Cross-Beta database variables description:

ID: ID in the database in the same format for each entry (« DB00000 »)
ID composed of the character « DB » followed by the numeric ID of the entry

LABEL: AMYLOID
It tells us if the sequence is forming amyloids or not. Mainly used for machine learning applications.

Entry ID: 4-character code (letter or integer) or format as « AP00000 »
The PDB access code of the protein or AmyPro access code.

Cluster_ID: Integer code
The code corresponds to the attributed cluster (cluster made using the « CD-Hit » program). Entries sharing the same code belong to the same cluster.

Other ID: list of IDs (DB-ID_CHAIN-ID or DB-ID if no information about the chain)
Regroup all the formatted IDs sharing the same sequence with the current entry.

Reference Title: title of the article
Name of the article used to justify/create the entry on the original database (e.g. PDB or AmyPro)

Reference Authors: list of author's names of the article
List of author's name for the article used to justify/create the entry on the original database (e.g. PDB or AmyPro)

Reference link / DOI: Link / DOI to access the article
URL link or DOI used to access the article used to justify/create the entry on the original database (e.g. PDB or AmyPro)

Amyloid DB: Database name
The database name where the entry is coming from (e.g. PDB or AmyPro). If the entry is not coming from another database, the value will be '-' as well as the value of 'Entry_ID'. However, the entry will still be related to an accessible reference publication.

Protein Name: protein name
The name of the amyloid protein is given by the original database. For some cases, the name has been modified to include the specific form of the amyloid (e.g. Alpha-synuclein -> phosphorylated Tyr39 Alpha-synuclein).

Species/organism: organisms providing the protein
Name (binomial nomenclature) of the species or organism from which the given protein is taken.

Experimental Amyloid Region: start position-end position
Start and end position according to the position in the original protein. This position has been determined using a BLAST alignment of the AR on the complete protein accessible using a database like Uniprot or NCBI).

Predicted EAR: start position-end position
Start and End position according to the position in the complete protein of the predicted Exposed Amyloid Region (using ArchCandy2.0 provided by TAPASS).

Predicted AR: start position-end position
Start and End position according to the position in the complete protein of the predicted Amyloid Region (using ArchCandy2.0 provided by TAPASS).

Chain ID: one-letter code
In some cases, the PDB code leads to a structure formed from different elements with different sequences. To differentiate them, if different elements are coming from the same entry, they will be recognized by their chain ID.

Number of Structures: Number of different structures for the same entry ID
In some cases, the PDB code leads to a structure formed from different elements coming from the same original protein. In this case, the number of structures tells us how many different structures are stored in the same Entry ID. For Entries coming from the AmyPro Database, this number corresponds to the number of regions having the same entry ID in the AmyPro database.

AR containing protein Database: name of the database
The database where the full protein sequence can be found (e.g. Uniprot, NCBI).

AR containing protein Accession Code: Specific accession code linked to the Database name value
The accession code leads to the full protein sequence using the corresponding database given in the 'AR containing protein Database' column. This provides us with the sequence appearing in the 'AR containing protein sequence' field.

AR Sequence: amyloid amino acids sequence (1 letter code)
The sequence corresponds to the amyloid-forming region of the AR-containing protein. Cleaned from the N-term and C-term disordered regions if there were.

AR IDR: amino acid & start position-amino acid & end position
Intrinsically Disordered Region (IDR) inside the amyloid sequence if its structure is resolved or if the information is available. The region is given by the start amino acid (1 letter code) and position as well as the end amino acid (1 letter code) and position. This region is manually entered based on the observation of inconstant amyloid structure resolved in ss-NMR methods.

AR unresolved structure: amino acid & start position-amino acid & end position
Some resolved amyloid structures exhibit regions where the structure remains unresolved. Those regions are identified by the start amino acid (1 letter code) and position as well as the end amino acid (1 letter code) and position.

AR Sequence Length: number of amino acids
The number of amino acids present in the amyloid sequence. Recompute after the sequence cleaning operation.

AR Molecular weight kDa: computed molecular weight (in kDa)
The molecular weight of the amyloid sequence in kDa. Recomputed after the sequence cleaning operation.

AR Amino acids composition (vector): a 21-dimension vector of float values
The computed proportion (from 0 to 1) of each amino acid in the amyloid sequence. The values in the vector follow the given order:

['A', 'R', 'N', 'D', 'C', 'Q', 'E', 'G', 'H', 'I', 'L', 'K', 'M', 'F', 'P', 'S', 'T', 'W', 'Y', 'V', 'X']

Ligand(s): name and/or formula of present ligand(s)
Mark the name and/or formula of ligand(s) if one (or more) are present for the resolution of the structure.

Protein obtention method: Synthesis, extraction (from culture cell or tissue), purchased
Gives details on the source of the proteins studied (e.g. *E. coli* (BL21-DE3), human brain tissue, synthesis, …). When the proteins come from extraction, the buffer and the fibrillation/aggregation condition might not be accurate as this phenomenon is not studied in the article.

(Minimum/optimum) concentration for aggregation: Float number or range
The minimum required concentration, or the used concentration of the protein to form amyloid structures. When the fibrils are directly extracted from tissue, the value corresponds to the concentration in the final solution and might not correspond to the aggregation condition concentration.

pH (window/optimum) for aggregation: Float number or range
The pH window (min and max) or optimum pH value in which the amyloid structure can be formed. When the fibrils are directly extracted from tissue, the value corresponds to the pH in the final solution and might not correspond to the aggregation condition concentration.

Temperature (°C): Integer number
Used temperature to stimulate the formation of amyloids or to extract them from tissue. When the fibrils are directly extracted from tissue, the value corresponds to the temperature in the final solution and might not correspond to the aggregation condition concentration.

Specific buffer: common name (optional) and composition of the used buffer
The necessary specific buffer (if there is one) for the amyloid structure formation. When the fibrils are directly extracted from tissue, the value corresponds to the buffer in the final solution and might not correspond to the aggregation condition concentration.

Mutation(s): Common mutation notation
The mutation was identified using blast with the amyloid protein region against the original full protein sequence.

AR containing sequence: original full protein amino acids sequence (1 letter code)
The sequence of the original protein has been gathered using the accession code in the corresponding database.

AR containing protein length: number of amino acids
The length of the original full protein sequence.

AR Coverage (include EAR): Proportion of the amyloid-forming region in the full protein (0 to 1)
The amyloid proportion in the original protein was computed using the length of both sequences. This value doesn't take into consideration other entries even if they are another region of the same protein.

Fibrils substructure state: The fibrils substructure represented in the entry
Show if the amyloid fibrils tend to assemble into a homo (protofibrils share the same sequence and structure), hetero (protofibrils don't share the same sequence or structure) mono/di/trimeric (depending on the number of protofibrils forming the fibrils)

Method used: methods used for the entry
The method used to prove that the protein is forming amyloid structures (cross-beta structure) (e.g. Electronic Microscopy, Magnetic Nuclear Resonance, X-ray, Thioflavin T, congo red)

Role: disease-related, functional or both
Express if the amyloid form of the protein is disease-related, functional, or can have both forms in the organism. Can be unknown if the reference publication gives no explicit information about the pathogenicity of the studied amyloid.

IUPred3_score: Float 0 to 1
Score expressing the disorder prediction provide by IUPred3 on the AR-containing sequence. The score corresponds to the mean of the AA score in the experimental AR.

ESMFold_score: Float 0 to 100
Score expressing the confidence of ESM Fold for the AR contained in the AR containing sequence. In other words, ESM Fold predicts the structure of the AR-containing protein, and the final score is the mean of the score obtained by the AA in the AR.