High-throughput Selection of Human *de novo*-emerged sORFs with High Folding Potential

Margaux Aubel [1], Filip Buchel[3,4], Brennen Heames[1], Alun Jones[1], Ondrej Honc[6], Erich Bornberg-Bauer [1,2,*], and Klara Hlouchova [3,5,*]

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

[2]Department of Protein Evolution, Max Planck-Institute for Biology Tuebingen, Tuebingen, Germany

[3]Department of Cell Biology, Faculty of Science, Charles University, Prague, Czech Republic

[4]Department of Biochemistry, Faculty of Science, Charles University, Prague, Czech Republic

[5]Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague, Czech Republic

[6]Imaging Methods Core Facility, BIOCEV, Prague, Czech Republic

*Corresponding authors: E-mail: ebb@uni-muenster.de; klara.hlouchova@natur.cuni.cz

Accepted: March 23, 2024

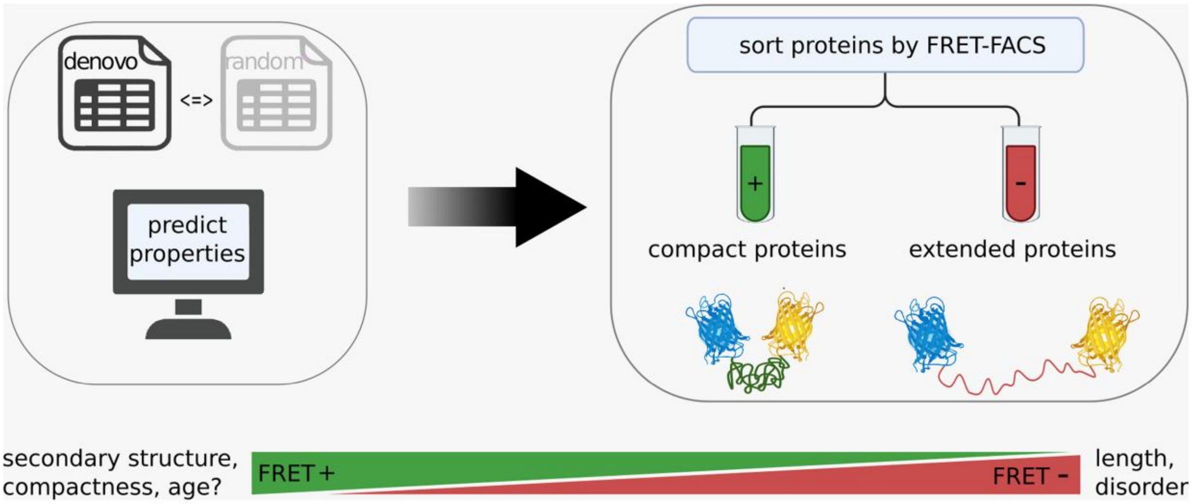
Abstract

De novo genes emerge from previously noncoding stretches of the genome. Their encoded *de novo* proteins are generally expected to be similar to random sequences and, accordingly, with no stable tertiary fold and high predicted disorder. However, structural properties of *de novo* proteins and whether they differ during the stages of emergence and fixation have not been studied in depth and rely heavily on predictions. Here we generated a library of short human putative *de novo* proteins of varying lengths and ages and sorted the candidates according to their structural compactness and disorder propensity. Using Förster resonance energy transfer combined with Fluorescence-activated cell sorting, we were able to screen the library for most compact protein structures, as well as most elongated and flexible structures. We find that compact *de novo* proteins are on average slightly shorter and contain lower predicted disorder than less compact ones. The predicted structures for most and least compact *de novo* proteins correspond to expectations in that they contain more secondary structure content or higher disorder content, respectively. Our experiments indicate that older *de novo* proteins have higher compactness and structural propensity compared with young ones. We discuss possible evolutionary scenarios and their implications underlying the age-dependencies of compactness and structural content of putative *de novo* proteins.

Significance

Akin to random sequences, proteins that emerge *de novo* from noncoding DNA are usually predicted to be highly disordered as they inherit no structure from their ancestor, unlike proteins evolving through duplication. Most studies to-date have relied on computational prediction of the structural properties of *de novo* proteins. Here, we show experimentally that, while most of our putative human *de novo* proteins are highly disordered, some contain propensity for structure and globularity, seen most clearly in the oldest *de novo* proteins in our dataset. With that we aim to lay the groundwork for experimental verification of hypotheses regarding the structural evolution of *de novo* proteins.

Graphical Abstract



Introduction

De novo protein emergence provides the genome with great innovative potential to explore the hitherto unexplored sequence space (Tautz and Domazet-Lošo 2011; McLysaght and Hurst 2016; Weisman and Eddy 2017; Rödelsperger et al. 2019; Van Oss and Carvunis 2019). In comparison to other known mechanisms of protein emergence that rely on recycling of conserved genetic elements, like duplication (Ohno 1970; Sikosek and Bornberg-Bauer 2010) or gene fusion (Dohmen et al. 2020), *de novo* genes emerge from noncoding regions of the genome (Bornberg-Bauer et al. 2021). *De novo* proteins are shorter on average than new proteins emerging through duplication (Montañés et al. 2023) and have low expression that is often restricted to specific tissues or conditions (Wu and Knudson 2018; Heames et al. 2020; Schmitz et al. 2020). Therefore, many potential *de novo* proteins are overlooked by classical annotation methods focusing on longer and highly expressed proteins. Due to the same challenges, short open reading frames (sORFs) and their microproteins remain understudied. Recently, sORFs and microproteins have gained attention (Pueyo et al. 2016) and are proposed to serve as a reservoir for *de novo* proteins (Vakirlis et al. 2022; Sandmann et al. 2023). Additional *de novo* proteins are continuously detected across many species, including, e.g. several plants (Zhang et al. 2019; Marsch-Martínez et al. 2022), fruit flies (Heames et al. 2020), and humans (Guerzoni and McLysaght 2016; Sandmann et al. 2023). Many of these studies focus on detection and functional characterization of the *de novo* proteins, but few report structural characterization (Bungard et al. 2017; Her et al. 2019; Lange et al. 2020; Matsuo et al. 2021). Computationally, *de novo* proteins are mostly predicted to contain high structural disorder (Wilson et al. 2017; Dowling et al. 2020; Aubel et al. 2023; Peng and Zhao 2024). They may assume molten globule-like structures containing secondary structure elements but lacking the stable tertiary fold of a globular protein (Bungard et al. 2017; Lange et al. 2020). Similarly, random-sequence proteins have been shown to contain secondary structure elements, but are best tolerated *in vivo* when they have a higher amount of disordered regions (Tretyachenko et al. 2017).

In a previous study comparing *de novo* protein candidates to random-sequence proteins (Heames et al. 2023), we showed experimentally that both sets of proteins are on average highly similar to each other concerning their solubility, interaction with chaperones and protease resistance. However, the putative *de novo* proteins showed slightly higher solubility, yet at the same time higher degradability when exposed to a bacterial Lon protease (Niwa et al. 2019) than their random-sequence counterparts. Their higher solubility combined with more degradation by the protease points to overall higher disorder content of *de novo* proteins compared with the random-sequence proteins. Corresponding to the experimental findings on both *de novo* and random proteins, solubility and prevention against aggregation seem to be the main bottleneck for newly emerging proteins to avoid purging by natural selection (Ángyán et al. 2012; Agozzino and Dill 2018; Monti et al. 2021; Vakirlis et al. 2022).

Here, we aim to select candidate *de novo* proteins originating from sORFs with high potential for compactness, and accordingly with a lower amount of disorder and increasing potential for folding. Our goal is to investigate whether and how frequently compact sORFs have the propensity to form secondary structure elements and potentially stable folds. We apply a high-throughput assay based on fluorescent-activated cell sorting (FACS) of *Escherichia coli* cells to select top candidates from a library of 3,750 putative *de novo* proteins (Fig. 1a–b). The putative *de novo* sORFs in the library are of differing lengths between 32 and 59 amino acids (aa). Compared to the amino acid frequencies reported in UniProt (EMBL-EBI 2024), the library sequences contain several more disorder promoting amino acids (Uversky 2013), like proline and serine, but also slightly more cysteine and tryptophan (supplementary material figure S1a–b, Supplementary Material online). All putative *de novo* sORF proteins show evidence of translation in different ribosome sequencing (ribo-seq) experiments (Olexiouk et al. 2016). The design of the assay, based on previous work by Philipps et al. (2003), makes use of Förster resonance energy transfer (FRET) between two fluorescent proteins with spectral overlap (Förster 1948). Efficiency of the transfer is inversely dependent on the distance of the fluorescent proteins, thereby offering a way to measure proximity of two molecules or to study intramolecular conformation states. By making use of the latter, we aimed to develop a high-throughput assay capable of screening for compact protein variants. In this assay, the library target protein is expressed in fusion with the FRET pair, serving as a linker between the fluorophores. The yellow fluorescent protein (mVenus) serving as an acceptor is fused to the C-terminus of the target protein, while the donor cyan fluorescent protein (mTurquoise2) is at the N-terminus of the target protein (Fig. 1c). In case the two fluorescent proteins are connected by a stable and compact target protein, which places them within the FRET radius, the energy transfer can happen. Increased FRET signal reflects N- to C-terminal distance or persistence length of the linker proteins and can be interpreted as a measure of compactness (Krishna and Englander 2005; van Rosmalen et al. 2017). Target proteins that are more flexible and do not provide a structurally stable link between the two fluorescent proteins, result in *E. coli* cells without FRET signal. The cells expressing the target protein with fluorescent proteins can be sorted accordingly using FACS. Thereby, we can select single protein sequences that have a low N- to C-terminal distance, are more compact and less disordered (Fig. 1d).

Results

The library of 3,750 putative *de novo* sORF proteins (DN) and corresponding random- sequence proteins (R) ranging from 32 to 59 amino acids were amplified from an oligonucleotide library synthesis (OLS) pool. DN and R libraries were separated in PCR by using library-specific primers and then sequenced using Illumina NGS. The PCR product of both libraries covered 90% (3365 DN and 3,367 R) of the initial libraries (supplementary material figure S2, Supplementary Material online). Library sequences of DN and R were then cloned into pETMF plasmid, placing the single library sORF proteins as a linker between two fluorescent proteins. Upon expression in *E. coli* cells the library proteins are tagged with mTurquoise2 at the N-terminus and mVenus at the C-terminus resulting in a fusion protein (Fig. 1c). More compact linker proteins lead to a stronger FRET signal between the two fluorescent tags and can be sorted using FACS on the *E. coli* expressing the fusion protein. Each library was sorted into FRET-positive and FRET-negative categories based on controls (GS-linkers and previously characterized proteins of the same length, see Table 1) in two sequential rounds starting from the presorted samples. After completing all FACS rounds, cells from the different samples were recovered separately, barcoded, and their DNA sequenced using NGS. This Figure was created with BioRender.

