarare**	**Rare**	**Ultrarare**
SOD1	1	1	1	0.23
FUS	0.01	0.01	0.01	0.77
TARDBP	0.08	0.18	0.39	1
KIF4A	0.56	0.65	0.27	1
UTP14C	0.4	0.07	0.94	0.69
TTC3	0.01	0.34	0.94	0.97
UNC13C	0.09	0.72	1	1
RAC1	1	-	1	-
MAP2K1	1	-	0.37	-

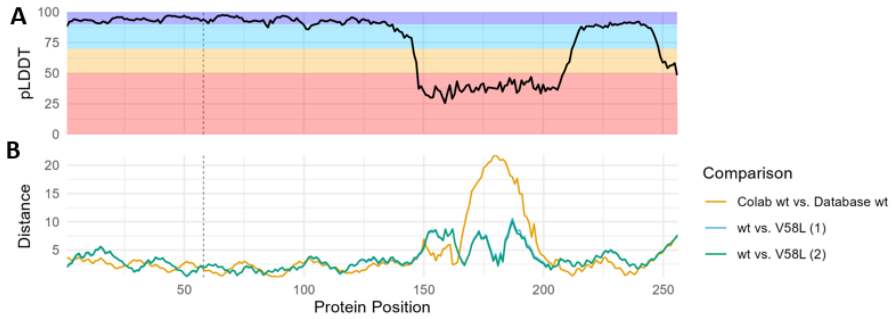
6.3 Supplementary figures



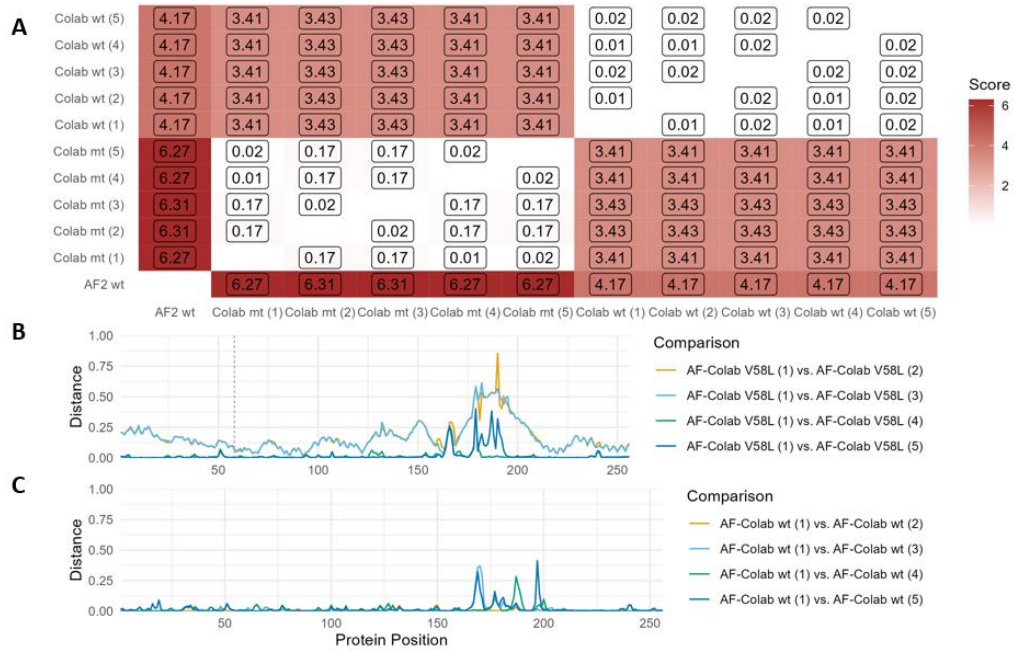
Supplementary Figure 1. Comparison between AlphaFold2 structures of *FUS*. The confidence score (pLDDT) of the wildtype protein structure prediction (A) and the deformation scores between structure pairs (BC). The deformation scores are based on full structure alignments (B) and localised (only folded regions) alignments to filter out the deformation caused by disordered regions (C). The deformation score is the distance at each position between aligned structures. The position of the mutation is displayed by a vertical line.



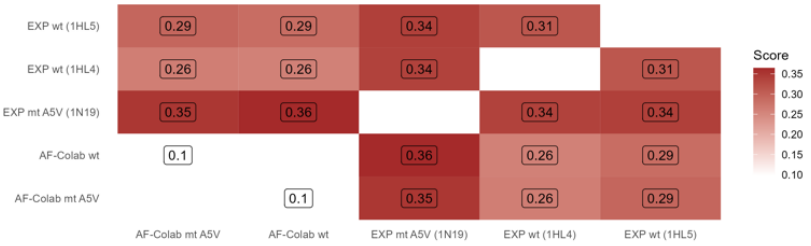
Supplementary Figure 2. Comparison between AlphaFold2 structures of *TARDBP*. The confidence score (pLDDT) of the wildtype protein structure prediction (A) and the deformation scores between structure pairs (BC). The deformation scores are based on full structure alignments (B) and localised (only folded regions) alignments to filter out the deformation caused by disordered regions (C). The deformation score is the distance at each position between aligned structures. The position of the mutation is displayed by a vertical line.



