


# ProtGPT2 is a deep unsupervised language model for protein design

Received: 1 April 2022

Accepted: 13 July 2022

Published online: 27 July 2022

 Check for updatesNoelia Ferruz <sup>1,3</sup>✉, Steffen Schmidt <sup>2</sup> & Birte Höcker <sup>1</sup>

Protein design aims to build novel proteins customized for specific purposes, thereby holding the potential to tackle many environmental and biomedical problems. Recent progress in Transformer-based architectures has enabled the implementation of language models capable of generating text with human-like capabilities. Here, motivated by this success, we describe ProtGPT2, a language model trained on the protein space that generates *de novo* protein sequences following the principles of natural ones. The generated proteins display natural amino acid propensities, while disorder predictions indicate that 88% of ProtGPT2-generated proteins are globular, in line with natural sequences. Sensitive sequence searches in protein databases show that ProtGPT2 sequences are distantly related to natural ones, and similarity networks further demonstrate that ProtGPT2 is sampling unexplored regions of protein space. AlphaFold prediction of ProtGPT2-sequences yields well-folded non-idealized structures with embodiments and large loops and reveals topologies not captured in current structure databases. ProtGPT2 generates sequences in a matter of seconds and is freely available.

Natural language processing (NLP) has seen extraordinary advances in recent years. Large pre-trained language models have drastically transformed the NLP field and with it, many of the tools we use in our daily lives, such as chatbots, smart assistants, or translation machines. Analogies between protein sequences and human languages have long been noted by us and others<sup>1,2</sup>. Protein sequences can be described as a concatenation of letters from a chemically defined alphabet, the natural amino acids, and like human languages, these letters arrange to form secondary structural elements (“words”), which assemble to form domains (“sentences”) that undertake a function (“meaning”). One of the most attractive similarities is that protein sequences, like natural languages, are information-complete: they store structure and function entirely in their amino acid order with extreme efficiency. With the extraordinary advances in the NLP field in understanding and generating language with near-human capabilities, we hypothesized that these methods open a new door to approach protein-related problems from sequence alone, such as protein design.

Although protein sequences and human languages are not without dissimilarities, their analogies have stimulated applying NLP

methods to solve protein research problems for decades<sup>2</sup>. Supervised NLP methods, where the input sequences are trained jointly with their labels to produce predictive models, have been applied to various tasks, such as detecting structural similarity or predicting stability<sup>3,4</sup>. A remarkable collection of supervised language models applied to biomolecules is available in the BioSeq-BLM platform<sup>5,6</sup>. Nevertheless, since the inception of the Transformer<sup>7</sup>, unsupervised learning, where the training occurs on unlabeled data, has emerged as a versatile tool for language modeling. Several Transformer-based models, such as TCR-BERT<sup>8</sup>, epiBERT<sup>9</sup>, ESM<sup>10</sup>, ProtTrans<sup>11</sup>, or ProteinBERT<sup>12</sup>, have shown to be very competitive with other methods<sup>13,14</sup>. Most of these models use BERT-like<sup>15</sup> architectures and denoising autoencoding training objectives, i.e., they are pre-trained by corrupting the input tokens in some way and trying to reconstruct the original sentence<sup>2</sup>. Although these models could be adjusted for generation<sup>16</sup>, their most direct application is sequence embedding.

Another important branch of language models benefits from autoregressive training, i.e., models are trained to predict subsequent words given a context. These models, the most well-known of which

<sup>1</sup>Department of Biochemistry, University of Bayreuth, Bayreuth, Germany. <sup>2</sup>Computational Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany.

<sup>3</sup>Present address: Institute of Informatics and Applications, University of Girona, Girona, Spain. ✉ e-mail: [noelia.ferruz-capapey@uni-bayreuth.de](mailto:noelia.ferruz-capapey@uni-bayreuth.de)

are possibly the GPT-x series<sup>17</sup>, excel at generating long, coherent text—sometimes to the extent that much debate has been raised about their potential misuse<sup>18</sup>. Protein autoregressive language models, such as ProGen<sup>19–21</sup>, RITA<sup>22</sup>, and DARK<sup>23</sup> have also been studied, and show the potential of autoregressive Transformers for protein design. Motivated by these works and the ever-increasing capabilities of English-speaking models such as the GPT-x series, we wondered whether we could train a generative model to (i) effectively learn the protein language, (ii) generate fit, stable proteins, and (iii) understand how these sequences relate to natural ones, including whether they sample unseen regions of the protein space.

Here, we introduce ProtGPT2, an autoregressive Transformer model with 738 million parameters capable of generating de novo protein sequences in a high-throughput fashion. ProtGPT2 has effectively learned the protein language upon being trained on about 50 non-annotated million sequences spanning the entire protein space. ProtGPT2 generates protein sequences with amino acid and disorder propensities on par with natural ones while being “evolutionarily” distant from the current protein space. Secondary structure prediction calculates 88% of the sequences to be globular, in line with natural proteins. Representation of the protein space using similarity networks reveals that ProtGPT2 sequences explore ‘dark’ areas of the protein space by expanding natural superfamilies. The generated sequences show predicted stabilities and dynamic properties akin to their natural counterparts. Since ProtGPT2 has been already pre-trained, it can be used to generate sequences on standard workstations in a matter of seconds or be further finetuned on sequence sets of a user’s choice to augment specific protein families. The model and datasets are available in the HuggingFace repository<sup>24</sup> at (<https://huggingface.co/nferruz/ProtGPT2>). Since protein design has an enormous potential to solve problems in fields ranging from biomedical to environmental sciences<sup>25,26</sup>, we believe that ProtGPT2 is a timely advance towards efficient high-throughput protein engineering and design.

## Results

### Learning the protein language

The major advances in the NLP field can be partially attributed to the scale-up of unsupervised language models. Unlike supervised learning, which requires the labeling of each data point, self-supervised (or often named unsupervised) methods do not require annotated data, thus promoting the use of ever-growing datasets such as Wikipedia or the C4 Corpus<sup>27</sup>. Given both the growth of protein sequence databases and the lack of annotation for a significant part of the protein space, protein sequences have become great candidates for unsupervised training<sup>4,10,11</sup> and now offer the opportunity to encode and generate protein sequences.

To achieve this goal, we trained a Transformer<sup>7</sup> to produce a model that generates protein sequences. Language models are statistical models that assign probabilities to words and sentences. We are interested in a model that assigns high probability to sentences ( $W$ ) that are semantically and syntactically correct or fit and functional, in the case of proteins. Because we are interested in a generative language model, we trained the model using an autoregressive strategy. In autoregressive models, the probability of a particular token or word ( $w_i$ ) in a sequence depends solely on its context, namely the previous tokens in the sequence. The total probability of a sentence ( $W$ ) is the combination of the individual probabilities for each word ( $w_i$ ):

$$p(W) = \prod_i^n p(w_i | w_{<i}) \quad (1)$$

We trained the Transformer by minimizing the negative log-likelihood over the entire dataset. More intuitively, the model must learn the relationships between a word  $w_i$ —or amino acid—and all the

previous ones in the sequence, and must do so for each sequence  $k$  in dataset ( $D$ ):

$$\mathcal{L}_{\text{CLM}} = - \sum_{k=1}^D \log p_{\theta}(w_i^k | w_{<i}^k) \quad (2)$$

To learn the protein language, we used UniRef50 (UR50) (version 2021\_04), a clustering of UniProt at 50% identity. We chose this dataset versus larger versions of UniParc (such as UR100) as it was previously shown to improve generalization and performance for the ESM Transformers<sup>10</sup>. Uniref50’s sequences populate the entire protein space, including the dark proteome, regions of the protein space whose structure is not accessible via experimental methods or homology modeling<sup>28,29</sup>. For evaluation, we randomly excluded 10% of the dataset sequences—these sequences are not seen by ProtGPT2 during the training process. The final training datasets contained 44.9 and 4.9 million sequences for training and evaluation, respectively. We tokenized our dataset using the BPE algorithm<sup>30</sup>. The final model is a decoder-only architecture of 36 layers and 738 million parameters.

