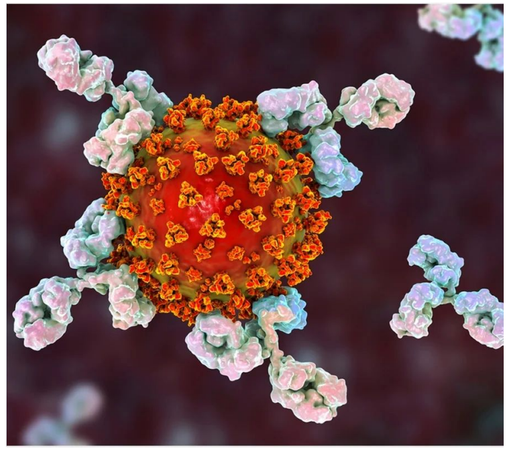
Learning inverse folding from millions of predicted structures

ICML 2022

Chloe Hsu‡, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer*, Alexander Rives*
Fundamental AI Research (FAIR) at Meta AI.
‡ University of California, Berkeley. Work performed during internship at FAIR.

Meta AI

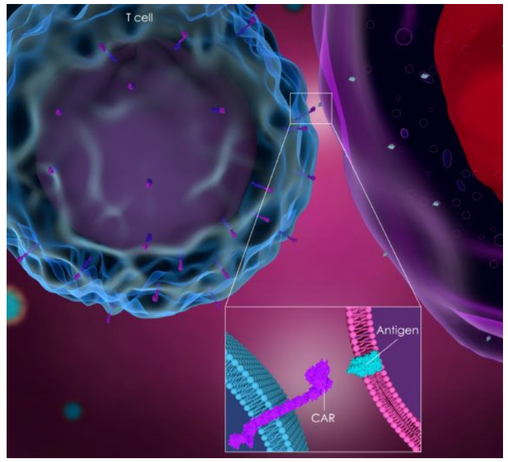
Exciting time for computational protein design and machine learning:



Antibodies and binders

Designed binders to bind with specific viruses or receptors.

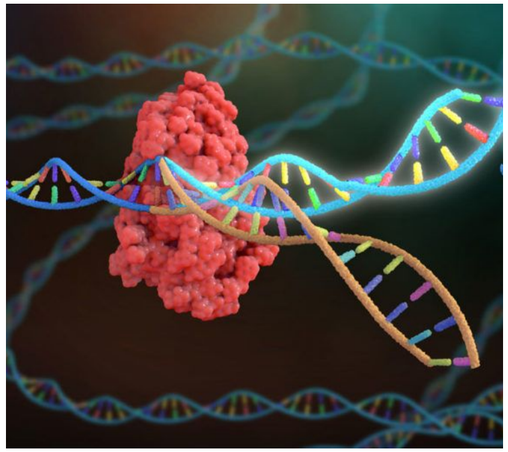
Shin, JE., Riesselman, A.J., Kollasch, A.W. et al. “Protein design and variant prediction using autoregressive generative models.” Nat Commun 12, 2403 (2021).



Cancer cell therapy

Cells with engineered receptors target and kill cancer cells.

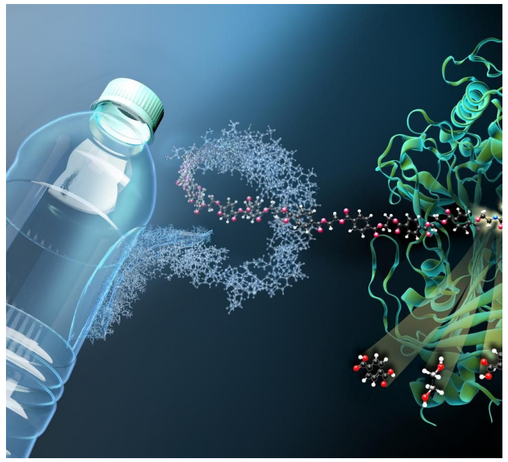
Sockolosky, Jonathan T., et al. "Selective targeting of engineered T cells using orthogonal IL-2 cytokine-receptor complexes." Science 359.6379 (2018): 1037-1042.



Gene editing

Engineered enzymes target and edit specific genetic sequences.

Thean, Dawn GL, et al. "Machine learning-coupled combinatorial mutagenesis enables resource-efficient engineering of CRISPR-Cas9 genome editor activities." Nature Communications 13.1 (2022): 1-14.