Fluorescence Lifetime Measurements

We chose mTurquoise2 and mVenus as a FRET pair, as both of these fluorescent proteins express well in *E.coli*, possess self-association preventing mutations, exhibit high brightness and, in the case of mTurquoise2, long fluorescence lifetime (Bajar et al. 2016). First, we generated a cassette consisting of the FRET pair genes separated by a Golden Gate assembly compatible cloning site and inserted the cassette into the backbone of the pET24a(+) vector. The resulting vector pETMF allows for T7 promoter controlled expression of the large fusion protein, where the protein of interest is connected to the FRET pair by one glycine–serine repeat and glutamate and leucine residues as part of the cloning process. To validate the performance of our compactness assay, we selected several well-characterized *E.coli* proteins with different properties and generated constructs with the FRET pair separated by GGS repeat linkers of various lengths (Table 1). Intensity based FRET efficiency measurements require purified samples of known concentrations. Therefore, we used time correlated single photon counting (TCSPC) to obtain a fluorescence lifetime of the donor molecule, which can be used to calculate FRET efficiency *in vivo*. Fluorescence lifetime is a property inherent to every fluorescent molecule, and it defines the time it takes the molecule to return to its ground state upon excitation. In the presence of a FRET acceptor, the lifetime of the donor shortens and by comparing it to the lifetime of the donor alone the FRET efficiency can be calculated. After exponential curve fitting, we deconvoluted the fluorescence lifetime of the donor alone and fusion protein constructs. The fluorescence lifetime of the donor, 3.79 ns, slightly deviates from lifetime of 4.0 ns reported in a

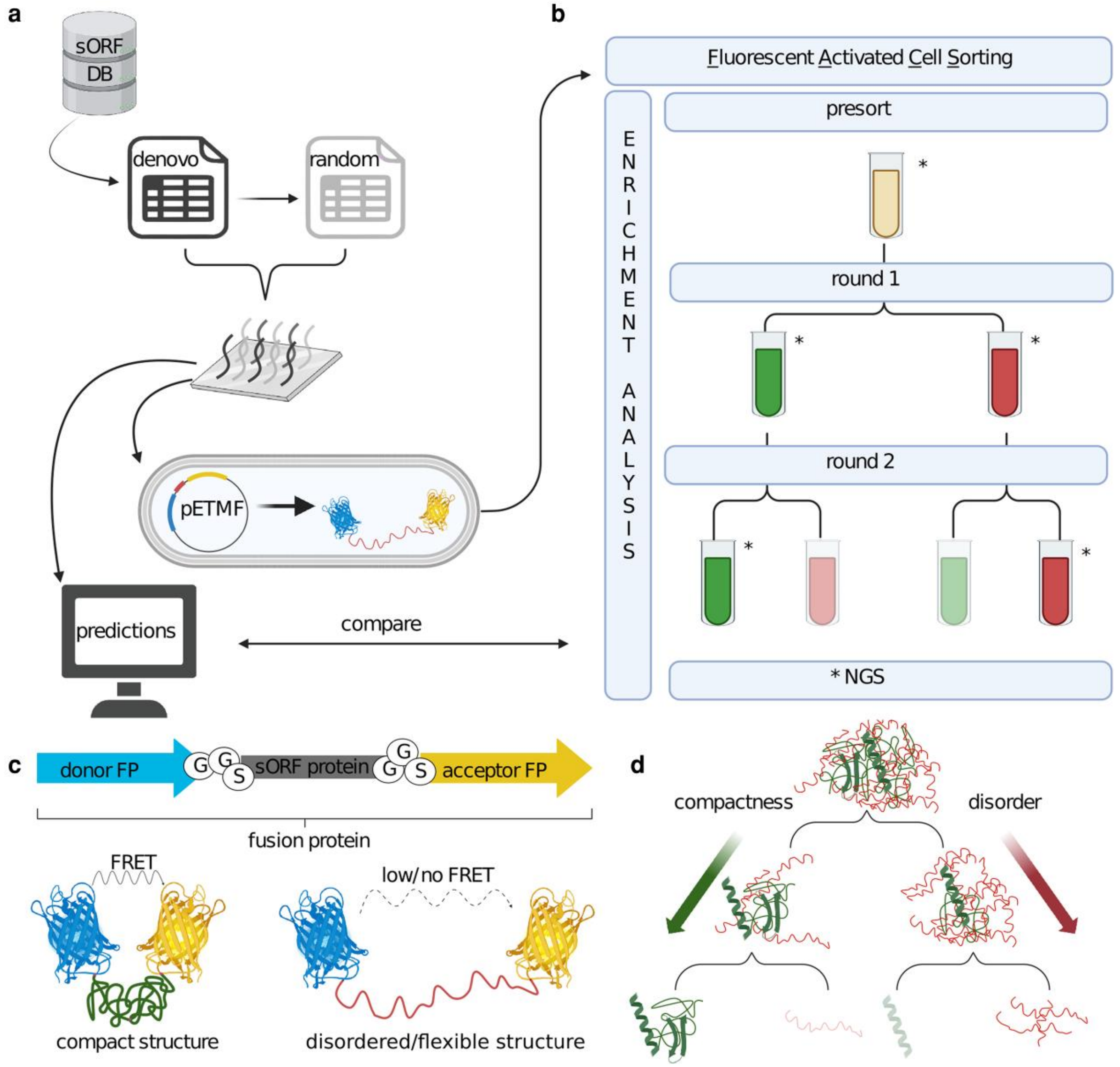


Fig. 1.—Workflow: a) We selected 3,750 *de novo* emerged human sequences from the sORFs database and generated a library of comparable random sequences. Protein structure properties were predicted computationally. After ordering the libraries as oligonucleotides they were cloned into the pETMF plasmid and transformed into *E. coli* for the FRET-FACS assay. b) The FACS was performed in two sequential rounds for both libraries separately. Presorted cells containing single library sequences were sorted into FRET-positive (on the left in green) and FRET-negative (on the right in red) samples. Samples with * were recovered, sent for next-generation sequencing (NGS) and used for enrichment analysis. c) The library sORF proteins (gray) are tagged with fluorescent proteins (FP) mVenus (acceptor FP in yellow) and mTurquoise2 (donor FP in blue) on the termini with GGS spacers. Compact library sORF proteins place the fluorescent proteins in close proximity and are expected to cause FRET. Disordered or fibrillar library proteins are expected to be FRET-negative. d) The presorted samples contain all library protein structures. The first round is expected to result in separation of compact and disordered structures, which becomes more pronounced in the second round. The FRET-negative proteins should have an increased N- to C-terminal distance and disorder, while the FRET-positive proteins show increased compactness and folding. Made with BioRender.

previous study (Goedhart et al. 2012), which can be attributed to fluorescence quenching of the cellular environment in bacteria. Nevertheless, we observed an increase in lifetime with increasing length of the linker in case of GS-linker controls. The various measured lifetimes for selected *E.coli* proteins corresponded to their overall flexibility and structure content, albeit invariant of the length (Table 1).

Flow Cytometry

Following the fluorescence lifetime measurements, we continued with flow cytometry experiments to verify that

Table 1

Fluorescence Lifetime of Different Control Proteins Measured *in vivo*

Protein name	Length (aa)	Lifetime measured (ns)	FRET efficiency	Properties
mTurquoise2	N/A	3.79	N/A	Donor
GS2	12	2.51	0.338	Shortest linker
GS3	18	2.85	0.248	Short linker
GS4	24	2.9	0.235	Medium length linker
1259	100	3.36	0.113	Secondary structure rich, insoluble
665	100	2.64	0.303	Moderate structure content, soluble
BolA	107	2.56	0.325	Globular, terminal disorder
IscA	107	2.76	0.272	Homodimeric
SodA	206	2.77	0.269	Homodimeric
CspA	69	2.8	0.261	β-barrel
FtsB	103	3.41	0.1	Coiled coil
YacG	68	3.56	0.061	Disordered

GS2/3/4, controls with corresponding number of GGSGGS repeats; 1259 and 665, characterized random proteins (Tretyachenko et al. 2017); BolA, transcriptional regulator; IscA, iron binding protein, forms homodimers around Fe atoms; SodA, superoxide dismutase; CspA, cold shock DNA binding protein; FtsB, transmembrane protein of bacterial divisome; YacG, DNA gyrase inhibitor. FRET efficiency is calculated as: $E = 1 - \frac{\tau_{DA}}{\tau_D}$, where τ_{DA} is the lifetime of the donor in presence of the acceptor and τ_D is the lifetime of the donor alone.

spectroscopic results correspond to the cytometric ones. The nature of our assay, being an intramolecular FRET sensor, does not require the tedious control process as in the case of detecting protein–protein interactions by FRET (Banning et al. 2010). Nevertheless, we employed three channels: donor excitation and emission (donor channel), acceptor excitation and emission (acceptor channel), donor excitation and acceptor emission (FRET channel). First, we used donor and acceptor channels to gate a population positive for both fluorophores (P1), suggesting presence of a complete fusion protein. To robustly detect the FRET signal, we derived a parameter of ratio between the FRET channel and the donor channel (FRET ratio). We observed that the FRET ratio of the control proteins were in line with the FRET efficiencies obtained from fluorescence lifetime measurements (supplementary material figure S5, Supplementary Material online). We then proceeded to sort the DN and R library samples. First, we gated approximately 30% of the events exhibiting maximum fluorescence for the donor and the acceptor channel (P1 in Fig. 2a). Apart from the main double fluorescent population, a smaller population showing only donor signal appeared, which can be attributed to spurious stop codons or variants difficult to translate. After projecting the P1 population to the FRET ratio histogram, we started with excluding the top 3% of the population, as the cells showing extreme fluorescence have lower viability. FRET-negative and FRET-positive gates were set by gating the bottom and the top 10% of the distribution, with 30,000 events sorted for each of these gates (Fig. 2a, left and middle). Sorted cells were recovered on LB-agar plates and new expression cultures prepared (see Methods). In subsequent round of FACS, the FRET-positive population was only sorted for high FRET signal and FRET-negative population for low FRET signal, to further strengthen the selection (Fig. 1). Overlaid FRET ratio projections of P1 populations from two rounds of sorting (Fig. 2a, right) showed an incremental shift of the FRET ratio signal away from the naive population (DN all). This trend was most prominent in the FRET-negative samples, however in the second round of FRET-positive sorting the FRET ratio drops with the median even below the naive population. One possible explanation for this phenomenon might be more structural variety of the selected variants, resulting in larger spread of the signal. The predictions of all sorted sequences from round one to two follow the expected trend with decreasing disorder and increasing compactness and secondary structure content (supplementary material figure S4, Supplementary Material online).

