Comprehensive in silico analysis of the TDP-43 protein variants related to Amyotrophic Lateral Sclerosis and Frontotemporal Dementia

Abrangente na análise silicoanalítica das variantes proteicas TDP-43 relacionadas à Esclerose Lateral Amiotrófica e Demência Frontotempor

DOI:10.34117/bjdv8n8-190

Recebimento dos originais: 21/06/2022
Aceitação para publicação: 29/07/2022

Juliana Pereira Loureiro[&]
M.Sc
Institution: Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Address: Avenida Pasteur, 296, CEP: 22290-250, Urca, Rio de Janeiro - RJ
E-mail: juliana.loureiro@edu.unirio.br

Gabriel Rodrigues Coutinho Pereira[&]
M.Sc
Institution: Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Address: Avenida Pasteur, 296, CEP: 22290-250, Urca, Rio de Janeiro - RJ
E-mail: gabrielkytz@hotmail.com

Leonardo Cardoso da Silva Bloise
Bacharelado em Ciências Biológicas
Institution: Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Address: Avenida Pasteur, 296, CEP: 22290-250, Urca, Rio de Janeiro - RJ
E-mail: leonardo.bloise2000@gmail.com

José Alexandre de Carvalho Salerno
M.Sc
Institution: Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Address: Avenida Pasteur, 296, CEP: 22290-250, Urca, Rio de Janeiro - RJ
E-mail: josealexandresalerno@gmail.com

Vinicius Abrantes Silvestre
MD
Institution: Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Address: Avenida Pasteur, 296, CEP: 22290-250, Urca, Rio de Janeiro - RJ
E-mail: abrantes_silvestre@hotmail.com

Joelma Freire de Mesquita
Ph.D
Institution: Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Address: Avenida Pasteur, 296, CEP: 22290-250, Urca, Rio de Janeiro - RJ
E-mail: joelma.mesquita@unirio.br

ABSTRACT

Amyotrophic lateral sclerosis (ALS) is a highly disabling neurodegenerative disorder characterized by the progressive loss of voluntary motor activity. ALS is currently the most frequent adult-onset motor neuron disorder, which is associated with a major economic burden. Two drugs have already been approved to treat ALS, but they confer a limited survival benefit. In turn, frontotemporal dementia (FTD) is an early-onset and fatal dementia characterized by deficits in behavior, language, and executive function. FTD is the most frequent cause of pre-senile dementia after Alzheimer's. Currently, FTD has no cure and the available treatments are merely symptomatic. Missense mutations in TDP-43, a nuclear RNA/DNA-binding protein, are among the main causes associated with ALS and FTD. Nonetheless, most of these mutations are not yet characterized. To date, no complete three-dimensional structure has already been determined for TDP-43. In this work, we characterized the impact of missense mutations in TDP-43 using prediction algorithms, evolutionary conservation analysis, and molecular dynamics simulations (MD). We also performed structural modeling and validation of the TDP-43 protein. Two hundred and seven TDP-43 mutations were compiled from the databases and literature. The predictive analysis pointed to a moderate rate of deleterious and destabilizing mutations. Furthermore, most mutations occur at evolutionarily variable positions. Combining the predictive analyses into a penalty system, our findings suggested that the uncharacterized mutations Y43C, D201Y, F211S, I222T, K224N, A260D, P262T, and A321D are considered the most-likely deleterious, thus being important targets for future investigation. This work also provided an accurate, complete, and unprecedented three-dimensional structure for TDP-43 that can be used to identify and optimize potential drug candidates. At last, our MD findings pointed to a noticeable flexibility increase in functional domains upon K263E, G335D, M337V, and Q343R variants, which may cause non-native interactions and impaired TDP-43 recognition, ultimately leading to protein aggregation.

Keywords: Amyotrophic lateral sclerosis, frontotemporal dementia, in silico analysis.

RESUMO

A esclerose lateral amiotrófica (ELA) é um distúrbio neurodegenerativo altamente incapacitante, caracterizado pela perda progressiva da atividade motora voluntária. A ALS é atualmente o distúrbio motor do neurônio adulto mais freqüente, que está associado a uma grande carga econômica. Dois medicamentos já foram aprovados para tratar a ELA, mas eles conferem um benefício de sobrevivência limitado. Por sua vez, a demência frontotemporal (FTD) é uma demência precoce e fatal caracterizada por déficits no comportamento, na linguagem e na função executiva. O FTD é a causa mais freqüente de demência pré-senil após a doença de Alzheimer. Atualmente, o FTD não tem cura e os tratamentos disponíveis são meramente sintomáticos. As mutações de Missense no TDP-43, uma proteína nuclear de ligação RNA/DNA, estão entre as principais causas associadas com ALS e FTD. No entanto, a maioria dessas mutações ainda não são caracterizadas. Até o momento, nenhuma estrutura tridimensional completa já foi determinada para o TDP-43. Neste trabalho, caracterizamos o impacto das mutações de missense no TDP-43 usando algoritmos de previsão, análise de conservação evolutiva, e simulações de dinâmica molecular (MD). Também realizamos a modelagem estrutural e validação da proteína TDP-43. Duzentas e sete mutações TDP-43 foram compiladas a partir dos bancos de dados e da literatura. A análise preditiva apontou para uma taxa moderada de mutações deletérias e desestabilizadoras. Além disso, a maioria das mutações ocorre em posições evolutivamente variáveis. Combinando as análises preditivas em um sistema de penalidades, nossas descobertas sugeriram que as mutações não caracterizadas Y43C, D201Y, F211S, I222T, K224N, A260D, P262T, e A321D são consideradas as mais prováveis deletérias, sendo assim alvos importantes para investigações futuras. Este trabalho também forneceu uma estrutura tridimensional precisa, completa e sem precedentes para o TDP-43 que pode ser usada para identificar e otimizar potenciais candidatos a drogas. Finalmente, nossas descobertas MD apontaram para um notável aumento de flexibilidade nos domínios funcionais nas variantes K263E, G335D, M337V e Q343R, o que pode causar interações não nativas e prejudicar o reconhecimento do TDP-43, levando finalmente à agregação de proteínas.

Palavras-chave: Esclerose lateral amiotrófica, demência frontotemporal, em análise siliciosa

1 INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a highly disabling neurodegenerative disorder that results in injury and death of upper and lower motor neurons. The disease is characterized by the progressive loss of voluntary motor activity, usually leading to death due to respiratory failure within two to five years of symptoms onset. ALS and other neurodegenerative diseases, which lead to patient's inability to work and perform daily activities, demand an average of 47hours/week of supervised care. Most of these hours are provided by family members who often have to give up their regular jobs (MARESOVA *et al*., 2020). It represents a major economic burden, estimated at over $ 1 billion/year for the United States (GLADMAN; DHARAMSHI; ZINMAN, 2014). ALS is currently the most frequent adult-onset motor neuron disorder, affecting 2/100.000 new individuals each year, which is projected to increase by 69% in the next 20 years due to population aging (ARTHUR et al., 2016). Although two drugs have been approved for the treatment of ALS, riluzole, and edaravone, they confer a slight survival benefit of two to three months (JAISWAL, 2018).

Frontotemporal dementia (FTD), in turn, is an early-onset and fatal dementia caused by the degeneration of the frontal and temporal lobes of the brain. The disease is characterized by deficits in behavior, language, and executive function. FTD usually results in the patient's death within 7 to 8 years after the disease onset. Approximately 40% of FTD patients have a familial story of the disease, which indicates a strong genetic contribution. FTD is the second most frequent cause of dementia in people under the age of 65 years (HODGES; PIGUET, 2018), representing a major psychological and economic burden for patients and families, particularly due to caregiver's distress and patient's inability to work (DIEHL-SCHMID et al., 2013). FTD has no cure, and the available treatments are merely symptomatic (HODGES; PIGUET, 2018). Recently, studies have revealed several clinical, pathological, and genetic overlaps between ALS and FTD, suggesting that these disorders are two ends in the spectrum of a single disease (JI et al., 2017).

Transactivation Response (TAR) DNA Binding Protein (TDP-43) is a nuclear RNA/DNA-binding protein involved in non-coding RNA biosynthesis and metabolism, transcription regulation, and splicing (SCOTTER; CHEN; SHAW, 2015). Missense mutations in TDP-43 are among the main causes associated with ALS and FTD (JI et al., 2017). These mutations are believed to lead to pathogenesis through abnormal post-translational modifications, such as phosphorylation or disulfide bond formation, which significantly increases TDP-43's propensity to form toxic aggregate. The accumulation of aggregated cytosolic inclusions of TDP-43 is considered a hallmark of both disorders (PESIRIDIS; LEE; TROJANOWSKI, 2009), being observed in approximately 97% of all sporadic ALS cases and 45% of the FTD cases (SCIALÒ et al., 2020). Interestingly, TDP-43 aggregates are also observed due to mutations in several other ALS/FTD-associated genes, including C9orf72, PFN1, UBQLN2, VCP, and FUS, highlighting the importance of TDP-43 protein in the pathogenesis of ALS and FTD (MARESOVA et al., 2020).

The computer approach, *in silico*, has become widely used to characterize the effects of missense mutations and to identify those potentially deleterious. Prediction algorithms allow filtering among a large dataset, the most probable deleterious mutations to be thoroughly analyzed by wet-lab experiments (PEREIRA; TELLINI; DE MESQUITA, 2019). The *in silico* methods are also important allies of the experimental methods in the field of structural-based drug design, allowing the accurate modeling of three-dimensional protein structures based on their amino-acid sequence at a relatively low-cost (KREBS; DE MESQUITA, 2016). Protein structures are used as targets to identify and optimize potential drug candidates, favoring the development of more effective drugs with fewer side effects and less toxicity (BONETTA et al., 2016).

To date, the complete three-dimensional structure of human TDP-43 protein is still unknown (NCBI, 2017). Therefore, we generated *in silico* an unprecedented, accurate, and complete three-dimensional structure of human TDP-43 protein, a promising pharmacological target for ALS/FTD (OJAIMI et al., 2022). We also characterized the structural and functional impact of missense mutations in TDP-43 using prediction algorithms and molecular dynamics simulations, following the methodology previously established by our group (PEREIRA; ABRAHIM-VIEIRA; DE MESQUITA, 2021; PEREIRA et al., 2022). This approach could contribute to the selection of more appropriate therapeutic interventions for patients with TDP-43 mutations and the design of more effective drugs to treat ALS and FTD. Our findings may also provide relevant information on the structural basis of ALS/FTD, guiding the development of future experiments (ROY CHOUDHURY et al., 2017).