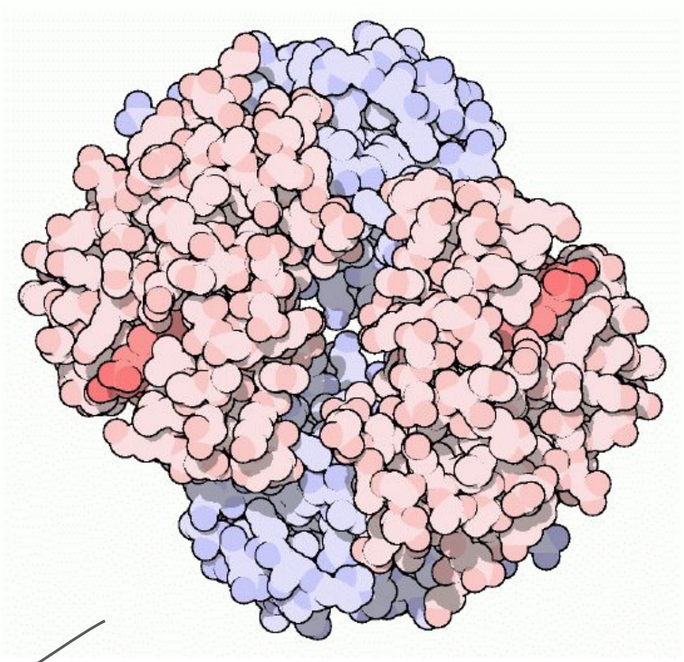
Plastic degradation

PETase enzymes eat plastic by catalyzing chemical reactions.

Lu, Hongyuan, et al. "Machine learning-aided engineering of hydrolases for PET depolymerization." Nature 604.7907 (2022): 662-667.

