Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Structural Biology

Foundation models of protein sequences: A brief overview

Andreas Bjerregaard[1,2,a], Peter Mørch Groth[1,3,a], Søren Hauberg[4], Anders Krogh[1,2] and Wouter Boomsma[1]

Protein sequence models have evolved from simple statistics of aligned families to versatile foundation models of evolutionary scale. Enabled by self-supervised learning and an abundance of protein sequence data, such foundation models now play a central role in protein science. They facilitate rich representations, powerful generative design, and fine-tuning across diverse domains. In this review, we trace modeling developments and categorize them into methodological trends over the modalities they describe and the contexts they condition upon. Following a brief historical overview, we focus our attention on the most recent trends and outline future perspectives.

Addresses
[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[2] Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark
[3] Novonesis, Kgs, Lyngby, Denmark
[4] Section for Cognitive Systems, Technical University of Denmark, Kgs, Lyngby, Denmark

Corresponding author: Boomsma, Wouter (wb@di.ku.dk)
[a] Equal contributions.

Current Opinion in Structural Biology 2025, **91:**103004

This review comes from a themed issue on **VSI:Sequences and Topology 2025**

Edited by **Arne Elofsson** and **Rachel Kolodny**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.sbi.2025.103004

Introduction

We have long modeled protein traits statistically. Statistical modeling aims to generalize from limited observed data to establish general relationships between entities of interest and make useful downstream predictions. To *train* a statistical model, we need examples of matching input and output data. In our context, the input data would be a protein sequence, and output values could be any measurable quantity of interest, e.g., protein stability. Often, our ability to train models is hampered by only having limited experimental data. However, this scarcity relates only to the *output* values—we typically have an abundance of protein sequence data available, for which we do not have corresponding experimental output observations. This problem of a lack of *labeled* data is common to many domains, and the machine learning community has developed various techniques for learning models from *only* input data. These approaches are typically designed to learn abstract vector *representations* or *embeddings* of the input data, aiming to capture inherent patterns in the input, making it easier to subsequently learn the output trait of interest. The primary technique used for learning representations is called *self-supervised* learning, where models are given only partial observations and are asked to predict the missing parts.

Following developments in natural language processing, self-supervised models of protein sequences have grown in complexity and size, requiring months of training time on large compute clusters [1]. Many varieties of these models now exist, differing in training objectives and input modalities. Simultaneously, their applicability has broadened beyond providing representations. Their generative capabilities are used directly in protein design [2], their likelihoods are used for *zero-shot* predictions [3],[1] and they are increasingly *fine-tuned* before being used for downstream prediction tasks [4—6]. For these reasons, the focus has moved from representations to the large self-supervised models that the representations can be extracted from. These models have already shown substantial real-world impact, e.g. in the field of protein design [7]. They are now often referred to as *foundation models*.

In this review, we outline the historical trajectory of generative protein models, emphasizing the crucial shift from local models over aligned sequences to global models over large sets of sequences, culminating in the ongoing surge of novel foundation models. We discuss key open questions, such as the role of representations,

[1] Zero-shot learning refers to the setting where a model is used to make predictions on a task that it was not trained on. In protein language modeling it is frequently used to describe the ability of models trained on sequence data to predict variant effects (e.g., change in stability) through the likelihood.

the potential for incorporating new modalities, and the future impact of scaling, to provide a qualitative overview of the state-of-the-art and offer perspectives on future directions.

Early generative models of protein sequences

Protein foundation models can be designed in different ways. For pure representation-learning purposes, one approach is to construct vector representations such that distances in vector space approximate a meaningful metric in output space [8], or adhere to contrastive learning objectives [9,10] whereby models learn to recognize similar examples as being related while distinguishing them from dissimilar ones. In the current literature, foundation models are often defined to entail the ability to generate sequences. We will follow this definition in our review. Such *generative* models are trained through self-supervised training, e.g. autoregressively (one position at a time, conditioned on all preceding positions) or through masked language modeling (predicting masked positions conditioned on unmasked positions). We can think of these models as approximating a distribution over protein sequences.[2] The difference between such models is how they define *context*, i.e., what the probability distribution is conditioned on.

To understand these differences, it is fruitful to consider the origins of generative models for protein sequences (see Figure 1). The simplest of such models focuses on proteins within a single family and rely on the sequences being *aligned* prior to modeling. Once equivalent positions in related proteins are placed neatly underneath each other in columns, the statistical modeling of amino acid propensities involves only simple counting statistics [12]. This leads to site-independent models (e.g. position-specific scoring matrices), which assign probabilities to each amino acid at a given position, but are limited by their assumption that positions evolve independently. To overcome this, pairwise models (e.g. Potts models) were introduced to capture coevolution between amino acid pairs, offering a more expressive view of sequences and forming the basis for contact prediction, which provided insights into the spatial arrangement of residues in protein structures [13]. Later, the DeepSequence model [14] demonstrated that it was possible to capture higher-order residue dependencies with a variational autoencoder and provided an early example of learned protein representations.

