Supplemental Materials

S.1 Statistics of classified domains in major resources for structure domain classification

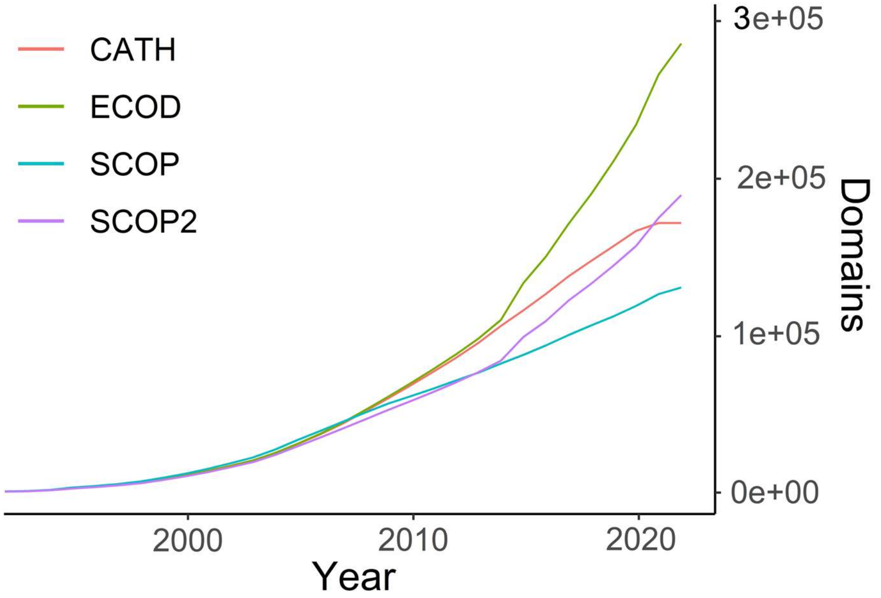

Figure S1. Classification of experimental structures in major domain classifications. Number of unique domains and release date of their parent PDB deposition cumulatively classified by SCOP/SCOPe, ECOD, SCOP2, and CATH between 1991-2021. Data acquired from RCSB web site. ECOD classifies more domains due to two reasons: first, it has been constantly updated with newly released structures with an attempt to classify every PDB entry, if possible; second, since ECOD captures more remote homology, it may tend to define smaller and more ancient evolutionary units as domains.

S.2 Properties of AlphaFold models that challenge our previous domain classification pipeline

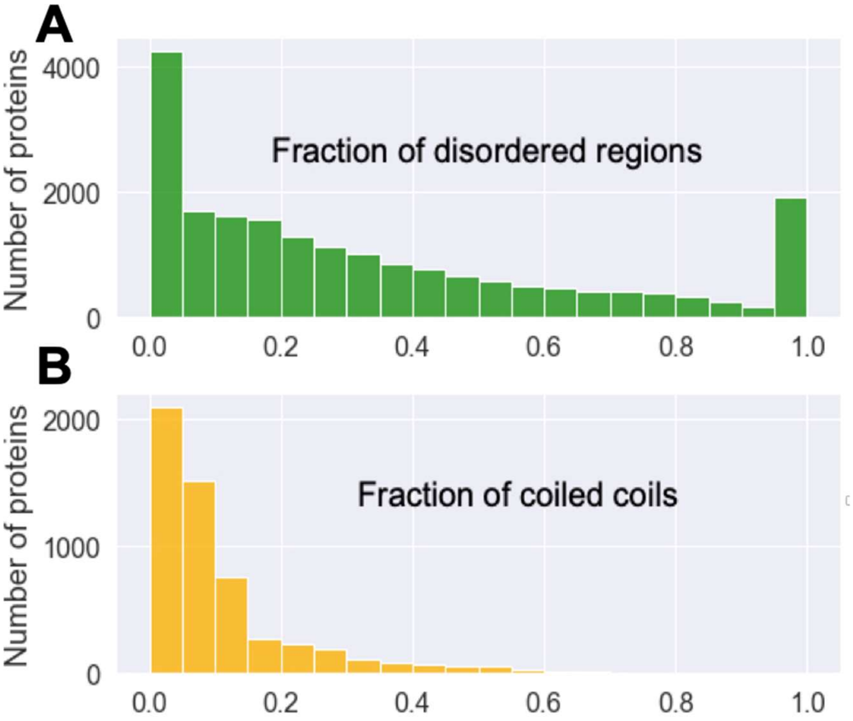

Figure S2. The fraction of (A) disordered regions (predicated by PAE) **and (B) coiled-coils** (predicted by NCOILS(1)) **in AF models of human proteins.** Proteins without coiled-coils are not included in **B**.

