ALIGN

Scaling and Data Saturation in Protein Language Models

Aviv Spinner[1], Erica DeBenedictis[1], Corey M. Hudson[1]
The Align Foundation | alignbio.org

1 Biological data is expensive and extremely useful for modeling sequence to function relationships.

Unlabeled Protein Data

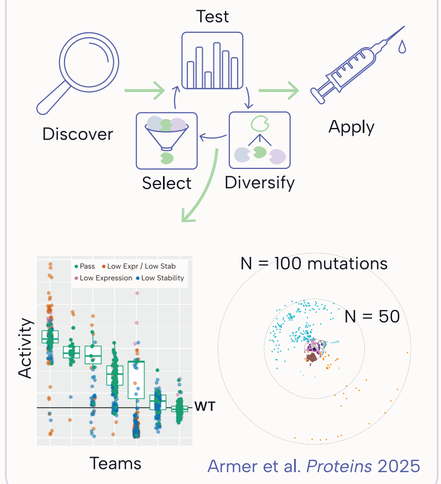
```
>UniRef100_XXXX1
MADEEILALVADNGAGMCKAGFA...
>UniRef100_XXXX2
CRDEEIARLVVDNGSGMCRSSFA...
>UniRef100_XXXX3
MRDEEIVRLVVDNGAGMCEAGFA...
>UniRef100_XXXX4
MRDEEIIRLVADNGAGMCKASGAR...

>UniRef100_YYYY1
MESRADHAVPADEGHGATEPPRG...
>UniRef100_YYYY2
MESRAAHAVPAEEGFGATEPPRGR...

>UniRef100_ZZZZ1
MSSKEAPMRAWLKMAGLAMSNKS...
>UniRef100_ZZZZ2
MSAKDAPMRAFLKMAGLAM
>UniRef100_ZZZZ3
MSSKEAPGLLAYS
>UniRef100_ZZZZ4
MSAKEAPMRAWLKMAGLAMSNKS...
>UniRef100_ZZZZ5
MSAKDAPMRAFLKMSGLAM
>UniRef100_ZZZZ6
MSAKDAPMRAWLKEAGLAMSNKA...
>UniRef100_ZZZZ7
MASKDAPMRAMAGLAMSNKSWLS...
```

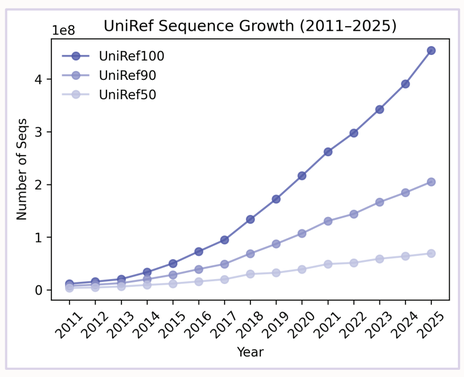
Labeled Protein Data

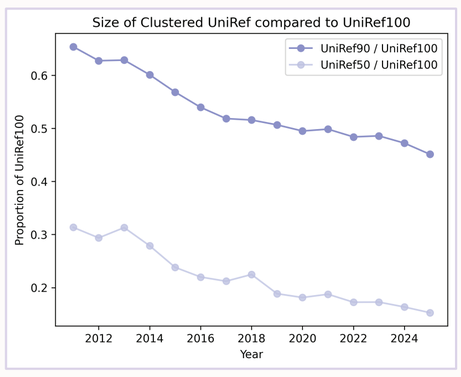
Sequence	Score
MADEEILALVADNGA...	0.123
CRDEEIARLVVDNGS...	1.421
MRDEEIVRLVVDNGA...	0.360
MRDEEIIRLVADNGAG...	0.2
MESRADHAVPADEGH...	1.0
MESRAAHAVPAEEGF...	1.3
MSSKEAPMRAWLKM...	0.744
MSAKDAPMRAFLKM...	0.529
MSSKEAPGLLAYS	0.01
MSAKEAPMRAWLKM...	1.01
MSAKDAPMRAFLKM...	1.322
MSAKDAPMRAWLKE...	0.051
MASKDAPMRAMAGL...	0.916