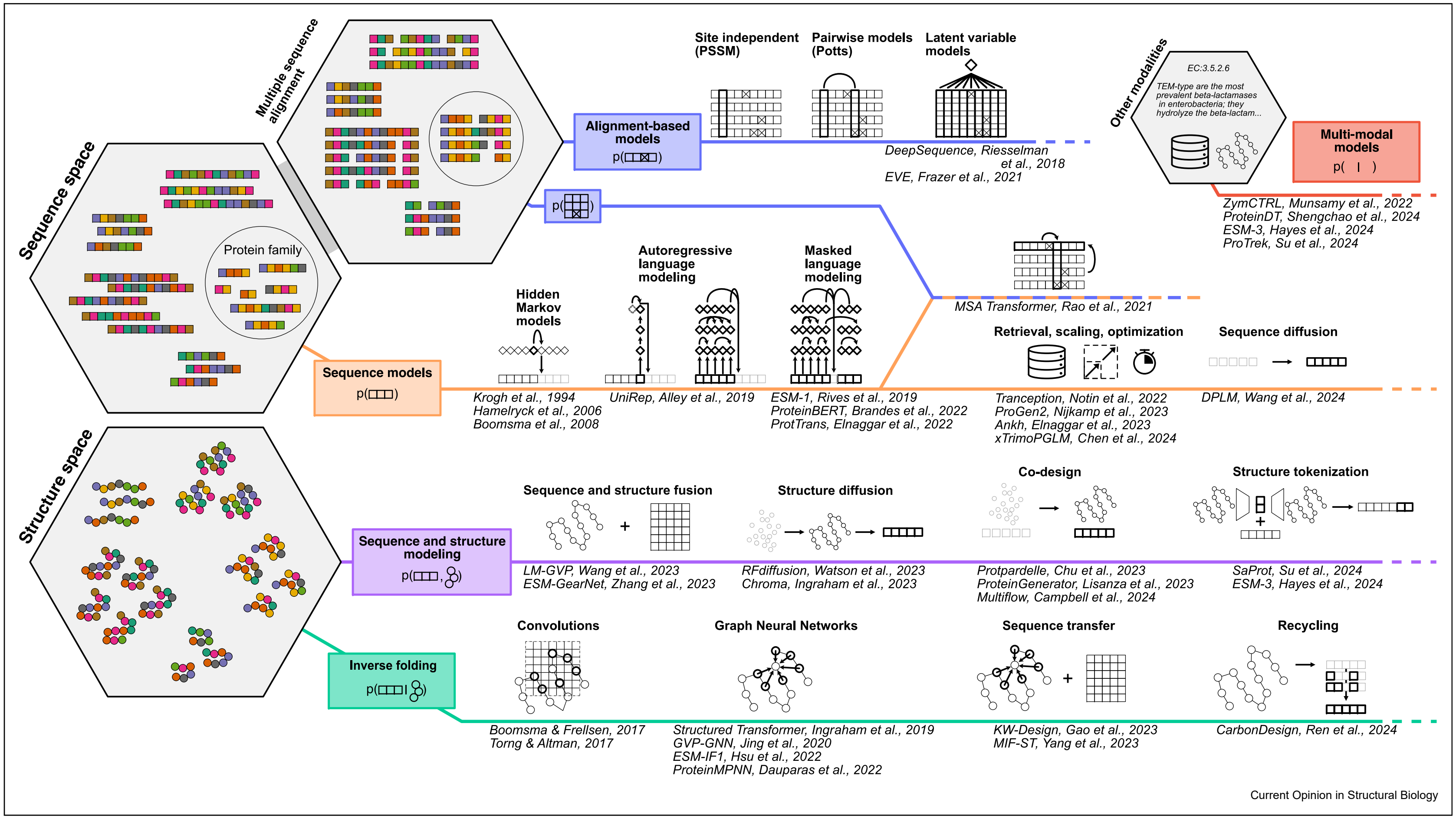
Despite an abundance of successful applications in bioinformatics, the reliance on aligned sequences constitutes a substantial modeling limitation. It restricts models to consider only proteins that are relatively similar, preventing generalization beyond single protein families. Furthermore, for proteins with few known relatives, it can be expensive or even impossible to find homologous sequences, making it difficult to find patterns for proteins that are functionally similar but divergent in sequence. *Sequential* models were proposed as a potential remedy to these issues. The first examples were hidden Markov models (HMMs). The HMM-framework provides efficient algorithms for evaluating conditional probabilities, for sampling, and for calculating most probable representations (state sequences), making it an attractive model class. Initially, these models were employed for single protein families [15], where they defined discrete states for the amino acid composition at each position and for insertions and deletions. HMMs were since developed for broader families such as transmembrane proteins and proteins with signal peptides (see, e.g. Ref. [16], for a review). In later work, the predefined state connectivity was replaced with learned transition probabilities, making it possible to scale to larger state spaces and allowing for expressive probabilistic models of local protein sequence and structure [17,18]. As the number of discrete states in these models increased, they became increasingly decoupled from predefined interpretations and more reminiscent of the abstract learned representations of current foundation models.

The Markov property[3] of HMMs is both a blessing and a curse. On one hand, the assumption results in a flexible and convenient model class providing grammatical rigor, interpretability, and feasibility for small datasets. On the other hand, it places strong limitations on the expressivity of the model. The rise of deep learning brought ways to expand the contextual information. Early methods such as ProtVec [19] directly predicted amino acid propensities in the context of a surrounding *k*-mer, and later, recurrent neural networks such as UniRep [20] provided consistent models for entire sequences, replacing HMMs as the preferred sequential modeling architecture. A further expansion of context was seen in models that conditioned on 3D structure; early efforts in this direction used convolutional neural networks of the structural environment around a site to predict its amino acid propensities [21,22] This was later extended to entire sequences in *inverse folding* models [23,24]. Common to these methods was that they were trained on large databases of diverse proteins. This made it possible to consider the internal vectors within these models as universal protein *representations*, which could be used as input when training prediction algorithms for downstream tasks. With these developments, it became natural to invest compute resources to *pretrain* a model once, and then make it available as a resource to the

[2] Although masked language modeling does not trivially correspond to a maximum likelihood over protein sequences, it has been shown to be an approximation [11].

[3] The Markov property dictates that the context at one position in the chain can only depend on the context at the previous position.

