Data-driven large-scale genomic analysis reveals an intricate phylogenetic and functional landscape in J-domain proteins

Duccio Malinverni[a,b*], Stefano Zamuner[c*], Alessandro Barducci[d], Nadinath B. Nillegoda[e,1], Paolo De Los Rios[c,f,1]

[a] MRC Laboratory of Molecular Biology, Cambridge, UK

[b] Department of Structural Biology and Center for Data Driven Discovery, St Jude Children's Research Hospital, Memphis, TN, USA

[c] Institute of Physics, School of Basic Sciences, École Polytechnique Fédérale de Lausanne – EPFL, CH-1015 Switzerland.

[d] Centre de Biochimie Structurale (CBS), INSERM, CNRS, Université de Montpellier, Montpellier, France.

[e] Australian Regenerative Medicine Institute (ARMI), Monash University, Melbourne, Victoria, Australia

[f] Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne – EPFL, CH-1015 Switzerland.

[*] The authors contributed equally to this work

[1] Correspondence: nadinath.nillegoda@monash.edu , paolo.delosrios@epfl.ch

Abstract

The 70 kDalton Heat shock protein (Hsp70) chaperone system is emerging as a central node of the proteostasis network that supervises protein homeostasis in all organisms. Hsp70 recruitment to perform a vast array of different cellular functions is regulated by a family of co-chaperones known as J-domain proteins (JDP) that bear a small namesake J-domain, which is required to interact and drive the ATPase cycle of Hsp70s. Both prokaryotic and in eukaryotic JDPs display a staggering diversity in domain architecture (besides the ubiquitous J-domain), function and cellular localization. This is mirrored by a comparatively small number of Hsp70 paralogs, suggesting a degree of specificity, but also of promiscuity, in their mutual partnering. Very little is known about this family of cochaperones, despite their essential role in cellular proteostasis, development and the link to a broad range of human diseases. The numbers of JDP sequences identified as a consequence of efficient sequencing techniques has exponentially increased where it is now beyond the ability of careful manual curation. In this work, we first provide a broad view of the JDP repertoire accessible from public databases, and then we then use an automated classification scheme, based on Artificial Neural Networks (ANNs), to demonstrate that the sequences of the J-domains carry sufficient information to recover with high reliability the phylogeny, localization and domain composition of the corresponding full length JDP. Exploiting the interpretability of the ANNs employed, we find that several of the sequence positions that are most useful for the classification match to residues that form the interaction interface between the J-domain

and Hsp70, revealing that a set of residues within the J-domains have co-evolved with their obligatory Hsp70 partners to build chaperone circuits for specific functions.

Keywords: J-domain proteins, Hsp40 Chaperones, Protein Homeostasis, Large-scale Data Analysis, Artificial Neural Networks

Introduction

The Hsp70 is an essential and highly ubiquitous chaperone involved in a plethora of constitutive and stress-related cellular processes including promoting protein folding (*1*, *2*), preventing and reverting the formation of cytotoxic aggregates(*3–7*), importing polypeptides into cellular compartments (*8*, *9*), driving the (dis)assembly of oligomeric protein complexes (*10–12*), regulating protein activity (*13*), and targeting terminally damaged proteins for degradation (*14–16*). Underlying these diverse functions is the fundamental ability of Hsp70 to bind substrate proteins upon a conformational change that depends on the nature of the bound nucleotide, which alternates between ATP and ADP. This biochemical cycle is stringently regulated by JDPs that, together with the substrate, accelerate ATP hydrolysis in Hsp70s by orders of magnitude (*17*, *18*), which results in ultra-affinity binding, namely a significant, non-equilibrium enhancement of the Hsp70 affinity for its substrates (*19*). Subsequently, Nucleotide exchange factors (NEFs) of Hsp70 binds and accelerate the release of ADP thus allowing rebinding of ATP and client protein release (*20*) (Fig.1A).

Besides catalyzing ATP hydrolysis, JDPs are responsible for the diversity of the functional roles of Hsp70s owing to their multidomain architecture (*21*, *22*). They invariably contain their J-domain namesake, which is a ~63 residue-long domain formed by a well-defined arrangement of four alpha helices (Fig.1B) with a highly-conserved Histidine-Proline-Aspartate (HPD) motif between helices 2 and 3. The HPD motif forms the primary interface for the interaction and stimulation of ATP hydrolysis in Hsp70s (*17*, *18*, *23–26*). The remaining subdomains help target JDPs to specific locations, client proteins or protein conformations (*e.g.* unfolded, misfolded or aggregated polypeptides), thereby recruiting the action of Hsp70s to selected substrates.

A classification scheme for JDPs has been suggested based on the overall domain architecture, resulting in three main classes (*21*, *27*, *28*). Class A JDPs share the domain architecture of the prototypical DnaJ member of *Escherichia coli*. Members of this class are characterized by a canonical domain architecture comprising of an N-terminal J-domain, followed by a glycine and phenylalanine (G/F) rich region and two successive C-terminal β-sandwich domains (CTDs), with a zinc-finger-like region (ZFLR) domain protruding from the first CTD, and are terminated by a short dimerization domain at the C-terminal (Fig.1C). The presence of an N-terminal J-domain fused with a G/F region was used as the primary domain signature to segregate some of the remaining JDPs to form the so-called class B. This relatively

lose classification step, probably stemming from limited availability of JDPs sequences at the conception of this class, comprises both the "canonical" type members that share the same domain architecture of class A with the exception of missing the ZFLR (Fig.1D), alongside more diverse ones, that are characterized by an N-terminal J-domain followed by a G/F rich region-only (*28*). The G/F-rich region of both classes is typically considered to be largely unstructured, although it has been recently observed in class B that a small helix in the G/F region docks on the J-domain, likely with functional consequences (*29–32*). Furthermore, it has been recently observed that some class A and class B JDPs can undergo liquid-liquid phase separation to some degree with the help of the G/F regions (*33*).These results suggest that the G/F region might play one, if not several, functional roles. All the JDPs belonging to class A and some of the ones belonging to class B (also known collectively as Hsp40s owing to their approximate 40 kDa molecular mass) constitutively form homo-dimers through the C-terminal dimerization domain (Fig.1C and D), and bind a large variety of unfolded, misfolded or aggregated substrates through one or multiple interaction sites located on their CTDs and, possibly for class A, on the ZFLR (*34*, *35*). Thus, class A and B JDPs mediate the localization of Hsp70s onto their misfolded/aggregated substrates and contribute to Hsp70 crowding on the substrate (*4*). In contrast, class C JDPs display a high variability in domain architectures (two examples are shown in Fig.1E and F) and have been historically classified by what they are not, namely neither A nor B, rather than by what they are. Thus, members of class C share with each other little or no common functional or phylogenetic features, except for the presence of a J-domain.

Here we leveraged the recent rapid growth in available genomic data and the improvements in homology detection tools to explore the relations between JDP sequences and their phylogenetic and functional classification at various resolutions, with particular emphasis on subclasses of class C JDPs, uncovering their most common domain architectures. We then used Artificial Neural Networks (ANNs) to show that even the relatively short and simple sequences of the J-domains generally carry enough phylogenetic and functional signatures to reliably identify the architecture of the corresponding full length JDP, its cellular localization and the taxonomic clade. These findings allowed us to use amino acid sequence variations embedded in the J-domain to develop an ANN-based automated solution to curate existing and new JDP sequences reported in various organisms with high precision. By further identifying the residues that are most important for the classification, we pinpointed the structural positions where the class- and organism-specific signatures localize.

From a more general perspective, we highlight how the combination of interpretable large-scale sequence analysis techniques and of coarse-grained annotations can unlock the wealth of information that would otherwise remain hidden in exponentially growing, and thus manually unmanageable, sequence databases.

Results

Construction of an extensive annotated J-domain protein sequence dataset

The number of deposited JDP sequences in the UniProt repository (*36*) has been exponentially growing over the past two decades, increasing by an order of magnitude every 6-7 years (Fig.2A), ranging from 10^3 in year 2003 to 10^5 in year 2018, and projected to reach 10^6 sequences in years 2023-2024 and 10^7 by year 2030. Using stringent homology-search criteria, we extracted a total 223'124 JDP sequences from existing genome databases (see Methods). The associated annotations allowed us to assign all JDPs to different phylogenetic classes at various taxonomic levels. Using sequence domain models from the Pfam database (*37*), we further characterized the domain composition of the full-length sequences, assigning to each JDP a tag describing the domain architecture of the protein. Protein-localization prediction algorithms were used to infer the most probable organelle of residency of eukaryotic JDPs (see Methods). This compound procedure allowed us to assign to each JDP sequence in the dataset a set of *ground-truth* attributes comprising phylogenetic, architectural and cellular localization features. In this context, the features describing for instance the human DnaJA1 are, alongside its sequence, that it belongs to class A JDPs (based on the full-length domain composition), that it is a eukaryotic/metazoan/mammalian protein, and that it is predominantly cytoplasmic.

The correct identification and proper functional assignment of such a plethora of JDPs is, and will increasingly be, incompatible with any careful manual curation. Instead, it will require rigorous, but automated characterization and analysis approaches, trained on extensive datasets endowed with some reliable ground truth, defined according to different, complementary criteria. Being an instance of such dataset, the vast and richly annotated JDP repertoire collected in this work allowed us to investigate the ability of these automated approaches to extract and rationalize the information it contains, while gaining new insights into this complex chaperone family.

Overall characterization of the diversity of the J-domain protein dataset

Based on the overall domain composition of JDPs, the three established functional classes were clearly identifiable: Class A and Class B (members with the canonical Hsp40 domain architecture, Figs.1C and D) represented 24.4% and 13% of the dataset, respectively (Fig.2B), with the remaining JDPs assigned to class C. We found that there are 1945 different domain architectures that compose the ~140’000 class C JDPs, with roughly half of them (78’282) have a domain-architecture consisting, according to the Pfam-based annotation, only of the J-domain (Fig.2F). Collectively, the 12 most populated domain-architectures (including class A and B JDPs) made up roughly 90% of the total dataset (196’719 sequences). Visual inspection of class A and B JDP sequences showed that beyond the clear difference due to the presence of the ZFLR, we could distinguish sequences belonging to the two classes by the composition of their G/F-rich regions. Indeed, the amino-acid profiles of residues in the G/F regions of classes A and B are remarkably different, with the class A G/F-rich regions showing a more marked

145 segregation of phenylalanine, arginine, glycine, and glutamate residues compared to the ones of class
146 B JDPs (Fig.2C and D). This was quantitatively confirmed by a simple statistical model which showed
147 that the amino-acid sequences of the G/F region alone were able to successfully discriminate class A
148 and B JDPs (data not shown). This significant difference further supports a functional role played by
149 the G/F-rich regions, beyond being simple flexible linkers.