2 MATERIALS AND METHODS

2.1 DATASET

The sequence of wild-type TDP-43 protein was retrieved from the UniProt (UniProt ID: Q13148), while its missense mutations were compiled from the dbSNP, OMIM, ALSOD, and ClinVar databases, in addition to a literature review on PubMed (NCBI, 2017).

2.2 PREDICTIVE ANALYSIS

The TDP-43 sequence and its missense mutations were submitted to prediction algorithms. The algorithms MutPred2, PMut, Fathmm, PolyPhen-2, PROVEAN, PredictSNP, SIFT, SNAP2, PhD-SNP, and SNPs&GO were used for the functional prediction. I-Mutant3.0 was used for the stability prediction. The algorithms WALTZ, TANGO, and LIMBO of SNPEffect4.0, were used for the prediction of amyloid propensity, protein aggregation, and chaperone binding, respectively (PEREIRA; ABRAHIM-VIEIRA; DE MESQUITA, 2021).

2.3 STRUCTURAL MODELING AND VALIDATION

Complete three-dimensional structures of wild-type TDP-43 were predicted using comparative, *ab initio*, and threading modeling strategies in the following algorithms: Rosetta, I-Tasser, and Phyre2 (KELLEY et al., 2015). Standard I-Tasser and Phyre2 modeling options were selected. The Rosettas' models were generated using both ab initio and comparative protocols in Robetta's server. Protein domains were built comparatively using experimental protein structures as the template. *Ab initio* was applied to model low identity regions. These regions were later assembled on a fully automated procedure (PEREIRA et al., 2019).

The generated structures were then refined by the GalaxyREFINE algorithm (HEO; PARK; SEOK, 2013) available online at the GalaxyWEB server, which is a structural refinement algorithm that employs a hybrid-type energy function on a local structure-modeling protocol to refine template-based models (HEO; PARK; SEOK, 2013). The output structures from GalaxyREFINE were further refined by the ModRefiner algorithm, which constructs the main chain from C-alfa trace and side-chain rotamers, and then is finally refined together with the backbone using physics- and knowledge-based force fields (XU; ZHANG, 2011).

The refined models underwent quality assessment using five validation algorithms: PROCHECK, RAMPAGE, VERIFY-3D, QMEAN, and PROSA-Web (PEREIRA et al., 2019). The model with the best quality assessment scores was selected to undergo structural alignment. A Protein BLAST (NCBI, 2017) was performed to select a suitable template structure for the structural alignment in the TM-Align server. The alignment returned root-mean-square deviation (RMSD) and TM-score values, which were used to measure the structural similarity between the theoretical models and experimental fragments of human TDP-43 protein (ZHANG; SKOLNICK, 2005).

To further validate the *in silico* modeled structure of TDP-43, we also performed predictions of secondary structure and intrinsically disordered regions (IDR). The secondary structure of TDP-43 was predicted using the algorithms Stride, Porter, JPred, PsiPred4, YASPIN, SYMPRED, and Scratch (WEI; THOMPSON; FLOUDAS, 2012). The IDRs of TDP-43 were predicted using the algorithms DisPro, Dismeta, DisoPred3 (MAKINO et al., 2014), Scratch, and IntFold4 (MCGUFFIN et al., 2015). The consensus results of the secondary structure and IDR predictions were then compared to the TDP-43 model. Finally, we compared the secondary structure of the theoretical model with the potential template structures identified by ProteinBlast.

2.4 EVOLUTIONARY CONSERVATION ANALYSIS

The validated model was submitted to the ConSurf algorithm (ASHKENAZY et al., 2016), which calculated the evolutionary conservation degree of each amino acid composing TDP-43, through multiple alignments of homolog sequence, taking into account phylogenetic relations between aligned proteins. The homologs were collected from the UniProt database using CS-BLAST as the search algorithm, with a maximum of 95% of identity between sequences and a minimum of 35% of identity for homologs. The conservation scores were calculated with the Bayesian method.

We then combined all previous predictive analyses, a literature review, and alterations in physicochemical properties (hydrophobicity, size, and charge) to rank all uncharacterized mutations in TDP-43 according to their deleterious propensity by using a penalty system. Alterations in physicochemical properties were predicted using the HOPE algorithm (VENSELAAR et al., 2010).

2.5 MOLECULAR DYNAMICS SIMULATIONS

In silico mutagenesis was performed using the mutator plugin of the VMD 1.9.3 software. The mutations K263E, G335D, M337V, and Q343R were individually induced in the validated model of wild-type TDP-43. Molecular Dynamics (MD) simulations were performed using the software GROMACS package version 5.0.7 for both wild-type and variants structures. Following the methodology previously described by our group (PEREIRA; ABRAHIM-VIEIRA; DE MESQUITA, 2021; PEREIRA et al., 2022), AMBER99SB-ILDN was selected as the force field of the simulations. The molecules were solvated in a cubic box filled with TIP3P water molecules and neutralized by the addition of Na^+ Cl^- ions. The energy minimization was carried out using the steepest descent method. After system minimization, NVT (constant number, volume, and temperature) equilibration was done with a constant temperature of 300K during 100ps, followed by NPT (constant number, pressure, and temperature) equilibration with a constant pressure of 1atm and a temperature of 300K during 100ps.

After the NVT and NPT ensembles, MD simulations were conducted for 40ns at 300K and 1atm. The MD trajectories were saved every 10ps simulated and analyzed using GROMACS distribution programs. The MD trajectories were analyzed comparatively for the following parameters: RMSD, root-mean-square fluctuation (RMSF), radius of gyration (Rg), number of intramolecular hydrogen bonds, solvent-accessible surface area (SASA), and B-factor. The data visualization was performed using the ggplot2 package in R software and UCSF chimera software.

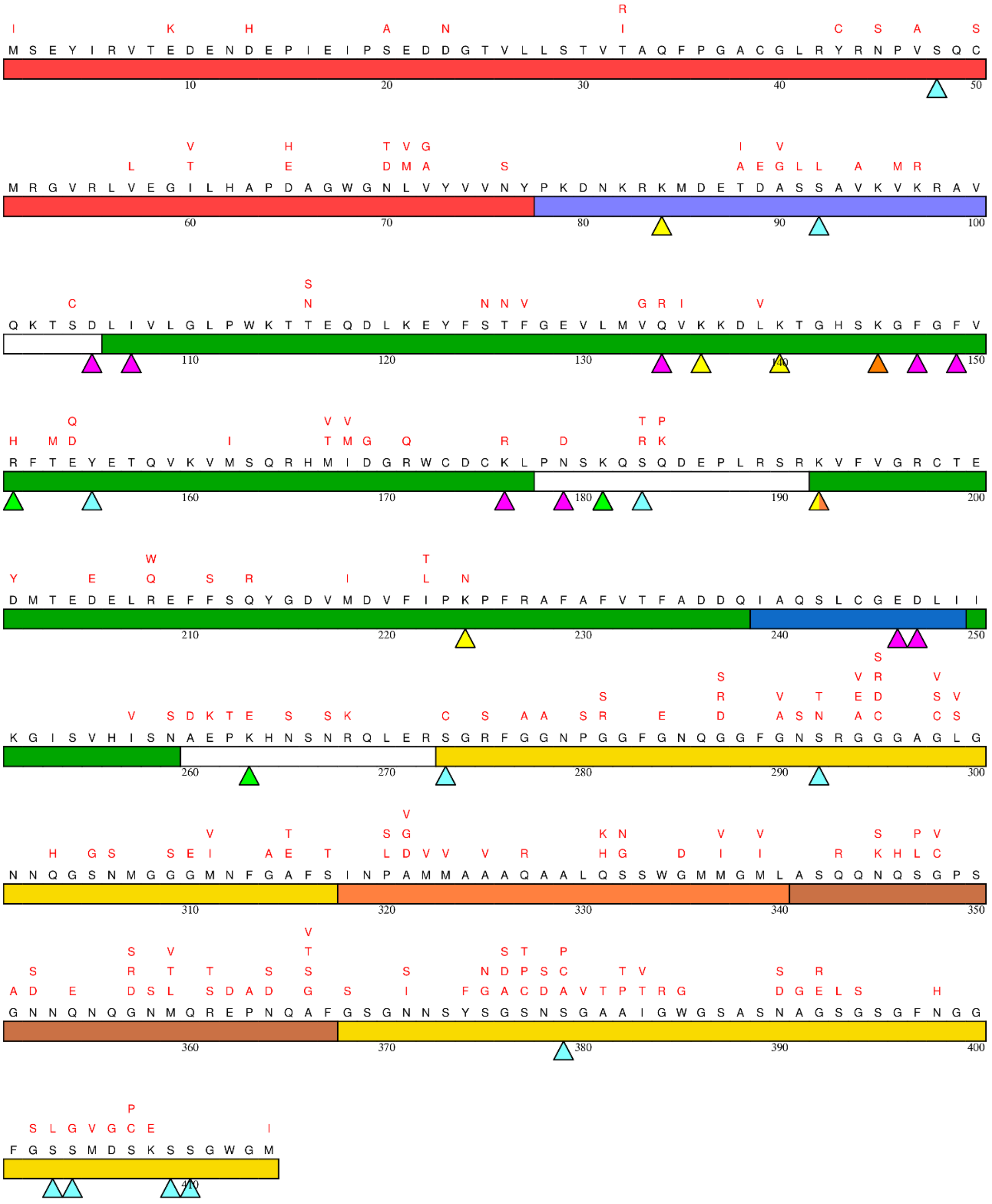
3 RESULTS AND DISCUSSION

3.1 DATASET

The sequence of TDP-43 obtained at the UniProt has 414 amino acids, which corresponds to the complete protein length. As shown in Fig 1, the protein can be divided into functional domains: N-terminal domain (NTD), RNA-recognition motifs (RRM1 and RRM2), and C-terminal domain (CTD) (BERNING; WALKER, 2019). The NTD is located between residues 1-77. This domain is responsible for TDP-43 oligomerization, which is required for protein's splicing activity (JO et al., 2020). This region also contains a key phosphorylation site, Ser48, and one of the three mitochondrial targeting sequences (residues 35-41). Phosphorylation of Ser48 is known to disrupt TDP-43 oligomerization, leading to impaired splicing function. Between the NTD and RRM1, there is a nuclear localization signal (NLS), which is required for the active transport of TDP-43 into the nucleus (residues 78-100). The RRM1 (residues 106-177) and RRM2 (192-259) are involved in nucleic acid binding and contain RNA recognition motifs, highly conserved and short sequences needed for nucleic acid binding. Specific amino acids in the RRM domains, highlighted in Figure 1 by magenta triangles, are known to be directly involved in nucleic acid-binding in such a way that mutations may decrease the specificity for RNA sequences. Among them residues 151, 181, and 263, which are highlighted in the figure by green triangles. RRM2 domain also contains a nuclear export signal (NES), involved in the active nuclear export of TDP-43 (residues 239-249) (FRANÇOIS-MOUTAL et al., 2019).