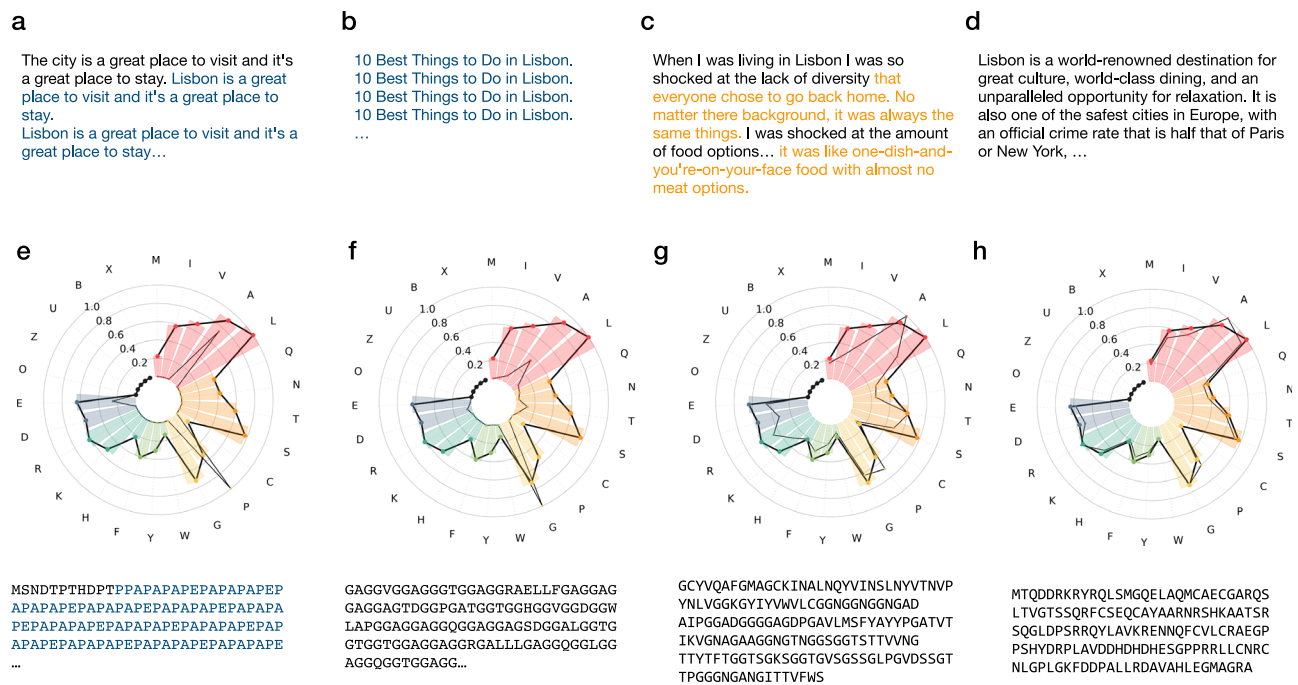
Analogous to the GLUE benchmark<sup>31</sup>—a collection of tools that computational linguists use to evaluate language models on different tasks such as question answering or translation—we also developed a series of extrinsic tests to assess the quality of ProtGPT2-generated sequences. The following sections elaborate on how ProtGPT2 generates de novo sequences with properties that resemble modern protein space.

### Statistical sampling of natural amino acid propensities

Autoregressive language generation is based on the assumption that the probability distribution of a sequence can be decomposed into the product of conditional next-word distributions (Eq. 1). However, there is still considerable debate about the best decoding strategy to emit sequences from a model<sup>32</sup>. It is not uncommon that well-trained generic language models that perform well in GLUE tasks generate incoherent gibberish or repetitive text depending on the sampling procedure<sup>32</sup>. We briefly summarize here the most used sampling strategies for language generation that we applied in this study.

Greedy search strategy selects the word with the highest probability at each timestep. Although algorithmically simple, the generated sequences are deterministic and soon also become repetitive (Fig. 1a). Beam search tries to alleviate this problem by retaining the most probable candidates, although the resulting texts still suffer from repetitiveness and are not as surprising as those from humans, which tend to alternate low and high probability tokens<sup>32</sup> (Fig. 1b). Lastly, random sampling moves away from deterministic sampling by randomly picking a word out of the top- $k$  most probable ones (Fig. 1c, d).

In a recent study, Holtzman et al.<sup>32</sup> investigated several sampling strategies to find the best parameters for text generation. Inspired by this work, we systematically generated sequences following different sampling strategies and parameters (Fig. 1). To assess what sampling procedure generates the most natural-like sequences, we compared the amino acid propensities of the generated set to that found in natural protein sequences (Methods). As stated by Hoffmann et al., we also observe greedy and beam search to produce repetitive, deterministic sequences, while random sampling dramatically improves the generated propensities (Fig. 1). Moreover, we also observe that high values of  $k$  are needed to generate sequences that resemble natural ones, i.e., our best results occur in the range of  $k > 800$  and we specifically chose  $k = 950$  in this work (Fig. 1h). As observed with other generative models<sup>33,34</sup>, our sampling improves when applying a repetition penalty of 1.2. Consequently, we used these sampling parameters for the rest of this work.



**Fig. 1 | Examples with different sampling parameters for GPT2-large after the context input: 'ten best things to do in Lisbon' (a–d) and ProtGPT2 without context (e–h).** While greedy and beam search produce repetitive sentences (a, b) and protein sequences (e, f), sampling generates creative texts, which, however,

can be degenerate (c) or not sample natural sequence propensities (g) for small values of  $k$ . Larger values of  $k$  produce quality text (d) and sequences whose propensities match natural ones. Repetitive and degenerate text are shown in blue and orange, respectively.

### ProtGPT2 sequences encode globular proteins

In order to evaluate ProtGPT2's generated sequences in the context of sequence and structural properties, we created two datasets, one with sequences generated from ProtGPT2 using the previously described inference parameters, and the other with randomly chosen sequences from UR50. Each dataset consists of 10,000 sequences. Since ProtGPT2 was trained in an unsupervised manner, i.e., without including functional annotations, our analyses focus on validating the structural and biochemical properties of ProtGPT2 sequences.

We first studied disordered and secondary structural content in the datasets. It has been previously shown that approximately 14% of the proteins found in bacteria and archaea are disordered<sup>28</sup>. To this end, we ran IUPred3<sup>35</sup> to analyze if the ProtGPT2-generated sequences are more prone to be disordered than a set of natural sequences. Interestingly, our analysis shows a similar number of globular domains among the ProtGPT2-generated sequences (87.59%) and natural sequences (88.40%). Several methods have been reported that detect short intrinsically disordered regions<sup>36</sup>. Since our goal is to provide high-level comparisons of globularity and prevalent disorder across datasets, we further performed an analysis of the protein sequences at the amino acid level using IUPred3. Remarkably, our results show a similar distribution of ordered/disordered regions for the two datasets, with 79.71 and 82.59% of ordered amino acids in the ProtGPT2 and natural datasets, respectively (Table 1).