Sorting of DN and R Libraries

Over two thirds of the presorted sORF proteins from the DN and R libraries were present at least once in the FRET-positive and FRET-negative samples of round one, indicating that a second round of sorting is needed to clearly separate the FRET-positive from FRET-negative proteins (Fig. 2b,c, left and supplementary material figure S3–S4, Supplementary Material online). After the second round of sorting, the number of sORF proteins decreased to around half of the presorted sORF proteins in FRET-positive and FRET-negative samples for the DN library. In contrast, two thirds of the presorted sORF proteins of the R library are still present in both samples after the second sorting. There is no significant difference in length between FRET-positive and FRET-negative sorted sORF proteins after the first round of sorting for both DN and R libraries. After the second sorting both libraries contain significantly shorter sequences on average in the FRET-positive samples than the FRET-negative samples. However, the whole length range of library sequences is covered in both FRET-positive and FRET-negative samples.

Enrichment Analysis

We calculated enrichment of single sequences in FRET-positive and FRET-negative samples between the rounds to determine the sORF proteins that are overrepresented in the different FRET categories. Overall, more sequences are significantly enriched (*P*-value < 0.05) in the FRET-negative sample compared to the FRET-positive for both DN (Fig. 2b) and R (Fig. 2c). A higher proportion of the library is enriched in the FRET-negative sorting in the DN library compared to R (0.12 and 0.08, respectively, after round two of sorting).

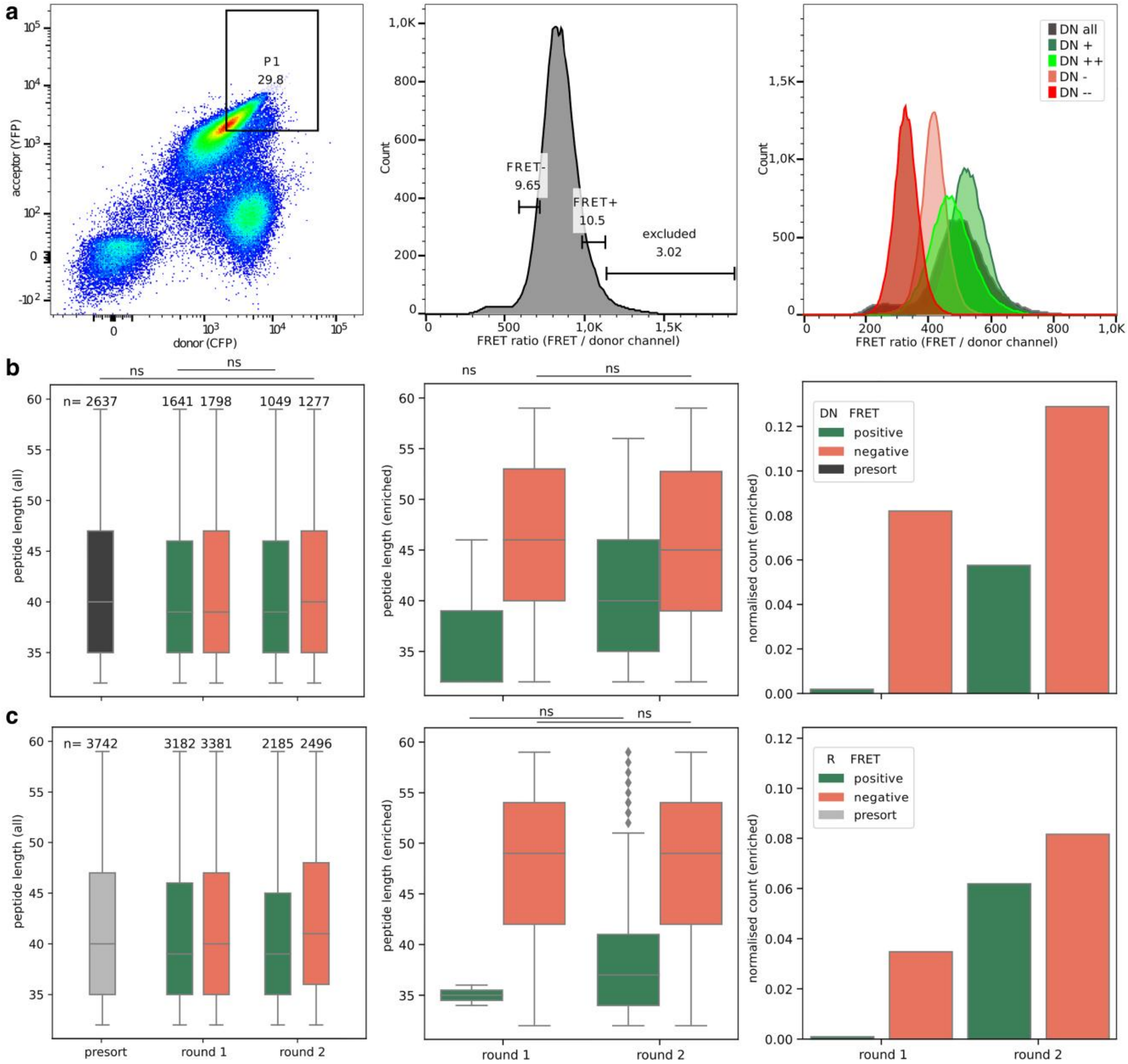


Fig. 2.—a) On the left, exemplary gating of the double fluorescence positive population P1 is shown. The central histogram shows the FRET ratio of the P1 population with arbitrary gate 'excluded' (see Results) and FRET-negative/FRET-positive gates. The right plot shows the FRET ratio of overlaid P1 populations after two rounds of sorting, recorded after the sorting experiment on LSRFortessa cytometer. b, c) Peptide lengths of all sequences of libraries DN (b) and R (c) in FRET-positive and FRET-negative samples across rounds including total number of unique sequences on top of the bars (left), peptide lengths of significantly enriched sequences (centre) and number of enriched sequences normalized by total sequences present for each library (right). The lengths are significantly different (*P*-value < 0.05) according to Tukey HSD test between all categories if not indicated otherwise (ns).

The proportion of enriched sequences increases substantially from <0.01 in the first round of sorting to 0.06 in the second round for both DN and R. Comparisons between FRET conditions of enriched proteins in round one are difficult because of very few significantly enriched sequences in the FRET-positive samples. From round one to two, the average length of the FRET-positive enriched proteins increases while the predicted disorder decreases. This increasing length could be causing the drop in FRET ratio observed for DN after the second round of sorting (see Fig. 2a). The FRET-negative samples do not change in length, but predicted disorder increases from round one to two for both libraries (significant for R). Four sORF proteins of the DN library are enriched in the FRET-positive samples as well as in the FRET-negative samples, and are discarded as false positives when choosing most compact candidates with folding potential.

For a comparison of protein properties between FRET-positive and FRET-negative samples, we took only the proteins that are significantly enriched from presort to round two. This way we use the highest number of sORF proteins and the most confidently enriched ones (see Fig. 2). As the predictions for random-sequence proteins are less reliable (Middendorf and Eicholt 2024) which could be explained partially by significantly higher sequence complexity (supplementary material figure S1c, Supplementary Material online), from hereon only the DN library is regarded for analysis (for library R see supplementary material figures S4, S9, S10, supplementary Material online). The protein sequences enriched in the FRET-negative samples are significantly longer on average than sequences that are enriched in the FRET-positive samples. To predict structural properties, we used ESMFold (Lin et al. 2023), which has been proposed to be more reliable on sequences without homology (Elofsson 2023) like *de novo* proteins, as well as random-sequence proteins (Liu et al. 2023). Taking the ESMFold predictions, we calculated the percentage of secondary structure (helix, sheet, coil, PP-II helix, turns), radius of gyration, average solvent accessibility per amino acid (asa) and N- to C-terminal distance. As a comparison we also predicted all structures with AlphaFold2 and extracted structural properties. All predicted protein properties, except turns, correlate significantly between ESMFold and AlphaFold2 for the random-sequence proteins. In contrast, the predicted protein properties correlate much less for the *de novo* sequences with lower Pearson's R and nonsignificant *P*-values (see supplementary material table S2, Supplementary Material online). The ESMFold predictions correspond more to our experimental findings, especially looking at the *de novo* sequences (supplementary material figure S6, Supplementary Material online). Sequences enriched in FRET-positive are predicted to have significantly lower radius of gyration, lower N- to C-terminal distance, lower average solvent accessibility, higher amount of secondary structure elements, and higher confidence of ESMFold predictions seen by the higher pLDDT (Fig. 3 and supplementary material figure S4, Supplementary Material online).

Predictive Models

We used predictive modeling to gain further insights into which predicted protein properties had an impact on whether *de novo* sORF proteins are sorted into the FRET-positive or FRET-negative samples. For this purpose, we split all enriched sequences into test and training data sets and used elastic net type regression (Tay et al. 2023) of the log2 fold change (LFC) from presorted samples to round two of sorting plotted against different protein properties (Fig. 4 and supplementary material figures S6–S9, Supplementary Material online). The protein properties with the biggest influence on FRET-positive sorting are predicted turns in the structure and disorder, while length of the protein sequence has only a very low impact (Fig. 4). The LFC is positively correlated with predicted turns after secondary structure elements, predicted beta-sheets and age of transcription but negatively correlated with length, predicted disorder and coils, radius of gyration and N- to C-terminal distance (supplementary material figure S7, Supplementary Material online). The length of the sORF proteins had a bigger impact (correlation 4x higher) on whether a protein is enriched in the FRET-negative samples compared to FRET-positive. Predicted disorder and age categories (transcription or BLAST) have the reverse effect on FRET-negative sorting compared to FRET-positive sorting, but of the same magnitude (−2.2 vs 2.6 and 0.06 vs −0.06).

Ages of *de novo* sORFs

Protein folding and compactness is usually considered to be a trait that needs long evolutionary timescales to develop, given that it is highly unlikely for a random sequence. Young *de novo* proteins and random-sequence proteins are therefore predicted to be highly disordered, lacking secondary structure elements and a stable tertiary fold. A major unresolved issue in understanding the evolution of *de novo* proteins is, whether older *de novo* proteins are more structured than younger ones. Our DN library contains putative *de novo* emerged sORFs of different ages. We grouped sequences into age categories 1–6 (supplementary material figure S11, Supplementary Material online, covering an evolutionary age range from 6 to 90 Million years (My). The age categories are based on: i) BLAST age, i.e. significant sequence similarity in outgroup species (chimpanzee, gorilla, orangutan, macaque, and mouse) or ii) RNA age, i.e. detectable transcription in human, macaque and mouse. There are no sequences of blast age 1 (human only), because we excluded sequences without any sequence similarity in related species in the library design as the mechanism of emergence could not be determined. The *de novo* sORFs specific to monkeys (ages 2–5) show similar behavior in the assay, with around 70% of sequences in the FRET-negative enriched, i.e. disordered proteins and 30% in the FRET-positive enriched, i.e. more compact sORF proteins (Fig. 5). The oldest age group with homologous sequences detectable up to mouse (age 6) has a higher percentage of FRET-positive enriched sequences (45%) than the younger sORFs. For the age groups based on transcription, we see a similar trend. The older *de novo* sORFs that are transcribed in macaque (age 5) or mouse (age 6) are more frequently represented in the FRET-positive enriched sequences than the younger ones which are transcribed in human only (age 1). The level of translation as reported in the sORFs database, according to different ribo-seq experiments, varies between age groups. The youngest age group shows the lowest translation for least disordered proteins of all age groups and steepest increase of translation with increasing disorder