S.3 Potentially incorrect AF models

AF predicted a number of proteins with beta-helices corresponding to sequence repeats. Some of these predictions have relatively low confidence scores (pLDDT score < 70), such as the two uncharacterized proteins FLJ40521 (**Figure S3A**) and FLJ45999 (**Figure S3B**) and some cysteine-rich keratin-associated proteins (**Figure S3C**, Keratin5-4). Multiple mucins were also predicted to possess beta-helices, e.g., Mucin-22 (**Figure S3D**). Mucins contain sequence repeats enriched with serine and threonine, which are considered disordered regions and are heavily O-glycosylated *in vivo* (2). The beta-helical AF-models of these regions may not be correct or not the preferred conformation under physiological conditions. For example, the AF model of Mucin-21 beta-helix (**Figure S3E**) exhibits the highest structural similarity to a bacteria antifreeze protein (PDB: 3P4G, **Figure S3F**) (3). Inspection of the bacterial protein revealed multiple glycine residues occurring at both ends of the beta sheets (shown in magenta in **Figure S3E** and **S3F**). However, only a single glycine is found at one end of each repeat of the AF model of Mucin-21 beta-helix, suggesting that the AF model of Mucin-21 might be incorrect.

Another example of a potentially incorrect AF model is the protein PALM2 (Paralemmin-2). PALM2 has a closely related paralog, PALM1 (**Figure S4A**). The AF2-model of PALM1 adopts a FKBP-like fold. It has a C-terminal beta-strand (colored magenta in **Figure S4B**) in the core of the beta-sheet. This core structural element is also found in a chitinase insertion domain adopting the same fold (**Figure S4C**, PDB: 4URI) (4). However, the region in PALM2 homologous to the PALM1 last core beta-strand was not modeled as a core beta-strand. Instead, it is modeled to be in the disordered region (**Figure S4A**, magenta region). Long insertions between this last beta-strand region and the other part of the domain possibly caused AF to mismodel this region in PALM2.