We next investigated whether the similarities in disorder are a consequence of equivalent secondary structure element content. To this end, we computed PSIPRED<sup>37</sup> predictions for the ProtGPT2 and natural sequence datasets. The natural sequences display alpha-helical, beta-sheet, and coil contents of 45.19, 41.87, and 12.93%, respectively. The ProtGPT2 dataset presented percentages of 48.64, 39.70, and 11.66%, respectively.

These results indicate that ProtGPT2 generates sequences that resemble globular domains whose secondary structure contents are comparable to those found in the natural space.

### Table 1 | Disorder and secondary structure predictions of the natural and ProtGPT2 dataset

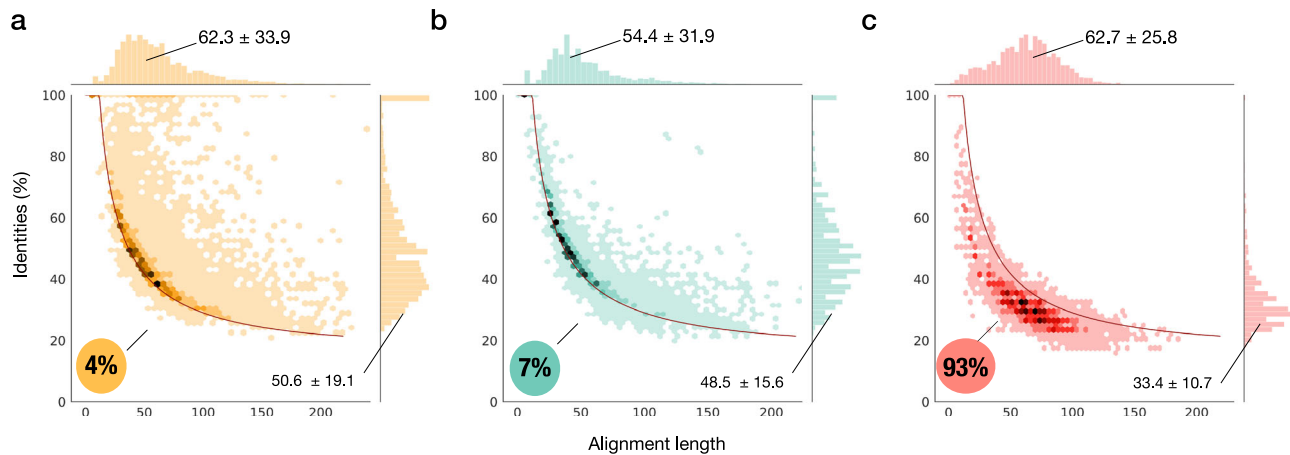
	Natural dataset	ProtGPT2 dataset
IUPred3 (globular domains)	88.40%	87.59%
Ordered content	79.71%	82.59%
Alpha-helical content	45.19%	48.64%
Beta-sheet content	41.87%	39.70%
Coil content	12.93%	11.66%

( $n = 10,000$  independent sequences/dataset).

### ProtGPT2 sequences are similar yet distant to natural ones

Proteins have diversified immensely in the course of evolution via point mutations as well as duplication and recombination. Using sequence comparisons, it is, however, possible to detect similarities between two proteins even when their sequences have significantly diverged. We wondered how related ProtGPT2 sequences are to natural ones. To this end, we utilized HHblits, a sensitive remote homology detection tool that uses profile hidden Markov models to search query sequences against a database<sup>38</sup>. We searched for homologs of the 10,000 sequences in ProtGPT2's dataset against the Uniclust30 database<sup>39</sup>. For comparison purposes, we also performed the same search with the natural dataset using the same settings. In addition, to analyze how completely random sequences would compare against ProtGPT2 ones, we also crafted a third dataset by randomly concatenating the 25 letters in the vocabulary.

Because we want to provide a quantitative comparison of the datasets' relatedness to modern protein space, we produced identity vs sequence length plots (Fig. 2). In detail, for each of the alignments found in Uniclust30, we depict the one with the highest identity and length. As a reference point in this sequence identity-length space, we use the HSSP curve<sup>40</sup>, a boundary set to define the confidence of



**Fig. 2 | Pairwise sequence identities vs. alignment length for each of the datasets (a: natural (yellow), b: ProtGPT2 (green), and c: random (red)) as computed with HHblits against the Uniclust30 database.** The lines depicted in red on each plot represent the HSSP curve, which we use as a reference to compare the three datasets<sup>40</sup>. Each plot shows a hexbin compartmentalization of the best-

scoring identities and their distributions. While natural (a) and ProtGPT2 (b) sequences show similar percentages below the curve, 93% of the sequences in the random dataset (c) do not have significantly similar sequences in the Uniclust30 database. Natural and ProtGPT2 datasets show significant differences in the high-identity range ( $n = 10,000$  independent sequences/dataset).

protein sequence relatedness. Proteins whose identity falls below this curve, an area known as the “twilight zone”, do not necessarily have similar 3D structures nor are likely homologous. Since the sequences in the ProtGPT2 and random datasets are not the consequence of protein evolution, we use the curve as a well-known threshold to compare the datasets.

When looking at the distribution of hits above and below the curve, we observe that HHblits finds many hits in the Uniclust30 database that are related to the dataset of natural sequences (Fig. 2a). Specifically, out of the 10,000 dataset sequences, 9621 (96.2%) showed identities above the HSSP curve. Similarly, 9295 ProtGPT2-generated sequences (93%) also have counterparts in the Uniclust30 database that align above the HSSP curve (Fig. 2b). Conversely, 93% of the randomly generated sequences fall below this threshold (Fig. 2c). Despite these similar patterns for the natural and ProtGPT2 datasets, the two datasets show differences in their distribution of hits. With a one-standard-deviation range of 31.5–69.7%, the natural dataset has a higher mean identity than the ProtGPT2 set, with a range of 32.9–64.1% (Fig. 2a, b). The differences between the natural and ProtGPT2 sequence distributions are not statistically significant ( $p$  value  $< 0.05$  Kolmogorov–Smirnov). However, substantial differences between the natural and ProtGPT2 datasets occur in the high-identity range ( $> 90\%$ ). Although 365 sequences in the ProtGPT2 dataset have high-identity sequences in Uniclust30, they correspond in all cases to alignments below 15 amino acids, whereas the natural dataset displays 760 sequences over 90% with an alignment length in the one-standard-deviation range of 14.8–77.3 amino acids. These results suggest that ProtGPT2 effectively generates sequences that are distantly related to natural ones but are not a consequence of memorization and repetition.