Folding

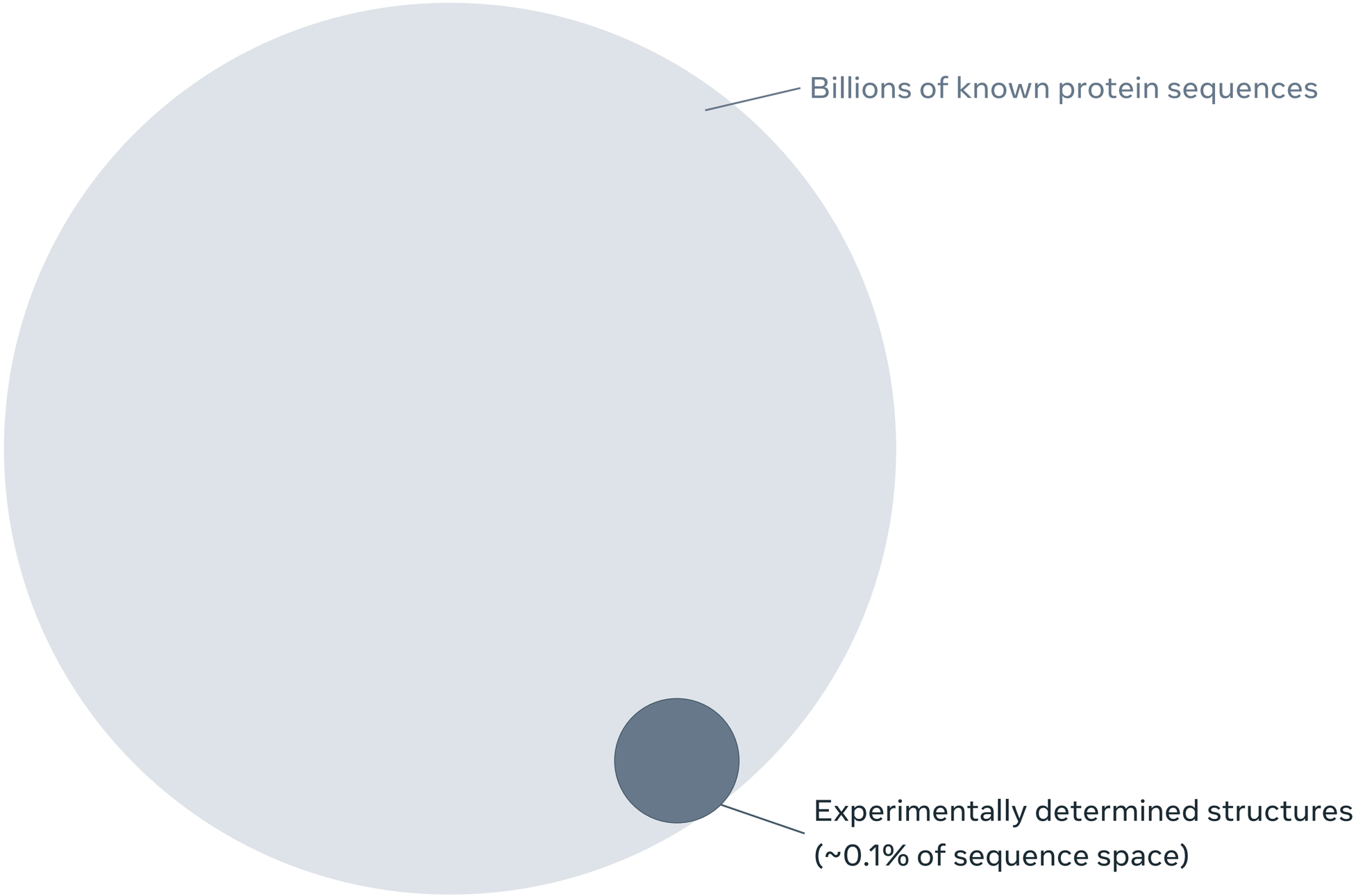
Protein 3D structure

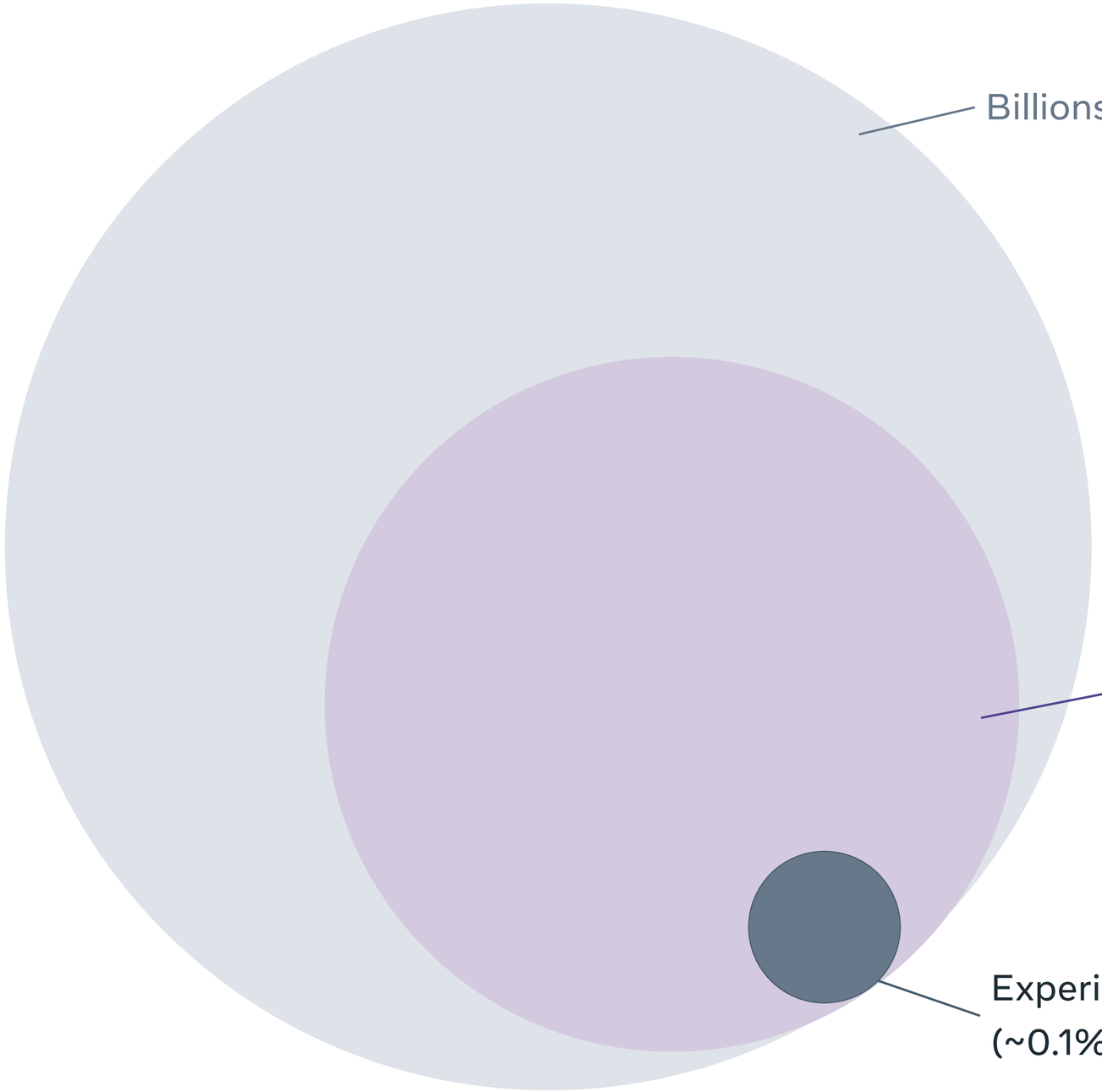


Protein amino acid sequence

MVLSPADKTNVKAAWGKVGAHAGEYGA
EALERMFLSFPTTKTYFPHFDLSHGSAQV
KGHGKKVADALTNAVAHVDDMPNALSAL
SDLHAHKLRVDPVNFKLLSHCLLVTLAAH
LPAEFTPAVHASLDKFLASVSTVLTSKYR