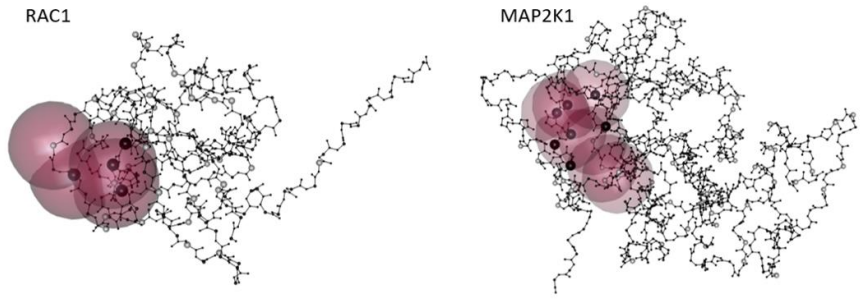
Supplementary Figure 3. Confidence score of *CFAP410* structures. The confidence score (pLDDT) of the wildtype protein structure prediction (A) and the deformation scores between structure pairs. The deformation scores are distances at each position between two aligned structures. The aligned structures are based on full structure alignment (B).



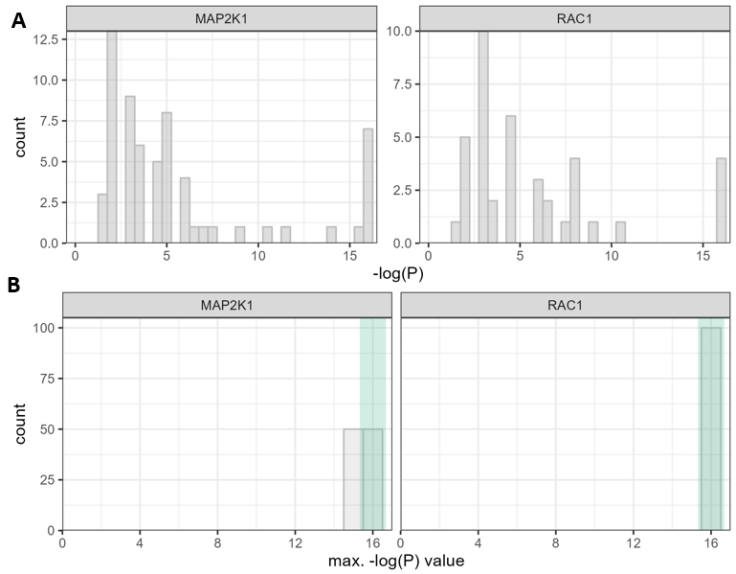
Supplementary Figure 4. Deformation scores for *CFAP410* structures. Pairs of structures are compared using a summary score that represents the mean deformation score. This score is calculated by dividing the cumulative distance between the two structures in a pair by the protein length (A). The deformation score is the distance at each position between the two aligned structures within a pair (BC). Deformation scores are calculated between V58L mutants (B) and wildtypes (C).



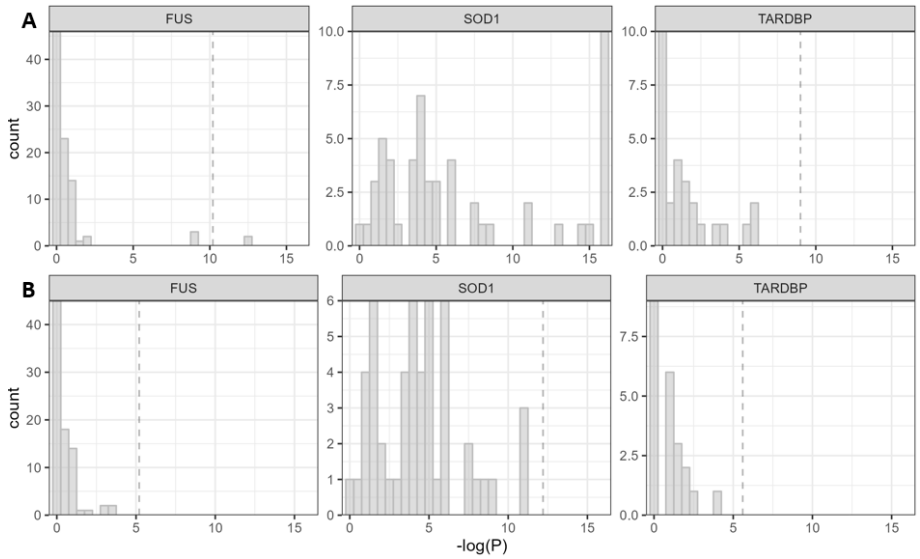
Supplementary Figure 5. Deformation scores for *SOD1* computational and experimental scores. Pairs of structures are compared using a summary score that represents the mean deformation score. The deformation score is the distance at each position between the two aligned structures within a pair. The mean score is calculated by dividing the cumulative distance between the two structures in a pair by the protein length.