### ProtGPT2 generates ordered structures

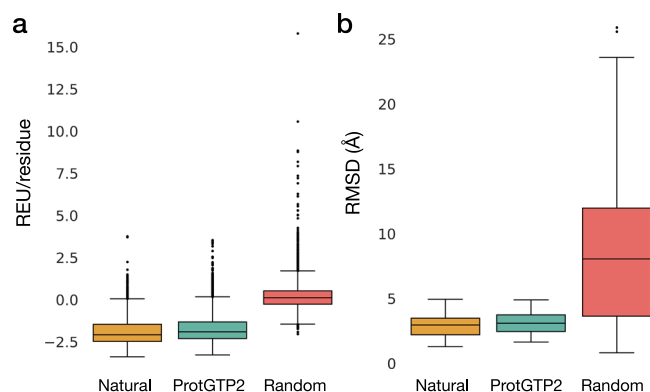
One of the most important features when designing de novo sequences is their ability to fold into stable ordered structures. We have evaluated the potential fitness of ProtGPT2 sequences in comparison to natural and random sequences in the context of AlphaFold predictions, Rosetta Relax scores, and molecular dynamics (MD) simulations.

AlphaFold<sup>41,42</sup> produces a per-residue estimate of its confidence on a scale from 0–100 (pLDDT). This score has been shown to correlate with order<sup>43</sup>: Low scores (pLDDT  $> 50$ ) tend to appear in disordered regions, while excellent scores (pLDDT  $> 90$ ) appear in ordered ones<sup>43</sup>.

Here we produced five structure predictions per sequence. The mean pLDDT of the dataset is 63.2 when taking the best-scoring structure per sequence and 59.6 when averaging across all five predictions per sequence. Moreover, 37% of sequences show pLDDT values over 70, in agreement with other recent studies<sup>23</sup>. A representation of all data points is shown in Supplementary Fig. 2a. Since pLDDT scores are a proxy for structural order, we turned to the natural and random datasets to see how they compare to ProtGPT2 sequences. In agreement with previous works, 66% of the sequences in the natural dataset were predicted with pLDDT values greater than 70<sup>43</sup>, giving an average value of 75.3 for the whole dataset (Supplementary Fig. 2b). In contrast, the predictions in the random dataset revealed a mean pLDDT value of 44, with only 7.4% of sequences with pLDDT values over 70 (Supplementary Fig. 2c).

To further validate the quality of the model, we performed Rosetta-RelaxBB runs on the three datasets<sup>44</sup>. Rosetta Relax performs a Monte Carlo optimization over the Rosetta energy function, which results in different backbone and rotamer conformations. Lower Rosetta Energy conformers correlate with more relaxed structures<sup>45</sup>. The most recent Rosetta Energy Forcefield (REF2015) strongly correlates with experimental variables such as heat capacity, density, and enthalpy<sup>46</sup>. This scoring function reflects the thermodynamic stability of one static protein conformation. Here we have performed Rosetta Relax experiments for the 30,000 sequences of the three datasets (Fig. 3a). A broad rule of thumb is that the total score (Rosetta Energy Units, REU) should lie between  $-1$  and  $-3$  per residue<sup>47</sup>. We observe such distribution in the natural and ProtGPT2 datasets, with averages of 1.90 and 1.73 REU/residue, respectively. As expected, the dataset of random sequences showed an average value of 0.13 REU/residue.

We further tested if ProtGPT2 sequences show similar dynamic properties as natural sequences. Proteins are dynamic entities; without their inherent flexibility, they would not be capable of interacting with other biomolecules and performing their functions in the cell<sup>48</sup>. To evaluate whether ProtGPT2 sequences show flexibility patterns in the same range as natural proteins, we randomly selected 12 sequences per dataset and ran three replicas of molecular dynamics (MD) of 100 ns each, totaling 108 trajectories and an aggregate time of 10.8 microseconds (Methods). To ensure that the dynamics observed during the simulations were not an artifact of different pLDDT values—and hence possible different disorder predictions—we made sure that differences among dataset-pLDDT mean values were not statistically different (Supplementary Fig. 3). The Root Mean Square Deviation means for



**Fig. 3 | Comparison of Rosetta and molecular dynamics calculations among the three datasets.** **a** Average Rosetta energy units per residue for the three datasets. AlphaFold prediction structures were used as input for the Rosetta RelaxBB protocol. 10,000 structures were run per dataset, one replica per system. **b** Root mean square deviation (RMSD) distribution for each MD dataset as computed by averaging RMSDs independently for each trajectory, represented as a boxplot. Twelve structures were simulated per dataset, three replicas per system. In both plots, the median is indicated as a black line; boxes depict the interquartile range (IQR), and whiskers represent 1.5 × IQR. Points outside this range are displayed as individual data points.

each of the trajectories in the natural and ProtGPT2 datasets resulted in average values of 2.93 and 3.12 Å, respectively (Fig. 3b). As expected, the random sequences showed significant deviations during the trajectories, with an average of 9.41 Å. While ProtGPT2 sequences showed higher values than the natural ones, the distributions are not significantly different (Mann–Whitney *U*-test, *p* value 0.39). The results indicate that ProtGPT2 sequences might have similar dynamic properties as proteins found in nature. The complete list of the trajectories' RMSD is presented in Supplementary Figs. 4, 5.

### ProtGPT2 transcends the boundaries of the current protein space

Several studies tried to reduce the large dimensionality of protein sequences into a few discernible dimensions for their analysis. Most representation methods consist of (i) hierarchical classifications of protein structures such as the ECOD and CATH databases<sup>49,50</sup>, (ii) Cartesian representations<sup>51</sup>, and similarity networks<sup>52,53</sup>. We recently represented the structural space in a network that showed proteins as nodes, linked when they have a homologous and structurally-similar fragment in common<sup>54</sup> and made the results available in the Fuzzle database<sup>55</sup>. The network represented 25,000 domains from the seven major SCOP classes and showed that the modern known protein space has both connected and “island-like” regions.

It is implausible that evolution has explored all possible protein sequences<sup>56</sup>. Therefore, the challenge has been posed whether we can design proteins that populate unexplored—or dark—regions of the protein space and if, by doing so, we can design novel topologies and functions<sup>56</sup>. Here, we integrated the ProtGPT2 sequences into our network representation of the protein space. To this end, we generated an HMM profile for each SCOPe2.07 and ProtGPT2 sequence, compared them in an all-against-all fashion using HHsearch and represented the networks with Protlego<sup>57</sup>. To avoid that specific sequences with several alignments end up represented by the same node in the network, we duplicate entries with two non-overlapping alignments, as previously described<sup>54</sup>.