CTD is a low complexity major intrinsically disordered region (IDR) highly associated with the pathogenesis of TDP-43 protein due to its aggregation-prone (BERNING; WALKER, 2019). CTD is also the region with more phosphorylation sites, as shown in Fig 1 (FRANÇOIS-MOUTAL et al., 2019). Furthermore, CTD is required for TDP-43 splicing activity, autoregulation, and interaction with several proteins (JO et al., 2020), including ubiquilin-2 (UBQLN2) and hnRNPs (heteronuclear ribonucleotide binding proteins). This domain can be subdivided into four subregions: Glycine aromatic serine-rich regions (GaroS1 and GaroS2), a hydrophobic region (Φ), and a glutamine-arginine-rich region (Q/N). GaroS1 (residues 273-317) and GaroS2 (368-414) interact with RNA granules and leads to the formation of hydrogels. The hydrophobic region is known to be structured into an alpha helix and composes the amyloidogenic core of TDP-43 together with the Q/N domain. Finally, the Q/N domain contains aggregation-prone segments that ultimately contribute to the pathogenic aggregation of TDP-43 (FRANÇOIS-MOUTAL et al., 2019).

Fig 1. Schematic representation for the functional regions and missense mutations of human TDP-43. Schematic representations of TDP-43 designed using the software Aline (https://bondxray.org/software/aline.html). The wild-type sequence (black) is shown along with the corresponding missense mutations (red). Below, the NTD is colored red, while the RRM domains are colored green. At the CTD, GaroS are colored yellow, Φ region is colored orange, and Q/N is colored brown. NLS and NES are highlighted in purple and blue, respectively. Post-translational modification and important sites are also shown in the figure. Phosphorylation sites are represented by cyan triangles, RNA-binding residues are represented by magenta triangles, sumoylation sites are represented by yellow triangles, and acetylation sites are represented in orange. Relevant sites are represented by green triangles.


R
I K H A N I C S A S
M S E Y I R V T E D E N D E P I E I P S E D D G T V L L S T V T A Q F P G A C G L R Y R N P V S Q C
10 20 30 40 50
V H T V G I V
L T E D M A S A E G L L A M R
M R G V R L V E G I L H A P D A G W G N L V Y V V N Y P K D N K R K M D E T D A S S A V K V K R A V
60 70 80 90 100
S
C N N N V G R I V
Q K T S D L I V L G L P W K T T E Q D L K E Y F S T F G E V L M V Q V K K D L K T G H S K G F G F V
110 120 130 140 150
Q V V T P
H M D I T M G Q R D R K
R F T E Y E T Q V K V M S Q R H M I D G R W C D C K L P N S K Q S Q D E P L R S R K V F V G R C T E
160 170 180 190 200
W T
Y E Q S R I L N
D M T E D E L R E F F S Q Y G D V M D V F I P K P F R A F A F V T F A D D Q I A Q S L C G E D L I I
210 220 230 240 250
S
S V R V
S R T E D S V
V S D K T E S S K C S A A S R E D A S N A C C S
K G I S V H I S N A E P K H N S N R Q L E R S G R F G G N P G G F G N Q G G F G N S R G G G A G L G
260 270 280 290 300
V
V T S G K N V V S P V
H G S S E I A E T L D V V V R H G D I I R K H L C
N N Q G S N M G G G M N F G A F S I N P A M M A A A Q A A L Q S S W G M M G M L A S Q Q N Q S G P S
310 320 330 340 350
V
S V T S T P
S R T T S S N D P S C T V S R
A D E D S L S D A D G S I F G A C D A V T P T R G D G E L S H
G N N Q N Q G N M Q R E P N Q A F G S G N N S Y S G S N S G A A I G W G S A S N A G S G S G F N G G
360 370 380 390 400
P
S L G V G C E I
F G S S M D S K S S G W G M
410

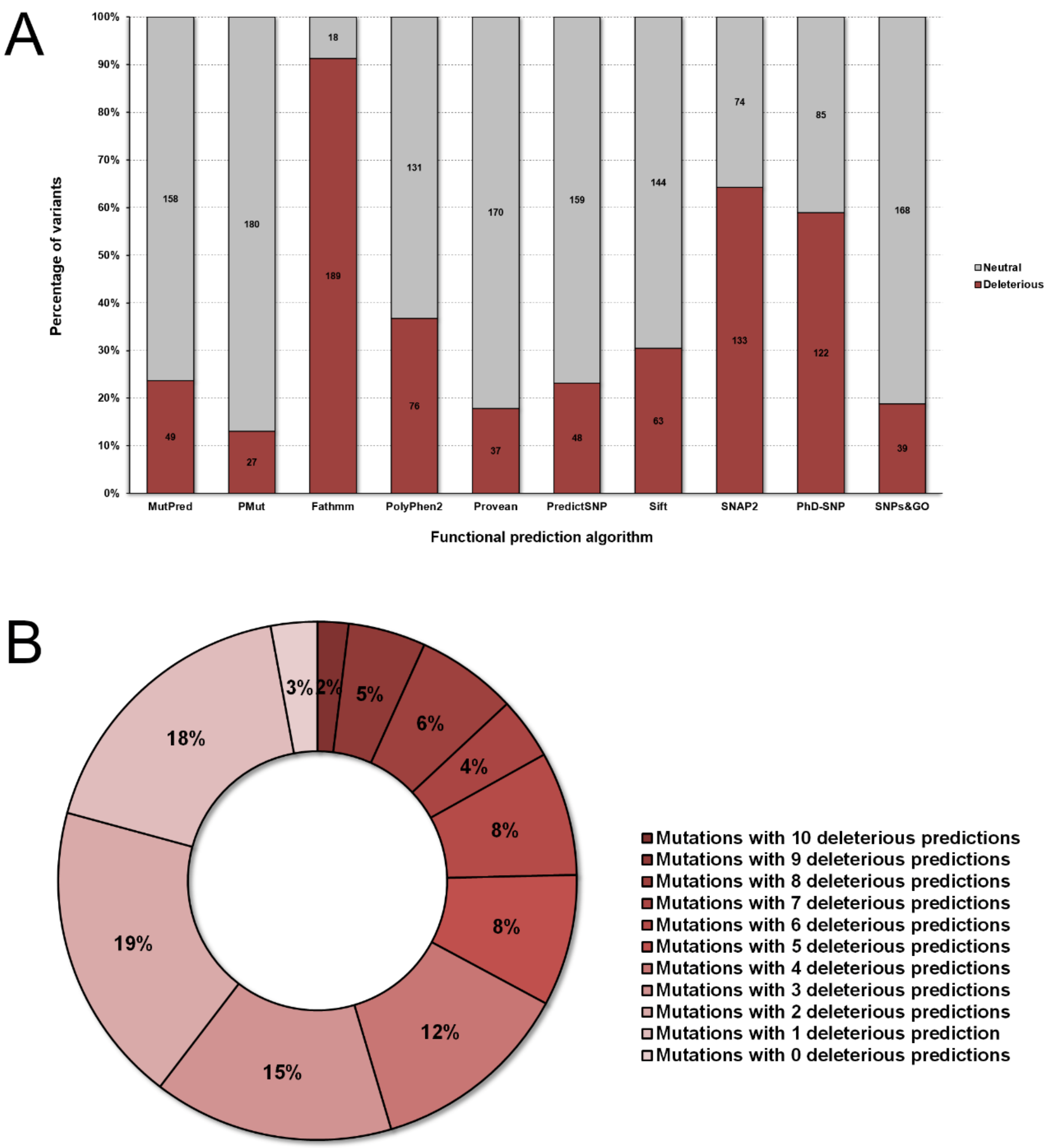
Two hundred and seven mutations were compiled for TDP-43 in the literature and databases consulted. Among them, only 50 have already been characterized, and all of them are considered deleterious for TDP-43 function. Thus, the effects of the other 157 mutations are still unknown (S1_Table). Aiming at a better comparison, important post-translational modifications and relevant residues for TDP-43 function or pathogenesis, which include phosphorylation, acetylation (HORNBECK et al., 2012), sumoylation, and nucleotide-binding sites (FRANÇOIS-MOUTAL et al., 2019), are shown in Fig 1 along with all the missense mutations compiled. As shown in Fig 1, these mutations occur all over the protein, but they are highly concentrated in the C-terminal domain. The following important residues for TDP-43 were also affected by at least one mutation: Ser92, Gln134, Lys176, Asp179, Ser183, Ser292, and Ser379, which may result in altered post-translational modifications.

3.2 PREDICTIVE ANALYSIS

Here, ten different algorithms were used to predict whether the missense mutation has deleterious or neutral effects on the TDP-43 function. The individual predictions for each algorithm are shown in Fig 2A. Among the algorithms used, Fathmm was the most accurate since it correctly classified 100% of the 32 already characterized missense mutations in TDP-43. Fathmm was followed by SNAP2 and PHD-SNP, with a 90% and 82% accuracy rate, respectively (S2_Table). Nonetheless, the elevated accuracy reached by these algorithms may be attributed to the high deleterious rates they had when predicting the entire set of mutations compiled for TDP-43 (Figure 2A), particularly considering that the 50 already characterized mutations are known to be deleterious.

Fathmm was the algorithm with the highest rate of deleterious predictions, classifying 91% of the analyzed mutations as damaging. On the other hand, PMut presented a deleterious rate of only 13%. These findings highlighted the importance of using multiple algorithms to predict the functional effects of missense mutations. To date, there is no gold standard method established for predicting the effects of missense mutations. The predictive algorithms currently available have been trained in different datasets using a distinct set of predictor variables and machine learning methods, resulting in individual accuracies ranging from 60 to 81% (LÓPEZ-FERRANDO et al., 2017; PEJAVER et al., 2017). As previously shown by BENDL *et al.*, 2014, combining the predictions of multiple functional algorithms can significantly increase the overall predictive performance (BENDL et al., 2014).