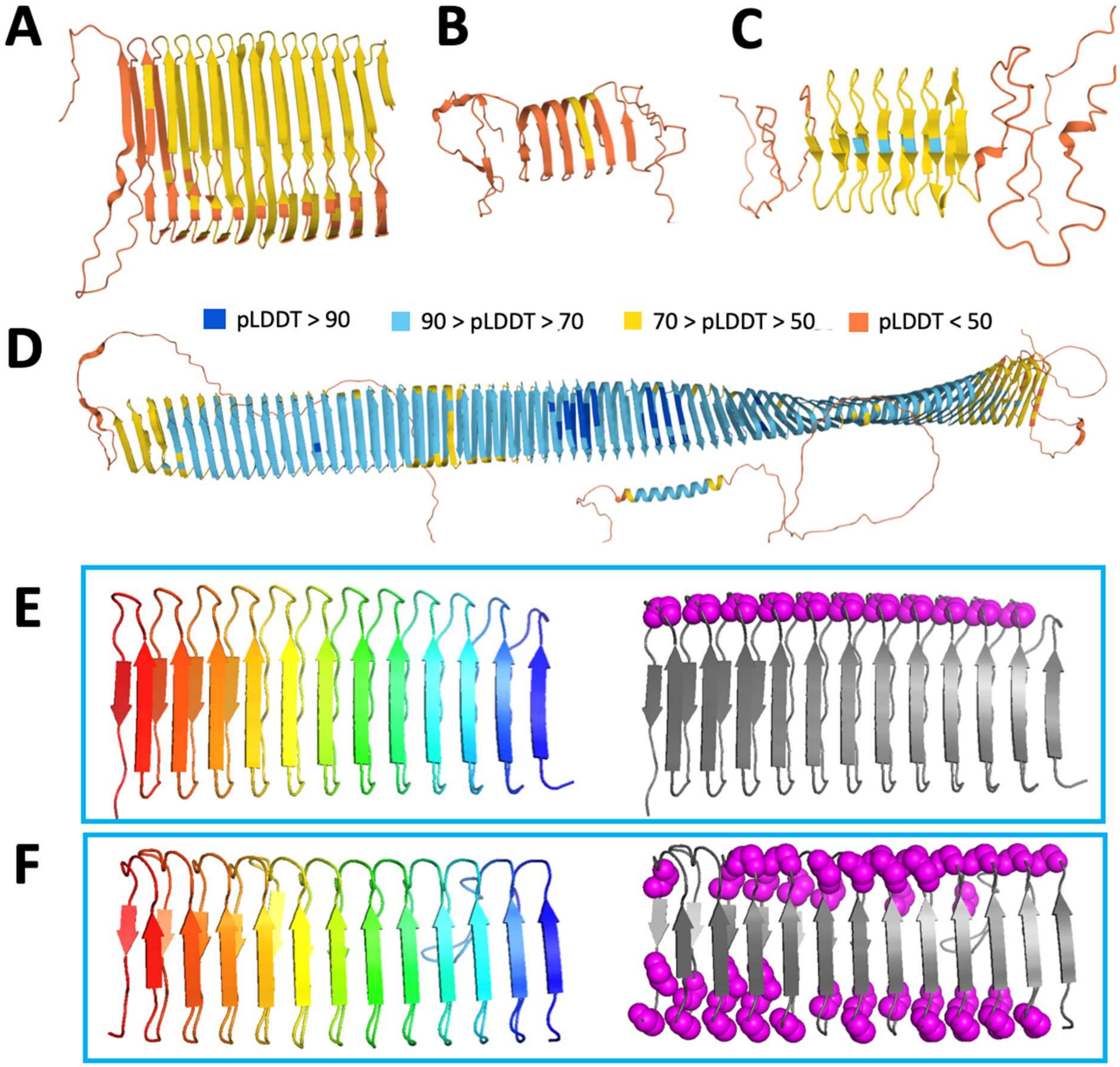

Figure S3. AF models showing beta-helical repeats. **(A)** Uncharacterized protein FLJ40521. **(B)** Uncharacterized protein FLJ45999. **(C)** Keratin-associated protein 5-4. **(D)** Mucin-22. These structures are colored according to the shown scale of AF pLDDT scores, which reflects per-residue confidence (pLDDT > 90 (dark blue): very high confidence; 90 > pLDDT > 70 (light blue): Confident; 70 > pLDDT > 50 (yellow): low confidence; pLDDT < 50 (orange): very low confidence). **(E)** Mucin-21 AF model colored in rainbow (left) and with glycines highlighted in magenta spheres (right). **(F)** The closest known structure (PDB: 3P4G) colored in rainbow (left) and with glycines highlighted in magenta spheres (right).

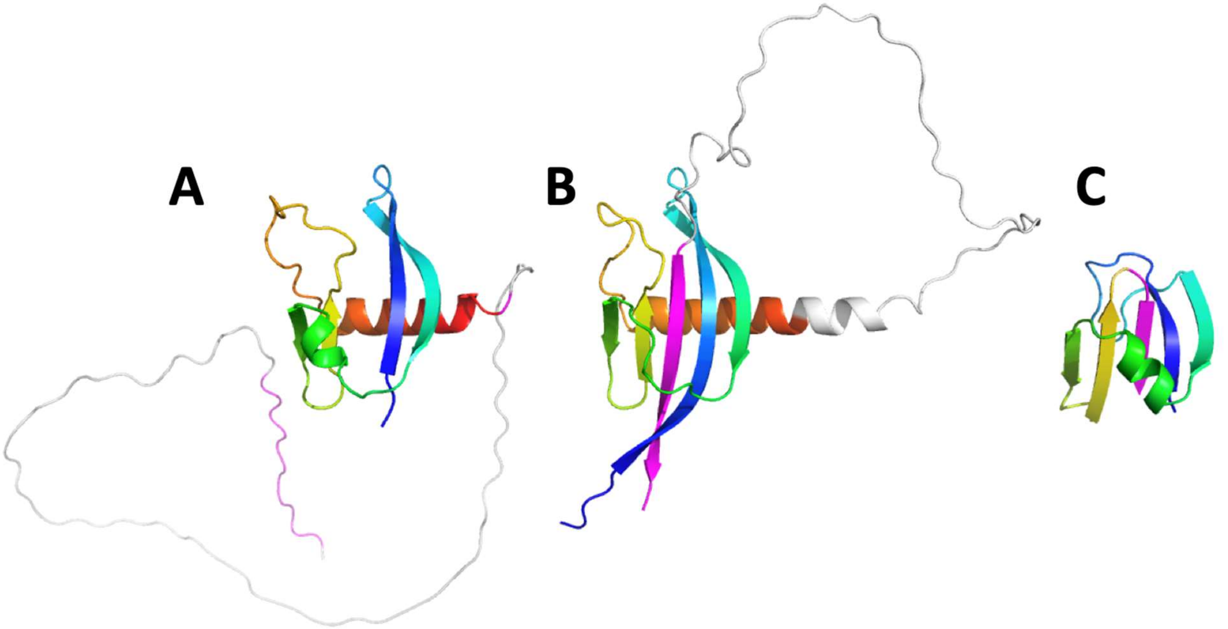

Figure S4. (A) The AF-model of PALM2. **(B)** The AF model of PALM1. **(C)** The structure of a Chitinase insertion domain from PDB: 4URI (chain B, residue range: 236-293).