Inverse folding


Billions of known protein sequences
Experimentally determined structures
(~0.1% of sequence space)



Billions of known protein sequences

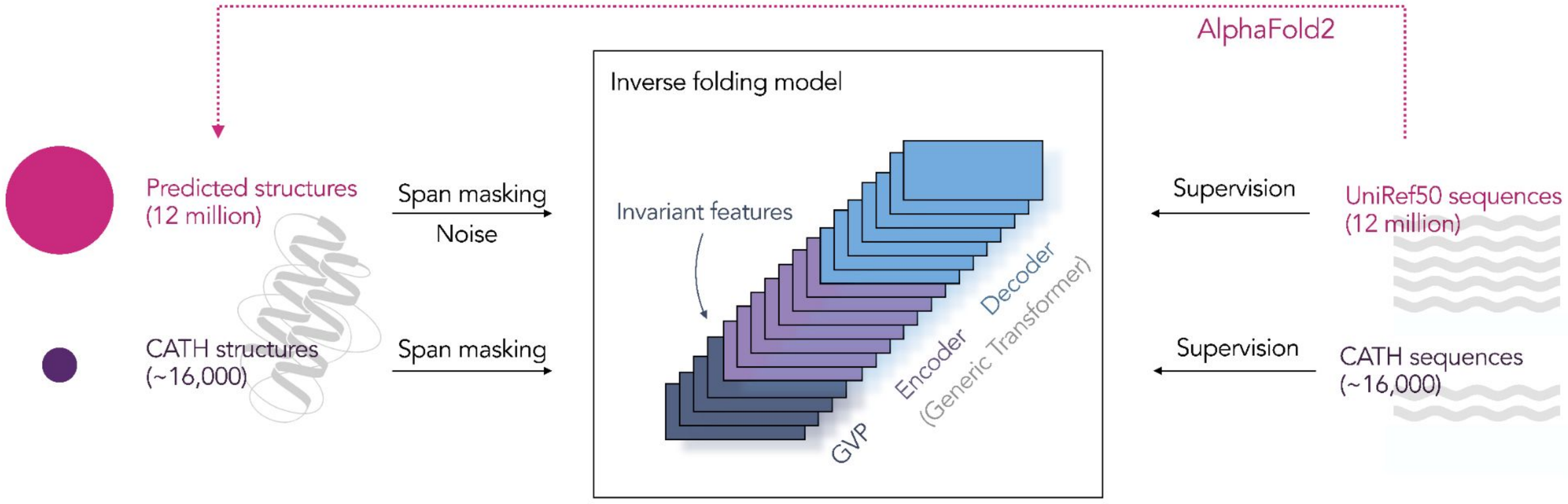
Our new data for training:

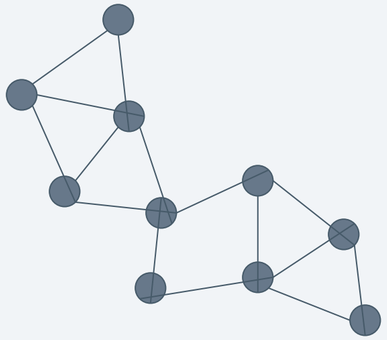
12 million predicted structures from AlphaFold2 (Jumper et al., 2021)

~3 orders of magnitudes more

Experimentally determined structures

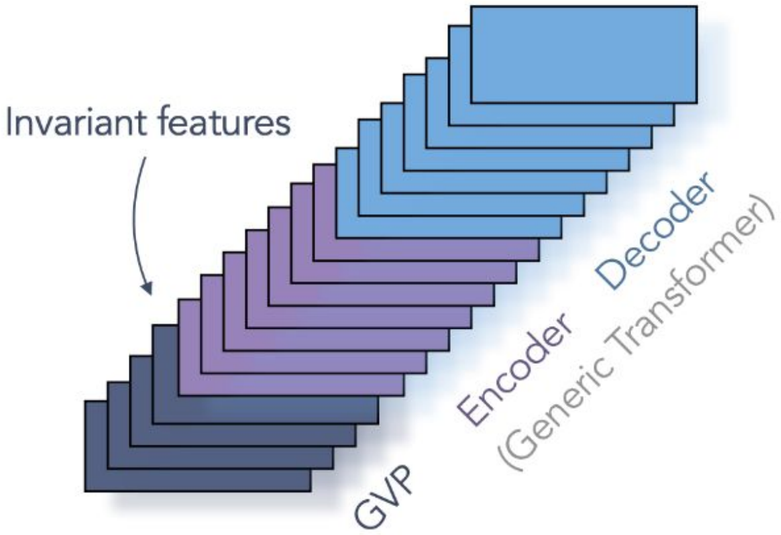
(~0.1% of sequence space)