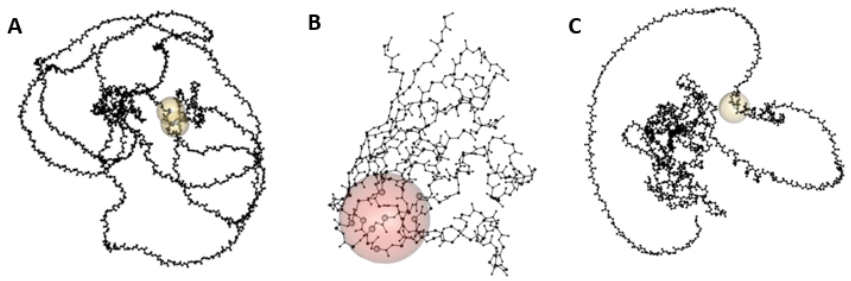
Supplementary Figure 6. Most significant spheres obtained with spherical clustering for *RAC1* and *MAP2K1*. Association with cancer was tested for spheres with a radius of 5Å. Only the spheres with an association score of -log(P) = 16 are displayed. The variant positions of the hotspots identified by Gao et al. (2017) are displayed by small black spheres.



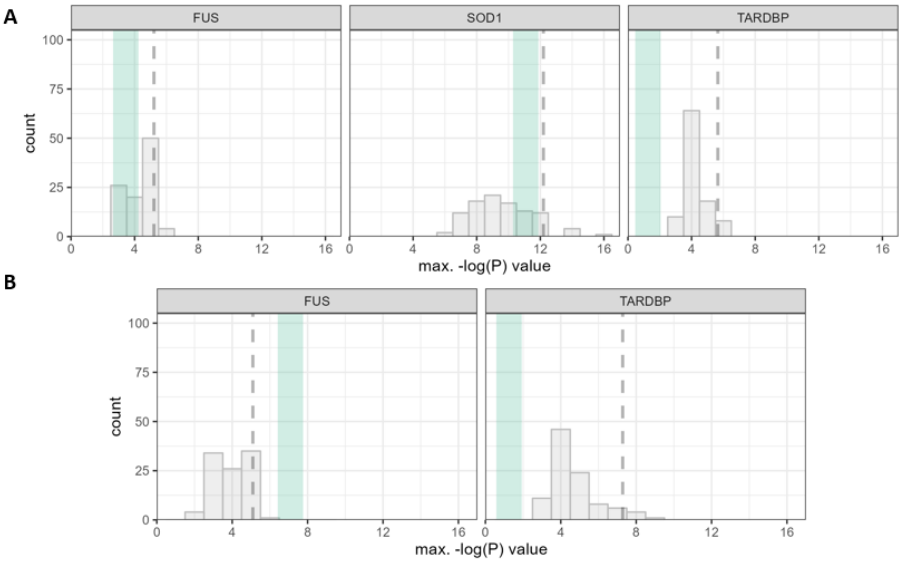
Supplementary Figure 7. Spherical clustering result for *MAP2K1* and *RAC1*. The number of spheres with a radius of 5Å that reach each association score (-log(P)) is represented by grey bars (A). For the permutation test result, the number of iterations in which each association score was the maximal -log(P) value obtained is displayed. The maximal -log(P) value obtained in the non-randomised variant distribution is visualized with a green line (B).



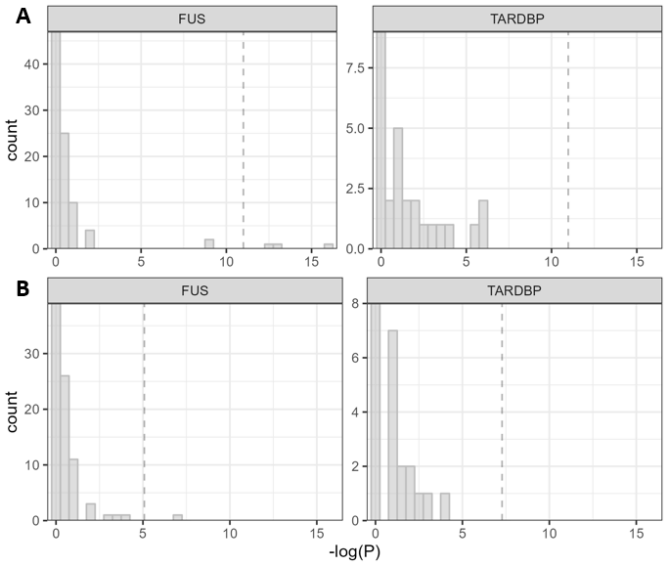
Supplementary Figure 8. Spherical clustering result for *SOD1, FUS* and *TARDBP*. Rare variant (A) and ultrarare variant (B) analysis results obtained for a window size of 5Å. The bars represent the number of clusters that reach each association score (-log(P)). The vertical dashed line represents a significant result for p-value=0.05 according to permutation testing.