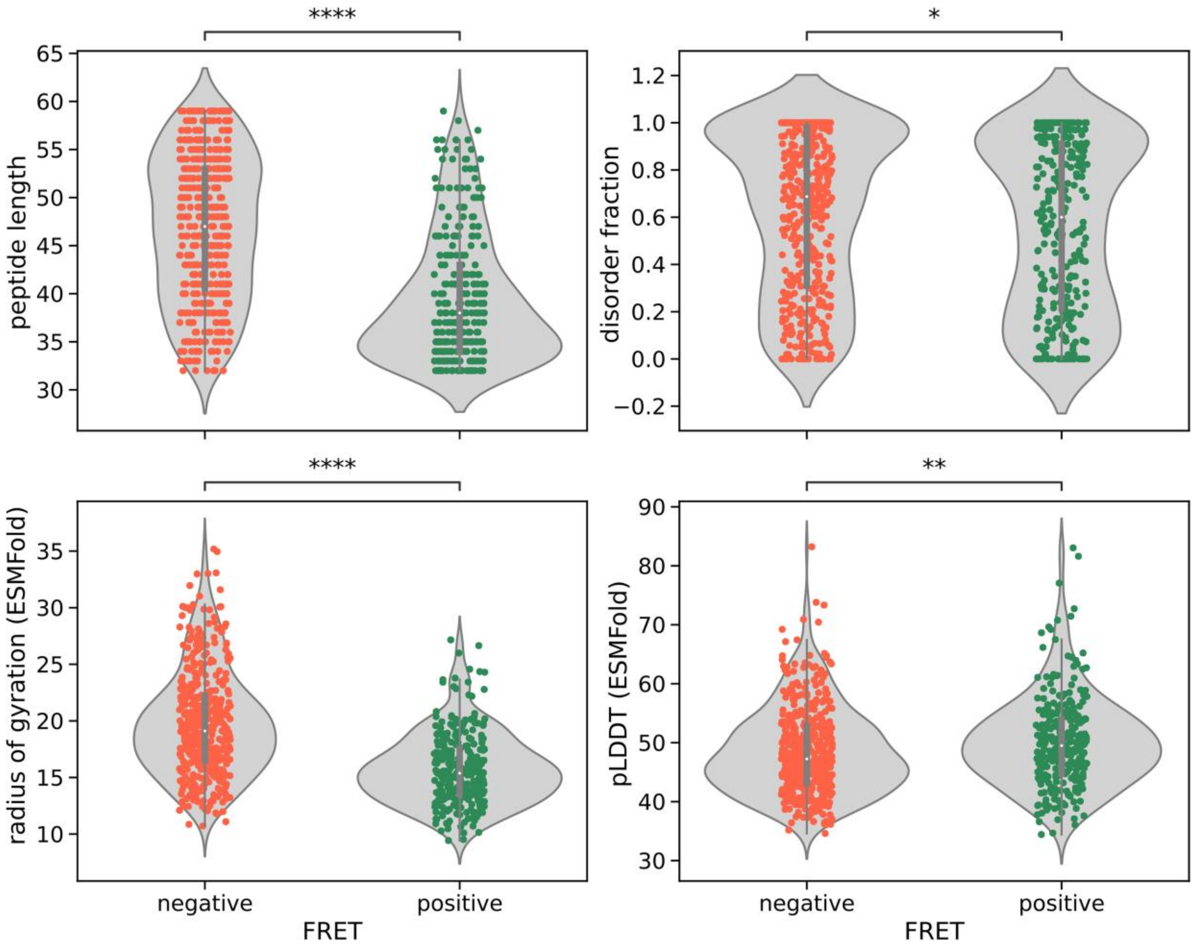


Fig. 3.—Protein property predictions (*y*-axis) of significantly enriched sequences in round two of sorting for FRET-positive (on the left colored in green) and FRET-negative (on the right colored in red) subsets of the DN library. Stars indicate significance calculated with t-test (*P*-value $<0.05^* < 0.01^{**} < 0.001^{***} < 0.0001^{****}$).

content. The oldest age group shows the most constant translation across all levels of disorder (supplementary material figure S1d, Supplementary Material online). The predictive models described above also predict a positive correlation between LFC and age of transcription for the sORF proteins enriched in the FRET-positive samples (Fig. 4). The observed trend deserves further verification using larger data sets and other species because only few enriched sequences ($n = 12$ and $n = 18$) belong to the oldest age group 6 and differences between age groups are not statistically significant (*P*-value >0.05, chi-square test).

Top Enriched *de novo* sORFs

The aim of this study was selection of putative *de novo* sORF proteins with the potential ability to fold into more compact structures. To verify the selection of appropriate candidates from the library of 3,750 sORF sequences, we used the ESMFold predictions of sequences with highest enrichment in the FRET-positive and FRET-negative samples after two rounds of sorting (Fig. 6). As an additional filter for FRET-positive sORF proteins, we took the candidates enriched in the first round as well as the second round, considering the drop in FRET ratio in the second round of sorting (Fig. 2a–c). The top enriched FRET-positive structures (a–d) contain a comparably high amount of predicted α-helices and display a relatively compact structure. The radius of gyration for the top structures is below the average, as calculated for the DN library based on ESMFold predictions. The top FRET-negative enriched sORF proteins (e-g) are predicted to consist of mainly disordered regions, with only one out of four structures having a small α-helix predicted. The confidence score (pLDDT) of predicted structures for the top enriched sORFs is low, within the range of the average for all putative *de novo* sORF library proteins. Although we observe the trend for more secondary structure and less disorder for FRET-positive sequences, a few FRET-positive sORF proteins still show high amounts of predicted disorder and low secondary structure content underlining the importance for additional filtering in future studies.

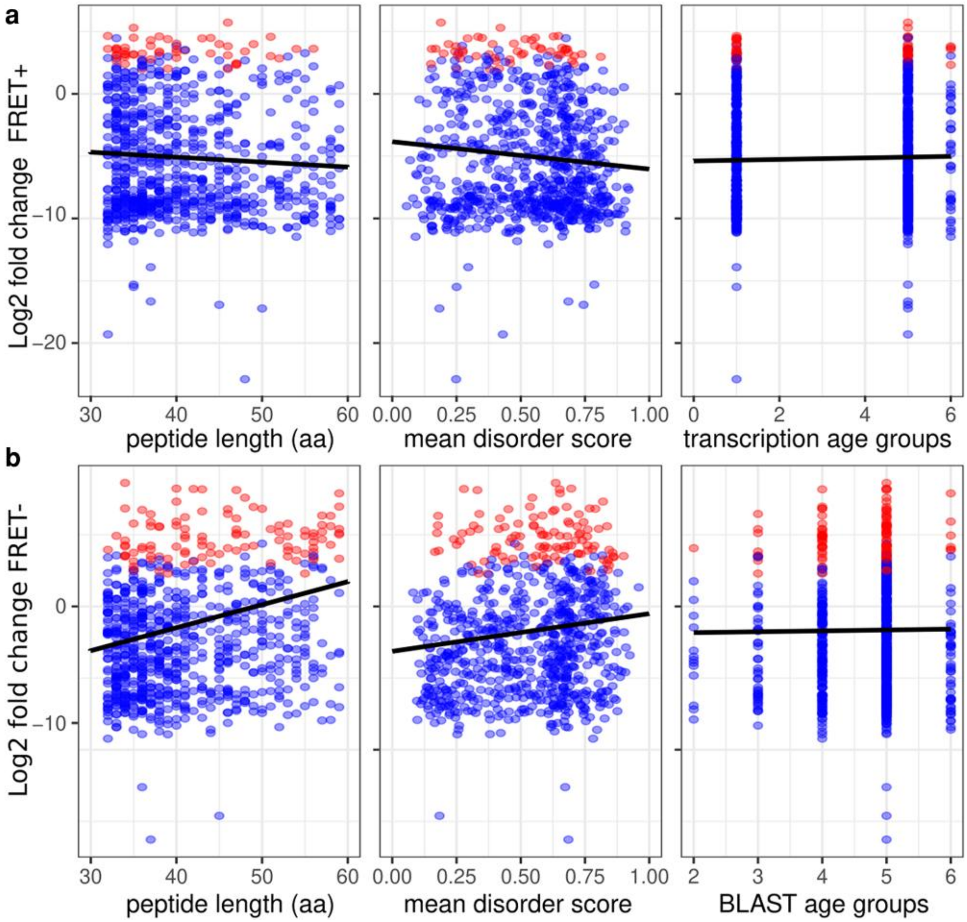


Fig. 4.—Predictive plots for FRET-positive FRET+) a) and FRET-negative FRET-) b) *de novo* sORF proteins enriched from presort to round two. The points are the test data set with significantly enriched sequences in red. The line is the coefficient of the particular predictor with all other variables either set to zero or to the median value. Note that in an an elastic net-type regression as used here, the uncertainty is not calculable in the same way as a normal regression, so no confidence intervals are added.

While the top enriched sORF proteins correspond to the overall expected predictions, they do not represent the best candidates according to the predicted protein properties, i.e. with the lowest radius of gyration and highest amount of secondary structure. All top enriched proteins, FRET-positive and FRET-negative, are indistinguishable from other enriched proteins according to predictions. However, looking at the predicted disorder in relation to length (Fig. 7), the FRET-positive enriched sORF proteins are scattered mostly at the lower edge of the plot with decreasing disorder for longer sORFs. The FRET-negative enriched sORF proteins do not display a trend in the distribution.