Armer et al. *Proteins* 2025

2 Biological data is growing at an ever-increasing rate...

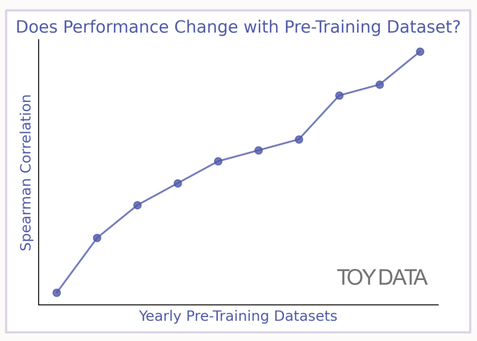


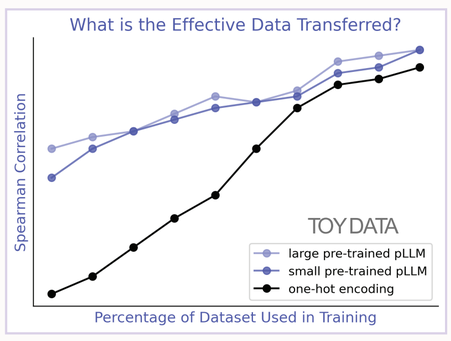


... and the data pose unique challenges

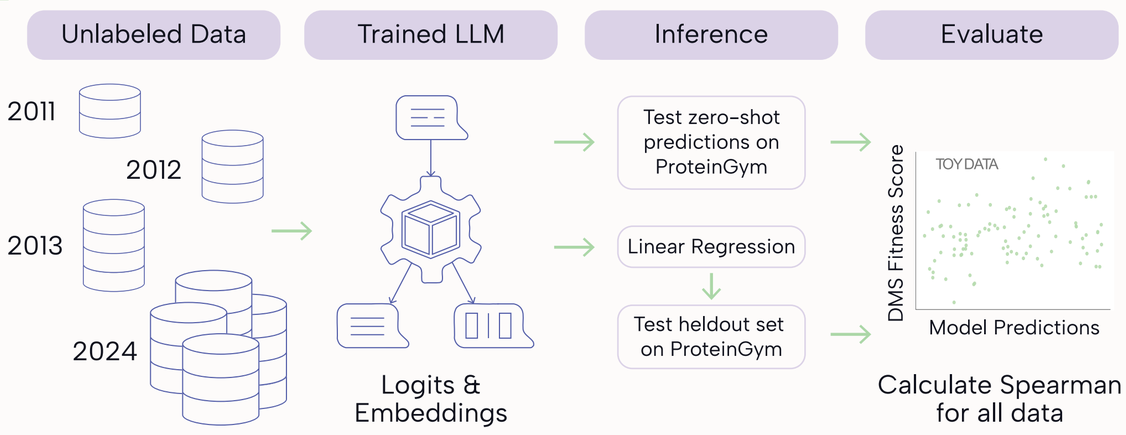
- Redundancy and imbalance
- Annotation sparsity
- Noisy and heterogenous data sources
- Proteins have multiple functions

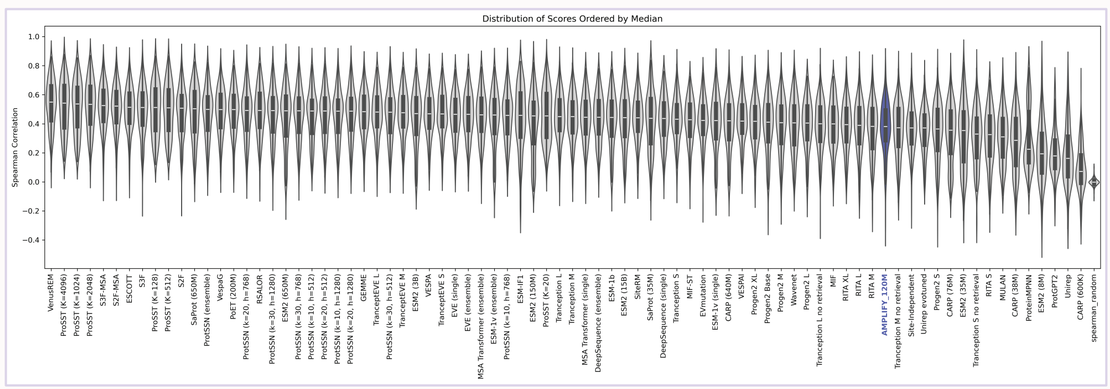
3 How does this data impact the performance of protein language models?





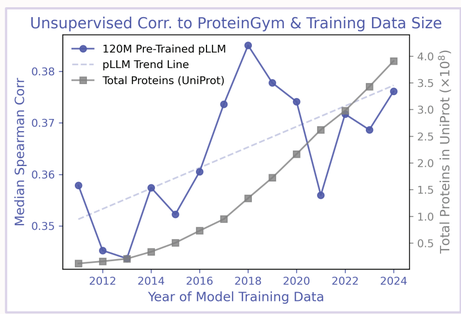
4 We use a suite of AMPLIFY models trained on time points of UniRef100 and evaluated with ProteinGym



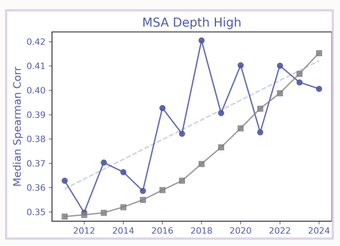


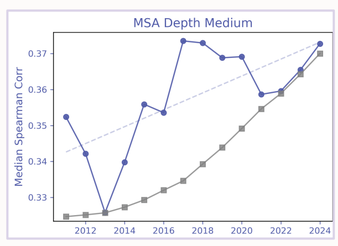
ProteinGym: Notin et al. NeurIPS Datasets and Benchmarks 2023
AMPLIFY: Fournier et al. bioRxiv 2024, https://github.com/chandar-lab/AMPLIFY