150 Agreeing with previous observations (*22*), class C members dominated the dataset (Fig.2B) mainly due
151 to their overwhelming abundance in eukaryotes (Fig.2E and Supplementary Fig.1A for a more detailed
152 analysis of class C abundance in the dataset). The 1945 different architectures present in class C were
153 built from 1558 distinct domains, often classified as Domain of Unknown Function (DUFs), with little
154 co-occurrence of the most common ones (Supplementary Fig.2). The only notable exceptions are JDPs
155 containing multiple tetratricopeptide-repeats (TPR) associated to several different Pfam families (in
156 Fig.2F and Supplementary Fig.2 all these TPR domains are collected into a single group). Inspection of
157 the most frequently observed JDP domain architectures within class C revealed a very disparate
158 phylogenetic makeup (Fig.2F). For instance, JDPs of the bacterial (*E. coli*) DjlA family, which formed
159 the most abundant class C-subtype architecture, are present only in bacteria. The HscB family (human
160 DnaJC20, Fig.1F), which participates in the universal iron-sulfur cluster assembly pathway, is instead
161 present in both prokaryotes and eukaryotes. Others, such as the endoplasmic reticulum (ER)-localized
162 yeast Sec63 (human DnaJC23), DnaJC18 (which shares almost the same domain composition as
163 DnaJB12 and DnaJB14), DnaJC11 (DUF3995) and DnaJC3/DnaJC7 were found only in eukaryotes.
164 Some highly represented architectures obeyed an even finer phylogenetic origin, with Djp1/Caj1, JJJ1
165 and JJJ3 (yeast names) present only in yeast (*e.g.* in *S. cerevisiae*).

166 We then analyzed the predicted subcellular localization of JDPs to assess if the classes are differently
167 represented in various cellular compartments. The analysis of the organellar makeup of the dataset
168 revealed that mitochondrial JDPs were mostly of class A and C (Supplementary Fig.1B) while
169 mitochondrial Class B JDPs were only about $1/10^{th}$ of the class A JDPs, compatibly with the precision
170 of the localization algorithm itself (see Methods). This suggested that mitochondrial Hsp40s belong to
171 class A, with possibly some exceptions to be manually controlled. ER-localized JDPs showed an
172 intriguing class distribution. Besides class C JDPs, the ER contains a resident Hsp40, which in Fungi
173 (and some other unicellular eukaryotes such as amoebas) was a canonical class A, whereas in Metazoa
174 (DnaJB11) and in plants (ErdJ3) it lacked the ZFLR, suggesting that it belonged to class B. Nonetheless,
175 a closer inspection revealed a non-canonical cysteine-rich region (which we call here *pseudoZFLR*)
176 where the customary ZFLR domain of class A JDPs is usually found. As these pseudoZFLR display
177 regularity and high degree of similarity in plants and metazoa (Fig.2G), they represent an interesting
178 evolutionary conundrum, given that the most accepted view of the evolution of eukaryotes posits that
179 the divergence between fungi and metazoa took place after the earlier divergence of plants.

180

181 ***J-domain sequences contain functional and phylogenetic information at multiple scales.***

182 Organisms ranging from viruses and prokaryotes to eukaryotes rely on multiple Hsp70-JDP pairs to
183 bring about a vast array of biological activities. What largely remains unclear is how a particular JDP
184 recognizes its partner Hsp70, an interaction which is largely mediated by the J-domain. It has been
185 speculated for many years that the specificity should thus be encoded in the sequences of both partners,
186 calling into question the degree to which J-domains can be interchanged between different JDPs with
187 minimal consequences to the J-domain's primary function. Indeed, not all J-domains indifferently
188 interact with Hsp70s. Exchanges of J-domains between two JDPs frequently result in non-functional
189 chimeras, suggesting that there is selectivity in JDP-Hsp70 pairing (*38–42*). Evolutionary factors such
190 as phylogeny, the necessity to select different Hsp70 paralogs, as well as the simple impact of the co-
191 localization of J- and non-J-domains in the same polypeptide, may have left traces in the sequence of
192 the J-domain. If this was the case, the sequence of a J-domain alone should carry relevant information
193 about the overall JDP it belongs to.

194 To test this hypothesis and provide a pictorial hint that the sequences of J-domains carry a rich source
195 of disparate evolutionary and functional signatures, we first extracted and aligned 223'124 J-domain
196 sequences from our JDP repertoire (see Methods for the detailed procedure). We then used the
197 unsupervised UMAP algorithm (*43*) to provide a (non-linear) projection of the sequences in two
198 dimensions, allowing for their direct visual inspection (Fig.3; see Methods). The projections revealed
199 that the J-domain sequences are gathered in clusters, and that the different groups, highlighted by
200 different colors, often associated well with various classifications schemes, such as their *i*) A, B or C
201 class (Fig.3A), *ii*) bacteria rather than eukaryotes (Fig.3B) *iii*) various C subclasses (only the five more
202 abundant ones are represented in Fig.3C for sake of clarity). To highlight the finer signatures that are
203 encoded in the J-domain sequences, we focused on HscB. Besides being well-clustered according to
204 the previous criteria (circles in Figs.3A, B and C), the HscB sequences from fungi, metazoa and plants
205 were also clearly individually well-grouped (Fig.3D). Thus, different, and independent layers of
206 information are simultaneously encoded in the sequences of the J-domains, even at finer scales or for
207 other criteria, such as subcellular localization (Supplementary Fig.3).

208 ***Artificial neural network classification of J-domains.***

209 Next, we used an approach based on Artificial Neural Networks (ANN) to find the detailed determinants
210 in the J-domain sequences that underpin the different classifications. We focused on a few broad
211 category discriminations, namely:

212 - Class A *vs.* class B (henceforth referred to as A/B);
213 - A *vs.* B *vs.* C (henceforth referred to as A/B/C);
214 - 12 different architectural classes, including the architectures of classes A and B JDP, the 9 most
215 populated C subclasses (excluding JDPs containing only the DnaJ domain) and an additional

"Other" class generically comprising all remaining JDPs (again not including JDPs containing only DnaJ domain) (henceforth referred to as Arch12);

- Bacteria *vs.* eukaryotes *vs.* archaea *vs.* Viruses (henceforth referred to as Bac/Euk/Arch/Vir);
- Alphaproteobacteria *vs.* gammaproteobacteria *vs.* firmicutes *vs.* metazoa *vs.* plants *vs.* fungi (henceforth referred to as Phylo);
- Endoplasmic reticulum *vs.* mitochondria *vs.* plastids (henceforth referred to as Localization).

Although these different groupings are far from exhaustive, we chose them as representatives of both complementary and orthogonal criteria

ANNs are often considered black-boxes with high performance for their assigned tasks, but poor interpretability. To circumvent this limitation, we employed a mathematical approach that makes ANNs interpretable, thus gaining a deeper insight into the information encoded in the J-domain sequences. The sequence of operations performed by ANNs is shown in Fig.4A (and further detailed in the Methods). The nodes (neurons) of the input layer receive the J-domain sequences in a suitable mathematical encoding, $\{x_i\}$. The nodes of the first *hidden* layer receive *affine transformations* $y_j^{[1]}$ of the input values, with weights $\left\{w_{ji}^{[in,1]}\right\}$ and offsets $\left\{b_j^{[in,1]}\right\}$: $y_j^{[1]} = \sum_i w_{ji}^{[in,1]} x_i + b_j^{[in,1]}$ (subscripts indicate the connected nodes in the two layers, and the superscripts in the square brackets indicate the corresponding layers). The output of the first hidden layer is then computed by applying a non-linear function to each of the $\left\{y_j^{[1]}\right\}$, resulting in a set of values $\left\{f\left(y_j^{[1]}\right)\right\}$, which is then passed to the nodes of the second hidden layer after undergoing the same set of operations. This procedure is repeated from layer to layer until the transferred values reach the nodes of the output layer that are in number equal to the different classes to be recognized (*e.g.* 3 output nodes for A/B/C, or 13 output nodes for Arch12). The output values are eventually transformed into the predicted probability that the input sequence $\{x_i\}$ belongs to each given class.

The training of the ANNs consisted in progressively modifying the network parameters (weights and offsets) until the output probabilities of the sequences of the dataset were, on average, as similar as possible to the ground truths established previously (Fig.4B). For example, in the A/B/C classification task, the J-domain of HscB belongs to class C with certainty, corresponding to probabilities (0,0,1), which is the set of values that the ANN should reproduce. To prevent the networks from simply memorizing features "by heart" and force the networks instead to learn useful information, the dataset was split in two parts. One part was used for the training, while the other (test set) was used to check the predictions on sequences not seen during the learning phase. An ANN could be reliably used as classification tool only if the performance over the training set was comparable to that over the test set.

For each one of the tasks (A/B, A/B/C, *etc.*), the output probabilities were used as coordinates for a low-dimensional representation of the dataset, similarly to the analysis performed in Fig.3. As an example, this kind of representation is shown in Fig.4C in the case of the classification task A/B/C

251 (including a further dimensionality reduction from three to two dimensions using the UMAP algorithm),
252 providing evidence that the ANN had found information embedded in the J-domain sequence that
253 allowed it to easily distinguish the three classes.

254 To move beyond a visual, but not quantitative appraisal of the performance of the ANNs, we computed
255 their *confusion matrices* (Fig.4D), which report the misclassification of each category with respect to
256 their *ground truth*. In the case shown in Fig.4D, for example, J-domains belonging to JDPs of class A
257 were correctly classified 97% of the times, and misclassified as class B and C members only 2% and
258 1% of the times, respectively. Similarly, class B J-domains were classified according to their ground
259 truth 96% of the times, whereas class C J-domains were the ones where the ANN was more *confused*
260 (hence the name of confusion matrices). More generally, the performance of the ANNs for all
261 classification tasks was excellent (Supplementary Fig.5). The only classification task that was slightly
262 less accurate was Localization, although it is important to recall that the accuracy in this case was
263 compatible with the prediction accuracy of the algorithms used to define the localization ground truth
264 (see Methods). Overall, the success of our ANN classification approach highlighted that a myriad of
265 phylogenetic and functional signatures is indeed encoded in the J-domain sequences alone, and that this
266 methodology can be used to tell J-domains apart from each other depending on different, possibly
267 orthogonal, criteria such as phylogeny, subcellular localization, and domain organization.

268 ***Most relevant discriminative sequence positions and their structural characterization.***
269 We next used the interpretability of the ANNs to identify the sequence positions that were most relevant
270 for each classification task (see Methods) by assigning them a *relevance score* (*44*). We uncovered that
271 although several of the sequence positions were expectedly task-specific, some were ubiquitously
272 important for all tasks (heatmap in Fig.5A). The average of the relevance score over different tasks
273 highlighted the positions that were most relevant to classify J-domains in general. We thus mapped the
274 five top-scoring positions on the structure of the J-domain of *E. coli* DnaJ in the DnaJ-DnaK (*E. coli*
275 Hsp70) chaperone complex (Fig.5B) (*24*). Importantly, we observed that the five top-scoring positions,
276 located on helix II and II of the J-domain, overlapped with the interaction surface between the J-domain
277 of DnaJ and DnaK (*24*), thus likely capturing some modulation of the binding of the J-domain to
278 Hsp70s, related to the HPD-dependent tuning of the Hsp70 ATPase activity. Other, slightly less
279 relevant, positions located on helices II and IV were also found in contact with DnaK (not shown in
280 Fig.5B for clarity).

