Sequence, Structure and Functional space of *Drosophila de novo* proteins

Lasse Middendorf [1], Bharat Ravi Iyengar [1, *], Lars A. Eicholt [1, *]

[1] Institute for Evolution and Biodiversity, University of Muenster, Muenster, Germany

[*] corresponding authors: Lars A. Eicholt & Bharat Ravi Iyengar, Huefferstrasse 1, 48149 Muenster, Germany, l.eicholt@uni-muenster.de & b.ravi@uni-muenster.de

Declaration of Interests

The authors declare no competing interests.

Abstract

During *de novo* emergence, new protein coding genes emerge from previously non-genic sequences. The *de novo* proteins they encode are dissimilar in composition and predicted biochemical properties to conserved proteins. However, many functional *de novo* proteins indeed exist. Both identification of functional *de novo* proteins and their structural characterisation are experimentally laborious. To identify functional and structured *de novo* proteins *in silico*, we applied recently developed machine learning based tools and refined the results for *de novo* proteins. We found that most *de novo* proteins are indeed different from conserved proteins both in their structure and sequence. However, some *de novo* proteins are predicted to adopt known protein folds, participate in cellular reactions, and to form biomolecular condensates. Apart from broadening our understanding of *de novo* protein evolution, our study also provides a large set of testable hypotheses for focused experimental studies on structure and function of *de novo* proteins in *Drosophila*.

keywords: *de novo* proteins, protein function, structural comparison, protein structure, structure predictions, sequence space

Introduction

Once considered impossible [Zuckerkandl, 1975, Jacob, 1977], many lines of evidence suggest that functional proteins can emerge from random sequences that have not been subjected to several generations of evolution [Keefe and Szostak, 2001, Hecht et al., 2004, Babina et al., 2023]. For example, high throughput selection experiments with a large number of random sequences have shown, that some random proteins can mitigate auxotrophy [the inability to metabolize nutrients; Knopp et al., 2021], provide resistance against toxins [Frumkin and Laub, 2023], and even catalyze biochemical reactions [Chao et al., 2013, Yamauchi et al., 2002]. In accordance with the fact that protein folding is often a critical requirement for protein function, many random proteins have been also shown to have secondary structures [Davidson and Sauer, 1994, Davidson et al., 1995, Tretyachenko et al., 2017, Surdo et al., 2004, Mansy et al., 2007]. *De novo* emergence is a phenomenon through which novel protein coding genes arise from non-genic regions of the genome [Tautz and Domazet-Lošo, 2011, Carvunis et al., 2012, Oss and Carvunis, 2019, Vakirlis et al., 2020a, Bornberg-Bauer et al., 2021, Schmitz and Bornberg-Bauer, 2017]. The *de novo* proteins thus emerged have been considered to be the natural equivalent of random sequences, because they emerge from supposedly "random" intergenic regions, and some of their predicted properties such as length, structural disorder and aggregation propensity, resemble that of random proteins, more than that of conserved proteins [Heames et al., 2023, Bornberg-Bauer et al., 2021, Ángyán et al., 2012, Bhave and Tautz, 2021, Castro and Tautz, 2021, Middendorf and Eicholt, 2024, Aubel et al., 2024]. For example, *de novo* proteins in *Drosophila*, are predicted to be more disordered than conserved proteins [Heames et al., 2020, Middendorf and Eicholt, 2024, Peng and Zhao, 2023], which can be partially explained due to higher GC content of the former [Landry et al., 2015, Zheng and Zhao, 2022]. While the structure of large sets of *de novo* proteins have been computationally analyzed [Schmitz et al., 2018, Heames et al., 2020, Peng and Zhao, 2023, Basile et al., 2017, Chen et al., 2023, Vakirlis et al., 2020b], the structures of only four *de novo* proteins have been experimentally approximated [Lange et al., 2021, Bungard et al., 2017, Baalsrud et al., 2018, Matsuo et al., 2021]. Determining the function of *de novo* genes and proteins is another challenging task. It involves identifying the cell types and stages in which *de novo* proteins may be involved and testing their phenotypic effects using genetic tools [Chen et al., 2010a, Gubala et al., 2017, Lange et al., 2021, Reinhardt et al., 2013]. Nonetheless, functional *de novo* proteins indeed exist and have been identified in organisms as diverse as insects, plants (*Arabidopsis thaliana*), fungi

(*Saccharomyces cerevisae*), arctic codfish , mice (*Mus musculus*) and humans (*Homo sapiens*) [McLysaght and Guerzoni, 2015, Li et al., 2009, Cai et al., 2008, Chen et al., 2010a, Gubala et al., 2017, Lange et al., 2021, Zhuang et al., 2019, Reinhardt et al., 2013, Heinen et al., 2009, Li et al., 2010a, Xie et al., 2019, Li et al., 2014, Vakirlis et al., 2022, Linnenbrink et al., 2024, Klasberg et al., 2018, Li et al., 2010b, Matsuo et al., 2021, Rivard et al., 2021, Begun et al., 2007]. Experimental structure determination is a laborious process that cannot be performed in a high throughput manner. This is especially difficult for *de novo* proteins because of high aggregation propensity and low solubility *in vitro* [Eicholt et al., 2022]. Despite the increasing numbers of solved structures, novel structures, whether they be folds or domains, were rarely ever found [Grant et al., 2004, Levitt, 2009, Tóth-Petróczy and Tawfik, 2014]. However, the recent advancements in high-throughput structure predictions through computational techniques, have led to discovery of novel folds [Durairaj et al., 2023]. Since *de novo* proteins are void of ancestry from conserved protein families, they could provide rare structural novelty [Bornberg-Bauer et al., 2021]. From another perspective, the occurrence of conserved or ancient structural folds in *de novo* proteins could suggest a high level of evolutionary accessibility in sequence space. This might explain the emergence of these folds during the early stages of protein evolution [Lupas et al., 2001, Kopec and Lupas, 2013, Alva et al., 2010, 2015, Romero Romero et al., 2016]. A protein's structure can provide some clues about its function [Orengo et al., 1999]. For example, one can reasonably guess the function of an uncharacterized protein by comparing its structure to that of a known functional protein [Nomburg et al., 2024]. Although, protein function is often attributed to its structure, and unfolded proteins were assumed to be toxic, many studies show that disordered proteins can be functional [Deiana et al., 2019, Jemth et al., 2018, Ali and Ivarsson, 2018]. For example, disordered proteins can help form intracellular condensates (or membrane less organelles) that have been shown to play a major role in the cellular physiology of diverse organisms [Lin et al., 2017, Hyman et al., 2014]. Because *de novo* proteins could be a source of novelty, with regards to both structure and function, we aimed to understand their structures and possible functions through computational analyses. To this end, we studied a previously characterized set of 2510 putative *de novo* proteins from the *Drosophila* clade [Heames et al., 2020, Middendorf and Eicholt, 2024]. We used a multi-faceted approach analyze these *de novo* proteins. First, we used Foldseek [van Kempen et al., 2023] to find experimentally known protein structures [Protein Data Bank, Berman et al., 2000] and predicted protein structures [AlphaFold database, Varadi et al., 2021] that are similar to the AlphaFold2 (AF2) [Jumper et al., 2021] predicted structures of our *de novo* proteins. Sec-

ond, we predicted the functions of our *de novo* proteins using DeepFRI [Gligorijević et al., 2021], a machine learning-based tool that predicts functional annotations (gene ontology terms) using protein structure and sequence features. Because many of our *de novo* proteins were predicted to be disordered *de novo* proteins, we hypothesized that they could form biomolecular condensates [Uversky, 2017]. To test this hypothesis, we predicted the condensate forming propensity of our *de novo* using PICNIC [Hadarovich et al., 2023], an algorithm that is based on predicted structure (AlphaFold2), predicted disorder (IUPred2A), as well as sequence complexity. Understanding the condensate forming behavior of *de novo* proteins would elucidate their potential involvement in the formation of membraneless organelles, offering an evolutionarily and biophysically feasible mechanism for their integration with the cellular physiology. Finally, we mapped the *de novo* proteins on the protein sequence space in relation to random and conserved proteins. To this end, we used protein language models that can predict several biophysical features from sequences, embedding their abstracted properties in the form of numerical values [Lin et al., 2023]. Our method allowed us to map different sequences with better resolution than by the analyses of individual properties separately [Weidmann et al., 2019, Agozzino and Dill, 2018, Heames et al., 2023, Aubel et al., 2024]. With these multi-faceted analyses we found that some *de novo* proteins can indeed adopt structures similar to known proteins and can have possible cellular activities including localization to specific organelles. We also found that some *de novo* proteins are likely to form biomolecular condensates. However, with our language model analysis we found that the majority of *de novo* proteins look distinct from conserved proteins of similar length, and resemble more the random proteins. Overall, our work enhances our understanding of how *de novo* proteins can not only develop features already known to the living systems, but can also be a source for evolutionary novelty.

Results

A few *de novo* proteins can indeed adopt known structures

To understand if *de novo* proteins can form known protein structures, we compared their predicted structure to that of conserved proteins. Recent studies have shown that structure predictions are not very reliable for *de novo* proteins [Middendorf and Eicholt, 2024, Aubel et al., 2023, Liu et al., 2023], and that many predicted structures are also thermodynamically unstable [Peng and Zhao, 2023]. Therefore, we refined the predicted structures of *Drosophila de novo* proteins from our previous study Middendorf and Eicholt [2024], using molecular dynamics simulations, performing 3 replicate simulations per protein for 100ns. We thus refined the predicted structures of 1,468 *de novo* proteins. Our MD simulations suggest that most *de novo* proteins exhibit structural flexibility, as indicated by the large root mean square deviation (RMSD) values (Figure Figure 1A and Figure S3). Next, we searched for conserved proteins that have predicted structures similar to those of *de novo* proteins, using Foldseek [van Kempen et al., 2023]. Specifically, with MD refined structures as queries, and the AFDB50 [Varadi et al., 2021] as the target, we observed that the majority of *de novo* proteins did not have a significant structural similarity to the conserved proteins in AFDB50 (TM score <0.5, Figure 1B). This was also the case for AF2 predicted structures of *de novo* and random proteins without MD simulations (Figure S1 and Figure S2). This observation, supports the *de novo* status of our proteins, aligning with the notion that structure is more conserved than sequence [Illergård et al., 2009]. To investigate whether these *de novo* proteins can adopt known structures, we performed structural mapping of *de novo* proteins with experimentally validated structures in the Protein Data Bank (PDB) [Berman et al., 2000], using Foldseek. We then extracted the ECOD domain annotations for matches found in the PDB [Cheng et al., 2014]. Out of the 1,468 *de novo* proteins analyzed, 42 showed structural alignment with proteins having an architecture annotation in ECOD (Figure 1C). Prior to MD simulation, 119 predicted structures were mappable to PDB structures (Figure S1). Figure 1D presents examples of these findings consisting of a structurally unalignable *de novo* protein, one similar to an SH3 fold, and another resembling an HTH fold. Both SH3 and HTH folds are considered highly conserved and ancient folds [Kishan and Agrawal, 2005, Alvarez-Carreño et al., 2021, Rosinski and Atchley, 1999, Grishin, 2000]. These three example proteins have emerged less than 5 million years ago (mya) [Heames et al., 2020]. Overall, our structure search analysis shows that, while most *de novo* proteins are likely to have

117 novel or uncommon structures, a minority of them can indeed adopt well known protein structures.