Fig 2. Functional prediction analysis of TDP-43 protein variants. (A) The number of deleterious (maroon), and neutral (gray) predictions per algorithm for the TDP-43 protein variants is shown as a bar plot. (B) The number of variants predicted as deleterious from zero (light maroon) to ten functional prediction algorithms (dark maroon) is shown as a donut plot.



We then analyzed the combined results from the functional prediction algorithms used, which are shown in Fig 2B. Twenty-five percent of all variants were consensus predicted as damaging, *i.e.*, more than half of all algorithms converge to the same response (> 5). These mutations mainly occur at the RRM domains, which are known to be involved in nucleic acid binding and recognition, key factors for TDP-43 function as a ribonuclear protein (BERNING; WALKER, 2019). Furthermore, the variants Y43C, D201Y, A260D, and K263E were predicted as deleterious by all the ten functional algorithms, regardless of the method used, which suggests that these mutations may be potentially harmful to TDP-43 (S2_Table).

The stability prediction, shown in S1_Fig, indicated that most TDP-43 mutations do not affect protein stability (68%). On the other hand, 25% of all mutations are destabilizing, and 7% are stabilizing. Increased TDP-43 stability upon ALS/FTD-related mutations could lead to malfunction (KLIM et al., 2021), possibly due to abnormal post-translational modifications and aberrant protein-protein-interactions (PPI) (LING et al., 2010), which ultimately results in protein aggregation, neuroinflammation and nuclear depletion (KLIM et al., 2021). The following mutations increase TDP-43 stability: N70T, S91L, S92L, E154D, D169G, D201Y, P320L, A321V, N345K, S347L, G348V, A366V, N371I, and D406G. These mutations mainly affect the C-terminal domain, which is involved in the majority of the TDP-43 PPI interactions and harbors most of the TDP-43 phosphorylation sites (JO et al., 2020).

The SNPEffect4.0 analyses are also shown in S1_Fig. The TANGO analysis suggested that six mutations reduce protein aggregation tendency (N70D, L71M, V72A, V72G, V133G, and Q134R), while three mutations increase this feature (T32I, N70T, and G314A). In turn, the WALTZ analysis indicated that eight variants increase the TDP-43 amyloid formation (V72A, V72G, T126N, R151H, R208Q, R208W, G314A, and S317T), while only three variants decrease this feature (I222T, A315E, and A315T). Finally, according to LIMBO, eight mutations were predicted to increase chaperone binding (E9K, L71V, V72A, V72G, N76S, V133G, and Q134R). No mutations were predicted to decrease chaperone binding.

Brain inclusions containing TDP-43 aggregates are a pathological hallmark of ALS and FTD. Similar inclusions were also observed in other neurodegenerative disorders, including Alzheimer's, Parkinson's, and Huntington's disease. Many of these TDP-43 toxic aggregates present short fibrillar structures with amyloid-like properties (LI; BABINCHAK; SUREWICZ, 2021). Thus, variants that increase protein aggregation and/or amyloid fibrils formation may contribute to TDP-43 toxicity.

3.3 STRUCTURAL MODELING AND VALIDATION

To date, no complete structure of human TDP-43 has already been experimentally determined (ROSE et al., 2021). The *in silico* methods allow the modeling of protein structures in an efficient and accurate approach for modeling protein structures. Among the available methods, comparative modeling is the most precise one. This method allows the generation of protein models with an overall quality range similar to that of experimental protein structures (KHAN et al., 2016). Using *ab initio*, comparative, and *threading* modeling methods, we generated eleven complete TDP-43 models, which underwent structural refinement and had their quality assessed by five validation algorithms, as shown in Table 1.

ProSa evaluates the potential energy of a protein structure and computes the Z-score value referring to its overall quality. The Z-score of a given structure is then plotted on a graph containing the Z-score values calculated for all experimental protein structures available at the PDB. QMEAN uses six different physicochemical descriptors to estimate the overall quality of a given structure (QMEAN-score). The QMEAN-score computed for the target structure is plotted on a graph containing the QMEAN-score values for 9766 high-resolution protein structures. For the specific size of the TDP-43 sequence, *i.e.*, 414 amino acids, the Z-score values for experimentally-determined structures range from approximately -3 to -13, whereas the normalized QMEAN-score values range from 0.7 to 0.9. QMEAN-score values approaching 0.75 indicates better physicochemical quality (PEREIRA et al., 2022a).

Table 1. Structural validation for the theoretical models of human TDP-43 protein.

Model	PROCHECK[1]		RAMPAGE[2]		VERIFY3D[3]		QMEAN[4]		PROSA-Web[5]	
	Before	After	Before	After	Before	After	Before	After	Before	After
ITASSER_1	60.7%	78.3%	68.2%	87.1%	78.5%	79.0%	0.32	0.51	-5.26	-4.98
ITASSER_2	57.8%	79.8%	69.4%	88.8%	82.6%	82.1%	0.44	0.57	-5.57	-5.49
ITASSER_3	62.2%	85.9%	73.8%	92.0%	79.2%	82.8%	0.47	0.64	-6.76	-6.44
ITASSER_4	62.8%	80.4%	70.4%	91.0%	83.6%	86.7%	0.43	0.59	-4.54	-4.65
ITASSER_5	57.8%	80.9%	68.7%	92.5%	79.5%	81.4%	0.43	0.62	-6.57	-6.55
Robetta_1	87.4%	91.5%	94.7%	98.3%	96.4%	93.0%	0.77	0.74	-8.58	-8.33
Robetta_2	85.9%	92.7%	94.9%	97.8%	87.9%	87.0%	0.78	0.76	-9.32	-9.19
Robetta_3	85.9%	94.4%	95.1%	98.3%	91.3%	91.8%	0.79	0.75	-7.95	-7.86
Robetta_4	86.5%	91.8%	94.2%	96.8%	91.8%	92.0%	0.78	0.73	-7.26	-7.43
Robetta_5	89.4%	93.8%	97.1%	98.8%	89.9%	90.8%	0.84	0.78	-8.17	-8.13
Phyre2_1	75.1%	84.2%	83.0%	92.5%	66.4%	77.0%	0.50	0.59	-4.31	-4.76

[1]Percentage of residues within most favorable regions in the PROCHECK Ramachandran plot; [2]Percentage of residues within favorable regions in the RAMPAGE Ramachandran plot; [3]Percentage of residues with 3D-1D compatibility score equal to or higher than 0.2; [4]The QMEAN-score estimated for the model normalized from a set of values computed for high-resolution protein structures; [5]The overall quality score (Z-score) estimated for the model is within the range of Z-score values computed for NMR or crystallographic structures.

Verify3D, in turn, evaluates the structure sequence compatibility for each protein residue, named 3D-1D–score. According to Verify3D, high-quality structures are expected to have more than 80% of their residues with a 3D-1D score equal to or higher than 0.2. PROCHECK and RAMPAGE evaluate the stereochemical quality of protein structures based on phi/psi angles arrangement. These algorithms generate Ramachandran plots in which protein residues are distributed in favored, allowed, or disallowed regions. High-quality protein structures are expected to have more than 90.0% of their residues in the most favored regions of PROCHECK's Ramachandran plot and around 98.0% of their residues in favored regions of Rampage's Ramachandran plot (OLIVEIRA et al., 2019).

As shown in Table 1, the structural refinement was responsible to increase the overall quality for most of the original models, leading to improved stereochemical quality (PROCHECK and RAMPAGE), structure-sequence compatibility (Verify-3D), and physicochemical quality (QMEAN). Only small differences in the Z-score values were observed upon refinement (ProSa-Web).

The models generated by I-TASSER and Phyre2 failed in PROCHECK and RAMPAGE evaluation even after structural refinement, suggesting that they presented low stereochemical quality. These models were then not considered for the validation assessment. The models generated by Robetta's algorithm were approved in all quality assessment analyses we carried out. Among them, model number three presented the greatest overall quality scores in the validation assessment (Table 1), being therefore selected for further analyses.

The individual validation results for the selected model, i.e., Robetta_3, are shown in S2_Fig and S3_Fig. The quality assessment analysis (S2_Fig and S3_Fig) indicated that the selected model presented an overall quality that is comparable to experimentally determined structures of the same length. The model also presented high stereochemical and physiochemical quality, in addition to elevated structure-sequence compatibility. Thus, being validated by the algorithms PROCHECK, PROSA, Verify-3D, QMEAN, and RAMPAGE.

The Protein Blast analysis returned three potential templates for the structural alignment: 4BS2, 6T4B, and 2N2C. All of them are structural PDB fragments of human TDP-43 protein, thus presenting 100% identity with the selected model's sequence. Moreover, these templates covered different protein regions, presenting an overall structural coverage of 65% (Table 2). The structural alignment returned RMSD and TM-score values, which are metrics of similarity between two structures. RMSD is a measure of the raw atomic deviation between all atoms within two structures. TM-score, on the other hand, is an atomic distance metric normalized by the protein length. Alignments with RMSD values lower than 2Å and TM-score greater than 0.5 indicates high structural similarity and, consequently, a non-random relationship between the model and template structure, which is indispensable to validate protein models predicted by comparative modeling like those generated by Robetta's algorithm. As shown in Table 2, the selected model presented RMSD and TM-score values within the desired range for templates 4BS2 and 6T4B, which cover most of the TDP-43 structure length (60%). The alignment with the 2N2C fragment, on the other hand, presented high RMSD values and low TM-score, suggesting that the selected model is not similar to this template structure (PEREIRA et al., 2018). The 2N2C fragment corresponds to the hydrophobic helix of the C-terminal domain, which is a low-ordered region (FRANÇOIS-MOUTAL et al., 2019). Thus, considering that most of the TDP-43 model is structurally similar to the available template structures, *i.e.*, 92% of the total coverage, the model was also validated by this criterion.

Aiming at a better visualization, we also provided a three-dimensional representation of the structural alignment between the theoretical model of TDP-43 and potential template structures 4BS2, 6T4B, and 2N2C, which is shown in S4_Fig. A visual inspection of the alignment corroborates the high structural similarities existing between the selected model and the experimentally determined fragments of TDP-43.