S.4 Implication about retroviral genes displaying human-specific integration and application

Endogenous retroviruses in the human genome have accumulated mutations over time and are thought to lack infectivity (5). However, HERVK_113 has intact open reading frames for all its viral proteins, lacks nonsynonymous substitutions, can produce viral particles, and might be capable of infecting humans today (6, 7). HERVK viral particles are implicated in autoimmune disease and cancer (8). Another prominent example of human-specific domains of viral origin is the SCAN domains (belonging to the Retrovirus capsid protein-C group), domain-swapped dimers related to the HIV capsid C-terminal domain (9). The SCAN domains found in metazoan species have expanded in humans, being identified in 57 proteins that include mainly zinc fingers functioning in transcription regulation. These identified relationships highlight the potential ability of modern-day vertebrate proteins to evolve rapidly from the remnants of ancestral retroviral infection.

S.5 Additional details about the neural network to evaluate candidate reference ECOD domains

The architecture and training routine of our neural network is shown in **Figure S5,** which contains two dense layers of neurons before the output layers with two neurons representing the two possible categories: (1) the query and reference ECOD domains are from the same T-group, (2) the query and reference domains are from different T-groups. In addition, we probed other architectures and tested the influence of input normalization on the performance (loss after 30 epochs of training) of the neural network (**Figure S6**). Using two layers of neurons and normalizing the input data improved the performance of this simple neural network. However, further improving the number of neurons in a layer did not obviously improve the performance. Therefore, we chose to use a light-weight 2-layered neural network with 64 and 16 neurons, respectively.

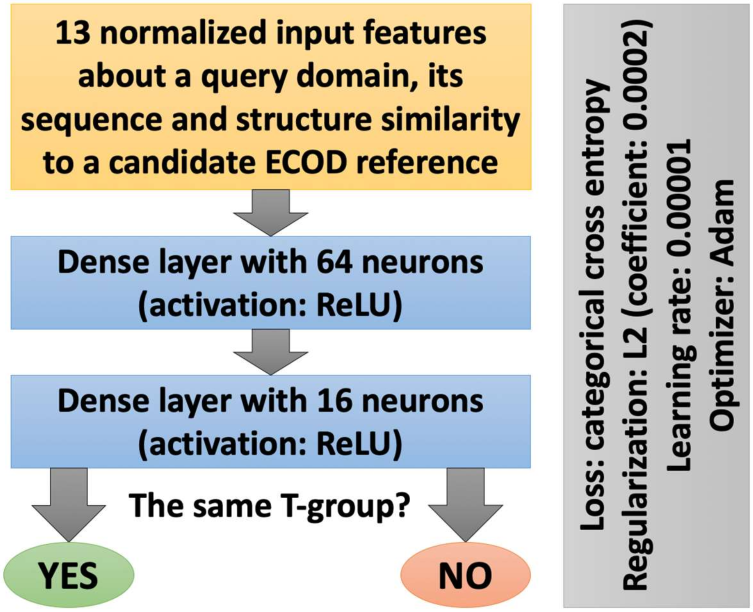
13 normalized input features about a query domain, its sequence and structure similarity to a candidate ECOD reference
Dense layer with 64 neurons (activation: ReLU)
Dense layer with 16 neurons (activation: ReLU)
The same T-group?
YES
NO
Loss: categorical cross entropy
Regularization: L2 (coefficient: 0.0002)
Learning rate: 0.00001
Optimizer: Adam

Figure S5. Architecture and training routine of DPAM classifier to assess if a reference ECOD domain is from the same T-group as the reference.