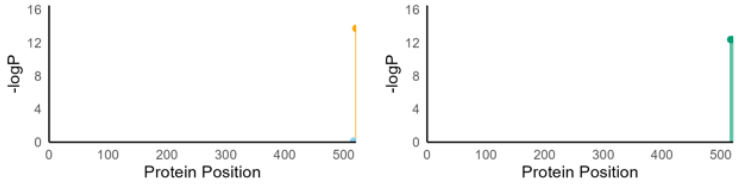
Supplementary Figure 9. Potential hotspot locations predicted by spherical ultrarare variant analysis. For *FUS* and *SOD1* only clusters that reach significance during permutation testing are displayed. For *TARDBP*, no significance is reached and the spheres with the highest -log(P) are displayed instead. The colour of the spheres indicates relative -log(P) values, where dark red represents high values and light yellow represents low values. All spheres have a window size of 5Å.



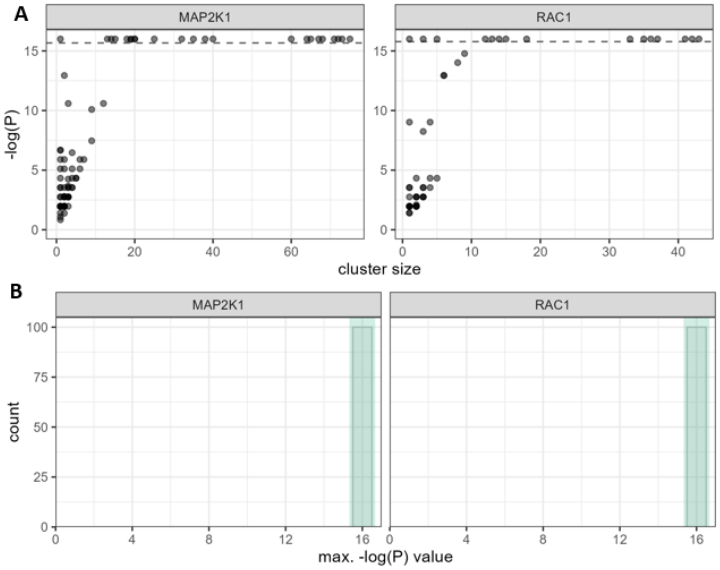
Supplementary Figure 10. Permutation test results for ultrarare variant analysis with spherical clustering. The spheres used during analysis have a radius of 5Å (A) and 10Å (B). The green line represents the highest -log(P) value for the real variant distribution, while grey bars represent the counts of highest -log(P) obtained in each iteration of randomised variant positions. The grey dashed line represents the significant threshold for p-value = 0.05, based on the permutation test. This line is only displayed when the threshold could be calculated.



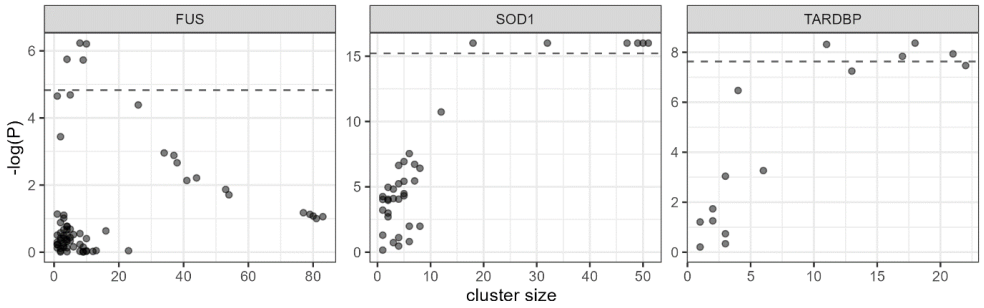
Supplementary Figure 11. Spherical clustering result for *FUS* and *TARDBP*. Rare variant (A) and ultrarare variant (B) analysis results obtained for a window size of 10Å. The bars represent the number of clusters that reach each association score (-log(P)). The vertical dashed line represents a significant result for p-value=0.05 according to permutation testing.



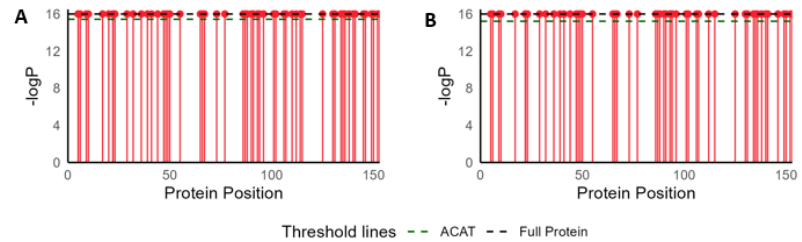
Supplementary Figure 12 Association score of rare variants of the 5Å hotspot in *FUS*. The association score (-log(P)) that is obtained for variants from position 521 (orange), 517 (blue) and the cluster which combines all of these variants (green) is displayed at the their respective positions within the *FUS* protein.