281 The relevance of residues that participate to the surface of interaction with DnaK, although not
282 completely unexpected, is an intriguing result, because at no step of our analysis we used information
283 about the binding of J-domains to Hsp70. It is thus tempting to hypothesize that the various, possibly
284 orthogonal, classifications are successful because, among other factors, J-domains co-evolved together
285 with Hsp70 in a concerted phylogenetic and functional divergence.

286 ***Classification of J-domain Proteins according to their J-domains.***

287 We next employed the ANNs as class predictors for the human JDPs. We first excluded all the human
288 JDPs, alongside the ones from other model organisms (see Methods) from the training set to prevent
289 any bias in their favor during the learning process. In Table 1 we report the probabilities that each
290 human J-domain is attributed to a given class or architectural group.

291 J-domains from human class A (DnaJA1-4) and "canonical" class B (DnaJB1, DnaJB4, DnaJB5) were
292 correctly classified according to the A/B, A/B/C and Arch12 tasks (Table 1 and Supplementary Table
293 2 for the A/B task), confirming that the ANNs were able to reliably learn their features during the
294 training phase. A more specific discussion is warranted for DnaJB11 and DnaJB13, whose J-domains
295 was consistently classified as class C according to the A/B/C classification task. They were however
296 both correctly classified by the Arch12 task as either "Other" (DnaJB11) or recognized as canonical
297 class B in the Arch12 task (Table 1). These results suggest that DnaJB11 and DnaJB13, despite a likely
298 class B origin, have further specialized enough that their sequences do not fully conform to the majority
299 of class B JDPs.

300 Class B JDPs with a non-canonical architecture (DnaJB2, DnaJB3, DnaJB6, DnaJB7, DnaJB8, DnaJB9,
301 DnaJB12, and DnaJB14) have historically been attributed to class B due to the presence of a G/F-rich
302 region N-terminal to the J-domain. We used both the A/B/C and the Arch12 classification tasks to try
303 and classify these proteins. According to the A/B/C task, none of them should be considered as genuine
304 class B. Cross-checking this prediction with the Arch12 task, we found that DnaJB12 and DnaJB14
305 belonged to class C and in particular to the same architectural class as DnaJC18, with whom they share
306 the presence of the DUF1977 domain. The Arch12 task assigned the other non-canonical class B JDPs
307 variably to Class A or to Class B. Taken together, these results suggest a complex evolutionary origin
308 of the non-canonical Class B JDPs, with some of them possibly descending from Class A JDPs (Table
309 1).

310 We next classified class C JDPs. They were all clearly predicted as class C by the A/B/C classification
311 task. The Arch12 task correctly classified all the human class C JDPs. The notable cases are DnaJC20
312 (HscB) and DnaJC23 (Sec63) (highlighted in orange in Table 1): since their architecture is not exactly
313 the canonical one based on Pfam-annotated domains, they were assigned to the *Others* group.
314 Nonetheless, the Arch12 task correctly assigned them to the HscB and Sec63 subclasses, hinting at the
315 ability of this automated approaches to correct annotations in the presence of a clear and large training
316 set.

317 Discussion

318 Genetic, biochemical, and structural data accumulated over the years provide a relatively
319 comprehensive map of the J-domain-Hsp70 interface (*23*, *24*, *45–50*), which serves a single task, *i.e.* to

320 establish the vital physical interaction that promotes ATP hydrolysis in Hsp70 (*17*, *18*, *23–26*) , thus
321 allowing the chaperone to capture client substrates (*19*). The proper docking of the J-domain onto ATP-
322 bound Hsp70 is ensured by a core set of intermolecular contacts occurring along the helix II--loop (HDP
323 motif)--helix III region of the J-domain, which is highly conserved from bacteria to human (*23*, *24*).
324 However, very little knowledge exists of the structural elements in J-domains that critically fine-tune
325 this interaction (*38*) despite such regulation being hypothesized over the years. Herein, we reported the
326 presence of phylogenetic and functional signatures buried within the amino acid sequences of J-domains
327 that seem to modulate the partnering of JDPs with Hsp70s to form chaperone machineries with varying
328 activities.

329 During evolution, the JDP family expanded both in number and structural diversity, and in eukaryotic
330 organisms there is a more than two-fold increase in the number of JDPs relative to Hsp70s, indicating
331 that the limited repertoire of Hsp70 isoforms is utilized by multiple distinct JDPs to generate different
332 chaperoning activities (*e.g.* unfolding and disaggregation of misfolded protein, disassembly of
333 functional oligomers, protein translocation into organelles and protein degradation). Our ANN-based
334 observations showed that functional subclasses of JDPs across different species indeed carry unique
335 sequence-based signatures in the J-domains that allowed their very precise discrimination according to
336 functional and phylogenetic criteria. In some cases, such as HscB, this can be understood because this
337 JDP is known to interact with its own Hsp70 chaperone (HscA in bacteria, the specialized Ssq in yeast
338 mitochondria, and mitochondrial Hsp70 in other eukaryotes), in a partnership dating back very early in
339 the history of life (*51*). Our analysis also allowed proposing JDP classifications different from the
340 currently accepted ones. In particular, the J-domains of non-canonical Class B JDPs (DnaJB2, DnaJB3,
341 DnaJB6-9, DnaJB12 and DnaJB14) turned out being very difficult to classify as genuine class B. The
342 presence of a G/F-rich region N-terminal to the J-domain surely suggest a Hsp40 ancestry, but possibly
343 from class A or from class B JDPs for different non-canonical class B JDPs, followed by some extensive
344 sequence divergence.

345 Class C JDPs have witnessed a great expansion in eukaryotes, and we have shown here that their
346 automated classification proposed here is already possible, and will improve in the coming years.
347 Indeed, the degree of uncertainty encountered in this work is mostly due to two factors. On the one
348 hand, several architectural C subclasses are too poorly represented in the dataset that it is difficult to
349 train a network to tell them apart from more abundant subclasses. This problem, though, is likely to be
350 solved through the exponential growth of the number of reported JDP sequences. On the other hand, a
351 large swath of class C JDPs comprise domains that have not yet been annotated in Pfam, or just some
352 disordered regions of various lengths. Yet, the advent of novel deep learning approaches on scales even
353 larger than the present ones are recognizing domains that were previously uncharacterized (*52*), paving
354 the way for a better assignment of class C JDPs to precise, well-populated architectural classes.

What is the basis for this functional and phylogenetic sorting? Remarkably, the most discriminative sequence positions structurally coincided either with the J-domain-Hsp70 binding interface, strongly hinting that some amino acids have a regulatory role to optimize the selection of unique functionally specific Hsp70-JDP pairs, or with some of the C-terminal residues (helix IV) that bridge the J-domain with the other C-terminal domains.

Our *in-silico* observations are strongly supported by indirect experimental evidence (measurements of JDP stimulated ATPase activity in Hsp70s) showing that J-domains are not fully interchangeable with respect to their collaboration with different Hsp70s within a given species. Yeast cytosol contains seven Hsp70 paralogs (Ssa1-4, Ssb1-2 & Ssz1). Both Ssa's and Ssb's bind to nascent polypeptide chains to promote *de novo* protein folding. However, only the ATPase activity of Ssa's, but not of Ssb's, is stimulated *in vitro* by the two major cytosolic yeast JDPs Ydj1 (class A) and Sis1 (class B) (*53*). Similarly, the ATPase activity of ER-localized yeast Hsp70 Bip/Kar2 was shown to be stimulated by the J-domain of Sec63 (ER; class C) 16-fold higher than by the J-domain of ERdj3 (ER; class C) (*54*). Furthermore, as expected, speciation also left a clearly discernible signature on the J-domain sequence, which points towards some incompatibilities in JDP-Hsp70 paring across species. If this is true, one might expect Hsp70-JDP pairs constituted from evolutionary distinct organisms to be unable to show similar activities. In agreement, the class A JDP bacterial DnaJ stimulates the ATPase activity of bacterial DnaK and human HSPA8 (Hsc70) to significantly varying degrees (*55*). Similarly, mouse DnaJC7 (DjC7; class C JDP) stimulates the ATPase activity of bacterial DnaK, and murine HSPA8 and HSPA5 (BiP) to varying levels (*56*). The greater diversity of J-domains with respect to their Hsp70 partners had, to some extent, to be expected. Indeed, mutations of their interaction surface on the Hsp70 site would require independent compensatory mutations of a plethora of J-domains. This is statistically very unlikely, thus imposing a greater evolutionary pressure on Hsp70s. By converse, mutations on the J-domain side would be useful to fine tune their interactions and the associated functional network.

From a more general perspective, the recent advent of ANN-based approaches in structural biology, epitomized by the spectacular and widely-advertised success of AlphaFold for protein folding (*57*, *58*), highlights that the exponentially growing number of available sequences holds a yet unexplored wealth of information that must be exploited. At the same time, this also creates a tension between the indiscriminate use of sequences from all sources and the more cautious desire to use only (possibly manually) carefully curated sequences. The new era of exponentially increasing genomic datasets and their Machine Learning exploration calls for an intermediate approach, where stringent, conservative automated criteria go hand-in-hand with manual curation at higher scales (*e.g.* domain composition). In this respect the present work can be considered as an example where this strategy allows for an unbiased, semi-unsupervised coupled functional/phylogenetic-structural investigation: it is supervised for the classification task, but unsupervised for the emergence of the relevant structural features.