Table 2. Alignment between the theoretical model of TDP-43 protein and the structural templates selected by the Protein Blast algorithm.

PDB ID	**Sequence identity**	**Query cover**	**Alignment score**	**Aligned length**	**RMSD**	**TM-score**
4BS2	100%	40%	354	174	1.78	0.92
6T4B	100%	19%	176	76	1.91	0.78
2N2C	100%	6%	38	35	3.10	0.39

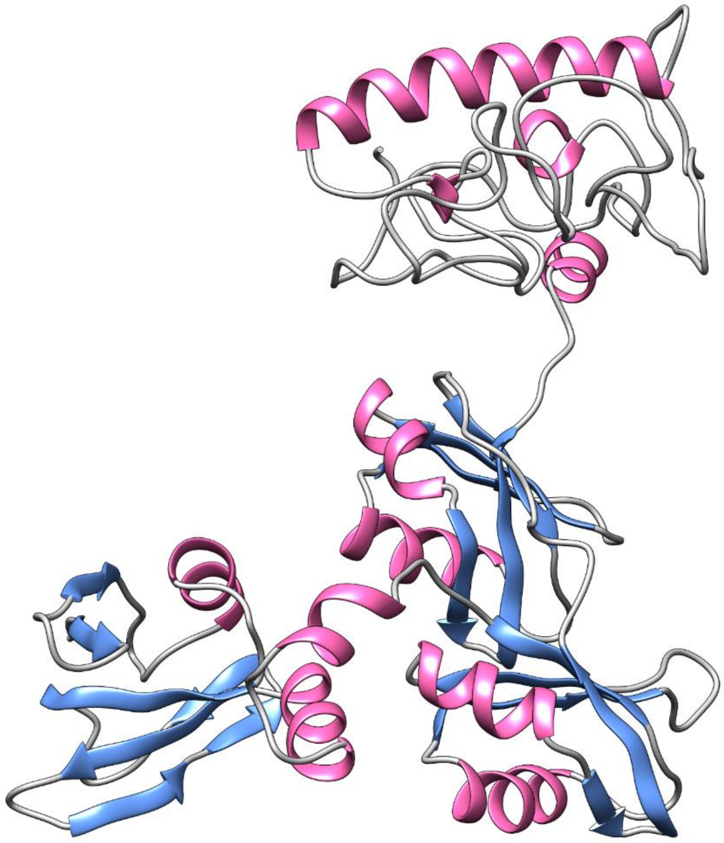
To further validate the *in silico* modeled structure, we performed secondary structure predictions of TDP-43 protein using seven algorithms, in addition to IDR predictions using five algorithms. The consensus results from these analyses are shown in S5_Fig. The secondary structure of potential template structures 4BS2, 6T4B, and 2N2C are also shown in S5_Fig for further comparison. The comparative analysis suggested that the theoretical model presented 83% of secondary structure similarity with the consensus prediction, in addition to 80% similarity with the IDR prediction. Furthermore, the theoretical model presented 81% of secondary structure similarity with potential template structures. The structure comparison analyses, therefore, pointed to the high similarity between the theoretical model, consensus predictions, and template structures, reaffirming the model's accuracy.

Most of the intrinsically disordered regions (IDRs) predicted for TDP-43 coincide with the region not yet experimentally determined for this protein, especially in the C-terminal domain, which is known to be a low-ordered region (FRANÇOIS-MOUTAL et al., 2019). IDRs do not fold into stable secondary structures and often only form a defined structure upon binding to partner proteins. These regions are inherently flexible, which usually leads to difficulties in protein crystallization (ATKINS et al., 2015) and purification, especially if they are located at the N and C terminus of the protein (LINDING et al., 2003). The lack of a complete experimentally determined structure of

TDP-43 might be related to the IDRs, especially because they are mainly located at the C terminal domain.

The quality assessment analysis (S2_Fig and S3_Fig), structural alignment (Table 2), and secondary structure comparisons (S5_Fig), thus, suggested that the complete theoretical model of TDP-43 protein is accurate and reliable. The validated model of TDP-43 is shown in Fig 3. The atomic coordinates of this model were provided in S1_File. The validated model may be used to identify and optimize potential drug candidates targeting TDP-43 through a structure-based drug design approach, possibly favoring the development of more effective anti-ALS/FTD drugs with fewer side effects and less toxicity (BONETTA et al., 2016).

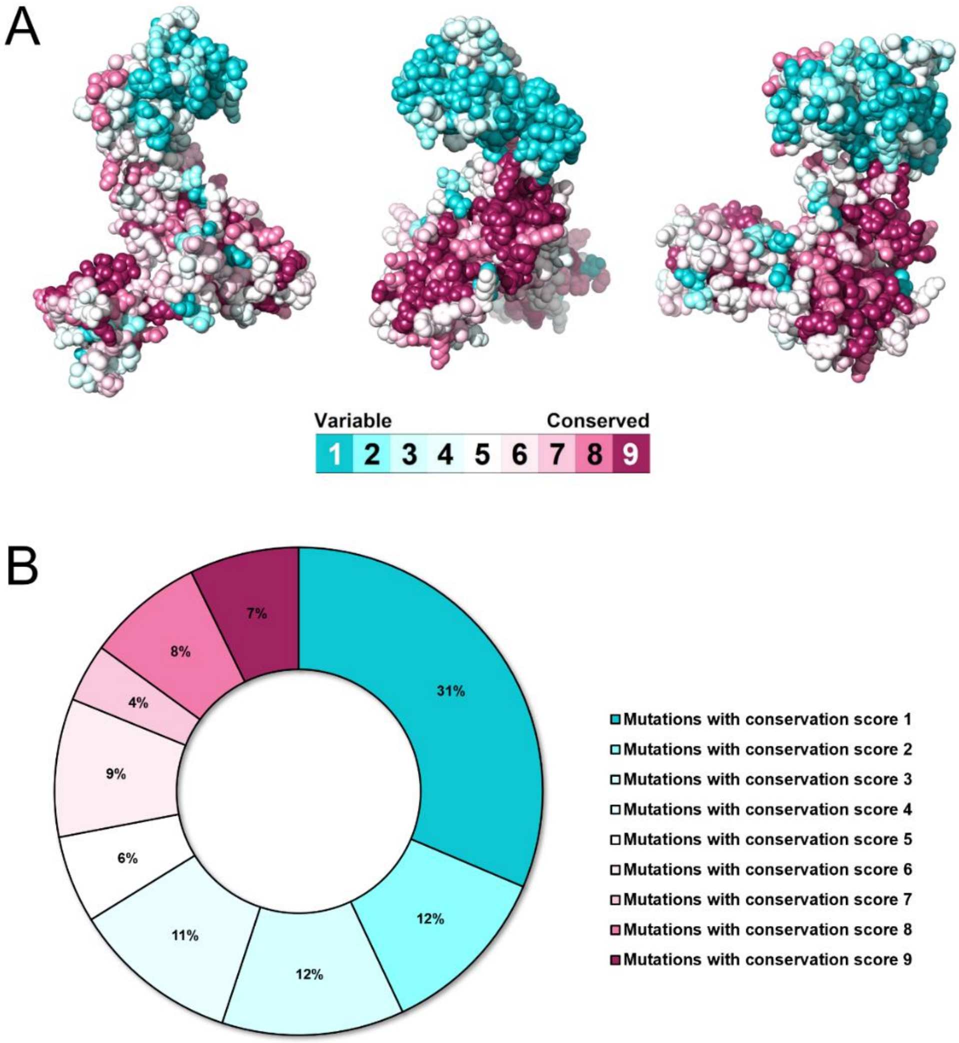
Fig 3. *In silico* modeled structure of TDP-43. TDP-43 is represented as a tridimensional model with the α-helices regions colored in magenta, β-strand regions colored in blue, and the coiled regions colored in gray.



3.4 EVOLUTIONARY CONSERVATION ANALYSIS

The validated model was submitted to the ConSurf server, which estimated the evolutionary conservation level of each amino acid of TDP-43, whose scores were projected on the protein's surface (Fig 4 A). Key amino acids within a protein, *i.e.*, those with structural or functional relevance, are usually conserved throughout the evolution due to highly selective pressure. Thus, the biological importance of a given amino acid could be associated with its evolutionary conservation level (ASHKENAZY et al., 2016). Furthermore, mutations affecting conserved positions within the protein are most likely to lead to deleterious effects (PEREIRA et al., 2022a). Detailed information on evolutionary conservation analysis is shown in S4_Table.

Fig 4. Evolutionary conservation analysis of TDP-43. Each amino acid of TDP-43 was colored according to the ConSurf coloring scheme and projected on the protein's surface. The color-coding bar shows the ConSurf coloring scheme, which ranges from cyan and variable to magenta and conserved. (A) TDP-43 is represented as a space-filling model, which is shown in three different angles rotating 90º from each other. (B) The number of mutations affecting amino acids with conservation scores from one (cyan) to nine (magenta) is shown as a donut plot.



The ConSurf analysis suggested that the RRM domains are highly conserved, while the C-terminal domain is highly variable (Fig 4A). As shown in Fig 4B, TDP-43 mutations mainly affect variable positions (ConSurf-score ≤3), most of them located in the C-terminal domain (S4_Table). On the other hand, 19% of all variants occur at conserved positions (ConSurf-score ≥7), from which 7% affect highly conserved amino acids, *i.e*, those with maximum ConSurf-score. Overall, fifteen variants affected highly conserved and, possibly, biologically relevant amino acids of TDP-43: M1I, E9K, D23N, N76S, Q134R, R151H, I168M, I168V, R171Q, F211S, I222L, I222T, A260D, P262T, and K263E.

To facilitates the prioritization of most-likely deleterious mutations for TDP-43, we adopted a penalty system combining all previous predictive analyses, a literature review, and alterations in physicochemical properties. Only uncharacterized mutations were considered for this analysis. The maximum penalty a mutation can receive is 27, in which: i) 0-10 points account for the number of deleterious predictions; ii) 1-9 points correspond to the ConSurf-score. Mutations affecting conserved regions are more likely to be deleterious (ASHKENAZY et al., 2016); iii) 0-4 points were assigned to increased aggregation tendency, increased amyloid propensity, decreased chaperone binding, and increased protein stability. These alterations are known to be related to TDP-43 toxicity (KLIM et al., 2021; LI; BABINCHAK; SUREWICZ, 2021); iv) 0-1 point was attributed to whether the mutation affects a residue involved in post-translational modification or nucleotide-binding (functional residues), which are central to TDP-43 function and pathogenesis (FRANÇOIS-MOUTAL et al., 2019); v) 0-3 points accounts for any physiochemical property altered upon the amino acid substitution, *i.e.*, hydrophobicity, charge, size. Radical amino acid substitutions, i.e., those between amino acids with very different physicochemical properties, are more often deleterious, possibly due to a stronger negative selective pressure (WEBER; WHELAN, 2019).