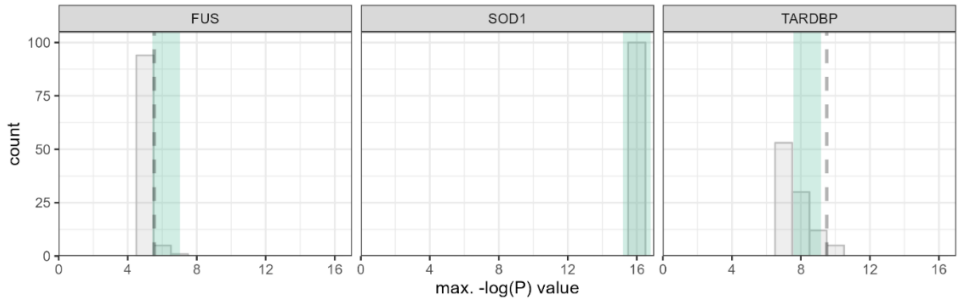
Supplementary Figure 13. Results of PSCAN based analysis for *MAP2K1* and *RAC1*. The size of clusters, i.e. the number of variant positions included within the cluster, and the corresponding association scores (-log(P)) is displayed. A horizontal dashed line is added to show the association score obtained with ACAT-O (A). The permutation test results are summarised by visualizing the maximal association score obtained in each iteration as a histogram with grey bars. The association score of the non-randomised variant distribution is displayed by the green line on top of these permutation test results (B).



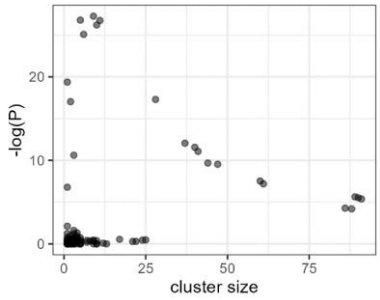
Supplementary Figure 14. Association scores of PSCAN clusters for ultrarare variant analysis. The size of the tested cluster, i.e. the number of variant positions included within the cluster, is plotted against the association score (-log(P)). A horizontal dashed line is added to show the association score obtained with ACAT-O.



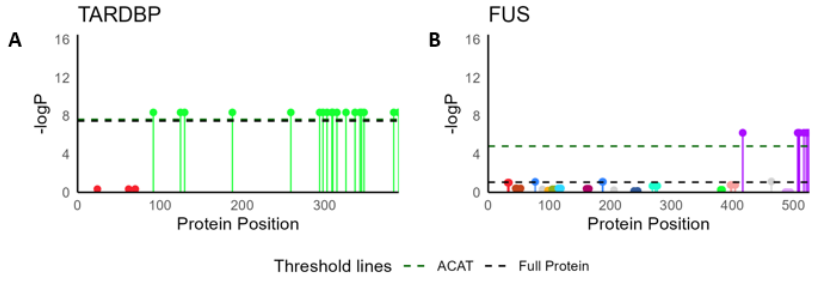
Supplementary Figure 15. Most informative cluster for *SOD1* according to PSCAN analysis. A 1D representation of the variants that make up the most informative cluster for rare variant (A) and ultrarare variant (B) analysis.



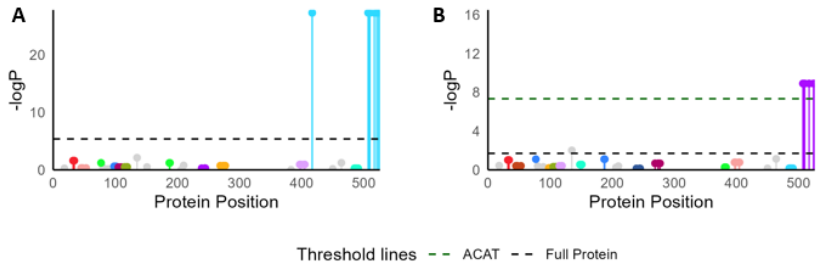
Supplementary Figure 16. Permutation test results for ultrarare variant PSCAN analysis. Grey bars represent a histogram of the maximal association score (-log(P)) in each iteration of the permutation test. A green line is added to the plot to show the -log(P) value of the potential hotspot, i.e. the cluster with the highest -log(P) value obtained in the non-randomised variant distribution. The grey dashed line represents the significant threshold for p-value = 0.05, based on the permutation test. This line is only displayed when the threshold could be calculated.