391 # Methods

392 ## *Construction of a curated J-protein dataset*

393 We constructed a JDP sequence database by querying the Uniprot database (union of Swissprot and
394 TrEMBL, release 2021_04) (*36*) to extract all sequences annotated to have a J-domain according to the
395 PFAMfam domain annotation (PF00226) (*37*). For each JDP sequence, we recovered its full-length
396 sequence as well as the full taxonomic lineage from Uniprot. We then extracted the J-domain sequence
397 of each JDP by aligning the full-length sequences to the Hidden Markov Model (HMM) of the Pfam
398 J-domain (PF0226) family using the *hmmalign* utility of the *hmmer* package (*59*). The full domain-
399 architecture was obtained by querying the PFAM database for each JDP entry in our database and the
400 domain order validated using the domain boundaries reported in PFAM.
401 The domain architectures were used to classify all sequences as either belonging to canonical class A
402 (PFAM architecture "*DnaJ;DnaJ_C;DnaJ_CXXCXGXG"*), canonical class B (PFAM architecture
403 "*DnaJ;DnaJ_C;"*) and class C (all other sequences).
404 We further identified the G/F-rich region in all JDPs as being the stretch of 40 amino-acids directly
405 following the JD sequence, if at least 25% of its amino-acids are either Glycines or Phenylalanines.
406 For ease of analysis, we also grouped all domains composing TPR repeat proteins (annotated as
407 TPR_XX) into a virtual TPR domain. Thus, in all proteins containing one or several TPR repeats, we
408 replaced all the TPR domains by a single effective TPR domain (as shown in Fig.2).
409 Subcellular localizations were predicted for eukaryotic sequences using the TargetP software (*60*),
410 version 2.0.
411 This resulted in an initial dataset containing 232'516 annotated JDP sequences.
412 We then curated the dataset by first removing all sequences for which the Uniprot and PFAM
413 annotations were not consistent and also removed all sequences which did not have a full characteristic
414 HPD motif. We further filtered out sequences containing annotated ANTI-TRAP domains (according
415 to Pfam), because in all but one case it simply recognized the canonical ZFLR domain. Furthermore,
416 we also removed sequences containing multiple J-domains. Finally, all sequences associated to
417 unclassified organisms were also removed from our final set.
418 This resulted in a curated JDP dataset comprising 223'124 sequences.
419

420 ## *Annotation of a novel pseudo Zinc-Finger Like Region (pseudoZFLR) domain in JDPs*

421 We created a new domain definition to annotate the pseudo zinc-finger like domains similar to the ones
422 observed in DnaJB11 and ErdJ3. These domains are characterized by one CNC and one CxxC motifs,
423 in contrast to the 4 pairs of CxxC motifs of the canonical Zinc-Finger motif as annotated in the PFAM
424 DnaJ_CXXCXGXG domain. We thus built a multiple sequence alignment containing 35 manually
425 selected and curated JDPs containing a pseudoZFLR domain, from organisms including vertebrates,
426 invertebrates and plants. We then aligned this seed using MAFFT (v7.487) (*61*) and manually identified

427 a region comprised of 55 positions defining the pseudoZFLR domain. We then constructed a hidden-
428 markov model profile of the novel domain using the *hmmbuild* utility from the hmmer suite (*59*) and
429 used to profile to query all the 223'124 JDP sequences in our curated dataset to identify JDPs containing
430 a pseudoZFLR domain. Because of the similarity between this domain and the canonical ZFLR domain,
431 we used a more stringent E-value $=10^{-4}$ cutoff to identify pseudo ZF domains (see Supplementary Figure
432 4). This resulted in a total of 1421 JDPs identified as containing a pseudoZFLR domain, of which 1336
433 possess a DnaJ-DnaJ_C-pseudoZF architecture.
434

435 ***Dimensionality reduction of JDP amino-acid sequences by UMAP***

436 To apply the UMAP algorithm (*43*) to the amino-acid sequences of our J-domain repertoire for
437 visualization, we first used a one-hot-encoding technique to numerically embed the JD sequences in a
438 high-dimensional space, consisting of 63x21=1323 dimensions (20 amino-acid + 1 gap for each of the
439 63 positions in the JD aligned sequences). We then applied the UMAP dimensionality reduction
440 technique with standard parameters as provided by the original publication.
441

442 ***Preparation of JDP datasets for training Neural Networks***

443 To avoid training the Neural Networks on sequences for which we want to predict and interpret the
444 class labels, sequences from model organisms (H. Sapiens, M., S. Cerevisiae, R. Norvegicus,
445 Melanogaster, D. Rerio, C. Elegans, A. Thaliana) were removed from the dataset prior to training. To
446 interpret the predictions of the trained models, we also removed non-canonical class B proteins from
447 our training set. These were identified as JDPs containing a DnaJ domain, an annotated G/F region (see
448 above) but no DnaJ_C domain. This procedure resulted in a reduced set of 205'252 sequences used for
449 building the classification models.
450 To reduce the effects of sampling bias during the training phase, we used as simple sequence
451 reweighting scheme previously used in several sequence analysis pipelines (*62*, *63*). We assigned to
452 each sequence a weight of $1/n_{80}$, where n_{80} denotes the number of sequences with sequence identity
453 greater than 80% from it. These weights were calculated for each classification task separately, as some
454 tasks are only trained on a subset of sequences, which can potentially slightly modify the weights.
455 We randomly split the processed dataset into training, validation and test set in such a way that
456 sequences from the same organism always belong to the same set and that the sum of the weights of
457 each dataset is respectively 0.7, 0.15, 0.15 times that of the total.
458

459 ***Feed-Forward Artificial Neural Network classification***

460 To feed the sequences to a neural network we encoded them into binary vectors. Every amino acid is
461 mapped to an integer *A(a)* in increasing alphabetical order, with the gap symbol mapped to zero. We
462 therefore can encode every amino acid with a 21-dimensional vector in which all elements are zeros

and element *A(a)* is one (the so called *one-hot encoding*). An amino acid sequence is then obtained by concatenating all 21-dimensional vectors together. To pass the sequences to a neural network, we concatenated the one-hot encoding of each position in the alignment, resulting in a 1323-dimensional binary vector (63x21 entries).

A Dropout layer has been used between the input and first hidden layer to randomly mask a fraction of inputs and has been found to improve the generalization accuracy of the network.

For each task described in the main text, we trained a shallow feed forward neural network to predict the corresponding label of each sequence. We trained the network by minimizing the categorical cross entropy between the output layer and the categorical encoding of the label. All weights have been regularized using an L2 regularizer. In all cases we used the Adam protocol during the minimization procedure and a learning decay equal to 0.01. A batch size of 1024 was used and SoftMax activation were used for all hidden layers. The remaining hyperparameters (number of layers, units per layer, dropout rate, weight regularizations and initial learning rate) were optimized by performing a grid search over parameter space, testing 216 combinations of hyperparameters per task. The best performing model was selected based on the maximum accuracy on the validation set. The model performances are reported on the hold out test set.

After the training is completed, we analyzed the trained networks using the technique described in (*44*) in order to obtain 21x63 coefficients describing how the networks use every bit in the input in order to classify the sequence. The relevance scores for each residue is finally computed as the Frobenius norm (sum of the squares) of its corresponding elements in the coupling matrix (see (*44*) for details on the relevance score calculations).

Author Contributions and Notes

SZ and DM designed the experiments, prepared the dataset, performed the analysis, and wrote the manuscript. NBN designed the experiments and wrote the manuscript. PDLR designed the experiments, performed the analysis, wrote the manuscript, and supervised the project. AB designed the experiments. All authors discussed the results and revised the manuscript.

Acknowledgments

DM thanks the Swiss National Science Foundation (SNFS) for financial support under grant number P2ELP3_181910. NBN thanks National Health and Medical Research Council of Australia Investigator Grant APP1197021 and Recruitment Grant from Monash University Faculty of Medicine Nursing and Health Sciences with funding from the State Government of Victoria and the Australian Government.

References

1. A. K. Mandal, N. B. Nillegoda, J. A. Chen, A. J. Caplan, Ydj1 Protects Nascent Protein Kinases from Degradation and Controls the Rate of Their Maturation. *Mol. Cell. Biol.* **28**, 4434–4444 (2008).
2. J. Behnke, M. J. Mann, F. L. Scruggs, M. J. Feige, L. M. Hendershot, Members of the Hsp70 Family Recognize Distinct Types of Sequences to Execute ER Quality Control. *Mol. Cell*. **63**, 739–752 (2016).
3. N. B. Nillegoda, J. Kirstein, A. Szlachcic, M. Berynskyy, A. Stank, F. Stengel, K. Arnsburg, X. Gao, A. Scior, R. Aebersold, D. L. Guilbride, R. C. Wade, R. I. Morimoto, M. P. Mayer, B. Bukau, Crucial HSP70 co-chaperone complex unlocks metazoan protein disaggregation. *Nature*. **524**, 247–251 (2015).
4. A. S. Wentink, N. B. Nillegoda, J. Feufel, G. Ubartaitė, C. P. Schneider, P. De Los Rios, J. Hennig, A. Barducci, B. Bukau, Molecular dissection of amyloid disaggregation by human HSP70. *Nature*. **587**, 483–488 (2020).
5. K. L. Howe, P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M. Ridwan Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, M. Charkhchi, C. Cummins, L. da Rin Fioretto, C. Davidson, K. Dodiya, B. El Houdaigui, R. Fatima, A. Gall, C. G. Giron, T. Grego, C. Guijarro-Clarke, L. Haggerty, A. Hemrom, T. Hourlier, O. G. Izuogu, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. G. Martinez, J. C. Marugán, T. Maurel, A. C. McMahon, S. Mohanan, B. Moore, M. Muffato, D. N. Oheh, D. Paraschas, A. Parker, A. Parton, I. Prosovetskaia, M. P. Sakthivel, A. I. Abdul Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, E. Steed, M. Szpak, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, B. Walts, A. Winterbottom, M. Chakiachvili, A. Chaubal, N. de Silva, B. Flint, A. Frankish, S. E. Hunt, G. R. IIsley, N. Langridge, J. E. Loveland, F. J. Martin, J. M. Mudge, J. Morales, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, F. Cunningham, A. D. Yates, D. R. Zerbino, P. Flicek, Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
6. J. Hageman, M. a Rujano, M. a W. H. van Waarde, V. Kakkar, R. P. Dirks, N. Govorukhina, H. M. J. Oosterveld-Hut, N. H. Lubsen, H. H. Kampinga, A DNAJB chaperone subfamily with HDAC-dependent activities suppresses toxic protein aggregation. *Mol. Cell*. **37**, 355–69 (2010).
7. J. Kumar, N. L. Kline, D. C. Masison, Human DnaJB6 Antiamyloid Chaperone Protects Yeast from Polyglutamine Toxicity Separately from Spatial Segregation of Aggregates. *Mol. Cell. Biol.* **38**, 1–15 (2018).
8. D. Sinha, S. Srivastava, P. D'Silva, Functional diversity of human mitochondrial j-proteins is independent of their association with the inner membrane presequence translocase. *J. Biol. Chem.* **291**, 17345–17359 (2016).
9. S. Hassdenteufel, N. Johnson, A. W. Paton, J. C. Paton, S. High, R. Zimmermann, Chaperone-Mediated Sec61 Channel Gating during ER Import of Small Precursor Proteins Overcomes Sec61 Inhibitor-Reinforced Energy Barrier. *Cell Rep.* **23**, 1373–1386 (2018).
10. K. Li, Q. Jiang, X. Bai, Y. F. Yang, M. Y. Ruan, S. Q. Cai, Tetrameric Assembly of K+ Channels Requires ER-Located Chaperone Proteins. *Mol. Cell*. **65**, 52–65 (2017).
11. A. Fotin, Y. Cheng, N. Grigorieff, T. Walz, S. C. Harrison, T. Kirchhausen, Structure of an auxilin-bound clathrin coat and its implications for the mechanism of uncoating. *Nature*. **432**, 649–653 (2004).
12. P. Walsh, D. Bursać, Y. C. Law, D. Cyr, T. Lithgow, The J-protein family: modulating protein assembly, disassembly and translocation. *EMBO Rep.* **5**, 567–71 (2004).
13. F. Rodriguez, F. Arsène-Ploetze, W. Rist, S. Rüdiger, J. Schneider-Mergener, M. P. Mayer, B. Bukau, Molecular Basis for Regulation of the Heat Shock Transcription Factor σ32 by the DnaK and DnaJ Chaperones. *Mol. Cell*. **32**, 347–358 (2008).
14. B. Westhoff, J. P. Chapple, J. Van Der Spuy, J. Höhfeld, M. E. Cheetham, HSJ1 is a neuronal shuttling factor for the sorting of chaperone clients to the proteasome. *Curr. Biol.* **15**, 1058–1064 (2005).
15. D. E. Grove, C. Y. Fan, H. Y. Ren, D. M. Cyr, The endoplasmic reticulum-associated Hsp40 DNAJB12 and Hsc70 cooperate to facilitate RMA1 E3-dependent degradation of nascent CFTRΔF508. *Mol. Biol. Cell*. **22**, 301–314 (2011).
16. Y. H. Yamamoto, A. Kasai, H. Omori, T. Takino, M. Sugihara, T. Umemoto, M. Hamasaki, T. Hatta, T. Natsume, R. I. Morimoto, R. Arai, S. Waguri, M. Sato, K. Sato, S. Bar-Nun, T. Yoshimori, T. Noda, K. Nagata, ERdj8 governs the size of autophagosomes during the formation process. *J. Cell Biol.* **219** (2020), doi:10.1083/jcb.201903127.
17. W. C. Suh, C. Z. Lu, C. A. Gross, Structural features required for the interaction of the Hsp70 molecular chaperone DnaK with its cochaperone DnaJ. *J. Biol. Chem.* **274**, 30534–30539 (1999).
18. J. Jiang, E. G. Maes, A. B. Taylor, L. Wang, A. P. Hinck, E. M. Lafer, R. Sousa, Structural basis of J cochaperone binding and regulation of Hsp70. *Mol. Cell*. **28**, 422–33 (2007).
19. P. De Los Rios, A. Barducci, Hsp70 chaperones are non-equilibrium machines that achieve ultra-affinity by energy consumption. *Elife*. **3**, e02218 (2014).
20. A. Bracher, J. Verghese, The nucleotide exchange factors of Hsp70 molecular chaperones. *Front. Mol. Biosci.* **2**, 1–9 (2015).
21. H. Kampinga, E. Craig, The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nat. Rev. Mol. Cell Biol.* **11**, 579–92 (2010).
22. E. A. Craig, J. Marszalek, How Do J-Proteins Get Hsp70 to Do So Many Different Things? *Trends Biochem. Sci.* **42**, 355–368 (2017).