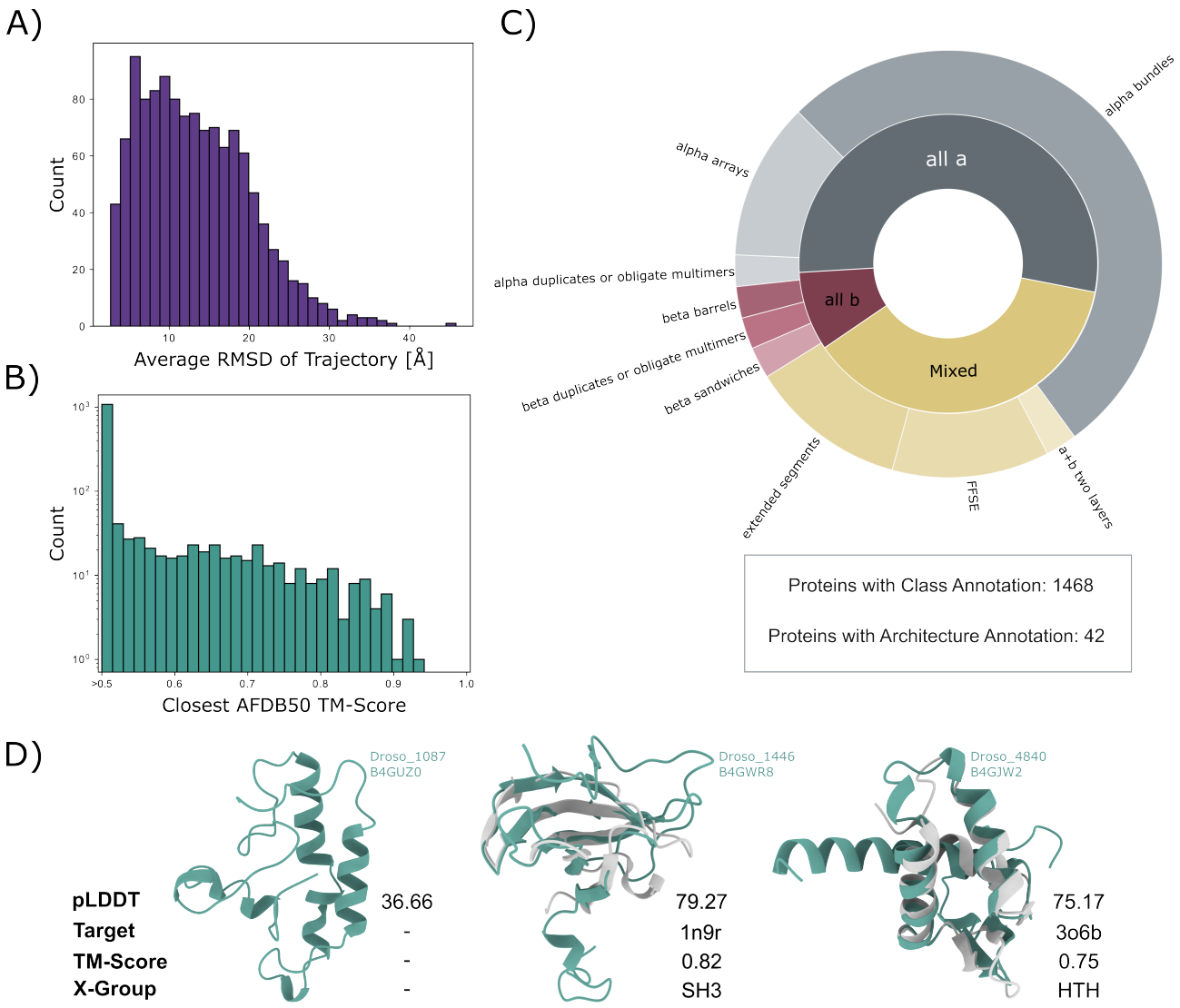

Figure 1: Structural diversity of *de novo* evolved proteins. (**A**) Distribution of the average root mean square deviation (RMSD, horizontal axis) per MD simulation trajectory. We display the average RMSD of three MD simulation replicates per *de novo* protein, only for proteins with i) less than 30% disorder predicted by flDPnn, and ii) less than 95% of their residues annotated as α-helices via DSSP (1468 of 2510 proteins). (**B**) Distribution of the TM-score (horizontal axis) for the mapping of *de novo* proteins (MD-refined structures) to the most similar protein structure in the AlphaFold database (AFDB50), excluding proteins from *Drosophila*. TM-scores below 0.5 indicate no similarity to any protein structure in the AlphaFold database. (**C**) Structural classification of *de novo* proteins. We assigned a structural class to each of the 1468 *de novo* proteins based on the DSSP annotations of their predicted structures (inner circle). To identify annotated protein domains in *de novo* proteins, we aligned their MD refined structures to structures in the PDB. We assigned each *de novo* protein with the ECOD domain of its highest scoring hit from the PDB, given the TM-score was greater than 0.5 and the alignment covered at least 80% of the PDB target. We assigned the 42 *de novo* proteins, that qualified the above criteria, with an ECOD domain from multiple domain architectures (outer circle). (**C**) Examples of *de novo* proteins without structural similarity to proteins in the AlphaFold database (Droso_1087), or with similar structure to an ECOD X-group (Droso_1446 & Droso_4840; aligned with their closest hit in the PDB).

Some *de novo* proteins may bind to nucleic acids, and are predicted to have enzymatic activities

Information on biological activities and functions, is available for only a handful of *de novo* proteins [Bornberg-Bauer et al., 2021, Weisman, 2022]. The existence and gain of biological activity would be critical factor determining the evolutionary fixation of *de novo* proteins. However, the lack of homology, makes functional annotation challenging. Therefore, we used DeepFRI to functionally annotate *de novo* proteins with Gene Ontology (GO) terms. Unlike homology based techniques, DeepFRI combines a protein language model, trained on the sequences of PFAM domains, and a graph convolutional network that represents amino acid interactions derived from protein structure [Gligorijević et al., 2021]. DeepFRI is also trained on the GO terms associated with different structures. We did not filter protein sequences according to any structural criteria, because DeepFRI can de-noise predicted protein structures [Gligorijević et al., 2021]. We summarized and clustered the predicted GO terms based on their semantic similarity, and projected them in a 2-dimensional semantic space using REVIGO [Supek et al., 2011] (Figure 2A & B). We identified these GO term clusters visually and manually annotated them based on the GO terms within the cluster. We performed this analysis for both *de novo* and random proteins. With our analysis, we found that a small fraction of *de novo* and random proteins could be confidently annotated with GO terms for all the three GO classes (Molecular Function, Biological Process, and Cellular Component; Figure 2C). The GO term class *Cellular Component* had the highest fraction of confident predictions with ≈31% and ≈17% for *de novo* and random proteins, respectively. However, we could not find any overarching GO terms within the cellular component category, for both *de novo* and random proteins. This suggests that both these kind of proteins can localize to many different cellular compartments. Specifically, we found that these proteins, can possibly localize to the following compartments: nucleus (GO:0005634), mitochondrion (GO:0005739), vesicles (GO:0031982), and membranes (GO:0016020).

Both *de novo* proteins and random sequences both show a broad variety of GO terms in other two GO classes with only a few prominent clusters within the semantic space (Figure 2A & B). Interestingly, *de novo* proteins and random sequences appear to have similar molecular functions and to be involved in similar categories of biological processes. Regarding their molecular function, they both showed multiple GO terms in relation to “hydrolase activity”, “transferase activity”, and “nucleic acid binding”. The biological processes in which *de novo* proteins and random sequences

are both predicted to be involved were “stimuli response”, “regulation” and “transport”. Next, we analyzed the impact of evolutionary age on functional annotation using GO terms. As young *de novo* proteins were more frequent than older proteins in the dataset, we normalized the number of proteins with predicted GO terms to the number of proteins in the respective age group. In all three categories of GO terms, the oldest *de novo* proteins (emerged >30 Mya) were more often predicted with a GO term, than younger proteins (Figure 2D). Only for the GO term category *Cellular Component*, old *de novo* proteins were annotated more frequently than expected by chance (Pearson's χ^2-Test; $P < 10^{-10}$).

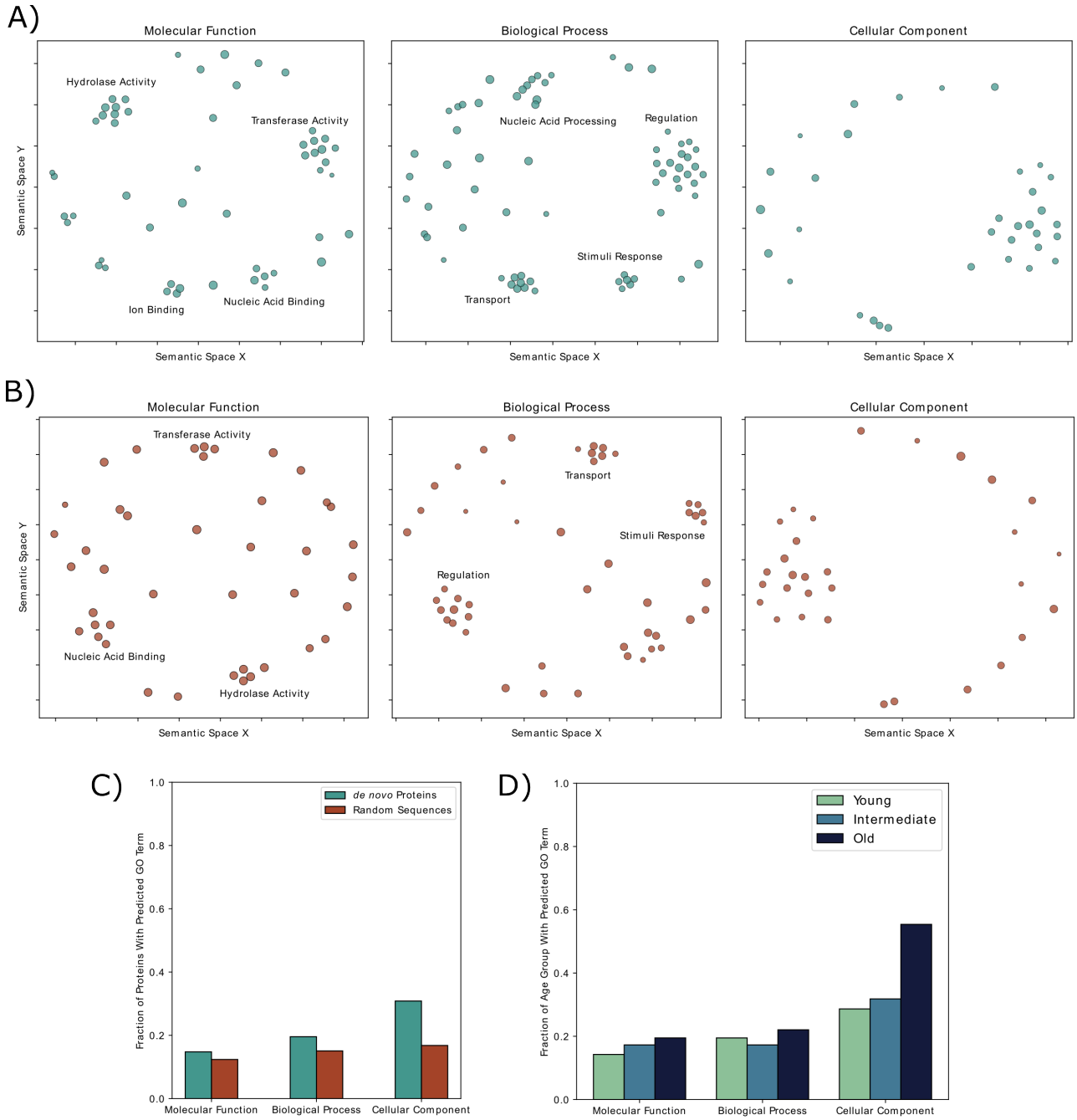

Figure 2: GO terms of random and *de novo* proteins predicted with DeepFRI
We predicted GO terms of *de novo* proteins (**A**) and random sequences (**B**) with DeepFRI and clustered them based on semantic similarity with REVIGO. We visually identified GO term clusters manually annotated with a generic term that describes all the GO terms within the respective cluster. (**C**) Fraction of *de novo* and random proteins (vertical axis) predicted with a GO term per GO term category (horizontal axis). (**D**) Fraction of *de novo* proteins in different age groups (vertical axis) with a predicted GO term (horizontal axis). *Old de novo* proteins were significantly more often annotated with a GO term in the *Cellular Component* category than expected by chance (Pearson's χ^2-Test; $P < 10^{-10}$).