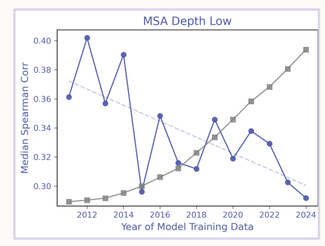
5 More unlabeled sequence data does not yield monotonically better performance



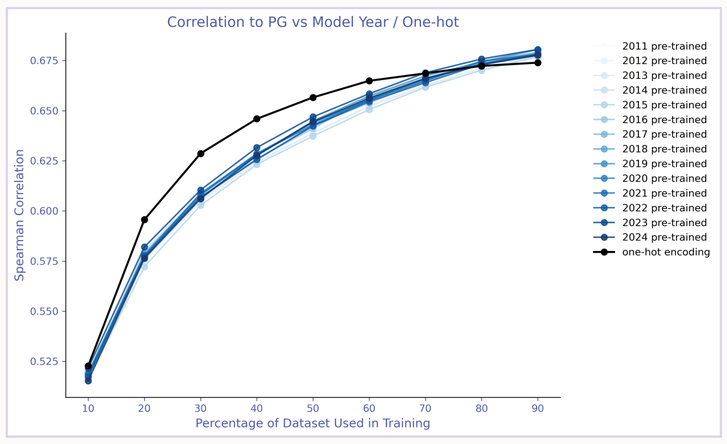
- Model performance is not a function of how many sequences pLM is pre-trained on
- The trend changes depending on how well-represented the protein is in nature
- There is insufficient data to say that we have reached measurable “scaling laws”

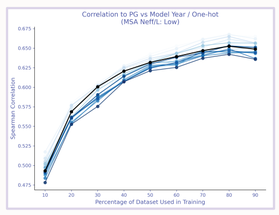


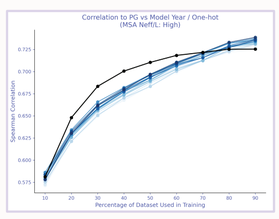




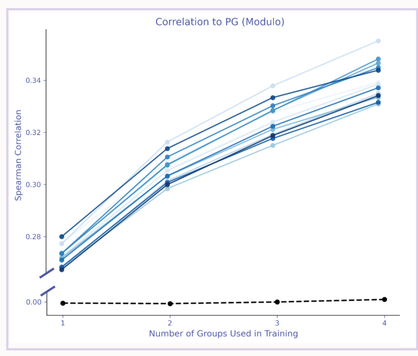
6 Including labeled data boosts performance, one-hot encodings are helpful

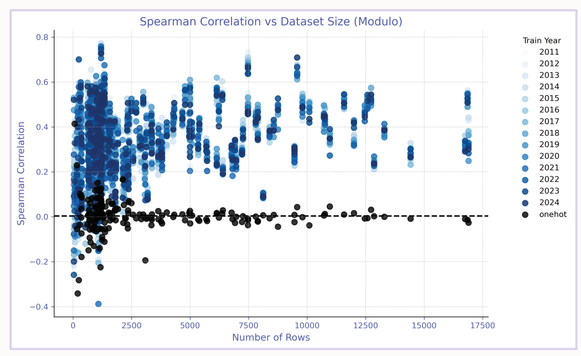




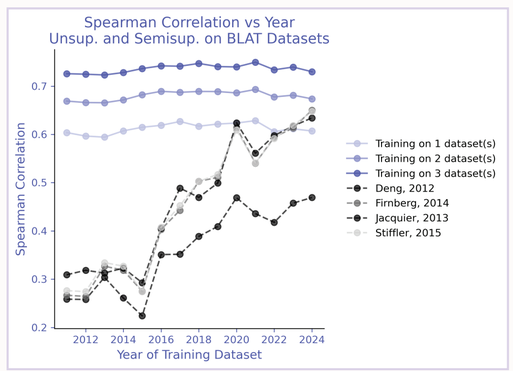


- In random splits, correlation with experimental data is strengthened by a little training data (i.e. 10% -> 0.52, 0-shot -> 0.38).
- Correlation is overall worse in position-based splits, especially for one-hot encodings





7 Deep dive on beta-lactamase suggests some emergent scaling laws



- Grey traces = zero-shot predictions
- Blue traces = train on dataset(s), test on heldout dataset(s)
- Training on any one dataset gives similar performance as a model trained on all sequences in UniRef100 from 2020
- Training on two or three datasets gives unbeatable performance

Note: these three experimental datasets are nearly identical

Future Directions

- Train a suite of models to answer this question of scaling more directly **(Want to collaborate?)**
- Semi-supervised data splits *between* datasets to better understand transfer learning between different proteins
- Expanding semi-supervised models to include ProteinNPT, Kermut, etc.

Align’s 2025 Protein Engineering Tournament

Whose protein model reigns supreme?
You generate. We test. **Science wins.**

register ›

alignbio.org