23. D. Malinverni, A. Jost Lopez, P. De Los Rios, G. Hummer, A. Barducci, Modeling Hsp70/Hsp40 interaction by multi-scale molecular simulations and co-evolutionary sequence analysis. *Elife*. **6**, e23471 (2017).

24. R. Kityk, J. Kopp, M. P. Mayer, Molecular Mechanism of J-Domain-Triggered ATP Hydrolysis by Hsp70 Chaperones. *Mol. Cell*. **69**, 227-237.e4 (2018).

25. B. Tomiczek, W. Delewski, L. Nierzwicki, M. Stolarska, I. Grochowina, B. Schilke, R. Dutkiewicz, M. A. Uzarska, S. J. Ciesielski, J. Czub, E. A. Craig, J. Marszalek, Two-step mechanism of j-domain action in driving hsp70 function. *PLoS Comput. Biol.* **16**, 1–29 (2020).

26. J. Tsai, M. G. Douglas, A conserved HPD Sequence of the J-domain Is Necessary fr YDJ1 Stimulation of Hsp70 ATPase ACtivity at a Site Distinct from Substrate Binding§. *J. Biol. Chem.* **271**, 9347–9354 (1996).

27. V. B. V. Rajan, P. D'Silva, Arabidopsis thaliana J-class heat shock proteins: Cellular stress sensors. *Funct. Integr. Genomics*. **9**, 433–446 (2009).

28. M. E. Cheetham, A. J. Caplan, Structure, function and evolution of DnaJ: Conservation and adaptation of chaperone function. *Cell Stress Chaperones*. **3** (1998), pp. 28–36.

29. B. D. Ryder, I. Matlahov, S. Bali, J. Vaquer-Alicea, P. C. A. van der Wel, L. A. Joachimiak, Regulatory inter-domain interactions influence Hsp70 recruitment to the DnaJB8 chaperone. *Nat. Commun.* **12**, 1–16 (2021).

30. O. Faust, M. Abayev-Avraham, A. S. Wentink, M. Maurer, N. B. Nillegoda, N. London, B. Bukau, R. Rosenzweig, HSP40 proteins use class-specific regulation to drive HSP70 functional diversity. *Nature*. **587**, 489–494 (2020).

31. H. Y. Yu, T. Ziegelhoffer, E. a. Craig, Functionality of Class A and Class B J-protein co-chaperones with Hsp70. *FEBS Lett.*, 1–6 (2015).

32. T. K. Karamanos, V. Tugarinov, G. M. Clore, Unraveling the structure and dynamics of the human DNAJB6b chaperone by NMR reveals insights into Hsp40-mediated proteostasis. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 21529–21538 (2019).

33. J. Gu, Z. Liu, S. Zhang, Y. Li, W. Xia, C. Wang, H. Xiang, Z. Liu, L. Tan, Y. Fang, C. Liu, D. Li, Hsp40 proteins phase separate to chaperone the assembly and maintenance of membraneless organelles. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 31123–31133 (2020).

34. T. Xie, T. Saleh, P. Rossi, C. G. Kalodimos, Conformational states dynamically populated by a kinase determine its function. *Science (80-.).* **370**, 1–16 (2020).

35. I. Baaklini, M. J. H. Wong, C. Hantouche, Y. Patel, A. Shrier, J. C. Young, The DNAJA2 substrate release mechanism is essential for chaperone-mediated folding. *J. Biol. Chem.* **287**, 41939–41954 (2012).

36. A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L. G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, J. Zhang, UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

37. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, A. Bateman, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

38. F. Hennessy, W. S. Nicoll, R. Zimmermann, M. E. Cheetham, G. L. Blatch, Not all J domains are created equal: implications for the specificity of Hsp40-Hsp70 interactions. *Protein Sci.* **14**, 1697–709 (2005).

39. X.-B. Qiu, Y.-M. Shao, S. Miao, L. Wang, The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. *Cell. Mol. Life Sci.* **63**, 2560–70 (2006).

40. C. J. Kluck, H. Patzelt, P. Genevaux, D. Brehmer, W. Rist, J. Schneider-Mergener, B. Bukau, M. P. Mayer, Structure-function analysis of HscC, the Escherichia coli member of a novel subfamily of specialized Hsp70 chaperones. *J. Biol. Chem.* **277**, 41060–9 (2002).

41. G. Schlenstedt, S. Harris, B. Risse, R. Lill, P. A. Silver, A yeast DnaJ homologue, Scj1p, can function in the endoplasmic reticulum with BiP/Kar2p via a conserved domain that specifies interactions with Hsp70s. *J. Cell Biol.* **129**, 979–988 (1995).

42. C. S. Sullivan, P. Cantalupo, J. M. Pipas, The Molecular Chaperone Activity of Simian Virus 40 Large T Antigen Is Required To Disrupt Rb-E2F Family Complexes by an ATP-Dependent Mechanism. *Mol. Cell. Biol.* **20**, 6233–6243 (2000).

43. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2018) (available at http://arxiv.org/abs/1802.03426).

44. S. Zamuner, P. D. L. Rios, Interpretable Neural Networks based classifiers for categorical inputs. *Arxiv* (2021) (available at http://arxiv.org/abs/2102.03202).

45. C. S. Gässler, A. Buchberger, T. Laufen, M. P. Mayer, H. Schrder, A. Valencia, B. Bukau, Mutations in the DnaK chaperone affecting interaction with the DnaJ cochaperone. *Biochemistry*. **95**, 15229–15234 (1998).

46. M. K. Greene, K. Maskos, S. J. Landry, Role of the J-domain in the cooperation of Hsp40 with Hsp70. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6108–13 (1998).

47. W. C. Suh, W. F. Burkholder, C. Z. Lu, X. Zhao, M. E. Gottesman, C. A. Gross, Interaction of the Hsp70 molecular chaperone, DnaK, with its cochaperone DnaJ. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 15223–8 (1998).

48. P. Wittung-Stafshede, J. Guidry, B. E. Horne, S. J. Landry, The J-domain of Hsp40 couples ATP hydrolysis to substrate capture in Hsp70. *Biochemistry*. **42**, 4937–4944 (2003).

49. S. J. Landry, Structure and energetics of an allele-specific genetic interaction between dnaJ and dnaK: correlation of nuclear magnetic resonance chemical shift perturbations in the J-domain of Hsp40/DnaJ with binding affinity for the ATPase domain of Hsp70/DnaK. *Biochemistry*. **42**, 4926–36 (2003).

50. A. Barducci, P. De Los Rios, Non-equilibrium conformational dynamics in the function of molecular chaperones. *Curr. Opin. Struct. Biol.* **30**, 161–169 (2015).

51. B. Schilke, B. Williams, H. Knieszner, S. Pukszta, P. D'Silva, E. A. Craig, J. Marszalek, Evolution of Mitochondrial Chaperones Utilized in Fe-S Cluster Biogenesis. *Curr. Biol.* **16**, 1660–1665 (2006).

52. M. L. Bileschi, D. Belanger, D. H. Bryant, T. Sanderson, B. Carter, D. Sculley, A. Bateman, M. A. Depristo, L. J. Colwell, Using deep learning to annotate the protein universe. *Nat. Biotechnol.* (2022), doi:10.1038/s41587-021-01179-w.

53. P. Lopez-Buesa, C. Pfund, E. A. Craig, The biochemical properties of the ATPase activity of a 70-kDa heat shock protein (Hsp70) are governed by the C-terminal domains. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 15253–15258 (1998).

54. S. S. Vembar, Y. Jin, J. L. Brodsky, L. M. Hendershot, The mammalian Hsp40 ERdj3 requires its Hsp70 interaction and substrate-binding properties to complement various yeast Hsp40-dependent functions. *J. Biol. Chem.* **284**, 32462–32471 (2009).

55. Y. Minami, J. Höhfeld, K. Ohtsuka, F. U. Hartl, Regulation of the heat-shock protein 70 reaction cycle by the mammalian DnaJ homolog, Hsp40. *J. Biol. Chem.* **271**, 19617–19624 (1996).

56. B. Kroczynska, S. Y. Blond, Cloning and characterization of a new soluble murine J-domain protein that stimulates BiP, Hsc70 and DnaK ATPase activity with different efficiencies. *Gene*. **273**, 267–274 (2001).

57. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* (2021), doi:10.1038/s41586-021-03819-2.

58. K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, Highly accurate protein structure prediction for the human proteome. *Nature* (2021), doi:10.1038/s41586-021-03828-1.

59. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**, 29–37 (2011).