AlphaFold2
Predicted structures
(12 million)
Span masking
Noise
CATH structures
(~16,000)
Span masking
Inverse folding model
Invariant features
GVP
Encoder
Decoder
(Generic Transformer)
Supervision
UniRef50 sequences
(12 million)
Supervision
CATH sequences
(~16,000)



GVP-GNN

A rotation-invariant graph neural network with geometric vector perceptron (GVP) layers handling both scalar and vector features on nodes and edges (Jing et al., 2021).



GVP-Transformer

A more flexible architecture with GVP-GNN encoder layers to extract geometric features, followed by a generic autoregressive encoder-decoder Transformer.

AlphaFold2-predicted structures improves inverse folding

Model	Data	Perplexity			Recovery %		
		Short	Single-chain	All	Short	Single-chain	All
Natural frequencies		18.12	18.03	17.97	9.6%	9.0%	9.5%
Structured GNN	CATH	7.91	6.48	6.49	31.5%	37.1%	37.1%
GVP-GNN	CATH	7.14	5.36	5.43	34.0%	42.7%	42.2%
	+ AlphaFold2	8.55	6.17	6.06	29.5%	38.2%	38.6%
GVP-GNN-large	CATH	7.68	6.12	6.17	32.6%	39.4%	39.2%
	+ AlphaFold2	6.11	4.09	4.08	**38.3%**	50.8%	50.8%
GVP-Transformer	CATH	8.18	6.33	6.44	31.3%	38.5%	38.3%
	+ AlphaFold2	**6.05**	**4.00**	**4.01**	38.1%	**51.5%**	**51.6%**

Prior SOTA

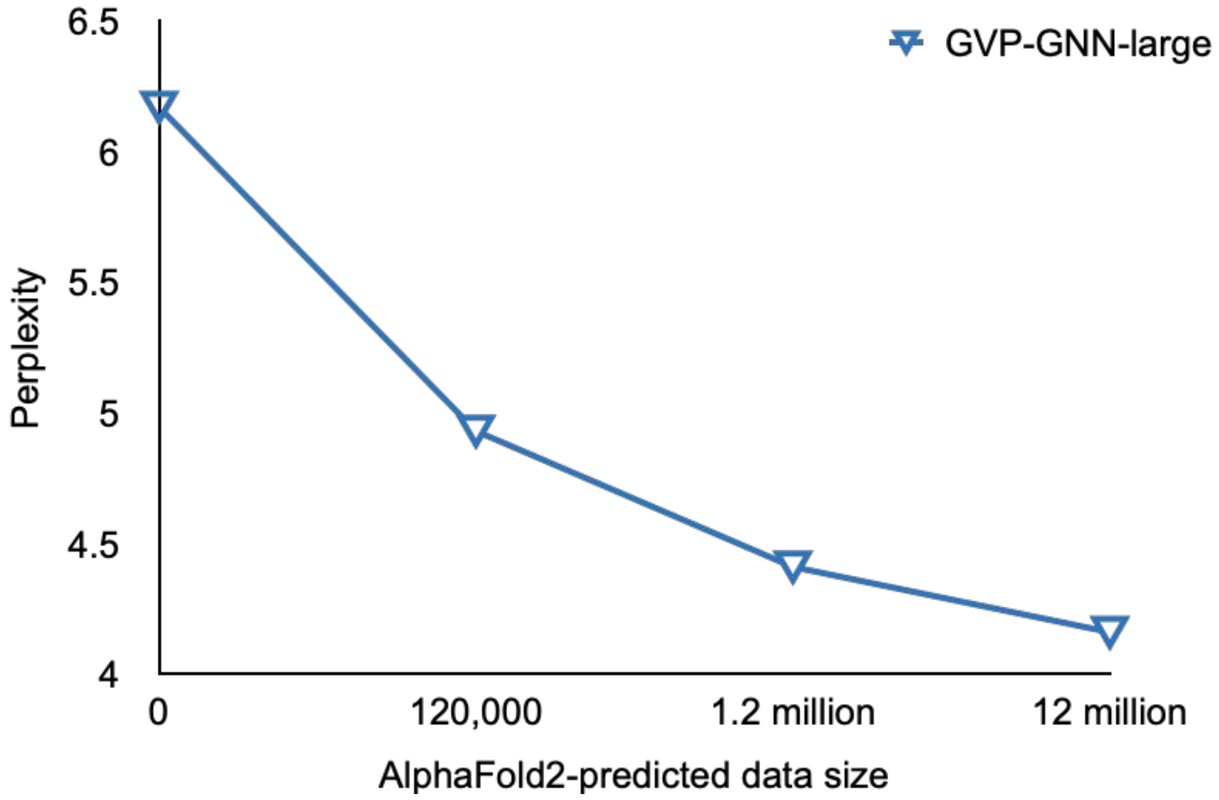
New

New

Fixed backbone sequence design evaluation on the CATH v4.3 topology split test set.