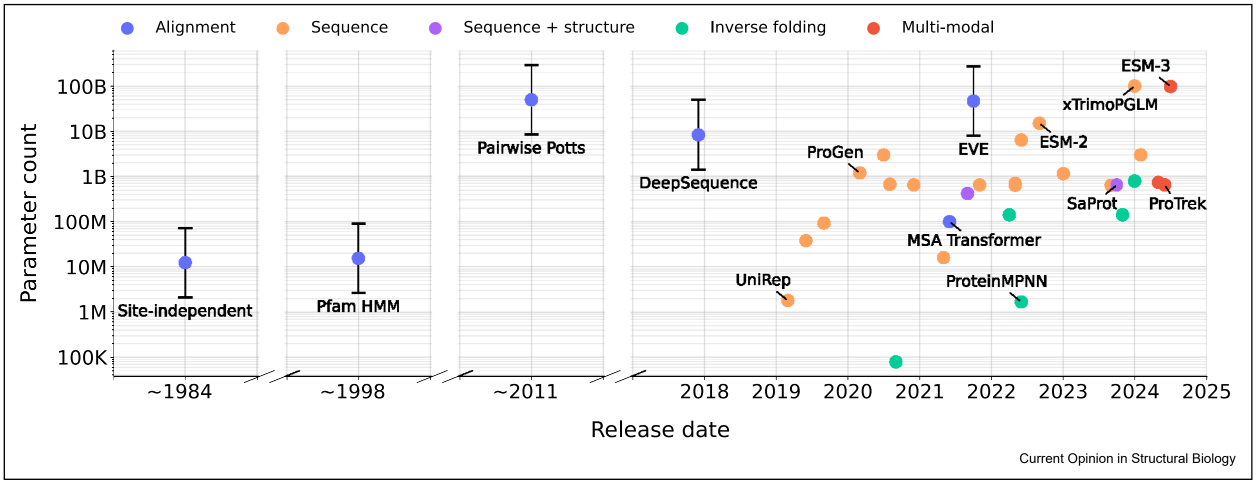
Figure 1


Sequence space
Multiple sequence alignment
Protein family
Structure space
Other modalities
EC:3.5.2.6
TEM-type are the most prevalent beta-lactamases in enterobacteria; they hydrolyze the beta-lactam...
Alignment-based models
p( )
Site independent (PSSM)
Pairwise models (Potts)
Latent variable models
DeepSequence, Riesselman et al., 2018
EVE, Frazer et al., 2021
Multi-modal models
p(|)
ZymCTRL, Munsamy et al., 2022
ProteinDT, Shengchao et al., 2024
ESM-3, Hayes et al., 2024
ProTrek, Su et al., 2024
p( )
MSA Transformer, Rao et al., 2021
Sequence models
p( )
Hidden Markov models
Autoregressive language modeling
Masked language modeling
Krogh et al., 1994
Hamelryck et al., 2006
Boomsma et al., 2008
UniRep, Alley et al., 2019
ESM-1, Rives et al., 2019
ProteinBERT, Brandes et al., 2022
ProtTrans, Elnaggar et al., 2022
Retrieval, scaling, optimization
Tranception, Notin et al., 2022
ProGen2, Nijkamp et al., 2023
Ankh, Elnaggar et al., 2023
xTrimoPGLM, Chen et al., 2024
Sequence diffusion
DPLM, Wang et al., 2024
Sequence and structure modeling
p(,)
Sequence and structure fusion
LM-GVP, Wang et al., 2023
ESM-GearNet, Zhang et al., 2023
Structure diffusion
RFdiffusion, Watson et al., 2023
Chroma, Ingraham et al., 2023
Co-design
Protpardelle, Chu et al., 2023
ProteinGenerator, Lisanza et al., 2023
Multiflow, Campbell et al., 2024
Structure tokenization
SaProt, Su et al., 2024
ESM-3, Hayes et al., 2024
Inverse folding
p(|)
Convolutions
Boomsma & Frellsen, 2017
Torng & Altman, 2017
Graph Neural Networks
Structured Transformer, Ingraham et al., 2019
GVP-GNN, Jing et al., 2020
ESM-IF1, Hsu et al., 2022
ProteinMPNN, Dauparas et al., 2022
Sequence transfer
KW-Design, Gao et al., 2023
MIF-ST, Yang et al., 2023
Recycling
CarbonDesign, Ren et al., 2024
+

Current Opinion in Structural Biology

Conceptual overview of the historical progression in protein modeling approaches, categorized by the targeted data distributions and the context on which they are conditioned.

Figure 2



Current Opinion in Structural Biology

Comparison of protein models based on parameter count. To facilitate comparison, we imagine a scenario where family-specific, alignment-based models are trained on all known protein families, thus constituting the equivalent of a foundation model. In doing so, we assume an average protein length of 200 amino acids and a number of families as given by Pfam at 1998 (lower bar) and 2021 (upper bar). This illustrates that non-Markovian alignment-based models like the Potts model and the DeepSequence model are as "overparameterized" as many current protein language models.

community, thus laying the groundwork for the notion of a *foundation* model.

Foundation models of protein sequences

The last few years have produced a wealth of new self-supervised models of proteins. While potentially overwhelming, they can generally be understood as natural extensions of the earlier trends presented above. We will describe the main directions below and refer to Figure 1 for an overview.

Protein sequences as sentences

Protein language models (pLMs) have emerged from developments in the rich natural language processing literature and have proved to be both powerful and flexible alternatives to traditional statistical approaches. Considering protein sequences as sentences and individual amino acids as words, these models are typically tasked with either predicting amino acids masked at random or causally masked.[4] Compared with the UniRep model described above, the main developments concern the choice of neural network architecture. While UniRep summarizes the amino acid sequence into a single representation, transformer-based models specifically refer to other parts in the sequence through an *attention* mechanism. Notable early transformer-based pLMs include ProtTrans [25], ProteinBert [26], and the first iteration of the ESM series of models [3,27,28]. These demonstrate impressive capabilities across multiple domains, from capturing biophysical features to encoding remote homology and predicting protein properties and variant effects.

Subsequent advances for protein language models include training on larger and more diverse sets of protein sequences and increasing model capacity as visualized from the sequence models of Figure 2. Examples are ESM-2 [1] and ProGen2, where the latter is also trained on sequences from the BFD metagenomic database in addition to non-redundant sets from UniProtKB [2]. Pushing the envelope of model capacity, recent protein language models with up to 100 billion parameters are being introduced, such as xTrimoPGLM [29] and ESM-3 [30], achieving improved performance across various discriminative and generative tasks.

As a complement to these high-capacity models, several recent works [31—34] examine the training methodology of pLMs and provide guidance on including biologically aligned modeling assumptions (*inductive biases*), whereas others investigate compute-optimality [35,36]; Ankh [37] notably provided state-of-the-art results while decreasing model size.

Other innovations include language models over entire multiple sequence alignments with the MSA transformer [28], inference-time retrieval for improved zero-shot fitness predictions with Tranception [38], and more recently, sequences-of-sequences modeling of whole protein families with retrieval-augmented capabilities with PoET [39]. To handle long sequences, ProtHyena [40] and linear state-space approaches like PTM-Mamba [41] aim to develop efficient alternatives to attention. While these models currently do not reach

[4] Auto-regressive models typically employ a causal masking strategy, where the model is trained to predict the next sequence element, conditioned on preceding ones.

the state-of-the-art in established protein evaluation settings, they may prove essential for scaling to longer proteins or other biological sequences like DNA [42].

Finally, after diffusion-based models have become ubiquitous in the image domain (and in the protein 3D structure domain—see below), we now see such models applied to protein sequences [43–45]. While diffusion models display competitive performance in generation, they generally lack explicit extraction of representations, currently making them less generally applicable as foundation models. However, recent models such as DPLM [46] have proposed potential solutions to this problem, and we are likely to see further exploration of this model class in the immediate future.

From structure to sequence

With the advent of AlphaFold2 [47], the availability of structural data has sharply increased, converting structures from a scarce data modality into a standard feature for protein modeling. As a consequence, structure-based self-supervised learning can now be conducted on a scale similar to that of sequence-based models. Many models follow the inverse folding strategy, letting structure serve as the context for sequence prediction, including ProteinMPNN [48] and ESM-IF1 [49]. Newer instances of inverse folding models such as KW-Design [50] and MIF-ST [51] incorporate pretrained pLMs for improved performance, while others use a recycling method similar to that of AlphaFold2 for refining sequence generation, e.g. CarbonDesign [52].

How best to combine sequence and structure is an open area of research [53–57]. Recent models [30,58,59] incorporate 3D structure by expanding the amino acid vocabularies with structural tokens such as those of FoldSeek's 3Di vocabulary [60]. A notable example is the SaProt model [59], which is the current state-of-the-art for both zero-shot predictions and supervised training of downstream tasks.

Joint sequence–structure models

Constructing generative models of 3D protein structure was long considered a highly challenging problem, due to the many angular and distance constraints in such systems. Diffusion models have in the last few years proven themselves as an elegant solution to this problem, making it possible to generate coherent protein backbone structures. When combined with a pretrained inverse folding model, this provides a path toward *joint* distribution of protein sequence and structure. Prominent early examples of this approach include the Chroma [61] and RFdiffusion [62] models.

More recent methods in *de novo* protein generation jointly generate structure and sequence (coined "co-design"). This emerging area of research has produced a number of recent contributions. One example is ProteinGenerator [63] which offers flexible sequence conditioning as a natural complement to the structure guidance capabilities of models like Chroma and RFdiffusion by gradually denoising the sequence and iteratively predicting the structure with a folding component. Other examples are Protpardelle [64] that denoise the structure and iteratively predicts the sequence with an inverse folding model, while Multiflow [65] combines them with a discrete flow-based approach.