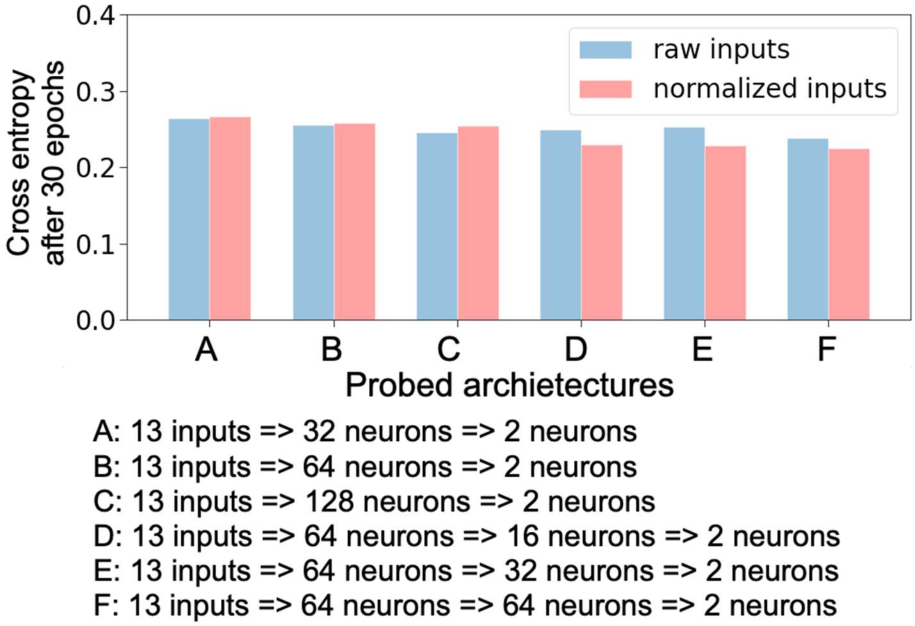

Figure S6. Performance of different architectures and influence of input normalization.

We performed 4-fold cross validation to train (with 3 quarters of data) and test (with 1 quarter of data) the neural network to distinguish a true reference ECOD domain from the same T-group as the query domain and a false reference ECOD domain from a different T-group as the query domain. For each of the testing set, the testing loss is always highly similar to the training loss, suggesting that over-fitting is not a problem (**Figure S7**). The output layer of our neural network contains two values, representing the probability for the hit ECOD domain to be a true and false reference for the query, respectively; we referred to the former as DPAM (domain parser for AlphaFold models) probability, which can be used to rank the candidate ECOD references detected through sequence and/or structure similarity searches. DPAM probability correlates with the fraction of correct ECOD reference domains at each DPAM probability range (**Figure S8**). We compared the performance of DPAM probability in ranking the true ECOD reference domains above false ECOD reference domains (**Figure S9**), and it shows superior performance than the well-established HHsuite probabilities and DALI Z-scores (components of our neural network).

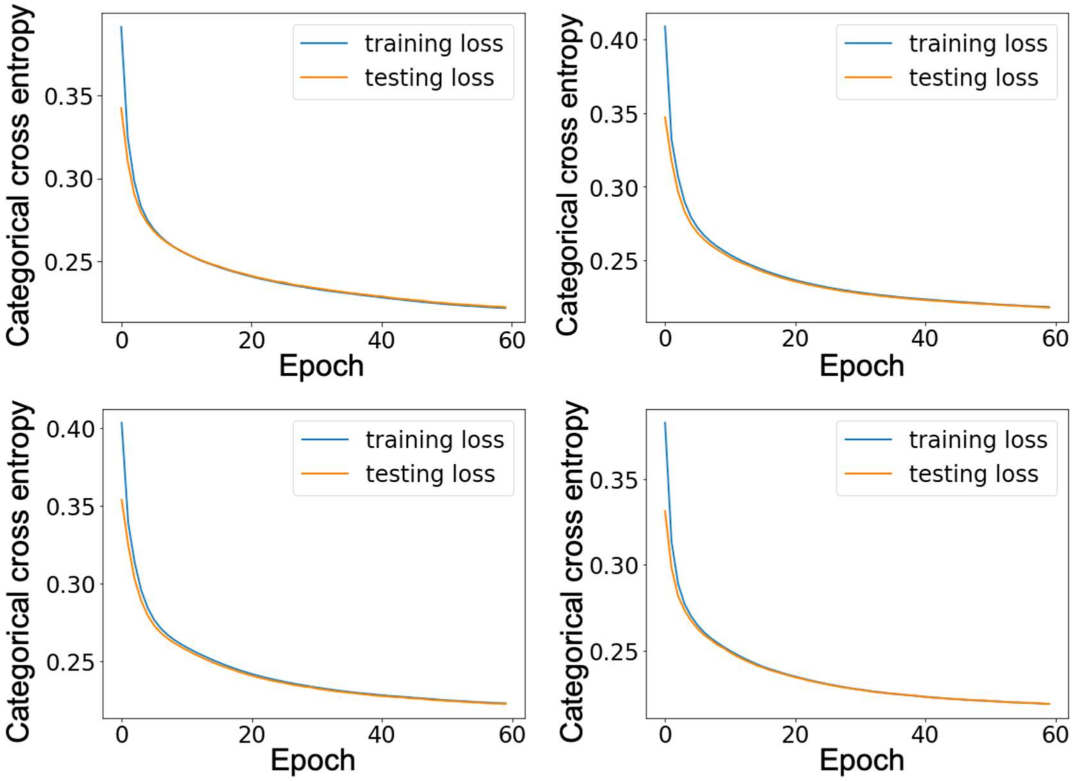

Figure S7. Training and testing loss for our neural network to evaluate whether an ECOD domain found by sequence or structure similarity searches is from the same ECOD T-group as the query domain. Each of plot corresponds to the result from one of the four-fold cross-validations. The training and testing loss at each epoch are highly similar, indicating the lack of over-training.