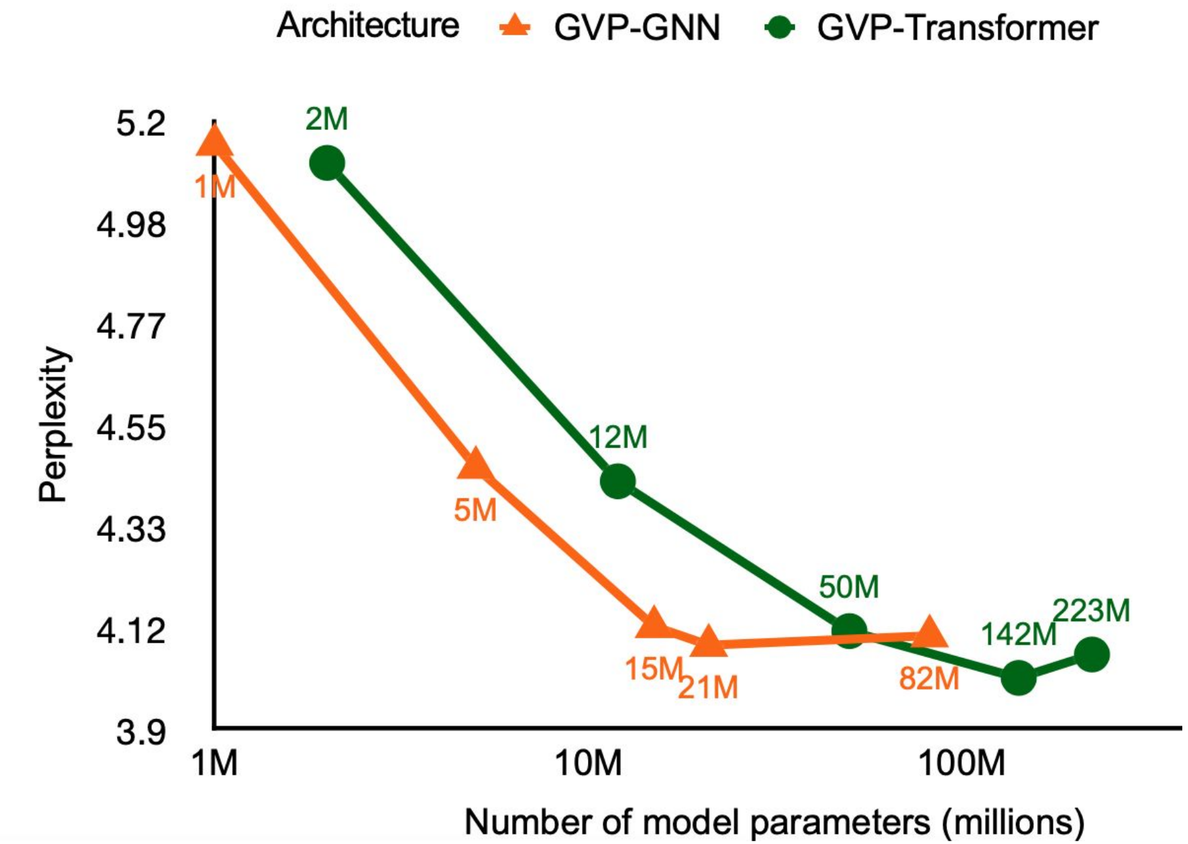
The power of data scaling

On an existing model architecture class (Jing et al., 2021), simply scaling up the training data size by 750x from under 20,000 to 12 million substantially improves the model, increasing sequence recovery rate (accuracy) from 42% to 51%.



Model scaling

While data scaling seems to keep improving the model performance up to 12 million predicted structures, model scaling hits a ceiling for a given data scale.



Wait.. Why would training on **predicted** structures help?

Wait.. Why would training on **predicted** structures help?

Hypotheses:

Wider sequence distribution

With the predicted structures, the model sees a much larger set of real UniRef50 sequences, which allows the model to better capture the output distribution.

Evidence from contemporary work:

Yang, Kevin K., Niccolò Zanichelli, and Hugh Yeh. "Masked inverse folding with sequence transfer for protein representation learning." bioRxiv (2022).