Multi-modal protein models

New modalities beyond primary and tertiary structures are increasingly being explored, such as the combination of protein language modeling with knowledge graphs of gene ontologies [66,67]. Other approaches utilize alternate modalities for conditioning, for example ZymCTRL [68] that directly conditions on enzyme commission numbers for guided generation. Several recent models combine protein language models with sequence- and residue-level annotations from databases like Swiss-Prot [69] for additional context during training and inference. These methods often utilize separate text encoders like SciBERT [70] to embed the textual protein descriptions. The embeddings are subsequently combined with protein sequence representations either via contrastive approaches such as in ProtST [71], ProteinDT [72], and ProTrek [73] or protein language modeling frameworks such as in PAIR [74] and Prot2Text [75]. Inspired by current trends in multi-modal language processing, the recent ESM-3 model [30] fuses distinct modality tracks into a single latent space, providing rich representations while maintaining a high degree of flexibility.

Multi-modal foundation models belong to a nascent model category and have yet to be tested as extensively and rigorously as earlier models, in part due to the variety of novel problems they aim to solve. However, they have already begun to show competitive results across diverse tasks and will likely dominate in the coming years.

Discussion and future perspectives

Changing perspective from representations to foundation models

The focus of the community is shifting from *representations* to the *foundation models* that harbor them as illustrated in the overview above. Five years ago, embeddings from pretrained models were envisioned as *universal* representations of proteins [20], and the ubiquitous applicability of representations has indeed been confirmed by the wealth of openly available foundation models and downstream tools. However, rather than static representations of proteins, it seems more fruitful today to think of them as baselines or priors that can be optimized further for downstream tasks and specific subdomains. This trend is particularly evident in supervised transfer learning, where fine-tuning methods have become the *de facto* standard, showing significant quantitative improvements when compared with fixed embeddings [4–6]. The likelihoods and generative capabilities of foundation models themselves are also increasingly used as a goal in itself [61,62] rather than as a means to obtain an informative representation.

Despite the focus on foundation models, we should not understate the usefulness of the notion of a protein representation. Even static representations will remain of interest as low-cost, strong baselines for downstream predictions tasks, and they can provide functionality that is difficult to obtain otherwise, e.g. for remote homology detection or interpretability through visualization. However, if we are to take representations seriously, it could require actively considering the properties that we desire from them, rather than extracting them in an *ad hoc* fashion from the layers of a foundation model. We have previously argued that it might be beneficial to construct representations that support a predefined set of useful operations [76], which relates to discussions of disentangled and decomposed representations [77]. As a field, it would be useful to consider which features in protein space can be meaningfully disentangled.

Managing richer contexts

Although current foundation models encompass many aspects of proteins, they are not yet *complete* descriptions of proteins. There are multiple features annotated in databases that are not yet systematically incorporated into the models. Examples include posttranslational modifications or information about binding partners. We begin to see such features incorporated in pretrained models [41] and anticipate many more models with rich combinations of modalities. It has been intriguing to observe that *predictions* of 3D structure have been placed on almost equal footing as experimental data for the purpose of training foundation models [30,59]. It will be interesting to see whether a similar approach will be helpful for other modalities that are expensive to probe experimentally. One example could be structural dynamics, which is currently largely ignored in foundation models, but for which MD simulations could provide simulated data. In particular for multi-conformer proteins, providing the full thermodynamical ensemble as a context for the amino acid propensity will likely prove beneficial.

The role of scaling

Protein foundation models are well-known for their vast number of parameters. However, it is important to note that this is primarily a consequence of the ambition to model all protein families in a single model, and that earlier family-specific models would have comparable parameters counts if they were trained on all families (Figure 2). Nevertheless, we have seen that the advances in foundation models over the last years have been closely related to scaling. This growth goes beyond what can be explained by the simultaneous exponential growth in the size of sequence databases, and one might reasonably ask if we can expect these trends to continue.

The answer is not clear as the relationship between model scaling and its impact on model performance remains underexplored. It has been shown that test set perplexity, a measure of a protein language model's ability to describe the underlying data distribution of protein sequences, systematically improves with model scaling [1,2,78]. While the resulting higher fidelity representations can lead to improved downstream performance for supervised tasks [6,25,78], the opposite has been observed in the unsupervised setting, where smaller, less flexible models tend to outperform their scaled counterparts [2,79]. A recent study further showed that downstream performance in both supervised and unsupervised settings is uncoupled from pretraining efforts [80], emphasizing the need for novel pretraining methods.

While the largest models currently available do still show benefits over earlier models, such as the increasing emergence of structure [1], there are signs that we are entering a phase of diminishing returns from scaling alone—even in terms of perplexity—as demonstrated in recent works on compute-optimal pLMs [35,36]. To obtain substantial leaps in performance in the coming years, we will therefore likely require dramatic gains in data availability, entirely new modalities, or architectures that more explicitly incorporate our biological priors. As scaling continues, the rising computational costs of fine-tuning and inference poses a potential challenge to accessibility and utility. It is crucial that we continue to develop and adapt efficient fine-tuning schemes [5] and that models are released at various capacities to ensure that research and innovation are not limited by compute access.

In summary

The field of protein sequence modeling has transitioned from explainable and simple statistical models to large-scale self-supervised foundation models. To effectively accommodate varied downstream tasks, there has been a shift from using static embeddings to applying efficient fine-tuning schemes for increased utility and performance. How best to manage even richer contexts, e.g. via the incorporation of molecular dynamics or functional annotations, remains an open challenge and could pave the way for an even wider range of real-world applications. High-quality, multi-modal data will be essential in training and evaluating these increasingly capable models—as well as taking a step back and re-evaluating which modeling techniques and objectives fit proteins best.

Declaration of competing interest

None.

Acknowledgements

This work was funded by the Novo Nordisk Foundation (NNF) through the MLLS Center (Basic Machine Learning Research in Life Science, NNF20OC0062606) and the Pioneer Centre for AI (DNRF grant number P1). The research was further supported by Innovation Fund Denmark (1044-00158A; PMG, WB), the Villum Foundation (40526; AK, WB), and the Novo Nordisk Foundation (NNF20OC0059939 and NNF20OC0063268; AK).

Data availability

No data was used for the research described in the article.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
* * of outstanding interest

1. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A: **Evolutionary-scale prediction of atomic-level protein structure with a language model**. *Science* 2023, **379**:1123–1130.
* *

Introduces the ESM-2 protein language model (and the ESMFold structure predictor). Slight upscaling from predecessor ESM-1b with model sizes range from 8M to 15B, thereby providing a large degree of flexibility, which together with the repository's ease-of-use has led to ESM-2 becoming the arguably most-used protein language model prior to archivation in August 2024.