Discussion

In this study, we screened putative *de novo* sORF proteins of different ages for candidate sequences with high compactness and with that, the capability for folding. The dataset of *de novo* sORFs does not overlap with other reported and better characterized human *de novo* proteins (Broeils et al. 2023) but merely consists of putative *de novo* proteins. The putative *de novo* sORF proteins used in this study have the potential of becoming *de novo* proteins, and some of them might already have gene-like properties. Indeed, the underlying *de novo* sORFs taken from the sORFs database all show evidence for translation based on different ribo-seq experiments (Olexiouk et al. 2016). While the younger *de novo* sORF proteins show increased evidence of translation with higher disorder content, the oldest sORFs show no increased translation with disorder and also show less variation overall in translation levels. The candidate sORFs are of varying lengths (32–59 aa) to screen as many putative *de novo* sequences as possible (see supplementary material figure S1a, Supplementary Material online) within the maximum length feasible by oligonucleotide library synthesis (OLS). For comparison, we took the same number of randomly generated sequences with the same lengths and amino acid frequencies as the *de novo* library sequences. Because the predictions for random-sequence proteins have been previously shown

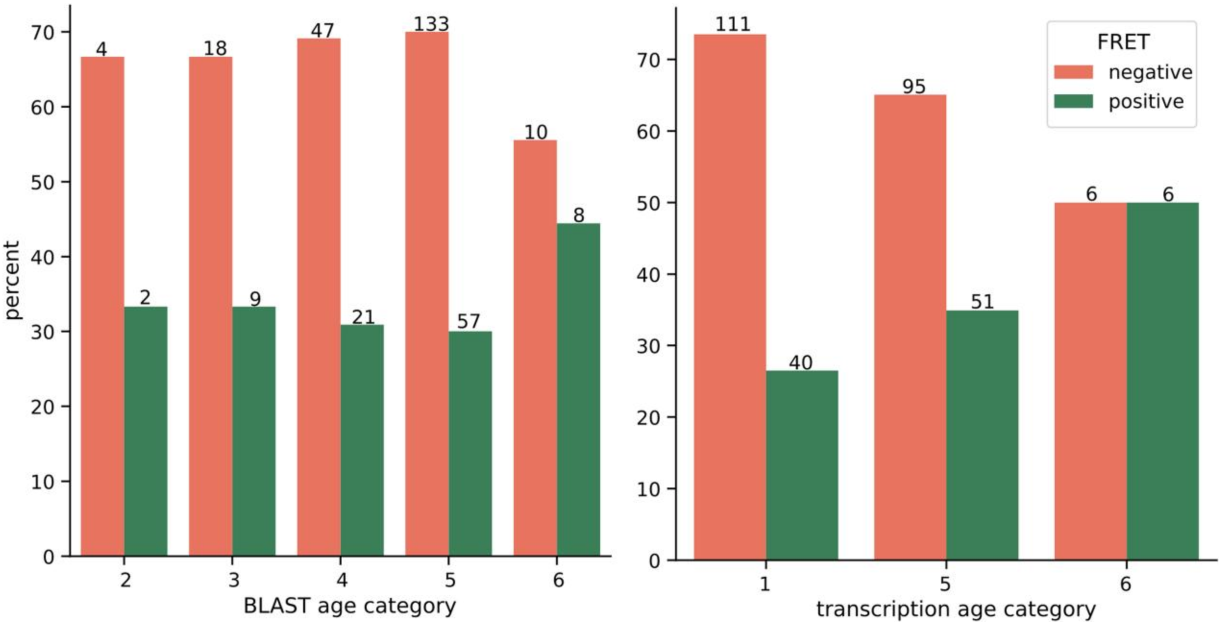


Fig. 5.—Percentage of all significantly enriched sequences in the DN library separated by FRET-positive and FRET-negative enrichment sorted by BLAST age categories (left) and transcription age categories (right).

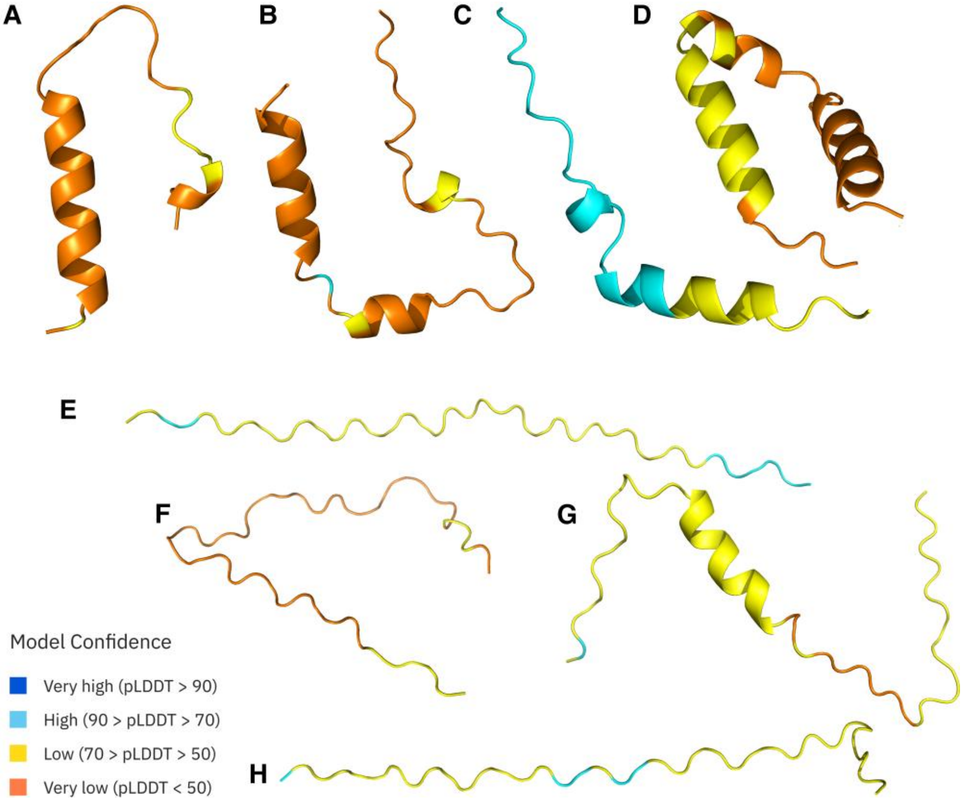


Fig. 6.—Examples for predicted *de novo* sORF protein structures of the four most confidently enriched FRET-positive a–d) and FRET-negative e–h) sequences colored by model confidence.

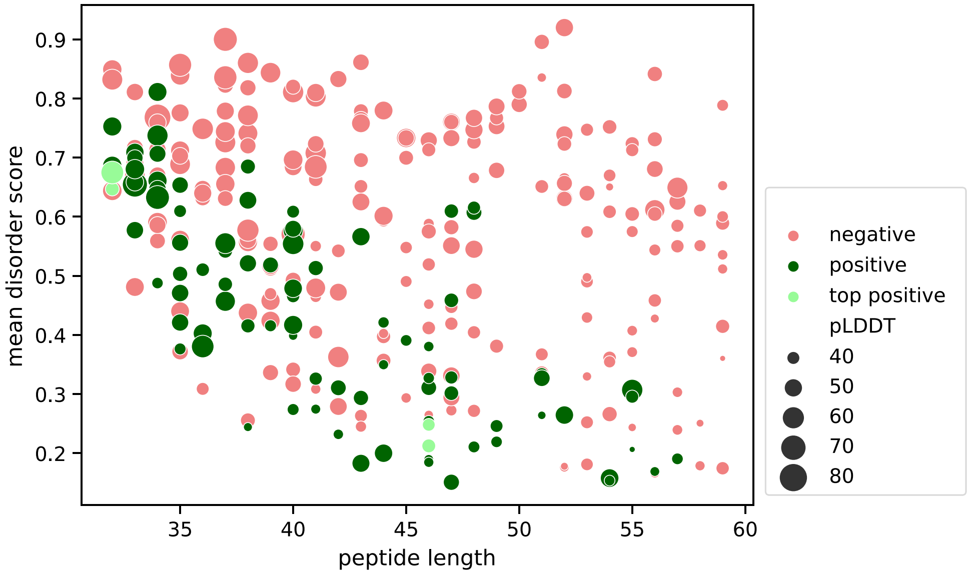


Fig. 7.—Scatterplot of significantly enriched sORF proteins looking at predicted mean disorder and length. Each dot represents a single sORF protein, the size of the dots corresponds to the confidence measure pLDDT of ESMFold predictions.

to be less reliable than predictions for natural protein sequences (Liu et al. 2023; Middendorf and Eicholt 2024) and do not correspond to the experimental results as well as the putative *de novo* proteins do, our analyzes mainly focus on the putative *de novo* sORFs. One possible explanation for their different behavior could be the higher sequence complexity of the random sequences compared to the *de novo* sequences (supplementary material figure S1c, Supplementary Material online).

Using an assay that combines FRET between the N-terminal mVenus (yellow fluorescent protein) and the C-terminal mTurquoise2 (cyan fluorescent protein) of the target protein with FACS allows for high-throughput screening of thousands of candidate sequences. We expected the FRET-positive sORF proteins to have a low N- to C-terminal distance, resulting in a higher probability for compactness and folding. The FRET-negative sORF proteins, on the other hand, are expected to have a higher N- to C-terminal distance and higher amount of disorder in their structure. Due to the distance-based nature of the assay, shorter sequences are more likely to trigger the FRET than longer sequences, and are therefore present in higher numbers in the FRET-positive samples. We still observed some of the longer sequences enriched in the FRET-positive sorted cells, especially after the second round of screening for the *de novo* library, which might explain the shift towards a lower FRET ratio (Fig. 2a). Another possible explanation for the lower FRET ratio in the second round of FRET-positive sorting could be better chances of survival in *E. coli* for more disordered proteins, as has been reported previously in Tretyachenko et al. (2017). This might cause the FRET-positive populations to be skewed towards false positives that have better chances of survival. To avoid false positives in our top candidates, we introduced the additional filter that best sequences should be enriched in all rounds and also have the highest LFC.

According to predictive modeling, the length of sORF *de novo* proteins has little impact on sorting for the FRET-positive enrichment. For the FRET-negative enrichment, the length has a coefficient four times higher than for the FRET-positive sequences indicating a higher impact of length for the sorting of FRET-negative sequences. As expected, the length is correlated negatively with LFC for FRET-positive enriched sequences and positively for FRET-negative ones (Fig. 4). The disorder, as predicted, influences both FRET-positive and FRET-negative enrichment in the same order of magnitude but with opposite signs, as expected. For FRET-positive sequences, there is a negative correlation between LFC and predicted disorder while there is a positive correlation for FRET-negative sequences. In line with the literature on *de novo* proteins and random sequence proteins, a higher percentage of both libraries is FRET-negative and with that, predicted to be rich in disordered regions and lacking a stable tertiary fold (Bornberg-Bauer et al. 2021; Heames et al. 2023). In a previous study on *de novo* and random proteins (Heames et al. 2023), we observed that *de novo* proteins tend to be more soluble than random proteins of same lengths and amino acid frequency, mainly because random proteins have higher propensity for secondary structure. Here, we observe a similar trend with a slightly higher percentage of *de novo* sORF proteins enriched in the FRET-negative samples with high disorder predicted, compared to the random ones (Fig. 2b-c). This can be explained by the evolutionary pressure on newly arising proteins to be soluble, to not disturb the function of the cell and to not cause harmful aggregation (Ángyán et al. 2012; Agozzino and Dill 2018; Monti et al. 2021; Vakirlis et al. 2022). A similar trend can be observed when comparing the level of translation in different age groups in relation to their predicted disorder content. The younger proteins show higher levels of translation when more disordered, whereas in the older age groups the trend is not as prominent or not there at all (see supplementary material figure S1d, Supplementary Material online).