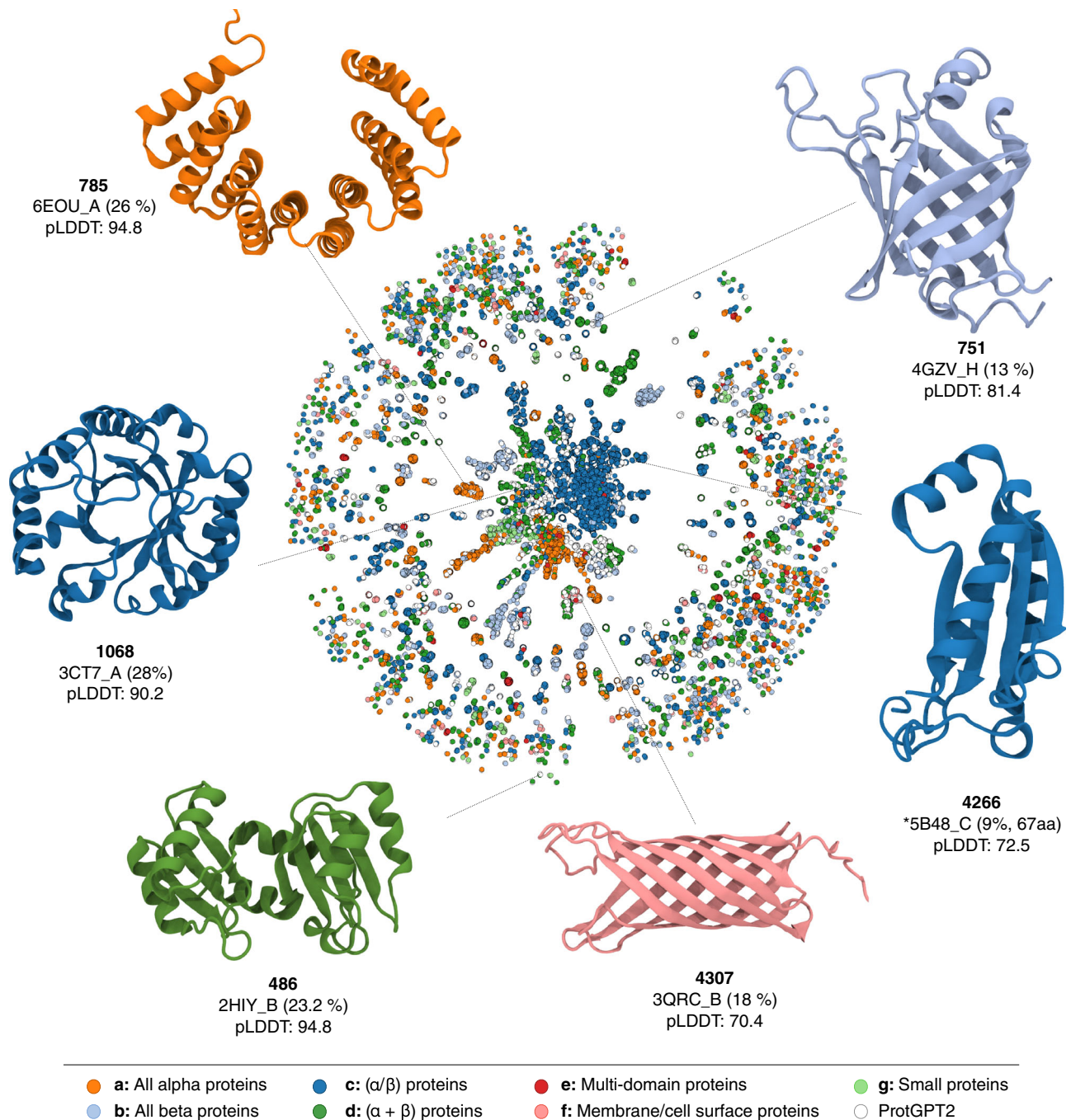
The network contains 59,612 vertices and 427,378 edges, comprising 1847 components or ‘island-like’ clusters (Fig. 4). The major component accumulates more than half of the nodes (30,690)—a

number significantly higher than the number observed in a network produced with the same settings but excluding ProtGPT2 sequences (Supplementary Fig. 6)—strongly suggesting that ProtGPT2 generates sequences that bridge separate islands in protein space. We select six examples across different areas of the network from topologically different SCOPe classes to showcase ProtGPT2 sequences at the structural level (Fig. 4). In particular, we report an all-β (751), two α/β (4266, 1068), one membrane protein (4307), an α + β (486) and all-α (785) structures. These structures illustrate ProtGPT2's versatility at generating de novo structures. For each case, we searched the most similar protein structure found in the PDB database using FoldSeek<sup>58</sup>. ProtGPT2 generates well-folded all-β structures (751, 4307), which despite recent impressive advances<sup>59</sup>, have for long remained very challenging<sup>60</sup>. ProtGPT2 also produces membrane proteins (4307), which pose a difficult target for protein design due to the challenges at specifying structure within the membrane and the laborious experimental characterizations<sup>61</sup>. Besides the generation of natural fold representatives, ProtGPT2 also produces previously unreported topologies. For example, we report protein 4266, whose topology does not match any of the currently reported structures in the PDB, with a low DALI Z-score of 5.4 and an RMSD of 3.0 Å to PDB 5B48 over 67 residues (identity 9%).

Nevertheless, possibly the most remarkable property of ProtGPT2 sequences is their significant deviation from all previously designed de novo structures, which often feature idealized topologies with loops and minimal structural elements. De novo proteins have the advantage of not carrying any evolutionary history and are thus amenable as a scaffold for virtually any function, but in practice, the lack of embodiments and longer loops hamper the design of crevices, surfaces, and cavities—necessary for the interaction with other molecules and function realization. ProtGPT2 sequences resemble the complexity of natural proteins, with multifaceted surfaces capable of allocating interacting molecules and substrates, thus paving the way for functionalization. In Fig. 4, we show structures 486 and 1060, two examples of such complex structures. In particular, 1068 shows a TIM-barrel fold, a topology which to date has met impressive success in de novo design<sup>62–64</sup>, but whose idealized structure has nevertheless proven challenging to extend via additional secondary elements and longer loops<sup>65,66</sup>.

### Preserved functional hotspots

Visual inspection of the structural superimposition of the best hits found with FoldSeek revealed several instances where the sidechains of ligand-interacting residues are conserved. Two examples are shown in Fig. 5. The natural structure most similar to sequence 357 (Fig. 5a) corresponds to PDB code 1XOP (chain A), a blue-light sensor domain that binds FAD. When superimposing the structures, we observe that 357 has retained the sidechain binding hotspots, with three residues identical (D169, Q150, and N131) and two different but capable of forming the same interactions, Lysine at position R165 and Histidine at position K127. Sequence 475 (Fig. 5b) is most similar to PDB code 5MIT (chain A), a phosphodiesterase that folds into a TIM-barrel and binds to the bacterial second messenger cyclic di-3',5'-guanosine monophosphate (PDB three-letter code C2E). Out of the five sidechain-interacting residues, the ProtGPT2 sequence preserves three residues (Q455, R473, and E469), and includes one substitution for another residue capable of hydrogen-bonding (aspartic acid for Q513). It is remarkable to note that ProtGPT2 has generated these sequences in a zero-shot fashion, i.e., without further finetuning in these two particular folds. These results have impactful consequences for protein engineering because ProtGPT2 appears to preserve binding positions in the generated sequences, despite the low identities (31.1 and 29.2% for 357 and 45, respectively), and can be used to augment the repertoires of specific folds and families.