2. Nijkamp E, Ruffolo JA, Weinstein EN, Naik N, Madani A: **ProGen2: exploring the boundaries of protein language models**. *Cell syst* 2023, **14**:968–978.
*

ProGen2 is an autoregressive protein language model trained at different model sizes and on different datasets to investigate both scaling and the impact of training distributions. For larger model sizes, it was observed that the test perplexity decreased with model size, i.e., that the larger models were better able to capture the training data distributions. However, the correlation between likelihood scores and experimental assays were observed to decrease for larger models, indicating a mismatch between training objective and downstream performance. This latter result stands in contrast to RITA (Hesslow et al., 2022), where both perplexity and zero-shot fitness predictions increased with model capacity.

3. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A: **Language models enable zero-shot prediction of the effects of mutations on protein function**. *Adv Neural Inf Process Syst* 2021, **34**: 29287–29303.

4. Zhou Z, Zhang L, Yu Y, Wu B, Li M, Hong L, Tan P: **Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning**. *Nat Commun* 2024, **15**:5566.
*

Introduces the Few-Shot Learning for Protein Prediction (FSFP) training strategy for transferring pre-trained pLMs to low-N fitness predictions tasks via model-agnostic meta-learning. The models are meta-trained on supervised ranking tasks similar to the task of interest, where LoRA is applied to both reduce overfitting and the computational burden. After meta-training, the model is transferred to the few-shot learning task of predicting variant effects with improved performance.

5. Sledzieski S, Kshirsagar M, Baek M, Dodhia R, Lavista Ferres J, Berger B: **Democratizing protein language models with parameter-efficient fine-tuning**. *Proc Natl Acad Sci USA* 2024, **121**, e2405840121.
*

Parameter-efficient fine-tuning (LoRA) is used in two downstream tasks: predicting protein–protein interactions, and predicting the symmetry of homooligomer quaternary structures. While requiring substantially fewer parameters, their method is competitive with traditional fine-tuning. The authors lay important groundwork for best practices of efficient fine-tuning in proteomic modeling, aiming to make an easily applicable blueprint for groups with limited resources.

6. Schmirler R, Heinzinger M, Rost B: **Fine-tuning protein language models boosts predictions across diverse tasks**. *Nat Commun* 2024, **15**:7407.
*

Parameter-efficient fine-tuning (specifically LoRA) is extensively applied to a set of state-of-the-art protein language models (Ankh, ESM-2, and ProtT5) on a variety of downstream tasks, consistently resulting in relatively low-cost yet significant performance improvements.

7. Notin P, Rollins N, Gal Y, Sander C, Marks D: **Machine learning for functional protein design**. *Nat Biotechnol* 2024, **42**: 216–228.
*

In this review article, the authors establish a unifying framework for machine learning for protein design, which categorizes machine learning models based on their use of three data modalities: sequences, structures, and functional labels. The article includes numerous references to related works with an emphasis on application, e.g., enzyme and antibody design. The review provides an extensive overview of this core application while outlining future directions and challenges.

8. Bepler T, Berger B: **Learning protein sequence embeddings using information from structure**. In *International conference on learning representations*; 2019.

9. Heinzinger M, Littmann M, Sillitoe I, Bordin N, Orengo C, Rost B: **Contrastive learning on protein embeddings enlightens midnight zone**. *NAR genomics and bioinformatics* 2022, **4**:lqac043.

10. Singh R, Sledzieski S, Bryson B, Cowen L, Berger B. In *Contrastive learning in protein language space predicts interactions between drugs and protein targets*, **120**. Proceedings of the National Academy of Sciences; 2023, e2220778120.

11. Moreno-Muñoz P, Recasens PG, Hauberg S: **On masked pre-training and the marginal likelihood**. In *Neural information processing systems*. NeurIPS; 2023.

12. Krogh A: **An introduction to hidden Markov models for biological sequences**. In *New comprehensive biochemistry*, **32**. Elsevier; 1998:45–63.

13. Hopf TA, Schärfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, Bonvin AM, Marks DS: **Sequence co-evolution gives 3D contacts and structures of protein complexes**. *Elife* 2014, **3**, e03430.

14. Riesselman AJ, Ingraham JB, Marks DS: **Deep generative models of genetic variation capture the effects of mutations**. *Nat Methods* 2018, **15**:816–822.
*

DeepSequence is a variational autoencoder with a multi-layer perceptron backbone, trained on multiple sequence alignments of protein families. The authors show how latent variable models with nonlinear, high-order residue dependencies obtain significantly better mutational effect predictions compared to previous site-independent and pairwise models. As such, DeepSequence became one of the earliest examples of learning neural representations of sequences, and provided an easily accessible framework to apply such latent variable models in other analyses.

15. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D: **Hidden Markov models in computational biology: applications to protein modeling**. *J Mol Biol* 1994, **235**:1501–1531.

16. Bystroff C, Krogh A: **Hidden Markov models for prediction of protein features**. In *Protein structure prediction*, **413**. Springer: Humana Press; 2007:173–198.

17. Hamelryck T, Kent JT, Krogh A: **Sampling realistic protein conformations using local structural bias**. *PLoS Comput Biol* 2006, **2**, e131.

18. Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T: **A generative, probabilistic model of local protein structure**. *Proc Natl Acad Sci USA* 2008, **105**:8932–8937.

19. Asgari E, Mofrad MR: **Continuous distributed representation of biological sequences for deep proteomics and genomics**. *PLoS One* 2015, **10**, e0141287.

20. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM: **Unified rational protein engineering with sequence-based deep representation learning**. *Nat Methods* 2019, **16**: 1315–1322.

21. Boomsma W, Frellsen J: **Spherical convolutions and their application in molecular modelling**. *Adv Neural Inf Process Syst* 2017, **30**.

22. Torng W, Altman RB: **3d deep convolutional neural networks for amino acid environment similarity analysis**. *BMC Bioinf* 2017, **18**:1–23.

23. * Ingraham J, Garg V, Barzilay R, Jaakkola T: **Generative models for graph-based protein design**. *Adv Neural Inf Process Syst* 2019, **32**.

Seminal paper in applying machine learning to the inverse folding problem. The Structured Transformer consists of a GNN-based structure encoder and a GNN-based autoregressive decoder, utilizing self-attention and causal self-attention, respectively. Both the overall architecture and graph featurization have been subsequently used in other notable works, such as ProteinMPNN, with a number of changes.

24. Jing B, Eismann S, Suriana P, Townshend RJL, Dror R: **Learning from protein structure with geometric vector perceptrons**. In *International conference on learning representations (ICLR)*; 2021.

25. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B: **ProtTrans: toward understanding the language of Life through self-supervised learning**. *IEEE Trans Pattern Anal Mach Intell* 2021, **44**:7112–7127.

26. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M: **ProteinBERT: a universal deep-learning model of protein sequence and function**. *Bioinformatics* 2022, **38**:2102–2110.

27. Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A: **Transformer protein language models are unsupervised structure learners**. In *International conference on learning representations*; 2021. https://openreview.net/forum?id=fylclEqgvgd; 2021.

28. Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, Sercu T, Rives A: **MSA transformer**. In *International conference on machine learning*. PMLR; 2021:8844–8856.