One of the remaining questions in *de novo* protein research is, if older *de novo* proteins are more structured and contain fewer disordered regions than newly emerging *de novo* proteins. A stable tertiary fold is difficult to attain from scratch (as in *de novo* emergence) and probably needs time to be formed by evolution (Bornberg-Bauer et al. 2021). Here we observed a trend for older *de novo* sORF proteins to be enriched in the FRET-positive samples compared to the younger ones (Figs. 4, 5). This trend might indicate a higher propensity for folding in the older *de novo* proteins as has been hypothesized previously (Wilson et al. 2017; Chen et al. 2023; Middendorf and Eicholt 2024). To best of our knowledge for the first time, we observe this trend experimentally, though further verification with a higher number of older and longer *de novo* proteins is needed. Overall, these results taken together with our earlier study (Heames et al. 2023), demonstrate how large libraries comprising proteins with random sequences or putative *de novo* proteins can be studied in a tractable experimental setup. While our results regarding structural properties are generally in good agreement with computational predictions, several outliers demonstrate that, wherever possible, experimental confirmations are recommended, particularly for single protein studies (Terwilliger et al. 2023). Akin to earlier computational studies on age stratification of *de novo* proteins we find that older *de novo* proteins, i.e. those which have orthologs and transcription in several outgroup species, are generally more compact and have a higher propensity for folding (Figs. 4, 5). This may hint at an evolutionary process which favors *de novo* proteins that survive purging by drift to become longer and to assume a more foldable scaffold. Almost all *de novo* proteins are much shorter than the optimal globular protein domain of an average 165 aa (Shen et al. 2005), often containing fewer than 100 aa (Heames et al. 2020; Blevins et al. 2021). Therefore, an optimal surface to volume ratio is difficult to accomplish with the hydrophobic-polar pattern along the peptide chain. To generate such a hydrophobic-polar pattern from a peptide chain which emerges from a DNA stretch which has most likely been subject to a strong and largely unconstrained drift for a long evolutionary time, positive selection would be necessary (Shen et al. 2005; Agozzino and Dill 2018). While many studies look for selection on *de novo* protein-encoding sequences (Zhao et al. 2014; Zhang et al. 2019; Heames et al. 2020), only few studies so far have found convincing evidence for *de novo* proteins being under positive selection (Gubala et al. 2017). Furthermore, computational studies on the extension of ORFs by read-through and loss of stop codons (Klasberg et al. 2018; Kleppe and Bornberg-Bauer 2018) also suggested that extending the peptide chain from previously noncoding regions will result in a higher degree of disorder, at least in first place. Alternatively, it is conceivable that older *de novo* proteins are not or only marginally extended because they are "born this way", as suggested recently by Peng and Zhao (2024). This hypothesis would be supported by the loss dynamics of *de novo* proteins which suggests that *de novo* genes emerge more often than duplicates but are lost much faster such that, after longer evolutionary time scales, the fraction of *de novo* emerged genes among novel genes has become much smaller (Schmitz et al. 2018; Grandchamp et al. 2023; Montañés et al. 2023). Our data do not allow for differentiating between these scenarios, but our results lay a foundation for future studies.

Methods

Identification of *de novo* sORFs

Human sORFs ($n = 2{,}626{,}006$) were downloaded from http://sorfs.org (now https://sorfs.ugent.be/) (Olexiouk et al. 2016). sORFs less than 30 aa long were discarded and the longest isoforms selected. The resulting set of 53,670 sORFs were used as a query against the NCBI-nr database. diamond was used with arguments: blastp-p28-e0.001–sensitive-b2.75. sORFs with significant hits (threshold 1e-3) in a genus other than *Homo* were discarded as non human-specific. The apparent emergence mechanism of the remaining (human-specific) orphan sORFs was subsequently obtained by mapping of each ORF to the outgroup genomes of four primate genomes as well as the *Mus musculus* genome (see supplementary material table S1, Supplementary Material online for accession numbers). First, BLAT was used to map each ORF against the six-frame translations of all five genomes. BLAT hits were filtered to only include the best hit for each reference sequence using the script pslCDnaFilter with arguments -maxAligns=1. Subsequently, we took a conservative approach to find the highest ranking annotation that overlapped with the set of BLAT hits across all outgroups for each sORF. Only if all BLAT hits across all outgroup genomes did not overlap with any annotated features (as defined in the corresponding Ensembl gtf files) did we define a sORF as 'intergenic *de novo*'. Alternately, if any of the BLAT hits overlapped with a gene feature but no CDS feature, we defined it as 'intronic *de novo*' or 'UTR *de novo*'. All sORFs overlapping with any annotated CDS features were discarded as non *de novo* resulting in a total number of 6,649 *de novo* sORFs.

Library Design and Oligo Specification

Two libraries, DN (*de novo*) and R (random) were designed *in silico*. We used the set of 6,649 *de novo* sORFs identified (see) as a starting point for library design. First, any homologous *de novo* sORFs were discarded using cd-hit (see script remove_similar.py). Coding sequences (CDSs) for library R were then generated by random selection of amino acids using the frequency of amino acids in library DN, with sequence lengths also matched to those of library DN. We then compared the amino acid frequencies to those of all proteins in UniProt (EMBL-EBI 2024). Final oligonucleotides were specified by addition of upstream and downstream barcodes, allowing each library to be PCR amplified separately from the oligo pool. DnaChisel (Zulkower and Rosser 2020) was subsequently used to codon optimize CDS regions for protein expression in *E. coli*, while avoiding introduction of unwanted restriction sites (inc. BsaI). Codon optimization of the target ORF by selecting the best target species codon possible, given additional optimization constraints. These two subpools of 3,750 sequences were specified within a pool of 7,500 oligos in total. For each ORF, flanking regions encoding BsaI sites and overhangs were added, and start codons were replaced with the canonical ATG start if it was not already present. Primers (16 bp) unique to each subpool were then added, and if the oligo length was less than 230 bp (the maximum available from Agilent), randomly generated filler sequences were added evenly up- and downstream of the target ORF (keeping primers at the extremities) to maintain 230 bp final length.

Scripts (see data availability) were used with the following arguments:

```
  python  remove_similar.py  human_denovo_
sorfs.csv
```

The resulting filtered file was used as input to generate a single subpool of oligos as follows: python build_oligos.py -i human_denovo_sorfs.unique.csv -s e_coli -n 3 750 -l 230 -r 1 -d primers.db -rf 1 -fL GGTCTCCA -fR GGCTCCCG AGACC The two resulting .csv files containing separate oligo pools were concatenated and oligos ordered from Agilent.

Age Group Classification

Ages of the *de novo* sORF proteins were assigned either based on sequence homology within outgroup genomes as described in above termed "BLAST age" or based on transcription termed "transcription age". The age of each sORF protein corresponds to the hit in the furthest species from human on the evolutionary timescale. Age one corresponds to human specific *de novo* sORF proteins, two to sORFs present up to chimpanzee, three up to gorilla, four up to orang utan, five up to macaque and six up to mouse (see supplementary material figure S11, Supplementary Material online). To determine the transcription age, we used the locations of each sORF in the outgroup genomes to search the RNA-seq dataset from Wang et al. (2020) covering brain, testis and liver tissues in human, macaque and mouse. We calculated TPM (transcript per million) values for each sORF in each species-tissue combination by counting reads with HTseq (Anders et al. 2015). TPM values were calculated using the region defined by each BLAT hit in a given species. Where more than one BLAT hit was kept in a given species, the highest TPM value across all hits was kept.

Prediction of Protein Properties

A number of sequence properties were predicted for the translated products of all library variants. Intrinsic structural disorder (ISD) was calculated using flDPnn (Hu et al. 2021) with a disorder threshold of 0.5 and mean, median, and percentage of disordered residues were extracted using multifasta_fldpnn.py. Secondary structure content was predicted using NetSurfP 3.0 (Høie et al. 2022) and percentages for secondary structure elements extracted using netsurf_predictions.py. For full structural prediction we used ESMFold (Lin et al. 2023) and AlphaFold2 (Jumper et al. 2021). Based on the top ranked model we calculated N- to C-terminal distance, average solvent-accessible surface area (ASA), and secondary structure content using DSSP (Kabsch and Sander 1983), and radius of gyration (see script esm_predictions.py). Sequence complexity was predicted by calculating Shannon entropy (Strait and Dewey 1996) (see script shannon.py).

pETMF Vector Construction

To generate the FRET folding sensor, we started with an in-house prepared pET09 backbone. pET09 is a modification of pET24a(+) (Milipore-Sigma, Burlington, MA) backbone with MCS switched for cloning cassette containing BsaI recognition sites. Plasmid harboring mVenus was obtained from Addgene (catalogue no.: 103,986) and mTurquoise2 was a gift from Ondrej Havranek (coding sequence corresponding to Addgene catalog no.: 61,602). To make a FRET pair fusion cassette, the genes of fluorescent proteins (FPs) were furnished with cloning elements by PCR. BsmBI sites were added to 5′ end of mTurquoise2 and 3′ end of mVenus sequence creating matching overhangs to the BsaI sites in pET09 backbone. The opposite termini of FPs were rigged with: (i) Eco53kI recognition sites and GGS motive to link the insert with the FPs; (ii) BsaI sites to be used for cloning of the libraries; (iii) BsmBI sites to assemble the FRET cloning cassette. pET09 vector was opened with BsaI enzyme and dephosphorylated with rSAP enzyme (NEB, Ipswich, MA), PCR amplicons of furnished FPs were digested with BsmBI enzymes and following clean-up all three elements were ligated with T4 DNA ligase (NEB, Ipswich, MA).

Construction of Controls

We generated plasmids for single expression of donor/acceptor and fusion expression of the FRET pair separated by either glycine/serine linkers of different lengths, two characterized random proteins, or six well-characterized *E.coli* proteins with diverse length and properties. In the case of GS controls, we used PCR with long protruding primers to add GGS sequences and BsaI sites directly to the genes of mTurquoise2 and mVenus and assembled them with the pET09 via NEBridge Golden Gate Assembly Kit (BsaI-HF v2) (NEB, Ipswich, MA). Control proteins were PCR amplified from either *E.coli* DH5 alpha genome or in-house generated constructs, to contain BsaI sites compatible with pETMF vector and subcloned via GGA. *E. cloni* 10G strain (LGC, Biosearch Technologies, Hoddesdon, UK) was used for all the cloning steps and plasmid amplification. Once confirmed by Sanger sequencing, constructs were retransformed into *E.coli* BL21 (DE3) for expression and cultures were stored as glycerol stocks at −80 °C.

Library Synthesis and Cloning

To obtain oligo pools of the libraries, we used Agilent's SurePrint OLS. Individual libraries were PCR amplified from the pool to contain BsaI sites compatible with pETMF vector. Purified PCR product and vector were added to 20 μl Golden Gate assembly reaction in 2:1 molar ratio and run overnight in 5-min cycles of 16 °C and 37 °C. The reaction mix was purified with Monarch PCR & DNA Cleanup Kit (NEB, Ipswich, MA). Next, 1 μl of purified reaction mix was electroporated to 25 μl of in-house prepared *E. cloni* 10G cells and incubated overnight at 37 °C on an LB-agar plate supplemented with 50 μg/ml kanamycin (same for all following growth media). The colonies were pooled and plasmid DNA isolated with Zyppy Plasmid Miniprep Kit (ZymoGenetics, Inc, Seattle, WA). In total, 50 ng of pETMF carrying either of the libraries was electroporated to 25 μl of *E.coli* BL21 and incubated overnight at 30 °C on an LB-agar plates containing kanamycin. The plates were washed with cold PBS containing kanamycin to collect the colonies. Finally, the mixture of scraped cells was used for both plasmid DNA isolation (serving as "presort" sample in NGS analysis) and to prepare 1 ml glycerol stocks (20% v/v) with OD600 adjusted to 1.