As shown in S5_Table, variants Y43C, D201Y, F211S, I222T, K224N, A260D, P262T, and A321D received the highest penalties. According to the criteria discussed earlier, these mutations are most likely to be deleterious to TDP-43, thus, being important targets for future investigation *in vitro* and *in vivo*.

3.5 MOLECULAR DYNAMICS SIMULATIONS

Molecular dynamics simulations (MD) is an *in silico* method of solving Newtonian equations of motions for a given group of atoms, which are useful to reproduce the protein behavior in its biological environment. The atomic coordinates and velocities computed for the simulated system are registered over time in the trajectory file, providing detailed information on changes in protein conformations and fluctuations. The trajectory file is then analyzed to assess biochemical and structural parameters like structural flexibility (KHAN et al., 2016). Four TDP-43 mutations were chosen to undergo MD simulations based on their pathogenic potential and phenotype peculiarities: K263E, G335D, M337V, and Q343R. All of them are ALS/FTD-associated mutations.

K263E variant occurs at the RRM2 domain and impairs RNA binding (CHEN et al., 2019). K263E variant is also related to a dramatic increase in TDP-43 ubiquitination and has been associated with atypical symptoms of ALS and FTD, including supranuclear gaze palsy, hyperkinetic chorea-form movements, motor stereotypes, and primitive reflexes (HANS et al., 2014). K263E increases TDP-43 stability and reduces its solubility, being more likely to form hyperphosphorylated intra-nuclear aggregates (CHEN et al., 2019). Furthermore, this mutation reduces TDP-43 ability to bind to RNA, and affects RNA processing, in addition to present neuron-specific loss of function (IMAIZUMI et al., 2022).

G335D and Q343R affect the amyloidogenic core region, which includes the Φ and Q/N regions of CTD (FRANÇOIS-MOUTAL et al., 2019). This region is believed to be involved in the formation of cytoplasmic inclusions and has been reported to form amyloid-like β-sheet structures, particularly between amino acids 331 and 369 (GAO et al., 2018). These mutations are responsible for important transformations in CTD, resulting in TDP-43 aggregation and inclusion formation (JIANG et al., 2016).

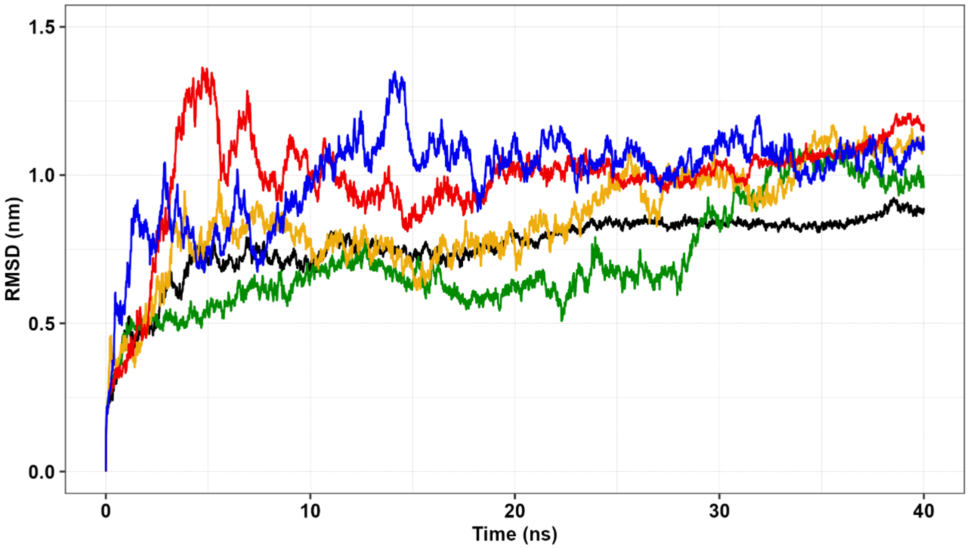
M337V is the most frequent TDP-43 missense mutation (PRASAD et al., 2019). Similar to G335D and Q343R, M337V also occurs in the amyloidogenic core region (FRANÇOIS-MOUTAL et al., 2019). This mutation leads to glial cell inclusions and a clinical phenotype of initial bulbar symptoms (TAMAOKA et al., 2010). M337V was found to impair mRNA splicing regulated by TDP-43 *in vivo* (WATANABE et al., 2020).

RMSD is a useful measure of the atomic displacement between two protein structures during an MD simulation. RMSD is computed at each simulation step, comparing the actual conformation with the initial one. As shown in Fig 5, TDP-43 wild-type and variants K263E and G335D presented a steady behavior after an initial moment of structural instability. The establishment of a plateau in the RMSD values after approximately 20ns suggested that these proteins float around average stable conformation, thus indicating system equilibration (PEREIRA; ABRAHIM-VIEIRA; DE MESQUITA, 2021). The MD simulations of M337V and Q343R do not seem to stabilize within the simulated time, pointing to structural instability. No differences in the average RMSD values were observed for the analyzed variants (S6_Table).

The radius of gyration (Rg) is a measure of the atomic distances between all protein atoms in a given protein and their common center of mass. RG thus provides information on protein's overall dimensions, or volume, over time. The Rg values computed for TDP-43 wild-type and variants are shown in S6_Fig. TDP-43 wild-type and variants K263E, Q343R, and M337V presented steady behavior after 10ns, suggesting stable protein folding (PEREIRA et al., 2022b). Variant G335D, on the other hand, presented a decreasing tendency in the Rg values after approximately 25ns. Furthermore, variants K263E (3.45±0.18nm) and G335D (3.06±0.23nm) presented increased Rg values when compared to the wild-type (2.74±0.04nm) (S6_Table), which indicates that these mutations could affect TDP-43 volume.

Solvent Accessible Surface Area (SASA) is a measure of the protein exposed surface, providing information on its ability to interact with the solvent over time (BONET et al., 2021). As shown in S6_Fig, all the analyzed proteins presented a decreasing SASA tendency throughout the simulations. Variants K263E (248.59±6.49nm²), G335D (244.57±8.65nm²), Q343R (245.25±5.68nm²), and M337V (244.40±7.25nm²) presented similar average SASA values to that of wild-type TDP-43 (238.10±10.12nm²) (S6_Table), indicating that they may not alter the protein's exposed surface.

Fig 5. The RMSD values calculated for TDP-43 wild-type and variants at 300K are shown as a function of time. TDP-43 wild-type is colored black, variant M337V is colored green, variant Q343R is colored orange, variant G335D is colored red, and variant K263E is colored blue.



The Rg and SASA analyses pointed to increased TDP-43 volume upon K263E and G335D mutations but no alterations in the accessible surface area (S6_Table). This finding suggests that variants K263E and G335D could affect the surface-to-volume ratio of TDP-43, which is associated with the protein's ability to interact (PEREIRA et al., 2020).

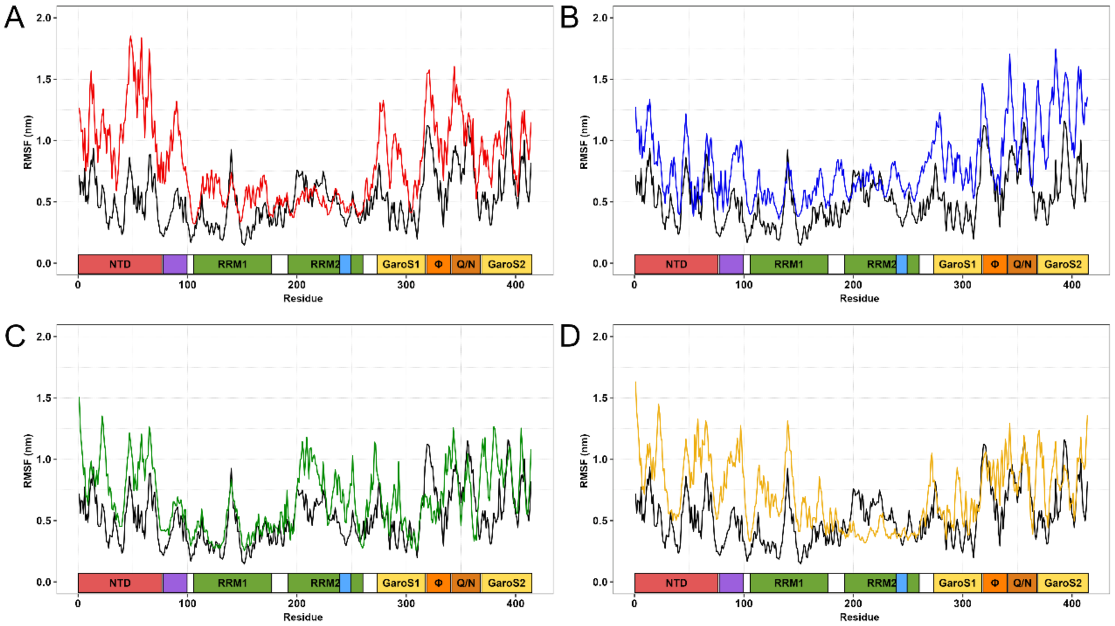
The stability of protein structures is determined by the number of interactions they formed, including hydrogen bonding, hydrophobic, and electrostatic interactions (PIKKEMAAT et al., 2002). We then analyzed the number of hydrogen bonds to better understand the structural impact of TDP-43 variants. As shown in S6_Fig, TDP-43 wild-type and variants presented an increasing tendency of hydrogen bonding formation during the simulations. No differences in the average number of these interactions were observed for the variants when compared to the wild-type (S6_Table).