29. Chen B, Cheng X, Li P, Geng Y-a, Gong J, Li S, Bei Z, Tan X, Wang B, Zeng X, Liu CL, Zeng A, Dong Y, Tang J, Song L: **xTrimoPGLM: unified 100B-scale pre-trained transformer for deciphering the language of protein**. *arXiv preprint arXiv: 2401.06199* 2024.

30. ** Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M, Badkundri R, Shafkat I, Gong J, Derry A, Molina RS, Thomas N, Khan YA, Mishra C, Kim C, Bartie LJ, Nemeth M, Hsu PD, Sercu T, Candido S, Rives A: **Simulating 500 million years of evolution with a language model**. *Science* 2025, **0**, eads0018, https://doi.org/10.1126/science.ads0018.

Introduces the ESM-3 suite of multi-modal protein language models (sizes 1.4B, 7B, 98B) with separate token tracks for structure, sequence, and function, effectively resulting in a joint distribution over the modalities. Structures are tokenized using a vector-quantized VAE, whereby each residue is associated with one of 4096 tokens. ESM-3 uses functional annotations at both residue and sequence levels, derived from InterPro. Synthetic sequences from inverse folding increase the training data from 2.78B sequences to 3.15B sequences, 236M protein structures, and 539M proteins with function annotations, totaling 771B tokens. ESM-3 can follow complex prompts with a combination of partial inputs across modalities.

31. * Bepler T, Berger B: **Learning the protein language: evolution, structure, and function**. *Cell syst* 2021, **12**:654–669.

Comprehensive synthesis of the state of protein language models at the time of publication, with perspectives on future directions. Some of the raised points, such as the need for models combining sequence and structure, are currently receiving much interest from the field, while others, such as the inclusion of biophysical priors have yet to become core components in foundation model architectures and methods. Many of the raised challenges of aligning methods originally developed to handle natural text to biological data persist.

32. Vu MH, Akbar R, Robert PA, Swiatczak B, Sandve GK, Greiff V, Haug DTT: **Linguistically inspired roadmap for building biologically reliable protein language models**. *Nat Mach Intell* 2023, **5**:485–496.

33. * Valeriani L, Doimo D, Cuturello F, Laio A, Ansuini A, Cazzaniga A: **The geometry of hidden representations of large transformer models**. *Adv Neural Inf Process Syst* 2024, **36**.

The authors systematically analyze geometric properties of representations in a state-of-the-art protein language model — intrinsic dimensionality and neighbor composition — and learn that patterns arise when moving through layers or training iterations. They find that large transformers first encode the data into a low-dimensional and abstract representation, and successively decode from it; they find that intermediate representations are the most semantically rich.

34. Gong C, Klivans A, Loy JM, Chen T, Liu Q, Diaz DJ: **Evolution-Inspired loss functions for protein representation learning**. In *Proceedings of the 41st international conference on machine learning, volume 235 of* Proceedings of machine learning research. PMLR; 2024:15893–15906.

35. Serrano Y, Serrano AC, Molina A: **Are Protein Language models compute optimal?**. In *ICML 2024 workshop on efficient and accessible foundation models for biological discovery*; 2024.

36. Cheng X, Chen B, Li P, Gong J, Tang J, Song L: **Training Compute-Optimal Protein Language models, In**. *Adv Neural Inf Process Syst* 2024, **37**:69386–69418.

37. ** Elnaggar A, Essam H, Salah-Eldin W, Moustafa W, Elkerdawy M, Rochereau C, Rost B: **Ankh: optimized Protein Language Model unlocks general-purpose modelling**. *arXiv preprint arXiv:2301.06568* 2023.

The Ankh collection of protein language models focused on optimization and accessibility and reached state-of-the-art performance with fewer parameters than competing methods, showing that considerations such as model implementation and computational costs are important factors to consider rather than only scaling model size.

38. ** Notin P, Dias M, Frazer J, Hurtado JM, Gomez AN, Marks D, Gal Y: **Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval**. In *International conference on machine learning*. PMLR; 2022:16990–17017.

Tranception is an autoregressive protein language model which introduces a biologically motivated attention mechanism operating on k-mers of size 1, 3, 5, and 7. An inference-time retrieval mechanism is used to leverage the evolutionary information encoded in a multiple sequence alignment by altering its zero-shot scores to be a weighted average of log-likelihoods from the model and the MSA-derived empirical distribution of amino acids, resulting in markedly improved results.

39. Truong Jr T, Bepler T: **PoET: a generative model of protein families as sequences-of-sequences**. *Adv Neural Inf Process Syst* 2024, **36**:77379–77415.

40. Zhang Y, Okumura M: **ProtHyena: a fast and efficient foundation protein language model at single amino acid Resolution**. *bioRxiv* 2024, https://doi.org/10.1101/2024.01.18.576206.

41. Peng Z, Schussheim B, Chatterjee P: **PTM-Mamba: a PTM-aware protein language model with bidirectional gated Mamba blocks**. *bioRxiv* 2024, https://doi.org/10.1101/2024.02.28.581983.

42. Marin FI, Teufel F, Horlacher M, Madsen D, Pultz D, Winther O, Boomsma W: **BEND: benchmarking DNA Language Models on biologically meaningful tasks**. In *The twelfth international conference on learning representations*; 2023.

43. Alamdari S, Thakkar N, van den Berg R, Lu A, Fusi N, Amini A, Yang K: **Protein generation with evolutionary diffusion**. In *NeurIPS 2023 generative AI and biology (GenBio) workshop*; 2023.

44. Gruver N, Stanton S, Frey N, Rudner TG, Hotzel I, Lafrance-Vanasse J, Rajpal A, Cho K, Wilson AG: **Protein design with guided discrete diffusion**. *Adv Neural Inf Process Syst* 2024, **36**.

45. Liu G, Wang Y, Feng Z, Wu Q, Tang L, Gao Y, Li Z, Cui S, Mcauley J, Yang Z, Xing EP, Hu Z: **Unified generation, reconstruction, and representation: generalized diffusion with adaptive latent encoding-decoding**. In *Proceedings of the 41st international conference on machine learning, volume 235 of* Proceedings of machine learning research. PMLR; 2024: 31964–31993.

46. ** Wang X, Zheng Z, Ye F, Xue D, Huang S, Gu Q: **Diffusion Language models are versatile protein learners**. In *Proceedings of the 41st international conference on machine learning, volume 235 of* Proceedings of machine learning research. PMLR; 2024:52309–52333.

Introduces the Diffusion Protein Language Model (DPLM) based on discrete diffusion, enabling strong representations and generation simultaneously. Matches the state-of-the-art structure-aware SaProt in downstream performances using learned representations.

47. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**: 583–589.

48. * * Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N, Leung PJY, Huddy TF, Pellock S, Tischer D, Chan F, Koepnick B, Nguyen H, Kang A, Sankaran B, Bera AK, King NP, Baker D: **Robust deep learning–based protein sequence design using ProteinMPNN**. *Science* 2022, **378**:49–56.