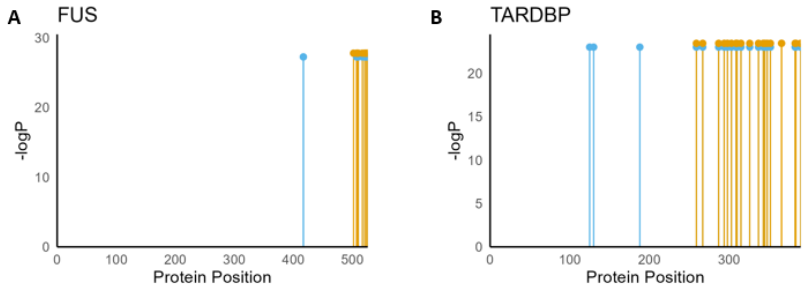
Supplementary Figure 17. Association scores of PSCAN analysis with SKAT burden testing for *FUS*. The clusters obtained in rare variant analysis. The size of the tested cluster, i.e. the number of variant positions that in each cluster, is plotted against the association significance (-log(P)).



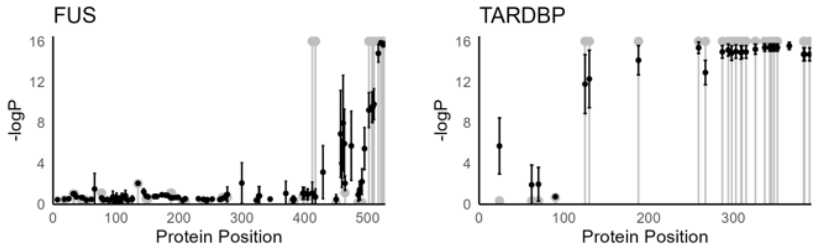
Supplementary Figure 18. Most informative division of ultrarare variant analysis with PSCAN. A 1D representation of the ultrarare variants that make up the most informative clusters for *TARDBP* (A) and *FUS* (B). Each colour represents a separate cluster while grey coloured positions represent positions that have not been clustered. The cluster with the highest level of association (-log(P)) is the predicted hotspot.



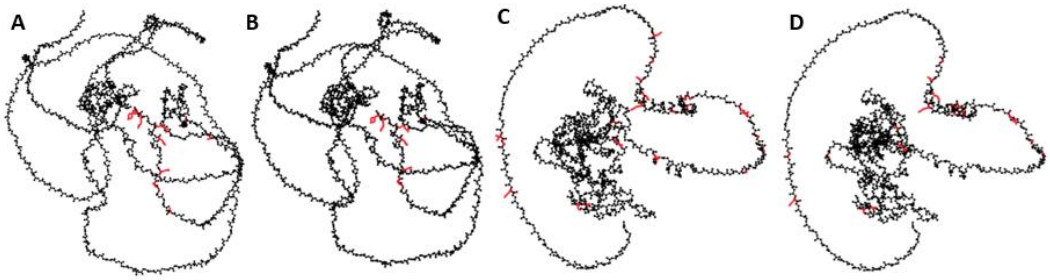
Supplementary Figure 19. Most informative division of PSCAN analysis results for *FUS*. A 1D representation of the rare variants that make up the most informative clusters obtained with SKAT burden testing (A) and firth burden testing on clusters that were made without included position 521 during clustering (B). Each colour represents a separate cluster while grey coloured positions represent positions that have not been clustered. The cluster with the highest level of association (-log(P)) is the predicted hotspot.



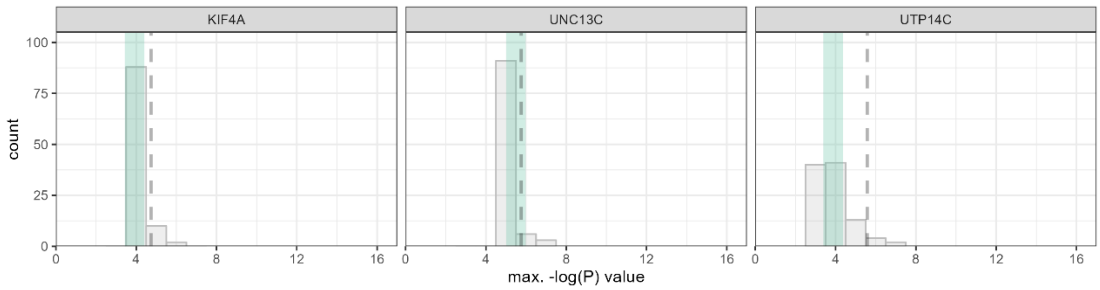
Supplementary Figure 20. Analysis on the contribution of unexpected hotspot variants to hotspot results. The association score of a rare variant cluster with (blue) and without (orange) the unexpected variants is performed using SKAT burden testing. The analysis is performed for both *FUS* (A) and *TARDBP* (B).