Protein Expression

Cells carrying control protein plasmids were inoculated from glycerol stocks, grown overnight at 37 °C, reinoculated to fresh media and grown at 37 °C. After reaching OD600=0.6, cells were cooled down and expression induced with IPTG to 0.5 mM concentration followed by expression at 25 °C for 16 h. Glycerol stocks of BL21 cells carrying library DNA were thawed on ice, diluted with fresh LB with kanamycin to OD600 = 0.2 and grown shaking at 37 °C until reaching cell density 0.6 (approximately 50 min). Cells were cooled down, IPTG added to 0.5 mM and expression carried out at 25 °C for 16 h.

Measurement of Donor Fluorescence Lifetime

Expressing cells were harvested (5 min, 4,000 × *g*, 4 °C) and washed three times with cold PBS. Density of the cultures was adjusted to approximately 1, while keeping the cells on ice. Fluorescence measurements were carried out on Photoluminescence Spectrometer FLS 1,000 (Edinburgh Instruments Ltd., Livingston, UK). We used Xenon lamp for steady-state measurements to obtain emission spectra of mTurquoise2. For time-resolved measurements a 405 nm picosecond diode laser with repetition rate set to 10 MHz was used to collect a fluorescence decay at 473 nm emission in three technical replicates.

Flow Cytometry

Following the overnight expression, the OD600 of cell cultures was adjusted to 1. The cells were collected (5 min, 4,000 × *g*, 4 °C), washed three times and diluted 30 times in cold, filtered PBS. The cytometer BD FACSAria Fusion was equipped with a 70 μm nozzle and ND (neutral density) 1.0 filter. The sample chamber temperature was set to 4 °C and in the case of sorting, the collection tube temperature was set to 30 °C. We used the 405 nm violet laser with 450/50 bandpass filter for donor emission, 540/40 filter for FRET emission and 488 nm blue laser with 530/30 filter for acceptor detection. SSC detection threshold was set to 300 to capture the size of *E. coli* cells. After initial gating for size and shape (SSC-A x FSC-A and FSC-H x FSC-A), we used donor versus acceptor channel plot to gate population positive for both fluorophores (P1). To set the gates for FRET-positive/negative sorting, the P1 population was projected as a histogram of a parameter derived from the ratio of the FRET channel and the donor channel (FRET ratio). The cells were sorted to rich recovery medium (LGC, Biosearch Technologies, Hoddesdon, UK) in a 1.5 ml centrifugation tube. Following the sorting, the cells were plated on a LB-agar+kanamycin plate and incubated at 30 °C for 16 h. Finally, the colonies were scraped to LB+kanamycin and used either for subsequent round of expression and sorting or the plasmid DNA was extracted. Stocks from all rounds of sorting were then cultured and expressed in a single experiment and recorded on BD LSRFortessa cytometer.

High-throughput Sequencing

The plasmid pools recovered from the FACS experiments were used to generate PCR amplicons for subsequent NGS analysis. We used Q5 High-Fidelity DNA Polymerase (NEB, Ipswich, MA) with 50 ng of plasmid DNA per 50 μL of PCR as a template and the reaction was run for eleven cycles. Primers were designed to anneal to the pETMF backbone and introduce nine sets of barcodes (see Zenodo repository). Following a clean-up, the amplicon size distribution for selected samples was obtained by the Agilent 2,100

Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA). In total, four sequencing libraries were generated by NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, Ipswich, MA). Sizes and concentrations of the sequencing libraries were again verified by the Agilent 2,100 Bioanalyzer. Finally, the samples were pooled as a part of a larger run on Illumina NextSeq platform. NextSeq 1,000/2,000 P1 Reagents with 600 cycles were used and we dedicated 340–440 K reads per sample. Reads were merged, trimmed and filtered to remove low quality reads using the fastp suite (Chen et al. 2018). Reads were mapped to CDS sequences of library DN and R, respectively, using the Burrows–Wheeler Alignment (BWA) MEM algorithm (Li and Durbin 2009). SAMtools was used for conversion to SAM file format, sorting, and indexing (Li et al. 2009). Reads mapped to each variant were then counted using the HTSeq python module (Anders et al. 2015) using the script sam_counter.py.

Enrichment Analysis

Library sequences were filtered out before enrichment analysis if the total read count number across replicates was below 50 and if the sequence was not present in at least two replicates including the presorted library. Enrichment of single sequences was calculated using the python implementation of DESeq2 (Love et al. 2014) comparing sorted samples after round one and two to the presorted samples. For analysis only the positive log fold changes with a significant adjusted *P*-value below 0.05 were used. To check for statistical differences we used SciPy (Virtanen et al. 2020). To analyze differences between the FRET-positive and FRET-negative groups we applied standard t-test, for differences between multiple groups we used Kruskal–Wallis test with Dunn post hoc test. To check whether the numbers of sequences sorted into FRET-positive or FRET-negative were nonrandom between age groups we applied Chi2 test.

Statistical Modelling

All statistical modeling was done using R version 4.3.1 (Ihaka and Gentleman 1996). The data were split first into *de novo* and random sequences and then into positive and negative FRET data sets. These data sets were further divided randomly in half into a training and test data set. Models were then created using the glmnet package (Tay et al. 2023) on the respective training data set. Cross-validation was run with 10-folds using type-measure deviance. This was done with four different alpha values; 1, 0.7, 0.5, and 0.3. The chosen model was the minimum lambda model with the lowest deviance out of the four alpha values. Predictive plots were then created using a custom link function from the coefficients of the training set models and plotted against the test data. For random sequences, the model intercept was included while for the *de novo* sequences, it was not.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgements

We thank Tereza Neuwirthova for running ESMFold predictions using computational resources provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. We thank Gustavo Fuertes Vives and Tatsiana Charnavets for their help with the fluorescence measurements and Ondrej Havranek for kindly providing the mTurquoise2 construct and discussing the flow cytometry results. We acknowledge the Imaging Methods Core Facility at BIOCEV for their support with obtaining flow cytometry data presented in this paper. We thank Max Gantz and Lars Eicholt for discussion of the results. We acknowledge support from the Open Access Publication Fund of the University of Münster. Figure 1 was created with BioRender.com.

Funding

M.A., F.B., E.B.B., and K.H. received funding from Volkswagen foundation grant code 98183. M.A. and E.B.B.received funding from HFSP grant HFSP - RGP004/2023. F.B. and K.H. received funding from Primus grant PRIMUS/20/SCI/012 from Charles University. The project received funding from the EU under the Horizon 2020 Research and Innovation Framework Programme No. 722610 and the Erasmus+ programme.

Data Availability

Code used for library design, prediction of protein properties and analysis is available at: https://zivgitlab.uni-muenster.de/ag-ebb/de-novo/fret-facs/

All used protein sequences, result files and raw NGS reads are available at Zenodo under: https://zenodo.org/doi/10.5281/zenodo.10498066

Literature Cited

Agozzino L, Dill KA. Protein evolution speed depends on its stability and abundance and on chaperone concentrations. Proc Natl Acad Sci USA. 2018:115(37):9092–9097. https://doi.org/10.1073/pnas.1810194115.

Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015:31(2):166–169. https://doi.org/10.1093/bioinformatics/btu638.

Ángyán AF, Perczel A, Gáspári Z. Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? FEBS Lett. 2012:586(16):2468–2472. https://doi.org/10.1016/j.febslet.2012.06.007.

Aubel M, Eicholt L, Bornberg-Bauer E. Assessing structure and disorder prediction tools for de novo emerged proteins in the age of machine learning. F1000Res. 2023:12:347. https://doi.org/10.12688/f1000research.

Bajar BT, Wang ES, Zhang S, Lin MZ, Chu J. A guide to fluorescent protein FRET pairs. Sensors (Basel). 2016:16(9):1488. https://doi.org/10.3390/s16091488.

Banning C, Votteler J, Hoffmann D, Koppensteiner H, Warmer M, Reimer R, Kirchhoff F, Schubert U, Hauber J, Schindler M, et al. A flow cytometry-based FRET assay to identify and analyse protein-protein interactions in living cells. PLoS ONE. 2010:5(2):e9344. https://doi.org/10.1371/journal.pone.0009344.

Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Cañas JL, Espinar L, Díez J, Carey LB, Albà MM. Uncovering de novo gene birth in yeast using deep transcriptomics. Nat Commun. 2021:12(1):604. https://doi.org/10.1038/s41467-021-20911-3.

Bornberg-Bauer E, Hlouchova K, Lange A. Structure and function of naturally evolved de novo proteins. Curr Opin Struct Biol. 2021:68:175–183. https://doi.org/10.1016/j.sbi.2020.11.010.

Broeils LA, Ruiz-Orera J, Snel B, Hubner N, van Heesch S. Evolution and implications of de novo genes in humans. Nat Ecol Evol. 2023:7(6): 804–815. https://doi.org/10.1038/s41559-023-02014-y.

Bungard D, Copple JS, Yan J, Chhun JJ, Kumirov VK, Foy SG, Masel J, Wysocki VH, Cordes MHJ. Foldability of a natural de novo evolved protein. Structure. 2017:25(11):1687–1696.e4. https://doi.org/10.1016/j.str.2017.09.006.

Chen J, Li Q, Xia S, Arsala D, Sosa D, Wang D, Long M. One million years of solitude: the rapid evolution of de novo protein structure and complex. 2023. https://doi.org/10.1101/2023.12.24.573215.

Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018:34(17):i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Dohmen E, Klasberg S, Bornberg-Bauer E, Perrey S, Kemena C. The modular nature of protein evolution: domain rearrangement rates across eukaryotic life. BMC Evol Biol. 2020:20(1):30. https://doi.org/10.1186/s12862-020-1591-0.

Dowling D, Schmitz JF, Bornberg-Bauer E, Aoife M. Stochastic gain and loss of novel transcribed open reading frames in the human lineage. Genome Biol Evol. 2020:12(11):2183–2195. https://doi.org/10.1093/gbe/evaa194.

Elofsson A. Progress at protein structure prediction, as seen in CASP15. Curr Opin Struct Biol. 2023:80:102594. https://doi.org/10.1016/j.sbi.2023.102594.

EMBL-EBI. Uniprotkb/trembl protein database release $2024_0$1 statistics. 2024.

Förster T. Zwischenmolekulare energiewanderung und fluoreszenz. Ann Phys. 1948:437(1-2):55–75. https://doi.org/10.1002/andp.19484370105.

Goedhart J, von Stetten D, Noirclerc-Savoye M, Lelimousin M, Joosen L, Hink MA, van Weeren L, Gadella TWJ, Royant A. Structure-guided evolution of cyan fluorescent proteins towards a quantum yield of 93%. Nat Commun. 2012:3:751. https://doi.org/10.1038/ncomms1738.