Introduces ProteinMPNN, a widely used model for inverse protein folding. It was introduced with outstanding performances in both in-silico and experimental tests, emphasizing the validity of deep learning-based protein design. Since its introduction in 2022, the ProteinMPNN framework has been used extensively in pipelines which concentrate on generating de-novo proteins, improving the efficacy of these works (e.g., RFdiffusion (Watson et al., 2023)).

49. * * Hsu C, Verkuil R, Liu J, Lin Z, Hie B, Sercu T, Lerer A, Rives A: **Learning inverse folding from millions of predicted structures**. In *International conference on machine learning*. PMLR; 2022:8946–8970.

The ESM-IF1 (ESM-inverse folding 1) model was trained on a subset of CATH which was expanded with millions of predicted protein structures from UniRef50. The model utilizes geometric vector perceptrons \citep {jing2020learninggvp} for structure encoding with either GVP-GNN or generic autoregressive encoder-decoder transformer layers for sequence decoding. In addition to its primary functionality as a structure-conditioned sequence generator, zero-shot estimates from ESM-IF have been found to correlate highly with protein stability.

50. Gao Z, Tan C, Chen X, Zhang Y, Xia J, Li S, Li SZ: **KW-design: pushing the limit of protein design via knowledge refinement**. In *The twelfth international conference on learning representations*; 2024.

51. Yang KK, Zanichelli N, Yeh H: **Masked inverse folding with sequence transfer for protein representation learning**. *Protein Eng Des Sel* 2023, **36**:gzad015.

52. * Ren M, Yu C, Bu D, Zhang H: **Accurate and robust protein sequence design with CarbonDesign**. *Nat Mach Intell* 2024, **6**: 536–547.

CarbonDesign leverages several concepts from AlphaFold2 such as recycling to achieve state-of-the-art performance in inverse folding, surpassing ESM-IF1 and ProteinMPNN. Its backbone consists of several novel InverseFormer layers, followed by a Markov random field sequence decoder. During recycling, the pre-trained ESM-2 model is used to incorporate evolutionary constraints.

53. Wang Z, Combs SA, Brand R, Calvo MR, Xu P, Price G, Golovach N, Salawu EO, Wise CJ, Ponnapalli SP, Clark PM: **LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction**. *Sci Rep* 2022, **12**:6832.

54. * Zhang Z, Wang C, Xu M, Chenthamarakshan V, Lozano A, Das P, Tang J: **A systematic study of joint representation learning on protein sequences and structures**. *arXiv preprint arXiv: 2303.06275* 2023.

A systematic study of how to combine sequence and structure signals, where a recent protein language model (ESM-2) is combined with structure-encoding models (GVP, GearNet, and CDConv) through various fusion methods. The results show benefits from combining signals from both modalities by achieving state-of-the-art results on a number of downstream tasks.

55. Zheng Z, Deng Y, Xue D, Zhou Y, Ye F, Gu Q: **Structure-informed language models are protein designers**. In *International conference on machine learning*. PMLR; 2023:42317–42338.

56. Wang X, Yin X, Jiang D, Zhao H, Wu Z, Zhang O, Wang J, Li Y, Deng Y, Liu H, Luo P, Han Y, Hou T, Yao X, Hsieh C-Y: **Multi-modal deep learning enables efficient and accurate annotation of enzymatic active sites**. *Nat Commun* 2024, **15**:7348.

57. Groth PM, Kerrn MH, Olsen L, Salomon J, Boomsma W: **Kermut: Composite kernel regression for protein variant effects**. *Adv Neural Inf Process Syst* 2024, **37**:29514–29565.

58. * * Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Mirdita M, Steinegger M, Rost B: **Bilingual Language model for protein sequence and structure**. *NAR Genomics and Bioinformatics* 2024, **6**.

ProstT5 is a protein language model which incorporates structural 3Di tokens from Foldseek. The model is trained as a ``translator" which allows for either generating sequence from structure, or structure from sequence, i.e., by converting between amino acid tokens and 3Di tokens. It is emphasized that ProstT5 is not a general purpose pLM and is subject to information loss due to catastrophic forgetting during fine-tuning of the protein-encoding ProtT5 model. It instead serves as a proof-of-concept of the viability of leveraging a structure-encoding alphabet.

59. * * Su J, Han C, Zhou Y, Shan J, Zhou X, Yuan F: **SaProt: protein Language modeling with structure-aware vocabulary**. In *The twelfth international conference on learning representations*; 2024.

SaProt is a general purpose protein language model which uses a structure-aware vocabulary by leveraging 3Di tokens from Foldseek. Various pre-training setups are investigated, resulting in a masking strategy, where only residue tokens (and not structure tokens) are corrupted and recovered. SaProt is applied to a range of both supervised and unsupervised tasks, achieving state-of-the-art results in both settings.

60. * Van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CL, Söding J, Steinegger M: **Fast and accurate protein structure search with Foldseek**. *Nat Biotechnol* 2024, **42**: 243–246.

Application of vector-quantized VAEs to describe protein structure as sequences over a structural alphabet named 3Di. Rather than describing the protein backbone, the vocabulary describes tertiary interactions. While the authors use this to create a highly efficient database for looking up structures, this work has had a large impact with their tokenized structure proxies used in following multi-modal works like ProstT5 (Heinzinger et al., 2023), SaProt (Su et al., 2024), and ESM-3 (Hayes et al., 2024).

61. * Ingraham JB, Baranov M, Costello Z, Barber KW, Wang W, Ismail A, Frappier V, Lord DM, Ng-Thow-Hing C, Van Vlack ER, Tie S, Xue V, Cowles SC, Leung A, Rodrigues JV, Morales-Perez CL, Ayoub AM, Green R, Puentes K, Oplinger F, Panwar NV, Obermeyer F, Root AR, Beam AL, Poelwijk FJ, Grigoryan G: **Illuminating protein space with a programmable generative model**. *Nature* 2023, **623**:1070–1078.

Chroma is a flexible and programmable generative model of protein structure and sequence. Using a diffusion process, a structure is generated, which subsequently is used by a design network to generate a protein sequence via inverse folding. In addition to novel design choices in its diffusion model such as its correlated noise process, Chroma allows for conditional structure generation by conditioning on geometric constraints without the need for retraining. To validate the generative capabilities, 310 proteins are generated both unconditionally and conditionally showing that the sampled proteins are highly expressed, fold, and show favorable biophysical properties.

62. * Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, Wicky BIM, Hanikel N, Pellock SJ, Coubert A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres SV, Lauko A, De Bortoli V, Mathieu E, Ovchinnikov S, Barzilay R, Jaakkola TS, DiMaio F, Baek M, Baker D: **De novo design of protein structure and function with RFdiffusion**. *Nature* 2023, **620**:1089–1100.

RFdiffusion uses a diffusion probabilistic model for de novo protein design, capable of generating novel protein backbones with desired structural and functional properties. The model, based on the RoseTTAFold architecture, is trained to iteratively denoise random noise into plausible protein structures, conditioned on user-specified constraints such as binding sites or functional motifs. By integrating the inverse folding model ProteinMPNN (Dauparas et al., 2022) into the pipeline, RFdiffusion translates generated backbones into amino acid sequences — thus being a joint structure and sequence generative model, similarly to Chroma (Ingraham, 2023). The approach has demonstrated the ability to design novel proteins that not only fold as predicted but also exhibit intended biochemical functions upon experimental validation and wet-lab success.