Subset of *de novo* proteins may form biomolecular condensates

Biomolecular condensates are membraneless compartments formed by proteins via liquid-liquid phase separation, and are involved in several biological processes such as stress response and regulation of transcription [Tsang et al., 2020, Hyman et al., 2014]. We observed that that GO terms concerning RNA binding, transferase activity, and hydrolase activity that predicted for *de novo* proteins (Figure 2), are also important features of condensate-forming proteins [Hadarovich et al., 2023]. Therefore, we predicted the propensity of *de novo* proteins for condensate-formation. To this end, we used another prediction tool called PICNIC [Hadarovich et al., 2023]. However, PICNIC uses AF2 predicted structures and a disorder prediction tool IUPred2A, to predict condensate formation propensity. It has been shown, that both AF2 and IUPred can make qualitatively discordant predictions of *de novo* proteins [Middendorf and Eicholt, 2024, Aubel et al., 2023]. Therefore, we performed additional analyses to ensure a high-confidence prediction of condensate-forming *de novo* proteins (Figure 3A). Specifically, we retrieved 175 known condensate-forming conserved proteins from the CD-CODE database [Rostam et al., 2023] and used them as a positive control dataset. For all these proteins, we calculated the sequence features that are associated with the biological function of their intrinsically disordered regions, e.g. amino acid homorepeats, sequence complexity, and net charge [Zarin et al., 2021]. We clustered sequences based on these sequence features using Uniform Manifold Approximation and Projection (UMAP) [McInnes et al., 2018], a commonly non-linear dimensionality reduction tool (in contrast to principal component analysis, which is linear; Figure 3B). We identified seven clusters of different sizes. Of these, cluster 1 and cluster 3 contained most proteins (88.6%) of the CD-CODE database that we used in our analysis (Figure 3C). The *de novo* proteins in cluster 1 and cluster 3 with a PICNIC score greater than 0.5 can be considered high-confidence condensate forming proteins, because they are not only predicted by PICNIC according to its own criteria, but they also have a similar sequence composition as experimentally validated condensate-forming proteins. In total, we identified 63 such high-confidence condensate-forming *de novo* proteins. We next analysed the age groups of these condensate forming *de novo* proteins. When normalized by the number of proteins per age group, we found intermediate and old *de novo* proteins to be 5.9- and 6.6-fold more often predicted to form condensates than young *de novo* proteins, respectively (Figure 3D). Furthermore, intermediate and old *de novo* proteins contained significantly more high-confidence condensate-forming proteins than expected by chance (Pearson's χ^2-Test; $P < 5 \times 10^{-54}$).

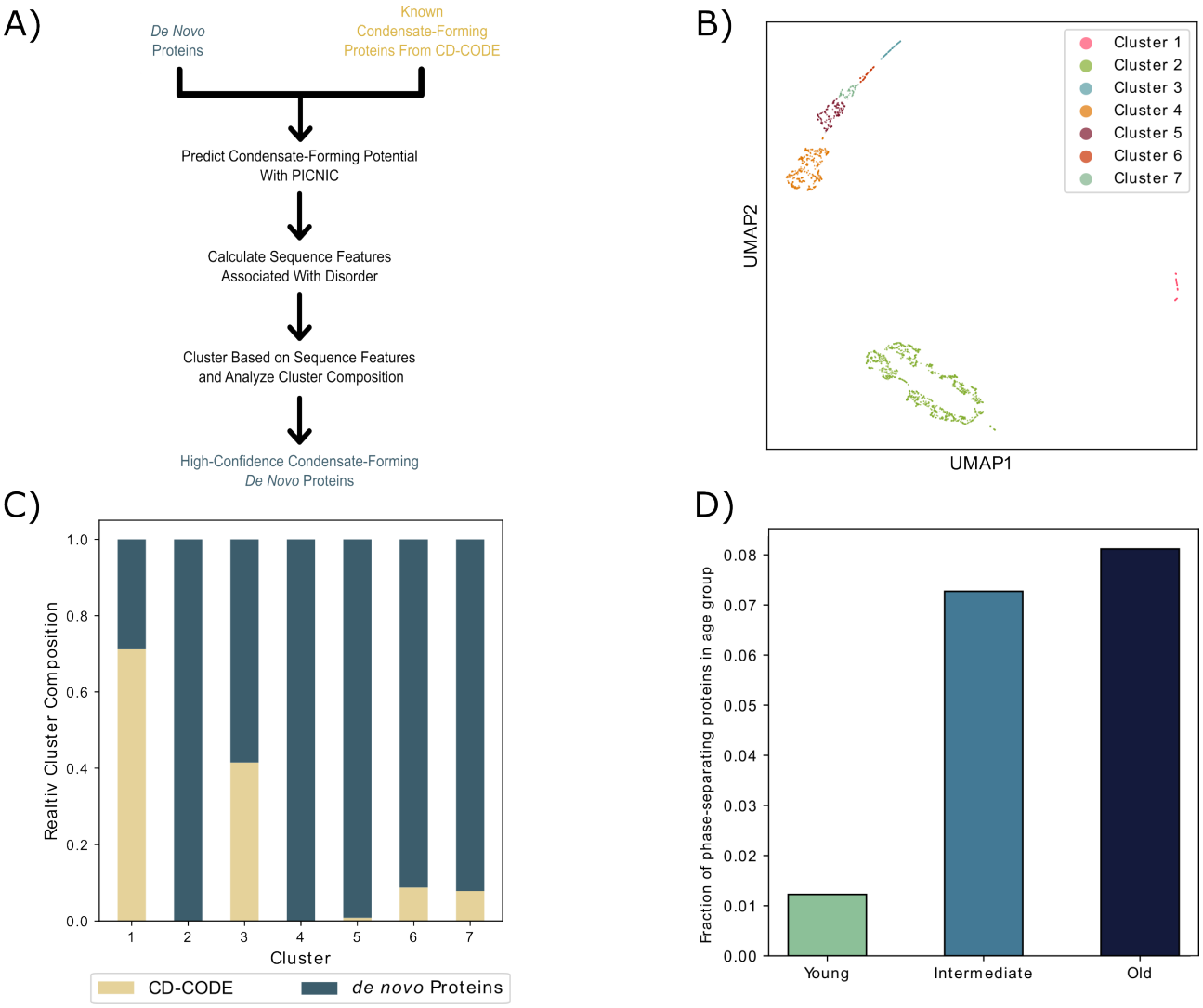

Figure 3: Identification of condensate-forming *de novo* proteins.
(**A**) Workflow for the identification of condensate-forming *de novo* proteins. We predicted condensate-forming potential of *de novo* proteins and known condensate-forming proteins from the CD-CODE database with PICNIC. For both groups of proteins, we calculated the sequence features associated with the functions of intrinsically disordered regions were calculated. Subsequently, we clustered all proteins based on these sequence features using hdbscan, and the analyzed the clusters for their constituent proteins. (**B**) Clusters of *de novo* proteins and known condensate-forming proteins based on sequence features associated with the function of intrinsically disordered proteins. (**C**) Constitution of the identified clusters based on protein type. We classified the 63 *de novo* proteins from clusters 1 and 3 were as high-confidence condensate-forming proteins. (**D**) Fraction of *de novo* proteins from the respective age groups that were classified as high-confidence condensate-forming proteins. The age groups *Intermediate* and *Old* contained significantly more high-confidence condensate-forming proteins than expected by chance (Pearson's χ^2-Test; $P < 5 \times 10^{-54}$).

Protein language models show that *de novo* and conserved proteins occupy distinct regions of the sequence space

Although we found that some *de novo* proteins may be structurally similar to known proteins, we don't yet know if evolutionary origin indeed determines the structural properties of a protein. Indeed, many studies have compared a handful of features such as structural disorder, protein composition, and aggregation propensity between *de novo* and conserved proteins [Knowles and McLysaght, 2009, Ekman and Elofsson, 2010, Landry et al., 2015, Wilson et al., 2017, Vakirlis et al., 2018, Klasberg et al., 2018, Schmitz et al., 2018, Heames et al., 2020, 2023, Peng and Zhao, 2023, Middendorf and Eicholt, 2024]. However, these analyses may not provide reliable inferences because they use tools depending on limited data (e.g. TANGO/IUPred) [Fernandez-Escamilla et al., 2004, Erdős et al., 2021], and because the different features are analysed in isolation. Language models use machine learning to analyse several hidden parameters (and their interactions) simultaneously using sequence information alone. Indeed, protein language models have proved extremely adept at predicting and designing protein structures [Heinzinger et al., 2019, Madani et al., 2023, Alley et al., 2019, Chowdhury et al., 2022, Ferruz and Höcker, 2022, Ferruz et al., 2023, Lin et al., 2023].Therefore, we used the ESM2 protein language model to compare the three different kinds of protein sequences in our dataset (random, *de novo* and conserved proteins). Specifically, we generated a numerical vector for each protein sequence using the ESM2 language model with 650 million parameters (ESM2-650M) [Lin et al., 2023]. Each vector contains 1280 elements, that denote an abstraction of different sequence features predicted by the model. We used UMAP [McInnes et al., 2018] to visualize the protein sequences in sequence space, and found that *de novo*, random, and conserved proteins indeed occupy distinct regions in the sequence space (Figure 4). To quantify these observations, we calculated the Manhattan distance (or L1 norm) between every pair of protein numerical sequences, a method particularly effective for multidimensional data with potential extreme outliers [Barrodale, 1968]. Our findings indicate that the distances between *de novo* and conserved proteins are generally larger than those between sequences within each of these categories (one-sided Mann-Whitney U test; $P < 10^{-99}$). We also found that the distances between the *de novo* and conserved proteins are generally larger than the distances between the *de novo* and the random proteins (one-sided Mann-Whitney U test; $P < 10^{-99}$). The generated random proteins were based on the same length and amino acid distributions as the *de novo* proteins [Middendorf and Eicholt, 2024, Heames et al., 2023]. Therefore, the nearness between these

two sets of protein sequences could be an artifact of our method. To verify if this is the case, we generated random protein sequences with same distribution of composition as our conserved sequences. We found that *de novo* proteins were closer to these new random proteins than with conserved proteins (one-sided Mann-Whitney U test; $P < 10^{-99}$; Figure S4). Overall our analyses suggest that despite certain structural similarities, *de novo* proteins are, distinct from conserved proteins at the sequence level, and bear a closer resemblance to random sequences.

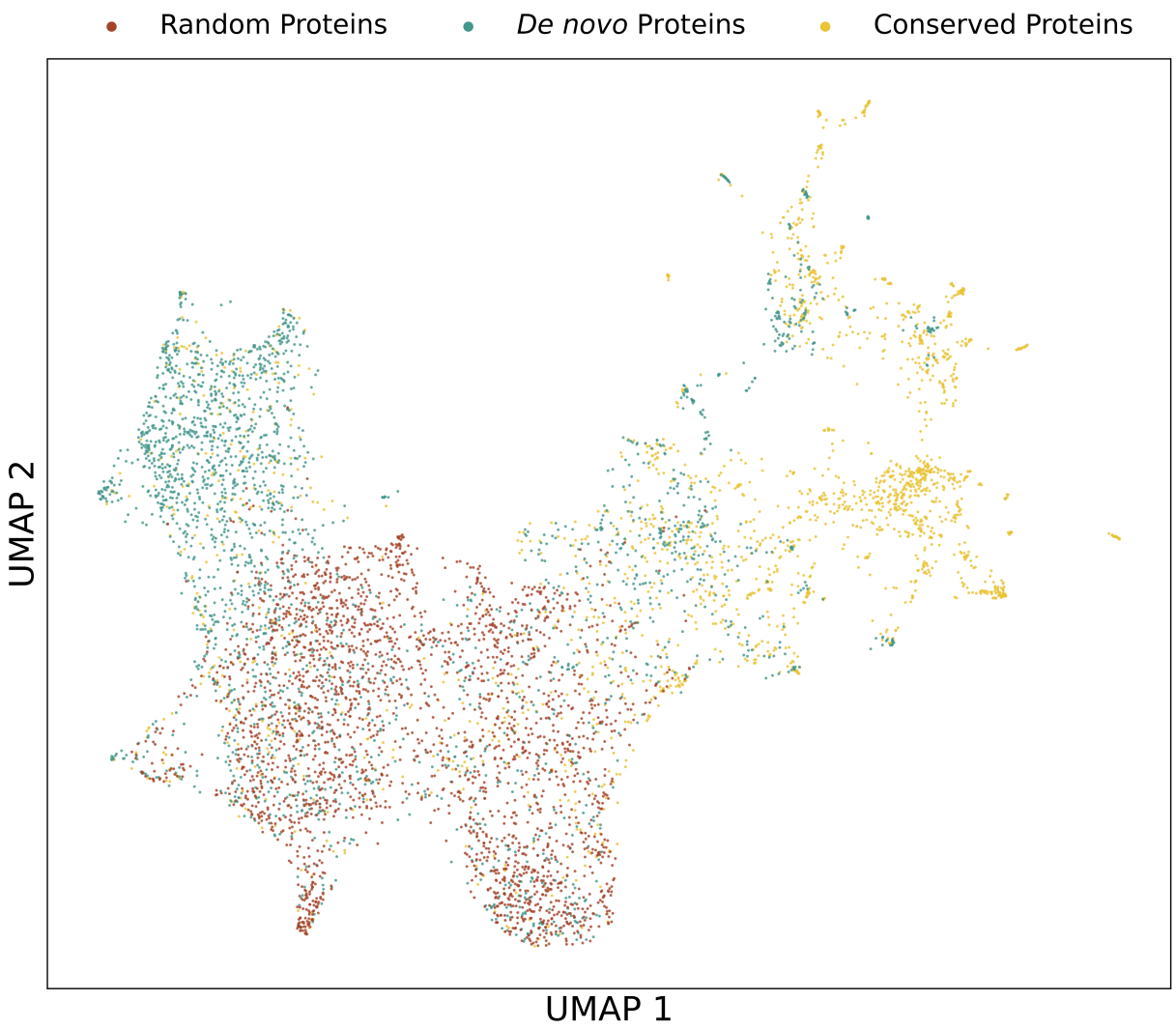

Figure 4: Location of our protein sequences in the sequence space
We used the protein language model ESM2-650M to generate a numerical representation of the *de novo*, random and conserved proteins sequences. We projected and plotted these numerical sequences into a two dimensional space using UMAP.