60. J. J. A. Armenteros, M. Salvatore, O. Emanuelsson, O. Winther, G. Von Heijne, A. Elofsson, H. Nielsen, Detecting sequence signals in targeting peptides using deep learning. *Life Sci. Alliance*. **2**, 1–14 (2019).

61. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

62. F. Morcos, T. Hwa, J. N. Onuchic, M. Weigt, in *Protein Structure Prediction*, D. Kihara, Ed. (Springer New York, New York, NY, 2014; https://doi.org/10.1007/978-1-4939-0366-5_5), pp. 55–70.

63. M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, *Phys. Rev. E*, in press, doi:10.1103/PhysRevE.87.012707.

Tables

Table 1. Classification of the 50 Human J-Domain Proteins. Architectures grouped into the "Other" class are indicated by *(Other)* in the third column. Class C JDPs mis-predicted in the Arch12 task are highlighted in orange.

	H. Sapiens name	Uniprot	Domain Architecture *(Ground truth)*	A/B/C Probabilities	Arch12 Probabilities
	Class A				
1	DnaJA1	P31689	DnaJ - DnaJ_C - DnaJ_CXXCXGXG	**0.94** / 0.06 / 0.00	DnaJ;DnaJ_C;DnaJ_CXXCXGXG **(0.98)**
2	DnaJA2	O60884	DnaJ - DnaJ_C - DnaJ_CXXCXGXG	**1.00** / 0.00 / 0.00	DnaJ;DnaJ_C;DnaJ_CXXCXGXG **(1.0)**
3	DnaJA3	Q96EY1	DnaJ - DnaJ_C - DnaJ_CXXCXGXG	**0.99** / 0.00 / 0.00	DnaJ;DnaJ_C;DnaJ_CXXCXGXG **(0.97)**
4	DnaJA4	Q8WW22	DnaJ - DnaJ_C - DnaJ_CXXCXGXG	**0.89** / 0.10 / 0.01	DnaJ;DnaJ_C;DnaJ_CXXCXGXG **(0.96)**
	Canonical Class B				
5	DnaJB1	P25685	DnaJ - DnaJ_C	0.00 / **1.00** / 0.00	DnaJ;DnaJ_C **(0.98)**
6	DnaJB4	Q9UDY4	DnaJ - DnaJ_C	0.00 / **1.00** / 0.00	DnaJ;DnaJ_C **(1.0)**
7	DnaJB5	O75953	DnaJ - DnaJ_C	0.00 / **0.99** / 0.01	DnaJ;DnaJ_C **(0.98)**
8	DnaJB11	Q9UBS4	DnaJ - DnaJ_C - pseudoZFLR *(Other)*	0.00 / 0.04 / **0.96**	Other **(0.89)**
9	DnaJB13	P59910	DnaJ - DnaJ_C	0.00 / 0.01 / **0.99**	DnaJ;DnaJ_C **(0.77)**
	Non-canonical Class B				
10	DnaJB2	P25686	DnaJ	0.00 / 0.03 / **0.96**	DnaJ;DnaJ_C **(0.92)**
11	DnaJB3	Q8WWF6	DnaJ	**0.45** / 0.11 / 0.44	DnaJ;DnaJ_C **(0.62)**
12	DnaJB6	O75190	DnaJ	**0.58** / 0.05 / 0.37	DnaJ;DnaJ_C;DnaJ_CXXCXGXG **(0.63)**
13	DnaJB7	Q7Z6W7	DnaJ	**0.76** / 0.11 / 0.14	DnaJ;DnaJ_C;DnaJ_CXXCXGXG **(0.79)**
14	DnaJB8	Q8NHS0	DnaJ	0.03 / 0.01 / **0.97**	DnaJ;DnaJ_C;DnaJ_CXXCXGXG **(0.58)**
15	DnaJB9	Q9UBS3	DnaJ	0.12 / 0.00 / **0.88**	DnaJ;DnaJ_C;DnaJ_CXXCXGXG **(0.82)**
16	DnaJB12	Q9NXW2	DnaJ - DUF1977	0.03 / 0.00 / **0.96**	DnaJ;DUF1977 **(1.0)**
17	DnaJB14	Q8TBM8	DnaJ - DUF1977	0.00 / 0.00 / **1.00**	DnaJ;DUF1977 **(1.0)**
	Class C				
18	DnaJC1	Q96KC8	DnaJ - Myb_DNA-binding *(Other)*	0.00 / 0.00 / **1.00**	Other **(1.0)**
19	DnaJC2	Q99543	DnaJ – RAC_head - Myb_DNA-binding - Myb_DNA-binding *(Other)*	0.00 / 0.00 / **1.00**	Other **(1.0)**
20	DnaJC3	Q13217	TPR_16 – TPR1 – TPR_8 - DnaJ	0.00 / 0.00 / **1.00**	DnaJ;TPR **(1.0)**
21	DnaJC4	Q9NNZ3	DnaJ	0.00 / 0.00 / **1.00**	Other **(1.0)**
22	DnaJC5	Q9H3Z4	DnaJ	0.00 / 0.01 / **0.99**	Other **(0.97)**
23	DnaJC5B	Q9UF47	DnaJ	0.00 / 0.00 / **1.00**	Other **(0.63)**
24	DnaJC5G	B5MCA9	DnaJ	0.00 / 0.00 / **1.00**	Other **(0.98)**
25	DnaJC6 (Auxilin)	O75061	PTEN_C2 – DnaJ *(Other)*	0.00 / 0.00 / **1.00**	Other **(1.0)**
26	DnaJC7	Q99615	TPR_11 – TPR_16 – TPR_8 - DnaJ	0.02 / 0.03 / **0.95**	DnaJ;TPR **(1.0)**
27	DnaJC8	O75937	DnaJ	0.00 / 0.00 / **1.00**	Other **(1.0)**
28	DnaJC9	Q8WXX5	DnaJ	0.00 / 0.00 / **1.00**	Other **(1.0)**
29	DnaJC10	Q8IXB1	DnaJ – Thioredoxin - Thioredoxin - Thioredoxin - Thioredoxin *(Other)*	0.00 / 0.00 / **1.00**	Other **(0.99)**
30	DnaJC11	Q9NVH1	DnaJ - DnaJ-like_C11_C	0.00 / 0.00 / **1.00**	DnaJ;DnaJ-like_C11_C **(1.0)**
31	DnaJC12	Q9UKB3	DnaJ	0.00 / 0.00 / **1.00**	Other **(1.0)**
32	DnaJC13	O75165	RME-8_M - GYF_2 - DnaJ *(Other)*	0.00 / 0.00 / **1.00**	Other **(1.0)**
33	DnaJC14	Q6Y2X3	DnaJ - Jiv90 *(Other)*	0.00 / 0.00 / **1.00**	Other **(1.0)**
34	DnaJC15	Q9Y5T4	DnaJ	0.00 / 0.00 / **1.00**	Other **(1.0)**
35	DnaJC16	Q9Y2G8	DnaJ - Thioredoxin	0.05 / 0.03 / **0.92**	Other **(0.61)**
36	DnaJC17	Q9NVM6	DnaJ - RRM_1 *(Other)*	0.00 / 0.00 / **1.00**	Other **(1.0)**
37	DnaJC18	Q9H819	DnaJ - DUF1977	0.00 / 0.00 / **1.00**	DnaJ;DUF1977 **(1.0)**
38	DnaJC19 (TIM14)	Q96DA6	DnaJ	0.00 / 0.00 / **1.00**	Other **(1.0)**
39	DnaJC20 (HSCB)	Q8IWL3	HscB_4_cys - DnaJ - HSCB_C	0.00 / 0.00 / **1.00**	DnaJ;HSCB_C **(0.99)**
40	DnaJC21	Q5F1R6	DnaJ - zf-C2H2_jaz	0.00 / 0.00 / **1.00**	DnaJ;zf-C2H2_jaz **(1.0)**
41	DnaJC22	Q8N4W6	TM2 – DnaJ *(Other)*	0.00 / 0.00 / **1.00**	Other **(1.0)**
42	DnaJC23 (SEC63)	Q9UGP8	DnaJ - Sec63 - Sec63 *(Other)*	0.00 / 0.00 / **1.00**	DnaJ;Sec63 **(0.97)**
43	DnaJC24	Q6P3W2	DnaJ - zf-CSL	0.00 / 0.00 / **1.00**	DnaJ;zf-CSL **(1.0)**
44	DnaJC25	Q9H1X3	DnaJ	0.00 / 0.00 / **1.00**	Other **(0.95)**
45	DnaJC25-GNG10	A0A024R161	DnaJ - G-gamma *(Other)*	0.00 / 0.00 / **1.00**	Other **(0.98)**
46	DnaJC26 (GAK)	O14976	Pkinase - PTEN_C2 – DnaJ *(Other)*	0.00 / 0.00 / **1.00**	Other **(1.0)**
47	DnaJC27	Q9NZQ0	Ras – DnaJ *(Other)*	0.00 / 0.00 / **1.00**	Other **(1.0)**
48	DnaJC28	Q9NX36	DnaJ – DJC28_CD *(Other)*	0.00 / 0.00 / **1.00**	Other **(1.0)**
49	DnaJC29 (Sacsin)	Q9NZJ4	Ubiquitin – DnaJ – HEPN *(Other)*	0.00 / 0.00 / **1.00**	Other **(0.98)**
50	DnaJC30	Q96LL9	DnaJ	0.00 / 0.00 / **1.00**	DnaJ;DnaJ-X **(1.0)**

668 **Table 2. Optimal hyperparameters used for training the different classification networks.**

669

Task	Layers	Units	Kernel/Bias regularization	Learning Rate	Dropout
AB	4	128,128,128,128	0.0000001	0.3	0.2
ABC	2	128,128	0.000001	0.5	0.2
Localization	4	128,128,128,128	0.000001	0.3	0.2
Phylo	2	128,128	0.000001	0.5	0.2
Bac/Euk/Arch/Vir	4	128,128,128,128	0.00001	0.4	0.2
Top12Archs	4	128,128,128,128	0.00001	0.4	0.1