At last, we analyzed root-mean-square fluctuation (RMSF) and B-factor of TDP-43 wild-type and variants. RMSF is a measure for the atomic deviations in protein residues related to their average position during a simulation, thus being useful to investigate local flexibility. B-factor, in turn, is a measure of the atomic displacement of amino acids due to thermal vibrations. B-factor is usually plotted on the protein's surface, providing an interesting three-dimensional representation of protein flexibility (OLIVEIRA et al., 2019).

Among the analyzed variants, G335D was the variant that most affected TDP-43 flexibility (Fig 6). This variant leads to an important flexibility increase in the N-terminal domain (NTD), nuclear localization signal (NLS), GaroS1, hydrophobic region, and Q/N, more than doubling the RMSF values in these regions. G335D mutation also results in increased flexibility at the RRM1, but to a lesser extent. Variant K263E leads to important flexibility alterations at the C-terminal domain, RRM2, and nuclear localization signal, in addition to a small flexibility increase in TDP-43 as a whole. M337V, in turn, was the variant that least affected this parameter. M337V slightly increases protein flexibility, except for the hydrophobic region, where it greatly reduces flexibility, and part of NTD, RRM2, and GaroS2, where it importantly increases this parameter. At last, Q343R greatly increases protein flexibility at the NTD, NLS, and part of GaroS2. This variant slightly increases the flexibility in other TDP-43 regions, except for the RRM2 domain, where Q343R decreased this parameter.

As shown in S7_Fig, the B-factor analysis pointed to flexibility alterations upon mutations at the same regions found altered in the RMSF analysis, corroborating this finding.

Fig 6. The RMSF values calculated for each residue of TDP-43 wild-type and variants at 300K are shown as a function of time. (A) Wild-type TDP-43 is colored black, while variant G335D is colored red. (B) Wild-type TDP-43 is colored black, while variant K263E is colored blue. (C) Wild-type TDP-43 is colored black, while variant M337V is colored green. (D) Wild-type TDP-43 is colored black, while variant Q343R is colored orange. A schematic representation for the functional domains of TDP-43 is also shown for further comparison. The NTD is colored red, while the RRM domains are colored green. At the CTD, GaroS are colored yellow, Φ region is colored orange, and Q/N is colored brown. NLS and NES are highlighted in purple and blue, respectively.



Structural flexibility is a determinant factor for protein binding affinity and specificity. Thus, flexibility alterations can lead to strong and non-intuitive consequences for protein interactions (FORREY; DOUGLAS; GILSON, 2012). The flexibility alterations observed in the RMSF and B-factor analyses for variants G335D, K263E, M337V, and Q343R may lead to defective TDP-43 interactions, especially in the most affected regions.

The study by Bhandare & Ramaswamy (2018) indicated that variant K263E impairs TDP-43 binding to RNA. The authors also observed overall increased flexibility in TDP-43 protein, especially in the RRM2 domain, similar to our findings (BHANDARE; RAMASWAMY, 2018). Chen *et al.* (2019) observed that in addition to disrupting RNA-binding, K263E also presents an increased ability to interact with other TDP-43 proteins within the cytoplasm or nucleus, sequestrating them into toxic aggregates (CHEN et al., 2019). TDP-43 self-interaction and co-aggregation are often attributed to the Q/N region (HANS et al., 2014), which presented an important flexibility alteration in the RMSF and B-factor analyses we carried out (Fig 6 and S7_Fig).

The study by Feneberg *et al.* (2020), in turn, pointed to aberrant protein-protein interactions (PPI) with significant functional impairment upon the M337V mutation.

These abnormal interactions could affect TDP-43 recognition by the endosomal complex required for transport and secretion, which is believed to be potentially toxic due to reduced TDP-43 cytoplasmic trafficking, a likely factor in the formation of ALS-related protein aggregates. The study by Feneberg *et al.* (2020) also suggested that specific RNA-dependent TDP-43 interactions were also affected by the M337V variant (FENEBERG et al., 2020). Interestingly, our study pointed to important flexibility alterations in the hydrophobic and GaroS2 region of the CTD, in addition to the RRM2 domain, which are key regions for TDP-43 interaction with other proteins and RNA recognition, respectively (FRANÇOIS-MOUTAL et al., 2019).

Conicella *et al.* (2020) analyzed the impact of the ALS-related mutation G335D on TDP-43. The authors observed that this mutation enhances helix-helix interactions within the hydrophobic region of TDP-43 (CONICELLA et al., 2020). This G335D-induced alteration makes the hydrophobic region, also known as the amyloidogenic core fragment, significantly more prone to aggregate or form amyloid-like fibers (JIANG et al., 2016). Our MD findings suggested that the variant G335D presented a great flexibility increase within the hydrophobic helix and adjacent regions of the C-terminal domain, which may be related to the enhanced helix-helix interactions observed by Conicella *et al.* (2020).

The study by Mompeán *et al.* (2014) suggested that variant Q343R formed aberrant intramolecular interactions within the C-terminal domain (MOMPEÁN et al., 2014). This mutation also decreases TDP-43 nuclear localization (WOOD et al., 2021). Our findings pointed to greatly increased flexibility for Q343R in part of GaroS2, located near the mutation spot at the C-terminal domain, in addition to NLS, which is the region directly recognized by Importin-α for the active transport of TDP-43 into the nucleus (Fig 6). Alterations at the NLS can lead to TDP-43 accumulation in the cytoplasm, an event often associated with the formation of protein aggregates (FRANÇOIS-MOUTAL et al., 2019).

Extensive flexibility alterations, such as those observed in our MD analyses for TDP-43, particularly at functional domains, can have perceptive consequences for protein binding properties (FORREY; DOUGLAS; GILSON, 2012). According to the literature consulted, previously described in the last four paragraphs, mutations K263E, G335D, M337V, and Q343R cause non-native interactions and impair protein recognition mainly associated with functional regions of TDP-43. These alterations ultimately lead to protein aggregation and amyloid-fiber formation, which are central events in the pathophysiology of ALS and FTD (MOMPEÁN et al., 2014; CONICELLA et al., 2020; FENEBERG et al., 2020; WOOD et al., 2021).

4 CONCLUSIONS

Two hundred and seven TDP-43 mutations were compiled from the databases and literature, mainly affecting its C-terminal domain. The predictive analysis pointed to a moderate rate of deleterious and destabilizing mutations. Furthermore, most mutations occur at variable positions of TDP-43, according to the ConSurf analysis. We also adopt a penalty system to prioritize the most likely damaging mutations of TDP-43 by combining the predictive analyses, evolutionary conservation, literature review, and alterations in physicochemical properties of amino acid substitution. This analysis indicated that mutations Y43C, D201Y, F211S, I222T, K224N, A260D, P262T, and A321D received the highest penalties, thus being important targets for future investigation *in vitro* and *in vivo*. The present work also provided an accurate, complete, and unprecedented three-dimensional structure for TDP-43 protein, which can be used to identify and optimize potential drug candidates, favoring the development of more effective drugs with fewer side effects and less toxicity. At last, our MD findings pointed to a noticeable flexibility increase in the CTD, NTD, NLS, and RRM domains upon K263E, G335D, M337V, and Q343R variants, which may cause non-native interactions and impaired TDP-43 recognition. These alterations can ultimately lead to protein aggregation and amyloid fiber formation, central events in ALS and FTD pathogenesis.

FUNDING

This study was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Financiadora de Estudos e Projetos, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Universidade Federal do Estado do Rio de Janeiro, and NVIDIA Corporation. The funders had no role in study design and manuscript preparation.

COMPETING INTERESTS

The material received as support from NVIDIA for this study does not alter our adherence to the journal policies on sharing data and materials.

REFERENCES

ARTHUR, K. C. et al. Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. **Nature Communications**, v. 7, p. 1–6, 2016.

ASHKENAZY, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. **Nucleic Acids Research**, v. 44, n. W1, p. W344–W350, jul. 2016.

ATKINS, J. D. et al. Disorder prediction methods, their applicability to different protein targets and their usefulness for guiding experimental studies. **International Journal of Molecular Sciences**, v. 16, n. 8, p. 19040–19054, 2015.

BENDL, J. et al. PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. **PLoS Computational Biology**, v. 10, n. 1, p. 1–11, 2014.

BERNING, B. A.; WALKER, A. K. The pathobiology of TDP-43 C-terminal fragments in ALS and FTLD. **Frontiers in Neuroscience**, v. 13, n. APR, p. 1–27, 2019.

BHANDARE, V. V.; RAMASWAMY, A. The proteinopathy of D169G and K263E mutants at the RNA Recognition Motif (RRM) domain of tar DNA-binding protein (tdp43) causing neurological disorders: A computational study. **Journal of Biomolecular Structure and Dynamics**, v. 36, n. 4, p. 1075–1093, 2018.

BONET, L. F. S. et al. Molecular dynamics and protein frustration analysis of human fused in Sarcoma protein variants in Amyotrophic Lateral Sclerosis type 6 : An In Silico approach. **PLOSONE**, v. 16, n. 9, p. e0258061, 2021.

BONETTA, R. et al. Role of Protein Structure in Drug Discovery Molecular Control of Globin Gene Switching View project Role of Protein Structure in Drug Discovery. **Xjenza Online - Journal of The Malta Chamber of Scientists**, v. 4, n. January, p. 126–130, 2016.

CHEN, H. J. et al. RRM adjacent TARDBP mutations disrupt RNA binding and enhance TDP-43 proteinopathy. **Brain**, v. 142, n. 12, p. 3753–3770, 2019.

CONICELLA, A. E. et al. TDP-43 α-helical structure tunes liquid–liquid phase separation and function. **Proceedings of the National Academy of Sciences of the United States of America**, v. 117, n. 11, p. 5883–5894, 2020.

DIEHL-SCHMID, J. et al. Caregiver Burden and Needs in Frontotemporal Dementia. **Journal of Geriatric Psychiatry and Neurology**, v. 26, n. 4, p. 221–229, 2013.

FENEBERG, E. et al. An ALS-linked mutation in TDP-43 disrupts normal protein interactions in the motor neuron response to oxidative stress. **Neurobiology of Disease**, v. 144, n. March, 2020.

FORREY, C.; DOUGLAS, J. F.; GILSON, M. K. The fundamental role of flexibility on the strength of molecular binding. **Soft Matter**, v. 8, n. 23, p. 6385–6392, 2012.

FRANÇOIS-MOUTAL, L. et al. Structural Insights Into TDP-43 and Effects of Post-translational Modifications. **Frontiers in Molecular Neuroscience**, v. 12, n. December,

p. 1–22, 2019.