Discussion

Most proteins can be grouped into families based on their sequence similarity, evolutionary ancestry, structural folds, and biochemical functions [Chothia, 1992]. *De novo* proteins are exceptions as they do not belong to any established protein family, because they not only originate from non-genic DNA sequences (lack of ancestry), but also lack sequence and structural homology to other proteins [Bornberg-Bauer et al., 2021, Schlötterer, 2015]. This makes it challenging to annotate functions to *de novo* proteins based on our knowledge of conserved proteins. Despite their dissimilarity with known proteins, *de novo* proteins have been shown to perform biological functions and improve the survival and fitness of the organisms that express them [McLysaght and Guerzoni, 2015, Li et al., 2009, Cai et al., 2008, Chen et al., 2010a, Gubala et al., 2017, Lange et al., 2021, Zhuang et al., 2019, Reinhardt et al., 2013, Heinen et al., 2009, Li et al., 2010a, Xie et al., 2019, Li et al., 2014, Chen et al., 2010b]. Advanced computational methods using deep learning have been able to solve problems at an unprecedented scale. For example, AlphaFold2 resulted in an exponential increase in the number of computationally predicted protein structures [Varadi et al., 2021] Therefore, we applied some of these deep learning based tools to elucidate the possible structure and function of *de novo* proteins.

First, we searched for conserved proteins that may be structurally similar to *de novo* proteins using Foldseek. Most *de novo* proteins did not bear a significant resemblance to known protein structures, in accordance with their non-genic evolutionary origin, and distinctiveness of their sequence and biophysical properties as shown by previous studies [Heames et al., 2023, Aubel et al., 2024]. This lack of resemblance could exist because *de novo* proteins are highly disordered [Middendorf and Eicholt, 2024, Peng and Zhao, 2023], and can contain rare secondary structures like 3_{10}- or π-helices [Chen et al., 2023], that could make structural alignment complicated.

While we attempted to refine AF2 predicted structures of *de novo* proteins through molecular dynamics (MD) simulations, it is important to note that many *de novo* proteins may reside in non-aqueous environments such as cell membranes (Figure Figure 2) [Vakirlis et al., 2020b], may only fold upon interaction with other proteins [Chen et al., 2023], and may be part of multimers [Lynch, 2012, Schulz et al., 2022, Jayaraman et al., 2022, Malik et al., 2024]. We did not consider all these possibilities in our MD simulations due to computational limitations. Nonetheless, the majority of individual *de novo* proteins were predicted to be disordered or, if structured, to predominantly form simple α-helices [Heames et al., 2023, Middendorf and Eicholt, 2024, Aubel et al., 2024, Peng and

Zhao, 2023], a trend attributed to many *de novo* proteins being too short to form globular structures [Aubel et al., 2024, Shen et al., 2005]. Our current study corroborates these observations. The frequent emergence of single α-helices in *de novo* proteins can be attributed to the lower stereochemical and thermodynamical requirements of α-helices [Barlow and Thornton, 1988, Greenwald and Riek, 2012]. On rare occasions where *de novo* proteins exhibit structural configurations beyond single α-helices, they can resemble common and ancient folds such as SH3 or HTH (Figure 1D). This observation implies that these widespread evolutionary folds, which are evolutionary successful and easily tolerated by cells, are more accessible in sequence space [Taverna and Goldstein, 2000, Shakhnovich et al., 2005, Goldstein, 2008], even for sequences that have not been shaped by millions of generations of evolution. Despite identifying some *de novo* proteins with structural homology to existing structures, we did not find any novel folds among our candidate proteins, unlike other studies that investigated a much larger set of proteins [Durairaj et al., 2023] (Figure 1B & D).

By employing the deep learning based functional annotation tool, DeepFRI [Gligorijević et al., 2021], we found that *de novo* proteins are associated with a wide array of Gene Ontology (GO) terms, spanning all three GO categories, with several distinct clusters emerging within the semantic field. We show that *de novo* proteins, despite their recent emergence and lack of evolutionary ancestry, are more often predicted to be functional than a comparable random set of sequences (Figure 2C). Their involvement in a range of molecular functions (like hydrolase activity, transferase activity, and nucleic acid binding) and biological processes (such as stimuli response, regulation, and transport) underscores their potential impact on the cellular physiology. Interestingly, the similarity in molecular functions and involvement in biological processes between *de novo* proteins and random sequences could imply a certain level of functional redundancy in the sequence space. This observation might suggest that the emergence of function from novel proteins, even through random sequences, could be a more probable phenomena than previously thought. Finally we emphasize that, while efforts to deduce protein function based on structural similarity is ongoing [Nomburg et al., 2024, Gligorijević et al., 2021], numerous instances exist where proteins with similar structures perform different functions, and *vice versa* [Finkelstein et al., 1993, Govindarajan and Goldstein, 1996, Galperin et al., 1998, Martin et al., 1998].

The association of *de novo* proteins with biophysical reactions such as RNA binding, and biochemical reactions similar to transferases, and hydrolases, presents an intriguing avenue for understanding their functional capacities and evolutionary significance. This is especially interesting because

RNA binding and hydrolase-activity are thought to be conferred even by primordial folds [Seal et al., 2022, Weil-Ktorza et al., 2023, Vyas et al., 2021, Longo et al., 2022], and could possibly been important during origin of life. Both these molecular activities, and a highly disordered structure, are also exhibited by condensate-forming proteins [Hadarovich et al., 2023]. Therefore, we investigated the possibility of *de novo* proteins to be involved in formation of biomolecular condensates. Biomolecular condensates, formed through liquid-liquid phase separation by proteins, are critical in various biological processes and such a propensity exists even for proteins with ancient and simple folds [Longo et al., 2020]. The use of PICNIC [Hadarovich et al., 2023] to predict the involvement of *de novo* proteins in biomolecular condensates represents an innovative approach, albeit with limitations. The reliance on AlphaFold2 predictions and IUPred2A as input requirements, introduces a degree of uncertainty, especially given the discordant predictions of these tools between *de novo* and conserved proteins [Middendorf and Eicholt, 2024]. This necessitated further analysis to establish a high-confidence set of condensate-forming *de novo* proteins, leveraging the CD-CODE database [Rostam et al., 2023] as a reference.

The identification of clusters based on sequence features associated with intrinsically disordered regions of proteins is particularly noteworthy. The fact that clusters 1 and 3, which have a high fraction of members from the CD-CODE database, include ≈12% of all *de novo* proteins with a PICNIC score greater than 0.5, is compelling. It suggests that these *de novo* proteins not only have the potential to form condensates but also share sequence composition with experimentally validated condensate-forming proteins. The discovery of 63 high-confidence condensate-forming *de novo* proteins contributes to our understanding of the functional diversity of these proteins. This finding expands the realm of *de novo* protein functionality beyond traditional views, indicating their potential involvement in complex cellular mechanisms like phase separation. Considering that phase separation is involved in spermatogenesis [Kang et al., 2022, Parvinen, 2005], and that *de novo* proteins show biased expression in testis [Levine et al., 2006, Heames et al., 2020, Zhao et al., 2014, Palmieri et al., 2014, Peng and Zhao, 2023, Nyberg and Carthew, 2017, Kondo et al., 2017, Neme and Tautz, 2013], being involved in biomolecular condensates suggests a possible mechanism by which *de novo* proteins could play a role in spermatogenesis [Lange et al., 2021, Gubala et al., 2017, Rivard et al., 2021]. Moreover, our analysis of the age groups of these *de novo* proteins revealed that intermediate and old *de novo* proteins are significantly more likely to form condensates than their younger counterparts. This observation is intriguing as it could imply two scenarios. First, as *de novo* protein evolve and mature, they acquire and refine their ability to par-

ticipate in cellular processes like biomolecular condensation and thereby their function. Under this scenario, the *de novo* proteins could be positively selected. Second, the ability to form biomolecular condensates could minimize toxic protein aggregation, and could protect *de novo* proteins from being purged by negative selection.

To understand if *de novo* proteins can indeed be a source of evolutionary novelty, we analyzed their distribution in the protein sequence space relative to that of conserved and random proteins, using the protein language model ESM2-650M. Our analysis shows that *de novo* proteins, arisen from non-coding sequences, have unique sequence characteristics that distinguish them from conserved proteins, but more similar to random proteins, as hypothesized before [Bornberg-Bauer et al., 2021]. Nevertheless, some *de novo* proteins indeed had a conserved protein, closely located to them in the sequence space (Figure 4). Together with our Foldseek analysis, this observation indicates an inherent capacity of amino acid sequences to adopt structures, and that a broad spectrum of sequence space is capable of evolving into foldable proteins [Tretyachenko et al., 2017, Heames et al., 2023, Aubel et al., 2024].

Our analysis is based on computational tools, which are always prone to some level of erroneous predictions. Furthermore, many of the deep learning based tools have not been trained on *de novo* proteins and can possibly make biased predictions [Middendorf and Eicholt, 2024]. Therefore, our study may not provide exact and perfect answers to the different open questions about *de novo* proteins. All computational predictions need experimental validation. Experimental studies, especially on *de novo* proteins are bottlenecked by serendipity, and labor intensive techniques that are not fully optimized for proteins with such an unusual biochemistry [Eicholt et al., 2022]. However, our exhaustive approach can help guide focused experimental studies, and can reduce the need for trial and error, and accidental discoveries. For example, the candidate *de novo* proteins with a possible structure, a specific molecular function (like hydrolysis, or RNA binding), and a propensity to form condensates, can be experimentally probed for these specific properties. Our sequence space analysis can also identify *de novo* proteins that are likely to adopt more conserved-protein-like properties, as a consequence of evolution. Overall, our study not only broadens our understanding of the dynamic nature of protein evolution but also serves as a guidebook for future experimental studies.

Materials & Methods

Dataset curation

We used the sequence datasets from our previous study [Middendorf and Eicholt, 2024]. Specifically, we first obtained 6716 orphan protein sequences from the *Drosophila* clade, and their corresponding evolutionary age, from Heames et al. [2020]. From this dataset, we discarded sequences that were annotated with the same FlyBase ID. Next, we extracted the sequences whose emergence origin was annotated as "*denovo*" (intergenic *de novo* protein) or "*denovo-intron*" (intronic *de novo* protein) by Heames et al. [2020], for further analysis. Out of the 2510 proteins sequences thus obtained, 1481 were annotated as "*denovo*," while 1029 were described as "*denovo-intron*". Based on their date of emergence, the *de novo* proteins were classified as young (<5 mya), intermediate (5-30 mya), and old (>30 mya) proteins [Heames et al., 2020, Middendorf and Eicholt, 2024]. In our filtered dataset, the three age groups consisted of 2205, 110, and 195 proteins, respectively.
We generated 2507 random sequences with the same distributions of amino acid composition and sequence length, as the 2510 *de novo* sequences set, using a technique used in previous studies [Heames et al., 2023, Middendorf and Eicholt, 2024]. We generated a set of conserved protein sequences with the same sequence length distribution as the *de novo* proteins, by randomly sampling protein sequences from the combined proteome of 12 *Drosophila* species. After removing sequences that were duplicated or were redundant with our set of *de novo* proteins, we obtained a set of 2235 unique conserved proteins.
We performed structure predictions using AlphaFold2 [v2.1.1, Jumper et al., 2021] on the High Performance Computing Cluster, PALMA II (University of Muenster). We used the predictions with the highest mean pLDDT for further analysis. We downloaded AlphaFold2 based structure predictions of conserved *Drosophila* proteins from the AlphaFold Protein Structure database [Varadi et al., 2021] for our initial analyses.