Knowledge distillation

AlphaFold2 is a more powerful model, especially with many recycling iterations. AlphaFold2 itself benefited from augmenting the training data with high confidence predicted structures for ~350,000 Uniclust30 sequences.

Co-evolutionary information

When predicting structures, AlphaFold2 makes use of co-evolutionary information in the form of multiple sequence alignments (MSAs), whereas inverse folding models do not make use of MSAs.

Wait.. Why would training on **predicted** structures help?

Hypotheses:

Wider sequence distribution

With the predicted structures, the model sees a much larger set of real UniRef50 sequences, which allows the model to better capture the output distribution.

Evidence from contemporary work:
Yang, Kevin K., Niccolò Zanichelli, and Hugh Yeh. "Masked inverse folding with sequence transfer for protein representation learning." bioRxiv (2022).

Knowledge distillation

AlphaFold2 is a more powerful model, especially with many recycling iterations. AlphaFold2 itself benefited from augmenting the training data with high confidence predicted structures for ~350,000 Uniclust30 sequences.

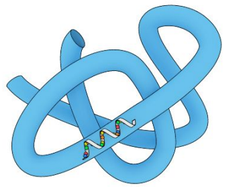
Co-evolutionary information

When predicting structures, AlphaFold2 makes use of co-evolutionary information in the form of multiple sequence alignments (MSAs), whereas inverse folding models do not make use of MSAs.

Caveat: training only on predicted structures does not work.

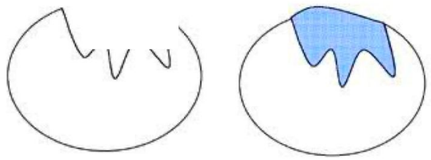
Expanding the set of structure-conditional protein design tasks:

From chain backbones



Existing benchmark for inverse folding on structurally split proteins (Ingraham et al., 2019).

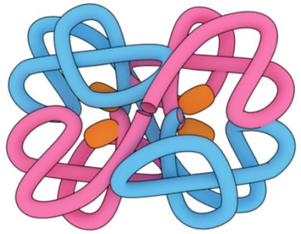
From multiple conformations



Conditioning sequence design on two conformations drives down sequence perplexity at flexible residues compared to using a single conformation.

(Structurally held-out proteins in the PDBFlex database)

From complex backbones



Performance substantially improves when given the full complex backbone coordinates as input, versus only the single chain as input.

(Complexes in the CATH4.3 topology split test set)