GAO, J. et al. Pathomechanisms of TDP-43 in neurodegeneration. **Journal of Neurochemistry**, v. 146, n. 1, p. 7–20, 2018.

GLADMAN, M.; DHARAMSHI, C.; ZINMAN, L. Economic burden of amyotrophic lateral sclerosis : A Canadian study of out-of-pocket expenses. **Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration**, v. 15, n. 5–6, p. 426–432, 2014.

HANS, F. et al. UBE2E Ubiquitin-Conjugating Enzymes and Ubiquitin Isopeptidase Y Regulate TDP-43 Protein Ubiquitination. **The Journal of Biological Chemistry**, v. 289, n. 27, p. 19164–19179, 2014.

HEO, L.; PARK, H.; SEOK, C. GalaxyRefine: Protein structure refinement driven by side-chain repacking. **Nucleic acids research**, v. 41, n. Web Server issue, p. 384–388, 2013.

HODGES, J. R.; PIGUET, O. Progress and Challenges in Frontotemporal Dementia Research: A 20-Year Review. **Journal of 'Alzheimer's Disease**, v. 62, n. 3, p. 1467–1480, 2018.

HORNBECK, P. V. et al. PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. **Nucleic Acids Research**, v. 40, n. D1, p. 261–270, 2012.

IMAIZUMI, K. et al. Pathogenic Mutation of TDP-43 Impairs RNA Processing in a Cell Type-Specific Manner: Implications for the Pathogenesis of ALS/FTLD. **eNeuro**, v. 9, n. 3, p. 1–12, 2022.

JAISWAL, M. K. Riluzole and edaravone: A tale of two amyotrophic lateral sclerosis drugs. **Medicinal Research Reviews**, n. July, 2018.

JI, A. et al. Genetics insight into the amyotrophic lateral sclerosis/frontotemporal dementia spectrum. **Journal of Medical Genetics**, p. jmedgenet-2016-104271, 2017.

JIANG, L.-L. et al. Two mutations G335D and Q343R within the amyloidogenic core region of TDP-43 influence its aggregation and inclusion formation. **Scientific Reports**, v. 6, p. 23928, mar. 2016.

JO, M. et al. The role of TDP-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies. **Experimental and Molecular Medicine**, v. 52, n. 10, p. 1652–1662, 2020.

KELLEY, L. a et al. Europe PMC Funders Group The Phyre2 web portal for protein modelling , prediction and analysis. **Nature protocols**, v. 10, n. 6, p. 845–858, 2015.

KHAN, F. I. et al. Current updates on computer aided protein modeling and designing. **International Journal of Biological Macromolecules**, v. 85, p. 48–62, 2016. Disponível em: <http://dx.doi.org/10.1016/j.ijbiomac.2015.12.072>.

KLIM, J. R. et al. Connecting TDP-43 Pathology with Neuropathy. **Trends in Neurosciences**, v. 44, n. 6, p. 424–440, 2021.

KREBS, B. B.; DE MESQUITA, J. F. Amyotrophic Lateral Sclerosis Type 20 - In Silico Analysis and Molecular Dynamics Simulation of hnRNPA1. **PLoS ONE**, v. 11, n. 7, p. e0158939, jul. 2016.

LI, Q.; BABINCHAK, W. M.; SUREWICZ, W. K. Cryo-EM structure of amyloid fibrils formed by the entire low complexity domain of TDP-43. **Nature Communications**, v. 12, n. 1, p. 1–8, 2021.

LINDING, R. et al. Protein disorder prediction: Implications for structural proteomics. **Structure**, v. 11, n. 11, p. 1453–1459, 2003.

LING, S. C. et al. ALS-associated mutations in TDP-43 increase its stability and promote TDP-43 complexes with FUS/TLS. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 30, p. 13318–13323, 2010.

LÓPEZ-FERRANDO, V. et al. PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update. **Nucleic Acids Research**, v. 45, n. Web Server, p. W222–W228, 2017.

MAKINO, S. et al. DisMeta – a Meta Server for Construct Design and Optimization Yuanpeng. **Methods Mol Biol**, v. 1091, n. 9, p. 3–16, 2014.

MARESOVA, P. et al. Activities of daily living and associated costs in the most widespread neurodegenerative diseases: A systematic review. **Clinical Interventions in Aging**, v. 15, p. 1841–1862, 2020.

MCGUFFIN, L. J. et al. IntFOLD: An integrated server for modelling protein structures and functions from amino acid sequences. **Nucleic Acids Research**, v. 43, n. W1, p. W169–W173, 2015.

MOMPEÁN, M. et al. Structural characterization of the minimal segment of TDP-43 competent for aggregation. **Archives of Biochemistry and Biophysics**, v. 545, n. January, p. 53–62, 2014.

NCBI. Database Resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 45, n. D1, p. D12–D17, 2017.

OJAIMI, Y. Al et al. TAR DNA-binding protein of 43 kDa (TDP-43) and amyotrophic lateral sclerosis (ALS): a promising therapeutic target. **Expert Opinion on Therapeutic Targets**, v. 1, n. Jun, 2022.

OLIVEIRA, C. C. S. De et al. In silico analysis of the V66M variant of human BDNF in psychiatric disorders : An approach to precision medicine. **PLoS ONE**, v. 14, n. 4, p. e0215508, 2019.

PEJAVER, V. et al. MutPred2: inferring the molecular and phenotypic impact of amino acid variants. **bioRxiv**, p. 134981, 2017.

PEREIRA, G. R. C. et al. In silico analysis and molecular dynamics simulation of human superoxide dismutase 3 (SOD3) genetic variants. **Journal of Cellular Biochemistry**, p. 1–16, 2018.

PEREIRA, G. R. C. et al. In silico analysis and molecular dynamics simulation of human superoxide dismutase 3 (SOD3) genetic variants. **Journal of Cellular Biochemistry**, v. 120, n. 3, p. 3583–3598, mar. 2019.

PEREIRA, G. R. C. et al. In silico analysis of the tryptophan hydroxylase 2 (TPH2) protein variants related to psychiatric disorders. **PLoS ONE**, v. 15, n. 3, p. 1–23, 2020.

PEREIRA, G. R. C. et al. Analysis of mutations in APP1 protein associated with development and protection against Alzheimer' s disease - an In silico approach. **Brazilian Journal of Development**, v. 8, n. 6, p. 46902–46924, 2022a.

PEREIRA, G. R. C. et al. In silico analyses of acetylcholinesterase (AChE) and its genetic variants in interaction with the anti - Alzheimer drug Rivastigmine. **Journal of Cellular Biochemistry**, n. May, p. 1–19, 2022b.

PEREIRA, G. R. C.; DE AZEVEDO ABRAHIM VIEIRA, B.; DE MESQUITA, J. F. Comprehensive in silico analysis and molecular dynamics of the superoxide dismutase 1 (SOD1) variants related to amyotrophic lateral sclerosis. **PLoS ONE**, v. 16, n. 2 February, p. 1–27, 2021.

PEREIRA, G. R. C.; TELLINI, G. H. A. S.; MESQUITA, J. F. De. In silico analysis of PFN1 related to amyotrophic lateral sclerosis. **PLOSONE**, v. 14, n. 6, p. e0215723, 2019.

PESIRIDIS, G. S.; LEE, V. M. Y.; TROJANOWSKI, J. Q. Mutations in TDP-43 link glycine-rich domain functions to amyotrophic lateral sclerosis. **Human Molecular Genetics**, v. 18, n. R2, p. 156–162, 2009.

PIKKEMAAT, M. G. et al. Molecular dynamics simulations as a tool for improving protein stability. **Protein engineering**, v. 15, n. 3, p. 185–192, 2002.

PRASAD, A. et al. Molecular mechanisms of TDP-43 misfolding and pathology in amyotrophic lateral sclerosis. **Frontiers in Molecular Neuroscience**, v. 12, n. February, p. 1–36, 2019.

ROSE, Y. et al. RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive. **Journal of Molecular Biology**, v. 433, n. 11, p. 166704, 2021.

ROY CHOUDHURY, A. et al. Supporting precision medicine by data mining across multi-disciplines: An integrative approach for generating comprehensive linkages between single nucleotide variants (SNVs) and drug-binding sites. **Bioinformatics**, v. 33, n. 11, p. 1621–1629, 2017.

SCIALÒ, C. et al. TDP-43 real-time quaking induced conversion reaction optimization and detection of seeding activity in CSF of amyotrophic lateral sclerosis and frontotemporal dementia patients. **Brain Communications**, v. 2, n. 2, p. 1–14, 2020.

SCOTTER, E. L.; CHEN, H. J.; SHAW, C. E. TDP-43 Proteinopathy and ALS: Insights into Disease Mechanisms and Therapeutic Targets. **Neurotherapeutics**, v. 12, n. 2, p. 352–363, 2015.

TAMAOKA, A. et al. TDP-43 M337V mutation in familial amyotrophic lateral sclerosis in Japan. **Internal medicine (Tokyo, Japan)**, v. 49, n. 4, p. 331–4, 2010.

VENSELAAR, H. et al. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. **BMC bioinformatics**, v. 11, n. 548, 2010. Disponível em: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-548>.

WATANABE, S. et al. ALS-linked TDP-43M337V knock-in mice exhibit splicing deregulation without neurodegeneration. **Molecular Brain**, v. 13, n. 1, p. 13–16, 2020.

WEBER, C. C.; WHELAN, S. Physicochemical amino acid properties better describe substitution rates in large populations. **Molecular Biology and Evolution**, v. 36, n. 4, p. 679–690, 2019.

WEI, Y.; THOMPSON, J.; FLOUDAS, C. A. CONCORD : a consensus method for protein secondary structure prediction via mixed integer linear optimization. **Proceedings of the Royal Society A**, v. 468, n. November 2011, p. 831–850, 2012.

WOOD, A. et al. Molecular Mechanisms Underlying TDP-43 Pathology in Cellular and Animal Models of ALS and FTLD. **International journal of molecular sciences**, v. 22, n. 9, 2021.

XU, D.; ZHANG, Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. **Biophysical Journal**, v. 101, n. 10, p. 2525–2534, 2011.

ZHANG, Y.; SKOLNICK, J. TM-align: A protein structure alignment algorithm based on the TM-score. **Nucleic Acids Research**, v. 33, n. 7, p. 2302–2309, 2005.