Molecular Dynamics simulations to refine structure predictions

To analyze the stability of the predicted structures of *de novo* proteins, we performed molecular dynamics (MD) simulations using a previously described method [Ferruz et al., 2022], with minor modifications. We only simulated protein structures with i) less than 30% disorder predicted by flDPnn [Hu et al., 2021], and ii) less than 95% of their residues predicted as α-helices by DSSP

[Kabsch and Sander, 1983] (1468 unique proteins). We constructed the MD model and performed the simulations using the HTMD python package [Doerr et al., 2016]. The model systems were built to form solvated all-atom cubic boxes. We centered our proteins at the origin of the simulation box coordinates. We used water as the solvent, and added NaCl ions to neutralize the system. We used the AMBER 14SB force field [Maier et al., 2015] for all simulations. We minimized, equilibirated, and simulated each system for 100 ns, using the ACEMD engine [Harvey et al., 2009] with the default settings in triplicates. We evaluated the simulations with the HTMD [Doerr et al., 2016] and MDAnalysis [Michaud-Agrawal et al., 2011] python packages. We calculated the average RMSD value per trajectory for every replicate simulation for a protein, and in turn calculated a single averaged value from three replicates.

Identifying similar protein structures using Foldseek

We searched the AlphaFold Protein Structure database [Varadi et al., 2021] clustered at 50% sequence identity (AFDB50), for structures similar to the predicted structures of our *de novo*, random, and conserved proteins, using Foldseek [v.8.ef4e960, van Kempen et al., 2023]. We applied the same filtering criteria our query proteins that we used for the MD simulations. For *de novo* proteins, we used the protein structures refined after 100ns of MD simulation. We downloaded pre-computed AFDB50 database via Foldseek's database module. We searched for similar structures using the "easy-search" module of Foldseek with the default settings. We did not filter the results or queries based on the pLDDT values. We discarded all hits to proteins within the *Drosophila* clade, to exclude hits to orthologous *de novo* proteins.

To identify and annotate potential known protein structural domains in the *de novo* proteins, we searched the protein data bank database [PDB, January 2024; Berman et al., 2000] for structures that were similar to that of *de novo* proteins (MD-refined). We used Foldseek for this analysis with the same settings as we did before for AFDB50. We discarded hits with a TM-score less than 0.5 [Xu and Zhang, 2010]. We retrieved the annotated ECOD domains of the highest scoring hits, from the ECOD database [Release: 20230309, Cheng et al., 2014] if the structural alignment of the *de novo* protein covered at least 80% of the target structure from the PDB. In all cases, we only used the highest scoring hit out of the three MD replicates for further analysis.

Predicting protein function using DeepFri

To understand the potential function of *de novo*, and random proteins, we predicted their gene ontology (GO) terms using DeepFRI [Gligorijević et al., 2021]. We used their AlphaFold2 predicted 3D-structures as the input and identified hits with a score ≥ 0.5. We summarized the predicted GO terms to a small list of terms using using REVIGO [Supek et al., 2011], and measured semantic similarity using SimRel [Sæbø et al., 2015]. We visually, identified clusters within the semantic space and annotated them with a term that summarizes the GO terms within them.

Analysis of *de novo* proteins that form biomolecular condensates

We predicted the potential of *de novo* proteins to form biomolecular condensates, using PICNIC [Hadarovich et al., 2023]. Because PICNIC makes predictions based on metrics derived from AlphaFold2 and IUPred2A predictions, we applied further filtering steps of the results in order to obtain a set of high-confidence condensate-forming *de novo* proteins. To this end, we retrieved all the proteins from the CD-CODE database [Rostam et al., 2023], that were experimentally shown from biomolecular condensates *in cellulo* or *in vivo*. This set of 175 proteins served as our positive control. Next, we retrieved sequence features associated with the biological functions of intrinsically disordered regions of proteins [Zarin et al., 2021], using the scripts provided in the idr.mol.feats GitHub repository. We discarded the specific features – *aromatic_spacing*, *omega_aromatic**, and *kappa**, and features that count the appearance of specific binding motifs. We normalized all the features that are directly influenced by the sequence length (e.g. amino acid counts), to the sequence length of the corresponding proteins. We subsequently clustered the sequences based on the computed features using hdbscan [McInnes et al., 2017] with a minimal cluster size of 100 the *min_samples* parameter set to a value of 50. We considered a *de novo* protein to be a high-confidence condensate-forming protein, if it shared a cluster with a large fraction of proteins from the CD-CODE database, and had a PICNIC score ≥ 0.5.

Mapping protein sequences to a numerical space using protein language model

To understand how *de novo* and random protein sequences are located within the protein sequence space relative to conserved proteins, we used the protein language model ESM2 with 650 million parameters (ESM2-650M) [Lin et al., 2023]. Specifically, we used the language model to convert

each sequence to a numerical vector with 1280 elements. More specifically, ESM2-650M assigns each amino acid residue in a protein sequence, a 1280-dimensional vector of “embeddings”. For each protein we calculated the multivariate mean of the embedding vectors from every amino acid residue. We calculated the Manhattan distance (or L1 norm) between the numerical sequences of every pair of proteins in our combined dataset of *de novo*, random and conserved proteins. We applied Mann-Whitney test to the pairwise distances to analyse if proteins of one class (*e.g. de novo*) are farther from that of another class (*e.g.* conserved), than with each other. For proteins of one class, we also used the pairwise distances to identify the nearest neighboring protein from the other class. To visualize the location of different proteins in the sequence space, we used UMAP to project and visualize the proteins (numerical sequence) in a two dimensional space [V 0.5.3, McInnes et al., 2018]. We used UMAP with the default settings (n_neigbours = 15, min_distance = 0.1), except for choosing Manhattan distance as the distance metric and optimizing the low dimensional embedding for 5000 instead of 200 epochs.

Data & statistical analysis

We analyzed most data with Python programming language (v3.9.18) [Van Rossum and Drake, 2009], using the following packages: Pandas (v1.5.3) [Reback et al., 2022], NumPy (v1.26) [Harris et al., 2020], SciPy (v1.11.3) [Virtanen et al., 2020], and BioPython (v1.80) [Cock et al., 2009]. We generated the plots using Matplolib (v3.4.3) [Hunter, 2007]. We performed the Pearson’s χ^2-tests using the "chisquare" from the scipy.stats package. We analyzed protein sequence space with Julia programming language using the packages Distances.jl (v0.10.11) and HypothesisTests.jl (v 0.11.0)

Acknowledgements

We thank Alun Jones for his advice on statistical tests.

Supporting information

Supporting information is available on Zenodo 10.5281/zenodo.10557890.

Code and Data Availability

Datasets are publicly available on Zenodo. All scripts are freely available on GitHub:
https://github.com/LasseMiddendorf/SequenceAndFunctionalSpaceOfDrosophilaDeNovoProteins

Supplementary Material

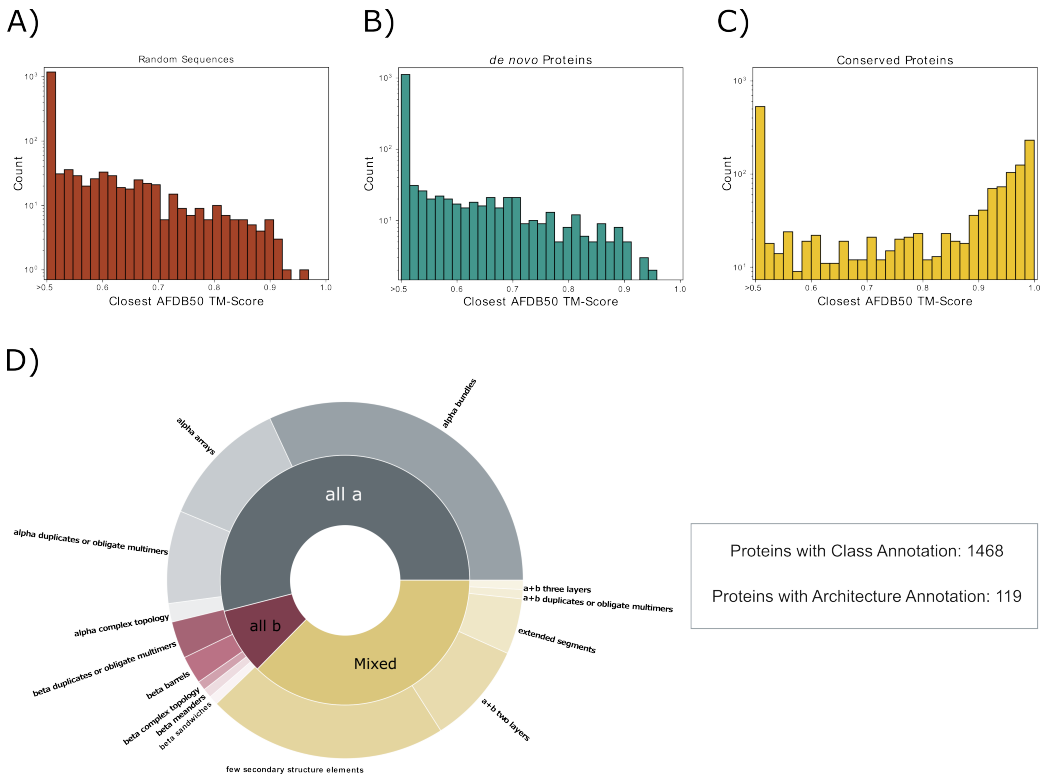

Figure S1: Structural diversity of *de novo* proteins, before MD refinement. The predicted protein structures of randomly generated sequences (**A**), *de novo* protein (**B**), and conserved proteins (**C**) were queried against the AlphaFold database (AFDB50) excluding proteins from *Drosophila*. Only proteins with less than 30% of their residues being predicted to be disordered and less than 95% with a DSSP annotation of being α-helical were considered for the analysis. Shown is the distribution of the highest TM-score found for each protein in the three datasets. (**D**) Overview of the structural classes and ECOD architectures of *de novo* proteins. The protein class (inner circle) was assigned to all *de novo* proteins queried against the AFDB50 based on the DSSP annotations of the predicted protein structures. Proteins containing no residues annotated as α-helices or β-sheets were classified as *all b* or *all a*, respectively. Protein structures containing residues annotated as α-helices and β-sheets were classified as *Mixed*. For the annotation of ECOD architectures in the predicted structures of *de novo* protein, the structures were queried against the PDB and assigned with the ECOD domain of the highest ranking hit if the alignment covered at least 80% of the target structure.

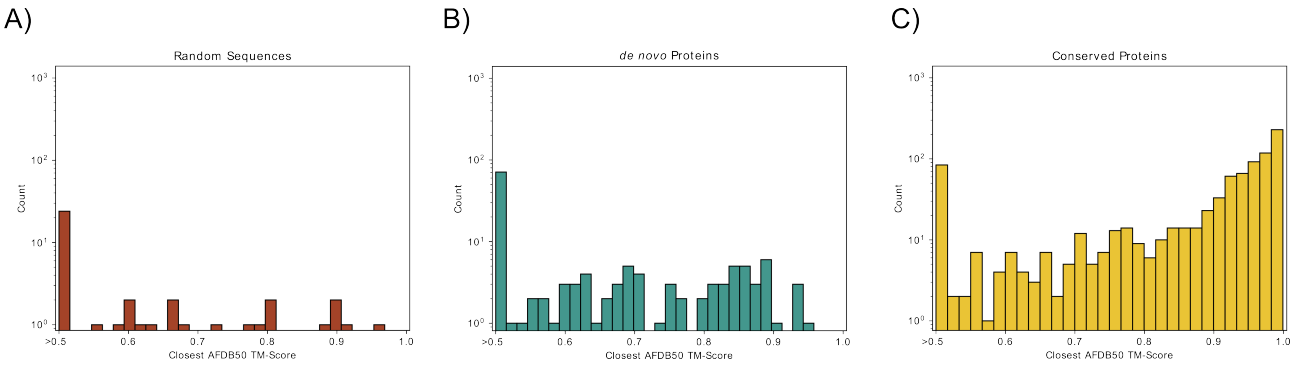

Figure S2: Structural similarity of high-pLDDT protein structures to AlphaFold database Similar structures in the AlphaFold database for high-pLDDT structure predictions only TM-Score distribution of predicted protein structures of (**A**) random, (**B**) *de novo*, and (**C**) conserved proteins with a pLDDT value >= 70 queried against the AlphaFold database (AFDB50) using Foldseek. The hit with the highest TM-score was chosen for each protein.