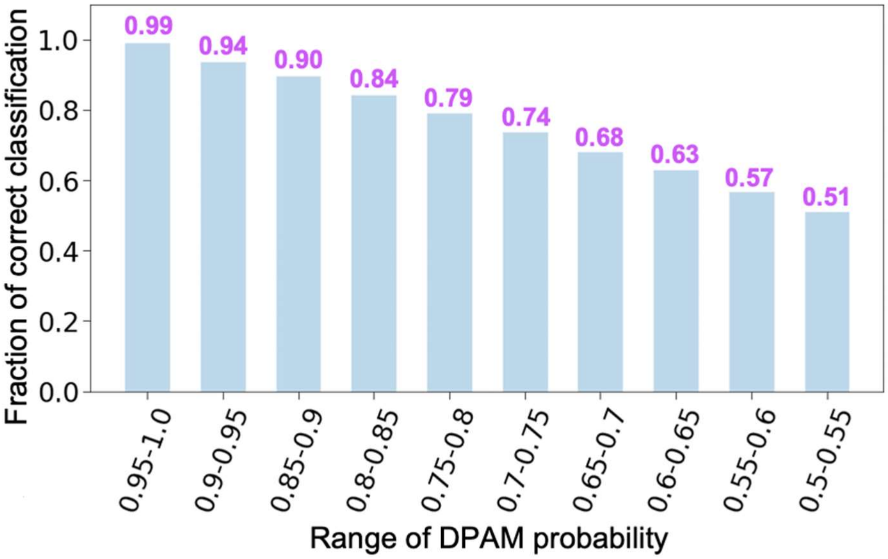

Figure S8. DPAM probability between a query domain and a candidate ECOD reference domain predicts the likelihood for this query domain to belong to the same T-group and the ECOD domain (and thus can be correctly classified based on this reference)**.**

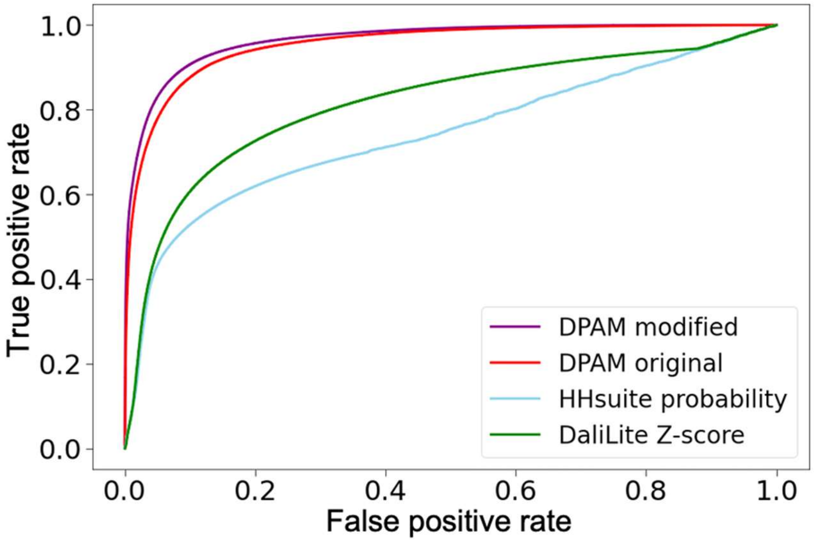

Figure S9. The performance (Receiver operating characteristic curves) **of DPAM probability** (blue), **HHsuite probability** (orange), **and Dali Z-score** (green) **in distinguishing correct ECOD reference domains** (belonging to the same T-groups as the query domains) **and incorrect ECOD reference domains** (belonging to different T-groups from the query domains)**.** The true positive and false positives were derived from the "acceptable HHsuite hits" and "acceptable Dali hits" we found for proteins in the DPAM benchmark set (details in (10)). If the hit and the query domains are from the same ECOD T-group, we consider the hit to be a true positive; if the hit and the query domains are from different ECOD T-groups, we consider the hit to be a false positive. "DPAM original" was the initial version of DPAM neural network we developed in the initial submission, which did not use normalized inputs and used 1 dense internal layer with 64 neurons. "DPAM modified" is the version we improved during the revision stage, which uses normalized inputs and 2 dense internal layers with 64 and 16 neurons, respectively.