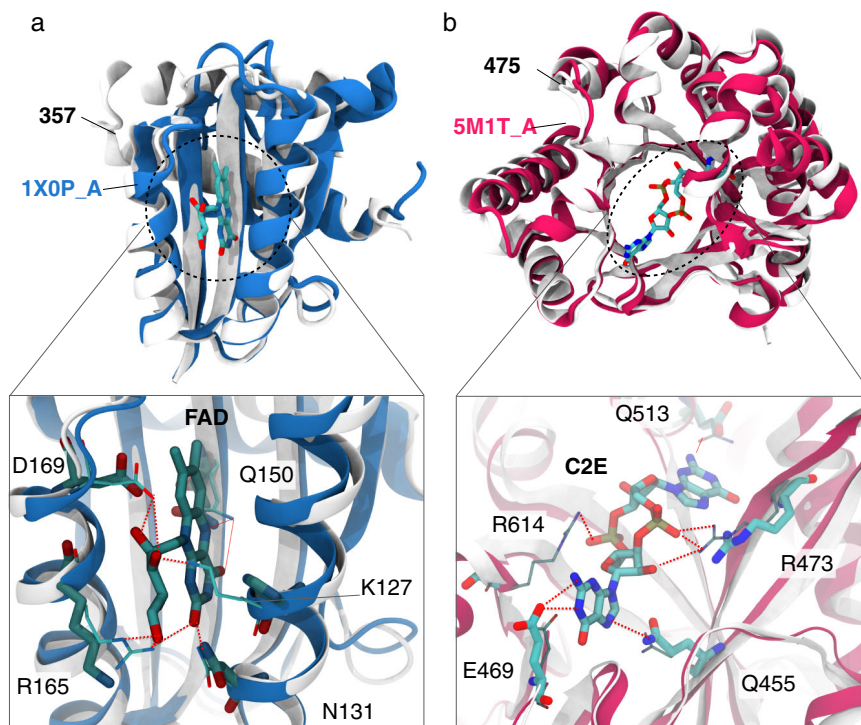
**Fig. 4 | An overview of the protein space and examples of proteins generated by ProtGPT2.** Each node represents a sequence. Two nodes are linked when they have an alignment of at least 20 amino acids and 70% HHsearch probability. Colors depict the different SCOPe classes, and ProtGPT2 sequences are shown in white. As examples, we select proteins of each of the major five SCOP classes: all- $\beta$  structures

(751),  $\alpha/\beta$  (4266 and 1068), membrane protein (4307),  $\alpha+\beta$  (486), and all- $\alpha$  (785). The structures were predicted with AlphaFold, and we indicate the code of the most similar structure in the PDB as found by FoldSeek<sup>58</sup>, except for protein 4266, where no structures were found.

## Discussion

The design of de novo proteins harnessing artificial intelligence methods has been meeting incredible success in the last 2 years<sup>10,67,68</sup>. Motivated by the unprecedented advances in NLP, we have implemented a generative language model, ProtGPT2, which has effectively learned the protein language. ProtGPT2 can generate sequences that are distantly related to natural ones and whose structures resemble the known structural space, with non-idealized complex structures. Since ProtGPT2 has been trained on the entire sequence space, the sequences produced by the model can sample any region, including

the dark proteome and areas traditionally regarded as very challenging in the protein design field, such as all- $\beta$  structures and membrane proteins. Visual superimposition of ProtGPT2 proteins with distantly related natural protein structures reveals that ProtGPT2 has also captured functional determinants, preserving ligand-binding interactions. As the design of artificial proteins can solve many biomedical and environmental problems, we see extraordinary potential in our protein language model. ProtGPT2 designs fit globular proteins in a matter of seconds without requiring further training on a standard workstation. ProtGPT2 can be conditioned towards a particular family, function, or



**Fig. 5 | Superimposition of the predicted structures for sequences 357 and 475 and the respective top scoring proteins in FoldSeek. a** Structural alignment of 357 with pdb 1XOP (chain A, blue). Shown are five residues in 1XOP that interact via their sidechains with the ligand FAD. Of these, three are identical in 357, and another two correspond to substitutions to the same amino acid type (R165 to

lysine and Q150 to histidine). **b** Structural alignment of 475 with pdb 5M1T (chain A) depicting five sidechain-interacting residues with ligand C2E. All amino acids in 475 are conserved except for residue R614, which was substituted by a glycine. The PDB structures are shown in color with their sidechains in a thinner representation.

fold by finetuning the model on a set of sequences of a user's choice. In this context, ProtGPT2 will enable the screening for proteins with similarities to natural proteins in order to improve, fine-tune or alter a specific biochemical function of a natural protein. Large-scale screening of ProtGPT2-designed protein libraries might identify proteins with folds not captured in structural databases and functions that have no related counterpart in the natural space. ProtGPT2 constitutes a big step forward towards efficient protein design and generation, and lays the groundwork for future experimental studies exploring the structural and functional parameters of designed proteins, and their subsequent real-world applications. Future efforts include the inclusion of conditional tags, which will enable the controlled generation of specific functions.

## Methods

### Vocabulary encoding

We use a BPE<sup>30</sup> tokenizer to train the vocabulary of our dataset. BPE is a sub-word tokenization algorithm that finds the most frequently used word roots, ensuring better performance than one-hot tokenization and avoiding the out-of-vocabulary problem. Given the size of Uniref50, we used Swiss-Prot (2021\_04) containing >0.5 M sequences to train our tokenizer. Following the training strategy of GPT2<sup>17</sup>, our final vocabulary contained 50,256 tokens that correspond to the most widely reused oligomers in protein space, with an average size of four amino acids per token (Supplementary Fig. 1). Learned positional embeddings were used as in the original GPT2.

### Dataset preparation

We took Uniref50 version 2021\_04 as the dataset for training, containing 49,874,565 sequences. 10% of the sequences were randomly selected to produce the validation dataset. The final training and validation datasets contained 44.88 and 4.99 million sequences,

respectively. We produced two datasets, one using a block size of 512 tokens, and another one with 1024 tokens. The results shown in this work correspond to a model trained with a block size of 512 tokens.

### Model pre-training

We use a Transformer decoder model as architecture for our training which processes input sequences tokenized with a BPE strategy. The model uses during training the original dot-scale self-attention as introduced by ref. 7. The model consist of 36 layers with a model dimensionality of 1280. The architecture matches that of the previously released GPT2-large Transformer<sup>17</sup>, which was downloaded from HuggingFace<sup>24</sup>. Model weights were reinitialized prior to training. The model was optimized using Adam ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ) with a learning rate of 1e-03. For our main model, we trained 65,536 tokens per batch (128 GPUs  $\times$  512 tokens). A batch size of 8 per device was used, totaling 1024. The model trained on 128 NVIDIA A100s in 4 days. Parallelism of the model was handled with DeepSpeed<sup>69</sup>.

### Model inference

We systematically sampled sequences using our main model using different inference parameters. In particular, we varied the repetition penalty from a range of 1.1 to 3.0 at each 0.1 units, top\_k from 250 to 1000 sampling every 50 units, and a top\_p from 0.7 to 1.0 with a window of 0.05 units. 100 sequences were produced for each sampling parameter set and the frequency of their amino acids compared to natural sequences. We observed which parameters produced fewer differences in the set of the seven most common amino acids in natural sequences. We also explored the beam search algorithm for beams in the range 50 to 100 using a window of 1 unit but it produced worse matches in all cases. To determine amino acid frequencies in natural sequences for comparison to ProtGPT2 samples, we

randomly picked 1 million sequences from the Uniref50 dataset. The best matching parameters were further downsampled with finer windows and their frequencies compared with radar plots, as shown in Fig. 1 in the main text. The best performing parameters in our dataset were top\_k 950, repetition penalty of 1.2, and default temperature and top\_p values of 1.

### Sequence dataset generation

Three sequence datasets were produced to compare their properties. The ProtGPT2 dataset was generated by sampling 1000 batches of 100 sequences, each with the selected inference parameters and a window context of 250 tokens. This step produced 100,000 sequences. We filtered from this set those sequences whose length had been cut due to the window context, giving a total of 29,876 sequences. From this set, we randomly selected 10,000 sequences. Their average length is  $149.2 \pm 50.9$  amino acids. The natural dataset was created by randomly sampling 100,000 sequences from Uniref50. 10,000 of these sequences were further chosen to ensure their average and standard deviation lengths matched that of the ProtGPT2 dataset sequences. The random dataset was created by concatenating the 25 amino acids that appear in UniRef50, which includes the 20 standard amino acids and other IUPAC codes such as “X”, “B”, “U”, “O”, and “Z”, by randomly concatenating them into sequences with a length taken from a normal distribution between 5 and 267 amino acids.