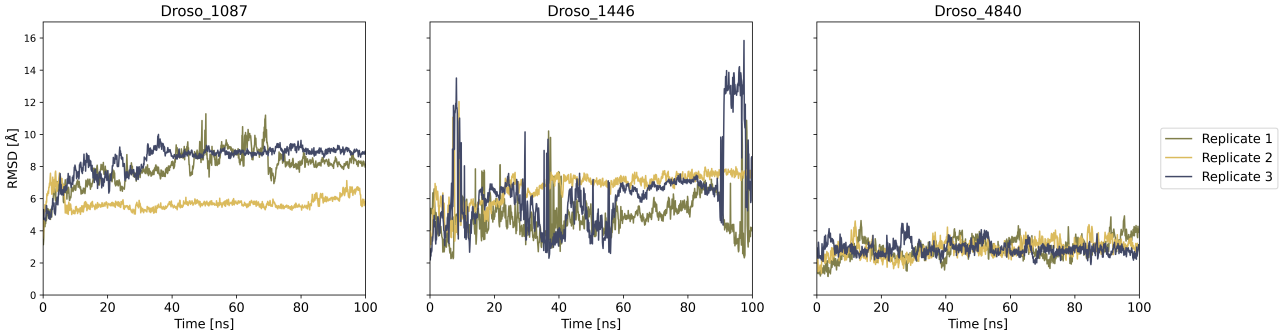

Figure S3: RMSD trajectories of selected *de novo* proteins Root mean square deviation (RMSD) of Droso_1087, Droso_1446, and Droso_4480 over 100 ns of molecular dynamics simulations. Simulations were performed as triplicates for all proteins.

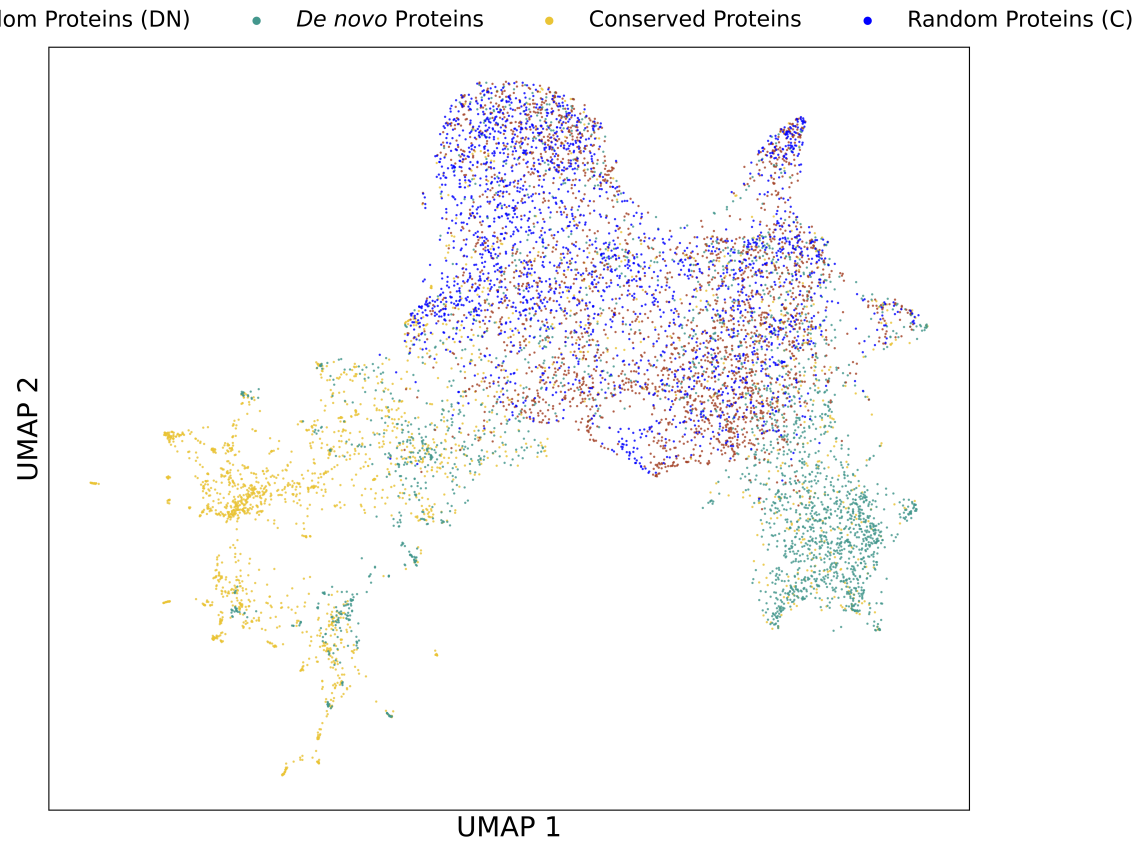

Figure S4: *De novo* proteins are closer to random sequences than conserved proteins. We used the protein language model ESM2-650M to represent the sequences of *de novo*, conserved, and random proteins as numerical vectors. In addition to random proteins that were generated to share the same amino acid distribution as *de novo* proteins (Random Proteins (DN)), we included a set of randomly generated sequences based on the properties of conserved proteins (Random Proteins (C)). We projected the representations into two dimensions using UMAP. The localization in sequence space shows that *de novo* proteins are closer to random proteins than conserved ones, regardless of the origin of the random sequences.

References

Emile Zuckerkandl. The appearance of new structures and functions in proteins during evolution. *Journal of molecular evolution*, 7(1):1–57, 1975.

François Jacob. Evolution and tinkering. *Science (New York, N.Y.)*, 196(4295):1161–1166, 1977.

Anthony D. Keefe and Jack W. Szostak. Functional proteins from a random-sequence library. *Nature*, 410(6829):715–718, April 2001.

Michael H Hecht, Aditi Das, Abigail Go, Luke H Bradley, and Yinan Wei. De novo proteins from designed combinatorial libraries. *Protein Science*, 13(7):1711–1723, 2004.

Arianne M Babina, Serhiy Surkov, Weihua Ye, Jon Jerlström-Hultqvist, Mårten Larsson, Erik Holmqvist, Per Jemth, Dan I Andersson, and Michael Knopp. Rescue of escherichia coli auxotrophy by de novo small proteins. *Elife*, 12:e78299, 2023.

Michael Knopp, Arianne M Babina, Jónína S Gudmundsdóttir, Martin V Douglass, M Stephen Trent, and Dan I Andersson. A novel type of colistin resistance genes selected from random sequence space. *Plos Genetics*, 17(1):e1009227, 2021.

Idan Frumkin and Michael T Laub. Selection of a de novo gene that can promote survival of escherichia coli by modulating protein homeostasis pathways. *Nature Ecology & Evolution*, pages 1–13, 2023.

Fa-An Chao, Aleardo Morelli, John C Haugner Iii, Lewis Churchfield, Leonardo N Hagmann, Lei Shi, Larry R Masterson, Ritimukta Sarangi, Gianluigi Veglia, and Burckhard Seelig. Structure and dynamics of a primordial catalytic fold generated by in vitro evolution. *Nature Chemical Biology*, 9(2):81–83, 2013.

Asao Yamauchi, Toshihiro Nakashima, Nobuhiko Tokuriki, Masato Hosokawa, Hideki Nogami, Shingo Arioka, Itaru Urabe, and Tetsuya Yomo. Evolvability of random polypeptides through functional selection within a small library. *Protein Engineering*, 15(7):619–626, 2002.

Alan R Davidson and Robert T Sauer. Folded proteins occur frequently in libraries of random amino acid sequences. *Proceedings of the National Academy of Sciences*, 91(6):2146–2150, 1994.

Alan R Davidson, Kevin J Lumb, and Robert T Sauer. Cooperatively folded proteins in random sequence libraries. *Nature structural biology*, 2(10):856–864, 1995.

Vyacheslav Tretyachenko, Jiří Vymětal, Lucie Bednárová, Vladimír Kopecký, Kateřina Hofbauerová, Helena Jindrová, Martin Hubálek, Radko Souček, Jan Konvalinka, Jiří Vondrášek, and Klára Hlouchová. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Scientific Reports*, 7(1):15449, 2017.

Paola Lo Surdo, Martin A. Walsh, and Maurizio Sollazzo. A novel adp- and zinc-binding fold from function-directed in vitro evolution. *Nature Structural &Molecular Biology*, 11:382–383, 2004.

Sheref S. Mansy, Jinglei Zhang, Rainer Kümmerle, Mikael Nilsson, James Jeiwen Chou, Jack W. Szostak, and John Charles Chaput. Structure and evolutionary analysis of a non-biological atp-binding protein. *Journal of molecular biology*, 371 2:501–13, 2007.

Diethard Tautz and Tomislav Domazet-Lošo. The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702, 2011.

Anne-Ruxandra Ruxandra Carvunis, Thomas Rolland, Ilan Wapinski, Michael A Calderwood, Muhammed A Yildirim, Nicolas Simonis, Benoit Charloteaux, César A Hidalgo, Justin Barbette, Balaji Santhanam, Gloria A Brar, Jonathan S Weissman, Aviv Regev, Nicolas Thierry-Mieg, Michael E Cusick, and Marc Vidal. Proto-genes and de novo gene birth. *Nature*, 487(7407): 370–374, 2012.

Stephen Branden Van Oss and Anne-Ruxandra Carvunis. De novo gene birth. *PLoS Genetics*, 15, 2019.

Nikolaos Vakirlis, Anne Ruxandra Carvunis, and Aoife McLysaght. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife*, 9:1–23, 2020a.

Erich Bornberg-Bauer, Klara Hlouchova, and Andreas Lange. Structure and function of naturally evolved de novo proteins. *Current Opinion in Structural Biology*, 68:175–183, 2021.

Jonathan F Schmitz and Erich Bornberg-Bauer. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding dna. 6:57, 2017.

Brennen Heames, Filip Buchel, Margaux Aubel, Vyacheslav Tretyachenko, Dmitry Loginov, Petr Novák, Andreas Lange, Erich Bornberg-Bauer, and Klára Hlouchová. Experimental characteri-

zation of de novo proteins and their unevolved random-sequence counterparts. *Nature Ecology & Evolution*, pages 1–11, 2023.

Annamária F Ángyán, András Perczel, and Zoltán Gáspári. Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? *FEBS letters*, 586(16):2468–2472, 2012.

Devika Bhave and Diethard Tautz. Effects of the expression of random sequence clones on growth and transcriptome regulation in escherichia coli. *Genes*, 13(1):53, 2021.

Johana F Castro and Diethard Tautz. The effects of sequence length and composition of random sequence peptides on the growth of e. coli cells. *Genes*, 12(12):1913, 2021.

Lasse Middendorf and Lars A Eicholt. Random, de novo, and conserved proteins: How structure and disorder predictors perform differently. *Proteins: Structure, Function, and Bioinformatics*, 2024.

Margaux Aubel, Filip Buchel, Brennen Heames, Alun Robert Claude Jones, Ondrej Honc, Erich Bornberg-Bauer, and Klara Hlouchova. High-throughput selection of human de novo-emerged sorfs with high folding potential. *bioRxiv*, 2024. doi: 10.1101/2024.01.22.576604. URL https://www.biorxiv.org/content/early/2024/01/24/2024.01.22.576604.

Brennen Heames, Jonathan Schmitz, and Erich Bornberg-Bauer. A continuum of evolving de novo genes drives protein-coding novelty in drosophila. *Journal of molecular evolution*, 88(4):382–398, 2020.

Junhui Peng and Li Zhao. The origin and structural evolution of de novo genes in drosophila. *bioRxiv*, 2023. doi: 10.1101/2023.03.13.532420. URL https://www.biorxiv.org/content/early/2023/03/15/2023.03.13.532420.