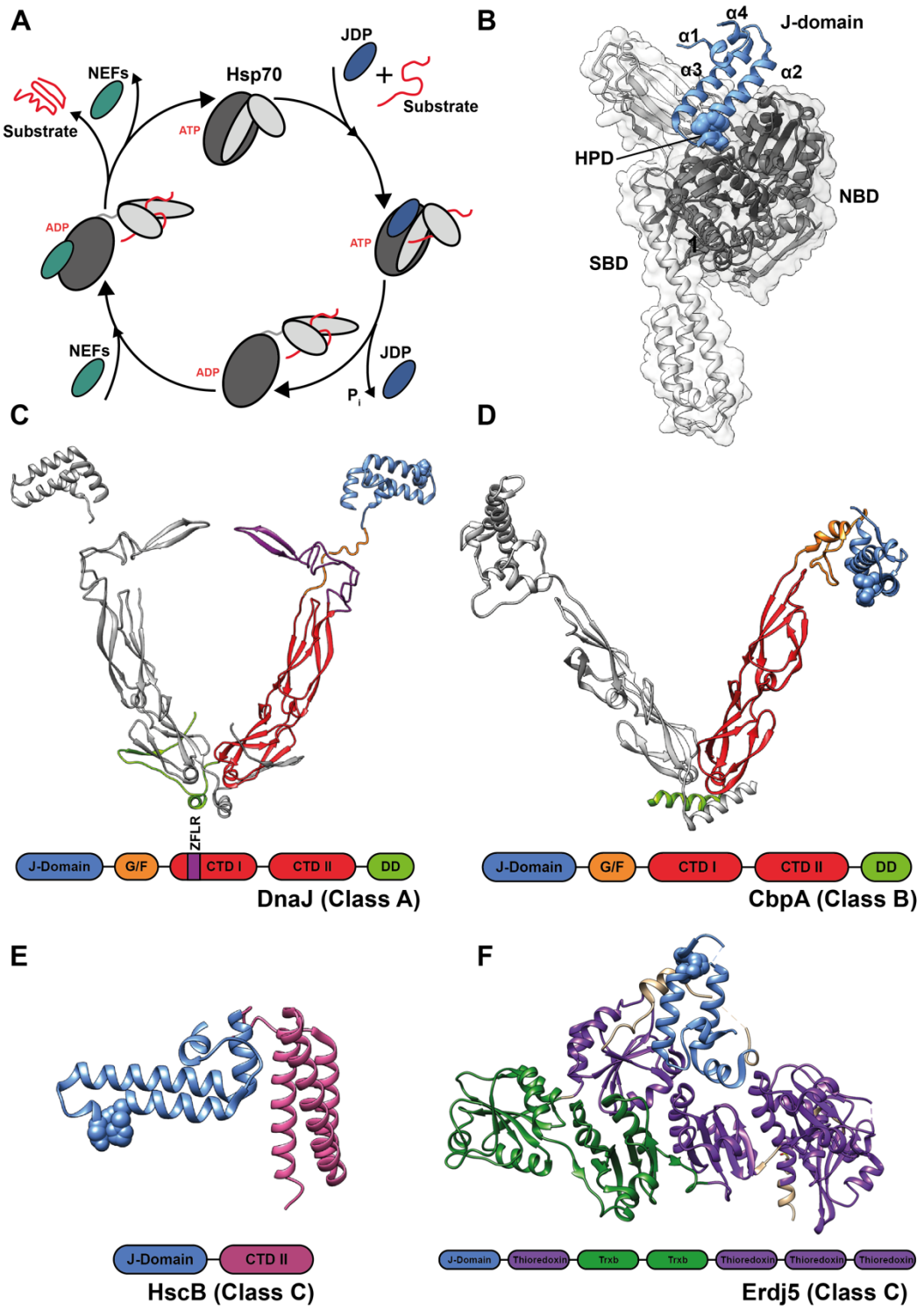
670

671

672

673 **Figures**

674



675

676 **Figure 1. Structural views of the J-domain and domain architectures of JDPs. A)** Schematic view
677 of the canonical Hsp70 cycle. **B)** Structural view of the Hsp70-J-domain complex (PDB ID: 5NRO).
678 NBD denotes the Hsp70 Nucleotide Binding Domain, SBD the Hsp70 Substrate Binding Domain. The
679 four helices (α1-α4) forming the J-domain and the characteristic HPD motif are highlighted. **C)**
680 Cytosolic class A JDP of *S. cerevisiae* (Ydj1) in its constitutive homo-dimeric form (the structure is a

combination of the separate J-domain, PDB ID 5VSO, CTDs PDB ID 1NLT, and dimerization domains, PDB ID 1XAO); in lack of a known experimentally determined structure, the G/F rich linker between the J-domain and the first C-terminal domain has been hand-drawn on one of the two protomers. The different domains have been colored only on one protomer. The general architecture of Class A JDPs is highlighted below the structure (G/F: glycine/phenylalanine rich linker; CTD: C-terminal, substrate binding domain; ZFLR: cysteine-rich, Zinc-Finger-Like Region; DD: dimerization domain). **C)** Class B JDP of *Thermus thermophilus* in its constitutive homo-dimeric form (PDB ID 1C3G). The different domains have been colored only on one protomer. The general architecture of Class B JDPs is highlighted below the structure. **D)** *Escherichia coli* HscB (Class C; PDB ID 1FPO) with its architecture (HSCB-C: C-terminal domain of HscB). **E)** *Mus musculus* Erdj5 (Class C; PDB ID 5AYK) and its architecture.

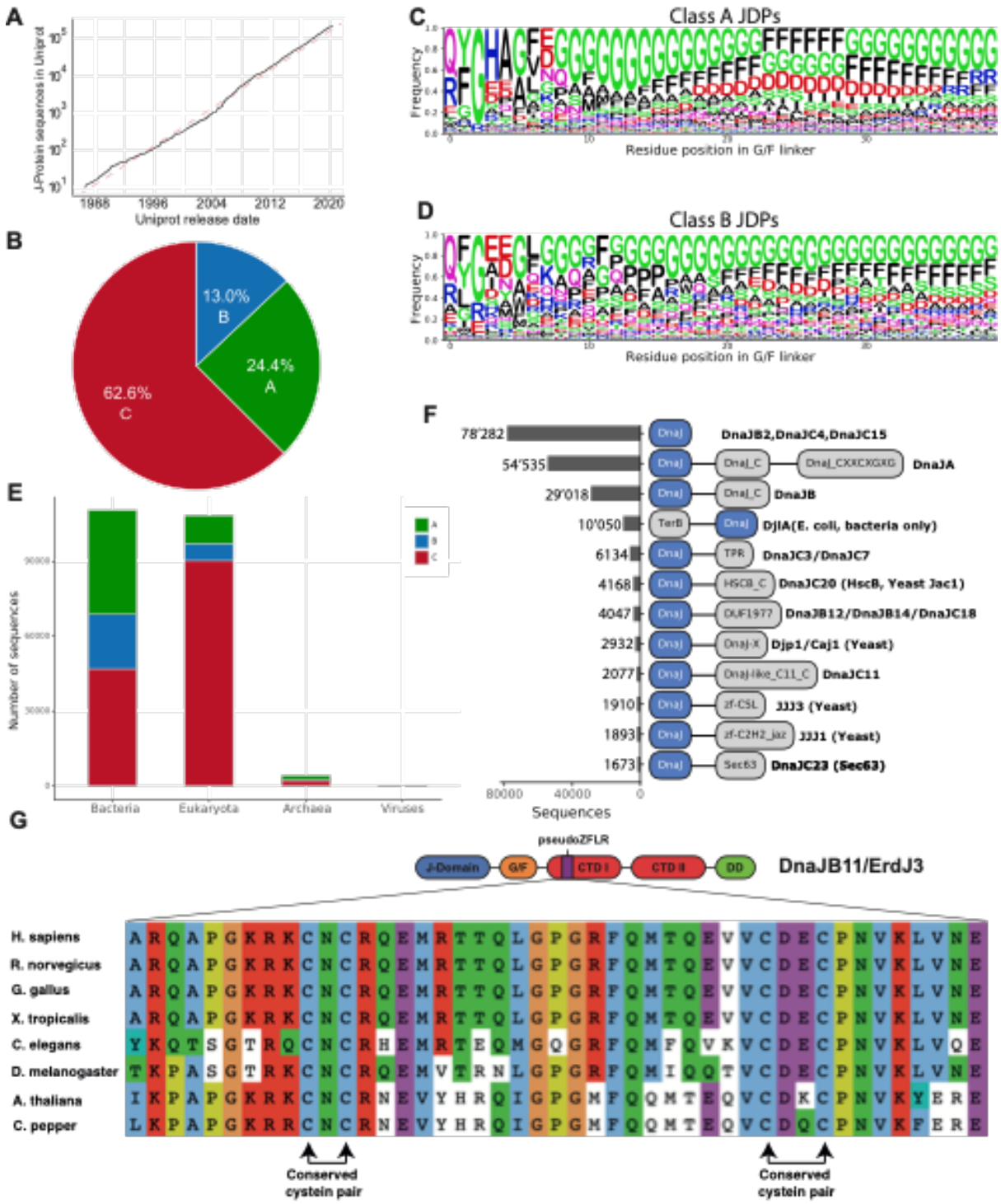
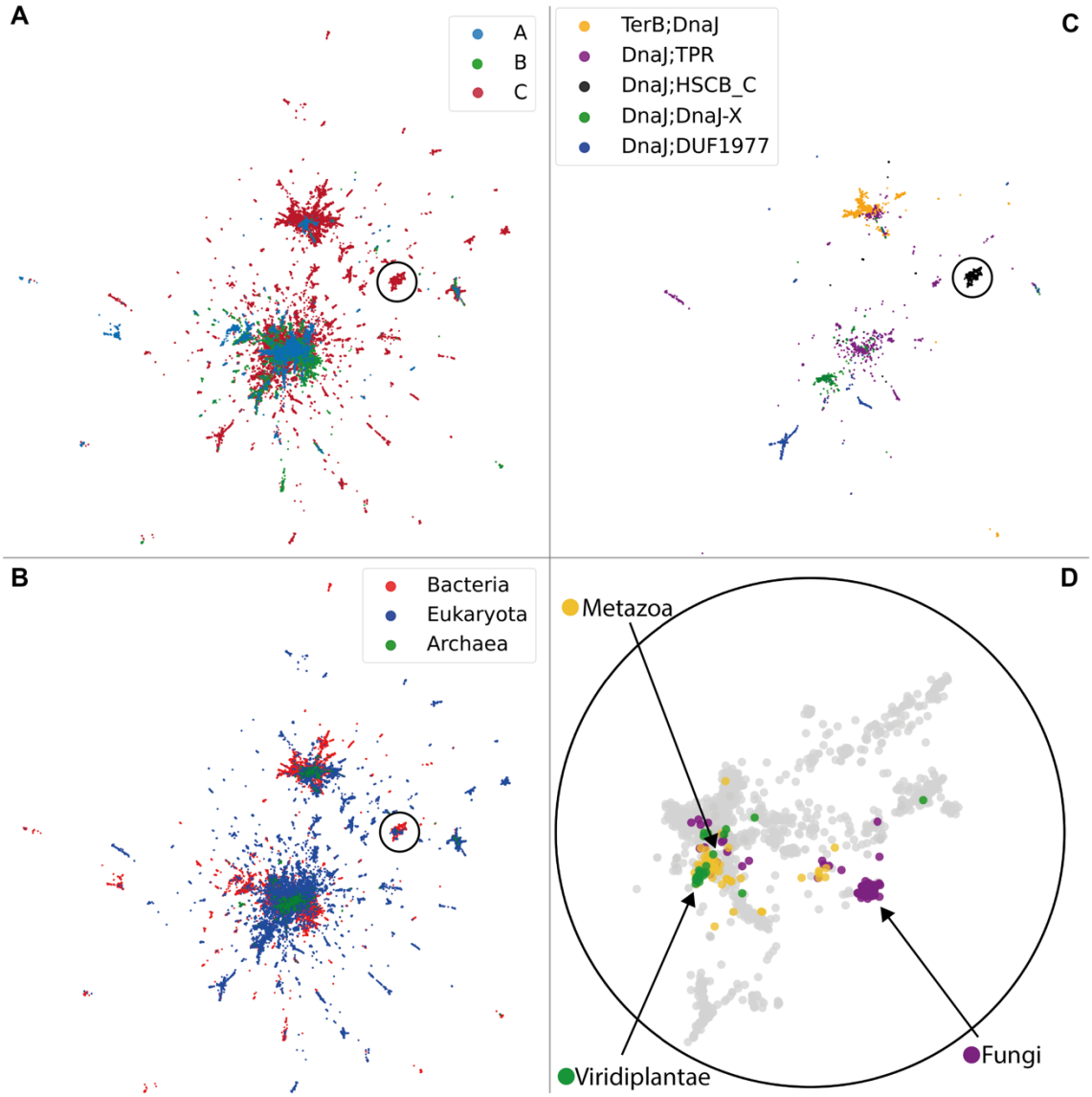