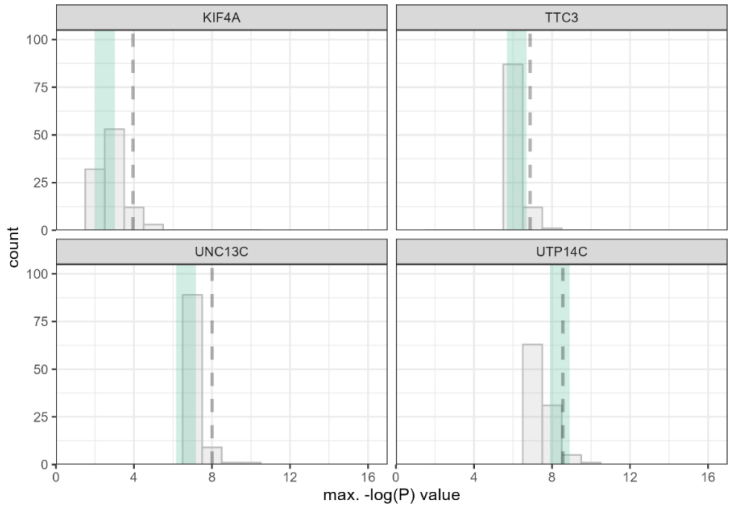
Supplementary Figure 21. Randomised structure analysis PSCAN on rare variants of *FUS* and *TARDBP*. The most informative division obtained with the AF2 models is displayed in grey. The results obtained with randomised structures are displayed with black error bars. Error bars are calculated based on the most association score (-log(P)) obtained in the most informative division of each randomisation iteration. Large error bars generally refer to positions that do not occur in the most informative division after filtering on minor allele count of 5.



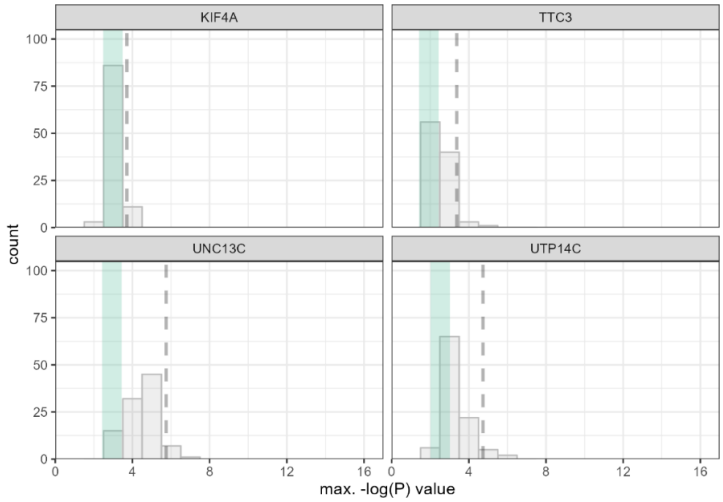
Supplementary Figure 22. Predicted hotspot result with PSCAN analysis. A 3D representation of the potential hotspot obtained with rare variant (AC) and ultrarare variant (BD) analysis for *FUS* (AB) and *TARDBP* (CD). Clusters are displayed on the 3D AlphaFold2 structure of the protein. Only the most significant cluster is displayed in each case.



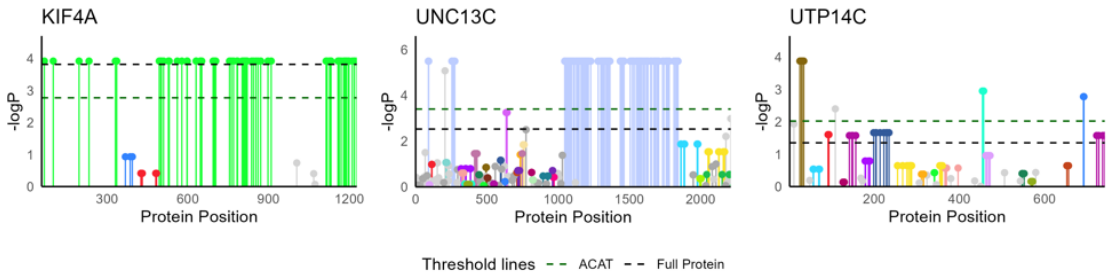
Supplementary Figure 23. Permutation test results for rare variant PSCAN analysis on candidate ALS genes. Grey bars represent a histogram of the maximal association score (-log(P)) in each iteration of the permutation test. A green line is added to the plot to show the -log(P) value of the potential hotspot, i.e. the cluster with the highest -log(P) value obtained in the non-randomised variant distribution. The grey dashed line represents the significant threshold for p-value = 0.05, based on the permutation test. This line is only displayed when the threshold could be calculated.



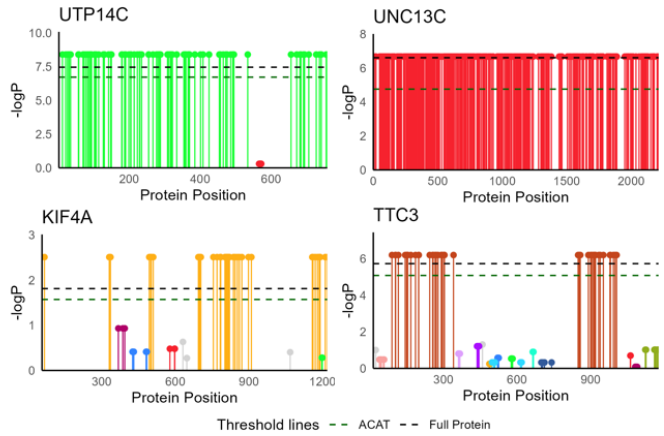
Supplementary Figure 24. Permutation test results for ultrarare variant PSCAN analysis on candidate ALS genes. Grey bars represent a histogram of the maximal association score (-log(P)) in each iteration of the permutation test. A green line is added to the plot to show the -log(P) value of the potential hotspot, i.e. the cluster with the highest -log(P) value obtained in the non-randomised variant distribution. The grey dashed line represents the significant threshold for p-value = 0.05, based on the permutation test. This line is only displayed when the threshold could be calculated.