S.6 Problems with our current DPAM pipelines and future directions

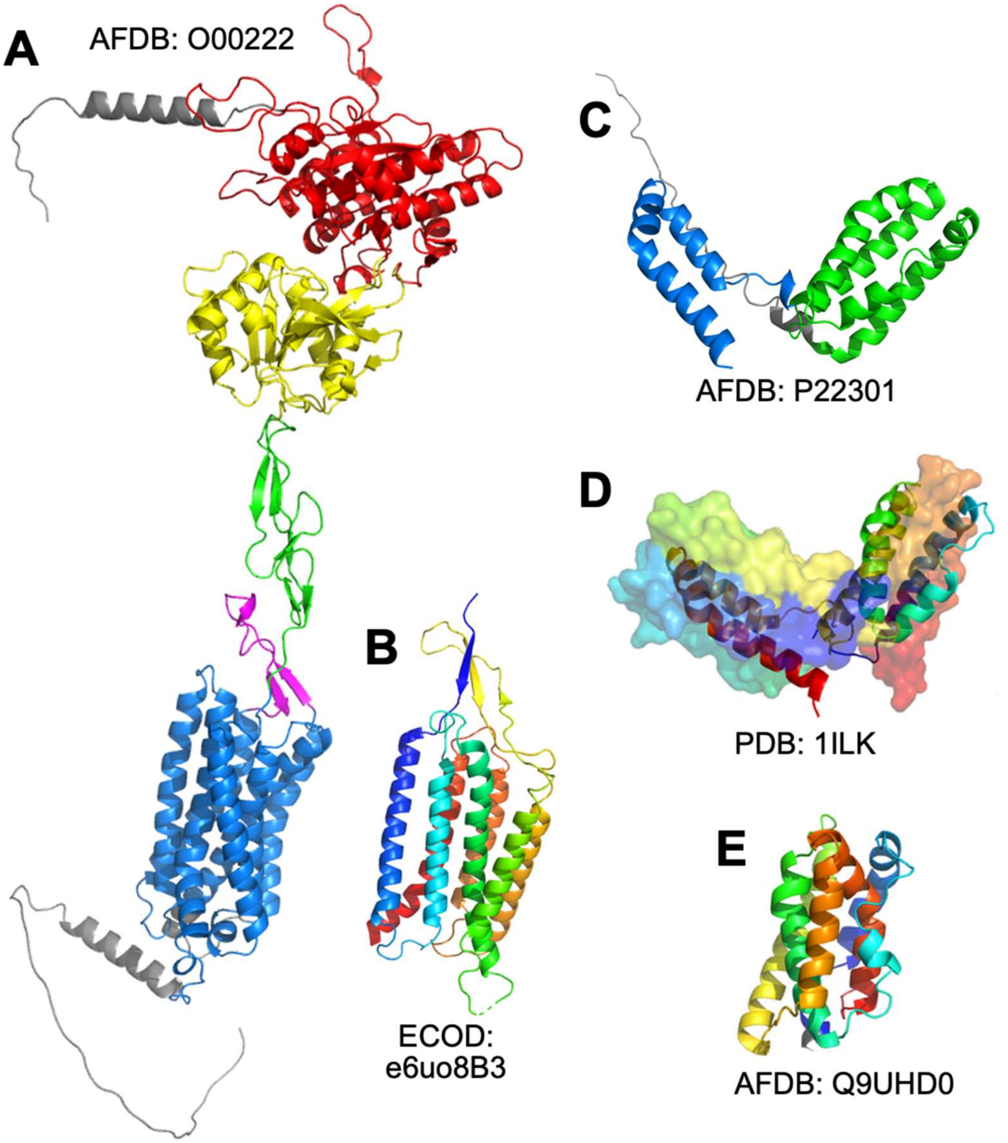

Figure S10. Examples where the DPAM procedure split one domain into two pieces because these domains are not structurally compact. (A) Parsed domains for an AF model of a glutamate receptor, and different domains are indicated by different colors. **(B)** A homologous domain from ECOD for the blue and magenta domains in **(A).** The blue and magenta domains in **(A)** should be merged to be consistent with the ECOD references; however, splitting them into two domains is sensible based on the 3D structure. **(C)** Parsed domains for an AF model of a cytokine, Interleukin-10. The two domains suggested by DPAM are colored in blue and green, respectively. **(D)** Experimental structure of Interleukin-10 reveals is a swapped dimer. Due to domain swapping, structure of Interleukin-10 is not as compact as other cytokines, such as Interlukin-19 in **(E).** DPAM split the domain-swapped Interleukin-10 into two domains based on structural features, but ECOD considered Interleukin-10 as a single-domain protein because the entire protein is a single evolutionary unit.