Figure 2. JDP Dataset characterization. A) Growth of the number of JDP sequences found in the Uniprot Database over the last 3 decades. The red dashed line shows an exponential fit. **B)** Class proportions in the dataset. **C)** Amino-acid frequencies of the G/F linker region of class A JDPs. **D)** Amino-acid frequencies of the G/F linker region of class B JDPs. **E)** Number of JDPs in the different kingdoms, highlighted by their class (A: green; B: blue; C: red). **F)** The 12 most abundant J-protein domain architectures. Numbers on the left denote the number of sequences with this architecture in our dataset, and, on the right, we report the corresponding human, yeast, plant or bacterial proteins. **G)**

Highlight of the evolutionary conservation of the cysteine-rich proto-ZFLR in ER-resident JDPs in 6 metazoa and 2 plants.



706
707

708 **Figure 3. Two-dimensional UMAP representations of J-domains.** JDP sequences are projected in
709 two dimensions by a non-linear transformation (UMAP) and colored according to different schemes.
710 **A)** J-domains are identified as being of class A (green), B (blue) or C (red) JDPs. **B)** J-domains are
711 identified as being of bacterial (red) or eukaryotic (blue) origin. **C)** The five most abundant C subclasses
712 are highlighted (see Fig.2F for representative members). The circles in each panel focus on the group
713 identified as HscB-like JDPs. **D)** Zoom on the encircled region in the 3 other panels, highlighting the
714 further differentiation of the eukaryotic ones into fungi, metazoa and plants (bacterial HscB sequences
715 are shown in grey background). For visual clarity, a random subset of 50% of all points are displayed.
716

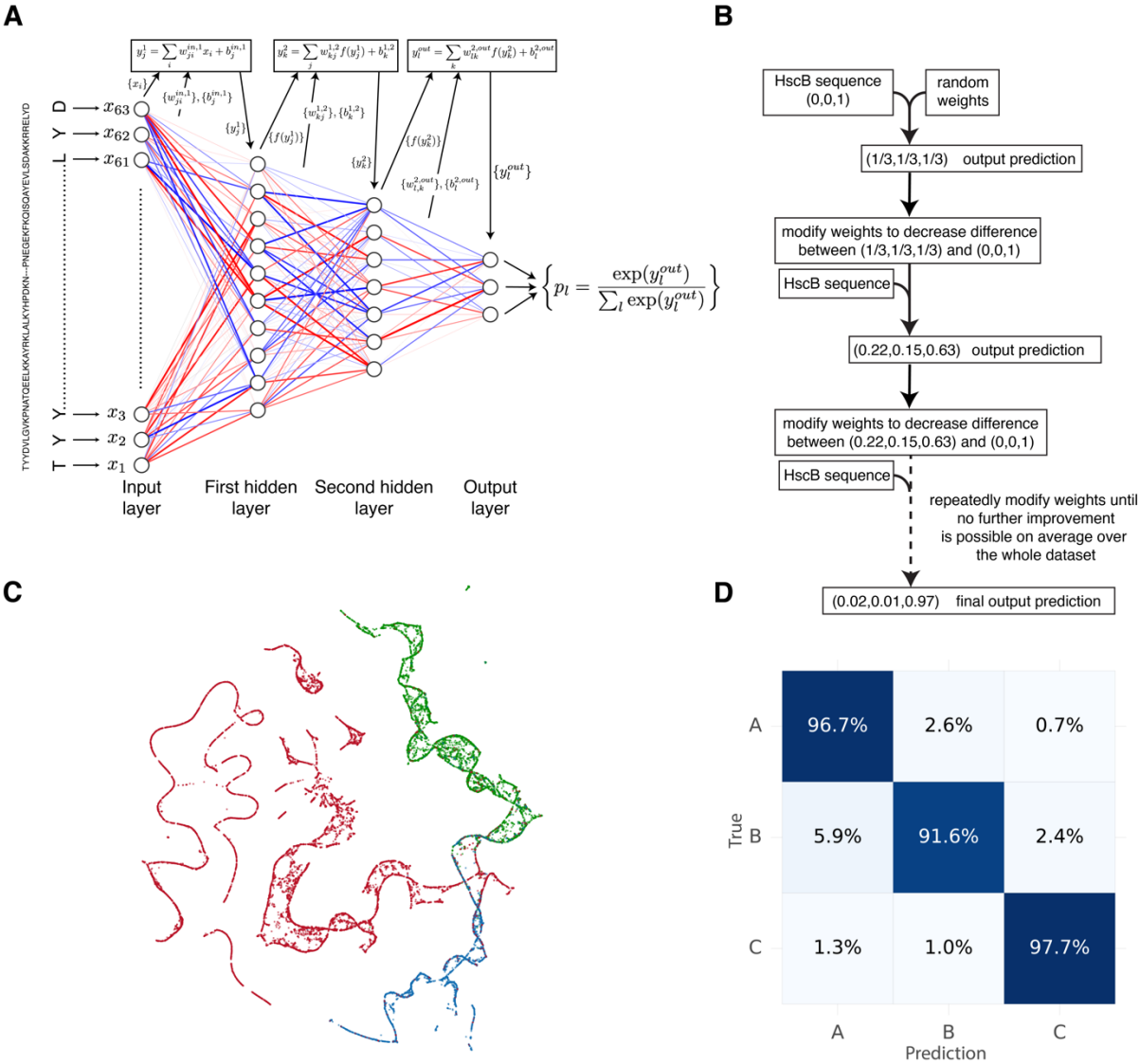


Figure 4. Structure, training and interpretation of Artificial Neural Networks. A) Structure of an ANN composed by input and output layers, separated by hidden layers (in this examples, two hidden layers); each node of a layer receives as input linear combinations of the values contained in the nodes of the previous layer, applies to them a non-linear function and the procedure is then repeated toward the next layer; the output layer transforms the received values into the probabilities for the input to belong to the various classes that must be distinguished. For visual clarity, the one-hot-encoding of the input sequences is not shown. The actual input size of all networks is 63*21=1323. **B)** Outline of the iterative procedure to change the ANN weights so that, on average over the training set, sequences are assigned as correctly as possible to their ground-truth class (defined according to domain composition). **C)** Two-dimensional embedding of the learnt probabilities for each J-domain sequence for the A *vs.* B *vs.* C classification task (colors according to the ground truth), providing a pictorial representation of

the quality of the distinction of the classes. **D)** Confusion matrix for the A *vs.* B *vs.* C classification task: each entry tells what fraction of a given class is classified to the possible classes, providing an intuitive appraisal of the ANN performance.

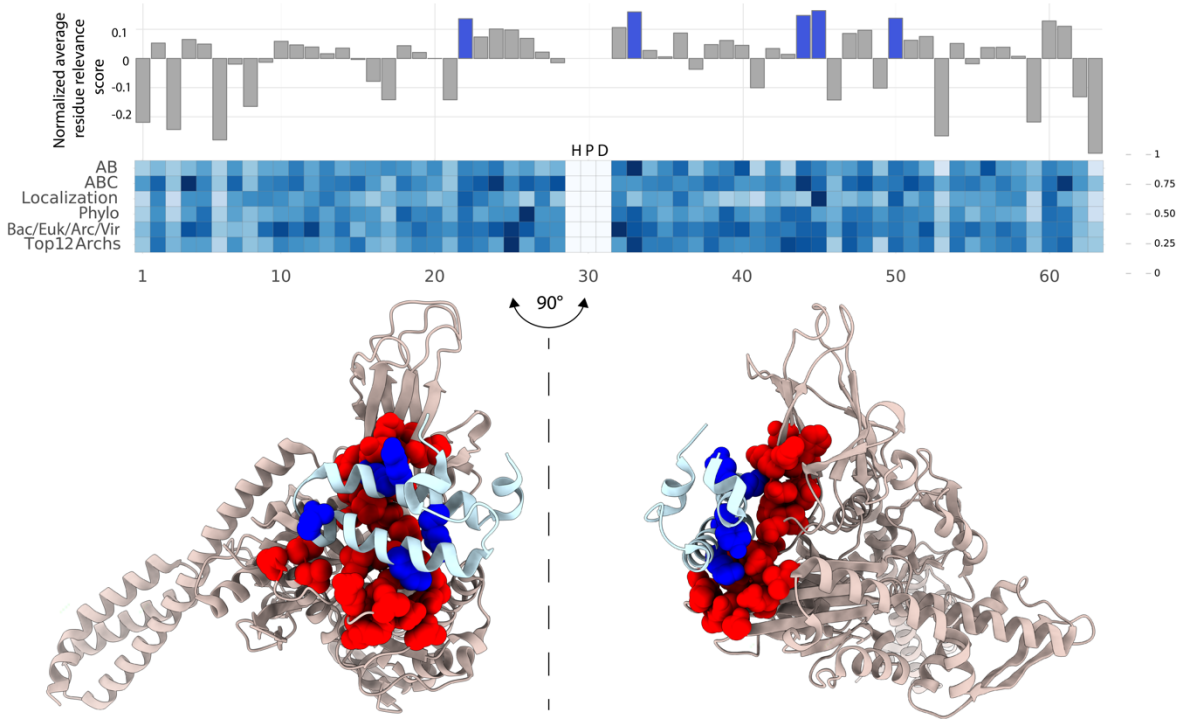


Figure 5. Identification of the most sequence and structure positions relevant for the classification. A) The relevance score for each position and for different classification tasks is represented in a heatmap. For each position, the relevance score averaged over the different tasks is represented, relative to the average over all positions, above the heatmap (HPD are not considered as they are completely irrelevant for all tasks). The top five positions are highlighted in blue. **B)** The five top scores identified in the top panel are highlighted in blue on the structure of the J-domain/DnaK complex (*24*), while the corresponding contacting residues on DnaK are represented in red, showing that the five most relevant positions on the J-domain participate in the interaction surface between the two proteins.