63. Lisanssza SL, Gershon JM, Tipps SWK, Sims JN, Arnoldt L, Hendel SJ, Simma MK, Liu G, Yase M, Wu H, Tharp CD, Li X, Kang A, Brackenbrough E, Bera AK, Gerben S, Wittmann BJ, McShan AC, Baker D: **Multistate and functional protein design using RoseTTAFold sequence space diffusion**. *Nat Biotechnol* 2024, https://doi.org/10.1038/s41587-024-02395-w.

64. Chu AE, Kim J, Cheng L, El Nesr G, Xu M, Shuai RW, Huang P-S: **An all-atom protein generative model**. *Proc Natl Acad Sci USA* 2024, **121**, e2311500121.

65. * Campbell A, Yim J, Barzilay R, Rainforth T, Jaakkola T: **Generative flows on discrete state-spaces: enabling multimodal flows with applications to protein Co-design**. In *Proceedings of the 41st international conference on machine learning, volume 235 of* Proceedings of machine learning research. PMLR; 2024: 5453–5512.

Multiflow is a flow-based generative model operating on continuous and discrete data simultaneously, e.g., protein structure and sequence. Multiflow is built on top of discrete flow models (DFMs), which use continuous-time Markov chains to simulate data-generating probability flows. Combining DFMs with a previous continuous-data flow method (FrameFlow, (Yim, 2023)) resulted in Multiflow. In contrast to Protein-Generator (Lisanza, 2023) and Protpardelle (Chu et al., 2024) — which iteratively generated either sequence or structure and used a prediction model to infer the missing modality — Multiflow operates more directly in the joint space. Their model achieves current state-of-the-art de novo protein generation.

66. Zhang N, Bi Z, Liang X, Cheng S, Hong H, Deng S, Zhang Q, Lian J, Chen H: **OntoProtein: protein pretraining with gene ontology embedding**. In *International conference on learning representations*; 2022.

67. Zhou H-Y, Fu Y, Zhang Z, Cheng B, Yu Y: **Protein representation learning via knowledge enhanced primary structure reasoning**. In *The eleventh international conference on learning representations*; 2023.

68. Munsamy G, Lindner S, Lorenz P, Ferruz N: **ZymCTRL: a conditional language model for the controllable generation of artificial enzymes**. In *NeurIPS machine learning in structural biology workshop*; 2022.

69. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Res* 2003, **31**:365–370, https://doi.org/10.1093/nar/gkg095.

70. Beltagy I, Lo K, Cohan A: **SciBERT: a pretrained language model for scientific text**. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics; 2019:3615–3620, https://doi.org/10.18653/v1/D19-1371.

71. Xu M, Yuan X, Miret S, Tang J: **ProtST: multi-modality learning of protein sequences and biomedical texts**. In *International conference on machine learning*. PMLR; 2023:38749–38767.

72. * Liu S, Li Y, Li Z, Gitter A, Zhu Y, Lu J, Xu Z, Nie W, Ramanathan A, Xiao C, Tang J, Guo H, Anandkumar A: **A text-guided protein design framework**. *arXiv preprint arXiv: 2302.04611* 2023.

ProteinDT merges a pre-trained pLM (ProtBERT) with a pre-trained text-encoder (SciBERT) using a contrastive learning approach to align the embeddings using their ProteinCLAP method. Their Protein-Facilitator maps natural text to a protein representation for improved sequence decoding. The model is trained using protein sequence and text pairs extracted from Swiss-Prot to enrich the latent space.

73. Su J, Zhou X, Zhang X, Yuan F: **ProTrek: navigating the protein universe through tri-modal contrastive learning**. *bioRxiv* 2024, https://doi.org/10.1101/2024.05.30.596740.

74. * Duan H, Skreta M, Cotta L, Rajaonson EM, Dhawan N, Aspuru-Guzik A, Maddison CJ: **Boosting the predictive power of protein representations with a corpus of text annotations**. *bioRxiv* 2024, https://doi.org/10.1101/2024.07.22.604688.

The PAIR model uses a sequence-to-sequence architecture where protein sequences (via a pre-trained protein language model) are encoded to a latent representation, which is then decoded (e.g., via SciBERT) to natural text. During training, pairs of proteins and different configurations of annotations from UniProt are used, which creates a shared latent space. By extracting the latent representations of protein sequences, the model is used for downstream tasks where it shows increased performance in supervised prediction and retrieval tasks using a subset of available annotations, highlighting the potential benefits of incorporating expert-curated text annotations.

75. Abdine H, Chatzianastasis M, Bouyioukos C, Vazirgiannis M: **Prot2Text: multimodal protein's function generation with GNNs and transformers**. In *Proceedings of the AAAI conference on artificial intelligence*, **38**; 2024:10757–10765.

76. Detlefsen NS, Hauberg S, Boomsma W: **Learning meaningful representations of protein sequences**. *Nat Commun* 2022, **13**: 1914.

77. Locatello F, Bauer S, Lucic M, Raetsch G, Gelly S, Schölkopf B, Bachem O: **Challenging common assumptions in the unsupervised learning of disentangled representations**. In *Proceedings of the 36th international conference on machine learning, volume 97 of* Proceedings of machine learning research. PMLR; 2019:4114–4124.

78. * Hesslow D, Zanichelli N, Notin P, Poli I, Marks D: **RITA: a study on scaling up generative protein sequence models**. *arXiv preprint arXiv:2205.05789* 2022.

RITA is a collection of autoregressive protein language models trained at different model sizes. Similar scaling laws as those of natural language processing (NLP) were established, where larger models resulted in lower perplexities. Similarly, the zero-prediction accuracy increased with model capacity which stands in contrast to the results of ProGen2 (Nijkamp et al., 2022).

79. Notin P, Kollasch A, Ritter D, van Niekerk L, Paul S, Spinner H, Rollins N, Shaw A, Orenbuch R, Weitzman R, Frazer J, Dias M, Franceschi D, Gal Y, Marks D: **ProteinGym: large-scale benchmarks for protein fitness prediction and design**. In *Advances in neural information processing systems*, **36**. Curran Associates, Inc.; 2023:64331–64379.

80. ** Li F-Z, Amini AP, Yue Y, Yang KK, Lu AX: **Feature reuse and scaling: understanding transfer learning with Protein Language models**. In *Proceedings of the 41st international conference on machine learning, volume 235 of* Proceedings of machine learning research. PMLR; 2024:27351–27375.

Comprehensive study of scaling in pLMs with the masked language model objective. Experiments are conducted on model size, model depth, and model training to shine light on a number of hypotheses regarding feature reuse, inductive biases, and weight statistics. It is shown that model scaling in general does not benefit zero-shot fitness predictions, nor the performance on downstream tasks even with some amount of fine-tuning, thereby decoupling scaling from downstream performance.