### Homology detection

Each sequence in the three 10k datasets was searched for similarity against the PDB70 and uniclust30 databases using HHblits<sup>70</sup>. We used the Uniclust30 database version 2018\_08 and the pdb70 version 2021\_04. As HHblits produces a list of alignments we selected all those over the HSSP curve as possible matches, and from these, selected the largest alignment. Thus, for each sequence in each dataset, the longest and the highest identity scoring alignment was selected and represented in Fig. 2.

### Disorder prediction

IUPred3 was run on ProtGPT2 and natural datasets using all three possible options to detect shorter (“short”) or longer (“longer”) unstructured regions, as well as structured regions (“glob”)<sup>35</sup>. Ordered content was determined with the “short” option. The output of the “glob” analysis also reports if any structured, globular domain was found, as shown in Table 1. We ran secondary structure prediction using PSIPRED v4.0 for each sequence in natural and ProtGPT2 datasets<sup>37</sup>. The alignments of the abovementioned HHblits searches were used as multiple sequence alignments. We computed the percentages for each secondary element by dividing the number of amino acids with a certain prediction by the total number of amino acids with a confidence value of 5 or more.

### AlphaFold2 structure prediction

We predicted five structures for each sequence in the ProtGPT2 dataset using AlphaFold ColabFold batch v1.2<sup>41</sup>.

### Network construction

Sequences in the ProtGPT2 and SCOP 2.07 filtered at 95% datasets were joined. For each sequence, we produced a multiple sequence alignment (MSA) using HHblits against the database Uniclust 2018\_08. Hidden Markov model profiles were produced for each MSA using HHblits<sup>70</sup>, and an all-against-all search for each profile was performed using HHsearch<sup>38</sup>. The network was constructed by representing every sequence as a node, and linking two nodes whenever they have an alignment of at least 20 amino acids with 70% HHsearch probability. Extensive details on the all-against-all comparison and network construction, and tools to generate the networks can be found in our

previous works Fuzzle<sup>54,55</sup> and Protlego<sup>57</sup>. Detection of similar topologies was determined with FoldSeek<sup>58</sup>.

### Molecular dynamics simulations

Simulation systems were built and run with the software HTMD<sup>71</sup>. In all cases, systems comprised solvated all-atom cubic boxes. Simulation boxes consisted of a protein centered at the origin of coordinates and explicit solvent molecules and neutralizing NaCl ions were added to each box. The Amber 19SB forcefield was used<sup>72</sup>. Three replicas were constructed per sequence. All systems were minimized, equilibrated, and run with ACEMD<sup>73</sup> using default parameters: each system was minimized and relaxed under NPT conditions for 1 ns at 1 atm and 300 K using a time-step of 4 fs, rigid bonds, cutoff of 9 Å, and PME for long-range electrostatics. Heavy protein and ligand atoms were constrained by a 10 kcal/mol/Å<sup>2</sup> spring constant. Production simulations were run in the NVT ensemble using a Langevin thermostat with a damping of 0.1 ps<sup>-1</sup> and a hydrogen mass repartitioning scheme to achieve timesteps of 4 fs<sup>74</sup>.

### Rosetta calculations

Rosetta Relax runs were produced with the Rosetta Software Suite v3.12<sup>44</sup> using as input structure the best-scoring prediction from AlphaFold.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

The model weights are publicly available in the HuggingFace repository: <https://huggingface.co/nferruz/ProtGPT2> and Zenodo: <https://doi.org/10.5281/zenodo.6796843> [<https://zenodo.org/record/6796843#.YswB9XbMIVA>]. The dataset for training is available at: [https://huggingface.co/datasets/nferruz/UR50\\_2021\\_04](https://huggingface.co/datasets/nferruz/UR50_2021_04). The three sequence datasets in this work are available at: [https://huggingface.co/datasets/nferruz/dataset\\_fastas](https://huggingface.co/datasets/nferruz/dataset_fastas). The AlphaFold predictions for the three datasets are available at [https://huggingface.co/datasets/nferruz/dataset\\_alphafold](https://huggingface.co/datasets/nferruz/dataset_alphafold). The Uniref50 original database version 21\_04 is available at [https://ftp.uniprot.org/pub/databases/uniprot/previous\\_releases/release-2021\\_04/](https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2021_04/). The Uniclust30 database version 2018\_08 is available at [http://gwdu11.gwdg.de/~compbiol/uniclust/2018\\_08/uniclust30\\_2018\\_08\\_hhsuite.tar.gz](http://gwdu11.gwdg.de/~compbiol/uniclust/2018_08/uniclust30_2018_08_hhsuite.tar.gz).

### Code availability

The model was trained with the HuggingFace transformers Trainer version 4.14.1. The code and documentation are available here: [https://huggingface.co/docs/transformers/main\\_classes/trainer](https://huggingface.co/docs/transformers/main_classes/trainer).

### References

1. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
2. Ferruz, N. & Höcker, B. Controllable protein design with language models. *Nat. Mach. Intell.* **4**, 521–532 (2022).
3. Bepler, T. & Berger, B. Learning the protein language: evolution, structure, and function. *Cell Syst.* **12**, 654–669.e3 (2021).
4. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
5. Li, H. L., Pang, Y. H. & Liu, B. BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. *Nucleic Acids Res.* **49**, e129–e129 (2021).
6. Liu, B., Gao, X. & Zhang, H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* **47**, e127–e127 (2019).