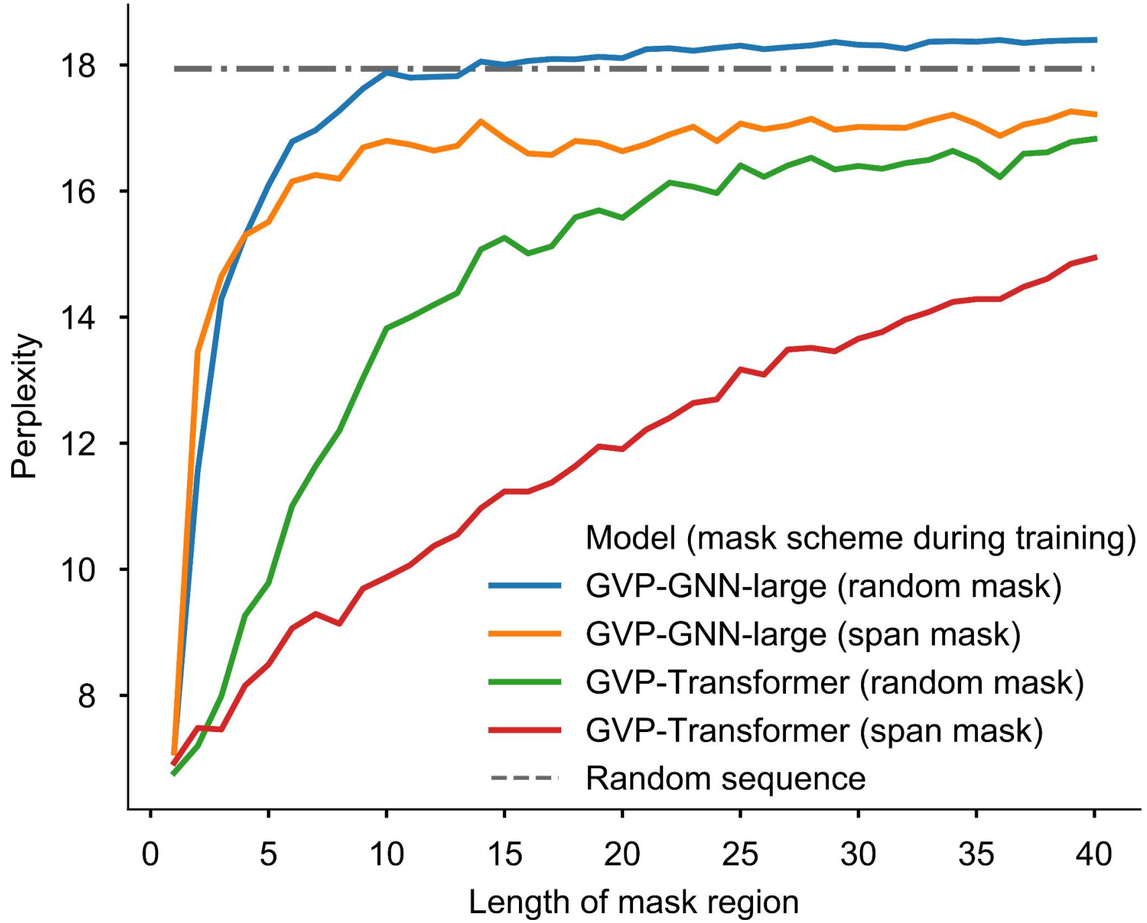
From partially masked backbones

Span masking during training improves the performance of the GVP-Transformer model on masked regions.

(Partially masked CATH4.3 topology split test set)

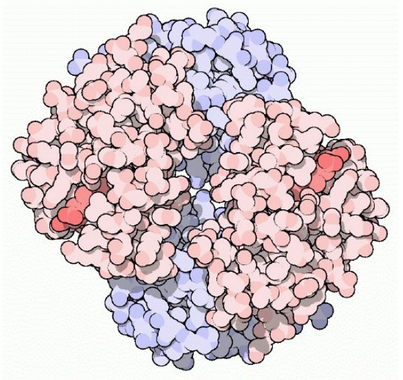
Inverse folding on partially masked structures

Both the more flexible GVP-Transformer architecture and span masking during training improve inverse folding performance on partially masked structures, although on longer mask regions all models suffer from degraded performance.



Perplexity on regions of masked coordinates of different lengths.

Protein 3D structure



P(Protein amino acid sequence

MVLSPADKTNVKAAWGKVGAHAGEYGA
EALERMFLSFPTTKTYFPHFDLSHGSAQV
KGHGKKVADALTNAVAHVDDMPNALSAL
SDLHAHKLRVDPVNFKLLSHCLLVTLAAH
LPAEFTPAVHASLDKFLASVSTVLTSKYR

|)

Zero-shot tasks

Training with predicted structures improves inverse folding model performance on the following zero-shot prediction tasks:

Complex stability (SKEMPI database)

De novo protein stability (Rocklin et al., 2017)

SARS-CoV-2 RBD binding (Starr et al., 2020)

Sequence inversion effects on AAV virus packaging (Bryant et al., 2021)

Prediction of mutational effects:

Variant sequences with higher conditional likelihoods are more likely to “fit” the given structure, e.g. implying higher stability.

Summary

Training with predicted structures improves inverse folding models.

In addition to the geometric inductive biases (which have been the major focus for existing work on inverse folding), finding ways to leverage more sources of training data is an equally important path to improved modeling capabilities.

Inverse folding as a pre-training task has diverse use cases.

Inverse folding as a general training objective enables many downstream structure-conditional design tasks.

By integrating span masking and using a sequence-to-sequence transformer, reasonable sequence predictions can be achieved for short masked spans.

GVP performs well on protein backbones at large model/data scales.

The geometric vector perceptron (Jing et al., 2021) is a scalable primitive for structural reasoning, although the GVP-GNN architecture has limitations with partially masked backbones.

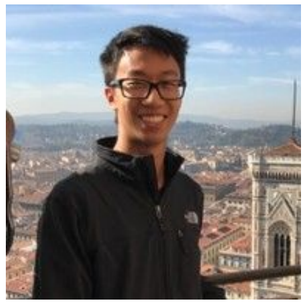
Further structural reasoning might be achieved by additional supervision during training, e.g. with structural completion or structural generation as a joint objective.

Try our open source code and Colab notebook for ESM-IF1:
https://github.com/facebookresearch/esm

Teamwork

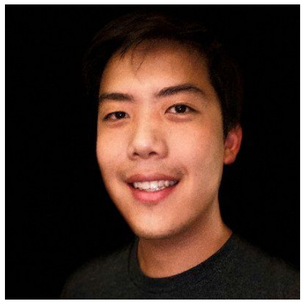
Robert Verkuil

Jason Liu



Zeming Lin

Brian Hie



Tom Sercu

Adam Lerer*

Alex Rives*