Christian R Landry, Xiangfu Zhong, Lou Nielly-Thibault, and Xavier Roucou. Found in translation: functions and evolution of a recently discovered alternative proteome. *Current Opinion in Structural Biology*, 32:74–80, 2015.

Eric B Zheng and Li Zhao. Protein evidence of unannotated orfs in drosophila reveals diversity in the evolution and properties of young proteins. *Elife*, 11:e78772, 2022.

Jonathan F Schmitz, Kristian K Ullrich, and Erich Bornberg-Bauer. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nature ecology & evolution*, 2(10):1626–1632, 2018.

Walter Basile, Oxana Sachenkova, Sara Light, and Arne Elofsson. High gc content causes orphan proteins to be intrinsically disordered. *PLOS Computational Biology*, 13(3):e1005375, 2017.

Jianhai Chen, Qingrong Li, Shengqian Xia, Deanna Arsala, Dylan Sosa, Dong Wang, and Manyuan Long. One million years of solitude: the rapid evolution of de novo protein structure and complex. *bioRxiv*, pages 2023–12, 2023.

Nikolaos Vakirlis, Omer Acar, Brian Hsu, Nelson Castilho Coelho, S Branden Van Oss, Aaron Wacholder, Kate Medetgul-Ernar, Ray W Bowman, Cameron P Hines, John Iannotta, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature communications*, 11(1):781, 2020b.

Andreas Lange, Prajal H Patel, Brennen Heames, Adam M Damry, Thorsten Saenger, Colin J Jackson, Geoffrey D Findlay, and Erich Bornberg-Bauer. Structural and functional characterization of a putative de novo gene in drosophila. *Nature communications*, 12(1):1–13, 2021.

Dixie Bungard, Jacob S Copple, Jing Yan, Jimmy J Chhun, Vlad K Kumirov, Scott G Foy, Joanna Masel, Vicki H Wysocki, and Matthew H J Cordes. Foldability of a natural de novo evolved protein. *Structure*, 2017.

Helle Tessand Baalsrud, Ole Kristian Tørresen, Monica Hongrø Solbakken, Walter Salzburger, Reinhold Hanel, Kjetill S Jakobsen, and Sissel Jentoft. De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. *Molecular Biology and Evolution*, 35(3):593–606, March 2018.

Tatsuhito Matsuo, Kazuma Nakatani, Taiki Setoguchi, Koichi Matsuo, Taro Tamada, and Yusuke Suenaga. Secondary structure of human de novo evolved gene product ncym analyzed by vacuum-ultraviolet circular dichroism. *Frontiers in Oncology*, page 3255, 2021.

Sidi Chen, Yong E. Zhang, and Manyuan Long. New genes in drosophila quickly become essential. *Science*, 330:1682 – 1685, 2010a. URL https://api.semanticscholar.org/CorpusID:7899890.

Anna M. Gubala, Jonathan F. Schmitz, Michael J. Kearns, Tery T. Vinh, Erich Bornberg-Bauer, Mariana F. Wolfner, and Geoffrey D. Findlay. The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen De Novo. *Molecular Biology and Evolution*, 34 (5):1066–1082, May 2017.

Josephine A. Reinhardt, Betty M. Wanjiru, Alicia T. Brant, Perot Saelao, David J. Begun, and Corbin D. Jones. De novo orfs in drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genetics*, 9, 2013. URL https://api.semanticscholar.org/CorpusID:14284334.

Aoife McLysaght and Daniele Guerzoni. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140332, 2015.

Ling Li, Carol Foster, Qinglei Gan, Dan Nettleton, Martha G. James, Alan M. Myers, and Eve Syrkin Wurtele. Identification of the novel protein qqs as a component of the starch metabolic network in arabidopsis leaves. *The Plant journal : for cell and molecular biology*, 58 3:485–98, 2009. URL https://api.semanticscholar.org/CorpusID:22373809.

Jing Cai, Ruoping Zhao, Huifeng Jiang, and Wen Wang. De novo origination of a new protein-coding gene in saccharomyces cerevisiae. *Genetics*, 179(1):487–496, 2008.

Xuan Zhuang, Chun Yang, Katherine R. Murphy, C.-H. Christina Cheng, and C. H. Christina Cheng. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *PNAS*, 116(10):4400–4405, 2019.

Tobias Heinen, Fabian Staubach, Daniela Häming, and Diethard Tautz. Emergence of a new gene from an intergenic region. *Current Biology*, 19:1527–1531, 2009. URL https://api.semanticscholar.org/CorpusID:12446879.

Dan Li, Yang Dong, Yu Jiang, Huifeng Jiang, Jing Cai, and Wen Wang. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Research*, 20:408–420, 2010a. URL https://api.semanticscholar.org/CorpusID:25017053.

Chen Xie, Cemalettin Bekpen, Sven Künzel, Maryam Keshavarz, Rebecca Krebs-Wheaton, Neva Skrabar, Kristian Karsten Ullrich, and Diethard Tautz. A de novo evolved gene in the house

mouse regulates female pregnancy cycles. *eLife*, 8:e44392, August 2019. Publisher: eLife Sciences Publications, Ltd.

Dan Li, Zhihui Yan, Lina Lu, Huifeng Jiang, and Wen Wang. Pleiotropy of the de novo-originated gene mdf1. *Scientific Reports*, 4, 2014. URL https://api.semanticscholar.org/CorpusID:13930352.

Nikolaos Vakirlis, Zoe Vance, Kate M Duggan, and Aoife McLysaght. De novo birth of functional microproteins in the human lineage. *Cell reports*, 41(12), 2022.

Miriam Linnenbrink, Gwenna Breton, Christine Pfeifle, Pallavi Misra, Julien Y Dutheil, and Diethard Tautz. Experimental evaluation of a direct fitness effect of the de novo evolved mouse gene pldi. *bioRxiv*, pages 2024–01, 2024.

Steffen Klasberg, Tristan Bitard-Feildel, Isabelle Callebaut, and Erich Bornberg-Bauer. Origins and structural properties of novel and de novo protein domains during insect evolution. *The FEBS journal*, 285(14):2605–2625, 2018.

Dan Li, Yang Dong, Yu Jiang, Huifeng Jiang, Jing Cai, and Wen Wang. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell research*, 20(4):408–420, 2010b.

Emily L Rivard, Andrew G Ludwig, Prajal H Patel, Anna Grandchamp, Sarah E Arnold, Alina Berger, Emilie M Scott, Brendan J Kelly, Grace C Mascha, Erich Bornberg-Bauer, et al. A putative de novo evolved gene required for spermatid chromatin condensation in drosophila melanogaster. *PLoS genetics*, 17(9):e1009787, 2021.

David J Begun, Heather A Lindfors, Andrew D Kern, and Corbin D Jones. Evidence for de novo evolution of testis-expressed genes in the drosophila yakuba/drosophila erecta clade. *Genetics*, 176(2):1131–1137, 2007.

Lars A Eicholt, Margaux Aubel, Katrin Berk, Erich Bornberg-Bauer, and Andreas Lange. Heterologous expression of naturally evolved putative de novo proteins with chaperones. *Protein Science*, 31(8):e4371, 2022.

Alastair Grant, David Lee, and Christine Orengo. Progress towards mapping the universe of protein folds. *Genome biology*, 5(5):1–9, 2004.

Michael Levitt. Nature of the protein universe. *Proceedings of the National Academy of Sciences*, 106(27):11079–11084, 2009.

Ágnes Tóth-Petróczy and Dan S. Tawfik. The robustness and innovability of protein folds. *Current opinion in structural biology*, 26:131–8, 2014.

Janani Durairaj, Andrew M Waterhouse, Toomas Mets, Tetiana Brodiazhenko, Minhal Abdullah, Gabriel Studer, Gerardo Tauriello, Mehmet Akdel, Antonina Andreeva, Alex Bateman, et al. Uncovering new families and folds in the natural protein universe. *Nature*, 622(7983):646–653, 2023.

Andrei N Lupas, Chris P Ponting, and Robert B Russell. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of structural biology*, 134(2-3):191–203, 2001.

Klaus O Kopec and Andrei N Lupas. β-propeller blades as ancestral peptides in protein evolution. *PLoS One*, 8(10):e77074, 2013.

Vikram Alva, Michael Remmert, Andreas Biegert, Andrei N Lupas, and Johannes Söding. A galaxy of folds. *Protein Science*, 19(1):124–130, 2010.

Vikram Alva, Johannes Söding, and Andrei N Lupas. A vocabulary of ancient peptides at the origin of folded proteins. *elife*, 4:e09410, 2015.

M Luisa Romero Romero, Avigayel Rabin, and Dan S Tawfik. Functional proteins from short peptides: Dayhoff's hypothesis turns 50. *Angewandte Chemie International Edition*, 55(52):15966–15971, 2016.

Christine A Orengo, Annabel E Todd, and Janet M Thornton. From protein structure to function. *Current opinion in structural biology*, 9(3):374–382, 1999.

Jason Nomburg, Nathan Price, and Jennifer A Doudna. Birth of new protein folds and functions in the virome. *bioRxiv*, 2024. doi: 10.1101/2024.01.22.576744. URL https://www.biorxiv.org/content/early/2024/01/23/2024.01.22.576744.

Antonio Deiana, Sergio Forcelloni, Alessandro Porrello, and Andrea Giansanti. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PLoS One*, 14(8):e0217889, 2019.

Per Jemth, Elin Karlsson, Beat Vögeli, Brenda Guzovsky, Eva Andersson, Greta Hultqvist, Jakob Dogan, Peter Güntert, Roland Riek, and Celestine N Chi. Structure and dynamics conspire in the evolution of affinity between intrinsically disordered proteins. *Science advances*, 4(10): eaau4130, 2018.

Muhammad Ali and Ylva Ivarsson. High-throughput discovery of functional disordered regions. *Molecular Systems Biology*, 14(5):e8377, 2018.

Yi-Hsuan Lin, Jianhui Song, Julie D Forman-Kay, and Hue Sun Chan. Random-phase-approximation theory for sequence-dependent, biologically functional liquid-liquid phase separation of intrinsically disordered proteins. *Journal of Molecular Liquids*, 228:176–193, 2017.

Anthony A Hyman, Christoph A Weber, and Frank Jülicher. Liquid-liquid phase separation in biology. *Annual review of cell and developmental biology*, 30:39–58, 2014.

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, pages 1–4, 2023.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A.

Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

Vladimir N Uversky. Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. *Current opinion in structural biology*, 44: 18–30, 2017.

Anna Hadarovich, Hari Raj Singh, Soumyadeep Ghosh, Nadia Rostam, Anthony A Hyman, and Agnes Toth-Petroczy. Picnic identifies condensate-forming proteins across organisms. *bioRxiv*, pages 2023–06, 2023.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Laura Weidmann, Tjeerd Dijkstra, Oliver Kohlbacher, and Andrei N Lupas. Minor deviations from randomness have huge repercussions on the functional structuring of sequence space. *bioRxiv*, page 706119, 2019.

Luca Agozzino and Ken A Dill. Protein evolution speed depends on its stability and abundance and on chaperone concentrations. *Proceedings of the National Academy of Sciences*, 115(37): 9092–9097, 2018.

Margaux Aubel, Lars Eicholt, and Erich Bornberg-Bauer. Assessing structure and disorder prediction tools for de novo emerged proteins in the age of machine learning. *F1000Research*, 12 (347):347, 2023.

Jing Liu, Rongqing Yuan, Wei Shao, Jitong Wang, Israel Silman, and Joel L Sussman. Do “newly born” orphan proteins resemble “never born” proteins? a study using three deep learning algorithms. *Proteins: Structure, Function, and Bioinformatics*, 2023.

Kristoffer Illergård, David H Ardell, and Arne Elofsson. Structure is three to ten times more conserved than sequence: Study of structural response in protein cores. *Proteins Struct. Funct. Bioinforma.*, 77(3):499–508, 2009.

H. Cheng, R.D. Schaeffer, Y. Liao, L.N. Kinch, J. Pei, S. Shi, and et al. Ecod: An evolutionary classification of protein domains. *PLoS Comput Biol*, 10(12):e1003926, 2014. doi: 10.1371/journal.pcbi.1003926. URL https://doi.org/10.1371/journal.pcbi.1003926.