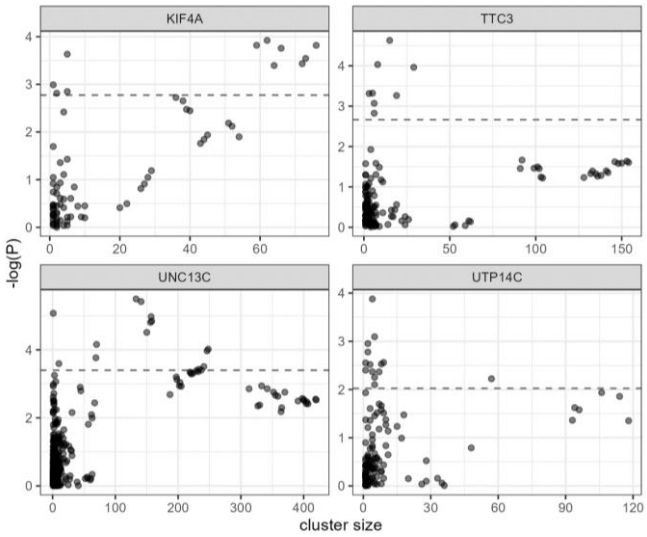
Supplementary Figure 25. Permutation test results for rare variant spherical clustering analysis on candidate ALS genes. Grey bars represent a histogram of the maximal association score (-log(P)) in each iteration of the permutation test. A green line is added to the plot to show the -log(P) value of the potential hotspot, i.e. the cluster with the highest -log(P) value obtained in the non-randomised variant distribution. The grey dashed line represents the significant threshold for p-value = 0.05, based on the permutation test. This line is only displayed when the threshold could be calculated.



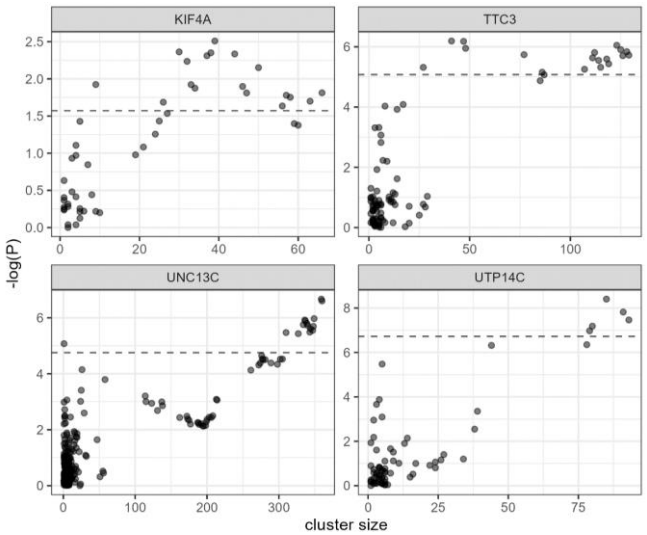
Supplementary Figure 26. Most informative division of rare variant analysis with PSCAN for candidate ALS genes. A 1D representation of the ultrarare variants that make up the most informative clusters for *KIF4A, UNC13C* and *UTP14C*. Each colour represents a separate cluster while grey coloured positions represent positions that have not been clustered. The cluster with the highest level of association (-log(P)) is the predicted hotspot.



Supplementary Figure 27. Most informative division of ultrarare variant analysis with PSCAN for candidate ALS genes. A 1D representation of the ultrarare variants that make up the most informative clusters for *KIF4A, UNC13C, TTC3* and *UTP14C*. Each colour represents a separate cluster while grey coloured positions represent positions that have not been clustered. The cluster with the highest level of association (-log(P)) is the predicted hotspot.



Supplementary Figure 28. Association scores of rare variant PSCAN clusters for candidate ALS genes. The size of the tested cluster, i.e. the number of variant positions included within the cluster, is plotted against the association score (-log(P)). A horizontal dashed line is added to show the association score obtained with ACAT-O.



Supplementary Figure 29. Association scores of ultrarare variant PSCAN clusters for candidate ALS genes. The size of the tested cluster, i.e. the number of variant positions included within the cluster, is plotted against the association score (-log(P)). A horizontal dashed line is added to show the association score obtained with ACAT-O.