7. Vaswani, A. et al. Transformer: attention is all you need. In *Advances in Neural Information Processing Systems* 5999–6009 (2017).
8. Wu, K. et al. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.11.18.469186> (2021).
9. Park, M., Seo, S., Park, E. & Kim, J. EpiBERTope: a sequence-based pre-trained BERT model improves linear and structural epitope prediction by learning long-distance protein interactions effectively. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.02.27.481241> (2022).
10. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**, e2016239118 (2021).
11. Elnaggar, A. et al. ProtTrans: Towards Cracking the Language of Life Code Through Self-Supervised Deep Learning and High Performance Computing. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.2021.3095381>.
12. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).
13. Yang, K. K., Lu, A. X. & Fusi, N. K. Convolutions are competitive with transformers for protein sequence pretraining. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.05.19.492714> (2022).
14. Rao, R. et al. Evaluating protein transfer learning with TAPE. *Adv. Neural Inf. Process. Syst.* **32**, 9689–9701 (2019).
15. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at *arXiv:1810.04805* (2018).
16. Johnson, S. R., Monaco, S., Massie, K. & Syed, Z. Generating novel protein sequences using Gibbs sampling of masked language models. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.01.26.428322> (2021).
17. Radford, A. et al. Language models are unsupervised multitask learners. <https://github.com/codelucas/newspaper> (2018).
18. OpenAI says its text-generating algorithm GPT-2 is too dangerous to release. <https://slate.com/technology/2019/02/openai-gpt2-text-generating-algorithm-ai-dangerous.html> (2019).
19. Madani, A. et al. ProGen: language modeling for protein generation. (2020).
20. Madani, A. et al. Deep neural language modeling enables functional protein generation across families. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.07.18.452833> (2021).
21. Nijkamp, E. et al. ProGen2: exploring the boundaries of protein language models. Preprint at *arxiv* <https://doi.org/10.48550/arxiv.2206.13517> (2022).
22. Hesslow, D. et al. RITA: a Study on Scaling Up Generative Protein Sequence Models. Preprint at *arXiv* 2205.05789 (2022).
23. Moffat, L., Kandathil, S. M. & Jones, D. T. Design in the DARK: learning deep generative models for de novo protein design. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.01.27.478087> (2022).
24. Wolf, T. et al. HuggingFace’s transformers: state-of-the-art natural language processing. Preprint at *arXiv* 1910.03771 (2019).
25. Campeotto, I. et al. One-step design of a stable variant of the malaria invasion protein RH5 for use as a vaccine immunogen. *Proc. Natl Acad. Sci. USA* **114**, 998–1002 (2017).
26. Lu, H. et al. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**, 662–667 (2022).
27. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020).
28. Perdigão, N. et al. Unexpected features of the dark proteome. *Proc. Natl Acad. Sci. USA* **112**, 15898–15903 (2015).
29. Perdigão, N., Rosa, A. C. & O’Donoghue, S. I. The Dark Proteome Database. *BioData Min.* **10**, 24 (2017).
30. Gage, P. A new algorithm for data compression. <https://doi.org/10.5555/177910.177914>.
31. Wang, A. et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding. Preprint at *arXiv* (2018).
32. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration. *CEUR Workshop Proc.* **2540**, (2019).
33. Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C. & Socher, R. CTRL: a conditional transformer language model for controllable generation. Preprint at *arxiv* (2019).
34. Madani, A. et al. ProGen: language modeling for protein generation. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.03.07.982272> (2020).
35. Abor Erd, G. , Os, , Atyásaty’atyás Pajkos, M. , Dosztányi, Z. & Dosztányi, D. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* **49**, W297–W303 (2021).
36. Tang, Y. J., Pang, Y. H. & Liu, B. DeepIDP-2L: protein intrinsically disordered region prediction by combining convolutional attention network and hierarchical attention network. *Bioinformatics* **38**, 1252–1260 (2022).
37. Buchan, D. W. A. & Jones, D. T. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).
38. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
39. Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170 (2017).
40. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.* **12**, 85–94 (1999).
41. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
42. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
43. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
44. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling symmetric macromolecular structures in Rosetta3. *PLoS ONE* **6**, e20450 (2011).
45. Sauer, M. F., Sevy, A. M., Crowe, J. E. & Meiler, J. Multi-state design of flexible proteins predicts sequences optimal for conformational change. *PLoS Comput. Biol.* **16**, e1007339 (2020).
46. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031 (2017).
47. Wedemeyer, M. J., Mueller, B. K., Bender, B. J., Meiler, J. & Volkman, B. F. Modeling the complete chemokine-receptor interaction. *Methods Cell Biol.* **149**, 289–314 (2019).
48. Miller, M. D. & Phillips, G. N. Moving beyond static snapshots: protein dynamics and the protein Data Bank. *J. Biol. Chem.* **296**, 100749 (2021).
49. Cheng, H. et al. ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.* **10**, e1003926 (2014).
50. Sillitoe, I. et al. CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).
51. Osadchy, M. & Kolodny, R. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proc. Natl Acad. Sci. USA* **108**, 12301–12306 (2011).
52. Alva, V., Remmert, M., Biegert, A., Lupas, A. N. & Söding, J. A galaxy of folds. *Protein Sci.* **19**, 124–130 (2010).
53. Nepomnyachiy, S., Ben-Tal, N. & Kolodny, R. Global view of the protein universe. *Proc. Natl Acad. Sci. USA* **111**, 11691–11696 (2014).
54. Ferruz, N. et al. Identification and analysis of natural building blocks for evolution-guided fragment-based protein design. *J. Mol. Biol.* **432**, 3898–3914 (2020).

55. Ferruz, N., Michel, F., Lobos, F., Schmidt, S. & Höcker, B. Fuzzle 2.0: ligand binding in natural protein building blocks. *Front. Mol. Biosci.* **8**, 805 (2021).
56. Huang, P. S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
57. Ferruz, N., Noske, J. & Höcker, B. Protlego: a Python package for the analysis and design of chimeric proteins. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btab253> (2021).
58. Kempen, M. van et al. Foldseek: fast and accurate protein structure search. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.02.07.479398> (2022).
59. Marcos, E. et al. De novo design of a non-local  $\beta$ -sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **25**, 1028–1034 (2018).
60. Pan, X. & Kortemme, T. Recent advances in de novo protein design: Principles, methods, and applications. *J. Biol. Chem.* **296**, 100558 (2021).
61. Xu, C. et al. Computational design of transmembrane pores. *Nature* **585**, 129–134 (2020).
62. Romero-Romero, S. et al. The Stability Landscape of de novo TIM Barrels Explored by a Modular Design Approach. *J. Mol. Biol.* **433**, 167153 (2021).
63. Huang, P. S. et al. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
64. Anand, N. et al. Protein sequence design with a learned potential. *Nat. Commun.* **13**, 1–11 (2022).
65. Kordes, S., Romero-Romero, S., Lutz, L. & Höcker, B. A newly introduced salt bridge cluster improves structural and biophysical properties of de novo TIM barrels. *Protein Sci.* **31**, 513–527 (2022).
66. Wiese, J. G., Shanmugaratnam, S. & Höcker, B. Extension of a de novo TIM barrel with a rationally designed secondary structure element. *Protein Sci.* **30**, 982–989 (2021).
67. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
68. Ferruz, N. & Höcker, B. Dreaming ideal protein structures. *Nat. Biotechnol.* **40**, 171–172 (2022).
69. Rasley, J., Rajbhandari, S., Ruwase, O. & He, Y. DeepSpeed: system optimizations enable training deep learning models with over 100 billion parameters. In *Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3505–3506 (Association for Computing Machinery, 2020).
70. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
71. Doerr, S., Harvey, M. J., Noé, F. & De Fabritiis, G. HTMD: high-throughput molecular dynamics for molecular discovery. *J. Chem. Theory Comput.* **12**, 1845–1852 (2016).
72. Tian, C. et al. Ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).
73. Harvey, M. J., Giupponi, G. & De Fabritiis, G. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* **5**, 1632–1639 (2009).
74. Ferruz, N., Harvey, M. J., Mestres, J. & De Fabritiis, G. Insights from fragment hit binding assays by molecular simulations. *J. Chem. Inf. Model.* **55**, 2200–2205 (2015).

## Acknowledgements

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High-Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under an early-access NHR project. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) —440719683. We thank Thomas Zeiser for his considerate support and Surbhi Dhingra for feedback on the manuscript. N.F. acknowledges support from an AGAUR Beatrui de Pinós MSCA-COFUND Fellowship (project 2020-BP-00130). The authors thank funding from the German Research Foundation (DFG) - 491183248 and the Open Access Publishing Fund of the University of Bayreuth.

## Author contributions

N.F. conceived the work, trained the model, analyzed the data, and wrote the manuscript. S.S. produced the IUPred3 disorder predictions and analysis and wrote the manuscript. B.H. analyzed the data and wrote the manuscript. The three authors discussed the results and supervised the work.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-32007-7>.

**Correspondence** and requests for materials should be addressed to Noelia Ferruz.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022