KV Kishan and Vishal Agrawal. Sh3-like fold proteins are structurally conserved and functionally divergent. *Current Protein and Peptide Science*, 6(2):143–150, 2005.

Claudia Alvarez-Carreño, Petar I Penev, Anton S Petrov, and Loren Dean Williams. Fold evolution before luca: Common ancestry of sh3 domains and ob domains. *Molecular Biology and Evolution*, 38(11):5134–5143, 2021.

James A Rosinski and William R Atchley. Molecular evolution of helix–turn–helix proteins. *Journal of molecular evolution*, 49:301–309, 1999.

Nick V Grishin. Two tricks in one bundle: helix–turn–helix gains enzymatic activity. *Nucleic acids research*, 28(11):2229–2233, 2000.

Caroline M. Weisman. The origins and functions of de novo genes: Against all odds? *Journal of molecular evolution*, 2022.

Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7):e21800, 2011.

Brian Tsang, Iva Pritišanac, Stephen W Scherer, Alan M Moses, and Julie D Forman-Kay. Phase separation as a missing mechanism for interpretation of disease mutations. *Cell*, 183(7):1742–1756, 2020.

Nadia Rostam, Soumyadeep Ghosh, Chi Fung Willis Chow, Anna Hadarovich, Cedric Landerer, Rajat Ghosh, HongKee Moon, Lena Hersemann, Diana M. Mitrea, Isaac A. Klein, Anthony A. Hyman, and Agnes Toth-Petroczy. Cd-code: crowdsourcing condensate database and encyclopedia. *Nature Methods*, Apr 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01831-0. URL https://doi.org/10.1038/s41592-023-01831-0.

Taraneh Zarin, Bob Strome, Gang Peng, Iva Pritišanac, Julie D Forman-Kay, and Alan M Moses. Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *Elife*, 10:e60220, 2021.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

David G Knowles and Aoife McLysaght. Recent de novo origin of human protein-coding genes. *Genome research*, 19(10):1752–1759, 2009.

Diana Ekman and Arne Elofsson. Identifying and quantifying orphan protein sequences in fungi. *Journal of Molecular Biology*, 396(2):396–405, 2010.

Benjamin A. Wilson, Scott G. Foy, Rafik Neme, and Joanna Masel. Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. *Nature ecology & evolution*, 1(6):0146, June 2017.

Nikolaos Vakirlis, Alex S Hebert, Dana A Opulente, Guillaume Achaz, Chris Todd Hittinger, Gilles Fischer, Joshua J Coon, and Ingrid Lafontaine. A molecular portrait of de novo genes in yeasts. *Molecular Biology and Evolution*, 35(3):631–645, 2018.

Ana-Maria Fernandez-Escamilla, Frederic Rousseau, Joost Schymkowitz, and Luis Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology*, 22(10):1302–1306, 2004.

Gábor Erdős, Mátyás Pajkos, and Zsuzsanna Dosztányi. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research*, 49(W1):W297–W303, 05 2021.

Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1):1–17, 2019.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pages 1–8, 2023.

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkare, Koushik Roye, Charlotte Rochereau, Gustaf Ahdritz, Joanna Zhang, George M Church, Peter K Sorger, and Mohammed AlQuraishi. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, pages 1–7, 2022.

Noelia Ferruz and Birte Höcker. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6):521–532, 2022.

Noelia Ferruz, Michael Heinzinger, Mehmet Akdel, Alexander Goncearenco, Luca Naef, and Christian Dallago. From sequence to function through structure: Deep learning for protein design. *Computational and Structural Biotechnology Journal*, 21:238–250, 2023.

I. Barrodale. L 1 approximation and the analysis of data. *Applied Statistics*, 17(1):51, 1968. ISSN 0035-9254. doi: 10.2307/2985267. URL http://dx.doi.org/10.2307/2985267.

Cyrus Chothia. One thousand families for the molecular biologist. *Nature*, 357(6379):543, 1992.

Christian Schlötterer. Genes from scratch–the evolutionary fate of de novo genes. *Trends in Genetics*, 31(4):215–219, 2015.

Sidi Chen, Yong E Zhang, and Manyuan Long. New genes in drosophila quickly become essential. *science*, 330(6011):1682–1685, 2010b.

Michael Lynch. The evolution of multimeric protein assemblages. *Molecular biology and evolution*, 29(5):1353–1366, 2012.

Luca Schulz, Franziska L Sendker, and Georg KA Hochberg. Non-adaptive complexity and biochemical function. *Current Opinion in Structural Biology*, 73:102339, 2022.

Vijay Jayaraman, Saacnicteh Toledo-Patiño, Lianet Noda-García, and Paola Laurino. Mechanisms of protein evolution. *Protein Science*, 31(7):e4362, 2022.

Saurav Malik, Johannes Venezian, Arseniy Lobov, Meta Heidenreich, Hector Garcia-Seisdedos, Todd O Yeates, Ayala Shiber, and Emmanuel D Levy. Structural determinants of co-translational protein complex assembly. *bioRxiv*, pages 2024–01, 2024.

Min-yi Shen, Fred P Davis, and Andrej Sali. The optimal size of a globular protein domain: A simple sphere-packing model. *Chemical Physics Letters*, 405(1-3):224–228, 2005.

DJ Barlow and JM Thornton. Helix geometry in proteins. *Journal of molecular biology*, 201(3): 601–619, 1988.

Jason Greenwald and Roland Riek. On the possible amyloid origin of protein folds. *Journal of molecular biology*, 421(4-5):417–426, 2012.

Darin M Taverna and Richard A Goldstein. The distribution of structures in evolving protein populations. *Biopolymers: Original Research on Biomolecules*, 53(1):1–8, 2000.

Boris E Shakhnovich, Eric Deeds, Charles Delisi, and Eugene Shakhnovich. Protein structure and evolutionary history determine sequence space topology. *Genome research*, 15(3):385–392, 2005.

Richard A Goldstein. The structure of protein evolution and the evolution of protein structure. *Current opinion in structural biology*, 18(2):170–177, 2008.

Alexei V Finkelstein, Alexander M Gutun, and Azat Ya Badretdinov. Why are the same protein folds used to perform different functions? *FEBS letters*, 325(1-2):23–28, 1993.

Sridhar Govindarajan and Richard A Goldstein. Why are some proteins structures so common? *Proceedings of the National Academy of Sciences*, 93(8):3341–3345, 1996.

Michael Y Galperin, D Roland Walker, and Eugene V Koonin. Analogous enzymes: independent inventions in enzyme evolution. *Genome research*, 8(8):779–790, 1998.

Andrew CR Martin, Christine A Orengo, E Gail Hutchinson, Susan Jones, Maria Karmirantzou, Roman A Laskowski, John BO Mitchell, Chiara Taroni, and Janet M Thornton. Protein folds and functions. *Structure*, 6(7):875–884, 1998.

Manas Seal, Orit Weil-Ktorza, Dragana Despotovic, Dan S Tawfik, Yaakov Levy, Norman Metanis, Liam M Longo, and Daniella Goldfarb. Peptide-rna coacervates as a cradle for the evolution of folded domains. *Journal of the American Chemical Society*, 144(31):14150–14160, 2022.

Orit Weil-Ktorza, Yael Fridmann-Sirkis, Dragana Despotovic, Segev Naveh-Tassa, Yaacov Levy, Norman Metanis, and Liam M Longo. Functional ambidexterity of an ancient nucleic acid-binding domain. *bioRxiv*, pages 2023–03, 2023.

827 Pratik Vyas, Olena Trofimyuk, Liam M Longo, Fanindra Kumar Deshmukh, Michal Sharon, and
828 Dan S Tawfik. Helicase-like functions in phosphate loop containing beta-alpha polypeptides.
829 *Proceedings of the National Academy of Sciences*, 118(16):e2016131118, 2021.

830 Liam M Longo, Rachel Kolodny, and Shawn E McGlynn. Evidence for the emergence of β-trefoils
831 by 'peptide budding'from an igg-like β-sandwich. *PLOS Computational Biology*, 18(2):e1009833,
832 2022.

833 Liam M Longo, Dragana Despotović, Orit Weil-Ktorza, Matthew J Walker, Jagoda Jabłońska, Yael
834 Fridmann-Sirkis, Gabriele Varani, Norman Metanis, and Dan S Tawfik. Primordial emergence of
835 a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conver-
836 sion. *Proceedings of the National Academy of Sciences*, 117(27):15731–15739, 2020.

837 Jun-Yan Kang, Ze Wen, Duo Pan, Yuhan Zhang, Qing Li, Ai Zhong, Xinghai Yu, Yi-Chen Wu,
838 Yu Chen, Xiangzheng Zhang, et al. Llps of fxr1 drives spermiogenesis by activating translation
839 of stored mrnas. *Science*, 377(6607):eabj6647, 2022.

840 Martti Parvinen. The chromatoid body in spermatogenesis. *International journal of andrology*, 28
841 (4):189–201, 2005.

842 Mia T Levine, Corbin D Jones, Andrew D Kern, Heather A Lindfors, and David J Begun. Novel
843 genes derived from noncoding dna in drosophila melanogaster are frequently x-linked and exhibit
844 testis-biased expression. 103(26):9935–9939, 2006.

845 Li Zhao, Perot Saelao, Corbin D Jones, and David J Begun. Origin and spread of de novo genes
846 in drosophila melanogaster populations. *Science*, 343(6172):769–772, 2014.

847 Nicola Palmieri, Carolin Kosiol, and Christian Schlötterer. The life cycle of drosophila orphan genes.
848 *elife*, 3:e01311, 2014.

849 Kevin G Nyberg and Richard W Carthew. Out of the testis: biological impacts of new genes. *Genes*
850 *& development*, 31(18):1825–1826, 2017.

851 Shu Kondo, Jeffrey Vedanayagam, Jaaved Mohammed, Sogol Eizadshenass, Lijuan Kan, Nan
852 Pang, Rajaguru Aradhya, Adam Siepel, Josefa Steinhauer, and Eric C Lai. New genes often
853 acquire male-specific functions but rarely become essential in drosophila. *Genes & development*,
854 31(18):1841–1846, 2017.

Rafik Neme and Diethard Tautz. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC genomics*, 14:1–13, 2013.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Gang Hu, Akila Katuwawala, Kui Wang, Zhonghua Wu, Sina Ghadermarzi, Jianzhao Gao, and Lukasz Kurgan. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nature Communications*, 12(1):4438, July 2021. Number: 1 Publisher: Nature Publishing Group.

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. doi: https://doi.org/10.1002/bip.360221211. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.360221211.

S Doerr, MJ Harvey, Frank Noé, and GHTMD De Fabritiis. Htmd: high-throughput molecular dynamics for molecular discovery. *Journal of chemical theory and computation*, 12(4):1845–1852, 2016.

James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713, 2015.

Matt J Harvey, Giovanni Giupponi, and G De Fabritiis. Acemd: accelerating biomolecular dynamics in the microsecond time scale. *Journal of chemical theory and computation*, 5(6):1632–1639, 2009.

Naveen Michaud-Agrawal, Elizabeth J Denning, Thomas B Woolf, and Oliver Beckstein. Mdanalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry*, 32(10):2319–2327, 2011.

Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.

Solve Sæbø, Trygve Almøy, and Inge S Helland. simrel—a versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. *Chemometrics and Intelligent Laboratory Systems*, 146:128–135, 2015.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), mar 2017. doi: 10.21105/joss.00205. URL https://doi.org/10.21105%2Fjoss.00205.

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

Jeff Reback, jbrockmendel, Wes McKinney, Joris Van den Bossche, Matthew Roeschke, Tom Augspurger, Simon Hawkins, Phillip Cloud, gfyoung, Patrick Hoefler, Sinhrks, Adam Klein, Terji Petersen, Jeff Tratner, Chang She, William Ayd, Richard Shadrach, Shahar Naveh, Marc Garcia, JHM Darbyshire, Jeremy Schendel, Torsten Wörtwein, Andy Hayden, Daniel Saxton, Marco Edward Gorelli, Fangchen Li, Matthew Zeitlin, Vytautas Jancauskas, Ali McMaster, and Thomas Li. pandas-dev/pandas: Pandas 1.4.4, August 2022. URL https://doi.org/10.5281/zenodo.7037953.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics.

Bioinformatics, 25(11):1422–1423, 03 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163. URL https://doi.org/10.1093/bioinformatics/btp163.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.