Grandchamp A, Kühl L, Lebherz M, Brüggemann K, Parsch J, Bornberg-Bauer E. Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in Drosophila melanogaster. Genome Res. 2023:33(6):872–890. https://doi.org/10.1101/gr.277482.122.

Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, Findlay GD. The goddard and saturn genes are essential for Drosophila male fertility and may have arisen de novo. Mol Biol Evol. 2017:34(5):1066–1082. https://doi.org/10.1093/molbev/msx057.

Guerzoni D, McLysaght A. De novo genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. Genome Biol Evol. 2016:8(4): 1222–1232. https://doi.org/10.1093/gbe/evw074.

Heames B, Buchel F, Aubel M, Tretyachenko V, Loginov D, Novák P, Lange A, Bornberg-Bauer E, Hlouchová K. Experimental characterization of de novo proteins and their unevolved random-sequence counterparts. Nat Ecol Evol. 2023:7(4):570–580. https://doi.org/10.1038/s41559-023-02010-2.

Heames B, Schmitz J, Bornberg-Bauer E. A continuum of evolving de novo genes drives protein-coding novelty in Drosophila. J Mol Evol. 2020:88(4):382–398. https://doi.org/10.1007/s00239-020-09939-z.

Her C, Yeh Y, Krishnan VV. The ensemble of conformations of antifreeze glycoproteins (AFGP8): a study using nuclear magnetic resonance spectroscopy. Biomolecules. 2019:9(6):235. https://doi.org/10.3390/biom9060235.

Høie MH, Kiehl EN, Petersen B, Nielsen M, Winther O, Nielsen H, Hallgren J, Marcatili P. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. Nucleic Acids Res. 2022:50:W510–W515. https://doi.org/10.1093/nar/gkac439.

Hu G, Katuwawala A, Wang K, Wu Z, Ghadermarzi S, Gao J, Kurgan L. flDPnn: accurate intrinsic disorder prediction with putative propensities of disorder functions. Nat Commun. 2021:12(1):4438. https://doi.org/10.1038/s41467-021-24773-7.

Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996:5(3):299–314. https://doi.org/10.2307/1390807.

Jumper JM, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zídek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021:596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983:22(12):2577–2637. https://doi.org/10.1002/bip.v22:12.

Klasberg S, Bitard-Feildel T, Callebaut I, Bornberg-Bauer E. Origins and structural properties of novel and de novo protein domains during insect evolution. FEBS J. 2018:285(14):2605–2625. https://doi.org/10.1111/febs.2018.285.issue-14.

Kleppe AS, Bornberg-Bauer E. Robustness by intrinsically disordered C-termini and translational readthrough. Nucleic Acids Res. 2018:46(19):10184–10194. https://doi.org/10.1093/nar/gky778.

Krishna MMG, Englander SW. The N-terminal to C-terminal motif in protein folding and function. Proc Natl Acad Sci USA. 2005:102(4):1053. https://doi.org/10.1073/pnas.0409114102.

Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, Findlay GD, Bornberg-Bauer E. Structural and functional characterization of a putative de novo gene in Drosophila. Nat Commun. 2020:12:1667. https://doi.org/10.1038/s41467-021-21667-6.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009:25(14): 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics. 2009:25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023:379(6637): 1123–1130. https://doi.org/10.1126/science.ade2574.

Liu J, Yuan R, Shao W, Wang J, Silman I, Sussman JL. Do "Newly Born" orphan proteins resemble "Never Born" proteins? A study using three deep learning algorithms. Proteins: Struct Funct Bioinform. 2023:91(8):1097–1115. https://doi.org/10.1002/prot.v91.8.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014:15(12):550. https://doi.org/10.1186/s13059-014-0550-8.

Marsch-Martínez N, Reyes-Olalde JI, Chalfun-Junior A, Bemer M, Durán-Medina Y, Ochoa-Sánchez JC, Guerrero-Largo H, Herrera-Ubaldo H, Mes J, Chacón A, et al. Twisting development, the birth of a potential new gene. iScience. 2022:25(12):105627. https://doi.org/10.1016/j.isci.2022.105627.

Matsuo T, Nakatani K, Setoguchi T, Matsuo K, Tamada T, Suenaga Y. Secondary structure of human de novo evolved gene product NCYM analyzed by vacuum-ultraviolet circular dichroism. Front Oncol. 2021:11:688852. https://doi.org/10.3389/fonc.2021.688852.

McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet. 2016:17(9):567–578. https://doi.org/10.1038/nrg.2016.78.

Middendorf L, Eicholt LA. Random, de novo, and conserved proteins: how structure and disorder predictors perform differently. Proteins. 2024:1–11. https://doi.org/10.1002/prot.26652.

Montañés JC, Huertas M, Messeguer X, Albà MM. Evolutionary trajectories of new duplicated and putative de novo genes. Mol Biol Evol. 2023:40(5):msad098. https://doi.org/10.1093/molbev/msad098.

Monti M, Armaos A, Fantini M, Pastore A, Tartaglia GG. Aggregation is a context-dependent constraint on protein evolution. Front Mol Biosci. 2021:8:678115. https://doi.org/10.3389/fmolb.2021.678115.

Niwa T, Uemura E, Matsuno Y, Taguchi H. Translation-coupled protein folding assay using a protease to monitor the folding status. Protein Sci Publ Protein Soc. 2019:28(7):1252–1261. https://doi.org/10.1002/pro.v28.7.

Ohno S. Evolution by gene duplication. London: George Alien & Unwin Ltd; 1970.

Olexiouk V, Crappé J, Verbruggen S, Verhegen K, Martens L, Menschaert G. sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2016:44(D1):D324–D329. https://doi.org/10.1093/nar/gkv1175.

Peng J, Zhao L. The origin and structural evolution of de novo genes in Drosophila. Nat Commun. 2024:15(1):810. https://doi.org/10.1038/s41467-024-45028-1

Philipps B, Hennecke J, Glockshuber R. FRET-based in vivo screening for protein folding and increased protein stability. J Mol Biol. 2003:327(1):239–249. https://doi.org/10.1016/S0022-2836(03)00077-9.

Pueyo JI, Magny EG, Couso JP. New peptides under the s(ORF)ace of the genome. Trends Biochem Sci. 2016:41(8):665–678. https://doi.org/10.1016/j.tibs.2016.05.003.

Rödelsperger C, Prabh N, Sommer RJ. New gene origin and deep taxon phylogenomics: opportunities and challenges. Trends Genet. 2019:35(12):914–922. https://doi.org/10.1016/j.tig.2019.08.007.

Sandmann C-L, Schulz JF, Ruiz-Orera J, Kirchner M, Ziehm M, Adami E, Marczenke M, Christ A, Liebe N, Greiner J, et al. Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Mol Cell. 2023:83(6):994–1011.e18. https://doi.org/10.1016/j.molcel.2023.01.023.

Schmitz JF, Chain FJJ, Bornberg-Bauer E. Evolution of novel genes in three-spined stickleback populations. Heredity. 2020:125(1-2):50–59. https://doi.org/10.1038/s41437-020-0319-7.

Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nat Ecol Evol. 2018:2(10):1626–1632. https://doi.org/10.1038/s41559-018-0639-7.

Shen M-Y., Davis FP, Sali A. The optimal size of a globular protein domain: a simple sphere-packing model. Chem Phys Lett. 2005:405(1-3):224–228. https://doi.org/10.1016/j.cplett.2005.02.029.

Sikosek T, Bornberg-Bauer E. In: Dittmar K, Liberles D, editors. Evolution After and Before Gene Duplication?. Wiley-Blackwell; 2010. p.105-131. https://doi.org/10.1002/9780470619902.ch6.

Strait BJ, Dewey TG. The Shannon information entropy of protein sequences. Biophys J. 1996:71(1):148–155. https://doi.org/10.1016/S0006-3495(96)79210-X.

Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011:12(10):692–702. https://doi.org/10.1038/nrg3053.

Tay JK, Narasimhan B, Hastie T. Elastic net regularization paths for all generalized linear models. J Stat Softw. 2023:106:1. https://doi.org/10.18637/jss.v106.i01.

Terwilliger TC, Liebschner D, Croll TI, Williams CJ, McCoy AJ, Poon BK, Afonine PV, Oeffner RD, Richardson JS, Read RJ, et al. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nat Methods. 2023:21(1):110–116. https://doi.org/10.1038/s41592-023-02087-4.

Tretyachenko V, Vymětal J, Bednárová L, Kopecký V, Hofbauerová K, Jindrová H, Hubálek M, Souček R, Konvalinka J, Vondrášek J, et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Sci Rep. 2017:7(1):15449. https://doi.org/10.1038/s41598-017-15635-8.

Uversky VN. The alphabet of intrinsic disorder. Intrinsically Disord Proteins. 2013:1(1):e24684. https://doi.org/10.4161/idp.24684.

Vakirlis N, Vance Z, Duggan KM, McLysaght A. De novo birth of functional microproteins in the human lineage. Cell Rep. 2022:41(12):111808. https://doi.org/10.1016/j.celrep.2022.111808.

Van Oss SB, Carvunis A-R. De novo gene birth. PLoS Genet. 2019:15(5):e1008160. https://doi.org/10.1371/journal.pgen.1008160.

van Rosmalen M, Krom M, Merkx M. Tuning the flexibility of glycine-serine linkers to allow rational design of multidomain proteins. Biochemistry. 2017:56(50):6565–6574. https://doi.org/10.1021/acs.biochem.7b00902.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020:17:261–272. https://doi.org/10.1038/s41592-019-0686-2.

Wang Z-Y, Leushkin E, Liechti A, Ovchinnikova S, Mößinger K, Brüning T, Rummel C, Grützner F, Cardoso-Moreira M, Janich P, et al. Transcriptome and translatome co-evolution in mammals. Nature. 2020:588(7839):642–647. https://doi.org/10.1038/s41586-020-2899-z.

Weisman CM, Eddy SR. Gene evolution: getting something from nothing. Curr Biol. 2017:27(13):R661–R663. https://doi.org/10.1016/j.cub.2017.05.056.

Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the preadaptation hypothesis of *de novo* gene birth. Nat Ecol Evol. 2017:1(6):0146. https://doi.org/10.1038/s41559-017-0146.

Wu B, Knudson A. Tracing the de novo origin of protein-coding genes in yeast. mBio. 2018:9(4):e01024-18. https://doi.org/10.1128/mBio.01024-18.

Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R. Rapid evolution of protein diversity by de novo origination in Oryza. Nat Ecol Evol. 2019:3(4):679–690. https://doi.org/10.1038/s41559-019-0822-5.

Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila melanogaster populations. Science (New York, N.Y.). 2014:343(6172):769–772. https://doi.org/10.1126/science.1248286.

Zulkower V, Rosser S. DNA Chisel, a versatile sequence optimizer. Bioinformatics. 2020:36(16):4508–4509. https://doi.org/10.1093/bioinformatics/btaa558.

Associate editor: Claudia Alvarez Carreño