In the process of manual analysis of automatic classification results from our domain parser and the newly developed domain classifying neural network, we observed some minor problems which we will highlights the future directions. First, when domains are not globular, the regions that sticks away from the globular region might be parsed into a separate domain due to the use of physical properties in DPAM (**Figure S10**). We added an *ad hoc* filter to identify such cases and merge them. Meanwhile, we are developing a deep-learning-based domain parser for AlphaFold models, which will likely be able to handle such problems.

Second, our neural network currently evaluates the probability, i.e., DPAM probability, for a query domain from AlphaFold model to be in the same T-group as an ECOD reference domain from experimental structures. We hypothesized that we may obtain better performance if we introduce a "reciprocal" criterion, where we evaluate the probability for a pair of domains to be in the same T-group twice (A as query and B as reference, as well as B as query and A as reference). We tested this idea using pairs of predicted AlphaFold domains that both can be mapped to existing ECOD domains (sequence identity > 95% and coverage > 60%) from our DPAM benchmark set (10). When requiring a pair to show DPAM probability above 0.8 in both directions, 99.9% of such pairs are from the same ECOD T-group. However, for pairs showing DPAM probabilities above 0.8 in only one direction, only 86.5% are mapped to the same ECOD T-groups (**Figure S11**).

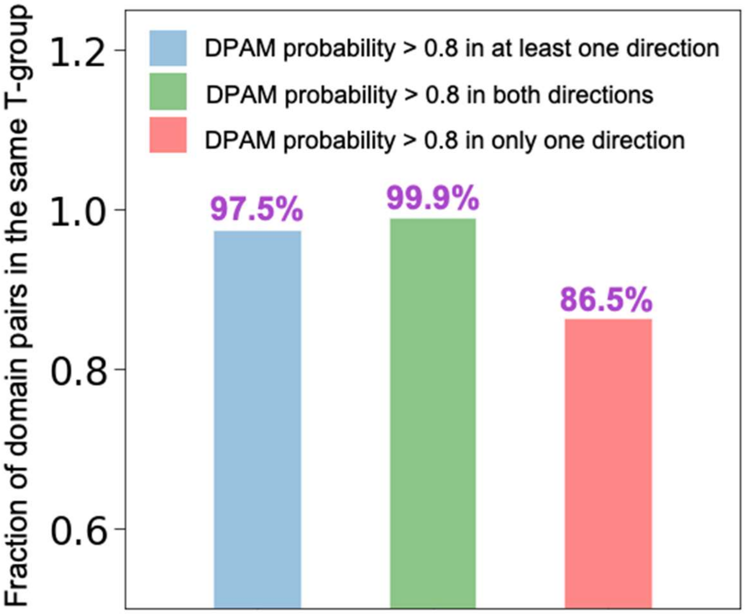

Figure S11. Application of DPAM classifier in a reciprocal fashion can improve the confidence of domain classification.

References

1. A. Lupas, M. Van Dyke, J. Stock, Predicting coiled coils from protein sequences. *Science* **252**, 1162-1164 (1991).
2. S. E. Baldus, K. Engelmann, F. G. Hanisch, MUC1 and the MUCs: a family of human mucins with impact in cancer biology. *Crit Rev Clin Lab Sci* **41**, 189-231 (2004).

3. C. P. Garnham, R. L. Campbell, P. L. Davies, Anchored clathrate waters bind antifreeze proteins to ice. *Proc Natl Acad Sci U S A* **108**, 7363-7367 (2011).
4. G. Sulzenbacher *et al.*, Structural basis for carbohydrate binding properties of a plant chitinase-like agglutinin with conserved catalytic machinery. *J Struct Biol* **190**, 115-121 (2015).
5. M. Barbulescu *et al.*, Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr Biol* **9**, 861-868 (1999).
6. G. Turner *et al.*, Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr Biol* **11**, 1531-1535 (2001).
7. K. Boller *et al.*, Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. *J Gen Virol* **89**, 567-572 (2008).
8. Y. Khalfallah, A. Genge, HERV-K inactive or potential pathogens from within? *J Neurol Sci* **423**, 117359 (2021).
9. D. Ivanov, J. R. Stone, J. L. Maki, T. Collins, G. Wagner, Mammalian SCAN domain dimer is a domain-swapped homolog of the HIV capsid C-terminal domain. *Mol Cell* **17**, 137-143 (2005).
10. J. Zhang, R. D. Schaeffer, J. Durham, Q. Cong, N. V. Grishin, DPAM: A Domain Parser for AlphaFold Models. *Protein Sci* 10.1002/pro.4548, e4548 (2022).